CHAPTER 1
INTRODUCTION

1.1 Basic Concepts

Today databases range in size that extends into terabytes. Within this mass of data lies hidden information of strategic importance. Then how do we draw meaningful details of importance from such data sets? The answer is data-mining, which is being used both to increase revenues and to reduce costs. The potential returns are enormous.

Data-mining refers to the task of discovering interesting patterns/rules/results from large amount of data stored in databases, data warehouses or in other information repositories. Data-mining takes advantage of the advances in the fields of artificial intelligence (AI) and statistics. Both disciplines are concerned with problems of pattern recognition and classification. Data-mining is a tool for increasing the productivity of people trying to build predictive models.

Data-mining offers pragmatic values across a broad spectrum of industries. Retail Industries perform data-mining on transactional database for their business promotions. Telecommunications and credit card companies are leaders in applying data-mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraudulent activities.

Medical applications are other fruitful area of data-mining, which can be used to predict the effectiveness of surgical procedures, medical tests or medications used. Pharmaceutical firms are mining large databases of
chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease.

Data-mining is just a tool. It does not eliminate the need to know your business, to understand your data, or to understand the analytical methods. Data-mining does not replace skilled business analysts or managers, but rather gives them a powerful new tool to improve the job they are doing. Various data-mining functionalities are:

- Characterization and Discrimination of Data
- Association Analysis
- Classification and Predication
- Cluster Analysis
- Outlier Analysis
- Trend and Evolution Analysis

When data-mining is performed, researchers or data miners are interested only in the pattern or general characteristics about the entire data set, not about the individual associated values. But the release of data for analysis sometimes leads to problems because of misuse. So, the data owners are usually unwilling to release their data. The results/rules/patterns of mining are based on the input data. If the given data is not a true data or clean data then predictions done are invalid.

Eventhough the focus on data-mining technology is on the discovery of general patterns (not on any specific information regarding individuals) some data-mining applications require access to individual’s records having sensitive privacy data. Data containing structured information on individuals is referred to as micro-data. Abundance of recorded, personal information available in electronic form coupled with increasingly powerful
data-mining tools prevailing, poses a threat to privacy and data security. The prime objective of this research is to find a tangible solution to this problem.

For example, a medical database may contain attributes such as name, social security number, address, age, gender, ethnicity, and medical history of each patient. It is desirable to provide aggregate knowledge about such databases for research or for other applications. If medical researchers have some form of access to this database, they can learn relationship between age (and/or ethnicity) and the risk of different diseases. Similarly by discovering the occurrence of communicable diseases being endemic in certain areas, they can predict the outbreak of an epidemic and thereby prevent it from spreading to other areas.

Governmental, public, and private institutions that systematically release data are increasingly concerned with possible misuses of their data that might lead to disclosure of confidential information. Confidentiality regulation requires that privacy of individuals represented in the released data must be protected. In the U.S., for example, privacy regulations promulgated by the Department of Health and Human Services as part of the Health Insurance Portability and Accountability Act (HIPAA) that came into effect in April 2003 in order to protect the confidentiality of electronic healthcare information.

Hence, the dissemination of data containing personal information is restricted in order to protect individual privacy. After micro data is masked to limit the possibility of disclosure and released for use by third parties, it is called masked or publishable micro data.

Perfect privacy can be achieved by refusing to publish any form of the data or to answer any queries about the data, but such data will not be
available for any useful research towards discoveries. Perfect utility can be achieved by publishing exact data or by answering all queries about the data exactly; but, this in turn may lead to legal issues, violating individual’s rights and privacy acts. There exists always a trade off between privacy and utility of data. Thus, there is a high need and niche for a method that preserves privacy while maintaining utility. This research work aims at developing a method and a tool to implement a strategy that helps in preserving privacy, while maintaining information and utility.

The initial research efforts in Privacy Preserving Data-mining refer to extending traditional data-mining techniques with the data modified to mask sensitive information. The key issues there of are how to modify the data and how to recover the data-mining result from the modified data. The solutions obtained are often tightly coupled with the relevant data-mining algorithms.

Two most important questions related to the release of actual data and the privacy of the individuals are: (i) Can general trends across individuals be determined without revealing information about individuals? And (ii) can highly private associations be extracted from the public data? In the former case, there is a need to protect individual data values while revealing associations or aggregation. In the latter case, there is a need to protect the associations and correlations between the data.

In this research work, transformation methods are proposed for preserving the privacy of the individuals. Two different types of transformation methods are proposed for two distinct data types. It is proposed to follow a categorical grading based transformation for numerical sensitive data and mapping-table based transformation for alphanumeric nominal sensitive data. It is proposed to develop a tool that performs the categorical grading and mapping-table based transformations on the micro data table. The transformed or publishable table can be released for research purpose without
any information loss. Any data-mining algorithm can be applied on the released table without any modification in the algorithm. And the results hence obtained are, as if mined from the original micro data table. Yet the privacy of the individual is preserved across the released table as well as in the mining results obtained.

Also, in real-life data publishing, a single organization often may not hold the complete data. Organizations need to share data for publishing to a third party for analysis. Data-mining performed on this kind of data is called **collaborative data-mining**. Mining this kind of collaborative data should preserve privacy of individual organization without disclosing sensitive information to other organizations involved in the collaboration. The proposed methods for micro-data release can be used for collaborative data mining also. Two different frame works are proposed to preserve the privacy of the horizontally and vertically partitioned data.

### 1.2 Related Works

Micro data in its original form often contains sensitive information about individuals; and, publishing such data as it is would violate individual privacy. The current practice primarily relies on policies and guidelines to restrict the types of publishable data, and agreements on the use and storage of sensitive data. The limitation of this approach is that it either distorts data excessively or requires a trust level that is impractically high in many data-sharing scenarios. For example, contracts and agreements cannot guarantee that sensitive data will not be carelessly misplaced and may end up in wrong hands.

Dalenius and Reiss [1] first proposed data swapping as a method to preserve confidentiality in data sets that contain alphanumeric nominal
attributes. The basic idea behind the method is to transform a database by exchanging values of sensitive attributes among individual records. Records are exchanged in such a way to maintain lower-order frequency counts or marginals. Such a transformation protects both confidentiality by introducing uncertainty about sensitive data values and maintains statistical inferences by preserving certain summary statistics of the data.

Method suggested in [1] maintains statistical properties but this kind of data swapping will result in inaccurate results for any mining tasks performed. Since this technique preserves the confidential information by introducing uncertainty about the sensitive values, it gives a wrong relationship among the other attributes. For example a person having the designation as manager will be related to the salary lower than a person designated as clerk. Hence, corresponding mining task performed would yield a wrong pattern.

The term, privacy-preserving data mining (PPDM) emerged in 2000 [2]. In this work, the authors have considered the concrete case of building a decision-tree classifier from training data in which the values of individual records have been perturbed. The resulting data records look very different from original records and the distribution of data values is also quite different from the original distribution. This technique has the limitation because of the fact that it is not possible to accurately estimate original values in individual data records after the perturbation.

The perturbation can be done on the input data or on the query output answers, but accordingly mining algorithms need to be modified. For example, when the input to the data is perturbed, it is a challenging problem to accurately determine the association rules on the perturbed data.
Any Privacy Preserving Data-mining (PPDM) technique can be classified into one of the two major frameworks, namely, interactive and non-interactive. In the interactive framework, the user (researcher) queries the database through a privacy mechanism, which may deny the query or alter the answer to the query in order to ensure privacy.

1.2.1 Interactive Techniques

In the interactive query model, the user can submit a sequence of queries based on previously received query results. Although this query model could improve the satisfaction on data recipients’ information needs, the dynamic nature of queries makes the returned results even more vulnerable to attacks.

The interactive privacy mechanism can be classified as query auditing and output perturbation techniques. In query auditing [3-5], a query is denied if the response could reveal sensitive information otherwise answered exactly. In output perturbation [5-9], the privacy mechanism computes the exact answer to the query and then outputs a perturbed result (for example, by adding noise) as the response to the query.

The results obtained in the interactive framework are expected to be of better quality since only queries of interest to the user are answered. But, query auditing and output perturbation methods are not preferred, whenever the underlying data mining task is inherently ad hoc. Further, for all interactive methods, collusion and denial of service are problems of concern. Also, there is an implicit assumption that all users can collude with each other and hence queries from all users are treated as if coming from a single user. Consequently, any one user has reduced utility.
While query auditing methods maintain consistency (i.e., if the same query is posed again, one gets the same answer), output perturbation does not. The query auditing method is useful in situations where exact answers to queries are necessary. In output perturbation schemes, privacy is obtained by perturbing the true answer to a database query by adding a small amount of Gaussian or exponentially distributed random noise. But, when the aggregate queries of interest are not known in advance, techniques such as query auditing, output perturbation may not provide an adequate solution.

1.2.2 Non-interactive Techniques

There is a need to release an anonymized view of the database that would enable the computation of non-sensitive query aggregates, perhaps with some acceptable error or uncertainty. Techniques under non-interactive framework such as input perturbation may not be suitable if one wants to draw inferences with 100% confidence.

Most methods for privacy computations use some form of transformation on the data in order to perform the privacy preservation. Typically, such methods reduce the granularity of representation in order to preserve the privacy. This reduction in granularity results in some loss of effectiveness of data management or mining algorithms. This is the natural trade-off between information gain and privacy.

The randomization method is a technique for privacy preserving data-mining in which noise is added to the data in order to mask the attribute values of records [2, 10, 11]. Therefore, techniques are designed to derive aggregate distributions from the perturbed records. The additive noise approach is used often because it is easy to implement and effective. There has been conjecture that, instead of adding noise, multiplying noise might better protect the confidentiality. Two forms of multiplicative noise are considered in
If the noise added, which may be either additive or multiplicative noise, is sufficiently large then individual record values cannot be recovered. But, smaller perturbation (or noise addition) always results in a strong violation of privacy and larger perturbation will affect the accuracy of mining results.

The other non-interactive technique is to anonymize the data and then publish it for mining. Relevant task of current interest is to develop methods and tools for publishing data in a more hostile environment so that the published data remains practically useful while individual privacy is preserved. This procedure is called privacy-preserving data publishing (PPDP).

PPDP may not necessarily tie to a specific data-mining task, and the data-mining task is sometimes unknown at the time of data publishing, in contrast to PPDM. Furthermore, some PPDP solutions emphasize on preserving truthfulness of the data at the record level; but, PPDM solutions often do not preserve such property. In this kind of non-interactive framework, the original database is first sanitized so as to preserve privacy and then the modified version is released to the public.

A Privacy Preserving Data Publishing approach that attempts to hide the identity and/or the sensitive data of record owners is referred to as Anonymization.

Typical data collection and data publication involve three different categories of people, namely, data owners, data publisher and data recipients. In the data collection phase, the data publishers collect the data from the data owners and in the data publication phase the data publisher publishes the data to the data recipients. For example, if a hospital publishes its patients’ data,
then the patients are the data owners, hospital is the data publisher and persons analyzing the data such as data miners, (data) researchers are data recipients.

The data publisher may be an expert or non-expert in data-mining. Even if the data publisher is an expert, it can be assumed that the publisher does not perform any mining operation on the data instead gives data to the data miner for analysis. In this scenario, since the publisher can judge what kind operation/analysis is to be performed by the data miner; more customized data that preserves pattern will be issued to miner. In short, data-mining tasks are predefined.

The non-expert data publishers also can be grouped into two types: (i) those interested and (ii) those non-interested in mining results. In this research the non-expert data publisher interested in mining results is considered. One more assumption made in the privacy preserving data publishing (PPDP) is that the data recipient may be an attacker; that is, an attacker is a person keen in knowing sensitive information about the individuals who are unwilling to disclose their information.

A majority of works as regard to disclosure limitation [16] focus on applying statistical disclosure limitation methods for micro-data sets. All such statistical disclosure limitation methods use techniques like cell suppression, data swapping, rounding, sampling, and generation of synthetic data as means of achieving statistical disclosure control. In limiting the disclosure, they may perturb the characteristics of the original dataset.

In the most basic form of PPDP, the data publisher has a table T of the form, \( T = \{t_1, t_2, t_3, \ldots t_n\} \). Each \( t_i \) is a tuple of attribute values representing some individual record. Let \( A = \{a_1, a_2, \ldots a_m\} \) be a set of attribute in T and \( t[a_i] \) represents the value of attribute \( a_i \) for tuple \( t \). The attribute set \( A \) can be
classified into four categories: Identifying Attributes $A^i$, Sensitive Attributes $A^s$, Quasi Identifying attributes $A^q$ and Neutral Attributes $A^n$.

Anonymization assumes that sensitive data must be retained for data analysis. Clearly, explicit identifiers of record owners must be removed. The identifying attribute such as Name, Social Security Number (SSN), which can uniquely identify an individual record in a table denotes identifying attribute ($A^i$).

An attribute is called Sensitive ($A^s$) if the individual (data owner) is not willing to disclose or an adversary must not be allowed to discover the value of that attribute. For example, the attributes Income, Disease etc are sensitive attributes. A sensitive attribute can be numerical data type ($A^{ns}$) or alphanumeric nominal data type ($A^{ns}$).

A set of attributes such as {Age, Gender and Zipcode}, which can identify an individual record in the table is called quasi-identifying attributes ($A^q$). The neutral attributes are neither sensitive nor quasi-identifying attributes. For example, the attribute Length-of-Stay in a patient’s data set is a neutral attribute.

1.2.3 Attacks on Released Data

Since, it is necessary to publish the micro-data for research purposes, any adversary who is a person interested in knowing others’ (Victim’s) sensitive data can easily get the sensitive information from the published table. This problem occurs by identity disclosure and is solved by not publishing the identifying attribute such as SSN (Social Security Number), Name etc.
The identity disclosure leads to the problem of (sensitive) attribute disclosure, since each tuple corresponds to an individual. Even with all explicit identifiers being removed (e.g. attributes like Name, Social Security Number), Sweeney [17] showed a real-life privacy threat on William Weld, former governor of the State of Massachusetts. In Sweeney’s example, an individual’s name in a public voter list was linked with his record in a published medical database through the combination of zip code, date of birth, and sex called Quasi-Identifier (QI) attributes. The association of quasi-identifiers with sensitive attributes in public records is known as sensitive attribute disclosure. Such sensitive attribute disclosure occurs by linking QI attributes with external data and hence this type of attack is called linking attack.

It is very easy to prevent sensitive attribute disclosure by simply not publishing Quasi-Identifiers and sensitive attributes together. But, the only reason to publish generalized Quasi-Identifiers and sensitive attributes together is to support data-mining tasks that consider both types of attributes in the sanitized database. Neutral attributes are neither sensitive nor quasi-identifiers for e.g attribute like Length-of-Stay and do not have much role to play.

To solve the problem of linking attack, K-anonymous micro-data table is published. The K-anonymity requirement is typically enforced through generalization, where real values of QI attributes are replaced with less specific but semantically consistent values [17]. The set of records or rows with the same values in their respective attributes is called equivalence class or QI group.

For example, if attribute age is one of QI attribute then the records with the values 21, 22, 25 respectively are grouped into one category and the
actual values are replaced with a single range as (21-25). Another member of QI attribute like gender is replaced with taxonomy structure i.e. values Male and Female are replaced with the value Person. Thus all the QI attributes are generalized and hence within each group (consists of ‘K’ tuples), values of all QI attributes are the same. A table $T$ is said $K$-anonymous, given a parameter ‘$K$’ and the quasi-identifier $QI = (A_{q1}, \ldots, A_{qL})$ if for each tuple $t \in T$, there exist at least another ($K-1$) tuples $t_1, \ldots, t_{k-1}$ such that those $K$ tuples have the same projection on the quasi-identifier. Tuple $t$ and all other tuples indistinguishable from $t$ on the quasi-identifier form an equivalence class.

Whenever an adversary is trying to find the sensitive values from the $K$-anonymous table, by linking with external data gets $K$ tuples as a result. Hence he/she cannot get the actual sensitive value with 100% confidence. Larger the value of $K$, greater is the implied privacy since no individual can be identified with a probability exceeding $1/K$ through linking attack. The process of $K$-anonymization sometimes involves cell value suppression. Suppression is the process of deleting cell values or the entire tuples. Different generalization techniques using taxonomy structure were also used. The major advantage is that, all the records within $K$-anonymized data set remain truthful. Numerous algorithms have been proposed to obtain $K$-anonymous table [17-23].

$K$-anonymity has several drawbacks. First drawback is that, a $K$-anonymous table may allow an adversary to derive sensitive information of an individual with 100% confidence, since $K$-anonymity requirement does not put any restriction on the sensitive attribute. This is because $K$-anonymity only prevents association between individuals and the tuples instead of association between individuals and sensitive values.
Second limitation is that, K-anonymous table may lose considerable information, when the generalization is applied to obtain K-tuples in an equivalence group. For example the values of attribute age 21, 23, 25 and 60 are generalized to a group as (20-60). An estimate for the number of patients above age 30, assuming a uniform age distribution, the value can be calculated as, 
\[
\frac{60 - 30}{60 - 20} = 3.
\]
But, this value is significantly deviates from the actual result 1. Thus, there is a loss of information.

Third limitation is that, K-anonymity does not take into account personal anonymity requirements. For example, a patient affected with flu may not mind releasing his actual disease. But, since such preference variations cannot be considered in K-anonymity he/she will be grouped with the persons having other diseases. There is a chance that the person with flu may be linked with Cancer, HIV etc. which the person may not prefer.

Also, once the anonymized table is released, further grouping is not possible on the anonymized data. For example, in a table the attribute age is anonymized by grouping the values as (35-45), (46-60), (61-70) etc. But the researcher who wants to analyse the problems of female in the age group of (40-50) cannot perform the analysis with the real data. He/she has to consider both the groups (35-45) and (46-60). And assuming uniform distribution of data among the group he/she can analyse data.

Also, Aggarwal [24] showed that it is not possible to create even a 2-anonymous table in a high-dimensional space without considerable information loss. Also, only when, the sensitive values in an equivalence class are different, the generalization done on QI attributes is said to be useful. But, depending on the data published, sometimes all the rows in an equivalence class may contain the same sensitive value. In such situation an adversary gets
the actual sensitive value of the victim with 100% confidence, even after generalization of QI attributes. This problem is known as homogeneity attack.

Machanavajhala et al. [25] gave a solution to homogeneity attack by proposing the principle of L-diversity in K-anonymous table. A QI group is L-diverse, if it contains at least L-well represented values for sensitive attribute. The adversary gets the actual value with only (1/L) confidence. The L-diversity principle is improvised in [26-28].

In general, stronger privacy protection is ensured, when a larger K or L is deployed. The notion of “well-represented value” can be interpreted in several ways [27]. A more robust diversity is achieved by enforcing entropy L-diversity, which requires every equivalence class to satisfy the condition [25] (in shannon sense), given in equation 1.1)

\[- \sum_{s \in S} p(e, s) \log p(e, s) > \log L \quad \ldots \quad \text{Eq. (1.1)}\]

where S is the domain of the sensitive attribute and p(e, s) represents the fraction of records in an equivalence class ‘e’ that has sensitive value ‘s’.

Although entropy L-diversity does provide stronger privacy, the requirement may sometimes be too restrictive. For instance, in order for entropy L-diversity to be achievable, the entropy of the entire table must also be greater than or equal to log L. The limitation of entropy L-diversity is that it does not provide a probability-based risk measure, which tends to be more intuitive to the human data publisher. Also L-diversity has the limitation of implicitly assuming that each sensitive attribute takes values uniformly over its domain, that is, the frequencies of the various values of a confidential attribute are similar. When this is not the case, achieving L-diversity may cause a large data utility loss.
For example, a table is released having 100 records with the sensitive attribute ‘disease’. It may have 80 records with flu, 15 records with respiratory problem and 5 records with cancer as disease. That is, the diseases flu, respiratory problem and cancer have the frequencies 0.8, 0.15 and 0.05 respectively. So, the impracticable assumption of equal probability of occurrence may lead to high information loss (because of generalization of QI and sensitive attributes) and inaccurate mining results, if failed.

Also, the diversified sensitive values in a particular group need not be diversified semantically. For example the different sensitive values such as flu, viral fever, fever are the same, semantically. That is the diversification is not checked semantically. So, the adversary gets almost the actual value of the victim individual. This problem is called as **semantic attack**.

Sometimes the background knowledge of the adversary about the victim helps him in getting the actual sensitive value of the victim, even when the victim’s record is in a group with many different sensitive values. For example, the adversary knows the group to which the victim belongs. Suppose the sensitive values in that group are flu, diarrhea and heart attack. Since the adversary has the background knowledge about the regular visit of victim to the hospital, he/she can conclude that the victim is a heart patient. This kind of attack is called **background knowledge attack**.

Li et al. [28] observed that L-diversity is neither sufficient nor necessary for protecting against attribute disclosure. They have proposed t-closeness as a stronger anonymization model, which requires the distribution of sensitive values in each equivalence class to be analogous to the distribution of the entire dataset.
The solution was given by considering only a single sensitive attribute. Also, the relationship between the value of $t$ and information is not clear. Since it is not possible to identify the optimal $t$, over generalization may happen which leads to high information loss.

The choice of the generalization and diversity principles depends on the needs of underlying application. But all of the above anonymization methods focus on universal approach that exerts the same amount of preservation for all persons without catering for their concrete needs. A new generalization framework based on concept of personalized privacy, which can maintain large amount of information from micro-data was presented in [29].

To reduce the loss of information, instead of global generalization on the entire table, local generalization can be performed after receiving preference/concern from the individual. In the method of personalized privacy preservation [29], the individual can decide the disclosure level of his/her sensitive value by choosing the level from the taxonomy tree. For example, a person with flu may not mind disclosing the actual value. So such a person’s record need not be generalized considering both the QI and sensitive attribute, instead, generalization can be on sensitive attribute alone or on QI alone.

But the sensitive attribute generalization introduced by this personalized method, even though retains more information on QI attribute results in less precise value on sensitive attribute. Also, sensitive attribute generalization is NP hard [29].

To increase the efficiency of anonymization method or to reduce the information loss, task-based anonymization principles are adopted such as for association rule mining [30-32], for clustering [33] and for classification
tasks [20, 21]. In most cases, data owners do not know the ultimate use of the released tables. Therefore, anonymization goal should not be associated with a specific data-mining task, but should minimize distortions in the released table.

Another issue is proximity breach, a natural privacy threat occurs in K-anonymous, L-diverse table published, if the sensitive attribute is numerical. **Proximity Breach** is a privacy threat specific to numerical sensitive attribute. It occurs when an adversary concludes with high confidence that the sensitive value of individual must fall in a short interval even though the adversary may have less confidence about his/her actual value.

This problem of proximity breach is handled by (e-m) anonymity method [34] which demands that, given a QI group G, for every sensitive value \( x \) in G, at most \( 1/m \) of the tuples in G can have sensitive values similar to \( x \) where the similarity is controlled by \( e \). The (e, m) method concentrated on micro data that contains only a single numerical sensitive attribute and assumed one time publication of a static data set.

Considering the nominal or categorical data type sensitive attribute, a new privacy breach is identified by this research work, in the released table, generated by K-anonymity, L-diversity principles and their improved versions. And this breach is termed as divergence breach. For example, in a released patient table, a patient with stomach cancer may be unwilling to disclose his actual disease but willing to disclose his disease as ‘stomach related disease’. He can do that by choosing the corresponding node from the taxonomy tree. So, his record should be generalized based on QI attributes. Assume that his record is grouped with other five records with values such as respiratory
disease, bronchitis, gastritis, flu and throat cancer and hence the privacy of the person with stomach cancer is preserved.

Even though a person with flu in the same group linked with Flu with the least probability (1/6), he will be linked with other irrelevant values with a high probability (5/6). That is, he will be linked even with the throat cancer with the probability of 1/6. This situation is more dangerous than revealing his actual disease flu. This is termed as **divergence breach**.

By K-anonymity and L-diversity methods, the individuals (records) are anonymized to be the one in a QI group, without asking for their willingness. They are forced to be in a particular group, based on the principle that less generalization/suppression to be performed. And so the records with close QI values are grouped. But the individuals may dislike the QI group to which they belong to, based the sensitive values available in the group. The new transformation based approach proposed does not do any kind of grouping, thereby the individuals will not have any dissatisfaction.

K-anonymity[17,18] and different methods of generating K-anonymous table [19-24], L-diversity[25] and its improved versions [26,27] deal with nominal data type sensitive attribute within the equivalence class (QI group).

The methods such as T-closeness [28], local recoding [35], variance control [36], (c,k) safety [37] deal with numerical sensitive attributes within the equivalence class. The choice of the principles such as local recoding, T-closeness, (c-k) safety etc. depends on the needs of underlying application. But, if the application or task for which the database required is not known in advance, all these above said methods lead to high information loss or privacy
loss. Also, various works in the literature have shown that finding the optimal anonymization is NP-hard [18, 38-40].

The proposed transformation based approach deals directly with the sensitive attributes of the original table without disturbing any other quasi identifying attributes. Hence maintains information for mining while preventing privacy violations.

1.2.4 Privacy of Collaborative Data

Data sets for analysis may be in a centralized server or in a distributed environment. In a distributed environment, the data may be horizontally or vertically partitioned. Several organizations own different sets of attributes on the same set of records (vertically partitioned) or own same sets of attributes on the different set of records (horizontally partitioned) and want to publish the integrated data for analysis. This scenario is called collaborative data publishing.

Organizations need to share data for publishing to a third party. For example, two credit card companies want to integrate their customer data for developing a fraud detection system, or for publishing to a bank. However, the credit card companies do not want to indiscriminately disclose their data to each other or to the bank for reasons such as privacy protection and business competitiveness. Such kind of data with same attributes at different sites is called as horizontally partitioned data.

If the researcher wants to find the association between the students’ character and their parents’ occupation or between medical diagnosis and attendance performance, the different databases like, academic, medical, personal data of the same set of students are to be combined for analysis and
such a kind of data set with a single join key (e.g., student id) is called vertically partitioned data.

The problem of preserving privacy of the distributed data (both horizontally and vertically partitioned data) overlaps closely with a field of cryptography. The techniques used to preserve privacy of the collaborative data are called secure multi-party computation [41 - 45].

Secure computation involves several parties with private inputs that wish to compute a function of their joint inputs, and require that the process of computing the function does not reveal to an adversarial party any information that cannot be computed using the input of the adversary and the output of the function.

It is proposed to find a solution to avoid the communication overhead of distributed protocol and to overcome the limitations of cryptographic approaches.

1.3 Motivation

The previous works on privacy preserving data publishing mainly focused on anonymization of QI attributes [17,18] by techniques such as generalization and suppression. Even though the main purpose of anonymization is not to disclose any sensitive information, Sweeney[17] and Samarati[18] did not put any restriction on sensitive attribute. K anonymity has several drawbacks:

- K-anonymity only prevents association between the individuals instead of association between the individuals and their sensitive values.
- K-anonymous table may lose considerable information
- K-anonymity does not take into account personal anonymity requirement.
- The researcher cannot do grouping of his/her choice from the given anonymized table.
- K-anonymous table is prone to homogeneity attack.

To overcome the homogeneity attack, L-diversity principle [25] was suggested. But, it is based on impracticable assumption of equal frequency for each sensitive value. Also, L-diverse table is prone to semantic attack, background knowledge attack and divergence breach. The methods such as t-closeness [28], local recoding [35], variance control [36], (c,k) safety [37] are prone to proximity breach.

All of the aforesaid anonymization techniques follow the basic principle of generalization of QI attributes. But there is no standard procedure available to find out the QI attributes. Unless and otherwise having knowledge about the type of data for which the adversary has access, it is very difficult to find the QI attributes. So far in the past research methods of identifying QI attributes are not discussed. If the QI attributes are not identified correctly, the entire process of anonymization becomes useless and results in leakage of privacy, information loss and utility loss. Also if the members of QI attributes are more, generalization level increases to achieve the expected privacy level which results in high information loss.

If anonymization is too strict, information loss is more but with the increased privacy; Otherwise, information can be preserved at a loss of required privacy. A table is optimally anonymous if it satisfies the given privacy requirement and contains most information according to the chosen information metric among all satisfying tables. But obtaining optimum
anonymization is NP hard problem. And all the anonymization techniques in literature assumed a single sensitive attribute which is impractical.

To perform the anonymization, to the required level, on the required attributes the data publisher should be an expert. Also to improve the accuracy of mining results, the data publisher is expected to do anonymization based on the task or application to be performed. When the task or application based anonymization is performed on the data and given to the miner, then accurate results maintaining privacy will be achieved. But if the miner uses the same data set for some other mining task, results may not be correct and there will be a disclosure risk.

This scenario is the motivation for this research work. It aims at solving the problems such as high information loss, homogeneity attack, semantic attack, proximity breach and divergence breach while preserving privacy. The proposed method overcomes the restriction on number of sensitive attributes to be handled and gives a task and application independent solution. Also, it avoids the problem of identifying Quasi-identifier attributes.

In the literature, cryptographic techniques have been applied for the privacy preserving collaborative data mining. These cryptographic approaches cause high overhead and depend on the type of partitioned data and type of mining task. So, this research work aims to find task- and application-independent solution to overcome these limitations.

This research work measures information content and data utility of publishable table by performing various data mining tasks on it and also develops an adversarial model to measure the privacy disclosure risk of the proposed methods.