References :

1) Hassan A. Afyouni, “Database Security and Auditing”, Third Edition, Cengage Learning, 2009

2) Charu C. Aggarwal, Philip S Yu, “Privacy Preserving Data Mining”: Models and Algorithms, Kluwer Academic Publishers, 2008

3) Ron Ben Natan, ”Implementing Database Security and Auditing”, Elsevier Digital Press, 2005.

4) Charu C. Aggarwal, Philip S Yu, “Privacy Preserving Data Mining”: Models and Algorithms, Kluwer Academic Publishers, 2008

5) http://charuaggarwal.net/toc.pdf

6) http://adrem.ua.ac.be/sites/adrem.ua.ac.be/files/securitybook.pdf

UNIT V - PRIVACY PRESERVING DATA MINING TECHNIQUES

- ✔ Introduction
- ✔ Privacy Preserving Data Mining Algorithms
- ✔ General Survey
- ✔ Randomization Methods
- ✔ Group Based Anonymization
- ✔ Distributed Privacy Preserving Data Mining
- ✔ Curse of Dimensionality
- ✔ Application of Privacy Preserving Data Mining

Introduction - privacy-preserving data mining

- ✔ The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information.

- ✔ The problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community.

- ✔ This tutorial will try to explore different topics from the perspective of different communities and give a fused idea of the work in different communities.

- ✔ The key directions in the field of privacy-preserving data mining are:
    - ✔ Privacy-preserving data publishing
    - ✔ Changing the results of data mining applications to preserve privacy
    - ✔ Query auditing
    - ✔ Cryptographic methods for distributed privacy
    - ✔ Theoretical challenges in high dimensionality

Privacy Preserving Data Mining Algorithms

- ✔ The problem of privacy-preserving data mining has become more important in recent years because of the increasing ability to store personal data about users, and the increasing sophistication of data mining algorithms to leverage this information.

- ✔ A number of techniques such as randomization and *k*-anonymity have been suggested in recent years in order to perform privacy-preserving data mining.

- ✔ Furthermore, the problem has been discussed in multiple communities such as the database community, the statistical disclosure control community and the cryptography community.

- ✔ The key directions in the field of privacy-preserving data mining are as follows:
  - Privacy-Preserving Data Publishing
  - Changing the results of Data Mining Applications to preserve privacy
  - Query Auditing
  - Cryptographic Methods for Distributed Privacy
  - Theoretical Challenges in High Dimensionality

Privacy Preserving Data Mining Algorithms …

Privacy-Preserving Data Publishing:

- ✔ These techniques tend to study
- ✔ different transformation methods associated with privacy
- ✔ These techniques include methods such as randomization , *k*-anonymity ,and *l*-diversity .
- ✔ Another related issue is how the perturbed data can be used in conjunction with classical data mining methods such as association rule mining
- ✔ Other related problems include that of determining privacy-preserving methods to keep the underlying data useful (utility-based methods), or the problem of studying the different definitions of privacy, and how they compare in terms of effectiveness in different scenarios.

Privacy Preserving Data Mining Algorithms …

Changing the results of Data Mining Applications to preserve privacy :

- ✔ In many cases, the results of data mining applications such as association rule or classification rule mining can compromise the privacy of the data.

- ✔ This has spawned a field of privacy in which the results of data mining algorithms such as association rule mining are modified in order to preserve the privacy of the data.

- ✔ A classic example of such techniques are association rule hiding methods, in which some of the association rules are suppressed in order to preserve privacy.

Privacy Preserving Data Mining Algorithms …

Query Auditing:

- ✔ Such methods are akin to the previous case of modifying the results of data mining algorithms
- ✔ Here, we are either modifying or restricting the results of queries.

Cryptographic Methods for Distributed Privacy:

- ✔ In many cases, the data may be distributed across multiple sites, and the owners of the data across these different sites may wish to compute a common function.
- ✔ In such cases, a variety of cryptographic protocols may be used in order to communicate among the different sites, so that secure function computation is possible without revealing sensitive information.

Theoretical Challenges in High Dimensionality:

- ✔ Real data sets are usually extremely high dimensional, and this makes the process of privacy-preservation extremely difficult both from a computational and effectiveness point of view.
- ✔ It has been shown that optimal *k*-anonymization is NP-hard. Furthermore, the technique is not even effective with increasing dimensionality, since the data can typically be combined with either public or background information to reveal the identity of the underlying record owners.

Privacy Preserving Data Mining Algorithms …

General Survey:

- ✔ There is a broad survey of privacy preserving data-mining methods.
- ✔ It provides an overview of the different techniques and how they relate to one another.
- ✔ The idea is to provide an overview of the field for a new reader from the perspective of the data mining community.
- ✔ However, more detailed discussions are deferred to future chapters which contain descriptions of different data mining algorithms.

Privacy Preserving Data Mining Algorithms – A General Survey

- ✔ Statistical Methods for Disclosure Control
- ✔ Measures of Anonymity
- ✔ The *k*-anonymity Method
- ✔ The Randomization Method
- ✔ Quantification of Privacy
- ✔ Utility Based Privacy-Preserving Data Mining
- ✔ Mining Association Rules under Privacy Constraints
- ✔ Cryptographic Methods for Information Sharing and Privacy
- ✔ Privacy Attacks
- ✔ Query Auditing and Inference Control
- ✔ Privacy and the Dimensionality Curse
- ✔ Personalized Privacy Preservation
- ✔ Privacy-Preservation of Data Streams
- ✔ Conclusions and Summary

Privacy Preserving Data Mining Algorithms – A General Survey

Statistical Methods for Disclosure Control

- The topic of privacy-preserving data mining has often been studied extensively by the data mining community without sufficient attention to the work done by the conventional work done by the statistical disclosure control community.

- Detailed methods for statistical disclosure control have been presented along with some of the relationships to the parallel work done in the database and data mining community.

- This includes methods such as *k*-anonymity, swapping, randomization, micro-aggregation and synthetic data generation.

- The idea is to give the readers an overview of the common themes in privacy-preserving data mining by different communities.

Privacy Preserving Data Mining Algorithms – A General Survey

Measures of Anonymity

- ✔ There are a very large number of definitions of anonymity in the privacy-preserving data mining field.
- ✔ This is partially because of the varying goals of different privacy-preserving data mining algorithms.
- ✔ For example, methods such as *k*-anonymity, *l*-diversity and *t*-closeness are all designed to prevent identification, though the final goal is to preserve the underlying sensitive information.
- ✔ Each of these methods is designed to prevent disclosure of sensitive information in a different way.

Privacy Preserving Data Mining Algorithms – A General Survey

The *k*-anonymity Method

- An important method for privacy de-identification is the method of *k*-anonymity.

- The motivating factor behind the *k*-anonymity technique is that many attributes in the data can often be considered pseudo-identifiers which can be used in conjunction with public records in order to uniquely identify the records.

- For example, if the identifications from the records are removed, attributes such as the birth date and zip-code an be used in order to uniquely identify the identities of the underlying records.

- For example, if the identifications from the records are removed, attributes such as the birth date and zip-code an be used in order to uniquely identify the identities of the underlying records.

Privacy Preserving Data Mining Algorithms – A General Survey

The Randomization Method

- ✔ The randomization technique uses data distortion methods in order to create private representations of the records
- ✔ In most cases, the individual records cannot be recovered, but only aggregate distributions can be recovered.
- ✔ These aggregate distributions can be used for data mining purposes. Two kinds of perturbation are possible with the randomization method:
  - ✔ **Additive Perturbation:**
    - ✔ In this case, randomized noise is added to the data records. The overall data distributions can be recovered from the randomized records.
    - ✔ Data mining and management algorithms re designed to work with these data distributions.
  - ✔ **Multiplicative Perturbation:**
    - ✔ In this case, the random projection or random rotation techniques are used in order to perturb the records.

Privacy Preserving Data Mining Algorithms – A General Survey

Quantification of Privacy

- ✔ A key issue in measuring the security of different privacy-preservation methods is the way in which the underlying privacy is quantified.

- ✔ The idea in privacy quantification is to measure the risk of disclosure for a given level of perturbation.

Privacy Preserving Data Mining Algorithms – A General Survey

Utility Based Privacy-Preserving Data Mining

- ✔ Most privacy-preserving data mining methods apply a transformation which reduces the effectiveness of the underlying data when it is applied to data mining methods or algorithms.

- ✔ There is a natural trade-off between privacy and accuracy, though this trade-off is affected by the particular algorithm which is used for privacy preservation.

- ✔ A key issue is to maintain maximum utility of the data without compromising the underlying privacy constraints.

Privacy Preserving Data Mining Algorithms – A General Survey

Mining Association Rules under Privacy Constraints

- ✔ Since association rule mining is one of the important problems in data mining
- ✔ There are two aspects to the privacy preserving association rule mining problem

1. When the input to the data is perturbed, it is a challenging problem to accurately determine the association rules on the perturbed data.

2. A different issue is that of output association rule privacy.
   In this case, to ensure that none of the association rules in the output result in leakage of sensitive data.

   This problem is referred to as *association rule hiding* by the database community, and that of *contingency table privacy-preservation* by the statistical community.

Privacy Preserving Data Mining Algorithms – A General Survey

Cryptographic Methods for Information Sharing and Privacy

- ✔ In many cases, multiple parties may wish to share *aggregate private data*, without leaking any sensitive information at their end

- ✔ For example, different superstores with sensitive sales data may wish to coordinate among themselves in knowing aggregate trends without leaking the trends of their individual stores.

- ✔ This requires *secure and cryptographic protocols* for sharing the information across the different parties. The data may be distributed in two ways across different sites:

  - ✔ **Horizontal Partitioning:** In this case, the different sites may have different sets of records containing the same attributes.

  - ✔ **Vertical Partitioning:** In this case, the different sites may have different attributes of the same sets of records.

Privacy Preserving Data Mining Algorithms – A General Survey

Privacy Attacks

- ✔ It is useful to examine the different ways in which one can make adversarial attacks on privacy-transformed data.
- ✔ This helps in designing more effective privacy-transformation methods.
- ✔ Some examples of methods which can be used in order to attack the privacy of the underlying data include SVD-based methods, spectral filtering methods and background knowledge attacks.

Privacy Preserving Data Mining Algorithms – A General Survey

Query Auditing and Inference Control

- ✔ Many private databases are open to querying. This can compromise the security of the results, when the adversary can use different kinds of queries in order to undermine the security of the data.

- ✔ For example, a combination of range queries can be used in order to narrow down the possibilities for that record. Therefore, the results over multiple queries can be combined in order to uniquely identify a record, or at least reduce the uncertainty in identifying it.

- ✔ There are two primary methods for preventing this kind of attack:

  - ✔ **Query Output Perturbation:** In this case, we add noise to the output of the query result in order to preserve privacy.

  - ✔ **Query Auditing:** In this case, we choose to deny a subset of the queries, so that the particular combination of queries cannot be used in order to violate the privacy

Privacy Preserving Data Mining Algorithms – A General Survey

Privacy and the Dimensionality Curse

- ✔ In recent years, it has been observed that many privacy-preservation methods such as *k*-anonymity and randomization are not very effective in the high dimensional case

Personalized Privacy Preservation

- ✔ In many applications, different subjects have different requirements for privacy.

- ✔ For example, a brokerage customer with a very large account would likely have a much higher level of privacy-protection than a customer with a lower level of privacy protection.

- ✔ In such case, it is necessary to *personalize* the privacy-protection algorithm.

Privacy Preserving Data Mining Algorithms – A General Survey

Privacy-Preservation of Data Streams

- A new topic in the area of privacy preserving data mining is that of data streams, in which data grows rapidly at an unlimited rate.

- In such cases, the problem of privacy-preservation is quite challenging since the data is being released incrementally.

- In addition, the fast nature of data streams obviates the possibility of using the past history of the data.

Privacy Preserving Data Mining Algorithms – A General Survey

Conclusions and Summary

✔ The broad areas of privacy are as follows:

Privacy-preserving data publishing:

This corresponds to sanitizing the data, so that its privacy remains preserved.

Privacy-Preserving Applications:

This corresponds to designing data management and mining algorithms in such a way that the privacy remains preserved. Some examples include association rule mining, classification, and query processing.

Utility Issues:

Since the perturbed data may often be used for mining and management purposes, its utility needs to be preserved. Therefore, the data mining and privacy transformation techniques need to be designed effectively, so to preserve the utility of the results.

Distributed Privacy, cryptography and adversarial collaboration:

This corresponds to secure communication protocols between trusted parties, so that information can be shared effectively without revealing sensitive information about particular parties.

Randomization Method

- ✔ 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.

- ✔ The noise added is sufficiently large so that individual record values cannot be recovered.

- ✔ Therefore, techniques are designed to derive aggregate distributions from the perturbed records.

- ✔ Subsequently, data mining techniques can be developed in order to work with these aggregate distributions.

Randomization Method …

The method of randomization can be described as follows.

- Consider a set of data records denoted by $X = \{x1 \ldots xN\}$
- For record $xi \in X$
- we add a noise component which is drawn from the probability distribution $fY(y)$.
- These noise components are drawn independently, and are denoted $y1 \ldots yN$.
- Thus, the new set of distorted records are denoted by

  $x1 + y1 \ldots xN + yN$.
- We denote this new set of records by

  $z1 \ldots zN$.
- In general, it is assumed that the variance of the added noise is large enough, so that the original record values cannot be easily guessed from the distorted data.
- Thus, the original records cannot be recovered, but the distribution of the original records can be recovered.

Randomization Method …

- ✔ Thus, if ***X*** be the random variable denoting the data distribution for the original record
- ✔ ***Y*** be the random variable describing the noise distribution
- ✔ ***Z*** be the random variable denoting the final record

We have:

$$Z = X + Y$$

$$X = Z - Y$$

- ✔ Now, we note that ***N*** instantiations of the probability distribution ***Z*** are known, whereas the distribution ***Y*** is known publicly.
- ✔ For a large enough number of values of ***N***, the distribution ***Z*** can be approximated closely by using a variety of methods such as kernel density estimation.
- ✔ By subtracting ***Y*** from the approximated distribution of ***Z***, it is possible to approximate the original probability distribution ***X***

Randomization Method …

- ✔ One key advantage of the randomization method is that it is relatively simple, and does not require knowledge of the distribution of other records in the data.

- ✔ This is not true of other methods such as *k*-anonymity which require the knowledge of other records in the data.

- ✔ Therefore, the randomization method can be implemented at *data collection time*, and does not require the use of a trusted server containing all the original records in order to perform the anonymization process.

- ✔ While this is a strength of the randomization method, it also leads to some weaknesses, since it treats all records equally irrespective of their local density.

Randomization Method …

Privacy Quantification

- ✔ The quantity used to measure privacy should indicate how closely the original value of an attribute can be estimated.

- ✔ A measure that defines privacy as follows:

  If the original value can be estimated with ***c*%** confidence to lie in the interval **[$\alpha 1$, $\alpha 2$]**, then the interval width **($\alpha 2 - \alpha 1$)** defines the amount of privacy at ***c*%** confidence level.

- ✔ For example,

  If the perturbing additive is uniformly distributed in an interval of width **2α**, then α is he amount of privacy at confidence level **50%** and **2α** is the amount of privacy at confidence level **100%.**

- ✔ However, this simple method of determining privacy an be subtly incomplete in some situations.

Randomization Method …

Randomization Methods for Data Streams

- ✔ The randomization approach is particularly well suited to privacy-preserving data mining of streams, since the noise added to a given record is independent of the rest of the data.

- ✔ However, streams provide a particularly vulnerable target for adversarial attacks with the use of PCA (Principle Component Analysis) based techniques because of the large volume of the data available for analysis.

Randomization Method …

Multiplicative Perturbations

- The most common method of randomization is that of additive perturbations.

- However, multiplicative perturbations can also be used to good effect for privacy-preserving data mining.

- Many of these techniques derive their roots in the work of which shows how to use multi-dimensional projections in order to reduce the dimensionality of the data.

- This technique preserves the inter record distances approximately, and therefore the transformed records can be used in conjunction with a variety of data mining applications.

Randomization Method …

- ✔ As in the case of additive perturbations, multiplicative perturbations are not entirely safe from adversarial attacks.

- ✔ In general, if the attacker has no prior knowledge of the data, then it is relatively difficult to attack the privacy of the transformation.

- ✔ However, with some prior knowledge, two kinds of attacks are possible

  - ✔ **Known Input-Output Attack:**
    - ✔ In this case, the attacker knows some linearly independent collection of records, and their corresponding perturbed version. In such cases, linear algebra techniques can be used to reverse-engineer the nature of the privacy preserving transformation.

  - ✔ **Known Sample Attack:**
    - ✔ In this case, the attacker has a collection of independent data samples from the same distribution from which the original data was drawn. In such cases, principal component analysis techniques can be used in order to reconstruct the behavior of the original data.

Randomization Method …

Data Swapping

- ✔ Noise addition or multiplication is not the only technique which can be used to perturb the data.
- ✔ A related method is that of data swapping, in which the values across different records are swapped in order to perform the privacy-preservation
- ✔ One advantage of this technique is that the lower order marginal totals of the data are completely preserved and are not perturbed at all.
  - ▪ Therefore certain kinds of aggregate computations can be exactly performed without violating the privacy of the data.
- ✔ This technique does not follow the general principle in randomization which allows the value of a record to be perturbed independent;y of the other records.
  - ▪ Therefore, this technique can be used in combination with other frameworks such as *k*-anonymity, as long as the swapping process is designed to preserve the definitions of privacy for that model.

Group Based Anonymization

- ✔ The randomization method is a simple technique which can be easily implemented at data collection time, because the noise added to a given record is independent of the behavior of other data records.

- ✔ This is also a weakness because outlier records can often be difficult to mask.

- ✔ Clearly, in cases in which the privacy-preservation does not need to be performed at data-collection time, it is desirable to have a technique in which the level of inaccuracy depends upon the behavior of the locality of that given record.

- ✔ Another key weakness of the randomization framework is that it does not consider the possibility that publicly available records can be used to identify the identity of the owners of that record.

- ✔ Therefore, a broad approach to many privacy transformations is to construct groups of anonymous records which are transformed in a group-specific way.

Group Based Anonymization …

The *k*-Anonymity Framework

- ✔ In many applications, the data records are made available by simply removing key identifiers such as the name and social- security numbers from personal records.

- ✔ However, other kinds of attributes (known as pseudo-identifiers) can be used in order to accurately identify the records.

  - For example, attributes such as age, zip-code and sex are available in public records such as census rolls.

  - When these attributes are also available in a given data set, they can be used to infer the identity of the corresponding individual.

  - A combination of these attributes can be very powerful, since they can be used to narrow down the possibilities to a small number of individuals.

Group Based Anonymization …

- ✔ In *k*-anonymity techniques, it reduce the granularity of representation of these pseudo-identifiers with the use of techniques such as *generalization* and *suppression*.

- ✔ In the method of *generalization*, the attribute values are generalized to a range in order to reduce the granularity of representation.
  - For example, the date of birth could be generalized to a range such as year of birth, so as to reduce the risk of identification.

- ✔ In the method of *suppression*, the value of the attribute is removed completely.

- ✔ It is clear that such methods reduce the risk of identification with the use of public records, while reducing the accuracy of applications on the transformed data.

- ✔ In order to reduce the risk of identification, the *k*-anonymity approach requires that every tuple in the table be indistinguishability related to no fewer than *k* respondents.

Group Based Anonymization …

- ✔ *k*-anonymity approach can be formalized as follows:
  - *Each release of the data must be such that every combination of values of quasi-identifiers (* are pieces of information that are not of themselves unique identifiers) *can be indistinguishably matched to at least k respondents.*

- ✔ The first algorithm for *k*-anonymity approach uses *domain generalization hierarchies* of the quasi-identifiers in order to build *k*-anonymous tables.

- ✔ The concept of *k*-minimal generalization has been proposed in order to limit the level of generalization for maintaining as much data precision as possible for a given level of anonymity.

- ✔ Subsequently, the topic of *k*-anonymity has been widely researched.

Group Based Anonymization …

- ✔ It was note that the problem of optimal anonymization is inherently a difficult one.
- ✔ It has been shown that the problem of optimal *k*-anonymization is NP-hard. Nevertheless, the problem can be solved quite effectively by the use of a number of heuristic methods.
- ✔ A method proposed by Bayardo and Agrawal is the *k-Optimize* algorithm which can often obtain effective solutions.
- ✔ The approach assumes an ordering among the quasi-identifier attributes.
- ✔ The values of the attributes are discretized into intervals quantitative attributes) or grouped into different sets of values (categorical attributes). Each such grouping is an *item*.
- ✔ For a given attribute, the corresponding items are also ordered. An index is created using these attribute-interval pairs (or items) and a set enumeration tree is constructed on these attribute-interval pairs.
- ✔ *k*-Optimize algorithm can use a number of pruning strategies to good effect.

Group Based Anonymization …

- ✔ A branch and bound technique can be used to successively improve the quality of the solution during the traversal process.

- ✔ *Incognito* method has been proposed for computing a *k*-minimal generalization with the use of bottom-up aggregation along domain generalization hierarchies.

- ✔ The Incognito method uses a bottom-up breadth-first search of the domain generalization hierarchy, in which it generates all the possible minimal *k*-anonymous tables for a given private table.

Group Based Anonymization …

- ✔ First, it checks *k*-anonymity for each single attribute, and removes all those generalizations which do not satisfy *k*-anonymity. Then, it computes generalizations in pairs, again pruning those pairs which do not satisfy the *k*-anonymity constraints.

- ✔ Incognito algorithm computes (i + 1)-dimensional generalization *candidates* from the *i*-dimensional generalizations, and removes all those generalizations which do not satisfy the *k*-anonymity constraint.

- ✔ This approach is continued until, no further candidates can be constructed, or all possible dimensions have been exhausted.

Personalized Privacy-Preservation

Not all individuals or entities are equally concerned about their privacy.

- For example, a corporation may have very different constraints on the privacy of its records as compared to an individual.

- This leads to the natural problem that we may wish to treat the records in a given data set very differently for anonymization purposes.

- From a technical point of view, this means that the value of *k* for anonymization is not fixed but may vary with the record.

- A condensation based approach has been proposed for privacy-preserving data mining in the presence of variable constraints on the privacy of the data records.

Personalized Privacy-Preservation…

- ✔ This technique constructs groups of non-homogeneous size from the data, such that it is guaranteed that each record lies in a group whose size is at least equal to its anonymity level

- ✔ Subsequently, pseudo-data is generated from each group so as to create a synthetic data set with the same aggregate distribution as the original data.

- ✔ Another interesting model of personalized anonymity is discussed in which a person can specify the level of privacy for his or her *sensitive values*.

- ✔ This technique assumes that an individual can specify a node of the domain generalization hierarchy in order to decide the level of anonymity that he can work with.

- ✔ This approach has the advantage that it allows for direct protection

- ✔ of the sensitive values of individuals than a vanilla *k*-anonymity method which is susceptible to different kinds of attacks.

Utility Based Privacy Preservation

- ✔ The process of privacy-preservation leads to loss of information for data mining purposes.
- ✔ This loss of information can also be considered a loss of *utility* for data mining purposes.
- ✔ Since some negative results on the curse of dimensionality suggest that a lot of attributes may need to be suppressed in order to preserve anonymity, it is extremely important to do this carefully in order to preserve utility.
- ✔ We note that many anonymization methods use cost measures in order to measure the information loss from the anonymization process.
- ✔ Examples of such utility measures include
  - Generalization height
  - Size of anonymized group
  - Discernability measures of attribute values
  - Privacy information loss ratio

Utility Based Privacy Preservation…

- ✔ A method for utility-based data mining using local recoding was proposed, The approach is based on the fact that different attributes have different utility from an application point of view.

- ✔ Most anonymization methods are *global*, in which a particular tuple value is mapped to the same generalized value globally.

- ✔ In local recoding, the data space is partitioned into a number of regions, and the mapping of the tuple to the generalizes value is local to that region.

- ✔ This kind of approach has greater flexibility, since it can tailor the generalization process to a particular region of the data set.

Utility Based Privacy Preservation…

- ✔ Another indirect approach to utility based anonymization is to make the privacy-preservation algorithms more aware of the workload.
- ✔ Typically, data recipients may request only a subset of the data in many cases, and the union of these different requested parts of the data set is referred to as the workload.
- ✔ A workload in which some records are used more frequently than others tends to suggest a different anonymization than one which is based on the entire data set.
- ✔ Another direction for utility based privacy-preserving data mining is to anonymize the data in such a way that it remains useful for particular kinds of data mining or database applications.
- ✔ In such cases, the utility measure is often affected by the underlying application at hand.
- ✔ There is a method has been proposed for *k*-anonymization using an information-loss metric as the utility measure.

Sequential Releases

- ✔ Privacy-preserving data mining poses unique problems for dynamic applications such as data streams because in such cases, the data is released sequentially.
- ✔ In other cases, different views of the table may be released sequentially.
- ✔ Once a data block is released, it is no longer possible to go back and increase the level of generalization.
- ✔ On the other hand, new releases may sharpen an attacker's view of the data and may make the overall data set more susceptible to attack.
- ✔ A technique discussed in relies on lossy joins in order to cripple an attack based on global quasi identifiers.
- ✔ The intuition behind this approach is that if the join is lossy enough, it will reduce the confidence of the attacker in relating the release from previous views to the current release.
- ✔ A new generalization principle called *m*-invariance is proposed, which effectively limits the risk of privacy-disclosure in re-publication.
- ✔ The broad idea in this approach is to progressively and consistently increase the generalization granularity, so that the released data satisfies the *k*-anonymity requirement both with respect to the current table, as well as with respect to the previous releases

The *l* -diversity Method

- The *k*-anonymity is an attractive technique because of the simplicity of the definition and the numerous algorithms available to perform the anonymization.
- The *k*-anonymity is an attractive technique because of the simplicity of the definition and the numerous algorithms available to perform the anonymization.
- Nevertheless the technique is susceptible to many kinds of attacks especially when background knowledge is available to the attacker
- Some kinds of such attacks are as follows:
    - **Homogeneity Attack:**
        - In this attack, all the values for a sensitive attribute within a group of *k* records are the same. Therefore, even though the data is *k*-anonymized, the value of the sensitive attribute for that group of *k* records can be predicted exactly.
    - **Background Knowledge Attack:**
        - In this attack, the adversary can use an association between one or more quasi-identifier attributes with the sensitive attribute in order to narrow down possible values of the sensitive field further

The *l* -diversity Method

- ✔ While *k*-anonymity is effective in preventing *identification* of a record, it may not always be effective in preventing inference of the sensitive values of the attributes of that record.

- ✔ Therefore, the technique of *l*-diversity was proposed which not only maintains the minimum group size of *k*, but also focuses on maintaining the diversity of the sensitive attributes.

- ✔ Therefore, the *l*-diversity model for privacy is defined as follows:

  - *Let a q*-block be a set of tuples such that its non-sensitive values generalize to q*.*
  - *A q*-block is l-diverse*
    - *if it contains l "well represented" values for the sensitive attribute S.*
    - *A table is l-diverse, if every q*-block in it is l-diverse.*

- ✔ when there are multiple sensitive attributes, then the *l*-diversity problem becomes especially challenging because of the curse of dimensionality.

The *t*-closeness Model

- The *t*-closeness model is a further enhancement on the concept of *l*-diversity.
- 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.
- A *t*-closeness model was proposed which uses the property that the distance between the distribution of the sensitive attribute within an anonymized group should not be different from the global distribution by more than a threshold *t*.

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 participant.
- ✔ 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 have 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.

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.

- ✔ 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 $h(x1 \ldots xk)$. 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 $h(x1 \ldots, xk)$, a *protocol* will have to designed for exchanging information in such a way that the function is computed without compromising privacy.

Distributed Privacy-Preserving Data Mining …

- ✔ 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.

The Curse of Dimensionality

- ✔ Many privacy-preserving data-mining methods are inherently limited by the curse of dimensionality in the presence of public information.

- ✔ For example, the technique in analyzes the *k*-anonymity method in the presence of increasing dimensionality.

- ✔ The curse of dimensionality becomes especially important when adversaries may have considerable background information, as a result of which the boundary between pseudo-identifiers and sensitive attributes may become blurred.

- ✔ This is generally true, since adversaries may be familiar with the subject of interest and may have greater information about them than what is publicly available.

- ✔ This is also the motivation for techniques such as *l*-diversity in which background knowledge can be used to make further privacy attacks.

Applications of Privacy-Preserving Data Mining

- ✔ The problem of privacy-preserving data mining has numerous applications in homeland security, medical database mining, and customer transaction analysis.

- ✔ Some of these applications such as those involving bio-terrorism and medical database mining may intersect in scope.

- ✔ Number of different applications of privacy-preserving data mining methods.

  - ▪ Medical Databases: The Scrub and Datafly Systems
  - ▪ Bioterrorism Applications
  - ▪ Homeland Security Applications
  - ▪ Genomic Privacy

Applications of Privacy-Preserving Data Mining …

Medical Databases: The Scrub and Datafly Systems

Scrub :

- ✔ The scrub system was designed for de-identification of clinical notes and letters which typically occurs in the form of textual data.
- ✔ Clinical notes and letters are typically in the form of text which contain references to patients, family members, addresses, phone numbers or providers.
- ✔ Traditional techniques simply use a global search and replace procedure in order to provide privacy.
- ✔ However clinical notes often contain cryptic references in the form of abbreviations which may only be understood either by other providers or members of the same institution.
- ✔ Therefore traditional methods can identify no more than 30-60% of the identifying information in the data
- ✔ The Scrub System uses local knowledge sources which compete with one another based on the certainty of their findings.
- ✔ Such a system is able to remove more than 99% of the identifying information from the data.

Applications of Privacy-Preserving Data Mining …

Datafly Systems:

- ✔ The Datafly System was one of the earliest practical applications of privacy-preserving transformations.
- ✔ This system was designed to prevent identification of the subjects of medical records which may be stored in multidimensional format.
- ✔ The multi-dimensional information may include directly identifying information such as the social security number, or indirectly identifying information such as age, sex or zip-code.
- ✔ The system was designed in response to the concern that the process of removing only directly identifying attributes such as social security numbers was not sufficient to guarantee privacy.

Applications of Privacy-Preserving Data Mining …

- ✔ Typically, the user of Datafly will set the anonymity level depending upon the profile of the data recipient in question.

- ✔ The overall anonymity level is defined between 0 and 1, which defines the minimum bin size for each field.

- ✔ An anonymity level of 0 results in Datafly providing the original data, whereas an anonymity level of 1 results in the maximum level of generalization of the underlying data.

- ✔ The Datafly system is one of the earliest systems for anonymization, and is quite simple in its approach to anonymization.

Applications of Privacy-Preserving Data Mining …

Bioterrorism Applications

✔ Often a biological agent such as anthrax produces symptoms which are similar to other common respiratory diseases such as the cough, cold and the flu.

✔ In the absence of prior knowledge of such an attack, health care providers may diagnose a patient affected by an anthrax attack of have symptoms from one of the more common respiratory diseases.

✔ In order to identify such attacks it is necessary to track incidences of these common diseases as well.

✔ Therefore, the corresponding data would need to be reported to public health agencies. However, the common respiratory diseases are not reportable diseases by law.

Applications of Privacy-Preserving Data Mining …

- ✔ **Homeland Security Applications**
  - A number of applications for homeland security are inherently intrusive because of the very nature of surveillance.
  - Some examples of such applications are as follows:
    - ✔ **Credential Validation Problem:**
      - Trying to match the subject of the credential to the person presenting the credential.
      - For example, the theft of social security numbers presents a serious threat to homeland security.
    - ✔ **Identity Theft:**
      - A related technology is to use a more *active* approach to avoid identity theft.
      - The *identity angel* system , crawls through cyberspace, and determines people who are at risk from identity theft.
      - This information can be used to notify appropriate parties.

Applications of Privacy-Preserving Data Mining …

Web Camera Surveillance:

- ✔ One possible method for surveillance is with the use of publicly available webcams which can be used to detect unusual activity.

- ✔ this is a much more invasive approach than the previously discussed techniques because of person specific information being captured in the webcams.

- ✔ The approach can be made more privacy-sensitive by extracting only *facial count* information from the images and using these in order to detect unusual activity.

Applications of Privacy-Preserving Data Mining …

Video-Surveillance:

- ✔ In the context of sharing video-surveillance data, a major threat is the use of facial recognition software, which can match the facial images in videos to the facial images in a driver license database.
- ✔ While a straightforward solution is to completely black out each face, the result is of limited new, since all facial information has been wiped out.
- ✔ A more balanced approach is to use selective downgrading of the facial information, so that it scientifically limits the ability of facial recognition software to reliably identify faces, while maintaining facial details in images.
- ✔ The algorithm is referred to as *k*-Same, and the key is to identify faces which are somewhat similar, and then construct new faces which construct combinations of features from these similar faces.

Applications of Privacy-Preserving Data Mining …

The Watch List Problem:

- ✔ The motivation behind this problem is that the government typically has a list of known terrorists or suspected entities which it wishes to track from the population.
- ✔ The aim is to view transactional data such as store purchases, hospital admissions, airplane manifests, hotel registrations or school attendance records in order to identify or track these entities.
- ✔ This is a difficult problem because the transactional data is private, and the privacy of subjects who do not appear in the watch list need to be protected.
- ✔ Therefore, the transactional behavior of non-suspicious subjects may not be identified or revealed.
- ✔ The watch list problem is currently an open problem.

Applications of Privacy-Preserving Data Mining …

Genomic Privacy

- Recent years have seen tremendous advances in the science of DNA sequencing and forensic analysis with the use of DNA.
- As result, the databases of collected DNA are growing very fast in the both the medical and law enforcement communities.
- DNA data is considered extremely sensitive, since it contains almost uniquely identifying information about an individual.
- As in the case of multi-dimensional data, simple removal of directly identifying data such as social security number is not sufficient to prevent re-identification.
- It has been shown that a software called *CleanGene* can determine the identifiability of DNA entries independent of any other demographic or other identifiable information.
- The software relies on publicly available medical data and knowledge of particular diseases in order to assign identifications to DNA entries.
- Another method for compromising the privacy of genomic data is that of *trail re-identification*, in which the uniqueness of patient visit patterns is exploited in order to make identifications.