# A Review on High Dimensional Data Publication

Prof.Deepa B. Mane

Assistant Professor,
Information Technology,
Anantrao Pawar College of Engineering and Reaserch, Pune,India

*Abstract :* In today's global world information is most demanded resource. globally networked society demands sharing of information. The Microdata to be published many times contains sensitive data, publishing such data without proper protection may risky to the individual's privacy, so must be preserved by Data Publisher before it is published. Research on protecting the privacy of individual's sensitive data has received contributions from many fields, such as computer science, economics, and social science. Privacy-preserving data publishing (PPDP) balances the fundamental trade-off between individual privacy and the utility of published data.

A number of different techniques have recently been used for privacy preserving of multi dimensional data. data anonymization is one of the most important nowadays . Data anonymization techniques, such as generalization, bucketization have been designed. Generalization losses large amount of information when it used for high dimensional data. Bucketization requires separation between quasi attributes with sensitive attributes. So, in this paper we introduce a novel technique called slicing which provides better data utility and preserves privacy**.**

***IndexTerms*** **– Anonymization, Microdata release, Data publishing, Data security, Privacy preservation.**

## I. INTRODUCTION

Government rules handling departments and other social and private organizations often need to publish Microdata for research and for data mining for extracting useful information from the data. The exploitation of Data Mining and Knowledge discovery has penetrated to a variety of Machine Learning Systems . A very important area in the field of privacy preserving is Text Categorization[4] to analyze the data characteristics. Typically, such data are stored in a table, and each table record (row) corresponds to one particular individual. Every record has lots of attributes, which can be categorized as the following three categories: Attributes that uniquely identify individuals. These can be termed as explicit identifiers e.g. Social Security Number. Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date, and Gender. Attributes that are represents sensitive information, such as Disease and Salary. When releasing Microdata, it requires to preserve the confidentiality of sensitive information [6] of the individuals. Two different type of information disclosure have been identified in the literature: identity disclosure and attribute disclosure. Identity disclosure is also known as link attack occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when additional information about some individuals is disclosed i.e the released data make it possible to infer the characteristics of an individual more accurately than it would be possible before the data publishing. Identity disclosure often results into attribute disclosure. So to preserve the data from these attacks it must be first anonymized the data before publishing. In anonymization first stage is to remove the unique identifiers. However, this is not sufficient, as an adversary may already know the quasi- identifier values of some individuals in the table, this knowledge can be either from the personal knowledge or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi- identifiers [13].Generalization is the most commonly used anonymization technique, which replaces quasi-identifier values with generalized values that are less-specific but semantically consistent. As a result, more records have the same set of quasi- identifier values. Another problem with privacy-preserving methods [10], in general, is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitiveGeneralization is the most common method used for de-identification of the data in k-anonymity based algorithms Generalization consists of substituting generalized value to the attributes with semantically consistent but less precise values. Generalization maintains the data preserved at the record level but results in less specific information that may affect the accuracy of algorithms applied on the k-anonymous dataset.

In this paper, we present a new technique called slicing for privacy-preserving data
Publishing [2]. Our contributions is the following. We introduce slicing as a new technique for data anonymization in privacy preserving data publishing. Slicing has several advantages when it is compared with generalization and bucketization. It has proved better data utility than generalization as this does not loss considerable amount of data when applied on high dimensional data. It achieves more attribute correlations with the Sensitive attributes than bucketization. It can also handle high-dimensional data and data without a clear separation of Quansi identifiers and Sensitive attributes. Secondly slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of *l*-diversity [1].

## II. BACKGROUND

### K-ANONYMITY:

The k-anonymity [8] model requires that within any equivalence class of the microdata there are at least k records such that k th attribute can not be distinguishable from k-1. The protection k anonymity provides is simple and easy to understand. K-anonymity cannot provide a safeguard against attribute disclosure in all cases. It does not give protection from Homogeneity attack [11] and the Background knowledge[3] attack. Limitations of k-anonymity are*:* (1) it does not protect from membership disclosure(2) it reveals individuals' sensitive attributes , (3) it does not protect against attacks based on background knowledge , (4) mere knowledge of the k-anonymization algorithm can violate privacy, (5) when it is applied on high dimensional data it losses complete data utility

### L-DIVERSITY:

l-diversity overcomes the limitations of the K-anonymity. l-diversity tries to put constraints on minimum number of distinct values that the sensitive attribute can have within an equivalence class. An equivalence class has l-diversity if there is l or more distinct values for the sensitive attribute. A table is said to be l-diverse[1] if each equivalence class of the table is l-diverse. Limitation of L-diversity While the 'l-diversity principle represents an important privacy model beyond k-anonymity in protecting against attribute disclosure, it has several drawbacks. 'l - Diversity may difficult to achieve and may provide insufficient privacy protection.

### GENERALIZATION:

Generalization replaces a Quansi-identifier values to generalized values that is with a "less-specific but semantically consistent" value.It causes too much information loss due to the uniform-distribution assumption.

### BUCKETIZATION:

Bucketization partitions tuples in the records into buckets and then it randomly permutes the sensitive attributes across the bucket. Bucketization does not preserves data from membership disclosure. Because bucketization publishes the Quansi Identifiers values in their original format, an adversary can find out whether an individual has a record in the published data[13] or not. A micro data usually contains many other attributes different from those three attributes. This means that the membership information of most individuals can be identified from the Bucketized table. Bucketization requires a clear separation between Quansi-Identifiers and Sensitive attributes. However, in many data sets, there is confusion about which attributes are Quansi identifiers and which are Sensitive attributes. By separating the sensitive attribute from the Quansi Identifiers 9+attributes, bucketization breaks the attribute correlations between the Quansi Identifiers and the Sensitive attributes. The Anonymized data consist of a set of buckets with permuted sensitive attribute values.

## SLICING

In this paper, we introduce a novel data anonymization technique called slicing[12] to improve the current limitations in the data anonymization techniques. Slicing partitions the dataset both vertically and horizontally. Vertical partitioning is done by grouping highly correlated attributes into columns. Each column

contains a subset of attributes that are highly correlated. Horizontal partitioning is done by grouping tuples into buckets. Finally, within each bucket, values in each column are randomly distributed (or sorted) to break the linking between two different columns. The basic idea of slicing is to break the association across the columns, but to preserve the association within the single column. This reduces the dimensionality of the data and preserves better data Utility [7] than generalization and bucketization. Slicing preserves better utility because it groups highly correlated attributes together, and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are ,uncorrelated ,infrequent and thus identifying. Note that when the dataset contains QIs and SA, bucketization needs to have separation between their correlation; slicing, on the other hand, can group some QI attributes with the SA, preserving attribute correlations with the sensitive attribute.

## III RELATED WORK

Privacy preserving data publishing has applications in Medical databases like in scrub systems and in bioterrorism applications. There are varieties of algorithm are proposed for data anonymization out of which k-anonymity is widely used in this paper we are using k-anonymity for generalization for comparing experimental results.

| Techniques | Dataset | Parameter Used | Advantages | Disadvantages |
|---|---|---|---|---|
| K-Anonymity | Market Basket Dataset | Number of data points, Dimensionality of data space | High correlation among the tuples | More Number of dimensions would be violated |
| ℓ-Diversity | Adult Database | Identifiers, Quasi-identifiers, Sensitive attribute | Sensitive attribute would have at most same frequency | Homogeneity and background knowledge attack has lacked |
| t-closeness | Pension Scheme Dataset | Identifiers, Quasi-identifiers, Sensitive attribute | Measure the distance between two probabilistic distribution that were indistinguishable from one another | Information gain was unclear |
| $K^m$ Anonymity | Market Basket Dataset | Distinct items, Maximum transaction size and Average transaction size on distinct items | Similar evaluated approach on k items | Loss of utility |
| Distributed K Anonymity framework (DKA) | Employee Dataset | Public-key, Secret-key, Encryption | Global anonymization to ensure privacy. | Utility and potential were misused. |
| Slicing | Adult Database | Identifier, Quasi-Identifier, Sensitive Attribute | Maintain trade-off between privacy and utility | Utility and risk measures not matched |

**K-anonymity** is a property that captures the protection of released data against possible re-identification of the respondents to whom the released data refer. To prevent identity linkage attacks through QID, Samarati and Sweeney proposed the notion of k-anonymity: if a record in the table has some value qid, then at least k − 1 other records should also have the value qid. In other words, the minimum group size on QID is at least k. A table satisfying this requirement is called k-anonymous. In a k-anonymous table, each record is indistinguishable from at least k − 1 other records with respect to QID. Consequently, the probability of linking a victim to a specific record through QID is at most 1/k.

**Definition 1:** Quasi-identifier

Given a population of entities U, an entity-specific table T ( A1 ,..., An), fc: U→ T and fg: T → U', where U U'. A quasi-identifier of T, written QT, is a set of attributes {Ai,...,Aj} { A1,... ,An }such that fg (fc (pi)[QT]) =Pi.

K-anonymity:

Let RT(A1,... ,An ) be a table and QIRT be the quasi-identifier associated with it. RT is said to satisfy k-anonymity if and only if each sequence of values in RT[QIRT] appears with at least k occurrences in RT[QIRT].

**l-diversity** A QI group is said to have ℓ-diversity if there are at least ℓ "well-represented" values for the sensitive attribute. A table is said to have ℓ-diversity if every QI group of the table has ℓ-diversity

**Probabilistic ℓ-diversity**. An anonymized table satisfies probabilistic ℓ-diversity if the frequency of a sensitive value in each group is at most 1/ℓ. This guarantees that an observer cannot infer the sensitive value of an individual with probability greater than 1/ℓ.

**Entropy ℓ-diversity**.

The entropy of an QI group E is defined to be attribute, and p(E,s) is the fraction of records in E that have sensitive value s. A table is said to have entropy ℓ-diversity if for every QI group E, Entropy(E) log ℓ. Entropy ℓ-diversity is strong than distinct ℓ-diversity, in order to have entropy ℓ-diversity for each QI group,

the entropy of the entire table must be at least log(ℓ). Sometimes this may too restrictive, as the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative notion of ℓ-diversity.

**t-closeness:**

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

the prior belief. The novelty of our approach is that we separate the information gain into two parts: that about the population in the released data and that about specific individuals.

**Definition 3** (The t-closeness Principle): An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness.

**$K^m$ Anonymity**

Km Anonymity has been proposed for an anonymized transactional database [4]. Km Anonymity aim at protect the database against an adversary who has knowledge about almost m items in the transaction [18]. The generalization was used to maintain the set valued data. For any transaction on K-1 records, other identical transaction would also appear. Km anonymity has been introduced via top down local generalization process to record the number of transaction records [4]. The partition based approach was used to group (partition) the similar items in a top down manner [4]. The km anonymity model would help to prevent privacy breaches raised from an adversary who would discovered m items in a transaction databases

**Distributed K-Anonymity framework (DKA)**

The collection of data from different sites cannot be shared directly. The key step was to anonymized the data in order to generalise a specific value. A secure 2-party framework was designed for multiparty computation that has been used to join the dataset from various sites. Distributed K-Anonymity (DKA) prevent identification of an individual by make use of global Anonymization in the encrypted form. DKA provide a secure framework between two parties. Two parties would agree on Global Anonymization algorithm that could produce local K-Anonymous dataset. In addition, DKA provide a secure distributed protocol which would require that two parties could mutually semi-honest. Still the trade-off between utility and potential of data was misused in DKA.

## IV. Pictorial Representation:

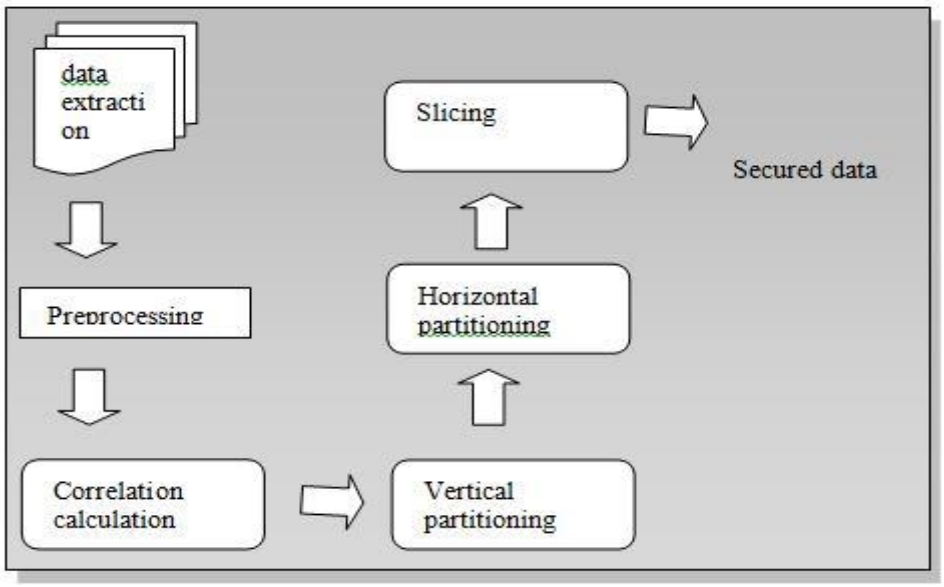


Figure1. Proposed architecture model of system

## REFERENCES


[1] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,2006 "*l-diversity: Privacy beyond k-anonymity*". In ICDE.

[2] B. C. M. Fung, K. Wang, R. Chen, and P. S. Yu. December 2010"*Privacy-preserving data publishing: A survey on recent developments.* ACM Computing Surveys,.

[3] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. 2007, "*Worst-case background knowledge for privacy-preserving data publishing*". In ICDE,.

[4] Dr.Emmanuel M, Saurabh Khatri ,Dr.Ramesh Babu D R, 2013 ,"*A Novel scheme for term weighting in text categorization : Positive Impact factor*",IEEE International Conference on systems, Man and Cybernetics.

[5] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao, February 2011," *Anonymous Publication of Sensitive Transactional Data*" in Proc. Of IEEE Transactions on Knowledge and Data Engineering.

[6] J. Brickell and V. Shmatikov, , 2008, "*The Cost of Privacy: Destruction of Data-Mining Utility in Anonymized Data Publishing*," Proc. ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD).

[7] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, 2005 "*Incognito: Efficient Full-domain k-Anony Anonymity*," in Proc.of ACM SIGMOD.

[8] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, 2006 "*Mondrian Multidimensional k-Anonymity*," in Proc. of ICDE,.

[9] Latanya Sweeney, 2002, *k-anonymity: a model for protecting privacy*. International Journal on Uncertainty, Fuzziness and Knowledge-Based Systems,.

[10] N. Li, T. Li, and S. Venkatasubramanian, 2007 ,"*t-Closeness: Privacy Beyond k-Anonymity and „-Diversity*," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE),.

[11] R. J. Bayardo and R. Agrawal, 2005, "*Data Privacy through Optimal k- Anonymization*," in Proc. of ICDE,.

[12] S. M. Metev and V. P. Veiko, 1998 "*Laser Assisted Microtechnology"*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag,.

[13] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy, MARCH 2012, "*Slicing: A New Approach for Privacy Preserving Data Publishing*" Proc. IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,.

[14] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati , 2007 "On K-Anonymity". In Springer US, Advances in Information Security.

[15] X. Xiao and Y. Tao, 2006, "*Anatomy: Simple and Effective Privacy Preservation*," Proc. Int'l Conf. Very Large Data Bases (VLDB).