International Journal of Research

International Journal of Research (IJR)
e-ISSN: 2348-6848, p- ISSN: 2348-795X Volume 2, Issue 07, July 2015
Available at http://internationaljournalofresearch.org

# Take a Look over Privacy preserving Data Publishing Techniques

**Prof.Rupali A. Wankhede**
Tulsiramji Gaikwad Patil College Of Engg. & Technology
Rupali1.wankhede@gmail.com

## ABSTRACT

*The process of extraction of useful data from a huge dataset is called as data mining. This process also includes the analysis of the extracted data. Interchange between data is needed nowadays. Recently, the process of interchanging and publishing data is become popular, leading to the need for data security and integrity. In data mining there are some challenges related to the security and privacy of the data. In many aggressive situations, the user's data needs it's privacy to be maintained. With the provision of privacy, people can share their social relations with each other. But sometimes, the users' data is entangled by the attacker by compromising the privacy of the data. Thus, privacy preserving is an important process to be scrutinized. The influences of privacy preserving process are important in data mining. There are many ways for ensuring the privacy of the data. This paper gives an excellent survey on privacy preserving techniques. It also reviews the privacy preserving data publishing techniques.*

***Keywords:*** Data mining; Data publishing; Privacy Preserving

## I. INTRODUCTION

Data mining is the integration and blend of various different fields like database, machine learning technology, information retrieval process, knowledge based systems, high performance computing and data visualization techniques, etc. Information of the organisations should be handled with attention, and the process of data mining plays a vital role in handling this information. Thus, data mining process is gathering the attention of many industries.

There is a huge amount of data available which includes useful information and knowledge. This information apprehension is useful for various real-time applications like market analysis, fraud detection, customer retention, etc [1].

In many situations, for improving the efficiency in business, many companies use a big amount of data for developing relationships and correlations between the data; provided by the data mining process. Also there are many organizations that have a widespread data, increasing a threat to the data privacy. Hence, privacy preservation is assimilated as a functional process with data mining. Privacy is provided to the gathered information and knowledge. Whenever a user shares or stores the data or information, it needs to be assured with the preservation of privacy of the data. Also while publishing the data; privacy is important to be provided.

The utilization of huge amount of data can be effective by preserving privacy of the data. As the data is stored in electronic format, no individual is disturbed. The privacy is preserved while collecting the data and while mining the data.

As discussed earlier, the data is scattered around which needs to be gathered. So the responsibility of the data publisher is to gather the information and to transform it into a standard format. This format is suitable for data recipient. Also while the data publisher processes the information, it must be preserved before publishing.

During this review, we first present the meaning of privacy preserving data publishing technique in section II, In section III the classification of Privacy Preservation Techniques has been discussed. In section IV, the focus of survey is to give various Privacy techniques that are available. We state our conclusion and future directions in section V.

## II. PRIVACY PRESERVING DATA PUBLISHING

There are two phases of privacy preserving data publishing: 1) Data collection 2) Data publish phase. There are three kinds of roles : 1) Data owner 2) Data publisher 3) Data recipient. The relationship between two phases and three roles involved in PPDP.

- Data collection phase: The data publisher collects dataset from data owner.
- Data publishing phase : The data publisher sends the processed dataset to data recipient.

The raw dataset is not directly sent to the data recipient. First it have to sent to data publisher and then to data recipient.

In [2], Data publisher divided into two types. 1) Untrusted model : Data publisher is tricky who is more likely to gain privacy from dataset 2) Trusted model : Data is safe and without any risk . The data publisher is reliable. Owing to the difference of data publishing scenarios affected by requirements to data publisher and by varying assumptions, data recipients purposes. We discussed four scenario are described in Table 1.

Table 1: Data publishing Scenario

| | | |
|---|---|---|
| **First Scenario** | Non-expert data publisher | Data publisher does not need to have specific knowledge about research fields. What they need to do is make data be published satisfying the requirements of data utility and information privacy |
| **Second scenario** | The data recipient could be an attacker | This scenario is more commonly-accepted and many proposed solutions make it as the requisite hypothesis. |
| **Third Scenario** | The publish data is not the data mining result | It indicates that dataset provided by data publisher in this scenario is not merely for data mining. That is to say, published dataset is not data mining result. |
| **Fourth Scenario** | Truthfulness at record level | Data publisher should guarantee the authenticity of data to be published whatever processing |

| | | |
|---|---|---|
| | | methods will be used. Thus, randomization and perturbation cannot meet the requirements in this scenario. |

## III. CLASSIFICATION OF PRIVACY PRESERVING TECHNIQUES

There are many techniques discussed in this section. Some techniques work on the original data for retaining the data sensitivity. The classification of PPDP is done in different categories as follows:

### 1. Randomization method

This Randomization process is used to add noise to the source data in order to mask attributes from disclosure [3,4]. There are different ways of randomization. Additive randomization is the simplest form of randomization. Let $X=\{x1,x2, . . .,xn\};$ where $X$ be a set of data records. The set $X$ contains data elements. Noise is taken from probability distribution $f(y)$ and are denoted by $y1,y2,\ldots , yn$. Various techniques related to randomization is researched by authors [5,6]. The accuracy of privacy preservation depends on how large the distribution y would be and the right amount of randomization. It can be done by multiplying or adding noise [7]. but the drawback is results are approximate and has huge information loss.

### 2. Data Swapping

In this method, the statistical inference of the relation is maintain in order to preserve privacy by swapping the values of records [8].

### 3. Cryptographic approach

Revolution of communication via internet has forced several areas one such is Distributed Data Mining. The data is distributed on multiple sites. whenever mining is required it must be retrieved securely[9,10,11]. There are two advantages of this approach. first, Well defined model for privacy and second, lots of algorithms are available to implement cryptographic methods.

### 4. Anonymization Approach

The process of anonymization is a very well known approach for preserving sensitivity of the data, before it is published. This process includes the modification to be made to the contents of the record before the data is published. The standard form of relational schema used by the data publisher is given as:

**R(U_ID,Q_ID_1,Q_ID_2,…,NQ_ID_1,NQ_ID_2,…,SV_1, SV_2,..)**

Here U_ID is the user identification values which are explicit values and are used directly for deducing the identity of the individual. Social security number is such an example of U_ID that is used to regain the identity of an individual. The other type of identifier is the Q_ID which is known as Quasi Identifiers. These identifiers can be used by a attacker for linking the value to the external database for gaining the identity of an individual. Examples of such identifiers are ZIP code and DOB. The pseudo SQL query is given as follows:

```
SELECT *
FROM VOTERS_TABLE AS V
WHERE V.ZIP=’&ZIP’ AND
V.GENDER=’&GENDER’ AND
V.DOB=’&DOB’;
```

Here SV stands for sensitive values. These sensitive values are used for mining and statistical analysis. Example of SV is, in many medical records, the sensitive values can be the disease of the patient. NQ_ID stands for non quasi identifier. These identifiers belong to the any of the categories mentioned above. These values are published only if they are related to data mining. For preventing the divulgence of the information caused by the attacker, the attacker modifies the relation R to R1 with the following schema:

**Rl(Q_ID_1,Q_ID_2,…,NQ_ID_1,NQ_ID_2,…,SV_1, SV_2,..)**

In the above schema, the U_ID in relation R1 is removed and the anonymization of Q_ID is done such that it supports the privacy model representation. This ensures the confidentiality. In this paper, we have discussed on anonymization approach for privacy preserving.

## IV. PRIVACY-PRESERVING DATA PUBLISHING TECHNIQUES

### 1. K-Anonymity

The data with higher dimensionality space must be handled with care. While dealing with such data, various events are detected. This data handling includes analysis of data and publishing the data. For dealing with this challenge a method is proposed known as K-anonymity [12]. The masking of the attribute to the exact value is achieved by applying generalization on K-anonymity [12]. The technique of perturbation applied on K-anonymity was beneficial for the distribution of an individual in aggregated manner rather than the inter attribute relations of an individual. There are other methods proposed on the concepts of K-anonymity known as 2-anonymity and Gaussian cluster methods.

### 2. ℓ-Diversity

By revealing sensitive attributes of an individual, information has been published[13]. K-Anonymization alone is not efficient to prevent the data from background knowledge attack and homogeneity attack. ℓ-Diversity technique outline that sensitive attributes would have at most same frequency. Consider an example, with positive disclosure, if Alice wants to discover Bob, Alice would determine Bob with high-probability distribution. The negative disclosure would happen when an adversary could correctly eliminate some possible value of the sensitive attributes .There could be a minimum difference between the posterior belief and prior belief [13].

### 3. t-closeness

Ravindar Mogili, Anil Prakash found that ℓ -Diversity and K-Anonymity was not used to prevent attribute disclosure [14]. ℓ-Diversity represented sensitive attribute value that was assigned only with certain number of limitations [13]. t-closeness has designed to distribute sensitive attribute with equivalence class. Earth Mover Distance was used to measure the distance between the two probabilistic distributions [14]. Conjuction has been designed to combine statistical analysis and machine learning. Aggregation is used to reduce closeness among the columns.

### 4. Km Anonymity

A technique proposed for anonymized transactional database is Km Anonymity [15]. This technique protects the database from an attacker who is having knowledge of m elements in the transaction [16]. For maintaining the set valued data, this generalization is used on any transaction with K-1 records or on the transaction with identical values. In Km anonymity process, the top down generalization approach is used for recording the transaction records number [15]. This process of Km anonymity will provide privacy against the attackers who identifies m items in a database.

## 5. Distributed K-Anonymity framework (DKA)

It is difficult to directly share the data which is gathered from various sites. Thus, for dealing with this situation the data must be anonymized for generalising it to a specific value. A new technique with a secured 2-party framework is proposed [17]. In this technique, the computation is carried on multiparty datasets joined from various sites [17]. The technique is called as Distributed K-Anonymity (DKA). It prevents the privacy of an individual's identification by using global anonymization in encrypted format. In fact it provides a secured framework in between two different parties [17]. These two parties agree on global anonymization algorithm and results into a local K-anonymity dataset. The DKA framework provides a protocol where the two parties mutually semi-honest.

## 6. K-Anonymity Clustering

Hierarichal clustering is commonly used clustering method to achieve KAnonymity [14]. To reduce the information distortion Weighted Feature C-means Clustering is used. It partition all the records into equivalence class and use class merging mechanism [18] to merge the class. The numerical values of quasi identifier were used to evaluate the Weighted Feature C-means Clustering technique. The authors also try to provide the dissimilarity evaluated approach which would take different types of feature values for class merging mechanism.

## 7. R-U Confidentiality Map

Normally, Anonymization would cause loss in potential utility gain. The most important factor that should be balanced are Risk(R) and Utility (U). The fundamental characteristic of generalization and bucketization was compared [19]. Privacy trade-off utility have fixed the privacy requirement with various privacy parameters and produceda n anonymized dataset. Generalization would work on three methods such as Apriori Anonymization (AA), the Constrained Based Anonymization Transaction (COAT) and Privacy-Constrained cluster Based Transaction Algorithm (PCTA).These three methods were used to maintain the association between Risk(R)-Utility (U). The trade-off was gained by combining the above three methods as (AA & COAT) and (COAT & PCTA) [19]. It was essential to maintain the histogram H for the occurrence of each sensitive item in the transaction database [19].

As per the authors [19], protection of the data requires:

(i) Security, to ensure that there was no loss of the stored information.

(ii) Confidentiality, to ensure that no one can drop data while the data were transmitted between authorized users.

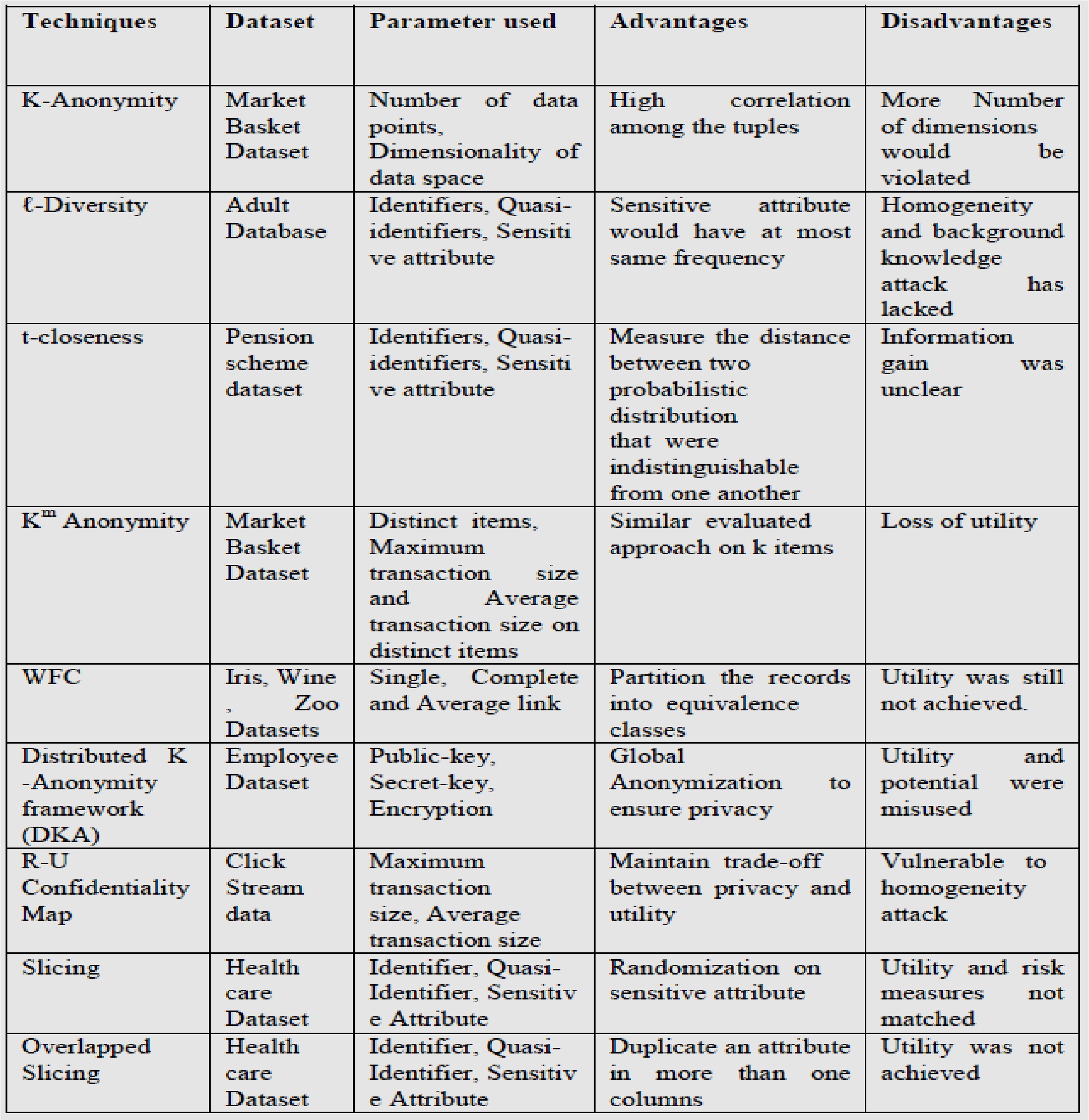
(iii) Anonymity, to ensure private and sensitive information about an individual was not disclosed when a record was released.

## 8. Slicing

K-Anonymity do not take guarantee the background knowledge attack of an adversary [20]. The data was partitioned in both vertical and horizontal direction to preserve the privacy by permuting the sensitive attribute [19].Slicing is popular technique which could be formalized by comparing bucketization [21] and generalization.

## 9. Overlapped Slicing

The problem is divided into the sub problems, which were reused several times. Overlapped slicing are used to duplicate an attribute in several column. Every column results in more attribute correlations [22]. By using Duplication of attributes in several columns, Vertical and Horizontal partition was performed. In vertical partitioning, attributes were correlated and In horizontal partitioning tuples were grouped together. Sensitive attributes are placed in each column and random permutation are applied on sensitive attributes column [22]. The authors tried to provide protection against membership disclosure by distinguishing the original tuples from the fake tuples where the disclosure risk has been occurred. In the overlapped slicing table the matching strategy is very low. Overlapped slicing enhanced with slicing provided the membership disclosure protection. But, it leads to more attribute correlations and there would be a secrecy loss of privacy in some extent.



| 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 |
| WFC | Iris, Wine , Zoo Datasets | Single, Complete and Average link | Partition the records into equivalence classes | Utility was still not achieved. |
| Distributed K -Anonymity framework (DKA) | Employee Dataset | Public-key, Secret-key, Encryption | Global Anonymization to ensure privacy | Utility and potential were misused |
| R-U Confidentiality Map | Click Stream data | Maximum transaction size, Average transaction size | Maintain trade-off between privacy and utility | Vulnerable to homogeneity attack |
| Slicing | Health care Dataset | Identifier, Quasi-Identifier, Sensitive Attribute | Randomization on sensitive attribute | Utility and risk measures not matched |
| Overlapped Slicing | Health care Dataset | Identifier, Quasi-Identifier, Sensitive Attribute | Duplicate an attribute in more than one columns | Utility was not achieved |

Table 1 : Comparative Study

## V. CONCLUSION

This paper, described about different data publishing techniques to preserve the privacy in data mining. Because of large collection of information, it is necessary to maintain the Privacy. The main task of data publisher is to fetch the information from various location and convert it into some standard format suitable for data recipient. Data publisher have to preserve the sensitive data before it has been published. This review paper helps to give future direction towards my future research.

## REFERENCES

[1] Han Jiawei, M Kamber. Data Mining: Concepts and Techniques, Beijing: China Machine Press, 2006, pp.1-40.

[2] Verykios V S, Bertino E, Fovino I N, Provenza L P, Saygin Y, Theodoridis Y. State-of-the-art in privacy preserving data mining, ACM SIGMOD Record, 2004.

[3] Agrawal D. On the Design and Quantification of Privacy- Preserving Data Mining Algorithms, ACM PODS Conference, 2002.

[4] Warner S L. Randomized response: A survey technique for eliminating evasive answer bias. Journal of the American Statistical Association,1965.

[5] Zhang P, Tong Y, Tang S, Yang D. Privacy-Preserving Naive Bayes Classifier, Lecture Notes in Computer Science, 2005, Vol 3584.

[6] Zhu Y, Liu L. Optimal Randomization for Privacy- Preserving DataMining, ACM KDD Conference, 2004.

[7] Agrawal R, & Srikant R. Privacy preserving data mining, Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'00), Dallas, 2000.

[8] Fienberg S,McIntyre J. Data Swapping: Variations on a Theme by Dalenius and Reiss, Technical Report, National Institute of Statistical Sciences, 2003.

[9] Pinkas B. CryptographicTechniques for Privacy-PreservingDataMining, ACM SIGKDD Explorations, 2002.

[10] Laur, H Lipmaa, and T Mieliainen. Cryptographically private support vector machines, In Twelfth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2006, pp. 618-624.

[11] Ke Wang, Benjamin C M, Fung and Philip S Yu. Template based privacy preservation in classification problems, In ICDM, 2005, pp. 466-473.

[12] Charu C. Aggarwal, (2005), ''On k-Anonymity and the Curse of Dimensionality", Proceedings of the 31st VLDB Conference, Trondheim, Norway, pp.901-909

[13] Ashwin Machanavajjhala , Daniel Kifer,Johannes Gehrke, Muthuramakrishnan Venkita Subramanian, (2006)," ℓ-Diversity : Privacy Beyond K-Anonymity", Proc.International conference on Data Engineering.(ICDE),pp.24.

[14] Anil Prakash, Ravindar Mogili ,(2012),‘‘Privacy Preservation Measure using t-closeness with combined l-diversity and k-anonymity”, International Journal of Advanced Research in Computer Science and Electronics Engineering (IJARC SEE)Volume 1, Issue 8,pp:28-33

[15] Yeye He, Jeffery Naughton .F, (2009), “Anonymization of Set Valued Data via Top Down Local Generalization”, Proc. International Conference on Very Large Databases (VLDB), pp.934-945.

[16] Tiancheng Li , Jian Zhang , Ian Molloy ,(2012),“Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transaction on KDD.

[17] Wei Jiang , Chris Clifton, (2006)‘‘ A secure distributed framework for achieving kanonymity”, the VLDB Journal , Vol.15, No.4, pp.316-333

[18] Tiancheng Li , Jian Zhang , Ian Molloy ,(2012),“Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transaction on KDD.

[19] Ravindra S, Wanjari Prof .Devi,(2013), “Improving the implementation of new approach for Data Privacy Preserving in Data Mining using slicing”. International Journal of Modern Engineering Research (IJMER), Vol. 3, Issue. 3.

[20] Agarwa, Srikan R., (2000) ‘‘Privacy Preserving Data Mining”, In Proc. ACM SIGMO, conference on management of data (SIGMOD’00), Dallas, TX,pp.439-450.

[21] Tiancheng Li , Jian Zhang , Ian Molloy ,(2012),“Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transaction on KDD.

[22] B.Vani, D.Jayanthi, (2013), “Efficient Approach for Privacy Preserving Microdata Publishing Using Slicing” IJRCTT.