International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.3, No.12, December 2014 DOI:10.15693/ijaist/2014.v3i12.14-23

A Survey of Latest Developments in Privacy Preserving Data Publishing

Ashoka K *
Department of CS&E
BIET, Davangere, Karnataka, India

Dr. Poornima B.
Prof. and Head, Departmen of IS&E,
BIET,Davangere, Karnataka, India

ABSTRACT: **Today most of the governments, public/private sector organizations and individuals are actively collecting digital information in large databases. Detailed person specific data may often contain sensitive information about individuals. While sharing such information one has to protect the violation of individual privacy. Privacy Preserving Data Publishing (PPDP) provides techniques and tools for publishing useful information while preserving data privacy. The complexity of its representation and the requirements of the current industry have driven a lot of research in this direction. Here in this paper we provide a brief review of various methods for Privacy Preserving Data Publishing. We have also highlighted on recent research on anonymization and discussed different attacks that may take place in the process of anonymization.**

Keywords: Privacy Preserving Data Publishing, Anonymization, Data Mining

I.INTRODUCTION

Data Mining is the process of extracting potentially useful, interesting and previously unknown information from huge amount of data. Sometimes it is also called as Knowledge Discovery from Data (KDD). This knowledge based decision making process is used by many top level executives for statistical or experimental analysis. Today Data Mining has been used successfully in various domains like Market Prediction, Medical Data Analysis, Weather Forecasting, Financial Fraud Detection and also in counter attacking Terrorism. Because of government regulations or for the mutual benefits the data will be published/shared among various parties. Typically the data will be collected from different locations in different format, and converted into the format that is suitable to store in Data Warehouse.

In this scenario the Data Warehouse is the data recipient who receives data from multiple data publishers. The data publisher usually an independent organization collects data from the actual users. Data publisher share data for research purpose or for mutual benefit or due to policy decisions by the government. Detailed person specific data collected by data publisher may contain sensitive information about individuals. Sharing of such information could potentially violate individual privacy. For example, Red Cross Blood Transfusion Service (BTS) is an organization that collects and examine the blood from the donors and distribute the blood to different public hospitals. Government Health Agency in United States of America periodically collects patient's data from public hospitals that contains patient specific medical data. This patient specific medical data is shared with Red Cross Blood Transfusion Service (BTS) for the purpose of auditing and data analysis which can improve the estimated future blood consumption at different hospitals and also makes recommendations on the blood usage medical cases.

Here the patient's privacy must be preserved while sharing data between Government Health Agency and the Red Cross BTS. Similarly AOL an American Online web service provider release its data containing the details of searches made by the individuals for research purpose. These data may have personal identification details which can be used for detecting individuals. Therefore privacy of the individuals is of great concern and is becoming an important chore of research [1]. The privacy is also more important in the current industry as most of the organizations store sensitive information about customers or the business related information. This data can be linked with external databases to retrieve the sensitivity of individuals. The present techniques and methods concentrates on policies and procedures that restrict the access of sensitive information in published data by Anonymization or by swapping. These techniques may result in huge information loss or greater data distortion which will affect the efficiency of data mining algorithms. There is a tradeoff among privacy and data utility, if privacy is high the data utility is low and vice versa.

The task of utmost important is to develop methods and tools for publishing data in hostile environment so that the published data remain practically useful without

revealing individual's sensitive information. This undertaking is called as Privacy Preserving Data Mining (PPDM).In the past few years the research community has contributed several methods and techniques for Privacy Preserving Data Mining. Majority of the research in this area also come from statistics, economics, Big Data Analytics and Cryptography. An initial survey on different methods of PPDM can be found in [2, 3]. There are various directions for implementing PPDM. Randomization, Cryptographic Techniques and PPDP. In the past research there is no clear distinction betweenPPDM and PPDP, but in recent research PPDP is different from PPDM in several ways.

i) PPDP deals with techniques for publishing data, not techniques for data mining. In deed it is expected that conventional data mining techniques are applied on the published data.
ii) The truthfulness of the data is not maintained in PPDM as it uses randomization or the Cryptographic techniques. The truthfulness is maintained in PPDP.
iii) PPDM focusses in performing some data mining task on the data where as PPDP doesn't perform the actual data mining tasks, but concentrates on how to publish the data so that the anonymous data is useful for data mining.

A. PPDP Model

PPDP model can be represented as shown in figure-1. In the lower layer there is the data publisher and the upper layer has the data recipient.

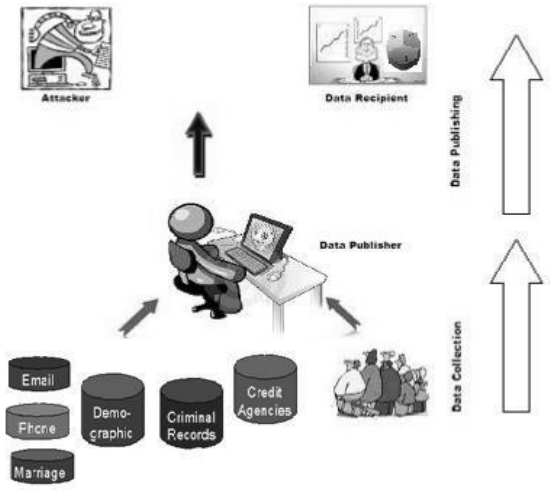

FIG-1: PPDP architecture

The overall model is divided into data collection phase and data publishing phase. In data collection phase the publisher collects data in its original form from the record holders. In the data publishing phase the data publisher releases the data to the data recipients by ensuring privacy. In this model we assume that the data publisher is trusted one with whom the record owners share their sensitive information. The data recipients are untrusted ones, so that sensitive data must be protected.

B. Outline

In this survey paper we focus on different privacy preserving techniques and algorithms. We have also discus about demerits of each of the techniques and algorithms. The rest of the paper is organized as follows. In section 2 some past research on PPDP and various classification of privacy preserving techniques are discussed with their notion and representation. Different types of data disclosures and attacks are discussed in section 3. Recent developments in anonymization techniques are discussed in section 4. Finally, the conclusion and future directions of the research are discussed in section 5.

II.EARLY RESEARCH ON PPDP

The main idea in PPDP is to develop methods and techniques that preserve the sensitivity of personal data. There are several techniques discussed in the past research, which are classified into following categories.

A. Data Perturbation

It is also called as Randomization method that adds noise component to the original data in order to disguise attributes from disclosure [4]. This approach can be classified into two main categories; the probability distribution approach and the value distortion approach. In probability distribution approach the original data is replaced by sample from the same (or estimated) distribution [5]. For example Let $X=\{x_1, x_2, \ldots x_n\}$ be the set of data records. For every element x_i of X, a noise is added which is the probability distribution f(y) and are denoted by $y_1, y_2, \ldots y_n$. The resulting distribution may be represented as $x_1+y_1, x_2+y_2, \ldots x_n+y_n$. Several perturbation techniques are discussed in literature [6, 7]. The value distortion approach perturbs data items or attributes directly by either additive noise, multiplicative noise or similar randomization

procedures [8].The additive and multiplicative perturbation normally applied to numerical data only. Perturbation for categorical data was first considered in [9]. Here a randomized response method was developed for the data collection purpose through interviews. The authors in [10] discuses about categorical data perturbation for association rule mining.

The accuracy of data perturbation techniques for privacy preserving depends on how large the distribution y would be and the correct amount of randomization. The disadvantage of this techniques is that results are approximate and has huge information loss.

B. Data Swapping

In this method the data tables are anonymized by exchanging values of sensitive attributes among individual records. This swapping maintains the low-order frequency counts or marginals in order to maintain privacy [10]. A well diverse refinements and applications of data swapping are discussed in [11].

C. Cryptographic Approach

Information integration is one of the active area of database research. Also the advancement in communication and internet technologies leads to Distributed Data Mining. In this scenario, data is distributed in multiple sites and the data must be securely retrieved for mining purpose [12, 13]. Another concept based on cryptographic approach is the Secure Multiparty Computation (SMC). It allows sharing of computed result (ex. A classification result) without sharing the actual data. Several cryptographic protocols like circuit evaluation protocol, commutative encryption, homomarphic encryption and oblivious transfer, serve as the building blocks of SMC. It can be shown by using a generic circuit evaluation protocol that any function which is expressed by an arithmetic circuit is privately computable.

But for large data sets it is infeasible due to communication and computational complexity. A collection of SMC tools and an overview of the state-of-the-art privacy preserving data mining techniques is presented in [14, 15].

D. Anonymization Approach

Anonymization is the process of removal of identifying information from data for protecting privacy of the data while allowing the modified data to be used for analysis purpose. It is the most common approach to PPDP that hides the identity and/or the sensitive data of record owners by assuming that anonymized data should be useful for data analysis. In Privacy Preserving Data Publishing, the data in its most basic form has the following relational schema, which is used by data publisher.

R (Explicit_Id, Quasi_Id, Sensitive_Attributes, Non_sensitive_Attributes)

Where Explicit_Id is a set of attributes that can be directly used to identify the individuals (Record owners). For example Social Security Number (SSN) can be used to access information of a person in USA. Quasi_Id is a set of attributes that can potentially identify record owners. These Quasi_Identifiers can be used by attackers to link this values to externally available database to retrieve the identity of the individual. For example gender, age and zip code can be used to link with external database like voters-list to identify the person. Sensitive_Attributes consist of person specific sensitive information like disease, income, and disability status. These attributes are useful for the purpose of data mining and statistical analysis. All the attributes that doesn't fall into the previous categories are called as non-sensitive_Attributes. They are published as it is if they are relevant for data mining. To prevent the disclosure of information, the data publisher will modify the relation R to R` as

R` (Quasi_Id`, Sensitive_Attributes, Non_sensitive_Attributes)

In R` Explicit_Id is removed and Quasi_Id are anonymized so that it satisfy the privacy and ensures the confidentiality. Alternatively, the anonymization operations my add noise to the original table R, or generates a synthetic data table R` based on the statistical properties of the original table R. In this paper we focus more on anonymization approaches for privacy preservation and provide our insights onto this topic.

III.ANONYMIZATION MODELS AND PRIVACY THREATS

In this section we explore the different representations of anonymization and the privacy threats on each of these algorithms. Based on the attack principle, we can broadly classify the privacy models into two categories. The first category threats occurs when an attacker can be able to link published record data to an external database and identify victim's

record. This type of attack is called as record linkage, attribute linkage, and table linkage. The data table is said to be privacy preserving if it can efficiently prevent the attacker from successfully performing these linkages. In second category the attacker knows some background knowledge to identify the sensitive information. We call this as probabilistic linkage attack.

A. Record Linkage Attack

In the Record Linkage Attack some value x on quasi identifier Quasi_Id identifies a small number of records in the published table R. In this circumstance the person having the value x is susceptible to being linked to some small number of records in R. Here the attacker faces only a small number of possibilities to identify victim's record, with the help of some additional knowledge. For example consider table 3.1 containing patient's record of a hospital published for the research purpose. Here the explicit identifiers like 'Name' and 'SSN' are removed. If the research center has access to the externally available voters data as shown in table 3.2

Name	Job	Age	Sex
Aruna	Engineer	35	F
Bhavana	Doctor	34	F
Chaithra	Doctor	31	F
David	Attender	40	M
Eshwar	Attender	40	M
Fred	Driver	36	M
Garry	Lawyer	39	M
Heena	Engineer	34	F

Table 3.1 Published patients' data by Hospitals

Job	Age	Sex	Disease
Engineer	35	F	HIV
Engineer	34	F	Flu
Lawyer	39	M	HIV
Attender	40	M	Arthritis
Attender	40	M	Cancer
Doctor	34	F	HIV
Doctor	31	F	Malaria

Table 3.2 external available voters' data

Joining these two tables on quasi identifier attributes Job, Age and Sex may link the identity of any person to his/her disease. For example, Garry a male lawyer of 39 years old is identified as a HIV patient with qid=<Lawyer, Male, 39> after join.

A.1 K-Anonymity:To thwart Record Linkage Attack, Samarthi and Sweeney [16, 17, 18] proposed the concept of 'k-Anonymity', a property that avoids possible re-identification of the record owners from published data. If one record in the table has some value for q-id, at least k-1 other records also have the value q-id. A table satisfying this constraint is called as k-Anonymous table. In a k-Anonymous table, each individual record is indistinguishable from at least k-1 other records with respect to QID.

Two important methods for implementing k-Anonymity on published data are Generalization or Suppression. Each Generalization or Suppression operation pelts some details in QID. For categorical attributes a specific value can be modified to a general value according to some predefined hierarchy. For example in figure 3.1, the parent node White-collar is more general than the children nodes Doctor and Engineer. The root node Any-Job represents the most general value for the attribute Job.

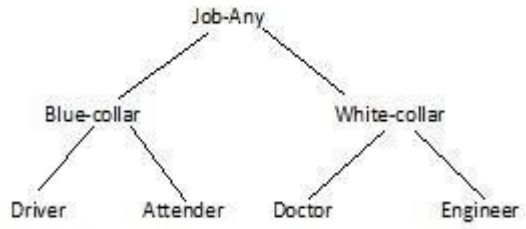

Fig 3.1 Generalization hierarchy for the attribute Job

For numerical attributes, exact values can be replaced by interval, or range of values. Figure 3.2 shows the generalization of attribute Age.

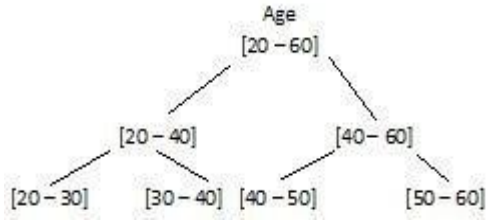

Fig 3.2 Generalization hierarchy for Age.

There are several generalization schemes discussed in the literature, like Full-Domain Generalization [19], Sub-Tree Generalization [20, 21, 22], Sibling Generalization [19], Cell Generalization [23, 24] and Multi-Dimensional Generalization scheme [25]. In Full-Domain Generalization all values of an attribute are

generalized to the same level in the hierarchy. In Sub-Tree Generalization either all child values or none are generalized at a non-leaf node. The process of sibling generalization scheme is same as that of Sub-Tree Generalization, except that few siblings may leftungeneralized. The above mentioned schemes are called as global recording because, if a value is generalized, every instance of it are generalized. The Cell Generalization scheme also called as local recording, where some instances of a value may remain ungeneralized while the other instances are generalized. Multi-Dimensional generalization method replaces entire record with another record. Overall Full-Domain generalization has the smaller search space and larger data distortion than the other methods because, each value is generalized to the same level in the hierarchy. Cell based generalization has the largest search space but with least distortion.

There are mainly three different suppression schemes. Record suppression [26] refers to suppressing the complete record. Value suppression [27] refers to suppressing all the instances of a value in a table. Cell suppression refers to suppressing few instances of a given value in a table. An improvement over k-anonymity is given by (X-Y)-anonymity [28] and MultiRelational k-anonymity [29] techniques. For disjoint set of attributes X and Y, (X-Y)-anonymity specifies that every value on X is linked to at least k distinct values on Y and this concept was motivated by sequential releases of data. When a database consist of multiple relational tables, Anonymization for these tables was done by MuliRelational k-anonymity technique.

Even though k-anonymity overcomes record linkage attack, it suffers from 'Homogeneity Attack' that occurs when the entire QID group has identical sensitive values and 'Background Knowledge Attack' that occurs when the attacker has some knowledge about sensitive attributes.

B. Attribute Linkage Attack

In Attribute Linkage the attacker may not precisely identify the record of the victim, but can identify his/her sensitive values from the published data, based on the set of sensitive values connected to the group that the victim belongs to. If some sensitive values predominates in a group, it is relatively easy to identify the victim even if k-anonymity is satisfied. For example consider the 3-anonymous patient data in table 3.3. The

the distribution of sensitive attribute values within attacker can say with 66.66% confidence that any woman with white collar job under the age group of [30 – 35] have HIV, because 2 out of 3 females with white-collar job in the age group [30 – 35] have HIV.

Job	Age	Sex	Disease
White-collar	[30-35]	F	Flu
Blue-collar	[35-40]	M	Cancer
White-collar	[30-35]	F	HIV
Blue-collar	[35-40]	M	Hepatitis
White-collar	[30-35]	F	HIV
Blue-collar	[35-40]	M	Flu

Table 3.3 A 3-Anonymous patient's Data

B.1 L-Diversity: To overcome the limitations of k-anonymity and to prevent the attribute linkage attack, Machnavajjhala et al.[30] propose the principle called as l-Diversity that requires each qid group to contain at least l "Well Represented" sensitive values. Here the meaning of "Well Represented" is that there are minimum of l distinct values for the sensitive attribute in each qid group. A 2-diversity patient's data is shown in table 3.4

Job	Age	Sex	Disease
White-collar	[30-35]	F	Flu
White-collar	[30-35]	F	HIV
White-collar	[30-35]	F	Flu
White-collar	[30-35]	F	Cancer
Blue-collar	[35-40]	M	Hepatitis
Blue-collar	[35-40]	M	HIV

Table 3.4 A 2-Diversity patient's Data

Even though l-Diversity overcomes attribute linkage attack, it suffers from (i) Skewness attack and (ii) Similarity attack. Skewness attack occurs when the attacker, based on the frequency distribution of the sensitive value he can derive it. Similarity attack occurs when all the sensitive attribute in a quasi-group are different, but semantically similar. For example cancer, Malignancy, Sarcoma, Tumor are semantically one and the same.

B.2 (k,e)-Anonymity: The k-anonymity and its variants assume categorical attributes. For numerical attributes such as salary, the concept of (k, e)-Anonymity was proposed in [31]. In (k, e)-Anonymity the set of records are partitioned into groups, such that each group contains a minimum of k different sensitive values with a minimum of e range. But (k, e)-Anonymity overlooks certain sub range λ. If some sensitive attribute values occur repeatedly with a sub range of λ, then the

opponent could confidentially infer the sub range in the group.

B.3LKC- Privacy: For high dimensional data, that is, when the number of QID attributes are large, majority of the data have to be suppressed to achieve k-anonymity. So there exists a significant degrade in the data quality. To overcome this problem authors in [32] proposed the concept of LKC-privacy, which is based on the concept of attacker's prior knowledge about victim's record is limited
to at most L values of the QID attributes. The LKC-privacy ensures that every combination of values in QID with highest length L in the data table T is shared by not less than K number of record, with the confidence of inferring the sensitive values is at most C, Where L, K, C are thresh hold values specified by data owners.

B.4 t-Closeness: To prevent the skewness attack and the similarity attack the author in [33] introduced t-Closeness. In this method, privacy is measured by adversary's information gain about sensitive attribute. Here the information gain is the difference between the beliefs about distribution of sensitive attributes before and after anonymization. This information gain (i.e. the difference) should not be more than certain thresh hold t. This is achieved by making the distribution of sensitive values in the publicly available database, same as that of the distribution of sensitive values in every QID group. t-Closeness requires Earth Mover Distance (EMD) function to measure the intimacy between the distribution of sensitive values in the original table and the anonymized table, and the closeness should be within t.

But t-Closeness has many drawbacks and flaws. There is no standard technique to impose t-Closeness. Different protection levels cannot be stated for different sensitive values. This technique cannot be applied for numerical attributes. Most importantly it greatly degrades the data utility because it needs the distribution of sensitive values to be the same in all qid groups.

B.5 ***Personalized Privacy****:* In this approach every record owner is allowed to specify his own level of privacy [34]. The assumption in this model is that there is a taxonomy tree for each sensitive attribute and the record owner can specify a guarding node in this tree. Violation of record owner's privacy takes place when the opponent is able to identify any domain sensitive value within the subtree of his guarding node. For example the sensitive attribute Disease can have taxonomy tree with Cancer and HIV as child nodes of Severe-Disease. An HIV patient Arun can set the guarding node to Severe-Disease meaning that he allows people to infer that he has sever disease, but not specific type of sever disease. Another HIV patient, Binay, doesn't mind unveiling his medical information, so that he won't set any guarding node for his sensitive attribute. An improvement of this technique is proposed in [35] where the authors use sensitivity flags as guarding elements for privacy preservation and they also considered the distribution of sensitive values in the qid groups. Based on several experiments it is found that this technique result in lower information loss and higher data utility other earlier techniques. But it is unclear that how individual record owners would set their guarding node, which varies from person to person.

C. Table Linkage Attack

In Record linkage and Attribute linkage the opponent assume that the victim's record is in the published table T. Sometimes the presence (or the absence) of the victim's record in T discloses the victim's sensitive information. If a hospital publishes a data table with a particular type of disease, identifying the presence of victim's record in the table discloses his sensitivity. If the opponent can confidently say the presence or the absence of victim's record in the published table, then we can say that 'Table Linkage' has occurred.

For example consider the 3-anonymized table T (table 3.3) published by hospital. If the opponent have access to external voter data (table 3.2), the table linkage on the target victim, for instance Bhavana, on T may occur. The probability that Bhavana is present in T is ¾=0.75, because there are 3 records in T and 4 records in the external voter data containing Qid < white-collar, F, [30-35]>

To overcome table linkage the author in [36] proposed δ-presence that limits the probability of inferring the victim's record within some indicated range δ_{min} to δ_{max}. But this model has an assumption that the data holder and the opponent has access to the same external table, is not to be a practical assumption in some situations.

D. Probabilistic Attack

In probabilistic attack the opponent will not infer sensitive information from the published data set. The opponent doesn't focus on exactly what records, to infer target victim, but concentrates on how his probabilistic beliefs will change after gaining access to the published data. The privacy model for this attack requires to ensure that the change of probabilistic confidence is relatively less after obtaining the published data. Few perceptive notions for probabilistic attack are (C, +)-Isolation [37], ε-differential privacy [38], (d, γ)-privacy [39], distributional privacy [40] etc. Different privacy preserving model has its own features determined by the spiteful attacks. Therefore the connected algorithms which belong to a specific privacy model are customized and targeted to overcome particular attack situation.

IV.SOME RECENT DEVELOPMENTS IN PPDP

Current research in the field of PPDP focus on Anonymization techniques. Most of them are an improvement/enhancement of the earlier PPDP methods that we discussed in this paper. In this section we briefly overview these recent developments.

A. Slicing Approach

Slicing [41] is a novel anonymization technique that partitions the data set both horizontally and vertically. In vertical partitioning the attributes are grouped into columns based on the associations among the attributes that is, highly associated subsets of attributes are put in one column. In horizontal partitioning the tuples are grouped into buckets based on q_id values. Finally, values in each column are arbitrarily permuted within each bucket so that the linking between different columns are eliminated. The essence of slicing is to disrupt the association across columns and preserve association inside each column. As it groups highly correlated attributes together, better utility is maintained and also it breaks the associations among uncorrelated attributes that enhances the privacy. Authors in [42] propose a similar concept called as Break-Merge. In Break-Merge the anonymized table is split into quasi-identifiers table and sensitive attributes table, which divides quasi identifier values and sensitive values in the anonymized table.

The drawback of slicing is that while performing random permutation within bucket, if more number of similar attribute values and the sensitive attribute values are present in different tuples,may result in original tuples. The utility of the data is lost by generation the counterfeit tuples. Even thoughthese techniques can be applicable to any number of sensitive attributes, the accuracy of user/data miners query against quasi-identifiers and more than one sensitive attributes is reduced. Because when the anonymized table is splitted, the probability of inference also increases. Once the database table is splitted the user/data miner has to join these splitted tables to perform data mining tasks, which will consume extra processor cycles.

To overcome these drawbacks authors in [43] proposed an enhanced slicing of two types. First one is suppression slicing in which slicing is performed after suppressing minimal attribute values in the tuples. Next one is Mondrian slicing that perform random permutation with all the buckets not inside the single bucket.

B. Utility Specification Technique

None of the abovementioned techniques have specific quality connected requirements of applications in their anonymization mechanisms. Authors in [44] suggested a technique called as PPDP based on utility specification that includes a method to specify haphazard requirements of general applications. The utility requirements of a data mining application are specified in the form of nested ordered list and may be used to preserve certain attributes and attribute values for some data mining tasks, such as decision tree learning.

Here the user specified attribute values are preserved while other attributes are generalized. If the distribution of sensitive attribute is not uniform then the accuracy of this technique will be reduced. The correlation among Quasi Identifiers and Sensitive Attributes is not eradicated, that may leads to privacy breach. As there may be infinite number of requirements and several data mining applications it is not to be practical to restrict on them. If there is a chance to specify correct level of utility/generalization, that may give more accurate result.

C. PPDP for Multiple Sensitive Attributes

All the works discussed above focus on single sensitive attribute. For the case of multiple sensitive attributes authors in [45] suggested a Multiple Sensitive Bucketization (MSB) approach. But the MSB technique is appropriate for micro data with few (2 to 3) sensitive attributes. Data with more number multiple sensitive attributes with MSB algorithm results in higher

suppression ratios. To overcome this problem author in [46] proposed SLOMS (**SL**icing **O**n **M**ultiple **S**ensitive) where the data table with multiple sensitive attributes is vertically partitioned into numerous sensitive attribute tables and one quasi identifier table. Tuples in individual table are partitioned into some equivalence classes. The q_id values of each equivalent class are generalized to the same value to satisfy k-anonymity principle. At the same time sensitive values of each partitioned sensitive tables are sliced and bucketized to satisfy the l-diversity principle.

SLOMS algorithm works better as the suppression ratio and the data distortion for this technique are less when compared to other techniques. But if the correlation among the sensitive attributes is important for the further data mining tasks, this technique fails as it partitions the sensitive attributes. It is not practical to assume the distribution of sensitive attributes as uniform.

V.CONCLUSION AND FUTURE RESEARCH DIRECTIONS

Sharing of data/knowledge is essential part of many individuals and organizations. As the data is distributed across various locations in different format, Privacy Preserving Data Publishing is a hope full methodology for preserving individual's privacy and defending sensitive information. In this survey we presented several types of attacks for Privacy Preserving Data Publishing and also discussed/reviewed various methods and techniques to circumvent from these attacks. We gave more emphasis on anonymization techniques. This paper can be used for researchers as quick reference of PPDP techniques.

Our future research directions on PPDP are shown below.

i) Existing research focuses on anonymizing a single sensitive attribute but for the case of multiple sensitive attributes, still there is a need for effective anonymization algorithm.

ii) Providing the privacy for continuous data streams is still in its infant state. Hence more stress need to be given on developing efficient algorithms for privacy preserving on continuous data streams.

iii) There are several performance measurement metrics for the evaluation of the PPDP techniques. We found that for different metrics fits for different consequences of PPDP. A new information metric for most of the PPDP scenarios will be part of the future research. iv) The privacy preserving technology solves only the technical side of the problem but, the nontechnical difficulties of the problem can be effectively solved with multidisciplinary research in collaboration with social scientists, psychologists and public policy makers.

REFERENCES

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

[2] V. S. Verykios, E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin, and Y. Theodoridis. "*State-of-the-art in privacy preserving data mining"*. ACM SIGMOD Record, 3(1):50-57, March 2004.

[3] Yang Xu, Tinghuai Ma, Meili Tang, "*A Survey of Privacy Preserving Data Publishing using Generalization and Suppression",*
Applied Mathematics & Information Sciences An International Journal, Vol. 3, pp 1103-1116 (2014)

[4] D. Agrawal and C. C. Aggarwal. "*On the design and quantification of privacy preserving data mining algorithms"*. In Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, pages 247–255, Santa Barbara, California, USA, May 21-23 2001.

[5] C.K. Liew, U.J. Choi, and C.J. Liew, "*A Data Distortion by Probability Distribution"*, ACM Trans. Database Systems (TODS), vol. 10, no. 3, pp. 395-411, 1985.

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

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

[8] Agrawal R. and Srikant R. "*Privacy preserving data mining"*, Proceedings of ACM SIGMOD Conference on Management of Data (SIGMOD'00), pp. 439-450, Dallas, 2000.

[9] S. L. Warner. "*Randomized response: A survey technique for eliminating evasive answer bias"*. Journal of the American Statistical Association, vol. 60, pp. 63–69, 1965.

[10] T. Dalenius and S.P. Reiss, "*Data-Swapping: A Technique for Disclosure Control"*, J. Statistical Planning and Inference, vol. 6, pp. 73-85, 1982.

[11] S.E. Fienberg and J. McIntyre, "*Data Swapping: Variations on a Theme by Dalenius and Reiss"*, technical report, Nat'l Inst. Of Statistical Sciences, Research Triangle Park, NC, 2003.

[12] B. Pinkas. "*Cryptographic techniques for privacy-preserving data mining"*. ACM SIGKDD Explorations Newsletter, 4(2):12–19, January 2002.

[13] 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.

[14] C. Clifton, M. Kantarcioglu, J. Vaidya, X. Lin, and M. Zhu, "*Tools for Privacy Preserving Distributed Data Mining",* ACM SIGKDD Explorations, vol. 4, no. 2, 2003.

[15] V.S. Verykios, E. Bertino, I.N. Fovino, L.P. Provenza, Y. Saygin, and Y. Theodoridis, "*State-of-the-Art in Privacy Preserving Data Mining",* ACM SIGMOD Record, vol. 3, no. 1, pp. 50-57, Mar. 2004.

[16] P. Samarati. "*Protecting respondents' identities in microdata release"*. IEEE Transactions on Knowledge and Data Engineering (TKDE), 13(6):1010–1027, 2001.

[17] P. Samarati and L. Sweeney. "*Generalizing data to provide anonymity when disclosing information"*. In Proc. of the 17th ACM SIGACTSIGMOD-SIGART Symposium on Principles of Database Systems (PODS), page 188, Seattle, WA, June 1998.

[18] P. Samarati and L. Sweeney. "*Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression"*. Technical report, SRI International, March 1998.

[19] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. "*Incognito: Efficient full-domain k-anonymity"*. In Proc. of ACM international Conference on Management of Data (SIGMOD), pages 49–60, Baltimore, ML, 2005.

[20] V. S. Iyengar. "*Transforming data to satisfy privacy constraints"*. In Proc. of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 279–288, Edmonton, AB, Canada, July 2002.

[21] B. C. M. Fung, K. Wang, and P. S. Yu. "*Top-down specialization for information and privacy preservation"*. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), pages 205–216, Tokyo, Japan, April 2005.

[22] B. C. M. Fung, KeWang, and P. S. Yu."*Anonymizing classification data for privacy preservation"*. IEEE Transactions on Knowledge and Data Engineering (TKDE), 19(5):711–725, May 2007.

[23] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W. C. Fu. "*Utility based Anonymization using local recoding"*. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, August 2006.

[24] R. C. W. Wong, J. Li., A. W. C. Fu, and K. Wang. "*(α, k)-anonymity: An enhanced k-anonymity model for privacy preserving data publishing"*. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 754– 759, Philadelphia, PA, 2006.

[25] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan. "*Workload-aware anonymization"*. In Proc. of the 12th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), Philadelphia, PA, August 2006.

[26] V. S. Iyengar. "*Transforming data to satisfy privacy constraints"*. In Proc. of the 8th ACM International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 279–288, Edmonton, Canada, July 2002.

[27] K. Wang, B. C. M. Fung, and P. S. Yu. "*Handicapping attacker's confidence: An alternative to k-anonymization"*. Knowledge and Information Systems (KAIS), 11(3):345–368, April 2007.

[28] K. Wang and B. C. M. Fung. "*Anonymizing sequential releases"*. In Proc. of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 414–423, Philadelphia, PA, August 2006.

[29] M. Ercan Nergiz, C. Clifton, and A. Erhan Nergiz. "*Multirelational k-anonymity"*. In Proc. of the 23rd International Conference on Data Engineering (ICDE), pages 1417–1421, Istanbul, Turkey, 2007.

[30] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam. "*l-diversity: Privacy beyond k-anonymity"*. In Proc. of the 22nd IEEE International Conference on Data Engineering (ICDE), Atlanta, GA, 2006.

[31] Q. Zhang, N. Koudas, D. Srivastava, and T. Yu. "*Aggregate query answering on anonymized tables"*. In Proc. of the 23rd IEEE International Conference on Data Engineering (ICDE), April 2007.

[32] N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. K. Lee. "*Anonymizing healthcare data: A case study on the blood transfusion service"*. In Proc. of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (SIGKDD), pages 1285–1294, France, June 2009.

[33] Ninghui. Li, T. Li, and S. Venkatasubramanian. "*t-closeness: Privacy beyond k-anonymity and l-diversity"*. In Proc. of the 21st IEEE International Conference on Data Engineering (ICDE), Istanbul, Turkey, April 2007.

[34] X. Xiao and Y. Tao. "*Personalized privacy preservation"*. In Proc. Of ACM International Conference on Management of Data (SIGMOD), Chicago, IL, 2006.

[35] Kiran, P.; Kumar, S.S.; Hemanth, S.; Kavya, N.P., "*Assignment of SW using statistical based data model in SW-SDF based personal privacy with QIDB-anonymization method"*, 2nd IEEE International Conference on Parallel Distributed and Grid Computing (PDGC), pp.816,821, Dec. 2012

[36] M. Ercan Nergiz, M. Atzori, and C. W. Clifton. "*Hiding the presence of individuals from shared databases"*. In Proc. of ACM International Conference on Management of Data (SIGMOD), pages 665–676, Vancouver, Canada, 2007.

[37] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee. "*Toward privacy in public databases"*. In Proc. of Theory of Cryptography Conference (TCC), pages 363–385, Cambridge, MA, February 2005.

[38] Dwork C., "*Differential privacy"*, In Proceedings of the 33rd International Colloquium on Automata, Languages andProgramming (ICALP), 1-12 (2006).

[39] Rastogi V., Suciu D., Hong S., "*The boundary between privacy and utility in data publishing"*, In Proceedings of the33rd International Conference on Very Large Data Bases (VLDB), 531-542 (2007).

[40] Blum A., Ligett K., Roth A., "*A learning theory approach to non-interactive database privacy"*, In Proceedings of the 40th Annual ACM Symposium on Theory of Computing (STOC), 609-618 (2008).

[41] Tiancheng Li; Ninghui Li; Jian Zhang; Molloy, "*Slicing: A New Approach for Privacy Preserving Data Publishing"*, IEEE Transactions on Knowledge and Data Engineering, vol.24, no.3, pp.561, 574, March 2012

[42] Nadimpalli S.V.; Vatsavayi V.K., "*BM (Break-Merge): An Elegant Approach for Privacy Preserving Data Publishing",* IEEE Third International Conference on Social Computing (SocialCom), pp.1202, 1207, Oct. 2011

[43] Kiruthika S.; Raseen M.M. "*Enhanced slicing models for preserving privacy in data publication",* International Conference on Current Trends in Engineering and Technology (ICCTET), pp. 406, 409, July 2013

[44] Hongwei Tian; Weining Zhang, "*Privacy-Preserving Data Publishing Based on Utility Specification",* International Conference on Social Computing (SocialCom) , pp.114,121, 8-14 Sept. 2013

[45] Jianmin han, Fangwei Luo, Jianfeng Lu, Hao Peng, "*SLOMS: A Privacy Preserving Data Publishing Method for Multiple Sensitive Attributes Micro data",* Journal of software, Vol. 8, No. 12, Dec. 2013

[46] Yang Xiao-chun, Wang Ya-zhe, Wang Bin. "*Privacy Preserving Approaches for Multiple Sensitive Attributes in Data Publishing"*. Chinese journal of computers, 31(04), pp. 574-587, 2008

Authors Profile

Ashoka K. is an assistant professor of Computer Sciences and Engineering Department at Bapuji Institute of Engineering and Technology, Karnataka, India. He received his M.Tech. Degree from Visvesvaraya Technological University, Karnataka, India, in 2004. Now, he is a doctoral candidate of Computer Science & Engineering, Visvesvaraya Technological University. His main research interests are in the areas of Data Processing, Big Data Analytics and Privacy Preserving Data Mining/Publishing.

B. Poornima is a professor and Head of Information Science and Engineering Department at Bapuji Institute of Engineering & Technology, Karnataka, India. She received her M.Tech. (Visvesvaraya Technological University, India, 2001) and PhD (Kuvempu University, India, 2012). Her research interests are in the areas of Data Mining, Big Data Analytics, Fuzzy Systems and Image Processing. She is the principle investigator of AICTE funded RPS project- "Design and Analysis of Efficient Privacy Preserving Data Mining Algorithms".