A Survey on Privacy Preservation in Data Publishing

Christy Thomas
Department of Computer Science
Rajagiri School of Engineering and Technology
Kochi, India
christythms@gmail.com

Diya Thomas
Department of Computer Science
Rajagiri School of Engineering and Technology
Kochi, India
diyat@rajagiritech.ac.in

Abstract— Privacy-maintaining data release is one of the most important challenges in an information system, because of the wide collection of sensitive information on the internet. A number of solutions have been designed for privacy-maintaining data release. This paper provides an inspection of the state-of-the-art methods for privacy protection. The paper discusses novel and powerful privacy definitions which can be categorized into microdata anonymity methods and differential privacy methods for privacy-maintaining data release. The methods include k-anonymity, l-diversity, t-closeness and js-reduce defense. This paper proposes an enhanced method which will prevents sequential background knowledge attack and provides some anonymization also.

Keywords-*:* K-Anonymity; L-Diversity; T-Closeness; JS-Reduce

I. INTRODUCTION

Information is the most significant resource today. Private, public and governmental institutions may often need to collect and publish data. This publishing of data may sometimes lead to mutual advantage. Main profit of large databases is market oriented and research, whether it be economic or scientific. A key problem that occurs in this massive collection of data is confidentiality. So there is a need of privacy.

Data is stored mainly in the form of table. The attribute are mainly divided into explicit identifiers eg. Name, Quasi-identifiers such as age and sensitive attribute eg. Medical data result. Privacy protection techniques prevent the association of sensitive information in sensitive database and explicit identifiers in external database.

Data can be published in two ways. In past data are published mostly in precomputed statistical and tabular form [1]. Such type of data is called macrodata. In a statistical database or microdata, it is often desired to allow query access only to aggregate data, not individual records. Securing such a database is a difficult problem, since intelligent users can use a combination of aggregate queries to derive information about a single individual.

Some common approaches are:

- Only allowing aggregate queries (SUM, COUNT, AVG, STDEV, etc.)
- Rather than returning exact values for sensitive data like income, only return which partition it belongs to (e.g. 35k-40k)
- Return imprecise counts (e.g. rather than 141 records met query, only indicate 130-150 records met it.)
- Don't allow overly selective WHERE clauses

Today data is published as specific stored data called microdata. Microdata increase the flexibility and availability of the information to the user. In order to protect data privacy, explicit identifiers such as name, ssn, address, phone number are removed before publishing data. But the individuals can be de-identified by linking with other publicly available databases. That is microdata can be linked with publically available data such as voter registers. So this removal does not protect privacy completely. Figure 1 shows an example of how two databases medical database and voters register database can be link together[2].

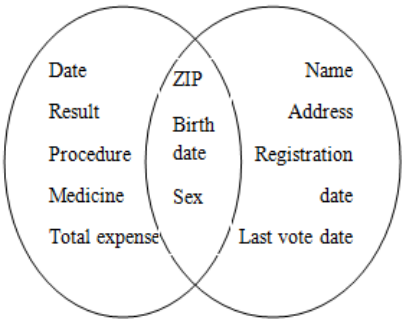

Figure1. Example: De-identification by linking databases [2]

Two type of attacks are identified, identity disclosure and attribute disclosure. Sometimes an individual is linked to a particular record in the released table. This type of attack is known as Identity disclosure. Attribute disclosure occurs when attribute or information about some individuals is revealed. In this attack, characteristic of an individual can be derived more accurate than that possible before releasing data. Identity disclosure often leads to attribute disclosure. Once identity disclosure occurs, individual is re-identified and sensitive attribute is revealed to adversary. But attribute disclosure can be occurs without identity disclosure also. The objective of the data publisher is to limit disclosure to affordable level while publishing data. In this paper, we will provide review of the different techniques for privacy-maintaining data publishing. We will provide new approach for data publishing which considers the background knowledge attack as well as provides anonymization.

The rest of the paper is organized as follows. Section II describes the k-anonymity in which each record is identical to k-1 other record. Section III discusses the second method L-Diversity in which each block contain L different sensitive item. Section IV describes a technique called T-Closeness which provides privacy better than K-Diversity and L-Diversity. JS-Reduce defense which prevents sequential background knowledge will be discussed in section V. Section VI provides proposed algorithm. Finally this paper is concluded in Section VII.

II. K-ANONYMITY

K-anonymity method is proposed by V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati. In k-anonymity, data is published in such a way that each record is identical to k-1 other records. That is data is published in a group of k records [2]. Two techniques which provide data anonymity are generalization and suppression [7]. The peculiarity of generalization and suppression is that they will maintain truthfulness of the information. Generalization is the method of substituting the given attribute with more general value. For this, the concept of domain, which is the set of values that an attribute can accept, is extended to a set of generalized domains. The original domains along with their generalizations are referred to as Dom. Each generalized domain contains generalized values and mapping between each domain and its generalizations

Another method to obtain k-anonymity is suppression which is applied along with generalization. This will moderate the generalization process when tuples with less number of occurrences undergo a greater amount of generalization.Therefore we can say that generalization is applied to attribute (column) level and suppression is applied to tuple (row) level. The generalization and suppression together provides more general table which provide more privacy to the individuals. Table 1 shows an example of 2-anonymous medical database.

TABLE I. 2-ANONYMOUS DATABASE

Age	Gender	ZIP	Result
[51,52]	F	41005*	Chest pain
[51,52]	F	41005*	Obesity
[50,56]	*	41004*	Short breath
[50,56]	*	41004*	Hypertension
[59,61]	M	41105*	Obesity
[59,61]	M	41105*	Short breath

Here each record is belongs to a block of 2 records. The quasi-identifiers of two records, age, gender, ZIP are generalized. Therefore confidentiality of association of particular record in the table and an individual is 0.5. For example an individual Alice of age 51 belongs to 410053. She has chest pain and obesity with a probability 0.5 and 0.5 respectively.

The benefit of this technique is that the method is simple and the individual cannot be easily identified because each individual is belongs to a group of k individual [3]. But the issue with this technique is that the method is not effective when adversary has background knowledge. In that case, the table may prone to background knowledge attack. This method is prone to homogeneity attack which occurs when all the values for a sensitive

attribute within a group of *k* records are same. While k-anonymity prevents against identity disclosure, it is insufficient to protest against attribute disclosure.

III. L-DIVERSITY

L-Diversity method is proposed by Ashwin Machanavajjhala, Johannes Gehrke and Daniel Kifer. L-diversity Principle is that "a q-block is l-diverse if contains at least l values for the sensitive attribute S. A table is l-diverse if every q-block is l-diverse" [4]. The definition means that each block should contain l different sensitive value. The parameter L can be set depends on how much protection the publisher wants. But the above distinct l-diversity does not prevent probabilistic attack. L-Diversity can be instantiated further as follows.

Entropy l-Diversity*:*

The entropy l-diversity is mathematically represented as follows [4].

$$\sum_{s\in S} p(q,s) * \log\left(p(q,s)\right) \geq \log(l)$$

where $p(q,s) = \frac{n(q,s)}{\sum_{s'\in S} n(q,s')}$ is the fraction of tuples in the q-block with sensitive attribute value equal to s. this equation represents the entropy of the l-diversity. The entropy gives the average information contained in table. One point that can be infer from above definition of Entropy l-Diversity is that in order to have entropy l-diversity for each equivalence class, the entropy of the entire table must be at least log(l).

Recursive *(c, l)****-Diversity:***

Let s1.. . sm be the possible values of the sensitive attribute S in a q-block. Sort the counts n(q, s1), ⋯ , n(q, sm) in descending order and name the elements and results in sequence r1, . . . , rm. Let Ti denote the number of times the i^{th} most frequent sensitive value appears in that q-block. Given a constant c, the q-block satisfies recursive (c, l)-diversity if $T1 < c(T_l + T_{r+1} + \cdots + T_m)$. That is, q-block satisfies recursive (c, l) - diversity if we can eliminate one possible sensitive value in the q-block and still have a $(c, l-1)$ -diverse block. A table T satisfies recursive (c, l)-diversity if every q-block satisfies recursive l-diversity. We say that 1-diversity is always satisfied.

TABLE II. SHOWS EXAMPLE OF 3-DIVERSITY

Age	Gender	ZIP	Result
[51,52,56]	*	4100**	Chest pain
[51,52,56]	*	4100**	Obesity
[51,52,56]	*	4100**	Short breath
[50,59,61]	*	4100**	Hypertension
[50,59,61]	*	4110**	Obesity
[50,59,61]	*	4110**	Short breath

The advantage of the method is that l-diversity does not require the knowledge about full distribution non-sensitive and sensitive attributes. L-parameter shield database from adversary. Instance level knowledge is covered automatically. L-Diversity addresses the homogeneity attack and background knowledge attack in the K-anonymity method. One issue of l-diversity is that it is limited in its assumption of adversarial knowledge [9]. It assumes all attributes to be categorical. The adversary either does or does not learn something and does not consider the numerical value of attribute.

IV. T-CLOSENESS

T-closeness is proposed by Ninghui L,i Tiancheng Li and Suresh Venkatasubramanian. This method proposed as solution to attribute disclosure. T-closeness can be defined as follows. Spreading of sensitive attributes in each quasi-identifier group should resemble to their distribution in whole original database [5]. That is the distance between distribution of the attribute in the whole table and distribution of a sensitive attribute in this class should not be more than a threshold. This method limits the correlation between sensitive attribute and quasi-identifier attribute.

The method is to find the distance between the two probability distributions. Two methods are used in the paper for measuring the distance. They are variational distance and Earth Mover's distance (EMD).

A. Variational distance

Consider two distributions P = (p1, p2, ..., pm), Q = (q1, q2, ..., qm). The variational distance is defined as [5]:

$$D[P,Q] = \sum_{i=1}^{m} \frac{1}{2} |pi - qi|$$

This distance measures do not reflect the semantic distance among values. Suppose Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}. Suppose that first equivalence class has distribution P1 = {3k, 4k, 5k} and the second equivalence class has distribution P2 = {6k, 8k, 11k}. Semantically P1 results in more information leakage than

P2 and therefore D[P1,Q] > D[P2,Q]. This is because the values in P1 are all in the lower end. But in this method 3*k* and 6*k* are just different points and have no other semantic meaning. So the paper move onto another method called Earth Mover's Distance (EMD).

B. Earth mover's distance (EMD).

The Earth Mover's Distance (EMD) is a method to evaluate dissimilarity between two multi-dimensional distributions in some feature space where a distance measure between single features, which we call the ground distance is given. The EMD ``lifts" this distance from individual features to full distributions.

$$D[P, Q] = \text{WORK}(P, Q, F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$

P = (p1, p2, ...pm), Q = (q1, q2, ...qm), and d_{ij} be the ground distance between element i of P and element j of Q

The EMD has the following advantages.

- Naturally extends the notion of a distance between single elements to that of a distance between sets, or distributions, of elements.
- Allows for partial matches in a very natural way. This is important, for instance, for image retrieval and in order to deal with occlusions and clutter.
- Is a true metric if the ground distance is metric and if the total weights of two signatures are equal. This allows endowing image spaces with a metric structure
- Is bounded from below by the distance between the centers of mass of the two signatures when the ground distance is induced by a norm. Using this lower bound in retrieval systems significantly reduced the number of EMD computations.
- Matches perceptual similarity better than other measures, when the ground distance is perceptually meaningful.

While EMD is the best measure found so far it is not perfect. In EMD the relationship between the value *t* and information gain is unclear. So there is a need for a measure that combines the distance-estimation properties of the EMD with the probability scaling nature of the KL distance.

V. JS-REDUCE DEFENSE

The methods such as k-anonymity, l-diversity and t-closeness do not consider the sequential background knowledge of the adversary. So another method called JS-Reduce has been proposed which considers adversaries background knowledge in serial microdata release that is background obtained by adversary in serial release of database [6]. The method was proposed by Daniele Riboni, Linda Pareschi and Claudio Bettini. In this method first a model created to find out the background knowledge, Posterior background knowledge and revised sensitive values background knowledge. Adversary's background knowledge is revised each time when data is released. The main goal of the method is to maximize the similarity of probability distribution of sensitive value. For that Jenson-Shannon divergence is used.

JS-Reduce defense is divided into two parts. First derive the background knowledge and second is apply Jenson Shannon divergence algorithm. In the first part first calculate the sensitive value background knowledge is calculated which is probability which associates an individual to a sensitive value. This is done by mining the background from available corpus of data. After that Sequential Background Knowledge is derived which is probability distribution after releasing a series of data is released.

Then posterior knowledge at a particular time is calculated which will give association between a respondent and sensitive values about the release at a particular time after release data at a particular time. This calculation is based on particular possible configuration between QI-group and sensitive information. Based on posterior knowledge, revised sensitive values background knowledge of the data to be released is calculated. The independent posterior knowledge PK^{sv} is calculated using RBK^{sv}, which is acquired by applying the conditional probability given by BK^{seq} on the sequence of sensitive values in the release history of the database.

After calculating all this data background knowledge Jenson - Shannon divergence is applied on the data. JS-reduce create QI-groups whose tuple respondents have similar RBK^{sv} distributions. The adversary cannot exploit background knowledge to perform the attack if the respondents of tuples in a QI-group are indistinguishable with respect to RBK^{sv}. Jenson - Shannon divergence [8] is used to quantify information disclosure because it will calculate the average information or entropy. The whole process can be modeled as in figure 2.

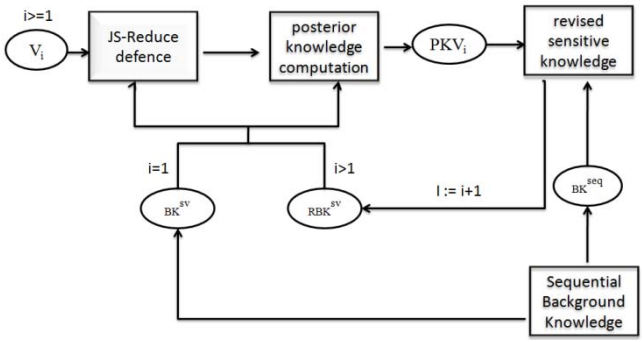

Figure 2. Defence model [6]

The advantage of JS-Reduce is that it is effective when adversary have sequential background knowledge.

VI. PROPOSED METHOD

JS-Reduce defence is efficient technique for privacy preservation in data publishing. The method uses Jensen-Shannon divergence for similarity measurement. The method does not provide any anonymization. The method can be improved by following technique.

There are many similarity measures in data mining with varying efficiency and consumption time. Instead of Jensen –Shannon Divergence for similarity measurement we use correlation technique for measuring similarity. The advantage of this method is that the equation is very simple. The correlation coefficient between two variables can be found out by following equation.

$$r = \frac{1}{n-1}\sum\left(\frac{x-\bar{x}}{s_x}\right)\left(\frac{y-\bar{y}}{s_y}\right)$$

Correlation gives better similarity among different similarity measures. The proposed method enforces anonymization using anonymization technique such as generalization or suppression. Anonymization is an important feature to provides full privacy for the individual without much loss of information.

VII. CONCLUSION

Privacy preservation in data publishing is one of the tedious tasks in data publishing. This survey describes several existing data publishing methods such as k-anonymity, l-diversity, t-closeness, JS-Reduce. Among these methods JS reduce is the only method which models sequential background knowledge attack. The proposed method provides better model which consider sequential background knowledge attack as well as anonymize data which provides better privacy protection to individual.

VIII. REFERENCES

[1] Dwork, "Differential Privacy," Proc. 33rd Int'l Colloquium on Automata Languages and Programming (ICALP '06)", pp. 1-12, 2006.
[2] L. Sweeney, "k-anonymity: a model for protecting privacy." ,international Journal on Uncertainty, Fuzziness and Knowledge –based Systems, 2002,pp. 557-570
[3] Meyerson and R. Williams, "On the Complexity of Optimal k- Anonymity", Proc. 23rd ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS '04), pp. 223-228, 2004.
[4] Ashwin Machanavajjhala, Daniel Kifer, Johannes Gehrke, and Muthuramakrishnan Venkitasubramaniam, "L-diversity: Privacy beyond k-anonymity.",ACM Trans. Knowl. Discov,2007.
[5] Ninghui Li, Tiancheng Li; Venkatasubramanian, S, "t-Closeness: Privacy Beyond k-Anonymity and l-Diversity", Data Engineering, ICDE 2007. IEEE 23rd International Conference , pp.106,115, 15-20 ,April 2007
[6] Riboni, D, Pareschi, L,Bettini, C, "JS-Reduce: Defending Your Data from Sequential Background Knowledge Attacks", Dependable and Secure Computing, IEEE Transactions,vol.9, no.3, pp.387,400, May-June 2012
[7] V. Ciriani, S. De Capitani di Vimercati, S. Foresti, and P. Samarati, ",k-Anonymity", Springer US, Advances in Information Security (2007)
[8] J. Lin, "Divergence Measures based on the Shannon Entropy," IEEE Trans. Information Theory, vol. 37, no. 1, pp. 145-151, 1991.
[9] T. Li, N. Li, and J. Zhang, "Modeling and Integrating Background Knowledge in Data Anonymization," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE '09), pp. 6-17, 2009