A New Secure Method of Privacy Publication For Data Publishing

Harsha Cherukuri [#1], Akash Kashyap [*2]

[#]B.Tech Scholar, *B.Tech Scholar
Department of CSE,
Jawaharlal Nehru Technological University,
Hyderabad, AP, India.

Abstract

Data Publishing is one of the leading publishers especially for a wide range of local and wide range of business environments. In order to publish micro data, we need at least of nearly k-records for maintaining privacy-requirements. Now a day's we have observed a rapid advance growth in privacy-preserving data publishing which have lead mainly to an majority of increase in the capability of any firm or companies for storing the data and also to record the personal data about the customers in an company. In order to maintain the privacy for the high dimensional database has become a very challenging aspect. Recent research work have undergone and identified two methods like k-anonymity and l-diversity, especially to limit only the identity disclosure, but these methods are not completely satisfied in their usage for limiting the attribute disclosure. Motivated by these limitations, in this paper we proposed a new technique for privacy preserving mechanism called as "closeness". Initially we present the main model in this paper like t-closeness, as it mainly requires the distribution of a very sensitive hidden attribute in any equivalence class is near to distributed attribute in the overall table .Later we also propose a more flexible privacy model called (n, t)-closeness that offers higher utility. We finally after several observations present two distance measures.

Keywords

Privacy Preservation, Data Anonymization.

1. Introduction

In our current research work, we have mainly found that our surroundings is greatly experienced by exponential growth in the number and variety of different type of data collections which mainly contains person's specific information, for example if we consider hospital as example, publishing medical data is very much crucial for analysis of medical data. For this purpose the data is stored in the form of tables, each table consists of individual rows and columns, where each row consists of records corresponding to one individual and each record has a number of distinct attributes, which can be further divided into the following 3 identifier categories.

1. **Explicit Identifiers:** Attributes that clearly identify individuals.

2. **Quasi Identifiers:** Attributes through which we can identify an individual by which values when taken together can potentially e.g., Zip-code, Birth-date, and Gender.

3. **Sensitive Identifiers**: Attributes that are considered as sensitive and very crucial is known as Sensitive Attributes

When we try to release micro data, it is very important for individual to prevent the valuable sensitive hidden information of the individuals from being disclosed by un authorized

users. We have identified two types of information disclosure in the literature [1], [2]: *identity disclosure method and attribute disclosure* method.

1) Identity Disclosure Method

This type of disclosure occurs mainly in a situation like individual is linked to a appropriate record in the released table. Once this identity disclosure is found, it is very easy for the identifier to identify the details of the particular individual.

2) **Attribute Disclosure Method**

This is another type of disclosure it occurs mainly when the new information of some individual is revealed.

With a lot of research work, we find that identity disclosure mainly leads to attribute disclosure but mostly the attribute disclosure may occur with the occurrence of identity disclosure or sometimes without the identity disclosure property. It is also recognized that it may cause harm [2] even if there is a disclosure of the false attribute information and also if the perception is incorrect.

As the researcher's gets almost very useful information with the released table, it also presents disclosure risk to the set of individual users who posted their data in the table. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table.

To effectively limit the data leverage, we mainly need to measure the leverage risk of an anonymized table. To this end, Samarati and Sweeney [3], [4] have mainly introduced a new property like *k-anonymity* for which each record is not able to distinguish with at least of *k-1 of all* other records with respect to the given quasi-identifier . It prevents the identity disclosure but it is insufficient to prevent attribute disclosure. To overcome this limitation of k-anonymity , Machanavajjhala [5] both have recently introduced l-diversity. This new implemented l-diversity also has some problem that which mainly deals with the limitation of assumption in adversarial knowledge i.e., it is possible for an adversary to gain information about the sensitive attributes as long as he has the information about the global distribution of this attribute.

Finally in this paper, we are going to propose a novel privacy notion called "Closeness". At first we are going to formalize the idea of the existing main model called " t-closeness" which is mainly required with the distribution of very valuable sensitive attributes in any equivalence class table to be very close to the total distribution of the whole attributes in the overall table .We are going to propose an flexible privacy model called (n,t)-closeness. With this new approach we are also going to find the average distance between the values of sensitive attribute by using a new metric called Earth Mover Distance (EMD).

2. Basic Definitions in Our Paper

In this section, we are going to find out some important notations used in this paper.

Quasi-Identifier Attribute Set

A Quasi-Identifier set Q is a defined as a very minimal set of attributes present in table T which can be joined with some additional external information to mainly re-identify individual records(with sufficiently high probability) [6] **.** K-Anonymity Property Relation T is said to satisfy the k-anonymity property (or to be k-anonymous) w.r.t attribute set Q if every count of attributes in the frequency set of table T w.r.t Q is greater than or equal to k. K-Anonymization a view V of a relation T is said to be a k-anonymization of T, if the view modifies the data of table T accordingly to some new mechanism such that variable or attribute V satisfies the k-anonymity property w.r.t the set of quasi-identifier attributes. T and V are assumed to be multi sets of tuples.

K-Anonymity

Let T(A1….. Am) be a table with a set of different attributes, and QI be quasi-identifier associated with the table T. T is said to satisfy k-anonymity with respect to QI iff the sequence of

distinct values in table w.r.t attributes T[QI] appears at least with k occurrences in T[QI] (T[QI] denotes the projection, by which maintain the most duplicate tuples, of attributes QI in T.

3. Related Work

In this section, we mainly deal with the problem of information disclosure. Till today a number of information leverage techniques have been designed for data publishing technique, including methods like Sampling of attributes present in a table, Cell Suppression, Rounding, and Data Swapping and Perturbation. Although different techniques were build, they even inserts noise to the data in the table. Two popular authors like Samarati [3] and Sweeney [4] introduced the new model called k-anonymity . Finally we classify them into two categories namely:

1. Privacy measurements
2. Anonymization techniques.

3.1 From k-Anonymity to l-Diversity

The protection method which is provided with the help of k-anonymity is simple and easy to understand. If a table satisfies k-anonymity for some distinct value k, then anyone in between who knows only the values of quasi-identifier of one individual may not able to identify the appropriate individual record corresponding to that individual with confidence greater than 1/k. This has been recognized by several authors [7, 8]. Two attacks were identified in the homogeneity attack and the background knowledge attack example.

Example 1:

Table 1 is the original data table, and Table 2 clearly shows that is an anonymized version of it satisfying 3-anonymity. From the two tables we found that disease attribute is very sensitive. Suppose Alice knows that Bob is a 27-year man who lives in ZIP 47678 and Bob's record is in the above table. We observe clearly from Table 2, Alice can conclude that Bob corresponds to one of the first three records, and thus must have heart disease. This is known to be as homogeneity attack.

TABLE 1
Original Patients Table

	ZIP Code	Age	Disease
1	47677	29	Heart Disease
2	47602	22	Heart Disease
3	47678	27	Heart Disease
4	47905	43	Flu
5	47909	52	Heart Disease
6	47906	47	Cancer
7	47605	30	Heart Disease
8	47673	36	Cancer
9	47607	32	Cancer

	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	4790*	$\geq$ 40	Flu
5	4790*	$\geq$ 40	Heart Disease
6	4790*	$\geq$ 40	Cancer
7	476**	3*	Heart Disease
8	476**	3*	Cancer
9	476**	3*	Cancer

TABLE 2
A 3-Anonymous Version of Table 1

TABLE 3
Original Salary/Disease Table

	ZIP Code	Age	Salary	Disease
1	47677	29	3K	gastric ulcer
2	47602	22	4K	gastritis
3	47678	27	5K	stomach cancer
4	47905	43	6K	gastritis
5	47909	52	11K	flu
6	47906	47	8K	bronchitis
7	47605	30	7K	bronchitis
8	47673	36	9K	pneumonia
9	47607	32	10K	stomach cancer

	ZIP Code	Age	Salary	Disease
1	476**	2*	3K	gastric ulcer
2	476**	2*	4K	gastritis
3	476**	2*	5K	stomach cancer
4	4790*	≥ 40	6K	gastritis
5	4790*	≥ 40	11K	flu
6	4790*	≥ 40	8K	bronchitis
7	476**	3*	7K	bronchitis
8	476**	3*	9K	pneumonia
9	476**	3*	10K	stomach cancer

TABLE 4
A 3-diverse version of Table 3

3.2 Limitations of l-Diversity

Although we know that the l-diversity principle represents an important step beyond k-anonymity, even it has several shortcomings that we now discuss. L-diversity may be difficult to achieve and may not provide sufficient privacy protection. l-diversity is not sufficient to prevent attribute leverage. By the below assumptions, we present two attacks on l-diversity attack, Skewness [a], Similarity attack[a].

Skewness Attack

This attack mainly occurs when the overall distribution of attributes present in the table is skewed, Satisfying „-diversity does not prevent attribute disclosure. Consider again Example 1. Suppose we assume that out of several equivalence classes we take into consideration, one equivalence class is assumed to has an equal number of positive records and same equal number of negative records. It satisfies distinct 2-diversity, entropy 2-diversity, and any recursive (c, 2)-diversity .However, this presents mainly a very serious privacy risk, because none of the person present in the class would be considered to have 60% possibility of being positive, as compared with the 1% of the total population.

Similarity Attack

This attack mainly occurs when the valuable sensitive attribute values in an equivalence class are distinct not same but semantically similar, an adversary can learn important information.

4. (n , t)-Closeness: A New Privacy Measure

A new privacy measure like t -closeness measure which mainly tells that the overall distribution of a valuable sensitive attribute in any equivalence class to the distribution of sensitive attribute in the overall table. It is an enhancement model of *l*-diversity.

The l-diversity requirement is mainly motivated by limiting the difference between the posterior and prior belief. Based on the research work, we proposed a very new and more flexible **privacy model class called (n, t) - closeness**, class which mainly requires the distribution in any equivalence class is near to the distribution in a very large enough equivalence class.

4.1 t-Closeness Principle Algorithm

Input:

P and Q is individually partitioned into r distinct partitions as {P1, P2,…… pr}and {Q1,Q2,….Qr} , EC is known as Each Class, where t represents the threshold value.

Output:

True if (n,t)-closeness is satisfied, otherwise false.

Where in this

P → Posterior Belief
Q →Prior Belief and
D →Difference

Information Gain = D[P,Q]

An EC is t-Closeness if $D[P,Q] \leq t$

An Table is said to be as t-Closeness if and only if all the EC has t-Closeness

If D[P,Q] is ↓, then the information gained by the observer will ↓ privacy risk will also get ↓

If D[P,Q] ↑, then the information gained by the observer will also ↑ the benefit of the published data

4.2 De-Anonymization Algorithm

De-anonymization is one among the best matching algorithm which uses matching function and scoring function. Scoring function initially assigns the numerical value of the data table and matching function mainly deals with the algorithm that is applied by the very adversary to determine the scores by using the different set of individual matches. Finally record selection selects one "best guess" record.

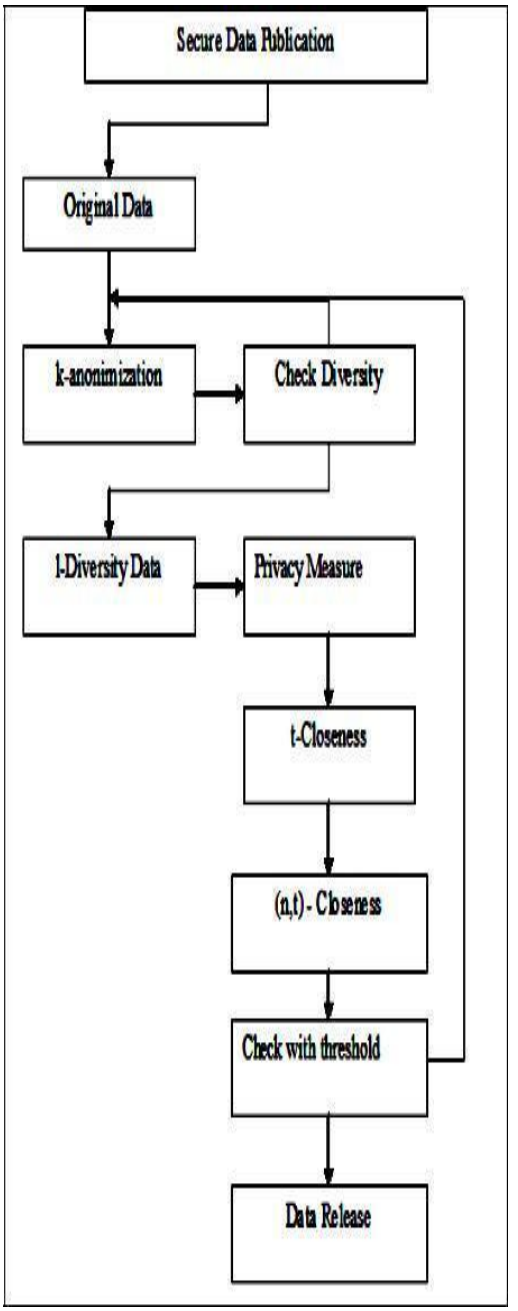

Figure 1 Represents architecture of Micro data Release

4.3 Earth Mover's Distance

The EMD method is mainly used to find the average distance between any two different distributions, which are mainly represented by signatures. The signatures are represented as the sets of weighted features that capture the distributions. The feature which we capture can be of any type and in any number of dimensions, and are defined only by the user.The notion of "work" is mainly based on principle of the user-defined ground distance which is the distance between two features of user selection. Here in this method, the size of the two signatures can also be different. Also, the sum of weights of individual signature can be almost different than the sum of weights of the other (partial match).

5. Conclusion and Future Work

In this paper, we finally proposed a very new method which will surely reduce the leverage risk and we also provide the very good high level security which is mostly useful in micro data publishing.t-closeness removes an outlier may smooth a distribution and it bring much closer to the overall distribution. For measuring the privacy we use similarity measure and the EMD metric for performing all this process.

6. References

[1] G. T. Duncan , D. Lambert, " Property of Disclosure-Limited Data Dissemination Technique, "Journal of The American Statistical Association method, vol. 81, pp. 10-28, 1986.

[2] D. Lambert, "Measures of Disclosure Risk and

Harm," Journal of Official Statistics Mechanism, vol. 9, pp. 313-331, 1993.

[3] P. Samarati,"Protecting Respondent‟s Privacy in Microdata Release", IEEE Trans. Knowledge and Data Eng., Vol. 13, No. 6, Nov./Dec. 2001.

[4]L. Sweeney,"k-Anonymity: A Model for Protecting Privacy", Int‟l,Vol. 10, No. 5, pp. 557-570, 2002.

[5]Machanavajjhala, and M. Venkitasubramaniam, "`-Diversity: Privacy Beyond k-Anonymity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 24, 2006.

[6]K. LeFevre, D. DeWitt, "Incognito: Efficient Full-Domain k-Anonymity," Proc. ACM SIGMOD,pp. 49-60, 2005.

[7]T. M. Truta and B. Vinay, "Privacy Protection: p-Sensitive" Proc. Int'l Workshop on Privacy Data Management ,2006.

[8]X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 229-240, 2006.

7. About the Authors

Harsha Cherukuri received his B.Tech. Degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, A.P in 2013.His research interests include data privacy and security, cryptography and data mining.

Akash Kashyap received his B.Tech. Degree in Computer Science and Engineering from Jawaharlal Nehru Technological University, Hyderabad, A.P in 2013.His research interests include Data Analytics, Algorithms.