International Journal of Modern Engineering Research (IJMER)
www.ijmer.com Vol.2, Issue.1, Jan-Feb 2012 pp-208-212 ISSN: 2249-6645

Secured Medical Data Publication & Measure the Privacy "Closeness" Using Earth Mover Distance (EMD)

Krishna.V[#], Santhana Lakshmi. S[*]
[#]PG Student, Department of Computer Science
[*] Department of Computer Science
[#*]Coimbatore Institute of Engineering and Technology
Coimbatore, India

Abstract: Privacy requirement for publishing microdata needs equivalent class contains at least k-records. Recent research for releasing microdata are k-anonymity and l-diversity this two method are used to limit only the identity disclosure it is not sufficient for limiting the attribute disclosure. so we are undergoing a new privacy technique known a t-closeness also we are going to measure the privacy that is finding the distance measure between the two probability distribution by using the Earth Mover's Distance a new distance measure technique.

I.INTRODUCTION

Data Publishing is the leading publisher for local communities and business. Data Publishing plays a vital role in hospitals, government agencies, insurance companies and all other business where data would like to release for the purpose of analysis and research purpose, Now a day's society is experiencing exponential growth in the number and variety of data collections containing persons specific information, for example publishing medical data is very much important for analysis and in research areas. Mainly for this purpose the data are stored in table, each table consists of rows and columns, each row consists of records corresponding to one individual and each record has a number of attributes. There are three types of attributes they are 1) **Explicit identifier** are the attribute that clearly identifies the individual, e.g., Patient name, Patient Number. 2) **Quasi identifier** are the attributes whose values that can be taken together can potentially identify an individual, e.g., Zip code, Date-of-birth, Gender. 3) **Sensitive attributes** are the attributes that are consider being more sensitive when releasing a micro data, it is very much necessary to prevent the sensitive information of the individuals from being disclosed, e.g., Disease of an individual is the more sensitive information in medical data publishing.

Privacy Preservation is the main aim of data publishing, however micro data contains more sensitive information it is very necessary for the owner to protect those information i.e.) guaranteeing the privacy of individual by ensuring that their sensitive information is not disclosed. Basically there are two types of disclosures they are as 1) **Identity disclosure** is a type of disclosure when an individual is linked to a particular record in the released table. Once if there is an occurrence of identity disclosure in the released table it is very easy to identify the details of the particular individual. 2) **Attribute disclosure** is another type of disclosure it occurs when the new information of some individual is revealed i.e., the released data make possible to know the characteristics of a particular individual more accurately. Identity disclosure mainly leads to attribute disclosure but attribute disclosure may occur with or without the occurrence of identity disclosure. It has been recognized that it may cause harm even if there is a disclosure of the false attribute information and also if the perception is incorrect.

Our main aim in data publishing is to limit the disclosure risk of the table that is to be published and this can be achieved by anonymizing the data before release. Anonymization is the technique in which the explicit identifiers are removed, but this is not enough for preserving the published data because the adversary is already having the quasi identifier values in the table. Generalization is the common anonymization approach.

It is necessary to measure the risk of an anonymized table to limit the disclosure risk. For this two of them Samarati and Sweeney introduced a technique known as k-anonymity, property that captures the protection of microdata table with respect to possible re-identification of respondents to which the data refer. It prevents the identity disclosure but it is insufficient to prevent attribute disclosure. To overcome this limitation of k-anonymity , Machanavajjhala introduced a notion of privacy called l-diversity i.e., which it requires that the distribution of sensitive attribute in each equivalence class has at least *l* well represented values. This 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.

In this we are going to propose a novel privacy notion called "Closeness" At first we are going to formalize the idea of the base model t-closeness requires that the distribution of sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., given that the distance between both the distribution should not be more than the threshold t) this can effectively reduce the amount of individual specific information that an observer can learn. We are going to propose an flexible privacy model called (n,t)-closeness. With this we are also going to find the distance between the values of sensitive attribute by using Earth Mover Distance metric(EMD).

II. K-Anonymity

Definition 1 (The k-anonymity principle). *Each release of data must be in such a way that every combination of the values of quasi-identifiers can be indistinctly matched to at least for k respondents.*

If the information for each person contained in the release cannot be distinguished from at least K-1 individuals whose information also appear in the release. It is a type of protection which is provided in a simple and easy way to understand. K-anonymity mainly deals with two factors Suppression and Generalization.

Suppression can replace individual attribute with a * and generalization will replace the individual attributes with a border category for example, consider the age of a person is 35 then in generalization it will be replaced as [30 – 40], while this K-anonymity is sufficient only to protect identity disclosure, but it is not sufficient to protect attribute disclosure. If a table satisfies K-anonymity for some value k, then if anyone knows the quasi identifier value of any particular person then it will be easy to find the details of that individual. Two attacks are identified in this k-anonymity they are Homogeneity attack and Background knowledge attack. K-Anonymity can create groups that leak the information due to the lack of diversity in the sensitive attribute, this may cause both the homogeneity and background knowledge attack.

Table 1 Original Patient Table

S.no	Zip code	Age	Nationality	Disease
1	13053	28	Russian	Heart
2	13068	29	Indian	Cancer
3	13068	21	Japanese	Viral
4	14858	50	American	Viral
5	14853	55	Russian	Cancer
6	14853	47	Indian	Heart
7	14850	33	American	Heart
8	14850	31	Indian	Cancer
9	14853	39	Russian	Heart

Example 1. From the above two table, Table 1 is the original data table and Table 2 is an anonymized version of it satisfying 3-anonymity, in this table disease is consider to be the sensitive attribute. Suppose Alice knows that Bob is 29 year old man living in ZIP 13068 and Bob's record is in table. From the Table 2 Alice can confirm that Bob corresponds to one of the first three records, and must have heart disease, this is called homogeneity attack.

Suppose by knowing Carl's age is 39 in the ZIP code 1305, Alice can conclude that Carl correspond to a record in the last equivalence class in the Table 2. Also Alice knows that Carl has very low risk of heart disease, this is background knowledge attack i.e., this enables Alice to conclude that Carl most likely has cancer.

Table 2 A 3-Anonymous version of Table

S.no	Zip code	Age	Nationality	Disease
1	130**	2*	*	Heart
2	130**	2*	*	Cancer
3	130**	2*	*	Viral
4	148**	[45-60]	*	Viral
5	148**	[45-60]	*	Cancer
6	148**	[45-60]	*	Heart
7	1485*	3*	*	Heart
8	1485*	3*	*	Cancer
9	1485*	3*	*	Heart

l-diversity

To address the limitation of K-Anonymity recently introduced a new notion of privacy known as *l*-diversity. which requires that the distribution of a sensitive attribute in each equivalence class has at least *l* "well represented" values.

Principle of *l*-diversity

A q-block is *l*-diverse if it contains at least *l* "well represented" values for the sensitive attribute S. A table is *l*-diverse if every q-block is *l*-diverse.

l-diversity Instantiations

l-diversity consists of certain instantiations and are stated by Machanavajjhala.

1. **Distinct *l*-diversity.** Each equivalence class has at least *l* "well represented" sensitive values. The drawback of this distinct *l*-diversity is it does not prevent probabilistic interface attacks. For example in an equivalent class. The table contains ten tuples in the sensitive area disease in that consider that one of them is "flu" and one is "Cancer" and the rest eight are "Heart Disease" this satisfies 3-diversity rule, but the attacker can still confirm that the target person's disease is "Heart Disease" with the accuracy of 80%.

2. **Entropy *l*-diversity.** Each equivalence class not only must have enough different sensitive values, but also the different sensitive values must be distributed evenly, it means the entropy of the distribution of sensitive values in each class is at least log(l). the disadvantage of this is sometimes it may be too restrictive. i.e., in a table some values are more common means then the entire table entropy is too low. This may cause less conservative notion of *l*-diversity.

3. **Recursive (c,*l*)-diversity.** This recursive (c,*l*)-diversity can be interpreted in terms of adversarial background knowledge. This mainly protect against all adversaries who posses almost *l*-2 diversity. The main drawback of this diversity is that the most frequent value won't appear more frequently, also the less frequent value does not appear too rarely. If $r1 < c(rl + rl+1 + \ldots + rm)$ then the table is said to have recursive (c,*l*)-diversity if all of its equivalence classes have recursive (c,*l*)-diversity.

Limitations of *l*-diversity

We are using *l*-diversity in order to overcome the disadvantage of k-anonymity beyond protecting against the attribute disclosure. the main disadvantage of *l*-diversity is that it won't consider the overall distribution of the sensitive values, *l*-diversity is difficult to achieve and also it does not provide sufficient protection against attribute disclosure.

Example 2. Suppose there are 10000 records in total in that if 99 percent is negative and only 1 percent is positive means then the two values have very difficult degrees of sensitivity. i.e., in this one won't mind for being known to test for negative, because one is same as 99 percent of the population, but one would not want to test for positive. In this case 2-diversity does not provide sufficient privacy protection.

This *l*-diversity is insufficient to provide attribute disclosure has two types of attack namely skewness attack and similarity attack.

Skewness attack.

Consider an example that one equivalent class has equal number of positive and negative records means it will satisfy distinct 2-diversity, entropy 2-diversity and (c,2)-diversity, by this we can consider that the 50 percent of the possibility be positive and the other 50 percent be negative.

Similarity attack

Sensitive attribute value in an equivalent class are said to be distinct but also semantically similar, in this case an adversary can learn the important information from the table and is said to be as similarity attack and is shown in table 3 and 4.

Table 3 Original Salary/Disease Table

	Zip code	Age	Salary	Disease
1	47671	25	4k	Gastric
2	47603	23	3k	Gastric
3	47677	24	5k	Cancer
4	47905	45	11k	Gastric
5	47980	53	6k	Bronchitis
6	47906	42	8k	Bronchitis
7	47603	32	10k	Flu
8	47609	39	9k	Pneumonia
9	47607	35	10k	Cancer

III. NEW PRIVACY MEASURE: t-closeness

t-closeness is said to be as a new privacy measure which is said to be as the distribution of a sensitive attribute in any equivalence class to the distribution of sensitive attribute in the overall table. It is an enhancement model of *l*-diversity. Generally privacy can be measured by the information gained by the observer and this can be measured or calculated by subtracting the posterior belief and the prior belif of the observer.

Table 4 A 3-Diverse Version of Table

	Zipcode	Age	Salary	Disease
1	476**	2*	4k	Gastric
2	476**	2*	3k	Gastric
3	476**	2*	5k	Cancer
4	479**	[40-50]	11k	Gastric
5	479**	[40-50]	6k	Bronchitis
6	479**	[40-50]	8k	Bronchitis
7	476**	3*	10k	Flu
8	476**	3*	9k	Pneumonia
9	476**	3*	10k	Cancer

The *l*-diversity requirement is motivated by limiting the difference between the posterior and prior belief. In this we are not going to limit the information gained by the observer about the whole population but we are going to limit the extent to which an observer can learn additional information about the specific individual. *An equivalence class is said to have t-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table should not be no more than threshold t.* Based on the analysis, we propose a more flexible ***privacy model called (n, t) - closeness***, which requires that the distribution in any equivalence class is close to the distribution in a **large-enough equivalence class** (contains at least n records) with respect to the sensitive attribute.

t-closeness can be performed based on the utility analysis by using the Greedy algorithm, the primary technique used for the generation of k-anonymous and l-diversity table from the original data set table are to generalize the quasi-identifier values and the sensitive values that are got from former and latter. This generalization can be performed in two ways namely they are of for the semantic data and the numeric data. In semantic data the data moves up by generalization hierarchy either by implied or by supplied. In numeric data a specific case of an implied generalization hierarchy is used.

Privacy can be measured by the information gain by an observer, i.e., at first the observer may have the prior belief about the sensitive attribute value before seeing the released table, and then after seeing the released table the observer may get the posterior belief about the sensitive attribute value. Prior belief is the distribution of sensitive attribute value in the whole table and it is represented by Q. Posterior belief is distribution of sensitive attribute value in a single class and this can be represented by P. From this we know that the information gain can be represented as the difference got from the posterior belief to the prior belief.

Information Gain = Posterior belief – Prior belief (or)
Information Gain = D[P,Q]

Architecture

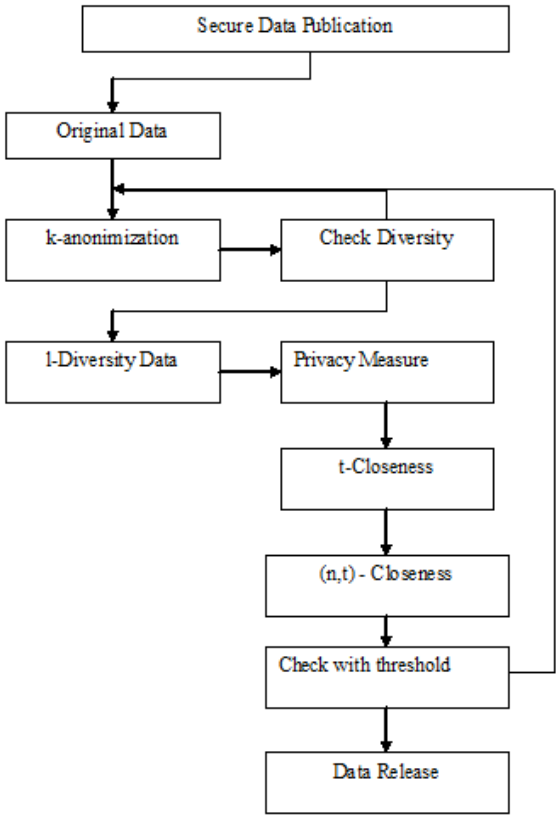

Figure 1Architecture of Microdata Release

t-Closeness Principle algorithm

input: P and Q is partitioned into r partitions as $\{P_1, P_2, \ldots\ldots p_r\}$and $\{Q_1,Q_2,\ldots.Q_r\}$, EC is Each Class, t is the threshold value.

Output: true if (n,t)-closeness is satisfied, otherwise false.

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

Where in this P → Posterior Belief

Q →Prior Belief and

D →Difference

IV.DISTANCE MEASURES

Distance Measure is the technique which is used for measuring the distance that is the security, which is must satisfy some five properties such as identifying of indiscernible, Non negativity, Probability Scaling, Zero Probability definability and semantic awareness. Identity of indiscernible is that no information is gained if an adversary does not change its belief it is generally represented as $D[P,Q] = 0$. Non negativity means if an adversary is gaining non negative information then it can be represented as $D[P,Q] \geq 0$. Probability Scaling means D[P,Q] should compulsorily reflect the difference. Zero probability distribution is the well defined zero probability values in P and Q and semantic awareness is that if the values of P and Q are of having semantic meaning means then D[P,Q] must reflect the semantic difference among that two different values.

In general we are planning to measure the distance with the help of two types of algorithms namely they are of two ways they are Similarity Measure and Earth Mover's Distance which is mainly used to calculate or to measure the privacy.

Similarity Measure:

Similarity Measure is the technique that which mainly allows similar evaluation of the encrypted policies. This technique will relies the existing encryption method that which mainly allows for the numerical data that are present in the table. Similarity Measure is the function that maps a pair of attributes to the interval [0,1]. It captures the intuitive notion of two values being "similar." Generally similar attributes will behave like an indicator function.

DE-ANONYMIZATION ALGORITHM:

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

Earth Mover's Distance:

The EMD computes the distance between two distributions, which are represented by signatures. The signatures are sets of weighted features that capture the distributions. The features can be of any type and in any number of dimensions, and are defined by the user. The EMD is defined as the minimum amount of work needed to change one signature into the other. The notion of "work" is based on the user-defined ground distance which is the distance between two features. The size of the two signatures can be different. Also, the sum of weights of one signature can be different than the sum of weights of the other (partial match). Because of this, the EMD is normalized by the smaller sum.

V. CONCLUSION

The method we used here will surely reduce the disclosure risk and also they provide the high level security which is very much useful in microdata publishing. t-closeness is the more flexible privacy model that which provide or achieves the better balance between the privacy and utility. t-closeness removing an outlier may smooth a distribution and it bring it much closer to the overall distribution. For measuring the privacy we use similarity measure and the Earth Mover's Distance for performing all this process we use generalization and suppression techniques.

REFERENCES

[1] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 901-909,2005.

[2] G. Aggarwal, T. Feder, K. Kenthapadi, S. Khuller, R. Panigrahy, D.Thomas, and A. Zhu, "Achieving Anonymity via Clustering,"Proc. ACM Symp. Principles of Database Systems (PODS), pp. 153-162, 2006.

[3] R.K. Ahuja, T.L. Magnanti, and J.B. Orlin, Network Flows: Theory, Algorithms, and Applications. Prentice-Hall, Inc., 1993.

[4] R.J. Bayardo and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), pp. 217-228,2005.

[5] F. Bacchus, A. Grove, J.Y. Halpern, and D. Koller, "From Statisticsto Beliefs," Proc. Nat'l Conf. Artificial Intelligence (AAAI), pp. 602-608, 1992.

[6] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure Anonymization for Incremental Datasets," Proc. VLDB Workshop Secure Data Management (SDM), pp. 48-63, 2006.

[7] B.-C. Chen, K. LeFevre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 770-781, 2007.

[8] G.T. Duncan and D. Lambert, "Disclosure- Limited Data Dissemination,"J. Am. Statistical Assoc., vol. 81, pp. 10-28, 1986.

[9] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets," Proc. ACM SIGMOD, pp. 217-228, 2006.

[10] D. Lambert, "Measures of Disclosure Risk and Harm," J. Official Statistics, vol. 9, pp. 313-331, 1993.

[11] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization," Proc. ACM SIGKDD, pp. 277-286, 2006.

[12] T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," Proc. Int'l Conf. Data Eng. (ICDE), 2008.

[21] T. Li and N. Li, "Towards Optimal k- Anonymization," Data and Knowledge Eng., vol. 65, pp. 22-39, 2008.

[22] T.M. Truta and B. Vinay, "Privacy Protection: P-Sensitive k-Anonymity Property," Proc. Int'l Workshop Privacy Data Management (ICDE Workshops), 2006.

[23] A. Asuncion and D.J. Newman, UCI Machine Learning Repository, 2007.

ABOUT THE AUTHORS

The Author, Krishna.V is a final year student doing Master of Engineering in Computer Science at Coimbatore Institute of Engineering and Technology and has a Bachelor of Engineering degree in Computer Science and Engineering from Ponjesly College of Engineering, Nagercoil. My Research area is Data Mining.

Ms. S. Santhana Lakshmi received her M.E., degree in Computer Science and Engineering from VLB Janakiammal college of Engineering .She is currently working as assistant professor in Coimbatore Institute of Engineering and Technology. Her Research area is Network Security and Data Mining.