A survey on data publishing of sensitive information through proximity in Cloud Database System

PROF.KALAM NARREN[1], Assistant Professor **PROF.V.VINAY KRISHNA**[2], Assistant Professor Department of Computer Science and Engineering

Christu Jyothi Institute of Technology & Science

***Abstract*- cloud computing has been the most adoptable technology current days and the database has also moved to cloud computing. In this paper includes the importance of considering the adversary's background knowledge when reasoning about privacy in data publishing. However, it is very difficult for the data publisher to know exactly the adversary's background knowledge. This paper presents a general framework for modeling the adversary's background knowledge using proximity, which allows the data publisher to enforce privacy requirements to protect the data against adversaries with different levels of background knowledge. Through an extensive set of experiments, we show the effects of probabilistic background knowledge in data anonymization and the effectiveness of our approach in both privacy protection and utility preservation.**

Keywords--- data anonymization, data security, cloud computing, virtualization, DBaaS.

1. INTRODUTION

A database can be accessed by the clients via the internet from the cloud database service provider and is deliverable to the users when they demand it. The cloud database is implemented using cloud computing that means utilizing the software and hardware resources of the cloud computing service provider. Many companies started cloud database as a service (Database as a Service). To publish the data to public we need certain privacy model to protect certain sensitive attributes fields in order to perform survey on the given data, example to know medical health, poverty issues, employee salary etc. By publishing the survey on data we need to protect the critical information of individuals. For this we have number of privacy models have been proposed for data anonymization, e.g., k-anonymity, l-diversity, t-closeness, and so on. A key limitation of these models is that they cannot guarantee that the sensitive attribute values of individuals are protected when the adversary has additional knowledge (called background knowledge). Background knowledge can come from diverse sources, such as well known facts, demographic information, public records, and information about specific individuals. As an example, consider that a hospital has the original patient table T in Table 1, which contains three attributes ZIP code, Age, and Disease. The hospital releases a generalized table in Table 2 which satisfies a 3-Anonymous.

When releasing critical data, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature [1], [2]: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data makes it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm [2].

Sno	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

Sno	ZIP Code	Age	Disease
1	476**	2*	Heart Disease
2	476**	2*	Heart Disease
3	476**	2*	Heart Disease
4	479**	≥40	Flu
5	479**	≥40	Heart Disease
6	479**	≥40	Cancer
7	476**	3*	Heart Disease

8	476**	3*	Cancer
9	476**	3*	Cancer

TABLE 2
A 3-Anonymous Version of Table 1

sease
Disease
Disease
Disease
Flu
Disease
ncer
Disease
ncer
ncer

An observer of a released table of cloud database may incorrectly perceive that an individual's sensitive attribute takes a particular value, and behave accordingly based on the perception. This can harm the individual, even if the perception is incorrect. While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective for cloud database is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly-available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasi-identifier values with values that are less specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. We define an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. To this end, Samarati and Sweeney [3], [4] introduced k-anonymity as the property that each record is indistinguishable with at least k-1 other records with respect to the quasi-identifier. In other words, k-anonymity requires that each equivalence class contains at least k records. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure.

Sno	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

TABLE 3
Original Salary/Disease Table

Sno	ZIP Code	Age	Salary	Disease
1	476**	2*	3k	gastric ulcer
2	476**	2*	4k	gastritis
3	476**	2*	5k	stomach cancer
4	479**	≥40	6k	gastritis
5	479**	≥40	11k	flu
6	479**	≥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

The techniques used for protecting critical information k-anonymity, l-diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least l-"well-represented" values. One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. As we shall explain below, it is possible for an adversary to gain information about a sensitive attribute as long as she has information about the global distribution of this attribute. This assumption generalizes the specific background and homogeneity attacks used to motivate l-diversity. Another problem with privacy preserving methods in general is that they effectively assume all attributes to be categorical; the adversary either does or does not learn something sensitive. Of course, especially with numerical attributes, being close to the value is often good enough. In this article, we propose a novel privacy notion called "proximity". We first formalize the idea of global background knowledge and propose the base model t-closeness which requires that the distribution of a sensitive attribute in any equivalence class to be close to the distribution of the attribute in the overall table (i.e., the distance between the two distributions should be no more than a threshold t). This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data. Based on the analysis, we propose a more flexible privacy model called proximity, which requires the distribution in any equivalence class is close to the

distribution in a large-enough equivalence class with respect to the sensitive attribute. This limits the amount of sensitive information about individuals while preserves features and patterns about large groups. Which shows that proximity achieves a better balance between privacy and utility than existing privacy models such as l-diversity and t-closeness. To incorporate distances between values of sensitive attributes, we use the Earth Mover Distance metric [5] to measure the distance between the two distributions. We also show that EMD has its limitations and describe our desiderata for designing the distance measure. We then propose a novel distance measure that satisfies all the requirements.

2. A PRIVACY MEASURE

Intuitively, privacy is measured by the information gain of an observer. Before seeing the released table from cloud database, the observer has some prior belief about the sensitive attribute value of an individual. After seeing the released table, the observer has a posterior belief. Information gain can be represented as the difference between the posterior belief and the prior belief.

2.1 t-Closeness: Base Model

To motivate our approach, let us perform the following thought experiment: First an observer has some prior belief B0 about an individual's sensitive attribute. Then, in a hypothetical step, the observer is given a completely generalized version of the data table where all attributes in a quasi-identifier are removed (or, equivalently, generalized to the most general values). The observer's belief is influenced by Q, the distribution of the sensitive attribute values in the whole table, and changes to belief B1. Finally, the observer is given the released table. By knowing the quasi-identifier values of the individual, the observer is able to identify the equivalence class that the individual's record is in, and learn the distribution P of sensitive attribute values in this class. The observer's belief changes to B2. The l-diversity requirement is motivated by limiting the difference between B0 and B2 (although it does so only indirectly, by requiring that P has a level of diversity). We choose to limit the difference between B1 and B2. In other words, we assume that Q, the distribution of the sensitive attribute in the overall population in the table, is public information. We do not limit the observer's information gain about the population as a whole, but limit the extent to which the observer can learn additional information about specific individuals. In some sense, the larger the difference between B0 and B1 is, the more valuable the data is. Since the knowledge gain between B0 and B1 is about the population the dataset is about, we do not limit this gain. We limit the gain from B1 to B2 by limiting the distance between P and Q. intuitively, if P = Q, then B1 and B2 should be the same. If P and Q are close, then B1 and B2 should be close as well, even if B0 may be very different from both B1 and B2.

Definition 1 (The t-closeness Principle)*:* 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 is no more than a threshold t. A table is said to have t-closeness if all equivalence classes have t-closeness. Requiring that P and Q to be close would substantially limit the amount of useful information that is released to the researchers. It might be difficult to assess a correlation between a sensitive attribute (e.g., disease) and some quasi identifier attributes (e.g., zip code) because by construction, partitions are selected to prevent such correlations from being revealed.

2.2 Privacy Model

We first illustrate that t-closeness limits the release of useful information through the following example.

Sno	ZIP Code	Disease	Age	Count
1	47673	Cancer	29	100
2	47674	Flu	21	100
3	47605	Cancer	25	200
4	47602	Flu	23	200
5	47905	Cancer	43	100
6	47904	Flu	48	900
7	47906	Cancer	47	100

8	47907	Flu	41	900
9	47603	Cancer	34	100
10	47605	Flu	30	100
11	47602	Cancer	36	100
12	47607	Flu	32	100

TABLE 6
An Anonymous Version of Table 5 violating 0.1-closeness

Sno	*ZIP Code*	*Disease*	*Age*	*Count*
1	476**	Cancer	2*	300
				0
				0
4	479**	Flu	4*	1800
5	476**	Cancer	3*	200
6	476**	Flu	3*	200

TABLE 5
Original Patients Table

An Anonymous Version of Table 5 violating 0.1-closeness *Count* that indicates the number of individuals. The probability of cancer among the population in the dataset is $\frac{700}{3000}$=0.23 while the probability of cancer among individuals in the first equivalence class is as high as $\frac{300}{600}$= 0.5. Since 0.5−0.23 > 0.1 (we will show how to compute the distance in Section 5), the anonymized table does not satisfy 0.1-closeness. To achieve 0.1-closeness, all tuples in Table 5 have to be generalized into a single equivalence class. This results in substantial information loss. If we examine the original data in Table 5, we can discover that the probability of cancer among people living in zip code 476** is as high as $\frac{500}{1000}$= 0.5 while the probability of cancer among people living in zip code 479** is only $\frac{200}{2000}$= 0.1. The important fact that people living in zip code 476** have a much higher rate of cancer will be hidden if 0.1-closeness is enforced. The t-closeness principle defines the large population to be the whole table; however, it does not have to be so. In the above example, while it is reasonable to assume that the distribution of the whole table is public knowledge, one may argue that the distribution of the sensitive attribute among individuals living in zip code 476** should also be public information.

Definition 3 (The proximity *Principle):* An equivalence class E_1 is said to have proximity if there exists a set E_2 of records that is a natural superset of E_1 such that both contains at least n records, and the distance between the two distributions of the sensitive attribute in E_1 and E_2 is no more than a threshold t. A table is said to have proximity if all equivalence classes have proximity with 0.1-closeness. The intuition is that it is okay to learn information about a population of a large-enough size .Assume that first equivalence class E_1 is defined by (zip code='476**', 20 ≤ Age ≤ 29) and contains 600 tuples. In another equivalence class E2 that contains 400 tuples it would be the one defined by (zip code='476**', 20 ≤ Age ≤ 39). If both of the two large equivalence classes E1's and E_2'Sdistribution is close to (i.e., the distance is at most 0.1) either of the two large equivalence classes, then E_1 satisfies proximity with 0.1-closeness. The first and the third equivalence classes also satisfy proximity because both have the same distribution (the distribution is (0.5, 0.5)) as the large group which is the union of these two equivalence classes and the large group contains 1000 individuals.

3. UTILITY ANALYSIS:

Consider T be the original dataset and {E1,E2, ...,Ep} be the anonymized data where Ei($1 \le i \le p$) is an equivalence class. Let H(T) denote the entropy of sensitive attribute values in T and H(Ei) denote the entropy of sensitive attribute values in Ei($1 \le I \le p$). Here, the total

information loss of the anonymized data is measured as:

$$IL(E1, ...,Ep) = \sum_{1 \le i \le p} \frac{|Ei|}{|T|} H(E_i)$$

while the utility of the anonymized data is defined as

$$U(E1, ...,Ep) = H(T) - IL(E1, ...,Ep).$$

Where Entropy (E) = $\sum_{s \in S} p(E,s) \log p(E,s)$ in which S is the domain of the sensitive attribute, and p(E, s) is the fraction of records in E that have sensitive value s.

4. DISTANCE MEASURES

4.1 Anonymization Algorithms

One challenge is designing algorithms for anonymizing the data to achieve proximity with 0.1 t-closeness. This algorithm consists of three components: (1) choosing dimension on which to partition, (2) choosing a value to split, and (3) checking if the partitioning violates the privacy requirement.

If Pi($1 \le i \le r$) contains less than n records, the algorithm computes the distance between Pi and each partition in Parent (P). If there exists at least one large partition (containing at least n records) in Parent (P) whose distance to Pi (D[Pi,Q]) is at most 0.1 t-closeness, then Pi satisfies the proximity requirement.

input: P is partitioned into r partitions {P1, P2, ..., Pr}

output: true if proximity with 0.1 t-closeness is satisfied, false otherwise

1. **for** every Pi
2. **if** Pi contains less than n records
3. find=false
4. **for** every Q ∈ Parent(P) and $|Q| \ge n$
5. **if** $D[Pi,Q] \le t$, find=true
6. **if** find==false, **return** false
7. **return** *true*

Fig. 1. The checking algorithm

Earth Mover's Distance

The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. Intuitively, one distribution is seen as mass of earth spread in the space and the other as a collection of holes in the same space. EMD measures the least amount of work needed to fill the holes with earth. A unit of work corresponds to moving a unit of earth by a unit of ground distance. EMD can be formally defined using the well-studied transportation problem. Let $P = (p_1, p_2, p_m)$, $Q = (q_1, q_2, q_m)$, and d_{ij} be the ground distance between element i of P and element j of Q. We want to find a flow $F = [f_{ij}]$ where f_{ij} is the flow of mass from element i of P to element j of Q that minimizes the overall work:

$$WORK(P,Q,F) = \sum_{i=1}^{m} \sum_{j=1}^{m} d_{ij} f_{ij}$$ subject to the following constraints:

$f_{ij} \ge 0$ $1 \le i \le m, 1 \le j \le m$ (c1)

$$p_i - \sum_{i=1}^{m} f_{ij} + \sum_{i=1}^{m} f_{ji} = q_i \qquad \text{(c2)}$$

$$\sum_{i=1}^{m}\sum_{j=1}^{m} f_{ij} = \sum_{i=1}^{m} p_i \sum_{i=1}^{m} q_i = 1 \qquad \text{(c3)}$$

These three constraints guarantee that P is transformed to Q by the mass flow F. Once the transportation problem is solved, the EMD is defined to be the total work i.e.,

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

For the disease attribute, we use the hierarchy in Figure 2 to define the ground distances. For example, the distance between "Flu" and "Bronchitis" is 1/3, the distance between "Flu "and "Pulmonary embolism" is 2/3, and the distance between "Flu" and "Stomach cancer" is 3/3 = 1. Then the distance between the distribution {gastric ulcer, gastritis, stomach cancer} and the overall distribution is 0.5 while the distance between the distribution {gastric ulcer, stomach cancer, pneumonia} is 0.278.

For the salary attribute, to show how t-closeness with EMD handles the difficulties of l-diversity. Recall that Q = {3k, 4k, 5k, 6k, 7k, 8k, 9k, 10k, 11k}, P1 ={3k, 4k, 5k}, and P2 = {6k, 8k, 11k}.We calculate D[P1,Q] and D[P2,Q] using EMD. Let v1 = 3k, v2 = 4k, ...v9 = 11k, we define the distance between vi and vj to be |i−j|/8, thus the maximal distance is 1. We have D[P1,Q] = 0.375, and D[P2,Q] = 0.167.
For example, One optimal mass flow that transforms P1 to Q is to move 1/9 probability mass across the following pairs: (5k → 11k), (5k →10k), (5k →9k), (4k →8k), (4k →7k), (4k →6k), (3k →5k), (3k →4k). The cost of this is 1/9 × (6 + 5 + 4 + 4 + 3 + 2 + 2 + 1)/8 = 27/72 = 3/8 = 0.375.

Sno	ZIP Code	Age	Salary	Disease
1	4767*	≤40	3k	gastric ulcer
3	4767*	≤40	5k	stomach cancer
8	4767*	≤40	9k	Pneumonia
4	4790*	≥40	6k	Gastritis
5	4790*	≥40	11k	Flu
6	4790*	≥40	8k	Bronchitis
2	4760*	≤40	7k	Gastritis
7	4760*	≤40	4k	Bronchitis
9	4760*	≤40	10k	stomach cancer

TABLE 7
Table that has 0.167-closeness w.r.t. Salary and
0.278-closeness w.r.t. Disease

Multiple Sensitive Attributes Multiple sensitive attributes present additional challenges. Suppose we have two sensitive attributes U and V . One can consider the two attributes separately, i.e., an equivalence class E has proximity if E has proximity of 0.1 t-closeness with respect to both U and V. Another approach is to consider the joint distribution of the two attributes. To use this approach, one has to choose the ground distance between pairs of sensitive attribute values. A simple formula for calculating EMD may be difficult to derive, and the relationship between proximity and the level of privacy become more complicated.

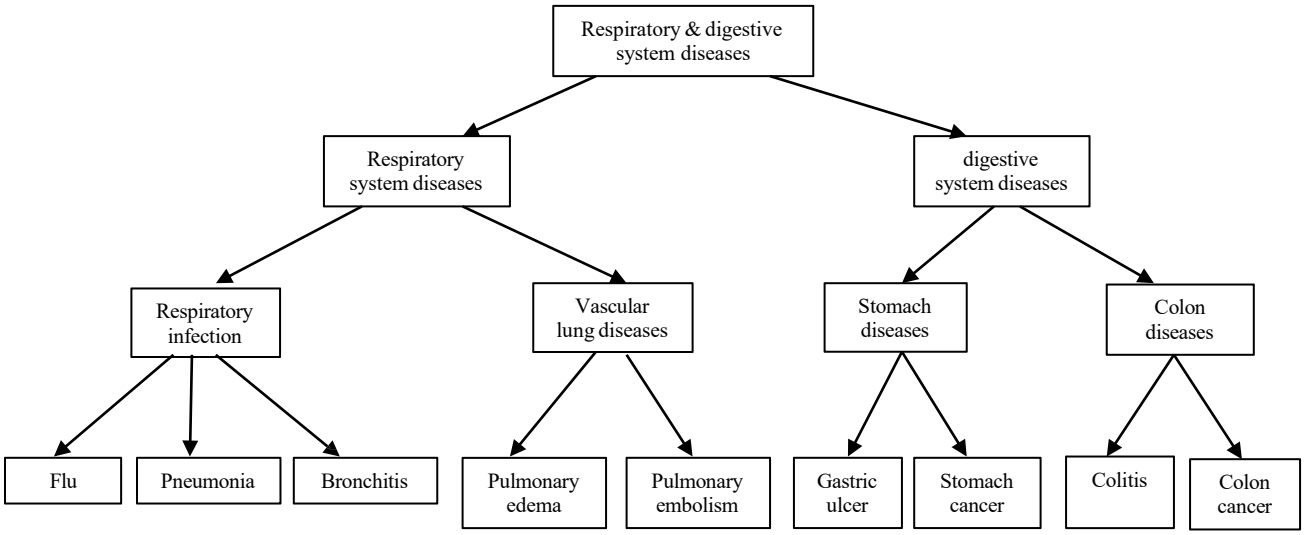

Fig 2. Hierarchy for categorical attributes *Disease.*

V. CONCLUSION

The method we used in cloud database system here will surely reduce the disclosure risk and also they provide the high level security which is very much useful in micro data 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 under cloud database system we use generalization and suppression techniques.

Other Anonymization Techniques: proximity allows us to take advantage of anonymization techniques other than generalization of quasi-identifier and suppression of records. For example, instead of suppressing a whole record, one can hide some sensitive attributes of the record; one advantage is that the number of records in the anonymized table is accurate, which may be useful in some applications. Because this technique does not affect quasi-identifiers, it does not help achieve k-anonymity and hence has not been considered before. Removing a sensitive value in a group reduces diversity and therefore, it does not help in achieving l-diversity. However, in t-closeness, removing an outlier may smooth a distribution and bring it closer to the overall distribution. Another possible technique is to generalize a sensitive attribute value, rather than hiding it completely. An interesting question is how to effectively combine these techniques with generalization and suppression to achieve better data quality in cloud database system.

REFERENCES

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

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

[3] P. Samarati, "Protecting Respondent's Privacy in Microdata Release," *IEEE Trans. on Knowledge and Data Engineering (TKDE)* vol. 13, no. 6, pp. 1010-1027, 2001.

[4] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," *Int'l J. Uncertain. Fuzz.,* vol. 10, no. 5, pp. 557-570, 2002.

[5] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a Metric for Image Retrieval," *Int.l J. Computer Vision,* vol. 40, no. 2, pp. 99-121, 2000.

[6] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," *Int'l J. Uncertain. Fuzz.,* vol. 10, no. 6, pp.571-588, 2002.

[7] Closeness: A New Privacy Measure for Data Publishing Ninghui Li, Member, IEEE, Tiancheng Li, and Suresh Venkatasubramanian. ieee transactions on knowledge and data engineering, vol. 22, no. 7, july 2010

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

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

[10] M.E. Nergiz, M. Atzori, and C. Clifton, "Hiding the Presence of Individuals from Shared Databases," Proc. ACM SIGMOD,pp. 665-676, 2007.