Closeness Privacy Measures Using Tree EMD for Data Disclosures

M.Srividya #1, J. Ranga Rajesh #2

#1 Student, Dvr & Dr. Hs Mic College Of Technology, Kanchikacherla,Krishna(dt)

#2 Assoc. professor, Dvr & Dr. Hs Mic College Of Technology, Kanchikacherla,Krishna(dt)

#1 msrividya12@gmail.com, #2 rangarajesh.j@gmail.com,

Abstract: Micro data publishing enables researchers and policy-makers to analyze the data and learn important information. But privacy is a key issue here. Existing privacy measures such as k-anonymity protects against identity disclosures, but it does not provide sufficient protection against attribute disclosures. Another privacy measure l-diversity attempts to solve this problem. But it is neither necessary nor sufficient to prevent attribute disclosures and fails at data utilization. So a base model called t-closeness and a more flexible privacy model called (n,t)-closeness was developed that achieves a better balance between privacy and utility. The base model t-closeness, which requires that the distribution of a sensitive attribute in any equivalence class is 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).(n,t)-closeness offers higher utility. These closeness measures require Probability distributions that are assessed using Earth Mover's Distance (EMD) measurement. We propose to use an efficient tree-based algorithm, Tree-EMD. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree. The number of unknown variables is reduced to O(N) from $O(N^2)$ of the original EMD.This paper introduces techniques that are used in the implementation of the Tree-EMD and performs extensive experiments to demonstrate its efficiency.

I INTRODUCTION

Government agencies and other organizations often need to publish micro data, e.g., medical data or census data, for research and other purposes. When releasing micro data, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified: 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.

Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is reidentified 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. To effectively limit disclosure, we need to measure the disclosure risk of an anonymized table. While k-anonymity protects against identity disclosure, it is insufficient to prevent attribute disclosure. A new notion of privacy, called l- diversity, which requires that the distribution of a sensitive attribute in each equivalence class has at least '"well represented" values. One problem with l-diversity is that it is limited in its assumption of adversarial knowledge. 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.

Later, a novel privacy notion called "closeness" was proposed. Based on 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. For privacy in Micro data publishing a base model called t-closeness and a more flexible privacy model called (n, t)-closeness were proposed. These closeness measures require Probability distributions that are assessed using Earth Mover's Distance (EMD) measurement. Side effects of using EMD include large number of unknown variables that are to be resolved and have high time complexities.

In this paper, we propose a new fast algorithm, *i.e.*, Tree-EMD Algorithm to compute EMD between histograms with $L1$ ground distance. The formulation of Tree-EMD is much simpler than the original EMD formulation. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree. It has only $O(N)$ unknown variables, which is significantly less than the $O(N2)$ variables required in the original EMD. The efficiency of this algorithm enables its application to handle problems that were previously prohibitive due to high time complexities. In order to reduce the computation times of the original distance, the proposed method uses the lower bounding distance.

II REALATED WORK

The problem of information disclosure has been studied extensively in the framework of statistical databases. A number of information disclosure limitation techniques have been designed for data publishing, including Sampling, Cell Suppression, Rounding, and Data Swapping and Perturbation. These techniques, however, insert noise to the data.

The first category of work aims at devising privacy requirements. The k-anonymity model assumes that the adversary has access to some publicly-available databases and the adversary knows who is and who is not in the table. A few subsequent works recognize that the adversary has also knowledge of the distribution of the sensitive attribute in each group. t-Closeness proposes that the distribution of the sensitive attribute in the overall table should also be public information.

We want to emphasize that l-diversity is still a useful measure for data publishing. l-diversity and our closeness measures make different assumptions about the adversary. l-Diversity assumes an adversity who has knowledge of the form "Carl does not have heart disease," while our closeness measures consider an adversary who knows the distributional information of the sensitive attributes. Our goal is to propose an alternative technique for data publishing that remedies the limitations of l-diversity in some applications. Privacy-preserving data publishing has been extensively studied in several other aspects.

First, background knowledge presents additional challenges in defining privacy requirements. Several recent studies have aimed at modeling and integrating background knowledge in data Anonymization. Second, several works considered continual data publishing, i.e., republication of the data after it has been updated. Nergiz et al. proposed σ-presence to prevent membership disclosure, which is different from identity/attribute disclosure. Wong et al. showed that knowledge of the anonymization algorithm for data publishing can leak extra sensitive information.

III BACK GROUND

The first category of work aims at devising privacy requirements. The k-anonymity model assumes that the adversary has access to some publicly-available databases and the adversary Knows who is and who is not in the table. A few subsequent works recognize that the adversary has also knowledge of the distribution of the sensitive

attribute in each group. T-Closeness proposes that the distribution of the sensitive attribute in the overall table should also be public information. Emphasize that L-diversity is still a useful measure for data publishing. L-diversity and our closeness measures make different assumptions about the adversary.

DEFINITION K-ANONYMITY Let T(A1….. Am) be a table, and QI be quasi-identifier associated with it. T is said to satisfy k-anonymity with respect to QI iff each sequence of values in T[QI] appears at least with k occurrences in T[QI] (T[QI] denotes the projection, maintaining duplicate tuples, of attributes QI in T.

THE L-DIVERSITY PRINCIPLE. An equivalence class is said to have L-diversity if there are at least "well-represented" values for the sensitive attribute. A table is said to have L- diversity if every equivalence class of the table has L-diversity. Machanavajjhala et al. gave a number of interpretations of the term "well represented" in this principle:

DISTINCT L-DIVERSITY The simplest understanding of "well represented" would be to ensure that there are at least 'distinct values for the sensitive attribute in each equivalence class. Distinct L-diversity does not prevent probabilistic inference attacks. An equivalence class may have one value appear much more frequently than other values, enabling an adversary to conclude that an entity in the equivalence class is very likely to have that value. This motivated the development of the following stronger notions of L-diversity.

PROBABILISTIC L-DIVERSITY. An anonymized table satisfies probabilistic L-diversity if the frequency of a sensitive value in each group is at most 1/l. This guarantees that an observer cannot infer the sensitive value of an individual with probability greater than 1/l.

ENTROPY L-DIVERSITY The entropy of an equivalence class E is defined to being 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.

A table is said to have entropy L-diversity if for every equivalence class E, Entropy (E) ≥ log L. Entropy L-diversity is stronger than distinct L-diversity. As pointed out in [23], in order to have entropy 'L-diversity for each equivalence class, the entropy of the entire table must be at least log (l). Sometimes, this may too restrictive, as the entropy of the entire table may be low if a few values are very common. This leads to the following less conservative notion of L-diversity.

L-diversity and Limitations. L-diversity requires that each equivalence class contains at least l "well-represented" values for the sensitive attribute. This is in contrast to the above definition of utility where the homogeneous distribution of the sensitive attribute preserves the most amount of data utility. In particular, the above definition of utility is exactly the opposite of the definition of entropy l-diversity, which requires the entropy of the sensitive attribute values in each equivalence class to be at least log l. '. Enforcing entropy l-diversity would require the information loss of each equivalence class to be at least log l. Also, as illustrated in [20], l'-diversity is neither necessary nor sufficient to protect against attribute disclosure.

While the 'L-diversity principle represents an important step beyond k-anonymity in protecting against attribute disclosure, it has several shortcomings that we now discuss. 'L-diversity may be difficult to achieve and may not provide sufficient privacy protection. The goal is to propose an alternative technique for data publishing that remedies the limitations of L-diversity in some application. An interesting question is how to effectively combine the existing techniques with generalization and suppression to achieve better data quality and privacy.

t-closeness. We show that t-closeness substantially limits the amount of useful information that the released table preserves. t-closeness requires

that the distribution of the sensitive attribute in each equivalence class to be close to the distribution of the sensitive attribute in the whole table. Therefore, enforcing t-closeness would require the information loss of each equivalence class to be close to the entropy of the sensitive attribute values in the whole table. In particular, a 0-close table does not reveal any useful information at all and the utility of this table is computed as Note that in a 0-close table, H(Ei) = H(T) for any Equivalence class Ei, ($1 \leq i \leq P$).

(n,t)-closeness. The (n, t) closeness model allows better data utility than t-closeness. Given an anonymized table {E1,……,Ep} where each Ei($1 \leq i \leq P$) is an equivalence class and another anonymized table {G1, . . .,Gd} where each Gj ($1 \leq j \leq d$) is the union of a set of equivalence classes in {E1, . . . ,Ep} and contains at least n records. The anonymized table {E1, . . . ,Ep} satisfies the (n, t)-closeness requirement if the distribution of the sensitive attribute in each Ei ($1 \leq i \leq p$) is close to that in Gj containing Ei. By the above definition of data utility, the utility of the anonymized table {E1, . . . ,Ep} is computed as

We are thus able to separate the utility of the anonymized table into two parts: 1) the first part U{G1; . . .;Gd} is the sensitive information about the large groups {G1; . . .;Gd}and 2) the second part is further sensitive information about smaller groups. By requiring the distribution of the sensitive attribute in each Ei to be close to that in the corresponding Gj containing Ei, the (n,t)-closeness principle only limits the second part of the utility function and does not limit the first part. In fact, we should preserve as much information as possible for the first part

IV Earth Mover's Distance

The earth mover's distance (EMD) was introduced in computer vision as an improved distance measure between two distributions, and it has been widely used in multimedia databases The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other.

EMD for numerical attributes: The EMD is based on the minimal amount of work needed to transform one distribution to another by moving distribution mass between each other. It is straightforward to verify that the ordered-distance measure is a metric. It is non-negative and satisfies the symmetry property and the triangle inequality. To calculate EMD under ordered distance, we only need to consider flows that transport distribution mass between adjacent elements, because any transportation between two more distant elements can be equivalently decomposed into several transportations between adjacent elements.

EMD for categorical attributes: a total order often does not exist. Two distance measures are considered. **Equal Distance:** The ground distance between any two values of a categorical attribute is defined to be 1. It is easy to verify that this is a metric. As the distance between any two values is 1, for each point that $pi - qi > 0$, one just needs to move the extra to some other points. **Hierarchical Distance:** The distance between two values of a categorical attribute is based on the minimum level to which these two values are generalized to the same value according to the domain hierarchy.

V Tree- Earth Mover's Distance

We introduce Tree-EMD with $L1$, a novel efficient formulation of EMD. We first show that, by using the $L1$ (Manhattan) distance as the ground distance. We designed a tree-based algorithm as an efficient discrete optimization solver, which extends the original simplex algorithm. The tree-based algorithm is significantly faster than the original simplex, and has a more intuitive interpretation as a network flow problem.

The simplex algorithm is a popular solution to linear programming problems because of its average polynomial time complexity. The simplex algorithm searches for the optimum solution among

the space of *basic feasible (BF) solutions*. A BF solution is a solution of such that only a fixed number of variables can be non-zero. These variables are called *basic variable* (BV) flows in our formulation.

The original EMD is solved by the transportation simplex (TS) algorithm by taking advantage of the special structure of the original EMD formulation. To exploit this similarity, we designed an *extended transportation simplex* (ETS) for EMD. The basic idea of ETS is to intelligently update the BF solution g, c and A during the simplex iteration.

A tree-based algorithm, *Tree-EMD*, can be naturally extended from ETS. First, an initial BFT is built. Then the BFT is iteratively replaced by a better BFT until the optimum is reached. Compared to ETS, Tree-EMD is more efficient because it avoids the brute force search used in the ETS algorithm and there is no need to update the whole vector.

VI PERFORMANCE

The Tree-EMD algorithm is presented in several issues:

(1) *The root of a BFT*: The root *r* is heuristically set to be the center of the graph. This is to make the tree as balanced as possible. Once *r* is fixed, the *u* value at *r* is fixed to 0.

2) *Build the initial BFT*: The nodes are considered sequentially, in a left-to-right and bottom-to-top order, i.e., starting from bottom-left node. When processing node *q*, all the flows connecting its lower and left neighbors are fixed. As a result, only one BV flow needs to be chosen between *q* and either its upper or right neighbor such that the flow makes the weight at *q* vanish.

The Tree-EMD algorithm can also be generalized to solve the original EMD problem (i.e. beyond histograms) for speedup. This is because the tree structure used in Tree-EMD is also true for the transportation simplex used in the original EMD. In addition, as indicated EMD can also be modeled as a network flow problem. This raises interest in the underlying relationship between the tree-based algorithm and network flow algorithms. It may be a key to find more efficient solutions the original EMD.

VII CONCLUSION

Existing privacy measures such as k-anonymity protects against identity disclosures, but it does not provide sufficient protection against attribute disclosures. Another privacy measure l-diversity attempts to solve this problem. But it is neither necessary nor sufficient to prevent attribute disclosures and fails at data utilization. So a base model called t-closeness and a more flexible privacy model called (n, t)-closeness were developed that achieves a better balance between privacy and utility.). (n, t)-closeness offers higher utility. These closeness measures require Probability distributions that are assessed using Earth Mover's Distance (EMD) measurement. We propose to use an efficient tree-based algorithm, Tree-EMD. Tree-EMD exploits the fact that a basic feasible solution of the simplex algorithm-based solver forms a spanning tree. The formulation of Tree-EMD is much simpler than the original EMD formulation. The efficiency of this algorithm enables its application to handle problems that were previously prohibitive due to high time complexities.

VIII REFERENCES

[1] 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.

[2] 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.

[3] G. T. Duncan, S. E. Fienberg, R. Krishnan, R. Padman, and S. F. Roehrig. Disclosure limitation methods and information loss for tabular data. In *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies*, pages 135–166. Elsevier, 2001.

[4] G. T. Duncan and D. Lambert. Disclosure-limited data dissemination. *J. Am. Stat. Assoc.*, pages 10–28, 1986.

[5] S. Cohen, L. Guibas. "The Earth Mover's Distance under Transformation Sets", *IEEE International Conference on Computer Vision*, II: 1076-1083, 1999.

[6] S. Cohen and L. Guibas, "The Earth Mover's distance: lower bounds and invariance under translation," Tech. Rep. CS-TR- 97-1597, Stanford University, 1997.

[7] I. Assent, A. Wenning, and T. Seidl, "Approximation techniques for indexing the earth mover's distance in multimedia databases," in *Proceedings of the 22nd International Conference on Data Engineering (ICDE '06)*, p. 11, usa, April 2006.

[8] Ling, H., Okada, K.: An e_cient earth mover's distance algorithm for robust histogram comparison. IEEE Trans. Pat. Anal. Mach. Intell. 29 (2007) 840-853.

[9] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.

[10] P. Samarati and L. Sweeney. Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression. Technical Report SRI-CSL-98-04, SRI Computer Science Laboratory, 1998.