A STUDY ON DATA ANONYMIZATION FOR PRIVACY- PRESERVING DATA PUBLISHING

Dr. B. Santhosh Kumar [1] , CH.Vivek Sumanth[2]
[1], Sr.Assistant Professor/CSE, GMR Institute of Technology, Razam, Andhra Pradesh, India
[2], UG Scholar/CSE, GMR Institute of Technology, Razam, Andhra Pradesh, India

ABSTRACT

We demonstrate that l-diversity has various limitations. In particular, it is neither essential nor sufficient to prevent attribute disclosure. Motivated by these limitations, we propose another notion of protection called "closeness." We present the base model t-closeness, which necessitates that the distribution of a multiple 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 close to a threshold t). We then propose an increasingly flexible protection model called on; t-closeness that offers higher utility. We portray our desiderata for structuring a distance measure between two probability distributions and present two distance measures. The k-anonymity protection requirement for publishing smaller scale data necessitates that every equivalence class (i.e., a set of records that are indistinguishable from one another with respect to certain "identifying" attributes) contains at least k records. Recently, several authors have perceived that k-anonymity cannot prevent attribute disclosure. The notion of l-diversity has been proposed to address this; l- diversity necessitates that every equivalence class has at least 'well-represented values for each multiple sensitive attribute. We propose a novel protection notion called "closeness." We formalize the possibility of global background knowledge and propose the base model t-closeness which necessitates that the distribution of a multiple 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 close to a threshold t).

Keywords: Anonimity, Diversity, Closeness, Privacy

INTRODUCTION

GOVERNMENT offices and other organizations often need to publish small scale data, e.g., medical data or enumeration data, for research and other purposes. Typically, such data are stored in a table, and each record (push) relates to one individual. Each record has various attributes, which can be separated into the following three categories: 1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as semi identifiers, and may include, e.g., Zip code, Birth-date, and Gender. 3) Attributes that are viewed as sensitive, for example, Disease and Salary.

While releasing miniaturized scale data, it is important to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature identity disclosure and attribute disclosure. Identity disclosure happens when an individual is linked to a particular record in the released table. The protection saving data mining problem has increased considerable importance in recent years be-reason for the vast amounts of personal data about individuals stored at critical rent commercial merchants and organizations.

Protection safeguarding data mining has risen as an important issue to be tended to in recent times. This is a direct result of the ability to store data of clients had expanded , utilization of social

networks helps in yielding personal information , sophisticated data mining algorithms and high computational forces available with the foe. All this makes it possible to leverage this information. Although most of the applications rest evacuate the records having sensitive information like name, social security numbers (or some other exceptional identification number), other kind of attributes like sex, age, stick codes, calling can be joined to shape a pseudo-identifier and the sensitive information would then be able to be retrieved from public data records like registration which contain all records.

TABLE 2
A 3-Anonymous Version of Table 1

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

There can be two wide approaches to accomplish the goal of security. First is to release limited data with the end goal that personal information cannot be extracted out of it but the overall heuristics are still close to original dataset. Furthermore, second is to pre-compute heuristics and release them instead of any data. Advantage of releasing some limited data instead of pre-computed heuristics is an expanded edibility and availability for the clients. So in Privacy Preserving Data Mining we look for methods to transform the original data to such an extent that heuristics determined from the transformed data are close to original heuristics and the security of clients isn't imperiled

K-ANONYMITY

The k-anonymity model requires that within any equivalence class of the micro data there are at least k records. In other words we should not be able to make ANY query to the database which returns less than k matches. Achieving k-anonymity is provided by use of generalization relationships between domains and between values that attributes can assume. Suppression is a complementary approach to providing k-anonymity.

Today's globally networked society places great demand on the dissemination and sharing of information, which is probably becoming the most important and demanded resource. While in the past released information was mostly in tabular and statistical form (macro data), many situations call today for the release of specific data (micro data). Micro data, in contrast to macro data reporting pre computed statistics, provide the convenience of allowing the final recipient to perform on them analysis as needed. To protect the anonymity of the entities, called respondents, to which Micro data undergoing public or semipublic release refer, data holders often remove or encrypt explicit identifiers such as names, addresses, and phone numbers. De-identifying data, however, provides no guarantee of anonymity. Released information often contains other data, such as race, birth date, sex, and ZIP code, which can be linked to publicly available information to re-identify (or restrict the uncertainty about) the data respondents, thus leaking information that was not intended for disclosure.

The large amount of information easily accessible today, together with the increased computational power available to the attackers, makes such linking attacks a serious problem. Indeed, the restricted access to information and its expensive processing, which represented a form of protection in the past, do not hold anymore. Information about us is collected every day, as we join associations or groups, shop for groceries, or executes most of our common daily activities; the amount of privately owned records that describe each citizen's finances, interests, and demographics is increasing every day. Information bureaus such as TRW, Equifax, and Trans Union hold the largest and most detailed databases on American consumers. Most municipalities sell population registers that include the identities of individuals along with basic demographics; examples include local census data, voter lists, city directories, and information from motor vehicle agencies, tax assessors, and real estate agencies.

The k-anonymity model assumes that person-specific data are stored in a table (or a relation) of columns (or attributes) and rows (or records). The process of anonymizing such a table starts with removing all the explicit identifiers, such as name and SSN, from it. However, even though a table is free of explicit identifiers, some of the remaining attributes in combination could be specific enough to identify individuals. For

example, as shown by Sweeney, 87% of individuals in the United States can be uniquely identified by a set of attributes such as {ZIP, gender, date of birth}. This implies that each attribute alone may not be specific enough to identify individuals, but a particular group of attributes could be. Thus, disclosing such attributes, called quasi-identifier, may enable potential adversaries to link records with the corresponding individuals.

Definition 1. (Semi identifier) A semi identifier of table T, denoted as QT, is a set of attributes in T that can be potentially used to link a record in T to a real-world identity with a significant probability. The primary objective of the k-anonymity problem is thus to transform a table with the goal that nobody can make high-probability associations between records in the table and the comparing entity instances by utilizing semi identifier.

Definition 2. (k-anonymity requirement) Table T is said to be k-mysterious with respect to semi identifier QT if and only if for each record r in T there exist at least (k − 1) other records in T that are indistinguishable from r with respect to QT. _ By implementing the k-anonymity requirement, it is guaranteed that despite the fact that a foe knows that a k-mysterious table T contains the record of a particular individual and furthermore knows the semi identifier value of the individual, he cannot determine which record in T compares to the individual with a probability greater than 1/k. The k-anonymity requirement is typically upheld through generalization, where real values are replaced with "less explicit but semantically consistent values". Given an area, there are different approaches to generalize the values in the space. Commonly, numeric values are generalized into intervals, and categorical values into a set of distinct values (e.g., {USA, Canada}) or a single value that represents such a set (e.g., North-America). A gathering of records that are indistinguishable from one another is often alluded to as an equivalence class

T-CLOSENESS: A NEW PRIVACY MEASURE

Intuitively, security is estimated by the information addition of a spectator. Prior to seeing the released table, the eyewitness has some earlier belief about the multiple sensitive attribute value of an individual. After observing the released table, the spectator has a posterior belief. Information addition can be represented as the contrast between the posterior belief and the earlier belief. The novelty of our methodology is that we separate the information gain into two parts: that about the whole population in the released data and that about explicit individuals.

LIMITATIONS OF T-CLOSENESS

There is no computational technique to authorize t-closeness followed in. There is effective path till now of joining with generalizations and concealments or slicing. Lost co-relation between different attributes: This is on the grounds that each attribute is generalized separately thus we lose their reliance on one another. Utility of data is harmed on the off chance that we utilize small t. (Also, small t will result in increment in computational time.

EARTH MOVER'S DISTANCE

In probability theory, the earth mover's distance (EMD) is a proportion of the distance between two probability distributions over a locale D. Informally, if the distributions are interpreted as two different methods for piling up a certain amount of dirt over the area D, the EMD is the base cost of turning one pile into the other; where the cost is thought to be amount of dirt moved times the distance by which it is moved.

The EMD depends on the minimal amount of work expected to transform one distribution to another by moving distribution mass between one another. Intuitively, one distribution is viewed as a mass of earth spread in the space and the other as a collection of holes in a similar space. EMD measures the least amount of work expected to fill the holes with earth. A unit of work compares to moving a unit of earth by a unit of ground distance. EMD can be formally characterized utilizing the well-studied transportation problem. Let

$$\mathbf{P} = (p_1, p_2, \ldots p_m), \mathbf{Q} = (q_1, q_2, \ldots q_m),$$

and dij be the ground distance between element i of P and element j of Q. We want to find a flow F ¼ ½fij_, where fij is the flow of mass from element i of P to element j of Q that minimizes the overall work:

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

subject to the following constraints:

$$f_{ij} \geq 0, \qquad 1 \leq i \leq m, 1 \leq j \leq m, \quad (c1)$$

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

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

The EMD is characterized as the base amount of work expected to transform one signature into the other. The notion of "work" depends on the client characterized ground distance which is the distance between two features. The measure of the two signatures can be different. Also, the aggregate of weights of one signature can be different than the total of weights of the other (partial match). Along these lines, the EMD is normalized by the smaller entirety.

RELATED WORK

In this section, we briefly overview existing literature that tends to data security. Instead of giving an extensive overview, we talk about different aspects of data security. Note that Ensuring security in published data has been a difficult problem for a long time, and this problem has been studied in different aspects. In, Lambert gives informative dialog on the risk and damage of undesirable disclosures and examines how to evaluate a dataset in terms of these risk and mischief. In, Dalenius represents the problem of re-identification in (supposedly) unknown enumeration records and firstly introduces the notion of "semi identifier". He also suggests a few thoughts, for example, concealment or encryption of data as possible solutions.

Data protection has been extensively tended to in statistical databases, which primarily go for preventing different induction channels. One of the basic techniques is data perturbation which mostly involves swapping data values or introducing clamor to the dataset. While the perturbation is applied in ways which jam statistical characteristics of the original data, the transformed dataset is useful only for statistical research. Another important technique is inquiry restriction, which restricts inquiries that may result in deduction. In this methodology, questions are restricted by different criteria, for example, inquiry set-estimate, inquiry history, and partitions. Although this methodology can be effective, it requires the protected data to stay in a dedicated database at all time.

CONCLUSION

We propose two instantiations: a base model called t-closeness and a progressively flexible security model called (n-t)- closeness. We explain the rationale of the (n-t)- closeness model and demonstrate that it accomplishes a better balance between protection and utility. To incorporate semantic distance, we utilize the Earth Mover Distance measure. We also point out the limitations of EMD, present the desiderata for planning the distance measure, and propose another distance measure that meets all the requirements. Finally, through experiments on real data, we demonstrate that similarity attacks are a real concern and the (n-t)- closeness model better protects the data while improving the utility of the released data.

(n-t)- Closeness allows us to take advantage of anonymization techniques other than generalization of semi identifier and concealment of records. For example, instead of smothering a whole record, one can shroud some sensitive attributes of the record; one advantage is that the quantity of records in the anonymized table is accurate, which might be useful in a few applications. Since this technique does not affect semi identifiers, it doesn't help accomplish k-anonymity, and consequently, has not been considered previously. Evacuating a sensitive value in a gathering diminishes diversity, and therefore, it doesn't help in accomplishing '- diversity. Be that as it may, in t-closeness, evacuating an outlier may smooth a distribution and convey it closer to the overall distribution. Another possible technique is to generalize a sensitive attribute value, rather than concealing it completely. An interesting question is the manner by which to effectively consolidate these techniques with generalization and concealment to accomplish better data quality.

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 Statistics to 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] B.C.M. Fung, K. Wang, and P.S. Yu, "Top-Down Specialization for Information and Privacy Preservation," Proc. Int'l Conf. Data Eng. (ICDE), pp. 205-216, 2005.
[10] C.R. Givens and R.M. Shortt, "A Class of Wasserstein Metrics for Probability Distributions," Michigan Math J., vol. 31, pp. 231-240,
1984.
[11] V.S. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc. ACM SIGKDD, pp. 279-288, 2002.
[12] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets," Proc. ACM SIGMOD, pp. 217-228, 2006.
[13] N. Koudas, D. Srivastava, T. Yu, and Q. Zhang, "Aggregate Query Answering on Anonymized Tables," Proc. Int'l Conf. Data Eng. (ICDE), pp. 116-125, 2007.
[14] S.L. Kullback and R.A. Leibler, "On Information and Sufficiency," Annals of Math. Statistics, vol. 22, pp. 79-86, 1951.
[15] D. Lambert, "Measures of Disclosure Risk and Harm," J. Official Statistics, vol. 9, pp. 313-331, 1993.