A Dynamic Privacy Preserving Data Publishing Using Three Metrics Technique

Allam Naga Sudha Madhuri[1], S. Mabjan[2]
[1]*PG Scholar, Department of CSE, Bharath College of Engineering and Technology for Women, Kadapa*
[2]*Assistant professor, Department of CSE, Bharath College of Engineering and Technology for Women, Kadapa*

Abstract **- In recent years, privacy preserving has seen rapid growth which leads to an increase in the capability to store and retrieve personal dataset without revealing the sensitive information about the individuals. Different techniques have been proposed to improve accuracy in a huge sourcing database. Anonimization techniques such as, generalization is designed for improving accuracy in privacy preserving method. But the malicious workers can hack the private information of the user and misuse it. Recent work has shown that anonymity for generalization loses significant amounts of information, especially for data of higher dimensionality. Collecting and publishing large amounts of individuals' data to public for purposes such as medical research, market analysis and economical measures has increased major privacy concerns about individual's sensitive information. Many Privacy-Preserving Data Publishing (PPDP) techniques have been proposed in literature to act. But the result is no proper privacy characterization and measurement. In this paper we introduce a novel technique called overlapped slicing, which partitions the data in both horizontal and vertical. Slicing preserves better data utility than generalization techniques. As an extension we proposed a technique called random attribute slicing, in which an attribute is divided into more than one column. Important advantage of this work is to handle high-dimensional data and also preserves better privacy than the previous techniques.**

Index Terms **- Data privacy, data security, data publishing, privacy leakages, slicing.**

I.INTRODUCTION

Nowadays, datasets are considered a valuable source of information for the medical research, market analysis and economical measures. These datasets can include information about individuals that contain social, medical, statistical, and customer data. Many organizations, companies and institutions publish privacy related datasets. while the shared data set provides researchers with valuable social knowledge, it also raises security risks and questions about confidentiality for individuals whose data are in the table. To avoid possible identification of individuals from records in published data, uniquely identifying information usually, names and social security numbers are omitted from the table. While the obvious personal identifiers are removed, the quasi-identifiers such as zip code, age and gender can still be used to uniquely identify a significant portion of the population since the released data allows individuals to infer and restrict the operations available without releasing the table. In fact, [1] showed that by correlating this data with the publicly available side information, such as information from voter registration list for Cambridge Massachusetts, medical visits about many individuals could be easily identified [2]. This study estimated that 87 percent of the U.S population could be uniquely identified using quasi-identifiers by side information-based attacks, including the Governor of Massachusetts medical records in medical data. The spate of events related to privacy has sparked a long line of information publishing and evaluation work into privacy concepts, such as k-anonymity, l-diversity and t-closeness. A table satisfies kanonymity if at least k-1 other quasi identifier attributes are indistinguishable from each quasi-identifier attribute in the table; such a table is called a k-anonymous table. Although k-anonymity prevents individual identity disclosure by linking attacks, it is not enough to prevent disclosure of attributes with side data. It makes it possible to infer the possible sensitive attributes corresponding to a person by combining the released data with side information. Once the correspondence between the identifier and the sensitive attributes is revealed for an

individual, it can affect the person and the table as a whole. Ldiversity was implemented in [4] to address this issue, l-diversity allows the responsive attributes to include, in each equivalence group, at least well-expressed values. As stated in [5], l-diversity has two major problems. One, is that it limits the adversarial knowledge, while it is possible to acquire knowledge of a sensitive attribute from generally available global distribution of the attribute. Another problem is that all attributes are assumed to be categorical, which assumes that the adversary either gets all the information or gets nothing for a sensitive attribute. In [5], researchers suggested a notion of confidentiality called t-closeness. First , they formalize the idea of global knowledge of background and propose the t-closeness base model. This model requires the distribution in any equivalence class of a sensitive attribute to be similar to the distribution of the attribute in the overall table (i.e, the difference should be no more than a threshold t of the two distributions). This distance was introduced to calculate the information gained by the Earth Mover Distance (EMD) metric between subsequent belief and prior belief [10], which is represented as the information gain for a specific individual over the entire population. Moreover, as we show in this paper, the distance between two distributions cannot be easily quantified by a single measurement, t-closeness also has many limitations. Research on data privacy has purely been focused on privacy definitions, like k-anonymity, l-diversity, tcloseness, etc. while these models consider only minimizing the amount of privacy leakage without directly measuring what the opponent might learn, there is a motivation to find consistent measurements of how much information is leaked to an opponent by publishing a dataset. In this paper, introduced our novel data publishing framework. The proposed framework consists of two steps. First, we model attributes in a dataset can be modeled as a multi-variable model. Based on this model, we are able to re-define the prior and posterior adversarial belief about attribute values of individuals. Based on the privacy risks attached we characterize the privacy of their individuals with combining different attributes. This precise model is to described as indeed privacy risk of publishing datasets. For a given dataset, we have to determine to what extent we can achieve privacy before it is released. Therefore, we introduced two privacy leakage measurements: distribution leakage and entropy leakage. They explain the reasoning for these two measures and use examples to illustrate and use examples to illustrate their benefits. We show how considering only one metric ignoring the effect of the other strongly contributes to the information leakage and in turn affects the privacy. An intuitive example for this problem is reviewing a blood tests. Based on only one measure patient's medical status can't be determined even if this particular measure is the most sensitive one. Instead, a physician has to review the relation between combinations of all measures in the blood test. We show that a minimized distribution leakage between sensitive attribute values distributions of the original and the published datasets does not essentially achieve the minimum entropy leakage that an adversary could gain. In fact, we show that distribution and entropy leakage are two different measures

II. LITERATURE SURVEY

Survey on "Data privacy through optimal kanonymization

The requirement for information deidentification to unlock data for research purposes and individuals request for privacy is reconciled. This paper proposes a system de-identification optimization algorithm known as k-anonymization. The property of an anonymized dataset is that each record cannot be separated from at least others. Optimized anonymization is NP-hard, leading to significant computational challenges. We present a new approach to explore the space of potential anonymizations which weaken the problem's combinatorics and establish data management strategies to minimize costly operations like sorting. Via experiments on real census data, we display the resulting algorithm in two representative cost measures and a wide range of. We also demonstrate that in situations where input data or input parameters prevent finding an optimal solution in a reasonable time, the algorithm may produce good anonymizations. Finally we use the slicing algorithm to investigate the impact on anonymization quality and performance of various coding methods and problem variations. To our knowledge, optimal anonymization under a general model of the problem is the first to demonstrate the outcome of a nontrivial dataset. Survey on "Top-down specialization for information and privacy preservation," The most sensitive state that poses a threat to individual privacy is a person-

specific data. This presents an efficient slicing algorithm for determining a generalized version that masks sensitive information for modeling classification. These generalization is implemented in a top-down manner by specializing or detailing the level of information until a minimum privacy requirement is violated. This specialization is natural and efficient for both categorical and continuous attributes to handle. Our approach has a fact that exploits the data usually contains redundant structures for classification and generalization may eliminate some structures. Our results show that quality of classification that can preserves for highly restrictive privacy requirements. This work is applicable to both public and private sectors which shares the information for productivity and mutual benefits. Survey on "t-closeness: Privacy beyond kanonymity and l-diversity" Each equivalence class (i.e, a collection of records that cannot be separated from each other's "identifying" attributes) includes at least k records for micro data publishing. Authors felt that k-anonymity could not prevent disclosure of attributes. The notion that each equivalence group has at least well-represented values for each responsive attribute has been proposed for l-diversity. We found that there are a range of drawbacks to l-diversity. In general, it is neither necessary nor sufficient to avoid disclosure of attributes. A novel definition of privacy called t-closeness has been introduced, requiring a responsive attribute in any equivalence class near the distribution of the attribute in the overall table (i.e, the distance between the two distributions must be no more than a threshold t). For our t-closeness requirement, we selected the Earth Mover Distance measure. Survey on "Robust de-anonymization of large sparse datasets" Attacks of de-anonymization against high dimensional micro-data, such as user preferences suggestions, payment information, etc. Our methods are versatile in order to interrupt the information and to accommodate certain errors in the background knowledge of the opponent. For the Netflix Prize dataset, which contains anonymous movie reviews from 500,000 Netflix subscribers, we submitted to the world's largest online video rental service. It demonstrates that by understanding only a little bit, an individual subscriber could easily identify the history of this subscriber in the dataset. By using the Internet Movie Database as the basis of background knowledge, we successfully identified Netflix accounts of verified users

III. PROPOSED WORK

We are introducing a novel technique called hybrid anonymization, the combination of generalization and slicing technique. By using the slicing Technique, data can be partitioned in to horizontally and vertically. Slicing can be used to shield membership transparency and also provides better data quality than generalization. The processing of high-dimensional data is another advantage of slicing. We demonstrate how to use slicing to secure disclosure attributes and create an effective slicing algorithm to compute the sliced information that obeys the diversity requirement. Our research confirms that slicing retains better utility than widespread use and is more effective in integrating the sensitive attribute.

1.We implement a new information anonymization technique called hybrid anonymization to improve the current state of the art.

2. We show that it is used efficiently to avoid disclosure of attributes on the basis of diversity's privacy criteria.

3. We create an effective algorithm for calculating the sliced table to satisfy ldiversity.

4. We perform detailed tests on the workload and confirm that our slicing findings maintain much better data usefulness than generalization.

Disadvantages:

For two attributes only, data can be seen more safely

Generalized Data

No other distribution assumption can be justified in order to perform generalization data analysis or data mining activities on the generalized table to make the uniform distribution assumption that each interval or array is equally possible.

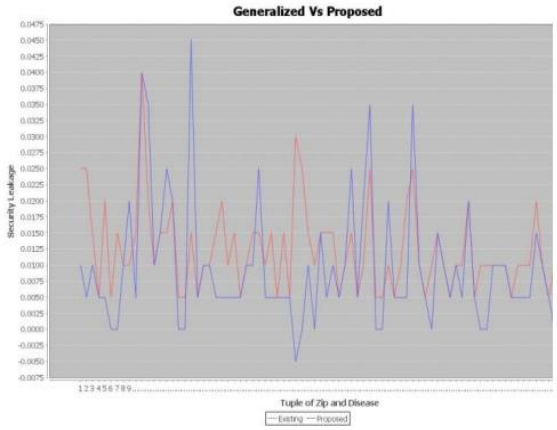
Sliced Data

Slicing is another important advantage for handling high-dimensional data. Slicing by partitioning into columns reduces the dimensionality of the dat. Each table can be presented as a sub-table with each column's lower dimensionality. The different approach that is related in slicing is to publish multiple independent sub-tables.

IV. METHODOLOGY

We compare the results of Generalized and Hybrid Anonymization in terms of data leakage of sensitive

data. X axis we taken distinct tuple of zip and disease data. In Y axis we taken the two techniques of Generalized and proposed slicing results with respect to same X-axis.



V.CONCLUSION

This paper presents a new approach to privacy-concerning micro data publishing called hybrid slicing. Slicing surpasses generalization limits and retains greater utility when defending against risks to privacy. Through explaining slicing, we demonstrated how to avoid disclosure of attributes and disclosure of membership. The experiment shows that the proposed slicing retains better data usefulness than generalization and is more efficient in workloads involving the critical attribute than generalization. In future research, this work motivates many ways. First, in this article, we find slicing in exactly one row where each of the two attributes are. It is a notion extension that overlaps a slicing, duplicating an attribute in more than one column. This will unlock further comparisons of attributes. Of example, the characteristic of the diseases can also be included in the first row. The two columns {Age, Sex, Disease} and { Zipcode, Disease} have better data usefulness, but the ramifications of confidentiality need to be researched and understood carefully. Studying the trade-off between privacy and utility is fascinating.

REFERENCE

[1] L. Sweeney, “k-anonymity: A model for protecting privacy,” International Journal on Uncertainty, Fuzziness and Knowledge based Systems, vol. 10, no. 5, pp. 557–570, 2002.

[2] L. Sweeney, “Uniqueness of simple demographics in the U.S. population,” 2000.

[3] A. Narayanan and V. Shmatikov, “Robust de-anonymization of large sparse datasets,” in Security & Privacy, pp. 111– 125, 2008.

[4] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam, “`-diversity: Privacy beyond k-anonymity,” ACM Trans. Knowl. Discov. Data, vol. 1, Mar. 2007.

[5] N. Li, T. Li, and S. Venkatasubramanian, “t-closeness: Privacy beyond k-anonymity and l-diversity,” in ICDE, pp. 106–115, 2007.

[6] N. Li, W. Qardaji, D. S. Purdue, Y. Wu, and W. Yang, “Membership privacy: A unifying framework for privacy definitions,” in CCS, (Berlin, Germany), 2013.

[7] I. Wagner and D. Eckhoff, “Technical privacy metrics: a systematic survey,” CoRR, vol. abs/1512.00327, 2015.

[8] J. Zhang, G. Cormode, C. M. Procopiuc, D. Srivastava, and X. Xiao, “Privbayes: Private data release via bayesian networks,” in Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data, SIGMOD ’14, (New York, NY, USA), pp. 1423– 1434, ACM, 2014.

[9] M. Gotz, S. Nath, and J. Gehrke, “Maskit: Privately releasing ¨ user context streams for personalized mobile applications,” in Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data, SIGMOD ’12, (New York, NY, USA), pp. 289– 300, ACM, 2012.

[10] Y. Rubner, C. Tomasi, , L. J., and Guibas, “The earth mover’s distance as a metric for image retrieval,” International Journal of Computer Vision, vol. 40, no. 2, pp. 99–121, 2000.

[11] D. Rebollo-Monedero, J. Forne, and J. Domingo-Ferrer, “From tcloseness-like privacy to post randomization via information theory,” IEEE Trans. on Knowl. and Data Eng., vol. 22, pp. 1623–1636, Nov. 2010.