Privacy Preservation Measurement through Diversity and Anonymity using Closeness

[1]Anil Prakash Dangi
RGPV, Bhopal
Asst. Prof. MIT Mandsaur

[2]Shayamlal Kumawat
RGPV, Bhopal
Asst. Prof. MIT Mandsaur

[3]Ravijeet Singh Chauhan
RGPV, Bhopal
Asst. Prof. MIT Mandsaur

ABSTRACT

Public survey data that may Increase the exposure of privacy and census information about the particulars is called the Data sensitivity. To maintain the privacy increase the similarity in the data item and introduce redundancy in such a way that information about individual users can not be disclose. This technique also desires the actual information of the data to not change. In this work we propose a unique method by combining two of the most widely used privacy preservation techniques: K-anonymity and l-diversity. The k-anonymity privacy requirement for publishing micro data requires that each equivalence class contains at least k records. Diversity requires that each equivalence class has at least well-represented values for each sensitive attribute. it is neither necessary nor sufficient to prevent attribute disclosure. Motivated by these limitations, we propose a new notion of privacy called k-anonymity,l-diversity, closeness. We first present 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). Based on entropy based closeness and distance measure between the class of data we propose a comprehensive technique to change the dataset to preserve the privacy while keeping the original meaning intact.

1. INTRODUCTION

Government agencies and other organizations often need to publish Microdata, typically, such data is stored in a table and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories.

1.1. Explicit Identifiers (EI):Attributes that clearly identify individuals, e.g., Pan Card no. or Enrolment number. ***2.2. Quasi-Identifiers (QI)***: Attributes whose values when taken together can potentially, may include, e.g., Village Name, Address. ***3.3. Sensitive-Identifiers (SI):***Attributes that are considered sensitive. Such as *salary.*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 in the literature:

1. Identity disclosure and

2. Attribute disclosure.

1. Identity Disclosure (ID): occurs when an individual is linked to a particular record in the released table. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is re-identified and the corresponding sensitive values are revealed.

2. Attribute Disclosure (AD): 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. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm [15].

An observer of a released table 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.

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 is to limit the disclosure risk to an acceptable level while maximizing the benefit.

Privacy preservation technique is a way to increase the similarity of the data item and introduce redundancy in such a way that information about individual users can not be revealed. This technique also desires the actual information of the data to not change.

For example government starts a new scheme for giving salary based pension for category for low salaried people and a person's name occurs in that table than easily his salary can be obtained. 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. 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 is to limit the disclosure risk to an acceptable level while maximizing the benefit.

1.2 Privacy Measurements

The first category of work aims at devising privacy requirements. The k-anonymity model [32], [31] assumes that the adversary has access to some publicly available databases (e.g., a voter registration list) and the adversary knows who is and who is not in the table. A few subsequent works [23], [33], [36], [38] recognize that the adversary also has knowledge of the distribution of the sensitive attribute in each group. T-Closeness [19] 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. -diversity and our closeness measures make different assumptions about the adversary. - 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 -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 [24], [7], [20], [22] have aimed at modeling and integrating background knowledge in data anonymization. Second, several works [6], [41] considered continual data publishing, i.e., re-publication of the data after it has been updated. Nergiz et al. [26] proposed σ-presence to prevent membership disclosure, which is different from identity/attribute disclosure. Wong et al. [37] showed that knowledge of the anonymization algorithm for data publishing can leak extra sensitive information.

1.3 Anonymization Techniques

Most anonymization solutions adopt generalization [1], [4], [9], [11], [16], [21], [17], [30], [31] and bucketization [13], [24], [39]. In this paper, we use the Mondrian algorithm [17] which partitions the high-dimensional space into regions and encodes data points in one region by the region's representation. There are several anonymization techniques, cluster- ing [2], marginals releasing [12], and data perturbation [28]. On the theoretical side, optimal k-anonymity has been proved to be NP-hard for $k \geq 3$ in [25], and approximation algorithms for finding the anonymization that suppresses the fewest cells have been proposed in [25], [27]. The notion of -diversity attempts to solve this problem. We have shown that -diversity has a number of limitations and especially presented two attacks on -diversity. Motivated by these limitations, we have proposed a novel privacy notion called "closeness". We propose two instantiations: a base model called t-closeness and a more flexible privacy model called (n, t)-closeness. We explain the rationale of the (n, t)-closeness model and show that it achieves a better balance between privacy and utility. To incorporate semantic distance, we choose to use the Earth Mover Distance measure. We also point out the limitations of EMD, present the desiderata for designing the distance measure, and propose a new distance measure that meets all the requirements. Finally, through experiments on real data, we show 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. Below, we discuss some interesting open research issues. 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 (n, t)-closeness if E has (n, 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 (n, t) and the level of privacy become more complicated. Other Anonymization Techniques (n, t)-closeness 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 tech-nique 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.

2. Literature Review/ Survey

Large amount of person-specific data has been collected in recent years both by governments and by private entities. Data and knowledge extracted by data mining techniques represent a key asset to the society by analyzing trends, patterns and formulating public policies. Laws and regulations require that some collected data must be made public for example, Census data.

2.1 What about Privacy

First thought: anonymized the data How?

1. Remove explicit identifier "personally identifying information" (PII) for example: Name, Social Security number, phone number, email, address etc.

2.Anything that identifies the person directly

Is this enough?

2.1.1 Key attributes:

Name, address, phone number - uniquely identifying! Always removed before release.

2.1.2Quasi-Identifiers:

(5-digit ZIP code, birth date, gender) uniquely identify

Can be used for linking anonymzed dataset with other datasets

2.1.3 Sensitive attributes:

Medical records, salaries, etc

2.2 K-Anonymity

The information for each person contained in the released table cannot be distinguished from at least k-1 individuals whose information also appears in the release.

Example: you try to identify a man in the released table, but the only information you have is his birth date and gender. There are k men in the table with the same birth date and gender.

Any quasi-identifier present in the released table must appear in at least k records.

Achieving k-Anonymity: Replace specific quasi-identifiers with less specific values until get k identical valuesPartition ordered-value domains into intervals

2.3 l-diversity

Sensitive attributes must be "diverse" within each quasi-identifier equivalence class.

Distinct l-Diversity-each equivalence class has at least l well-represented sensitive values. Sensitive attributes must be "diverse" within each quasi-identifier equivalence **class**.

Limitation of l-Diversity

1. It is neither necessary nor sufficient to prevent attribute disclosure.

2. l-diversity does not consider semantics of sensitive values.

2.4 t-Closeness:

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.

3. System Analysis/Design

Minimizing the data has been under research for quite some time. The main fundamental that is reached is the theory of closeness. Closeness is the concept where two elements are checked and if they are close they are reduced. Closeness always does not guarantee maximum distance in the result data set. The objective of the work is to develop a System using an unique method by combining two of the most widely used privacy preservation techniques so that we can prevent the sensitive information of the individuals from being disclosed and can generalize data using closeness of the attributes.

3.1 Proposed Design:

Privacy preservation technique is a way to increase the similarity of the data item and introduce redundancy in such a way that information about individual users can not be revealed. This technique also desires the actual information of the data to not change. This is achieved by anonymizing the data before release.

3.2 K-anonymity:

The k-anonymity privacy requirement for publishing micro data requires that each equivalence class (i.e., a set of records that are indistinguishable from each other with respect to certain "identifying" attributes) contains at least k record.

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

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.

Following are three steps of anonymization

3.2.1. Generalization: The entropy measure or the distance between row wise data is calculated, the data is reduced by appending special character like '*' till the data is reduced and entropy among the rows are low.Minimizing the data has been under research for quite some time. The main fundamental that is reached is the theory of closeness. Closeness is the concept where two elements are checked and if they are close they are reduced. Find those attributes where maximum repentance is observed. After finding most frequent value has to group them.

3.2.2. Grouped Data

Reduce rows by representing the data in aggregated format. Reduce rows by adding one extra column total and representing the data in aggregated format . This is the second anonymous process.

3.2.3. Last Generalization

This is third anonymous process which making the data anonymous .

3.3 l-Diversity:

Diversity requires that each equivalence class has at least well-represented values for each sensitive attribute. Closeness always does not guarantee maximum distance in the result data set. In this work we further extended the closeness with **Anti-closeness or l-diversity.** It is a concept by means of which after removing closeness from data, again data rows are re-grouped such that no similar data even after reduction appears together.

4. Methodology:

Following Steps are performed in first level :

Step 1: Take Complete Data

Step 2: Reduction of Each element using generalization based on closeness. So result having close data.

Step 3: Minimize columns through grouping of data

Step 4: Minimize rows through adding one extra column. So result having redundant data

Following Steps are performed in Second level:

Step 1: Take Complete Data and measure entropy

Step 2: If entropy is less than the threshold value minimizes each data element to increase entropy so getting same table with high entropy.

Step 3: Now minimize columns and add one extra column name count

Step 4: Reduce row by grouping them.

Step 5: Finally check the redundancy

4.1 Reduction Steps

Step1: Generate Bi- Gram of string 1 and string 2

Step 2: Find entropy

Step 3: Find average entropy.

Step 4: If the entropy is less than threshold value .3 than reduce data and again find the average entropy otherwise if entropy is greater than threshold value .3 than get finalize data form.

Step 5: Stop the process.

4.2 Entropy Calculate the amount of surprise in Information

1. The More Predictable, The lower the entropy

2. The less Predictable The higher the entropy

5. System Implementation

The following steps constitute whole process of implementing combined k-anonymity and l-diversity privacy preservation measurements are as follows:
First generalize each column which having less than threshold value. In this table first generalize first name, middle name, last Name and so on for data redundancy then Find those attributes where maximum repentance is observed So sensitive columns are removed and only repeated values are extracted. After finding most frequent values has to group them. After grouping of them reduce rows by adding one extra column. Now perform the last generalization on this table. Finally for more privacy combine diversity with anonymity.

6. Conclusion and future work:

There are many techniques proposed for preservation of privacy which includes anonymity based techniques, closeness based techniques, diversity based techniques. The literature reveals that each of these techniques have their advantage and disadvantage and no technique can be considered as completely full proof as par as privacy preservation is concerned. Therefore in this work we proposed a new technique by combining both diversity and anonymity based methods. First the entropy measure or the distance between row wise data is calculated, the data is reduced by appending special character like '*' till the data is reduced and entropy among the rows are low. Further closeness among the columns are calculated and columns are reduced by aggregation. This is followed by reducing number of rows also by first classifying the data into groups and replacing actual data with group aggregate. Therefore the end table has lesser generalized rows and columns with respect to the original one. Hence finding out any sensitive information is impossible from this data. The technique can be further improved if a new strategy can be devised to introduce extra rows in the data that increases the redundancy.

7. References:

[1] C. Aggarwal, "On k-Anonymity and the Curse of Dimensionality," Proc.of the Int'l Conf. on 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. of the ACM Symp. on 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 Engineering (ICDE), pp. 217- 228, 2005.
[5] F. Bacchus, A. Grove, J. Y. Halpern, and D. Koller, "From Statistics to Beliefs", Proc. of National Conference on Artificial Intelligence (AAAI), pp. 602-608, 1992.
[6] J.-W. Byun, Y. Sohn, E. Bertino, and N. Li, "Secure Anonymization for Incremental Datasets," Secure Data Management (SDM), pp. 4863, 2006.
[7] B.-C. Chen, K. Lefebvre, and R. Ramakrishnan, "Privacy Skyline: Privacy with Multidimensional Adversarial Knowledge," Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 770–781, 2007.
[8] G. T. Duncan and D. Lambert, "Disclosure-Limited Data Dissemination,"Journal of The American Statistical Association, 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 Engineering (ICDE), pp. 205216, 2005.
[10] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions," Michigan Math Journal, vol. 31, pp. 231-240, 1984.
[11] V. S. Iyengar, "Transforming Data to Satisfy Privacy Constraints," Proc of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD), pp. 279-288, 2002.
[12] D. Kifer and J. Gehrke, "Injecting Utility into Anonymized Datasets," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIG- MOD), 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 Engineering (ICDE), pp. 116-125, 2007.
[14] S. L. Kullback and R. A. Leibler, "On Information and Sufficiency,"Ann. Math. Stat., vol. 22, pp. 79-86, 1951.
[15] D. Lambert, "Measures of Disclosure Risk and Harm," Journal of Official Statistics, vol. 9, pp. 313-331, 1993.

[16] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-Domain k-Anonymity," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 49-60, 2005.
[17] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Mondrian Multidimen- sional k-Anonymity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 25, 2006.
[18] K. LeFevre, D. DeWitt, and R. Ramakrishnan, "Workload-Aware Anonymization," Proc. of the ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD), pp. 277-286, 2006.
[19] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and -diversity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 106115, 2007.
[20] T. Li and N. Li, "Injector: Mining Background Knowledge for Data Anonymization," In Proc. Int'l Conf. Data Engineering (ICDE), 2008.
[21] T. Li and N. Li, "Towards Optimal k-Anonymization," Data & Knowl- edge Engineering, vol. 65, 2008.
[22] T. Li, N. Li, and J. Zhang "Modeling and Integrating Backgroun Knowledge in Data Anonymization," To appear in Proc. Int'l Conf. Data Engineering (ICDE), 2009.
[23] A. Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam, " -Diversity: Privacy Beyond k-Anonymity," Proc. Int'l Conf. Data Engineering (ICDE), pp. 24, 2006.
[24] D. J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Y Halpern, "Worst-Case Background Knowledge for Privacy-Preserving Data Publishing," Proc. Int'l Conf. Data Engineering (ICDE), pp. 126135, 2007.
[25] A. Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," Proc. of the ACM Symp. on Principles of Database Systems (PODS), pp. 223-228, 2004.
[26] M. E. Nergiz, M. Atzori, C. Clifton, "Hiding the Presence of Individuals from Shared Databases," pp. 665-676, 2007.
[27] H. Park and K. Shim, "Approximate Algorithms for k-Anonymity," pp. 67-78, 2007.
[28] V. Rastogi, S. Hong, and D. Suciu, "The Boundary Between Privacy and Utility in Data Publishing," Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 531-542, 2007.
[29] 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.
[30] 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.
[31] L. Sweeney, "Achieving k-Anonymity Privacy Protection Using Generalization and Suppression," Int'l J. Uncertain. Fuzz., vol. 10, no. 6, pp. 571-588, 2002.
[32] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertain. Fuzz., vol. 10, no. 5, pp. 557-570, 2002.
[33] T. M. Truta and B. Vinay, "Privacy Protection: p-Sensitive k-Anonymity Property," Proc. Int'l Workshop on Privacy Data Management (ICDEWorkshops), 2006.
[34] A. Asuncion and D.J. Newman, "http://www.ics.uci.edu/~mlearn/ML-Repository.html", UCI Machine Learning Repository, 2007.
[35] M.P. Wand and M.C. Jones, "Kernel Smoothing (Monographs on Statistics and Applied Probability)," Chapman & Hall, 1995.
[36] K. Wang, B. C. M. Fung, and P. S. Yu, "Template-based Privacy Preservation in Classification Problems," Proc. Int'l Conf. Data Mining (ICDM), pp. 466-473, 2005.
[37] R. C.-W. Wong, A. W.-C. Fu, K. Wang, and J. Pei, "Minimality Attack in Privacy Preserving Data Publishing," Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp. 543-554, 2007.
[38] R. C.-W. Wong, J. Li, A. W.-C. Fu, and K. Wang, "(α, k)-Anonymity: an Enhanced k-Anonymity Model for Privacy Preserving Data Publishing," pp. 754-759, 2006.
[39] X. Xiao and Y. Tao, "Anatomy: simple and effective privacy preserva- tion," Proc. of the Int'l Conf. on Very Large Data Bases (VLDB), pp.139-150, 2006.
[40] X. Xiao and Y. Tao, "Personalized Privacy Preservation," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 229- 240, 2006.
[41] X. Xiao and Y. Tao, "m-Invariance: Towards Privacy Preserving Republication of Dynamic Datasets," Proc. of the ACM SIGMOD Int'l Conf. on Management of Data (SIGMOD), pp. 689-700, 2007.
[42] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A. W.-C. Fu, "Utility-based Anonymization using Local Recoding," Proc. of the ACM SIGKDD Int'lConf. on Knowledge Discovery and Data Mining (KDD), pp.785-790, 2006.