A Noval Privacy Preserving Method for Data Publication

Prof. R Amudha[1], Nisha S Udayan[2], Aishwaria E K[3], Alan K Thomas[4]

Assistant Professor, Department of CSE, JCT College of Engineering and Technology, Coimbatore[1]

U.G. Students, Department of CSE, JCT College of Engineering and Technology, Coimbatore[2,3,4]

Abstract: Privacy has received increasing concerns in publication of datasets that contain sensitive information. Providing useful information to users for data mining in the main aspect and goals. Generalization and randomized response methods were proposed in database community to tackle this problem. Both the methods has faced the same barriers. These Generalization and randomized response methods usually required to control the tradeoff between privacy and data quality, which may put the data publishers in a dilemma. In these paper, a novel privacy preserving method for data publication is proposed based on conditional probability distribution and machine learning techniques, which can in act different criteria for different transactions. A basic cross sampling algorithm and a complete cross sampling algorithm are designed respectively for the settings of single sensitive attribute and multiple sensitive attributes, and an improved complete algorithm is developed by using Gibbs sampling, in order to enhance data utility when data are not sufficient. Many other methods provide better and strong privacy and better data utility.

Keywords: Data publication, Privacy preservation, Data utility, Cross sampling, Gibbs sampling

I. INTRODUCTION

The main aspect of privacy preserving for data publication is to shield the privacy of individual data and retain the statistical patterns implied in the original datasets, it enables the multiple access to the published dataset. For example, hospitals have collected large volumes of medical records.

1.1 Preserving methods for Data publication

Additional knowledge can come from diverse sources, such as well known facts, public records, and information about specific individuals. For example, suppose a hospital releases an anonym zed table, in which only identities of patients have been removed. This provides some *Quasi-Identifier* (QID) attributes, such as Age, Sex, Occupation and Zip code. This process is known as a *linking attack*.

However, cryptographic primitives are not suitable for such data conversion, which are commonly used in privacy preserving for query processing as the published data should be accessible to the public. In this paper, we focus on the input perturbation solution, representative techniques of which are generalization and randomized response.

1.1.1Generalization

Generalization is a popular approach for publishing private datasets, the core idea of which is dividing the dataset into several groups by certain rules, and tuples in each group are indistinguishable from each other, in order to prevent the adversary from associating any individual to a particular transaction. *k*-anonymity is a typical privacy definition of generalization, which requires that the number of tuples in each group is no less than *k*. The groups here are called *equivalence classes*, As the improvement of *k*-anonymity, *l*-diversity and *t*-closeness restrict the distribution of *Sensitive Attribute* (SA) values in each equivalence class, so that the adversary cannot associate an individual to a particular sensitive value with high probability.

First, for the purpose of privacy protection, typical privacy definitions like *l*-diversity and *t*-closeness pose constraints on the distribution of SA values in each released data group.

Second, generalized datasets are released in a non-standard form that may require complicate analyses, disabling the use of many existing data mining tools. In particular, some variants of generalization may develop their own forms of released datasets. It is impossible to develop a new algorithm for every combination of an output form.

Third, the theoretical analysis for privacy guarantee is often subject to one-time publication. It is possible that a personal transaction is included in multiple datasets, which are released by different sources. Although a single released dataset does not reveal personal privacy, by combining the knowledge gained from multiple released datasets, the privacy could be completely exposed. Finally, a privacy controlling parameter is usually required to control the tradeoff between privacy and data quality. This serves flexibility to users, but at the same time, it may also put the users in a

DOI 10.17148/IJARCCE.2020.9305

dilemma, the data publisher tends to use very strong privacy guarantee, minimizing the legal risks, which results in very bad data utility.

1.1.2 *Randomized response*

Randomized response makes uneasy for each SA values in datasets in a certain way, and evaluates the query answers based on the perturbed datasets by likelihood-based analysis, so that personal privacy is concealed while the trend of the entire dataset is still recoverable, which can satisfy the primary target of privacy preserving for data publication. γ -amplification is a typical privacy definition for the data perturbation, which bounds the likelihood ratio between any two possible input values. First, the perturbation of SA values are random, which may result in large distribution difference between the published dataset and the original dataset, and may lead to poor data utility.

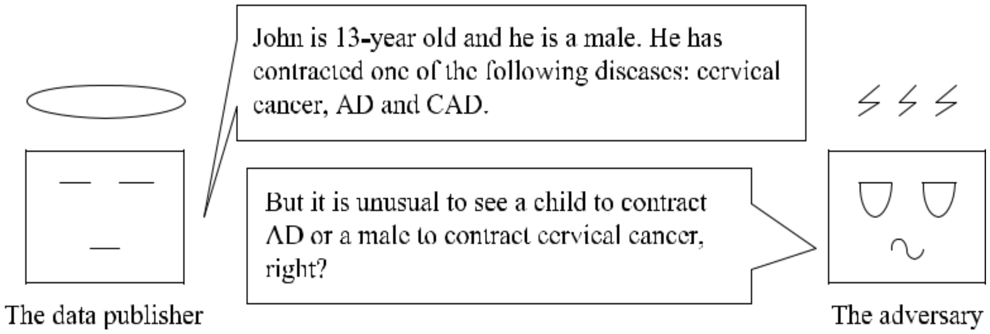
John is 13-year old and he is a male. He has contracted one of the following diseases: cervical cancer, AD and CAD.
But it is unusual to see a child to contract AD or a male to contract cervical cancer, right?
The data publisher
The adversary

Fig 1. Potential privacy breach of *l*-diversity and *t*-closeness.

II. THE CROSS SAMPLING ALGORITHM

Each transaction corresponds to a set of related data tuples in the dataset, and the benchmark distribution of each transaction can be regarded as a benchmark distribution combination of the data tuples associated with the transaction. Therefore, we select tuple as the unit to build the data distribution models, which can achieve different prior beliefs for different transactions, that is, non-homogeneous benchmark distribution.

Let *T* be an input dataset with *n* data tuples, which contain d_{QI} QID , . . . , A_{dQ} sensitiv
attributes *A* = {A_1 , A_{2I} e
attributes A^S={A^S_1,A^S_2, . . . ,A^S_{dS}}. Let A=$A^{Q\ I}$∪A^S and d=$d_{Q\ I}$+d_S. We denote by t[A] the value(s) of a tuple t for attribute(s∈ *A* or *A* ⊆*A*. The output dataset *T* is in the same form of the input dataset *T*, but with perturbed SA values. In this section, we will discuss our *cross sampling* data publication method as well as its improved version.

2.1 Basic Algorithm

For simplicity, we start with the simple case that there is only one sensitive attribute, i.e., d_S= 1. The basic algorithm of cross sampling method comprises two phases. At the first phase, we enumerate every tuple *t*, then learn a model $M^{(t)}$ for the conditional distribution $p(A^S|A^{QI})$ from the set *T* {*t*}. The predicted SA distribution $M^{(t)}(A^S|t[A^{QI}])$ forms *t*'s benchmark distribution, which is insensitive to *t*'s original SA value. At the second phase, we replace the SA value of every tuple *t* by a new sample drawn from *t*'s benchmark distribution, based on the pre-learnt models. The final dataset with the re-sampled SA values[1]is released.

Algorithm 1 The basic algorithm.

```
Input: $T$: a dataset
Output: $T$: a dataset
        for all $t \in T$ do
        $M^{(t)} \leftarrow$ learn $p(A^S|A^{Q\ I})$ from $T \setminus \{t\}$
        end for
$T$ = ∅
        for all $t \in T$ do
        $t\ [A^{Q\ I}] \leftarrow t[A^{Q\ I}]$

        $t\ [A^S] \leftarrow$ a sample drawn from $M^{(t)}(A^S|t\ [A^{Q\ I}])$
        $T = T \cup \{t\}$
        end for
return $T$
```

DOI 10.17148/IJARCCE.2020.9305

The algorithm comprises three phases. At the first phase, we randomly partition the dataset T into k groups of (almost) equal size. k ≥ 2 is a parameter that adjusts the efficiency of algorithm and the quality of released data. At the second phase, we learn models for computing the benchmark distribution. The computation is cross, meaning that to compute the benchmark distribution for a group, we learn models from tuples from other groups as training data

2.2 Improvement by Gibbs sampling

When there are multiple sensitive attributes, one issue of our method is that the order in which these attributes are taken into account matters, because when we are building a model for A^{S}_{i}, the attributes A^{S}_{i+1}, . . . , A^{S}_{dS} are not used. This issue may result in poor data utility when data are not sufficient. One strategy to resolve this issue is to use Gibbs sampling to post process the released dataset output by the original algorithm. the algorithm, for each sensitive attribute, we re-learn a model whose prediction makes use of all other attributes. Then, for each tuple we re-sample each of its SA values based on all other attribute values. The re-sample phase is repeated several time.

Attribute	Abbreviation	Type	Domain
Age	A	numeric	[16,94]
Gender	G	categorical	2
Education	E	numeric	[1,17]
Birth-place	B	categorical	51
Race	R	categorical	9
Work-class	W	categorical	7
Marital	M	categorical	6
Occupation	O	categorical	50
Salary	S	numeric	[0,49]

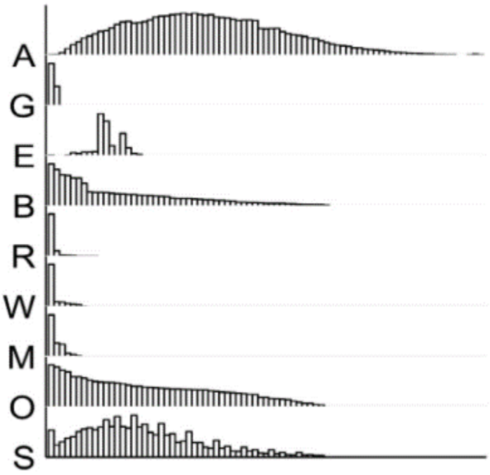

Fig. 4. Adult dataset.

III. EXISTING SYSTEM

Several anonymization techniques, such as generalization and bucketization, have been designed for privacy preserving microdata publishing. Recent work has shown that generalization loses considerable amount of information, especially for high-dimensional data. Bucketization, on the other hand, does not prevent membership disclosure and does not apply for data that do not have a clear separation between quasi-identifying attributes and sensitive attributes. In this paper, we present a novel technique called slicing, which partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. We show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement. Our workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute. Our experiments also demonstrate that slicing can be used to prevent membership disclosure.

Disadvantages:

1. Existing anonymization algorithms can be used for column generalization, e.g., Mondrian . The algorithms can be applied on the sub table containing only attributes in one column to ensure the anonymity requirement.
2. Existing data analysis (e.g., query answering) methods can be easily used on the sliced data.

IV. PROPOSED SYSTEM

We present a novel technique called slicing, which partitions the data both horizontally and vertically. We show that slicing preserves better data utility than generalization and can be used for membership disclosure protection. Another important advantage of slicing is that it can handle high-dimensional data. We show how slicing can be used for attribute disclosure protection and develop an efficient algorithm for computing the sliced data that obey the ℓ-diversity requirement. Our workload experiments confirm that slicing preserves better utility than generalization and is more effective than bucketization in workloads involving the sensitive attribute.

Advantages:

1. We introduce a novel data anonymization technique called slicing to improve the current state of the art.
2. We show that slicing can be effectively used for preventing attribute disclosure, based on the privacy requirement of ℓ-diversity.

Modules:

1. Dataset processing
2. Fixation of quasi-identifiers & sensitive attribute
3. Taxonomy tree structure of quasi-identifier
4. Execution of Anonymity Operations
5. Performance Evaluation

Module Description

1. Dataset processing

In this module we collect the data from the UCI bench mark resource in which we extract a particular dataset named, ADULT from the URI : https://archive.ics.uci.edu/ml/datasets/Adult , and finally this module loads the data attributes into the database. In this dataset, There are 32500+ records and fourteen attributes consisting of seven polynomials, one binomial and six continuous attributes. The nominal employment class attribute describes the type of employer such as self employed or federal and occupation describes the employment type such as farming or managerial. The education attribute contains the highest level of education attained such as high school graduate or doctorate. The relationship attribute has categories such as unmarried or husband and the marital status attribute has categories such as married or separated. The final nominal attributes are country of residence, gender and race. The continuous attributes are age, hours worked per week, education number (which is a numerical representation of the nominal education attribute), capital gain and loss and a survey weight attribute which is a demographic score assigned to an individual based on information such as area of residence and type of employment

1.1 Missing Value Estimation

An attribute table may contain multiple fields with null values and by default, these fields are populated with an empty space or with few special characters. These values are estimated and replaced with maximum repeated value of that attribute.

2. Fixation of quasi-identifiers & sensitive attribute

The attributes in the data set are categorized into personal identification attributes, quasi-identifiers and sensitive attributes. Personal identification attributes identify the persons directly. A set of attributes that can linked with external data to uniquely identify individuals in the population are called quasi-identifiers. Sensitive attributes hold sensitive information. Quasi-identifiers are pieces of information that are not of themselves unique identifiers, but are sufficiently well correlated with an entity that they can be combined with other quasi-identifiers to create a unique identifier. For ex. consider an employee dataset which contains the employee’s information with attributes Name, Zip code, Age, Sex & disease. In our dataset, Age, Native_Country, Race attribute is the personal identification, Sex is the sensitive attribute.

3. Taxonomy tree structure of quasi-identifier

Each quasi-identifier has a taxonomy tree structure of which generalization extent increases from leaf to root node. Empirically, every categorical quasi-identifier has a predetermined taxonomy tree, while, the taxonomy tree of numerical quasi-identifier will be dynamicall generated in the execution of anonymity algorithm. Below Figure is example for Taxonomy Tree, hierarchy tree and replacement presentation:

4. Execution of Anonymity Operations

There are two common methods for achieving privacy preservation.

1. Generalization: In this method, individual values of attributes are replaced by with a broader category. For example, the value '19' of the attribute 'Age' may be replaced by ' ≤ 20', the value '23' by '20 < Age ≤ 30' , etc.

2. Suppression: In this method, certain values of the attributes are replaced by an asterisk '*'. All or some values of a column may be replaced by '*'. In the anonymized table below, we have replaced all the values in the 'Name' attribute and all the values in the 'Religion' attribute with a '*'.

5. Performance Evaluation

In this module, Information Loss, Information Gain Values are calculated using the below formulas. A quick observation is that InfoGain(x) is not affected by applying Best -> child(Best) except that we need to compute InfoGain(c) for each value c in child(Best).

DOI 10.17148/IJARCCE.2020.9305

V. CONCLUSION

In this paper, we have presented a novel method *cross-sampling* for the privacy preserving data publication problem based on conditional probability distribution and machine learning techniques, which can overcome several shortcomings of existing generalization and randomized response methods. For the settings of single sensitive attribute and multiple sensitive attributes, our method designs the basic cross sampling algorithm and the complete cross sampling algorithm respectively, and by using Gibbs sampling, an improved complete algorithm is developed to advance the data utility when data are not sufficient. The merits of our method include:

(1) The released dataset is in the same form as the original dataset, the SA value of which is independent of the original value, while the data distribution of which is similar to the original distribution;
(2) Non-homogeneous benchmark distribution is obtained, which can accommodate different transactions;
(3) The difficulty of choosing privacy controlling parameter is circumvented. Theoretical analyses and extensive experiments show that our method can offer stronger privacy guarantee and retain better data utility than the existing methods.

REFERENCES

[1]. C. Liu, S. Chen and S. Zhou et al. / Information Sciences 501 (2019) 421–435
[2]. T. Li, C. Gao, L. Jiang, W. Pedrycz, J. Shen, Publicly verifiable privacy-preserving aggregation and its application in IoT, J. Netw. Comput. Appl. 126 (2019) 39–44, doi:10.1016/j.jnca.2018.09.018.
[3]. T. Li, Z. Huang, P. Li, Z. Liu, C. Jia, Outsourced privacy-preserving classification service over encrypted data, J. Netw. Comput. Appl. 106 (2018) 100–110, doi:10.1016/j.jnca.2017.12.021.
[4]. M. Song, W. Lin, B. Jiang, G. Deng, K-anonymity algorithm based on multi attributes generalization, J. Univ. Electron. Sci. Technol. China 46 (6) (2017)896–901.
[5]. C.C.Aggarwal, On k-anonymity and the curse of dimensionality, in: Proceedings of the VLDB, 2005, pp. 901–909.
[6]. R. Chen, Q. Xiao, Y. Zhang, J. Xu, Differentially private high-dimensional data publication via sampling-based inference, in: Proceedings of the SIGKDD, 2015, pp. 129–138, doi:10.1145/2783258.2783379.