Anonymization and Aggregation of Privacy Preserving of Personal data –using Slicing Technique

[1] M.Sheshikala, [2]P.Naveen Kumar

[1]Assistant professor,Dept. of CSE,SR Engineering College, Telangana, India, marthakala08@gmail.com
[2]PG Scholar,Dept. of CSE,SR Engineering College, ,Warangal,Telangana, India, polepallynaveen9@gmail.com

ABSTRACT: At present, most organisations are actively accumulating and storing data in large databases. Many of them have recognized in the expertise price of those data as an information source for making industry decisions. Privacy-preserving data publishing (PPDP) provides methods and tools for publishing priceless data even as keeping data privacy. In this paper, a short but systematic evaluate of a few Anonymization systems such as generalization and Bucketization, were designed for privacy maintaining micro data publishing. Contemporary work has proven that generalization loses huge amount of understanding, mainly for prime-dimensional information. On the other hand, Bucketization does not avert membership disclosure. Where as cutting preserves better data utility than generalization and in addition prevents membership disclosure. This paper makes a speciality of potent approach that can be used for supplying better data utility and might control excessive dimensional data.

KEYWORDS- Data Anonymization, PPDP, Privacy Preservation, Data publishing, Data Security.

I. INTRODUCTION

We exhibit how overlapping slicing can be used for attribute disclosure security and enhance an efficient algorithm for computing the sliced knowledge that obey the range requirement. Our workload test verify that overlapping cutting preserves better utility than generalization and is extra potent than bucketization in workloads involving the touchy attribute. Our experiments also demonstrate that overlapping cutting can be used to hinder membership disclosure. We consider the collaborative knowledge publishing crisis for anonymizing horizontally partitioned information at multiple information vendors. We consider a new variety of "insider assault" by way of colluding data providers who could use their own data documents (a subset of the total information) additionally to the external heritage abilities to infer the data records contributed by different information providers. The paper addresses this new risk and makes several contributions. First, we introduce the inspiration of m-privacy, which ensures that the anonymity knowledge satisfies a given privateness constraint towards any group of up to m colluding information vendors.second, we reward heuristic algorithms exploiting the equivalence group monotonicity of privacy constraints and adaptive ordering techniques for successfully checking m-privacy given a collection of files. Ultimately, we present a data provider-conscious anonymity algorithm with adaptive m-privacy checking procedures to be certain high utility and m-privacy of anonymity information with effectivity. Experiments on real-life datasets advocate that our procedure achieves better or related utility and efficiency than current and baseline algorithms while delivering m-privateness warranty.

Several micro data anonymity techniques have been proposed. The most popular ones are generalization for k-anonymity and bucketization for l-diversity. In both approaches, attributes are partitioned into three categories:

1) Some attributes are identifiers that can uniquely identify an individual, such as Name or Social Security Number.
2) Some attributes are Quasi Identifiers (QI), which the adversary may already know (possibly from other publicly available databases) and which, when taken

together, can potentially identify an individual, e.g., Birth date, Sex, and Zip code.
3) Some attributes are Sensitive Attributes (SAs), which are unknown to the adversary and are considered sensitive, such as Disease and Salary.

In both generalization and bucketization, one first removes identifiers from the data and then partitions tuples into buckets. The two techniques differ in the next step. Generalization transforms the QI-values in each bucket into "less specific but semantically consistent" values so that tuples in the same bucket cannot be distinguished by their QI values. In bucketization, one separates the SAs from the QIs by randomly permuting the SA values in each bucket.

Generalization for k-anonymity losses considerable amount of data, especially for high-dimensional data. Bucketization does not prevent membership disclosure. Because bucketization publishes the QI values in their original forms, an adversary can find out whether an individual has a record in the published data or not. Bucketization requires a clear separation between QIs and SAs. However, in many data sets, it is unclear which attributes are QIs and which are SAs. We assume the data providers are semi-honest, commonly used in distributed computation setting. They can attempt to conclude additional data about data coming from other providers by analyzing the data received during the anonymity. A data recipient, e.g. P0, could be an attacker and attempts to infer additional information about the records using the published data (T∗) and some background knowledge (BK) such as publicly available external data.

In the most basic form of privacy-preserving data publishing (PPDP), the data holder has a table of the form: D (Explicit Identifier, Quasi Identifier, Sensitive Attributes, non-Sensitive Attributes), where Explicit Identifier is a set of attributes, such as name and social security number (SSN), containing information that explicitly identifies record owners, Quasi Identifier is a set of attributes that could potentially identify record owners, Sensitive Attributes consist of sensitive person-specific information such as disease, salary, and disability status and Non-Sensitive Attributes contains all attributes that do not fall into the previous three categories.

Most works assume that the four sets of attributes are disjoint. Most works assume that each record in the table represents a distinct record owner.

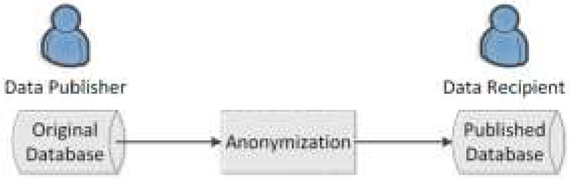

Figure 1. A Simple Model of PPDP [13].

II. RELATED WORKS

Two main Privacy preserving paradigms have been established: k-anonymity [7], which prevents identification of individual records in the data, and l-diversity [1], which prevents the association of an individual record with a sensitive attribute value. K-anonymity

The database is said to be K-anonymous where attributes are suppressed or generalized until each row is identical with at least (k-1) other rows. K-Anonymity thus prevents definite database linkages. K-Anonymity guarantees that the data released is accurate. K-anonymity proposal focuses on two techniques in particular: generalization and suppression. [2] To protect respondents' identity when releasing micro data, data holders often remove or encrypt explicit identifiers, such as names and social security numbers. De-identifying data, however, provide no guarantee of anonymity. Released information often contains other data, such as birth date, sex, and ZIP code that can be linked to publicly available information to re-identify respondents and to infer information that was not intended for release. One of the emerging concepts in micro data protection is k-anonymity, which has been recently proposed as a property that captures the protection of a microdata table with respect to possible re-identification of the respondents to which the data refer. K-anonymity demands that every tuple in the microdata table released be indistinguishably related. One of the interesting aspects of k-anonymity is its association with protection techniques that preserve the truthfulness of the data. The first approach toward privacy protection in data mining was to perturb the input (the data) before it is mined. The drawback of the perturbation approach is that it lacks a formal framework for proving how much privacy is guaranteed. At the same time, a second branch of privacy preserving data mining was developed, using cryptographic techniques. Thus, it falls short of providing a complete answer to the problem of privacy preserving data mining. One definition of privacy which has come a long way in the public arena and is accepted today by both legislators and corporations is that of k-anonymity [3]. The guarantee given by k-anonymity is that no information can be linked to groups of less than k individuals. Generalization for k-anonymity losses considerable amount of information, especially for high-dimensional data.

[4] Limitations of k-anonymity are: (1) it does not hide whether a given individual is in the database, (2) it reveals individuals' sensitive attributes , (3) it does not protect against attacks based on background knowledge , (4) mere knowledge of the k-anonymization algorithm can violate privacy, (5) it cannot be applied to high-dimensional data without complete loss of utility , and (6) special methods are required if a dataset is anonymized and published more than once.

l- Diversity

The next concept is “l-diversity”. Say you have a group of k different records that all share a particular quasi-identifier. That’s good, in that an attacker cannot identify the individual based on the quasi-identifier. But what if the value they’re interested in, (e.g. the individual’s medical diagnosis) is the same for every value in the group. The distribution of target values with in a group is referred to as “l-diversity”. [8] Currently, there exist two broad categories of l-diversity techniques: generalization and permutation-based. An existing generalization method would partition the data into disjoint groups of transactions, such that each group contains sufficient records with l-distinct, well represented sensitive items.

III. THE PROPOSED APPROACHES

Most commonly in privacy preservation there's a loss of safety. The privacy security is impossible due to the presence of the adversary’s historical past data in real life applications. Data in its original form includes sensitive data about contributors. These information when released violate the privacy. The current follow in data publishing depends most likely on insurance policies and recommendations as to what varieties of data may also be published and on agreements on using released data. The method alone may just result in excessive data distortion or inadequate protection. Privacy-preserving data publishing (PPDP) provides ways and tools for publishing valuable data even as preserving data privacy. Many algorithms like bucketization, generalization have tried to maintain privacy nonetheless they show off attribute disclosure. So that we can overcome this challenge by an algorithm called slicing.

Functional procedure:-

Step 1: Extract the data set from the database.

Step 2: Anonymity process divides the records into two.

Step 3: Interchange the sensitive values.

Step 4: Multi set values generated and displayed.

Step 5: Attributes are combined and secure data Displayed.

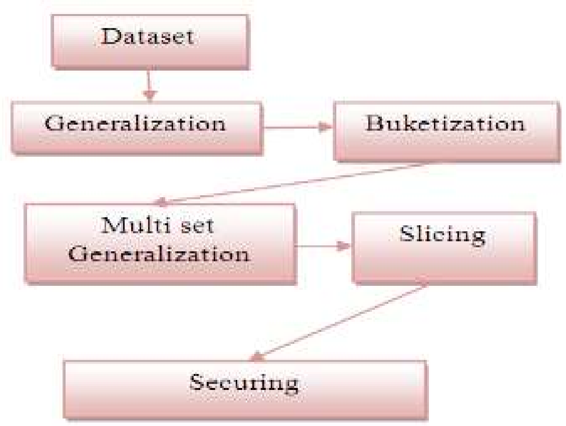

Figure.2 Slicing Architecture

Many algorithms like bucketization, generalization have tried to maintain privacy nevertheless they exhibit attribute disclosure. So that you can overcome this crisis by an algorithm called reducing is used. This algorithm contains three phases: attribute partitioning, column generalization,and tuple partitioning

Attribute Partitioning

This algorithm partitions attributes so that extremely correlated attributes are in the same column. That is just right for each utility and privateness. In terms of data utility, grouping totally correlated attributes preserves the correlations amongst these attributes. In terms of privateness, the association of uncorrelated attributes offers bigger identification risks than the organization of totally correlated attributes given that the associations of uncorrelated attribute values is much less established and accordingly more identifiable.

Column Generalization

Even though column generalization is not a required segment, it can be valuable in several features. First, column generalization may be required for identity/membership disclosure defense. If a column value is distinctive in a column (i.e., the column price seems simplest as soon as in the column), a tuple with this exact column value can handiest have one matching bucket.

This isn't just right for privacy security, as in the case of generalization/bucketization where every tuple can belong to just one equivalence-category/bucket. The foremost main issue is that this special column price can be making a choice on. In this case, it could be useful to apply column generalization to make certain that every column price appears with at the least some frequency. Second, when column generalization is utilized, to acquire the identical stage of privacy against attribute disclosure, bucket sizes can be smaller. Even as column generalization may just outcomes in know-how loss, smaller bucket-sizes allow better data utility. Hence, there is a alternate-off between column generalization and tuple partitioning.

Tuple Partitioning

The algorithm maintains two data structures: 1) a queue of buckets Q and 2) a set of sliced buckets SB. Initially, Q contains only one bucket which includes all tuples and SB is empty. For each iteration, the algorithm removes a bucket from Q and splits the bucket into two buckets. If the sliced table after the split satisfies l-diversity, then the algorithm puts the two buckets at the end of the queue Q. Otherwise, we cannot split the bucket anymore and the algorithm puts the bucket into SB. When Q becomes empty, we have computed the sliced table. The set of sliced buckets is SB.

IV. CONCLUSION

Reducing overcomes the barriers of generalization and bucketization and preserves better utility even as protecting in opposition to privateness threats. Cutting prevents attribute disclosure and membership disclosure. Chopping preserves higher data utility than generalization and is extra robust than bucketization in workloads involving the sensitive attribute. We remember cutting the place each and every attribute is in exactly one column. An extension is the thought of overlapping reducing, which duplicates an attribute in more than one column.

ACKNOWLEDGEMENT

Thanks for UGC for funding this project – UGC Reference No. MRP-6237/15(SERO/UGC).

REFERENCES

[1] Tiancheng Li, Ninghui Li, Senior Member, IEEE, Jia Zhang, Member, IEEE, and Ian Molloy "Slicing: A New Approach for Privacy Preserving Data Publishing" Proc. IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 3, March 2012.

[2] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[3] G.Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[4] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for PrivacyPreserving Data Publishing," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[5] P. Samarati, "Protecting Respondent's Privacy in Micro data Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010- 1027,Nov/Dec. 2001.

[6] A. Inan, M. Kantarcioglu, and E. Bertino, "Using Anonymized Data for Classification," Proc. IEEE 25th Int'l Conf. Data Eng. (ICDE), pp. 429-440, 2009–473.

[7] D. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J. Halpern. "Worst-case background knowledge for privacy-preserving data publishing". In ICDE, 2007.

[8] G.Ghinita, Y. Tao, and P. Kalnis, "On the Anonymization of Sparse High-Dimensional Data," Proc. IEEE 24th Int'l Conf. Data Eng. (ICDE), pp. 715-724, 2008.

[9] R. J. Bayardo and R. Agrawal, "Data Privacy through Optimal k- Anonymization," in Proc. of ICDE, 2005, pp. 217–228.

[10] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Incognito: Efficient Full-domain k-Anony Anonymity," in Proc. of ACM SIGMOD, 2005, pp. 49– 60.

[11] K. LeFevre, D. J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k-Anonymity," in Proc. of ICDE, 2006.

[12] Gabriel Ghinita, Member IEEE, Panos Kalnis, Yufei Tao," Anonymous Publication of Sensitive Transactional Data" in Proc. Of IEEE Transactions on Knowledge and Data Engineering February 2011 (vol. 23 no. 2) pp. 161-174.

[13] D.J. Martin, D. Kifer, A. Machanavajjhala, J. Gehrke, and J.Y. Halpern, "Worst-Case Background Knowledge for Privacy- Preserving Data Publishing," Proc. IEEE 23rd Int'l Conf. Data Eng. (ICDE), pp. 126-135, 2007.

[14] X. Xiao and Y. Tao, "Anatomy: Simple and Effective Privacy Preservation," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 139-150, 2006.

[15] Y. He and J. Naughton, "Anonymization of Set-Valued Data via Top-Down, Local Generalization," Proc. Int'l Conf. Very Large Data Bases (VLDB), pp. 934-945, 2009