ISSN 2348 – 9928
IJAICT Volume 2, Issue 8, December 2015 Doi:01.0401/ijaict.2015.08.01 Published on 05 (01) 2016

PRIVACY PRESERVING TECHNIQUE WITH PUBLIC DATABASE FOR DATA MINING

Ms. N. Gomathi *Ms. T. Cowsalya* *Mr. R. Arunkumar*

Department of Computer Science and Engineering,
SVS College of Engineering,
Coimbatore, Tamilnadu, India

Abstract– ***Publishing data for analysis from a table containing personal records, while maintaining individual privacy, is a problem of increasing important today. In most practical anonymization scenarios, there exists public knowledge that can be used by an attacker to breach privacy. The proposed work "Privacy Preserving Technique with Public Data Base For Data Mining" presents a novel K-Anonymity solution to capture ℓ -diversity with an availability of external data base information. It takes both micro data to be published and an external data Base for anonymity process. It protects published data from an external attack like Linking Attack and also reduces information loss by capturing diversity.***

Keywords– ***Privacy, Anonymity, Generalization, ℓ-diversity.***

I. INTRODUCTION

Information is to-day probably the most important and demanded resource. An Inter-networked society that relies on the dissemination and sharing of information in the private as well as in the public and Governmental sectors. The Problem is that once information is released, it may be impossible to prevent misuse. In order to protect the Anonymity of the entities to which information refers, data holders often remove or encrypt explicit identifiers such as Names, Addresses and Phone numbers.

Privacy is generally known as "Freedom from unauthorized intrusion." It aims at limiting the risk of linking published data to a particular person. Privacy preserving data mining techniques usually involve randomizing the data subjects' personal data. An enterprise wishing to carry out data mining obtains randomized user information which may be used for data mining but randomization ensures that the personal information is not disclosed.

There are several methods like Obscuring data, Anonymization to provide privacy on data mining. Obscuring data is an approach to provide privacy by making of private data available, but with enough noise added that exact value (or approximations sufficient to allow to misuse) cannot be determined. It is a technique which modifying the data values so real values are not disclosed. One approach, typically used in census data, is to aggregate items

Further organization of the paper is as follows. The Notations & Principles are mentioned in Section II, related works are discussed in Section III, existing system is analyzed in Section IV, Section V deals with the proposed algorithm, Results are shown in Section VI and conclusion is presented in Section VII.

II. NOTATIONS & PRINCIPLES

Three types of micro data attributes are relevant to privacy preservation: 1) identifiers (IDs); 2) quasi-identifiers (QIs); and 3) sensitive attributes (SAs). IDs (e.g., passport number, social security number, and name) can be used individually to identify a tuple.

Definition 1. (Quasi-identifier) *A set of non-sensitive attributes* {Q1, . . . , Qw} *of a table is called a* quasi-identifier *if these attributes can be linked with external data to uniquely identify at least one individual in the general population.* It is represented as QI. One example of a quasi-identifier is a primary key like social security number. Another example is the set like {Gender, Age, and Native Country}. QIs (e.g., zip code, gender, and birth date) are attributes that can be combined to act as IDs in the presence of external knowledge.

Definition 2. (**Sensitive Attribute**) A Sensitive Attribute is an attribute whose value for any particular individual must be kept secret from people who have no direct access to the original data. Let S denote the set of all sensitive attributes. An example of a sensitive attribute is salary class, occupation. SAs (e.g., disease, salary, and criminal offence) are fields that should be hidden so that they cannot be associated to specific persons.

Table 1. Describes the different symbols used in the following sections.

Symbol	Description
QI	Quasi Identifier
SA	Sensitive Attribute
ID	Identifier
MT	Micro Table
AT	Anonymized Table
JT^+	Join Table
PD^P	Public Data Base
q*	Quasi Identifier Block

Anonymization is removal of identifying information from data in order to protect privacy of the data subjects while allowing the data so modified to be used for secondary purposes like data mining. For example, anonymization may be used by a Web user to prevent collection of his/her information by hiding personal information like cookies and IP addresses

Anonymization may be based on a one-way hashing technique. Hashing allows different people to share information because when a hash function is applied to any information the resulting hashed value will always be the same if the same has function n is applied to the same information. It therefore allows sharing and matching information without disclosing personal information about a person.

Definition 3. (K-Anonymity) A k-Anonymized data set has the property that each record is indistinguishable from at least k-1 other records within the data set. A table T is said to be k-anonymous if each record is indistinguishable from at least k - 1 other tuples in T with respect to the QI set. To achieve k-Anonymized result there are some techniques like ,i.)Suppression- is the process of deleting cell values or entire tuples. ii.)Generalization- Generalization of the data, where low-level or "primitive" (Raw) data are replaced by higher-level concepts through the use of concept hierarchies.

Given MT, the anonymization process produces an anonymized table (or view) AT that contains all tuples and QI attributes, and preserves as much information as possible compared to the original table MT. A table AT is an anonymized instance of MT if: 1) AT has the same QI attributes as MT and 2) there is a one to-one and onto mapping (bijection) of MT to AT tuples. The most common method, i.e., mapping, for achieving anonymization is generalization. For numerical QIs, a generalization of a value is a range.

For categorical QIs, it is a higher level value in a given hierarchy (e.g., a city name is replaced with a state or country). Since categorical values can be trivially mapped to an integer domain. A generalized AT tuple is represented as an axis-parallel (hyper) rectangle, called G-box, in the QI space defined by the extent of its QI ranges. The goal of k-anonymity is to hide the identity of individuals by constructing G-boxes that contain at least k-MT tuples. An anonymized table AT of MT is k-anonymous if the mapping of each MT record is indistinguishable among the mappings of at least k- 1 other MT tuples.

An attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes. The technique of ℓ-diversity, which not only maintains the minimum group size of k, but also focuses on maintaining the diversity of the sensitive attributes. Therefore, the ℓ-diversity model for privacy is defined as follows:

Definition 4. (ℓ-diversity) Let a q*-block be a set of tuples such that its non-sensitive values generalize to q*. A q* -block is ℓ-diverse if it contains ℓ-"well represented" values for the sensitive attribute S. A table is ℓ-diverse, if every q*-block in it is ℓ-diverse.

III. RELATED WORKS

Several concepts have been proposed to achieve privacy like Obscuring data, Randomization, Anonymization and ℓ-diversity. Recoding is a technique which releases the generalized value. i.e., releasing the month and year of birth instead of the complete birth date [9].

Global recoding maps a given value in a single domain to another one globally. Global recoding [6] maps the domains of the quasi-identifier attributes to generalized or changed values. Local recoding maps individual tuple to generalized tuples. Global recoding may produce more information loss, in contrast local recoding may achieve less information loss in anonymization.

Table 2. and Table 3. Illustrates the global and local recoding respectively. Age and zip code fields are generalized in both tables.

Row-id	Age	Zip code
R1	[24-32]	[53712-53713]
R2	[25-30]	53711
R3	[25-30]	53711
R4	[25-30]	53711
R5	[24-32]	[53712-53713]
R6	[24-32]	[53712-53713]

Table 3. Anonymization by local recoding

Row-id	Age	Zip code
R1	[24-30]	[53711-53712]
R2	[24-30]	[53711-53712]
R3	[24-30]	[53711-53712]
R4	[30-32]	[53711-53713]
R5	[30-32]	[53711-53713]
R6	[30-32]	[53711-53713]

A *single-dimensional recoding* [7] is defined by a function Φ_i: DXi ->D' for each attribute X_i of the quasi-identifier. An anonymization V is obtained by applying each Φi to the values of Xi in each tuple of T.

Alternatively, a *multidimensional recoding* is defined by a *single* function Φ*:* D_{X1} x DX2 x DX3…D_{Xn} -> D', which is used to recode the domain of value *vectors* associated with the set of quasi-identifier attributes. Under this Model, *V* is obtained by applying *Á* to the vector of quasi identifier Values in each tuple of *T*.

K-anonymity has been proposed with primary goal as to protect the privacy of individuals to whom the data pertains. Single dimensional [6] model partitions the data with respect to single attribute. It performs mapping for each attribute individually. Multidimensional model partitions the domain into a set of non overlapping multi dimensional regions and maps the Cartesian product of multiple attributes.

Anonymization is defined as NP-hard problem [2, 5]. Optimal k-anonymity [2] achieves an approximation ratio independent of the size of the database, when k is constant. It is a O (k log k) approximation where the constant in big-O is no more than 4.

K-anonymity does not guarantee privacy against attackers using background knowledge and attacker can discover the values of sensitive attributes when there is little diversity in those sensitive attributes [1].

ℓ-diversity provides privacy even when the data publisher does not know what kind of knowledge is possessed by the adversary. The values of sensitive attributes are well represented in each group when ℓ-diversity applied. In [8] systematic clustering for ℓ-diversity is proposed. It increases the efficiency of the process by grouping the similar data together with ℓ-diverse sensitive values and then anonymizes each group individually.

IV. EXISTING SYSTEM

The concept of t-closeness requires that the distribution of values in each QI group is analogous to the distribution of the entire data set. Knowledge of the inner Mechanisms of the anonymization result in privacy breaches. To protect privacy, external knowledge database or public database can be considered [4] along with micro data to be published during anonymization process. There is a chance to have some sensitive attribute with diversity. The following system has the following two steps. First step is construction of G-Box and implementation of K-Join Anonymity. Construction of G-box can be done using the following two methods. These methods are implementation of Mondrian and top down.

Mondrian constructs QI groups than contain from k up to 2k - 1 tuples (when all QI values present in MT are distinct), following a strategy similar to the KD-tree space partitioning. In particular, starting with all MT records, it splits the d-dimensional space (defined by the d QI attributes) into two partitions of equal cardinality. The first split is performed along the first dimension (i.e., quasi-identifier QI1, according to the median QI1 value in MT. Each of the resulting groups is further divided into two halves according to the second dimension. Partitioning proceeds recursively, choosing the splitting dimension in a round-robin fashion among QI attributes. Mondrian terminates when each group contains fewer than 2k records. The resulting space partition is the anonymous version of MT to be published.

TopDown is a recursive clustering algorithm. Specifically, it starts with the entire MT and progressively builds tighter clusters with fewer points. Initially, the algorithm finds the 2 tuples that if included in the same anonymized group, they would result in the largest perimeter. TopDown considers the remaining records in random order, and groups them together with either any one of chosen tuple.

The goal of k-join-anonymity is to provide the same privacy guarantees with k-anonymity incurring. To achieve this, it shrinks the G-boxes using public knowledge about universe (U) tuples. PD^p should contain at least the QI attributes of MT. Extra attributes in PD^p are discarded. A PD^p that does not include all QIs is useless for KJA.

The anonymization process uses information from MT and PD^p. Let JT^+ denote the full outer join table of PD^p and MT^+, where MT^+ corresponds to the microdata augmented with the

ID attribute. JT refers to the join table without the ID and contains tuples that appear: 1) in both PD^p and MT; 2) in PD^p but not in MT; and 3) in MT but not in PD^p. The main difference of KJA from previous k-anonymity formulations is that an MT record may be anonymized/grouped with any JT tuple, as opposed to being restricted to MT records. Note that not all PD^p tuples may be needed during the anonymization process.

On the other hand, all MT records must be anonymized. It refers to a subset of JT, which contains all MT tuples, as proper.

V. PROPOSED SYSTEM

Anonymization is done with the consideration of availability of public database. External database helps to avoid back ground attacks, linking attacks. Anonymization via generalization or suppression usually causes information loss and now a natural question arises, how much information is lost due to anonymization. The problem of *k*-anonymization can also be considered as a clustering problem, where each equivalent class is a cluster and the size of each cluster is at least *k*.

The Mondrian and Top down model considers availability of public data base. Both can be easily adopted to capture diversity. The idea of information loss is used to measure the amount of information loss due to *k*-anonymization. There is a need to capture l-diversity to reduce information loss.

The requirement for ℓ-diversity model to satisfy at least ℓ distinct sensitive attribute values in each equivalent class. So the optimal solution of ℓ-diversity–KJA algorithm is to construct G-box such that it satisfies both *k*-anonymity and ℓ-diversity requirement with effective utility of public database and the total information loss will be as minimum as possible.

The Proposed algorithm ℓ-diversity KJA algorithm utilizes both Micro data (MT) and External Public Data Table (PT) for anonymization Process. It also applies ℓ-Diversity principle on both KJA- Mondrian and KJA-Top down model

The proposed system includes the G-box construction using Mondrian and top down, ℓ-Diversity KJA algorithm.

A table is Entropy ℓ-Diversity diverse if for every q* block

$$\sum_{s \varepsilon S} (p(q^*,s)\log(p(q^*,s')))\geq \log(l)$$

Where $p(q^*,s) = n(q^*,s) / \sum_{s' \varepsilon S} n(q^*,s'))$.

The above equation is the fraction of tuples in the q* block with sensitive attribute value equal to s. As a consequence of this condition, every q* block has at least ℓ distinct values for the sensitive attribute, xlog(x) is a concave function, it can be shown that if a q* block is split into two sub blocks q_a^* and q_b^* then entropy $(q^*) \geq \min(\text{entropy}(q_a^*), \text{entropy}(q_b^*))$. This implies that in order for entropy ℓ-Diversity to be possible, the entropy of the entire table must be at least log (ℓ).

All experiments are run under Windows XP. our experiments run on the Adult Database. The Adult Database contains 45,222 tuples from US Census data. Tuples with missing values are removed and adopted the same domain generalizations. Table 4. Provides a brief description of the data including the attributes used, the number of distinct values for each attribute. Salary Class and Occupation attributes are considered as Sensitive Attribute.

The performance Result of KJA and ℓ-diversity local recoding will be compared. Quality metrics for Information loss to be considered are Entropy, information loss ratio, information gain, accuracy.

Table 4. Adult Data Set

Attribute	Domain Size
Age	74
Gender	2
Race	5
Marital Status	7
Education	16
Native Country	41
Work Class	7
Salary Class	2
Occupation	14

VI. RESULTS

Accuracy achieved through ℓ- diversity is high. It is shown in the following figures.

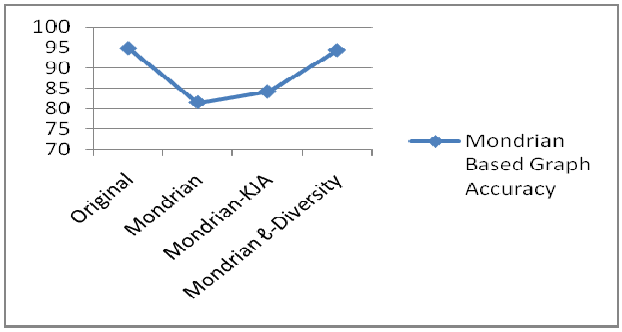

Fig 1: Mondrian Based graph

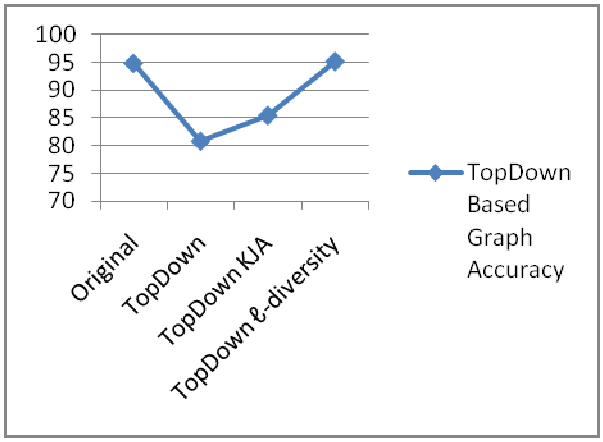

Fig 2: Top Down Based graph

VII. CONCLUSION

In this paper, Algorithms for ℓ-diversity-KJA model as an enhanced of simple *K*JA model. The proposed technique uses the idea of ℓ-diversity with KJA model and is implemented in two steps, namely k-join anonymity step for *k*-join anonymization and ℓ-diverse step to capture the diversity. The basic concepts of the proposed algorithms are discussed .The effect of the proposed approach can be useful for protecting private information of individuals as ℓ-diversity model is one of the most popular approaches for privacy preserving techniques.

References

[1] A.Machanavajjhala, J. Gehrke, D. Kifer, and M. Venkitasubramaniam," ℓ-Diversity: Privacy beyond k-Anonymity," Proc. IEEE Int'l Conf. Data Eng. (ICDE), p. 24, 2006.
[2] A.Meyerson and R. Williams, "On the Complexity of Optimal k-Anonymity," Proc. ACM SIGACT-SIGMOD-SIGART Symp. Principles of Database Systems (PODS), pp. 223-228, 2004.
[3] C.Bettini, X.S. Wang, and S. Jajodia, "The Role of Quasi-Identifiers in k-Anonymity Revisited," Technical Report abs/cs/0611035,Computing Research Repository (CoRR), 2006.
[4] Dimitris Sacharidis, kyriakos Mouratidis, and Dimitris Papadias, " K-Anonymity in the Presence of External DataBases" Proc. IEEE Int'l Conf. Data Eng. (ICDE), vol 22,No. 3, 2010.
[5] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy, D. Thomas, and A.Zhu, "Anonymizing Tables for Privacy Protection" Proc. Int'l Conf. Database Theory (ICDT), pp. 246-258, 2005.
[6] J. Xu, W. Wang, J. Pei, X. Wang, B. Shi, and A.W.-C. Fu, "Utility-Based Anonymization Using Local Recoding," Proc. ACM SIGKDD, pp. 785-790, 2006.
[7] K. LeFevre, D.J. DeWitt, and R. Ramakrishnan, "Mondrian Multidimensional k- Anonymity," Proc. IEEE Int'l Conf. Data Eng. (ICDE), p. 25, 2006.
[8] Md Enamul Kabir, Hua Wang, Elisa Bertino & yunxiang Chi, "Systematic Clustering Method for *l*- diversity Model.
[9] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6, pp. 1010-1027, Nov./Dec. 2001.
[10] R.J. Bayardo, Jr., and R. Agrawal, "Data Privacy through Optimal k-Anonymization," Proc. IEEE Int'l Conf. Data Eng. (ICDE), pp. 217-228, 2005.