

# Updating Anonymous Database While Maintaining Privacy and Confidentiality

Vidya Ingle<sup>1</sup>, Prof. Manjusha Deshmukh<sup>2</sup>

P.G. Student, Department of Information Technology Engineering , Pillai's Institute of Information Technology, Engineering, Media Studies and research ,New Panvel, Maharashtra, India<sup>1</sup>

Assistant Professor, Department of Computer Engineering, Pillai's institute of Information Technology, Engineering, Media Studies and Research, New Panvel Maharashtra, India<sup>2</sup>

**ABSTRACT:** Now a day's large amounts of data is collected by various agencies, such as health institution, government organization and market research etc. This data is able to be used for applications such as statistics, business intelligence, economics and various analyses. However sharing out such data is of enormous significance for statistical or investigations, it is especially essential to defend the confidentiality of individuals included in the data. Many organization and foundation are keen to release the data they collected to other parties for research and the formulation. Data frequently include individually specific information and therefore publishing such data may result in privacy violation. A number of anonymization techniques, resembling perturbation, swapping, suppression, generalization and bucketization are designed for privacy preserving data publishing. K-anonymity is a privacy preserving principal implemented with the help of two techniques: suppression and generalization. Generalization loses significant amount of information especially for large volume of data. Slicing is technique which provides improved data utility than generalization and is more efficient in workload pertaining to the sensitive attribute.

Suppose there is an anonymous database (e.g. including medical data of patients). Now target contains how to still preserve the privacy while updates are being made into the current anonymous database by preserving the privacy of the user and confidentiality of the database. It is required to ensure that the database is still anonymous past updating. K-anonymity and slicing approach is used to anonymize database and then updates are made to the anonymous database while preserving its privacy.

## I. INTRODUCTION

In today's world data is collected for various reasons. The collected data includes personal details or some other top secret information. Database stores records and facilitate access, update and recovery of data by the user. Providing security to these databases is important concern currently. Various methods are available to maintain the database secure. Merely protecting data and not making it available for research, analysis, data mining is big limitation on data use. But if the data is provided as it is, it can be misused by attackers. Nowadays information is stored for different things like education, stock, shopping, social sites etc. There is chance that this data can be used by insurance companies, drug manufacturing companies and various other marketing agencies without any notice to you. Illicit parties should be prohibited to access data in order to maintain confidentiality. It is crucial to uphold database integrity when data is transferred from source to target. Assume that data owner wants to distribute the data with researchers or analysts, in this scenario data owner should precisely assure that the data persist practically useful but people belonging to this data cannot be recognized.

The privacy-preserving is the method of protecting available information from the attackers for misuse. A number of privacy researches for protecting published data are available. Data anonymization is a process of removing personal identifiers to protect private information. Data anonymization facilitates transmit information between two organizations by converting text data into non human readable form [2]. It ensures confidentiality of original data by modification.

Any database table contains following types of attributes:

- I. Identifiers (I): These attributes are distinctively recognized. eg. Social Security Number which is unique and directly identifies an individual;
- II. Quasi-identifiers (QI): These are some attributes which the adversary may already know and can distinguish or identify a person. e.g., Date of Birth, Sex, and Zip code or PIN code.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

III. Sensitive attributes (SAs): These attributes are unknown to the adversary and are considered sensitive like Disease and Salary [2].

There are numerous techniques present for privacy preservation of data released socially named as K-anonymity, perturbation, swapping etc. Basically these techniques are used to convert data in such a format that illicit individual cannot trace the original data at the same time authoritative user also can discover the limited data for his use.

Data anonymization is widely used for converting data into non-human understandable form and then exchanged between two organizations. Data anonymization modifies the original data to protect its privacy.

**K-anonymity** - When some attributes of database are suppressed or generalized such that each row is identical with at least k-1 other rows in that database is said to be K-anonymous database. K-Anonymity is useful to prevent exact database linkages when multiple version of same database is released or an individual has record in multiple public databases. There are two techniques generalization and suppression for achieving K-anonymity. There are following type of information disclosure:

I. Identity disclosure II. Attribute disclosure III. Membership disclosure

K-anonymity protects against identity disclosure but cannot provide a safety against attribute disclosure. Generalization and suppression are the most common methods used for de identification of the data in k-anonymity based algorithms [2]. As the name indicates in suppression certain value is substituted a special value “?” or “\*” for its each occurrence and suggests that any figure can be positioned as a substitute. Suppression radically diminishes the data quality if not correctly used. Hence generalization is used more than suppression for K-anonymity.

In order to overcome limitations of generalization a new anonymization technique called slicing is used. By partitioning data both horizontally and vertically, slicing conserve enhanced information usefulness than generalization and can be used for membership disclosure protection

The overall focus of this project is to protect privacy of confidential database. The first part of this project focuses on privately updating confidential database without violating its K-anonymity. The second part focuses on overcoming various limitations of Generalization with a new privacy preserving technique called slicing and updating this sliced anonymous database without violating the privacy of database.

## II. RELATED WORK

Tiancheng Li, Ninghuli Li, Jian Zhang, Ian Molloy proposed a technique for data anonymization called slicing. Slicing gives attribute disclosure and membership disclosure protection and works for high volumes of data. Data slicing is efficient for high dimensional data and preserve improved data usefulness [1]. Slicing splits the relationship between different columns, but sustain the relationship of attributes inside each column.

Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi, Proposed a method for privately updating a K-anonymous database without revealing the contents of users record to database owner and contents of Database to user, to preserve data integrity by establishing the anonymity of Database. There are two methods of K-anonymity, suppression and Generalization. Suppression technique enables the database owner and data provider update anonymous database without violating privacy and confidentiality of database. Here database is updated without the owner gaining knowledge of tuple contents and without the user getting information of database contents. Here encryption is used in that sender and receiver secure their communication by encrypting them. To ensure the private authentication of the database anonymity after the inclusion of new record, sender and receiver uses a commutative and homomorphic encryption scheme. The second method is based on data anonymization using generalization. It implements secure set intersection protocol, for privately updating a confidential and anonymous database [2].

A. Macchanavajjala has brought in L-diversity. L-diversity is based on the principal that every associated group of records has at least l well-represented sensitive values [4]. L diversity is used with bucketization and slicing. It requires a clear separation between identifying attributes and susceptible attributes. Bucketization is not useful to protect for membership disclosure risk and sometimes it is very difficult to separate sensitive attributes.

L. Sweeney has demonstrated that to preserve privacy of person specific data, removing sensitive attributes is not enough. When multiple versions of same database were released it suffered from linkage attack. Research was done to protect against linkage attack and some solutions are projected in the literature to protect against linkage of public databases [5].

L. Sweeney and Samarathi offered privacy model called K-anonymity that makes use of value generalization. To accomplish K-anonymity every record must be impossible to differentiate from at least k-1 other records. K anonymity defends from harm against identity

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

## III. PROBLEM STATEMENT

Problem could be stated as: How updates could be made to anonymous databases thus ensuring the privacy of the user and the confidentiality of the database. The problem could be explained in more detail as if an anonymous database is owned by a database owner and suppose a user wants to insert certain tuple then in such a case the update should be made without revealing the contents to the database owner and the contents of the database is not revealed to the user [2]. Secrecy of database goes astray if database contents are released and disclosing the tuple contents breaks the privacy of the user.

The main aspect of this project is

- 1) To formulate anonymous database updating technique this preserves the notions of anonymity. First technique focuses on how to update k-anonymous database without violating k-anonymity of the database. K-anonymity has a weakness that it reduces the utility of the generalized data and correlations between different attributes are lost. To overcome these disadvantages new anonymity technique called Slicing is used. Slicing sustain data value because it groups very interrelated attributes together, and conserve the correlations between such attributes. Slicing divides the records vertically and horizontally. This System also proposes technique to update sliced database without violating its anonymity.
- 2) To devise techniques to deal with the tuple that fail to satisfy the anonymity.
- 3) To make certain more privacy and confidentiality of the database by introducing anonymity

## IV. METHODOLOGY

K-anonymity can be implemented using suppression and Generalization.

### I Suppression Based Algorithm [2]

once value suppression method is used for anonymizing we apply encryption method for tuple exchange between data provider and data owner and check if the new tuple breaks k-anonymity of original table before updating it into the table, if not allow to update else don't allow to update the table.

In suppression algorithm  $t$  stands for private tuple offered by Data provider,  $T$  symbolize secret and anonymous database,  $QI$  represents Quasi-Identifier which consist of some attribute of database that if utilized with several external information to discover a specific entity. Suppression algorithm works as follows:

Step 1: Data owner generates an encrypted copy of tuple from suppressed anonymous database, including merely the  $s$  non-suppressed attributes for that data provider.

Step 2: Data provider receives data from Data owner which is encrypted replica. Data provider encrypts the actual values of attributes in the tuple and sends it to data owner.

Steps 3-4: Data owner check the condition that non suppressed attributes received from data owner matches with one of the tuple in anonymous database.

If accurate, data provider's new tuple is inserted to table  $T$  else,  $t$  breaks  $k$ - anonymity when inserted to database [2].

### II Generalization Based algorithm [2]

In generalization algorithm attributes are replaced with general value based on Value Hierarchy Graph (VGH) [1]. Value generalization hierarchies (VGHs) are a set of possible values that can be replaced by a particular attribute value while implementing generalization. Table  $T$  is in anonymous database hold by data owner. To insert new tuple  $t$  into this anonymous database data provider will create this tuple having value for each attribute and sends this tuple to data owner. Suppose  $u$  is size of anonymous tuple which corresponds to Value Generalization Hierarchies subsequent to anonymous attributes acknowledged by data owner. As soon as request for new tuple insertion reaches to data owner he selects a tuple from table  $T$  randomly. Data owner figures out function  $\gamma = \text{GetSpec}(\delta)$  which gives bottom values from value generalization hierarchy. Now Data owner verifies that values contained in data providers tuple matches with values in the anonymous tuple which is randomly taken from anonymous database. If values match with any of the tuple randomly selected tuple  $t$  can be inserted to anonymous database and is properly generalized [2].

The protocol works as follows:

Step 1: Data owner arbitrary selects a tuple from witness set.

Step 2: Data owner calculates  $\gamma = \text{Get Spec}()$ .

Step 3: Data provider and Data owner mutually calculates  $s = \text{SSI}(\gamma, \tau)$ .

Step 4: Compare  $s$  and  $u$  if both are equal then  $t$ 's generalized form can be inserted to  $T$ .

Step 5: Or else, Data owner recursively do the above trial until either  $s=u$  or witness set is empty.

### III Slicing [1]

Slicing splits attribute uniformly horizontally and vertically. In vertical division significantly interrelated attributes are represented by one column and comparatively less interrelated attributes are assembled independently. Tuple are clustered forming buckets by partitioning horizontally. Subsequent to attribute grouping sensitive column values are randomly shuffled [1]. Slicing works in three main phases:

1. Attribute partitioning
2. Tuple partitioning
3. Column generalization

Functional procedure:-

Step 1: Obtain the data set from the database.

Step 2: Divide the records into columns and buckets.

Step 3: Swap the sensitive values.

Step 4: Multi set values generated and displayed.

Step 5: Dataset attributes are combined and safe and sound data presented.

When new record is inserted to sliced database tuple is checked if it violates the privacy of database. There is check for l-diversity to protect against homogeneity and background knowledge attack .

### V. IMPLEMENTATION

The proposed system contains a hospital database containing patient’s details such as name, disease, treatment. Doctors, patients and hospital staff are data provider in this case and data checker is to check whether a given record violates the privacy of anonymous database or not.

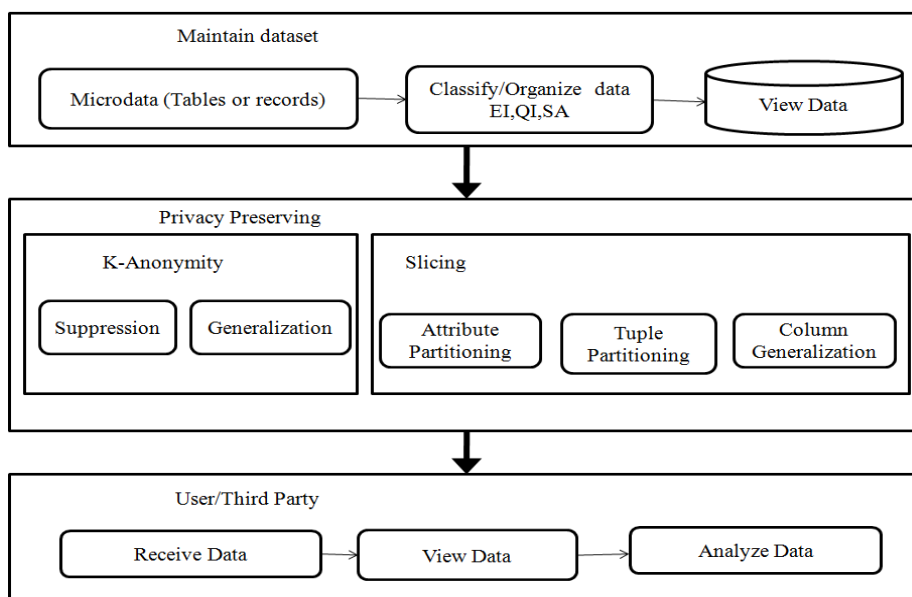


Fig 1 : Architecture of Privacy preserving data publishing

Architecture of the Privacy preserving data publishing is as shown in figure1. In first phase maintain dataset in the form of tables and records. Attributes are classified as explicit identifier, quasi identifier and sensitive attributes. Second step various data anonymization techniques are used on the data set such as K-anonymity and slicing. This anonymized dataset is published for analysis purpose.

# International Journal of Innovative Research in Science, Engineering and Technology

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 10, October 2015

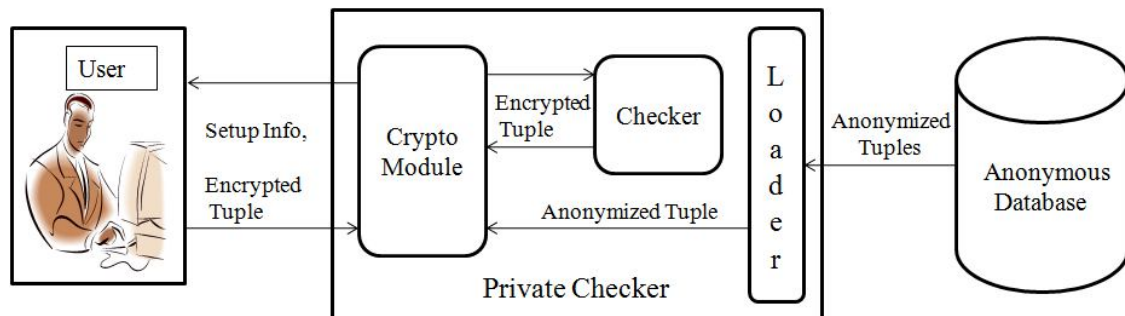
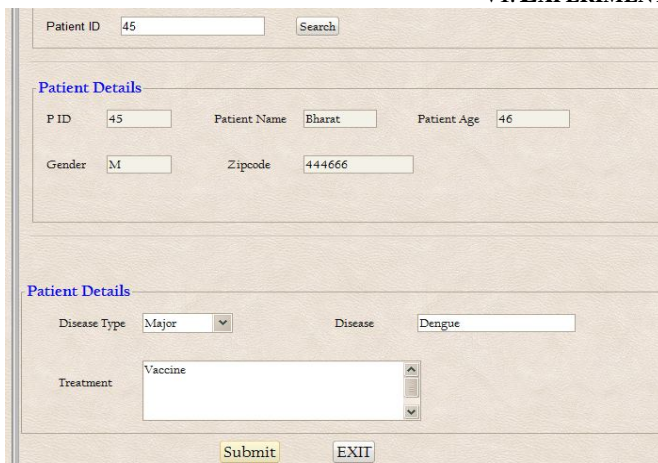


Fig 2: Proposed System Block Diagram

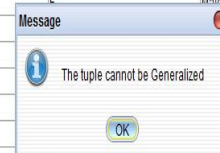
Proposed System is made of Private checker, User and database. Private checker is made up of Cryptography module, checker and Loader module. The data entered by user is stored in Cryptography module. This module is responsible for encrypting all tuple exchanged between User and the private Updater. Loader module fetches data or tuple from anonymous database. Checker module executes all the functions to ensure if the data being inserted equivalent with the data in the anonymous database by means of suppression and generalization or slicing method. The decision whether insertion of new record is feasible into anonymous database or not is taken by Private Checker module [2]. The projected system include module for patient data modification and module to deal with record which breaks anonymity of database.

## VI. EXPERIMENTAL RESULTS



The screenshot shows a web-based form for patient details. At the top, there is a search bar with 'Patient ID' set to 45 and a 'Search' button. Below this, there are two sections for 'Patient Details'. The first section includes fields for P ID (45), Patient Name (Bharat), Patient Age (46), Gender (M), and Zipcode (444666). The second section includes fields for Disease Type (Major), Disease (Dengue), and Treatment (Vaccine). At the bottom, there are 'Submit' and 'EXIT' buttons.

	AGE	GENDER	TYPE	DISEASE
01		M	Minor	Gastritis
01		M	Major	Hyperglycemia
01		F	Major	Cancer
01		F	Major	Cancer
01		C	Major	Brain
01				Flu
01				Flu
01				Flu
01				Throat Infection
01				Sexual
01		F	Major	Sexual
01		F	Major	Cancer
01		F	Major	Cancer
01		F	Minor	Flu
01		M	Minor	Respiratory Infection
01		M	Major	Hyperglycemia



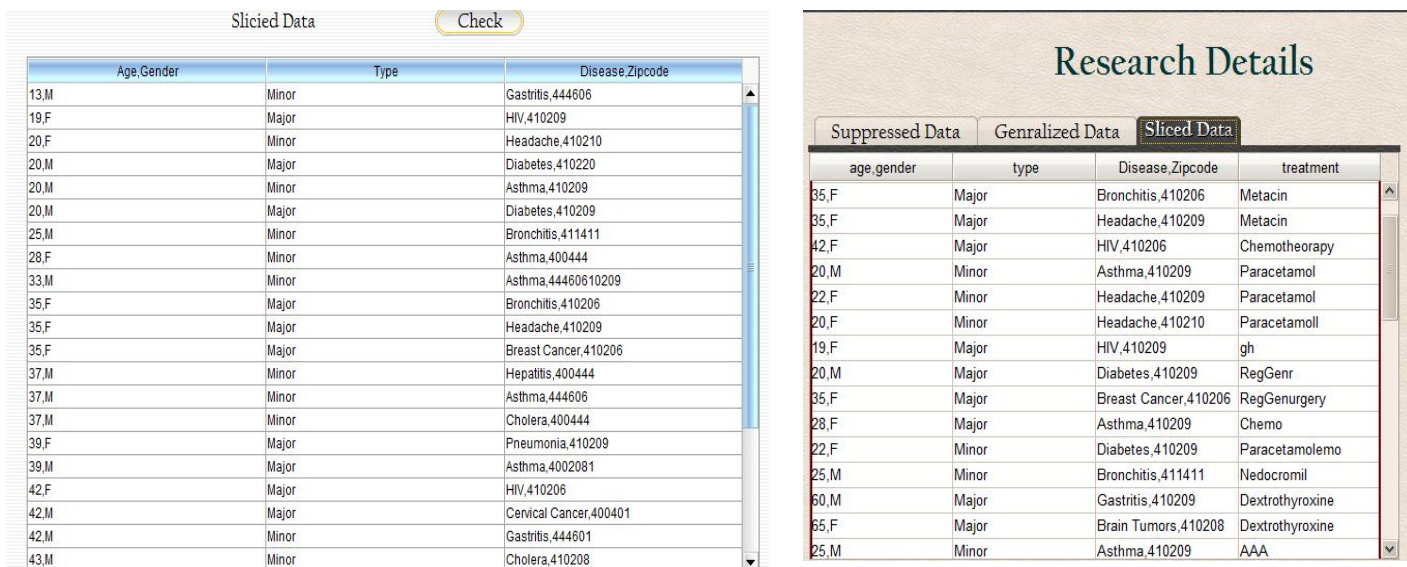


Fig 3: Screenshots of implemented system

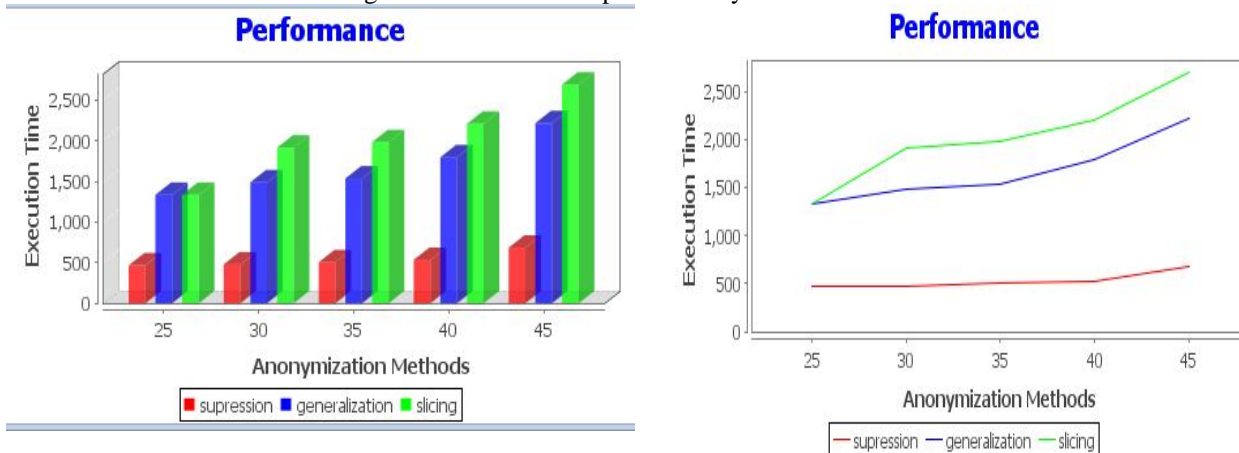


Fig: 4 Comparison Graphs.

## VII. CONCLUSION

Data anonymization technique such as K-anonymity and slicing is used when databases are made available for unhampered use. When it is required to update an anonymous database encryption is used for data exchange between data owner and data provider. If anonymity of database is dishonoured on insertion of new record, then update to database is disqualified. Suppose new record is declared to be suitable for insertion into anonymous database then such record is inserted in table. To get rid of limitations of K-anonymity slicing is used. Slicing uphold against attribute disclosure and membership disclosure. Slicing conserve better data utility than generalization and is more useful in high volume of data linking the sensitive attribute. This project gives method to update anonymous database without disclosing the database and tuple contents to data provider and data owner and thus maintains its anonymity.

## REFERENCES

- [1] Tiancheng Li, Ninghuli Li, Jian Zhang, Ian Molloy, Slicing: A new approach for privacy preserving data publishing, IEEE transactions on knowledge and data engineering, vol 24, no 3, march 2012.
- [2] Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi, Privacy -Preserving Updates to Anonymous and Confidential Databases, IEEE, 2011.



ISSN(Online) : 2319-8753  
ISSN (Print) : 2347 -6710

# International Journal of Innovative Research in Science, Engineering and Technology

*(An ISO 3297: 2007 Certified Organization)*

**Vol. 4, Issue 10, October 2015**

- [3] A. Trombetta, E. Bertino. Private updates to anonymous databases. In Proc. Int'l Conf. on Data Engineering (ICDE), Georgia, US, 2006.
- [4] Ashwin Machanavajjhala Johannes Gehrke Daniel Kifer Muthuramakrishnan Venkitasubramaniam,  $\ell$ -Diversity: Privacy beyond k-Anonymity, Proceedings of the 22nd International Conference on Data Engineering, 2006.
- [5] L. Sweeney. K-anonymity: a model for protecting privacy. International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10(5), 557–570, 2002.