Preserve the Privacy of Anonymous and Confidential Database using K-anonymity

Vishakha S. Kulkarni
M.E. Student, Dept. of Computer Science,
Alard College of Engineering & Management,
Pune, India

M. S. Gayathri
Asst.Professor, Dept.of Computer Science,
Alard College of Engineering & Management,
Pune, India

Abstract— **In field of IT sector to maintain privacy and confidentiality of data is very important for decision making. So there is requirement of certain data to be published and exchanging of the information is in demand. The data to be exchanged contains sensitive information which moves around various parties and this may violate individual's privacy. So to preserve information in its accurate form while moving among various parties, my aim is to provide mechanism known as k_anonymous technique that doesn't allow the unauthenticated user to modify the data. In this application two protocols that will solve this problem based on suppression and generalization k-anonymous and confidential databases are used. The protocols rely on well-known cryptographic assumptions, and it provides theoretical analyses to proof their experimental results to illustrate their efficiency.**

Keywords- Anonymity, data management, privacy, secure computation.

I. Introduction

The database is an important asset for many applications and thus their security is important. Data confidentiality is relevant because of the value that data have. As the medical data of patients collected by maintaining the history of patients over several years represent a valuable data that needs to be protected. Due to this requirement gave rise to a large variety of approaches that aim at better protecting data confidentiality and data ownership. Data confidentiality is the problems created by an unauthorized user to get the knowledge about data stored in the database. Access to individual's personal information is limited by privacy. It deals with the authorized access by authenticated users.

Database privacy should follow confidentiality, integrity, and availability of personal data, not only confidentiality alone. Anonymization is required to provide privacy. Anonymization means masking the data. In this identifying information is removed from the original data to protect personal or private information. Data Anonymization allows transferring of information between two organizations, by converting text data in to non-readable form using encryption method. K-Anonymization is one of the approaches that maintain privacy of data. In K-Anonymization approach, at least K-tuples should be indistinguishable by masking values.

The data providers are medical facilities (Hospitals) that provide sensitive information through anonymous authentication and connection. Authentication is done using user ID and password. The users shown in Fig. 1 can be the medical researchers who have the permission to access DB. The data provider's data privacy is protected from these researchers as the database is in anonymous form.

The existing system deals with difficulties concerning that the contents of tuples and DB is not revealed by users, how data integrity can be preserved by establishing the anonymity of DB. It deals with algorithms for database anonymization. It deals with how privacy of data of whole databases and their owner and also individual tuples and its owner is maintained without disclosing the contents.

The system consider suppression based anonymous database. A secure protocol is presented that privately checks whether K-anonymous database retains its anonymity even after insertion of a new tuple.

II. LITERATURE SURVEY

In references paper many fundamental methods and techniques are used to make maintain the data of database in anonymous form to provide privacy and confidentiality of data. By performing the literature survey, various issues and challenges are identified in existing system.

In 2013, secure protocol is presented for privately checking whether K-anonymous database remains anonymous even after insertion of new tuple. Quasi-Identifier (QI) [1]: QI is a set of attributes used to identify individual's information. To prevent the attack, masks the values of Quasi-Identifiers using either suppression based or Generalization based Anonymization methods. The Quasi –Identifiers for the below dataset is {Zip code, Age, Nationality}. So we must anonymize the Quasi-Identifiers value, because attacks come based on Quasi-Identifiers. Algorithm to compute an anonymized version of tuple T use encryption algorithm RSA (Rivest, Shamir, Aldemen) to encrypt the tuple T. RSA is the most common public key (Asymmetric key)algorithm. It uses two keys Private and Public key. It deals with algorithms for database anonymization. The problem is to check even after connecting the tuple the database is still k-anonymous, such that the actual data from, tuples or database can't be viewed [2]. The same amount of preservation is done for all persons, without considering their needs.

K-anonymity a formal protection model [3] that contains set of accompanying policies for deployment is proposed. K-anonymity protection is provided by a release if the information of each person in the release is indistinguishable from at least k-1 individuals whose information is also contained in the release. Some system proposed technique to satisfy everybody's requirement that performs the minimum generalization, and retains the large information from the micro data.

In 2012, Private Checker's prototype [4] is composed by the modules as: a crypto module that of encrypts all the tuples exchanged between user and the Private Updater, using the techniques a checker module that performs all the controls. The Private Checker prototype provides the functionality that check on whether insertion of tuple into the k-anonymous DB is possible. In 2012, the system is provided with facility for allowing the right users to access into the database by comparing existing data and the updates and make sure there is no redundancy and helps to analyses the data in database. K-Anonymization allows database to maintain a suppressed and generalized form of data such that data is much secured. The cryptography technique [5] is used to secure the saved data in database safely such that the information is encrypted, stored and can be retrieved and decrypted back to original with specific authorization.

In 2008, some simple protocols that are often used as basic building blocks, or primitives, of secure computation protocols. The protocols include oblivious transfer [6] and oblivious polynomial evaluation, which are two-party protocols, and homomorphic encryption, which is an encryption system with special properties. Oblivious transfer protocols have been designed based on virtually all known assumptions which are used to construct trapdoor functions, and also based on generic assumptions such as the existence of enhanced trapdoor permutations. A homomorphic encryption scheme is an encryption scheme which allows certain algebraic operations to be carried out on the encrypted plaintext, by applying an operation to the corresponding ciphertext.

In 2013, system is a new generalization framework based on the concept of personalized anonymity is described. To achieve personalized anonymity greedy Framework algorithm [7] is used. It works in two steps. In the first steps a generalization function for every QI attribute is chosen and the generalized value is obtained for all tuple t Є T. The Generalized tuple are divided into QI-Group. In the second step SA-generalization uses a different function for each group. This strategy achieves less information loss, by allowing each group to decide the amount of necessary generalization. SA-generalization results in less precise values on sensitive attribute, it retains more information on the QI attributes.

III. IMPLEMENTATION DETAILS

The information concerning a data provider is stored in a single tuple, and DB is kept confidentially at the server. Since DB is anonymous, the data provider’s privacy is protected from researchers. Such task is guaranteed through the use of anonymization. Preserving the privacy & confidentiality without revealing the contents of tuple and DB is done by establishing the anonymity of DB. A secure protocol is presented for privately checking whether K-anonymous database remains anonymous even after insertion of a new tuple. Suppressed the value of attribute by replacing “*” and Generalized the value with related possible general value to maintain the k-anonymity in database. Thus by making such k-anonymity in table it becomes complicated for third party to identify the record. In the system, before a tuple is inserted the data can be encrypted using shared secrete key AES algorithm. Based on a commutative encryption function the data provider can share a secrete key with each other using Diffie-Hellman Algorithm.

A. *Proposed Model*

As shown in Fig. 1, proposed system consists of following modules:

a. Login Module.

b. Data Provider for Suppression and Generalization.

c. Server for Suppression and Generalization.

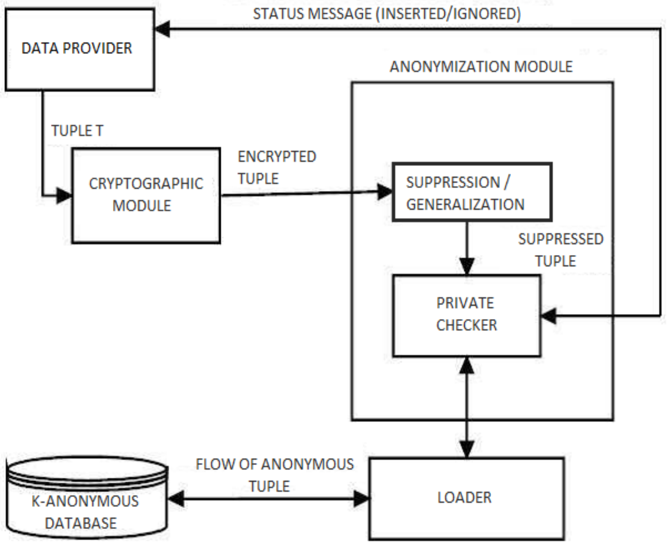


Figure1. Proposed System Architecture

In this proposed model a secure protocol is presented that privately checks whether database remains k-anonymous even after insertion of new tuple. Quasi-Identifier (QI): QI is a minimal set of attributes which is used to uniquely identify individuals. Attack is mainly using Quasi-Identifier. Attacks may be re-identification or linking attack. To prevent the attack, masks the values of Quasi-Identifiers using either suppression based or Generalization based Anonymization methods. In Suppression based anonymization method, mask the Quasi-Identifiers value using a special symbol like * and in Generalization based anonymization method, replace a specific value with a more general one using Value Generalization Hierarchies (VGH).

The diffie hellman key exchange algorithm is used to generate private secure key. Then AES algorithm is applied to encrypt and decrypt data by using the key generated by the diffie hellman key exchange algorithm. When user enters his information then this information is encrypted by using AES and also all data in table is encrypted using same algorithm. If information from user matches with table information the tuple will decrypted and inserted into table.

Let the data provider is X and Suppression & Checker module is Y. The flow of operation is given below:

a. X sends a tuple T in to cryptographic module.

b. The cryptographic module encrypts the tuple (encryption means to convert the plane text in to ciphertext) and send it to the Suppression &Checker module(Y)

c. Y, then compute.

d. The anonymized version of tuple T.

e. Check whether the data is matched with data's in the loader.

f. The loader reads chunks of anonymized tuple from the K-Anonymous database.

g. If the tuples are not matched, then the loader reads next chunks of anonymized tuple from the k-Anonymous database and checking can be performed.

h. If any match found, then the tuple t can be inserted in to the K-Anonymous database.

i. Finally we can send a message to the data provider about the status of the tuple T (status are INSERTED/ IGNORE).

j. According to the status, the data provider can decide further action.

B. *Proposed Methodology*

1) *Module1: Login Module*

Module 1 is the Login module. The user wanted to enter the data into the database is authenticated first. User enters username and password; if it is correct then user is validated and can proceed further. If user enters wrong information then user is invalid user and can't proceed further.

2) *Module2: Data Provider for Suppression Method*

a) *Data Provider for Generalization Method*

In the anonymous databases the meaning of Anonymization can be easily understand. Anonymization is technique which hides sensitive attribute value in such a way that it cannot be identified back. In k-anonymization approach the total number of rows is k and k cannot be differentiated with other k-1 rows by taking into account only a set of attributes, then this table is known as K-anonymized. Privacy preservation can be done by simply using k-anonymization approach on suppression and generalized techniques. In suppression method all data which is sensitive from database is suppressed by using "*", and in Generalization method a value is replaced with a "less-specific but consistent" value according to apriori established value generalization hierarchies (VGHs).

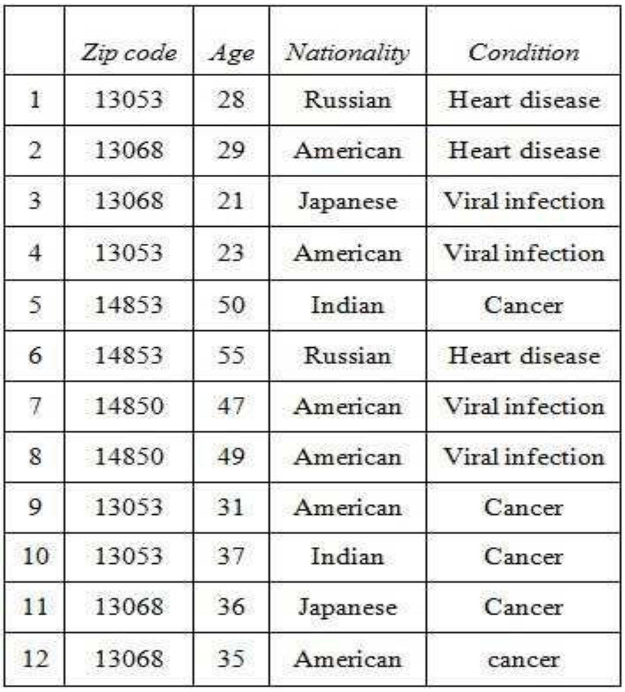
b) *Suppression Based Anonymous Technique*

When suppression-based anonymization method is used, consider a table $T= \{t1,\dots\dots\dots,\}$ tuples over the attribute set A. In suppression method, the values of some well-chosen attributes are masked to form subsets. It is mask with the special value „*". Forming the subset and classify that subsets by using Quasi-Identifier (QI). Quasi-Identifier (QI): Each record contains a number of attributes: some attributes are unique and personal attributes (such as disease and salary) and some may be repeated and general that is quasi-identifiers (called QI, such as zipcode, age, and gender) by taking this it can easily identify someone. Consider the example of patient As shown in table 1 which contains original database (Table T) having Quasi-Identifier QI={Zipcode, age, Nationality} or more sensitive three attributes value. After applying suppression based technique on original dataset the original dataset is anonymized and Table 2 shows a suppression based k-anonymization with k=2 it means that at least k=2 tuples should be indistinguishable by masking values.

c) *Generalization Based Anonymous Technique*

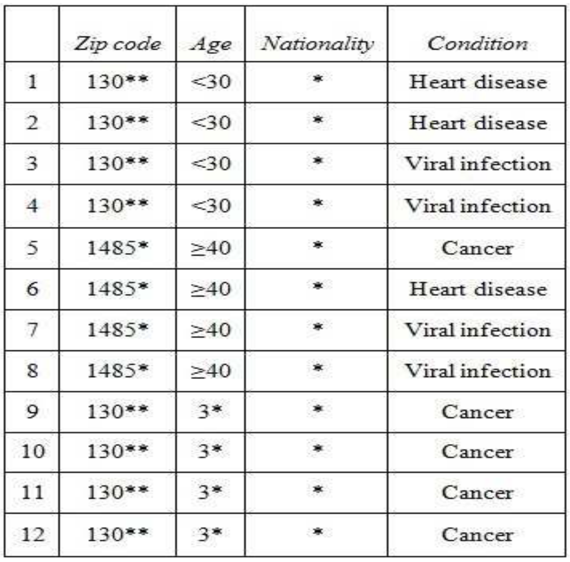
In generalization-based anonymization consists in substituting the values of a given attribute with more general values in the database, according to a priori established value generalization hierarchies (VGHs) with some Cryptographic Primitives. In Table 1 original information is stored and after performing generalization techniques on original dataset the original dataset is anonymized and table 2 gives generalized data with k=3. The Generalization is technique which replaces a value with a "less-specific but semantically consistent" value. It can be defined based on the VGH which specify how the data will be generalized. According to the VGH of DISEASE, say that the value of disease is generalized according to the disease causes. Like "HIV" cause by virus so it can be generalized to "Diseases Caused by virus". The attribute Age is generalized to the interval (30-39).

TABLE I Original Patient data



	Zip code	*Age*	*Nationality*	*Condition*
1	13053	28	Russian	Heart disease
2	13068	29	American	Heart disease
3	13068	21	Japanese	Viral infection
4	13053	23	American	Viral infection
5	14853	50	Indian	Cancer
6	14853	55	Russian	Heart disease
7	14850	47	American	Viral infection
8	14850	49	American	Viral infection
9	13053	31	American	Cancer
10	13053	37	Indian	Cancer
11	13068	36	Japanese	Cancer
12	13068	35	American	cancer

TABLE II Anonymous Patient data



	Zip code	Age	Nationality	Condition
1	130**	<30	*	Heart disease
2	130**	<30	*	Heart disease
3	130**	<30	*	Viral infection
4	130**	<30	*	Viral infection
5	1485*	≥40	*	Cancer
6	1485*	≥40	*	Heart disease
7	1485*	≥40	*	Viral infection
8	1485*	≥40	*	Viral infection
9	130**	3*	*	Cancer
10	130**	3*	*	Cancer
11	130**	3*	*	Cancer
12	130**	3*	*	Cancer

3) Module3: Server for Suppression Method.

d) Server for Generalization Method.

In this module, the suppressed tuple is compared by the tuple loaded from k-anonymous database in loader. Private checker compares this both the tuple, if they are same then the tuple is inserted. Otherwise the tuple is ignored. The system actually updates the database depends on the result of the anonymity checker. In some cases the insertion or updation failed in k-anonymous database then it waits until k-1 value becomes positive and other tuples fail the insertion.

C. Implementation of algorithm

1) AES algorithm: Advanced Encryption Standard

The AES algorithm is the algorithm based on permutations and substitutions of data. Permutations are rearranging of data, and in substitutions one unit of data is replaced with another unit of data. AES algorithm is a block cipher which has a block of length 128 bits. AES can be applied to three different key lengths: 128, 192, or 256 bits. In AES cipher key size used denotes the number of repetitions of transformation rounds which convert the plaintext, into ciphertext. To transform ciphertext back into the original plaintext a set of reverse rounds are applied that uses same encryption key.

2) Diffie-Hellman Key Exchange Algorithm:

In Diffie–Hellman algorithm a shared secret key is established that can be used for secret communications while exchanging data over a public network. The Diffie–Hellman is a key exchange method that allows two parties which does not have any information of each other and want to establish a shared secret key over communications channel which is insecure. Using a symmetric key cipher this key can then be used to encrypt subsequent communications. Diffie and Hellman uses a commutative function based on discrete logarithm.

IV. RESULT AND DISCUSSION

A Private Checker is composed by the following modules: a crypto module that is in charge of encrypting all the tuples exchanged between a user and the Private Updater, a checker module that performs all the controls, a loader module that reads chunks of anonymized tuples from the k-anonymous DB. The chunk size is fixed in order to minimize the network overload. The functionality provided by the Private Checker prototype regards the check on whether the tuple insertion into the k-anonymous DB is possible. The information flow across the above mentioned modules is as follows: after an initial setup phase in which the user and the Private Checker prototype exchange public values for correctly performing the subsequent cryptographic operations, the user sends the encryption of her/his tuple to the Private Checker; the loader module reads from the k-anonymous DB the first chunk of tuples to be checked with encrypted tuple. Such tuples are then encrypted by the crypto module. The checker module performs the above mentioned check one tuple at time in collaboration with the user. If none of the tuples in the chunk matches the User tuple, then the loader reads another chunk of tuples from the k-anonymous DB.

V. CONCLUSION

Data confidentiality and privacy is a challenging problem faced in case of security of database. In this work, two secure protocols are presented for privately checking whether a k-anonymous database retains its anonymity once a new tuple is being inserted to it. Since the proposed protocols ensure the updated database remains k-anonymous. The data provider's privacy cannot be violated from any user's updating the table. So the database is updated properly using the proposed protocols. This is useful in medical application. If insertion of record satisfies the k-anonymity then such record is inserted in table and suppressed the sensitive information attribute by "*" to maintain the k-anonymity in database. Thus, by making such k-anonymity in table that makes unauthorized user too difficult to identify the record.

ACKNOWLEDGMENT

I would like to express my sincere and profound thanks to Prof. M. S. Gayathri, Associate Professor of Alard College of Engineering & Management, Pune my project guide, for consistent support and valuable suggestions.

REFERENCES

[1] Sivasubramanian .R, K.P. Kaliyamurthie, "Privacy-Preserving Updates to Anonymous Databases", IJCSMC, Vol. 2, Issue. 4, April 2013, pg.582 – 587.

[2] Alberto Trombetta, Wei Jiang, Member, IEEE, Elisa Bertino, Fellow, IEEE, and Lorenzo Bossi, "Privacy-Preserving Updates to Anonymous and Confidential Databases" IEEE Transactions on Dependable and Secure Computing, VOL. 8, No. 4, July/August 2011.

[3] L. Sweeney, "K-Anonimity: A Model for Protecting Privacy", International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 10 (5), 2002.

[4] Dr.K.P.Thooyamani, Dr.V.khanaa, Er.M.R.Arun Venkatesh, "Privacy-Preserving Updates to Anonymous and Confidential Database", International Journal of Data Mining Techniques and Applications, Volume: 01 Issue: 01 January-June 2012.

[5] Ishwarya M. V, Dr. Ramesh Kumar. K, "Privacy Preserving Updates for Anonymous and Confidential Databases Using RSA Algorithm", International Journal of Modern Engineering Research (IJMER), Vol.2, Issue.5, Sep.-Oct. 2012.

[6] Yehuda Lindell, Benny Pinkas, "Secure Multiparty Computation for Privacy-Preserving Data Mining", May 6, 2008.

[7] Rajeshwari Suryawanshi, Sulabha Patil,"Privacy Preserving updates to Personalized Anonymity Based Anonymous and Confidential

Database", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, October 2013.

[8] Rakesh Agrawal, Alexandre Evfimievski, Ramakrishnan Srikant, "Information haring Across Private Databases", IBM Almaden Research Center June 9, 2003.

[9] Mahendrababu P, Rajarajan G,"Apprising in Secured Manner to Anonymous and Confidential Databases", International Journal of Engineering Research & Technology (IJERT), Vol. 2 Issue 1, January-2013.

[10] Ebin P.M, Brilley Batley. C, "Privacy Preserving Suppression Algorithm for Anonymous Databases", International Journal of Science and Research (IJSR), Volume 2 Issue 1, January 2013.

[11] Mr. Mahesh T.Dhande, Mrs. Neeta A.Nemade, "Performance Improvement of Privacy Preserving in K-anonymous Databases Using Advanced Encryption Standard Technique", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 6, June 2013.

[12] Rajeshwari Suryawanshi, Prof.Parul Bhanarkar, Rashtrasant Tukdoji Maharaj, "Survey On Privacy Preserving Updates On Anonymous Database", International Journal of Engineering Research & Technology (IJERT),Vol. 2 Issue 1, January- 2013.

[13] Neha Gosai, S.H.Patil, "Generalization Based Approach to Confidential Database Updates", International Journal of Engineering Research and Applications, Vol. 2, Issue 3, May-Jun 2012.

[14] Lavanya.Gunasekaran, R. Sujatha,"Enhanced Privacy Preserving Updates for Anonymous and Confidential Databases", International Journal of Computer Networks and Wireless Communications (IJCNWC), Vol.2, No.2, April 2012.

[15] Deepa.B, Meena.R, "Privacy Control Methods for Updating Confidential Databases", International Journal of Soft Computing and Engineering (IJSCE), Volume-1, Issue-ETIC2011, January 2012.