Privacy Preserving Suppression Algorithm for Anonymous Databases

Ebin P.M[1], Brilley Batley. C[2]

[1,2]AMIE, Assistant Professor
Department of Computer Science & Engineering,
Hindustan University, Chennai, India
pmebin74@gmail.com

Abstract: ***Suppose a medical facility connected with a research institution and the researchers can use the medical details of a patient without knowing the personal details. Thus the research data base used by the researchers must be anonymized (Sanitized). We can consider another problem in the area of census. Individuals give the private information to a trusted party (Census Bureau) and the census bureau must publish anonymized or sanitized version of data. So anonymization is done for privacy. Our works deals with privacy in database system.***

Keywords: Privacy, Anonymization, Secure Computation, Suppression.

1. Introduction

Today privacy or security has become crucial. So we mainly concentrate on privacy [2]. Privacy limits the access to individual's personal information. It deals with authorized access. The collection of data usually called as the database, contains large bodies of information. To provide security to these databases is big issue. Database privacy should fall on a balance between confidentiality, integrity, and availability of personal data, rather than confidentiality alone. Confidentiality means only authorized users can read the data. Usually confidentiality can be achieved by using some cryptographic tools. Not only confidentiality, but Anonymization [1] is still required to provide privacy.

Anonymization means masking. That is identifying information is removed from the original data to protect personal or private information. Data Anonymization enables transferring information between two organizations, by converting text data in to non-human readable form using encryption method [4]. There have been lots of approaches developed. K-Anonymization is one of the approaches. In K-Anonymization approach, at least K-tuples should be indistinguishable by masking values [3]. So the probability of linking a given data value to a specific individual is very small, and the individuals cannot be uniquely identified by linking attacks. The problem arises at the time of data updation. Without revealing the content of T (T is a tuple which is going to be inserted) and database How to preserve the privacy? How to privately check whether a K-anonymous database retains its anonymity once a new tuple 'T' is being inserted in to it. This paper will give the solution. Two approaches can be used for Anonymization. One is Suppression and the other is Generalization. In this paper we deals with Suppression based Anonymization approach.

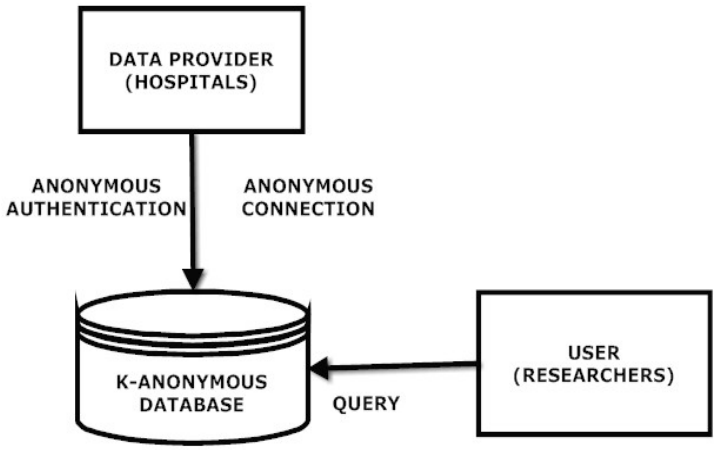

Figure 1: Anonymous Database

Figure: 1 shows, anonymous (sanitized) database system, which is used by the researchers. The data providers are medical facilities (Hospitals) through anonymous authentication and connection. Authentication can be done using user ID and password method. Anonymous connection can be done using Crowds or Onion routing protocols. Anonymous updation is proposed in this paper. Crowds increase the privacy of web transaction; the main idea is "blending in to the crowd". That is, hiding one's action within the actions of many others. A user first joins a crowd of other users. The users request to a web server is first passed to a random number of the crowd. That member can either submit the request directly to the end server or forward it to another randomly chosen member. In the latter case the next member chooses to submit or forward independently. When the request is finally submitted, it is submitted by a random member, thus preventing the end server from identifying its true initiator. Even crowd members cannot identify the initiator of the request. It is used for anonymous connection, it protect IP addresses and other sensitive information [7].

Onion routing supports private and anonymous connection/communication over a public network. Onion

routing is flexible communication infrastructure that is resistant to both eaves dropping and traffic analysis. It is a bi-directional, near real-time and can be used for both connection oriented and connection less traffic. When a packet is received by the first onion router, it is encrypted once for each additional router it will pass through. Each subsequent Onion router unwraps one layer of encryption until the message reaches its destination as plain text [6].

2. Existing System

There are various techniques to provide confidentiality and privacy to anonymous database like Data Reduction, Data perturbation and Secure Multiparty Computation etc.

The first approach is Data perturbation technique. The idea of protecting databases through data suppression or data perturbation has been extensively investigated in the area of statistical database. Basically there are two types of data perturbation. First type Probability distribution approach and the second type are called the value distortion approach. In the probability distribution, Original database is replaced by sample from distribution or by distribution itself and the value distortion approach perturbs data elements or attributes directly by either additive noise, multiplicative noise, or some other randomization procedures. Agrawal proposed a value distortion technique to protect the privacy by adding random noise from a Gaussian distribution to the actual data. They showed that this technique appears to mask the data while allowing extraction of certain patterns like the original data distribution and decision tree models with good accuracy.

Second research approach is Secure Multiparty Computation method consider problem of evaluating function of two or more parties‟ secret input in such a way that each party does not get anything else except specified output [8]. SMC represents an important class of techniques widely investigated in the area of cryptography. However, these techniques generally are not efficient.

The third research direction is related to the area of private information retrieval, which can be seen as an application of the SMC techniques to the area of data management. The problem of privately updating database has not been addressed in that these techniques only deal with data retrieval.

Finally, the fourth research direction is related to query processing techniques for encrypted data. The approaches do not address the K-anonymity problem since their goal is to encrypt data, so that their management can be outsourced to external entities. Most of privacy models developed are based on k-anonymity property-anonymity property deals the possibility of indirect identification of records form public databases-anonymity means each released record has at least (k-1) other records in the release whose values are indistinct. K-anonymity and SMC are used in privacy-preserving data mining, but they are quite different in terms of efficiency, accuracy.

3. Proposed System

The proposed system consider suppression based anonymous database. A secure protocol is presented for privately checking whether K-anonymous database retains its anonymity once a new tuple is being inserted.

Quasi-Identifier (QI): QI is a minimal set of attributes 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).

	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 1: A table contains patient's data

The Quasi –Identifiers for the above the dataset is {Zip code, Age, Nationality}. So we must anonymize the Quasi-Identifiers value, because attacks come based on Quasi-Identifiers.

	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

Table2: 4 - Anonymous patient data

We suppress (mask) the Quasi-Identifiers values. '*' denotes suppressed values. "Age = 3* means" that the Age is in the range [30-39]. Here 'Condition' is a non-suppressed attribute. We can say that it is a sensitive attribute.

A. Proposed system architecture

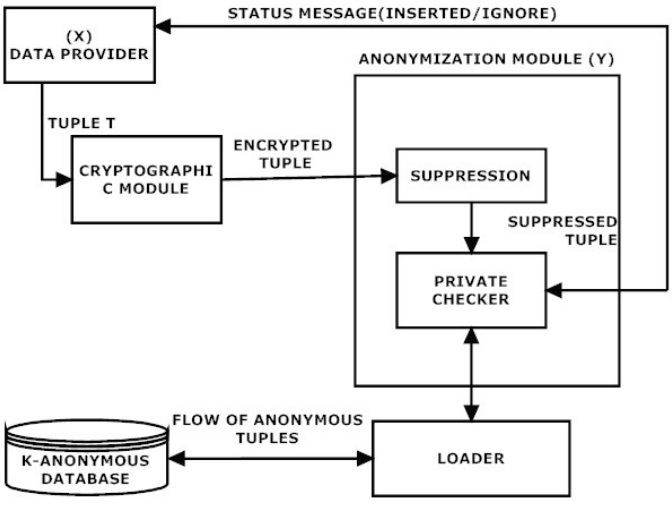

Figure 2: Proposed System Architecture

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 cipher text) 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. Otherwise can be ignoring.
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. Cryptographic primitives

Our algorithm to compute an anonymized (Sanitized) 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. The encryption scheme must be a commutative and product-homomorphic one. This encryption scheme allows performing mathematical operation over encrypted data.

Given a finite set K of keys and finite domain D, A Commutative and Product-homomorphic encryption scheme E is a polynomial time computable function

E: K×D -› D satisfying the following properties.

Commutativity : For all key pairs k1,k2 ϵ K and value d ϵ D ,then

Ek1 (Ek2 (d)) = Ek2 (Ek1 (d))

Product-homomorphism : For every k ϵ K and every value pairs d1, d2 ϵ D , the following equality holds:

Ek (d1). Ek (d2) = Ek (d1.d2)

Indistinguishability: It is infeasible to obtain data of plaintext from cipher text. The advantages are high privacy of data even after updating, and an approach that can be used is based on techniques for user anonymous authentication and credential verification.

The Diffie-Hellman key exchange algorithm allows the exchange of private encryption key. This algorithm can be used for key agreement, not for encryption and decryption. Here Diffie-Hellman is used to agree on shared secret key to exchange data between two parties.

4. Algorithm

STEP 1: X encrypt the tuple T, and send it to Y.

STEP 2: Y can decrypt tuple T and then suppress the personal identifiers in the tuple.

STEP 3: After the suppression check the nonsuppresed attributes in the tuple T and loaded tuples.

STEP 4: If any match found, insertion can be performed and send a status message "INSERTED".

STEP 5: If no match found, discard the tuple and send the status message "IGNORE".

5. Conclusion

In this paper, we have proposed secure protocol to check that if new tuple is being inserted to the database, it does not affect anonymity of database. It means when new tuple get introduced, k-anonymous database retains its anonymity. Database updates has been carried out properly using proposed protocol. 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.

6. Future Scope

The important issues in future will be resolved:

a. Implement database for invalid entries.
b. Solve problem of anonymity when initially table is empty.
c. When system fails to check tuple, it checks these tuple in wait state called hanging tuples. Try to resolve this problem.
d. Improving efficiency of protocol in terms of number of messages exchanged between user and database.
e. Implement real world database system.

References

[1] G. Aggarwal, T. Feder, K. Kenthapadi, R. Motwani, R. Panigrahy,D. Thomas, and A. Zhu, "Anonymizing Tables," Proc. Int'l Conf. Database Theory (ICDT), 2005.
[2] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Towards Privacy in Public Databases," Proc. Theory of Cryptography Conf. (TCC), 2005.
[3] L. Sweeney, "k-Anonymity: A Model for Protecting Privacy," Int'l J. Uncertainty, Fuzziness and Knowledge-Based Systems, vol. 10, no. 5, pp. 557-570, 2002.
[4] Privacy-Preserving Updates to Anonymous and Confidential Databases, Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi, Department of Computer Science and Communication, University of Insubria, Italy.
[5] Generalization Based Approach to Confidential Database Updates, Neha Gosai, S H Patil, Department of Computer Science, pune, Maharashtra, 2012
[6] Murdoch Steven J, Danezis G.low-cost traffic analysis
[7] TOR In: IEEE symposium on security and privacy May 2005.
[8] Brier, S. 1997. How to keep your privacy: Battle lines get clearer. The New York Times, January 13, 1997.
[9] G. Aggarwal, N. Mishra and B. Pinkas. Secure Computation of the k-th Ranked Element. In EUROCRYPT 2004, Springer-Verlag (LNCS 3027), pages 40{55, 2004}.
[10] J. Li, N. Li, W. Wins borough. Policy-hiding access control in open environment. In Proc of ACM Conf. on Computer and Communications Security (CCS), Alexandria, Virginia, 2005.
[11] N.R. Adam and J.C. Worthmann, "Security-Control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys (CSUR), vol. 21, no. 4, pp. 515- 556, 1989.

Author Profile

Mr Ebin P.M received bachelor's degree (AMIE) in computer science and Engineering from "The Institution of Engineers (INDIA)" of Kolkata in 2011 and doing master's degree in Computer Science and Engineering in Hindustan University, Chennai.