ISSN No. 0976-5697

Volume 4, No. 2, Jan-Feb 2013

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

An Approach to Deal the Untrusted Malicious Parties in Anonymous Database Systems

Ebin P.M (AMIE)*, Brilley Batley .C
Student*, Assistant Professor
Dept. of Computer Science& Engineering
Hindustan University, Chennai, India.
pmebin74@gmail.com*, cbrilleyb@hindustanuniv.ac.in

Abstract: Database is an important asset for many applications, so their security is most important one. Data confidentiality is particularly relevant because of the value, often not only monetary, that data have. For example, medical history of patients in a hospital over several years is an important asset, which must be protected. Consider a hospital is connected with a research center. It should be possible the researchers can use the medical details without knowing any personal information of patients .This type of information hiding can be called as anonymization (Sanitization).The sanitized version of data provides security and confidentiality. So the research data base used by the researchers must be anonymized (Sanitized). Data perturbation, Secure Multiparty Computation (SMC) and Anonymization are the different techniques used today. Here we consider an anonymous database system and an approach to manage the untrusted malicious parties in that system**.**

Keywords: Privacy, Anonymization, Secure Computation, Suppression.

I. INTRODUCTION

The collection of data, usually referred to as the database, contains informations relevant to an enterprise. Database systems are designed to manage large bodies of information. The database system must ensure the safety of the information stored, despite system crashes or attempts at unauthorized access. If data are to be shared among several users, the system must avoid possible anomalous results. A system security must consider the external environment. We say that a system is secure if its resource are used and accessed as intended under all circumstances. Security violations of the system can be categorized as intentional or accidential. It is easier to protect against accidential misuse than against intentional misuse. Types of security violation involve unauthorized reading of data (breach of confidentiality), unauthorized modification of data (breach of integrity) and unauthorized destruction of data (breach of availability).

The most common approach to authenticating a user identity is the use of password. If the user supplied password matches the password stored in the system, the system assumes that the account is being accessed by the owner of that account. To avoid problems of password sniffing, the system could use a set of paired passwords. User has a list of password and sequence usage is done. One password has a one-time usage. Challenge response string (questionary method), encrypted password, authentication using physical objects and authentication using Biometrics also exist. Message confidentiality or privacy means the transmitted message must make sense to only the intended receiver. The message must be rendered unintelligible to unauthorized parties. A good privacy technique guarantees to some extend that a potential intruder cannot understand the contents of the message. Encryption and decryption provide secrecy or confidentiality. In other words confidentiality means only authorized users can read the data. Not only confidentiality, but anonymization is still required to provide privacy.

Anonymization is a technique for masking. That is personal information is removed from the original data set to provide privacy and to protect private information. Data anonymization enables transferring information between two organizations, by converting text data in to non-human readable form using encryption method. There have been lots of approaches developed. K-Anonymization is one of the approaches [1]. In K-Anonymization approach, at least K-tuples should be indistinguishable by masking values. 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 privately check whether a K-anonymous database retains its anonymity once a new tuple 'T' is being inserted in to it[3].

Two approaches can be used for anonymization. One is Suppression and the other is Generalization [5]. To assure a higher level of anonymity to the party inserting the data, we require that the communication between this party and the database occurs through an anonymous connection, as provided by protocols like Crowds or Onion routing. Crowds increase the privacy of web transaction; the main idea is "blending in to the crowd". That is, hiding one's action with in 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 [9]. 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 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 [8]. 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.

a. 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).

Table 1: Original Dataset

Area	***Position***	***Salary***
Data mining	Associate professor	$90,000
Intrusion detection	Assistant professor	$78,000
Handheld system	Research assistant	$17,000
Handheld system	Research assistant	$15,500
Query processing	Associate professor	$100,000
Digital forensics	Assistant professor	$78,000

Table 2: Suppressed Dataset

Area	***Position***	***Salary***
*	Associate professor	*
*	Assistant professor	*
Handheld system	Research assistant	*
Handheld system	Research assistant	*
*	Associate professor	*
*	Assistant professor	*

Table 3: Generalized Dataset

Area	***Position***	***Salary***
Data mining	Associate professor	[90k-120k]
Intrusion detection	Assistant professor	[70k-90k]
Handheld system	Research assistant	[14k-20k]
Handheld system	Research assistant	[14k-20k]
Query processing	Associate professor	[90k-120k]
Digital forensics	Assistant professor	[70k-90k]

It is today well understood that databases represent an important asset for many applications and thus their security is crucial. Data confidentiality is particularly relevant because of the value, often not only monetary, that data have. The main aim is to increase the privacy in databases and to limit the access for protecting private data from intruders and hackers. The main problems addressing in this work are K-Anonymization techniques in databases for confidentiality and privacy and an approach to handle untrusted malicious parties in anonymized databases [4]. The work is organised into four sections as Existing system, proposed method, conclusion and references.

In the Existing system section, it involves the substantial finding and current knowledge related to the work as well as the theoretical and methodological contributions towards certain topics. The proposed method, gives the details about the system designed to overcome the existing hindrance. The conclusion section gives the summary about the basic understanding from the Existing system and also the work to be done to build an effective system. In references section, it list out all the researches papers and experiments conducted related to the work.

II. EXISTING SYSTEM

A number of methods were proposed to provide confidentiality and privacy to anonymous database. The first approach considers the problem of providing security to statistical databases against disclosure of confidential information. Security-control methods suggested in the literature are classified into four general approaches: conceptual, query restriction, data perturbation, and output perturbation. Criteria for evaluating the performance of the various security-control methods are identified [6]. Security-control methods that are based on each of the four approaches are discussed, together with their performance with respect to the identified evaluation criteria. To date no single security-control method prevents both exact and partial disclosures. There are, however, a few perturbation-based methods that prevent exact disclosure and enable the database administrator to exercise "statistical disclosure control."

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. SMC represents an important class of techniques widely investigated in the area of cryptography. However, these techniques generally are not efficient [2]. 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.

III. PROPOSED SYSTEM

The proposed system contains three databases. One is patient database which contains patient's medical details.

The second database is anonymous database, which contains anonymous data. That is data from patients database is anonymized and stored in to the anonymous database. The third database is Research database. The research database contains the researching details and the researcher's data. Researchers can use anonymous database for researching purpose.

The main modules are

a. Patients database management
b. Doctors details registration and login
c. Anonymous data management and updation by Admin
d. Anonymous and secure data transfer
e. Anonymous database allocation to research people
f. Generalization based user search
g. Medical search

The first step what patient and doctor to do is to register the personal details in the medical database and get the authentication processes to go forward. The patient and doctor want to give the database to medical admin all the registration process is done by a medical admin. After the registration process completed patient and doctor can get the authentication code and machine generated id, by using this only the patient and doctor can login to the medical.

The research people can see the data's send by medical. And allocate research peoples to each research data. And forward the data to research people. Here the research people can't do any changes or modifications in patient database; they only can use the database for reference purpose. Any user can get relevant information about the disease from the web which is newly updated by the research people and medical can see the research people newly updated data's from research database.

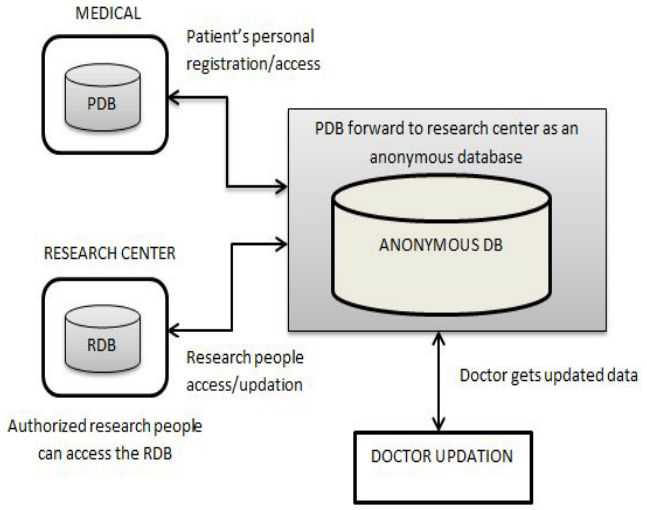

Figure 1: Proposed system architecture

A. Cryptographic primitive:

Our system use encryption algorithm RSA (Rivest, Shamir, and 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 [10]

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

a) Commutativity: For all key pairs k1,k2 □ K and value d □ D ,then

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

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

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

c) 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

B. RSA Algorithm:

The RSA scheme is a block cipher. Each plaintext block is an integer between 0 and n− 1 for some n, which leads to a block size≤ log2 (n). The typical size for n is 1024 bits. The details of the RSA algorithm are described as follows.

a. Key generation:

a) Pick two large prime numbers p and q, p 6= q;
b) Calculate n = p × q;
c) Calculate - (n) = (p − 1) (q − 1);
d) Pick e, so that gcd (e, - (n)) = 1, 1 < e < - (n);
e) Calculate d, so that d · e mod - (n) = 1, i.e., d is the multiplicative inverse of e in mod - (n);
f) Get public key as KU = {e, n};
g) Get private key as KR = {d, n}.

b. Encryption:

For plaintext block P < n, its ciphertext C = Pe mod n.

c. Decryption:

For cipher text block C, its plaintext is P = Cd mod n.

As we have seen from the RSA design, RSA algorithm uses modular exponentiation operation. For n = p · q, e which is relatively prime to - (n), has exponential inverse in mod n. Its exponential inverse d can be calculated as the multiplicative inverse of e in mod - (n). The reason is illustrated as follows.

Based on Euler's theorem, for y which satisfies y mod - (n) = 1, the following equation holds.

(xy) mod n = x mod n (12)

As d · e mod - (n) = 1, we have that Ped≡ P mod n. So the correctness of RSA cryptosystem is shown as follows.

d. Encryption: C = Pe mod n;

e. Decryption: P = Cd mod n = (Pe) d mod n = Ped mod n = P mod n = P.

The premise behind RSA's security is the assumption that factoring a big number (n into p, and q) is hard. And thus it is difficult to determine - (n). Without the knowledge of - (n), it would be hard to derive d based on the knowledge of e. Most of privacy models developed is based

on k-anonymity property. Anonymity property deals the possibility of indirect identification of records form public databases .The idea of protecting databases through data suppression or data perturbation has been extensively investigated in the area of statistical database.

Activity diagram are a loosely defined diagram to show workflows of stepwise activities and actions, with support for choice, iteration and concurrency. UML, activity diagrams can be used to describe the business and operational step-by-step workflows of components in a system. UML activity diagrams could potentially model the internal logic of a complex operation. In many ways UML activity diagrams are the object-oriented equivalent of flow charts and data flow diagrams (DFDs) from structural development.

The following Activity diagram shows how the anonymous database management system working flow. It shows in which order anonymous database updating and retrieval done by a medical environment.

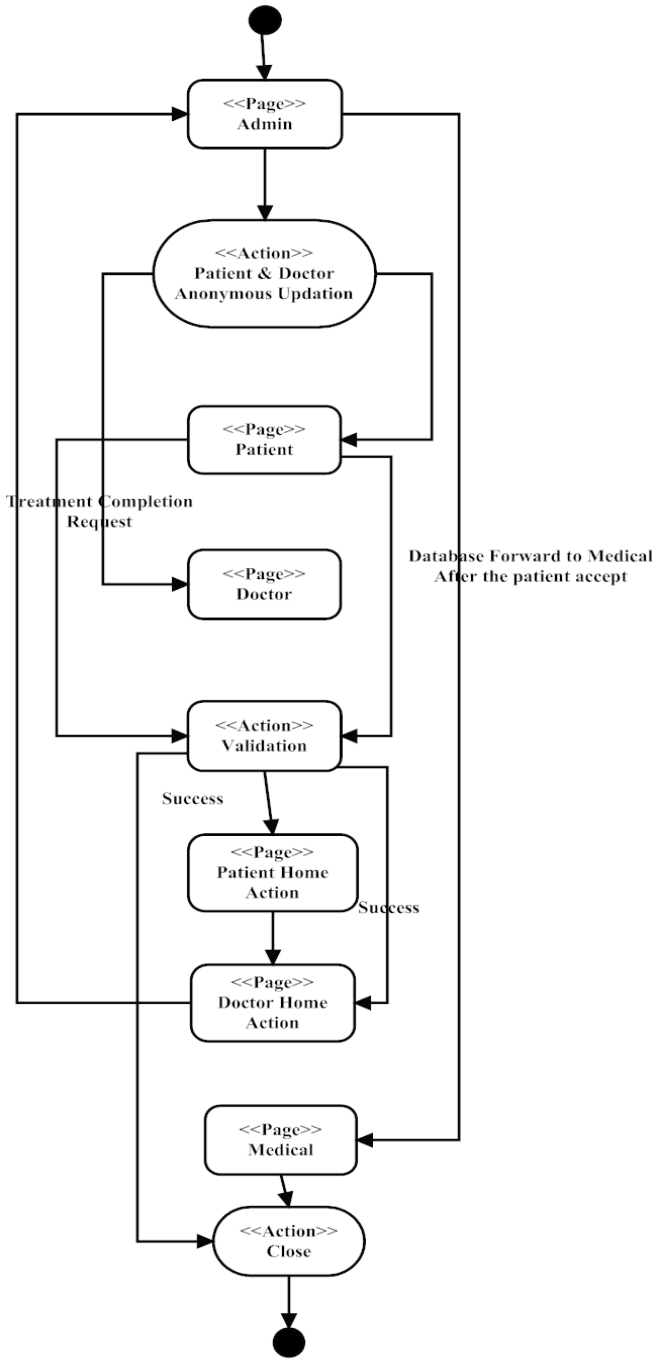

Figure 2: Activity diagram

C. Using the Chinese remainder algorithm:

For efficiency many popular crypto libraries (like Open SSL, Java and .NET) use the following optimization for decryption and signing: The following values are pre computed and stored as part of the private key:

a. p and q: the primes from the key generation,
b. $d_P = d \pmod{p-1}$,
c. $d_Q = d \pmod{q-1}$ and
d. $q_{Inv} = q^{-1} \pmod{p}$.

These values allow the recipient to compute the exponentiation $m = c^d \pmod{pq}$ more efficiently as follows:

a) $h = q_{Inv} * (m_1 - m_2) \pmod{p}$ (if $m_1 < m_2$ then some libraries compute h as $q_{Inv} * (m_1 + p - m_2) \pmod{p}$)

This is more efficient than computing $m = c^d \pmod{pq}$ even though two modular exponentiations have to be computed. The reason is that these two modular exponentiations both use a smaller exponent and a smaller modulus.

IV. CONCLUSION

In this paper, we have proposed to implement how a K-Anonymous database prevents a malicious party, by the introduction of an untrusted noncolluding third party without 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 system. 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-anonymous architecture makes unauthorized user too difficult to identify the record.

V. FUTURE WORKS

The important issues in future will be resolved:

a. Implement real world database system.
b. Solve problem of anonymity when initially table is empty.
c. Improving efficiency of protocol in terms of number of messages exchanged between user and database.
d. Implement database for invalid entries.

VI. 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 CryptographyConf. (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 TOR In: IEEE symposium on security and privacy May 2005.

[7]. Brier, S. 1997. How to keep your privacy: Battle lines get clearer. The New York Times, January 13, 1997.

[8]. 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}.

[9]. 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.

[10]. 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.