Privacy Maintenance to Anonymous and Restricted Databases Using Different method

Anup R.Maurya, Vaibhav C.Badbe, Shraddha Shetye

***Abstract*— Now a days there is an expanded concern for security in databases. So for that data privacy and confidentiality is a major issue. Suppose a person owns a k-anonymous database and needs to determine whether her database, when inserted with a tuple owned by another person, is still k- anonymous. Also, suppose that access to the database is strictly controlled, because for example data are used for certain experiments that need to be kept confidential. Allowing one person to directly read the contents of the tuple breaks the privacy of another person (e.g., a patient's medical record); Thus, the problem is to check whether the database inserted with the tuple is still k-anonymous, without letting them know the contents of the tuple and the database, respectively. In this paper, we propose two protocols solving this problem on suppression-based and generalization-based k-anonymous and confidential databases. The protocols rely on well known cryptographic assumptions. This paper suggests advancing the existing database systems and increasing the security and efficiency of the systems. This paper proposes a new concept to implement a real world anonymous database which improves the secure efficient system for protection of data, restricting the access to data even by the administrator thus maintaining the secrecy of individual patients. This paper considers the problem of providing security to statistical databases against disclosure of confidential information.**

***Index Terms*—Security database, tuple, anonymous, security.**

I. INTRODUCTION

Government agencies and other organizations often need to publish microdata, e.g., medical data or census data, for research and other purposes. Typically, such data are stored in a table, and each record (row) corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories: 1) Attributes that clearly identify individuals. These are known as explicit identifiers and include, e.g., Social Security Number. 2) Attributes whose values when taken together can potentially identify an individual. These are known as quasi-identifiers, and may include, e.g., Zip code, Birth-date, and Gender. 3) Attributes that are considered sensitive, such as Disease and Salary.

Today there is an increased concern for privacy. The availability of huge numbers of databases recording a large variety of information about individuals makes it possible to discover information about specific individuals by simply correlating all the available databases. Although confidentiality and privacy are often used as synonyms, they are different concepts: data confidentiality is about the difficulty (or impossibility) by an unauthorized user to learn anything about data stored in the database. Usually, confidentiality is achieved by enforcing an access policy, or possibly by using some cryptographic tools. Privacy relates to what data can be safely disclosed without leaking sensitive information regarding the legitimate owner

When releasing microdata, it is necessary to prevent the sensitive information of the individuals from being disclosed. Two types of information disclosure have been identified in the literature: identity disclosure and attribute disclosure. Identity disclosure occurs when an individual is linked to a particular record in the released table. Attribute disclosure occurs when new information about some individuals is revealed, i.e., the released data make it possible to infer the characteristics of an individual more accurately than it would be possible before the data release. Identity disclosure often leads to attribute disclosure. Once there is identity disclosure, an individual is identified and the corresponding sensitive values are revealed. Attribute disclosure can occur with or without identity disclosure. It has been recognized that even disclosure of false attribute information may cause harm. An observer of a released table may incorrectly perceive that an individual's sensitive attribute takes a particular value and behaves accordingly based on the perception. This can harm the individual, even if the perception is incorrect.

While the released table gives useful information to researchers, it presents disclosure risk to the individuals whose data are in the table. Therefore, our objective is to limit the disclosure risk to an acceptable level while maximizing the benefit. This is achieved by anonymizing the data before release. The first step of anonymization is to remove explicit identifiers. However, this is not enough, as an adversary may already know the quasi-identifier values of some individuals in the table. This knowledge can be either from personal knowledge (e.g., knowing a particular individual in person), or from other publicly available databases (e.g., a voter registration list) that include both explicit identifiers and quasi-identifiers. A common anonymization approach is generalization, which replaces quasiidentifier values with values that are less-specific but semantically consistent. As a result, more records will have the same set of quasi-identifier values. We define an equivalence class of an anonymized table to be a set of records that have the same values for the quasi-identifiers.

if one asks whether confidentiality is still required once data have been anonymized, the reply is yes if the anonymous data have a business value for the party owning them or the unauthorized disclosure of such anonymous data may damage the party owning the data or other parties. To better understand the difference between confidentiality and anonymity, consider the case of a medical facility connected with a research institution. Suppose that all patients treated at the facility are asked before leaving the facility to donate their personal health care records and medical histories (under the condition that each patient's privacy is protected) to the

Manuscript received June 18, 2014.
Anup R.Maurya, PG Student (Computer)
Vaibhav C.Badbe, PG Student (Computer)
Shraddha Shetye, PG Student (Computer)

research institution, which collects the records in a research database. To guarantee the maximum privacy to each patient, the medical facility only sends to the research database an anonymized version of the patient record.

Once this anonymized record is stored in the research database, the non-anonymized version of the record is removed from the system of the medical facility. Thus, the research database used by the researchers is anonymous. Suppose that certain data concerning patients are related to the use of a drug over a period of four years and certain side effects have been observed. and recorded by the researchers in the research database. It is clear that these data (even if anonymized) need to be kept confidential and accessible only to the few researchers of the institution working on this project, until further evidence is found about the drug. If these anonymous data were to be disclosed, privacy of the patients would not be at risk; however the company manufacturing the drug may be adversely affected. Recently, techniques addressing the problem of privacy via data anonymization have been developed, thus making it more difficult to link sensitive information to specific individuals.

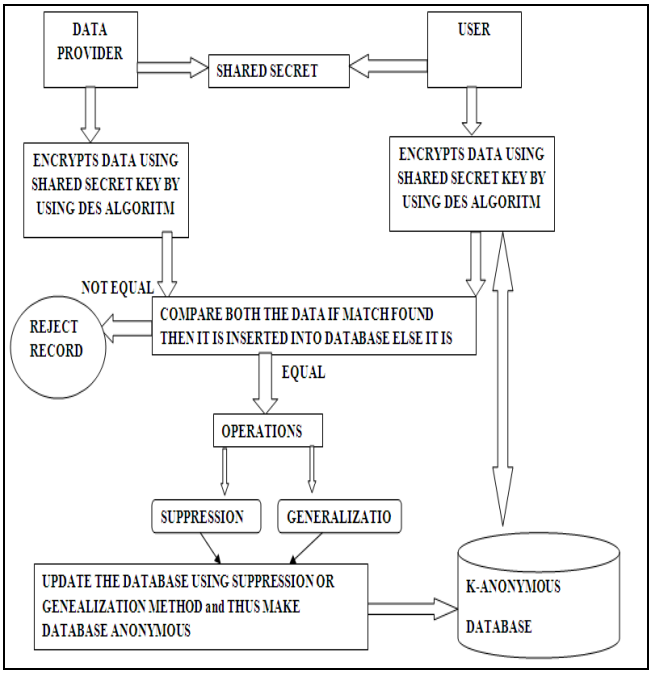

Fig. 1. Anonymous Database System.

A. *EXISTING METHODOLOGIES*

Number of approach has been proposed, these protocols have some serious limitations, in that they do not support generalization-based updates, which is the main strategy adopted for data anonymization Therefore, if the database is not anonymous with respect to a tuple to be inserted, the insertion cannot be performed. In addition one of the protocols is extremely inefficient.

The first research direction deals with algorithms for database anonymization. The idea of protecting databases through data suppression or data perturbation has been extensively investigated in the area of statistical databases. Aggarwal proposed the notion of k-anonymity for databases in the context of medical data. The problem of computing a k-anonymization of a set of tuples while maintaining the confidentiality of their content is addressed by Zhong et al. However, these proposals do not deal with the problem of private updates to k-anonymous databases. The problem of protecting the privacy of time varying data have recently spurred an intense research activity which can be roughly divided into two broad groups depending on whether data are continuously released in a stream and anonymized in an online fashion, or data are produced in different releases and subsequently anonymized in order to prevent correlations among different releases.

The second research direction is related to Secure Multiparty Computation (SMC) techniques. SMC represents an important class of techniques widely investigated in the area of cryptography. General techniques for performing secure computations are today available. However, these techniques generally are not efficient. Such shortcoming has motivated further research in order to devise more efficient protocols for particular problems.

The third research direction is related to the area of private information retrieval, which can be seen as an application of the secure multiparty computation techniques to the area of data management. Here, the focus is to devise efficient techniques for posing expressive queries over a database without letting the database know the actual queries the fourth research direction is related to query processing techniques for encrypted data. These 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.

B. *PROPOSED METHODOLOGIES*

In My paper, i have presented two secure protocols for privately checking whether a k-anonymous database retains its anonymity once a new tuple is being inserted to it.

The first protocol is aimed at suppression-based anonymous databases, and it allows the owner of DB to properly anonymize the tuple t, without gaining any useful knowledge on its contents and without having to send to t's owner newly generated data. To achieve such goal, the parties secure their messages by encrypting them. In order to perform the privacy-preserving verification of the database anonymity upon the insertion, the parties use a commutative and homomorphic encryption scheme. The second protocol is aimed at generalization-based anonymous databases, and it relies on a secure set intersection protocol, to support privacy-preserving updates on a generalization-based k-anonymous DB.

In particular, when using a suppression-based anonymization method, we mask with the special value *, the value deployed by Alice for the anonymization. When using a generalization-based anonymization method, original values are replaced by more general ones, according to a priori established value generalization hierarchies (VGHs).

II. DEVELOPMENT METHODS

A. *SUPPRESSION-BASED ANONYMOUS AND CONFIDENTIAL DATABASES*

- Where sensitive information and all informationthat allows the inference of sensitive information are simply not released.

- To achieve such goal, the parties secure their messages by encrypting them.
- To perform the privacy-preserving verification, the parties use cryptographic schemes.

In this suppression based anonymous method the original values are masked with the special value *.

B. A. I. ALGORITHM FOR SUPPRESSION METHOD

Consider Table T = {t1… tn} over the attribute set A. The idea of this algorithm is mask some attributes by special value *, the value employed by User A for the anonymization.In suppression based method, every attribute is suppressed by *.So third party cannot differentiate between any tuples. Here, k-anonymity indicates that is each row in the table cannot be distinguished from at least other k-1 rows by only looking a set of attributes. We assume that the database is anonymized using suppression based method.
The protocol works as follows:
Step1: User A sends User B an encrypted version containing only the s non-suppressed attributes.
Step2: User B encrypts the information received from User A and sends it to her, along with encrypted version of each value in his tuple t.
Steps3: User A examines if the non suppressed QI attributes is equal to those of t. If true, t can be inserted to table T. Otherwise, when inserted to T, t breaks k- anonymity.
In suppression algorithm t stands for private tuple provided by Data provider, T stands for Anonymous database, QI stands for Quasi-Identifier which consist of set of attributes that can be used with certain external information to identify a specific individual.

C. GENERALIZED–BASED ANONYMOUS AND CONFIDENTIAL DATABASES

It will rely on a secure set intersection protocol, to support privacy-preserving updates on a k-anonymous database.

- It may consist of following steps:
  - Random functions
  - GetSpec Function
  - SSI a secure protocol that computes the cardinality i.e.
- Private Matching and Set Intersection

which includes,

  - Encryption
  - Share public key
  - Private Matching and Set Intersection

- Intersection protocol which follows,
  - Hash function
  - Choose secrete key
  - Lexicographical order

For generalization-based anonymization, we assume that each attribute value can be mapped to a more general value. The main step in most generalization based k-anonymity protocols is to replace a specific value with a more general value.
WORKING OF GENERALIZED METHOD:

1. User A randomly chooses a $\delta \varepsilon T_w$ (Witness Set).
2. User A computes γ = GetSpec(δ) (bottom values of VGH(Value Generation Hierarchy)).
3. User A and User B collaboratively compute s= SSI(γ,τ) (cardinality of $\gamma \cap \tau$).
4. If s=u then t's generalized form can be safely inserted to T.
5. Otherwise, Alice computes $T_w \leftarrow T_w - \{\delta\}$ and repeat the above procedures until either s=u or $T_w = \varnothing$;

III. RESULT

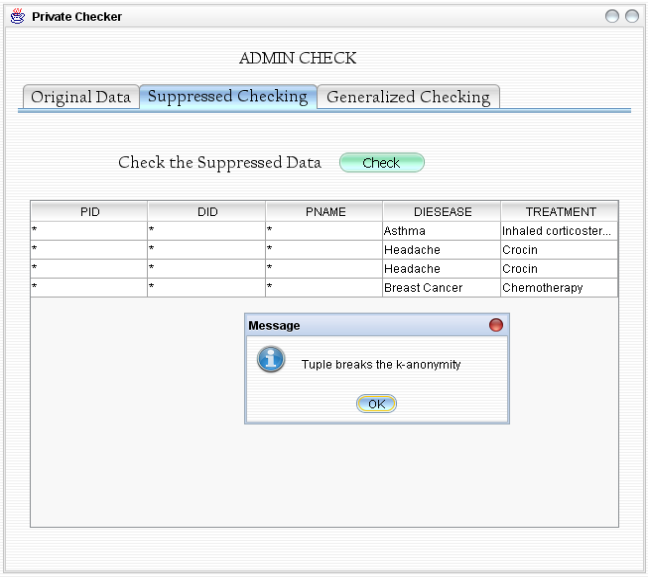

Fig 2. User View Suppressed Data

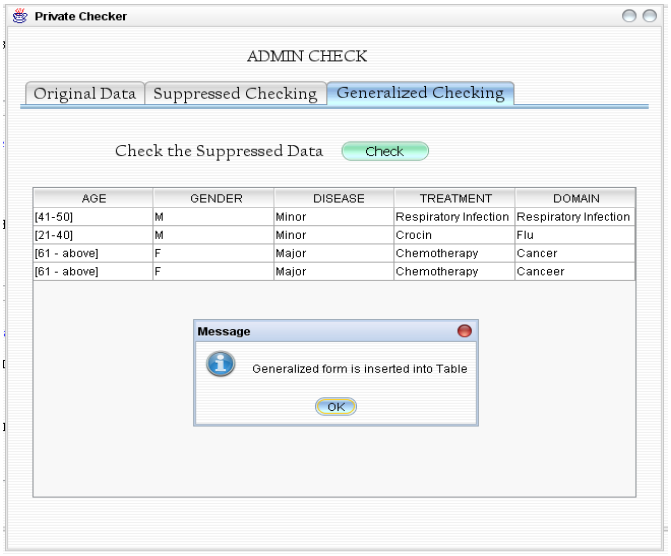

Fig 3. User view generalized Data

IV. CONCLUSIONS

In my paper, I have presented two secure protocols 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 results returned from a user's (or a medical researcher's) query are also k-anonymous. Thus, the patient or the data provider's privacy cannot be violated from any query. As long as the database is updated properly using the

proposed protocols, the user queries under our application domain are always privacy-preserving. In case of suppression based k-anonymity approach, suppressed the sensitive information attribute by * to maintain the k-anonymity in database.

In case of generalization based k-anonymity approach, specific or original values are replaced by more general values so that attacker cannot identify correct values. This is particularly applicable in military application or health care system.

The important issues for future work are as follows:

- In the case of malicious parties by the introduction of an untrusted third party, implementing a real-world anonymous database system.
- Improve the efficiency of protocols, by reducing the number of messages exchanged and sizes.
- How to initially populate an empty table.
- The integration with a privacy-preserving query system.

REFERENCES

[1] N.R. Adam and J.C. Wortmann, "Security-Control Methods for Statistical Databases: A Comparative Study," ACM Computing Surveys, vol. 21, no. 4, pp. 515-556, 1989.

[2] 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.

[3] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing across Private Databases," Proc. ACM SIGMOD Int'l Conf. Management of Data, 2003

[4] C. Blake and C. Merz, "UCI Repository of Machine Learning Databases," http://www.ics.uci.edu/mlearn/MLRepository. html, 1998.

[5] E. Bertino and R. Sandhu, "Database Security—Concepts, Approaches and Challenges," IEEE Trans. Dependable and Secure Computing, vol. 2, no. 1, pp. 2-19, Jan.-Mar. 2005.

[6] D. Boneh, "The Decision Diffie-Hellman Problem," Proc. Int'l Algorithmic Number Theory Symp., pp. 48-63, 1998.

[7] D. Boneh, G. di Crescenzo, R. Ostrowsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Eurocrypt Conf., 2004.

[8] S. Brands, "Untraceable Offline Cash in Wallets with Observers," Proc. CRYPTO Int'l Conf., pp. 302-318, 1994.

[9] J.W. Byun, T. Li, E. Bertino, N. Li, and Y. Sohn, "Privacy- Preserving Incremental Data Dissemination," J. Computer Security, vol. 17, no. 1, pp. 43-68, 2009.

[10] R. Canetti, Y. Ishai, R. Kumar, M.K. Reiter, R. Rubinfeld, and R.N. Wright, "Selective Private Function Evaluation with Application to Private Statistics," Proc. ACM Symp. Principles of Distributed Computing (PODC), 2001

[11] S. Chawla, C. Dwork, F. McSherry, A. Smith, and H. Wee, "Towards Privacy in Public Databases," Proc. Theory of Cryptography Conf. (TCC), 2005.

[12] U. Feige, J. Kilian, and M. Naor, "A Minimal Model for Secure Computation," Proc. ACM Symp. Theory of Computing (STOC), 1994.

[13] M.J. Freedman, M. Naor, and B. Pinkas, "Efficient Private Matching and Set Intersection," Proc. Eurocrypt Conf., 2004

[14] B.C.M. Fung, K. Wang, A.W.C. Fu, and J. Pei, "Anonymity for Continuous Data Publishing," Proc. Extending Database Technology Conf. (EDBT), 2008.

[15] O. Goldreich, Foundations of Cryptography: Basic Tools, vol. 1. Cambridge Univ. Press, 2001

[16] 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.

[17] S. Zhong, Z. Yang, and R.N. Wright, "Privacy-Enhancing k-Anonymization of Customer Data," Proc. ACM Symp. Principles of Database Systems (PODS), 2005.

[18] J.W. Byun, T. Li, E. Bertino, N. Li, and Y. Sohn, "Privacy-Preserving Incremental Data Dissemination," J. Computer Security,vol. 17, no. 1, pp. 43-68, 2009.

[19] O. Goldreich, Foundations of Cryptography: Basic Applications, vol. 2.Cambridge Univ. Press, 2004.

[20] R. Canetti, Y. Ishai, R. Kumar, M.K. Reiter, R. Rubinfeld, and R.N.Wright, "Selective Private Function Evaluation with Application to Private Statistics," Proc. ACM Symp. Principles of Distributed Computing (PODC), 2001.

[21] U. Maurer, "The Role of Cryptography in Database Security,"Proc. ACM SIGMOD Int'l Conf. Management of Data, 2004.

[22] H. Hacigu¨mu¨ s¸, B. Iyer, C. Li, and S. Mehrotra, "Executing SQL over Encrypted Data in the Database-Service-Provider Model,"Proc. ACM SIGMOD Int'l Conf. Management of Data, 2002.

[23] D.X. Song, D. Wagner, and A. Perrig, "Practical Techniques for Searches on Encrypted Data," Proc. IEEE Symp. Security and Privacy, 2000.

[24] M.K. Reiter and A. Rubin, "Crowds: Anonymity with Web Transactions," ACM Trans. Information and System Security (TISSEC), vol. 1, no. 1, pp. 66-92, 1998

[25] R. Agrawal, A. Evfimievski, and R. Srikant, "Information Sharing across Private databases," Proc. ACM SIGMOD Int'l Conf.Management of Data, 2003.

[26] K. Wang and B. Fung, "Anonymizing Sequential Releases," Proc. ACM Knowledge Discovery and Data Mining Conf. (KDD), 2006.

[27] Y. Han, J. Pei, B. Jiang, Y. Tao, and Y. Jia, "Continuous Privacy Preserving Publishing of Data Streams," Proc. Extending Database Technology Conf. (EDBT), 2008.

[28] D. Boneh, G. di Crescenzo, R. Ostrowsky, and G. Persiano, "Public Key Encryption with Keyword Search," Proc. Euro crypt Conf., 2004.

[29] M. Reed, P. Syverson, and D. Goldschlag, "Anonymous Connections and Onion Routing," IEEE J. Selected Areas in Comm., vol. 16, no. 4, pp. 482-494, May 1998.

[30] 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.