# Survey On Privacy Preserving Updates On Anonymous Database

Rajeshwari Suryawanshi[1]
[1]*Department of Computer Science & Engineering, Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur,INDIA*

Prof.Parul Bhanarkar[2]
[2]*Department of Computer Science &Engineering, Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur,INDIA*

Prof.Girish Agrawal[3]
[3]*Department of Computer Science &Engineering, Rashtrasant Tukdoji Maharaj Nagpur University, Nagpur,INDIA*

**Abstract**

*Privacy is main concern in the present technological phase in the world. Information security has become a critical issue since the information sharing has a common need. Thus privacy is becoming an increasingly important issue in many data mining applications in various fields like medical research, intelligence agencies, hospital records maintenance etc. The paper focuses on survey on privacy preserving on anonymous database and on devising private update techniques to database systems that supports notions of anonymity different than k-anonymity. The existing methods provides the same amount of privacy for all persons, and may be offering insufficient protection to a subset of people, while applying excessive privacy control to another subset. Motivated by these the concept of personalized anonymity is used which performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the microdata. To preserve privacy and confidentiality with minimum loss of information an approach on generalization based on personalized anonymity method to protect privacy of individual is proposed.*

## 1.Introduction

In today's world databases represent need for increases security. Data in the databases has its own relevant value. For example; medical data collected by over the history of patients over years is an invaluable asset, which needs to be secured and can be used by people in various related areas of work. [1] Nowadays, privacy accidents have become common problem in the information systems. For example, a hospital may have record of all the patients with various diseases critical and non-critical. If the hospital wishes to reveal the data to any pharmaceutical company or online market services, it should not be able to infer with particularity of patients with those diseases. It can give as a statistical view or just the superficial information such that privacy is not detained.

There are huge numbers of databases that hold numerous confidential information such that people access those data correlating various information from various databases. For example, assume that the hospital publishes the table, which does not explicitly indicate the names of patients. However, if an adversary has access to the voter registration list in table b, s/he can easily discover the identities of all patients by joining the two tables on {Age, Sex, Zipcode}. These three attributes are, therefore, the quasi-identifier (QI) attributes. The 2 anonymous tables for microdata are shown in table c of Fig.1.

The Personalization is an inherent notion of privacy preservation whose objective is to protect the interests of individuals at the first place. A well-known technique personalized anonymity, i.e., a person can specify the degree of privacy protection for her/his sensitive values. The operation of updating of an anonymous database e.g., by inserting a tuple containing information about a given individual, introduces two problems concerning both the anonymity and confidentiality of the data stored in the database and the privacy of the individual to whom the data to be inserted are related [1].The personalized privacy of individuals to whom data are referred is not only of interest to these individuals, but also to the organization owning the database.

### 1.1 Motivation

Generalization is a common approach to avoid the above problem, by transforming the QI values into less specific forms so that they no longer uniquely represent individuals. In particular, a table is k-anonymous if the QI values of each tuple are identical to those of at least k −1 other tuples. But k-anonymity has several

drawbacks as discussed in [2].A k-anonymous table may lose considerable information from the microdata and may allow an adversary to derive the sensitive information of an individual with 100% confidence.

Consider the tables in Fig 1.The Microdata for medical facility is given in Figure 1(a) and the other database for voter registration list is given in Fig 1(b). Assume that an adversary attempts to infer the disease of Samantha, knowing his age 18, sex, and zipcode 24000. From the published table in Fig1(c), s/he knows that Nick may correspond to tuple 5 or 6 (the QI values of the other tuples do not cover those of Nick). The diseases of both tuples are pneumonia; hence, the adversary can declare (with 100% confidence) that Samantha must have contracted pneumonia. Again it does not take into account personal anonymity requirements. A k-anonymity only prevents association between individuals and tuples, instead of association between individuals and sensitive values. Unfortunately, it is the second type of association that leads to privacy breach.

The paper proposes an update technique on personalized anonymous database. The Existing method based on K-anonymization exerts the same amount of preservation for all persons. The proposed system inserts an tuple concerning information about a person in personalized anonymous database and checks whether the database is still anonymous.

## 2. Literature Survey

In the paper [1] the author suggested paper deals with problems concerning that the users without revealing the contents of tuples and DB, how to preserve data integrity by establishing the anonymity of DB and if the anonymity is authorized then there is a concern of updating the data. It deals with algorithms for database anonymization. This paper shows how privacy is maintained without disclosing the contents of whole databases and their owner and individual tuples and its owner to each other. The problem is to check whether the database connecting the tuple is still k-anonymous, such that no one can view the actual data from, tuples or database. It exerts the same amount of preservation for all persons, without catering for their concrete needs. 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. 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.

| row # | Age | Sex | Zipcode | Disease | Guarding node |
|---|---|---|---|---|---|
| 1(John) | 4 | M | 12000 | gastric ulcer | Stomach disease |
| 2( Jill) | 8 | M | 14000 | Dyspepsia | Dyspepsia |
| 3(Ben) | 6 | M | 18000 | Pneumonia | Respiratory infection |
| 4( Nick) | 7 | M | 19000 | Bronchitis | Bronchitis |
| 5( Joel) | 13 | M | 22000 | Pneumonia | Pneumonia |
| 6(Samantha) | 18 | M | 24000 | Pneumonia | Pneumonia |
| 7( Lisa) | 24 | F | 58000 | Flu | Φ |
| 8(Jamie) | 27 | F | 36000 | Gastritis | Gastritis |
| 9(Sara) | 29 | F | 37000 | Pneumonia | Respiratory infection |
| 10( Margarette) | 55 | F | 33000 | Flu | Flu |

**(a) Microdata**

| Name | Age | Sex | Zipcode |
|---|---|---|---|
| John | 4 | M | 12000 |
| Jill | 8 | M | 14000 |
| Ben | 6 | M | 18000 |
| Nick | 7 | M | 19000 |
| Joel | 13 | M | 22000 |
| Samantha | 18 | M | 24000 |
| Lisa | 24 | F | 58000 |
| Jamie | 27 | F | 36000 |
| Sara | 29 | F | 37000 |
| Margarette | 55 | F | 33000 |

**(b)Voter Registration List**

| row # | Age | Sex | Zipcode | Disease |
|---|---|---|---|---|
| 1 | [1,10] | M | [10001,15000] | gastric ulcer |
| 2 | [1,10] | M | [10001,15000] | Dyspepsia |
| 3 | [1,10] | M | [15001,20000] | Pneumonia |
| 4 | [1,10] | M | [15001,20000] | Bronchitis |
| 5 | [11,20] | M | [20001,25001] | Pneumonia |
| 6 | [11,20] | M | [20001,25001] | Pneumonia |
| 7 | [21,60] | F | [30000-60000] | Flu |
| 8 | [21,60] | F | [30000-60000] | Gastritis |
| 9 | [21,60] | F | [30000-60000] | Pneumonia |
| 10 | [21,60] | F | [30000-60000] | Flu |

**(c) A 2-Anonymous Table**

Figure 1: Microdata, external source, and quasi-identifier generalization

In the paper [3] author proposed a formal protection model named *k*-anonymity and a set of accompanying policies for deployment. A release provides *k*-anonymity protection if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals whose information also appears in the release. This paper also examines re-identification attacks that can be realized on releases that adhere to *k*-anonymity unless accompanying policies are respected. The *k*-anonymity protection model is important because it forms the basis on which the real-world systems known as Data fly, Argus and *k*-Similar provide guarantees of privacy protection.

In the paper [4] the author proposed technique that performs the minimum generalization for satisfying everybody's requirements, and thus, retains the largest amount of information from the micro data. It illustrates how the k-anonymity requirement can be translated, through the concept of quasi-identifiers, in terms of a property on the released table. The authors illustrated how k-anonymity can be enforced by using generalization and suppression techniques. They have introduced the concept of generalized table, minimal generalization, and minimal required suppression, capturing the property of a data release to enforce k-anonymity while generalizing and suppressing only what strictly necessary to satisfy the protection requirement.

In the paper [5] the author proposed the techniques which address the problems of efficiently and privately computing set intersection database oriented operations. It formalize the notion of minimal information sharing across In these paper the author proposed protocols for three operations INTERSECTION , INTERSECTION SIZE and EQUIJOIN and proved that these protocols disclose minimal information apart from query result..It then gives cost analysis for these protocols and estimation of execution times of the application examples. It has two limitations. It do not address the problem of what the parties might learn by combining the results of multiple queries and how to find which database contains which tables and what are the attributes names.

In the paper [6] they discuss the relationship between privacy preserving and SMC and problems involved. It reviews definitions and constructions for secure multiparty computation and discusses the issue of efficiency and demonstrates the difficulties involved in constructing highly efficient protocols.

In the paper [7] the author discusses the privacy enhancing method for creating K-anonymous tables for distributed scenarios. The objective is to design protocols that allow miner, who wants to mine the entire table to obtain K-anonymous table representing the customer data in such a way that does not reveal any extra information that can be used to link sensitive attributes to corresponding identifiers For these they proposed two different formulations. In t he first formulations, the protocols needs to extract k-anonymous part and the sensitive attributes outside the k-anonymous part should be hidden from any individual participant including the miner. In the second formulations, the table is K-anonymized by suppressing some entries of the quasi identifier attributes and these entries should be hidden from any individual participant.

In this paper [8] the proposed 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 ineifficient.

In this paper [9] the anonymization tables were introduced. The issue of releasing tables mainly in relational database consisting of confidential data is considered and how it can be resolved, ensuring personal privacy and also maintaining integrity . One of the techniques proposed in the literature is k anonymization. They showed that k anonymization problem is NP-hard and gave an O (k)-approximation algorithm for the general k-anonymity problem with arbitrary alphabet size.

## 3. Proposed Work

The existing methods [1] focus on a universal approach that exerts the same amount of preservation for all persons, without catering for their concrete needs. The consequence is that we may be offering insufficient protection to a subset of people, while applying excessive privacy control to another subset. There may exist other attributes that can be used, in combination with an external database, to recover the personal identities.

To provide a higher level of anonymity to the Data Provider inserting the data, we require that the communication between this party and the database

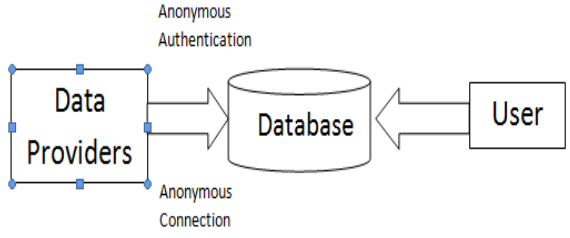


Figure 2. Anonymous Database

occurs through an anonymous connection as shown in Fig 2. Again, Sensitive information may be leaked from the access control policies adopted by the anonymous database System

The proposed system is a new generalization framework based on the concept of personalized anonymity, as k-anonymity has several drawbacks. A simple taxonomy on attribute Disease is accessible by the public, and organizes all diseases as leaves of a tree as shown in Fig.3. An intermediate node carries a name summarizing the diseases in its sub tree. Individual may specify node as the "guarding node" for his privacy, for sensitive attribute value. An individual may specify which implicit node of the taxonomy underneath all the leave is used. The empty-set preference implies that he is willing to release his actual diagnosis result for e.g. flu for Lisa in Fig 1; therefore it can be published directly.

Personalized privacy approach applies direct protection against the association between individuals and their sensitive values. This Paper proposes private updates techniques on a Database generalized using SA generalization algorithm [2] based on personalized concept that preserves a large amount of information in the microdata without violating any privacy constraint.

In SA-generalization,[2] generalization is performed 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.

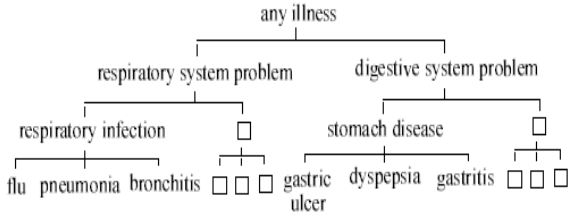


Figure 3: Taxonomy for disease

## 4. Conclusion

A Survey of privacy preserving for anonymous database and techniques to perform updates on the database is studied. The existing generalization methods using k-anonymity are inadequate because they cannot guarantee privacy protection in all cases, and often incur unnecessary information loss by performing excessive generalization. So the concept of Personalized Anonymity is becoming more important .In the paper, we propose work based on the concept of personalized anonymity, and updates will be performed on the personalized anonymous databases by using SA-generalization algorithm. So whenever a new tuple is inserted the individual will decide the level of privacy from taxonomy tree for sensitive attributes .Depending on that customized privacy requirement tuple will be inserted into table and checked whether the database is still personalized anonymous.

## References

[1] Alberto Trombetta ,Wei Jaing,Elisa Bertino and Lorenzo Bossi, "Privacy Preserving Updates to anonymous and Confidential database" IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 8, NO. 4, JULY/AUGUST 2011.

[2] Xiaokui Xiao, Yufei Tao "Personalized Privacy Preservation", *SIGMOD 2006*, June 27–29, 2006, Chicago, Illinois, USA. Copyright 2006 ACM 1595932569/ 06/0006

[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] P.Samarati, "Protecting Respondent's Privacy in Mirodata Release, "IEEE Trans.Knowledge and Data Eng.,vol. 13,no.6,pp. 1010-1027, Nov./Dec. 2001. W. and Marchionini, G. 1997.

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

[6] Yehuda Lindell and Benny Pinkasy,"Secure Multiparty Computation for Privacy-Preserving Data Mining" 2005

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

[8] Trombetta and E. Bertino, "Private Updates to Anonymous Databases," Proc. Int'l Conf. Data Eng. (ICDE), 2006.

[9] 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.