Personalized Privacy Preserving Updates to Anonymous Databases

Rajeshwari Suryawanshi
Rashtrasant Tukdoji Maharaj
Nagpur university
Abha Gaikwad Patil College of Engineering,Nagpur,India

Parul Bhanarkar
Rashtrasant Tukdoji Maharaj
Nagpur university
Abha Gaikwad Patil College of Engineering, Nagpur, India

Girish Agarwal
Rashtrasant Tukdoji Maharaj
Nagpur university
Abha Gaikwad Patil College of Engineering, Nagpur, India

ABSTRACT
Privacy of individual's information in datasets is main concern in the present technological phase. Thus it is becoming an increasingly important issue in many data mining applications in various fields like medical research, hospital records maintenance, intelligence agencies etc. Many previous works has focused on generalization and suppression based anonymity which provides same amount of privacy preservation to all individuals. The paper focuses on devising private update techniques to database systems that supports notions of anonymity different than k-anonymity. Therefore 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. Personalized Privacy is achieved by using SA (sensitive Attribute)-generalization to protect privacy of individual. In the paper, a method to perform updates on personalizes anonymity based database is proposed and its design view is explained.

Keywords
Anonymous database, Personalized Anonymity, K-Anonymity, Generalization, SA-generalization.

1. INTRODUCTION
It is today well understood that databases represent an important asset for many applications and thus their security is crucial. It is necessary to publish data for research purpose. Data in the databases has its own relevant value. Consider medical data collected by over years is an invaluable asset, which needs to be secured and can be used by people in various related areas of work. Data confidentiality is not only the issue that needs to be addressed. Nowadays, privacy accidents have become common problem in the information systems. If the hospital wishes to reveal the data to any pharmaceutical company or online market services, view of disease with its place and age detail can be provided. So it should not be able to infer with particularity of patients with those diseases.

Privacy and security of database is the main concern as there are huge numbers of databases that hold large amount of confidential information such that people access those data and correlate that released information with public records. 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, Zip code}. 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 objective of personalized privacy is to guard the interests of peoples at the primary place. An individual can specify the degree of privacy protection for her/his sensitive values by using the concept of personalized anonymity. When a database needs to be updated by inserting a tuple which contains information about an individual, introduces two problems concerning both the anonymity and confidentiality of the data stored in the database and the privacy of the individual whose data is to be inserted [6]. In general every individual is concerned about its privacy and also the organization to which the individual belongs to. So, the technique of personalized anonymity can be used to make the database anonymous with minimum loss of information.

The paper proposes an update technique on personalized anonymous database. The Existing method supports on K-anonymization that exerts the same amount of preservation for all persons. Section 2 gives the detailed literature survey. Section 3 explains the problem that need to be considered. Section 4 gives the proposed work with the design view of the system. The proposed system inserts a tuple concerning information about a person in personalized anonymous database and checks whether the database still satisfies personalized anonymity.

2. LITERATURE SURVEY
In the paper [1] author proposed a formal protection model named *k*-anonymity for privacy de-identification. It prevents the attack by suppressing and generalizing the Quasi-identifier attributes which can combine with public records and uniquely identify the records. A microdata release provides *k*-anonymity protection if the information for each person contained in the release cannot be distinguished from at least *k*-1 individuals in microdata. This paper also verifies re-identification attacks that can be realized on releases that adhere to *k*-anonymity .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 [2] author introduces a new notion of privacy called as l-diversity. As k-anonymity protects the microdata released table against identity disclosure, it is insufficient to provide attribute disclosure. L-diversity requires that each equivalence class of dataset should have at least l represented values for sensitive attribute. Its limitation is that it is possible for an adversary to gain information about the sensitive

attribute if the attacker has knowledge about global distribution of the attribute.

In the paper [3] author proposes novel privacy called t-closeness and showed that l-diversity has number of limitations as it is difficult to achieve and insufficient to prevent attribute disclosure. If the distance between the distribution of a sensitive attribute in an equivalence class and the distribution of the attribute in the whole table is less than or equal to threshold t then the equivalence class is having t-closeness. These highly limit the amount of individual specific knowledge an attacker can learn**.**

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 personalized anonymity concept which specify degree of privacy for his/her sensitive values. K-anonymity has several drawbacks as discussed in the paper [5]. 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 Joel, knowing his age 13, sex, and zip code 22000. 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 Joel must have contracted pneumonia. Again it does not take into account personal anonymity requirements. This paper also explains the algorithm for deriving generalized table. The two algorithms are the greedy framework and optimal SA-generalization.

In the paper [6] the author suggested paper deals with problems concerning privacy and confidentiality such that updates can be performed without revealing the contents of tuples and DB to the user or data provider. It deals with algorithms for database anonymization. The problem is to check whether the database connecting the tuple is still k-anonymous It exerts the same amount of preservation for all persons, resulting in more information loss in microdata release. The first protocol is aimed at suppression-based anonymous databases which allow the database owner to anonymize the tuple without gaining any information about the individual specific data and without sending new tuples owner newly generated data. The second protocol is aimed at generalization-based anonymous databases, and it works mainly on a secure set intersection protocol, to provide privacy-preserving updates on a generalization-based k-anonymous database.

In the paper [7] 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 results. 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.

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 [8] 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 this paper [9] the two protocols are proposed to perform private updates on anonymous database. But these protocols have limitations, of not supporting to generalization-based updates, which is the main strategy adopted for database anonymity. Therefore, if the database is not anonymous with respect to a tuple to be inserted, the insertion cannot be performed. One of the protocols proposed in the paper is not efficient.

3. PROBLEM STATEMENT

The existing methods [6] for privacy preserving updates focus on K-anonymity concept that exerts the same amount of preservation for all persons, without considering their personal requirements for privacy. The consequence is that we may be offering insufficient protection to a subset of people, while applying excessive privacy control to another subset. There exist other attributes that can be used, in combination with an external database, to recover the personal identities.

The Existing system [6] has K- anonymized Database DB by generalizing and suppressing the tuples before performing private updates. In addition, k-anonymity fails to guarantee safe publication, even in the scenario with no personal preferences. Assume that the information related to a patient is stored in a tuple t of Database and it is kept confidentially at the server as shown in fig.2. The insertion of information about new patient in the anonymous database DB can be performed if the updated database DB U t is still anonymous. Since Database contains privacy sensitive data, main concern is to protect the identity of patient. So the database is K-anonymized by performing Generalization and suppression. While inserting a tuple in anonymous database, the main concern is to protect the identity of patient. Therefore before inserting the tuple it is anonymized and then it is inserted in Anonymous database. But the existing method provides same amount privacy to all person which leads to unnecessary information loss.

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

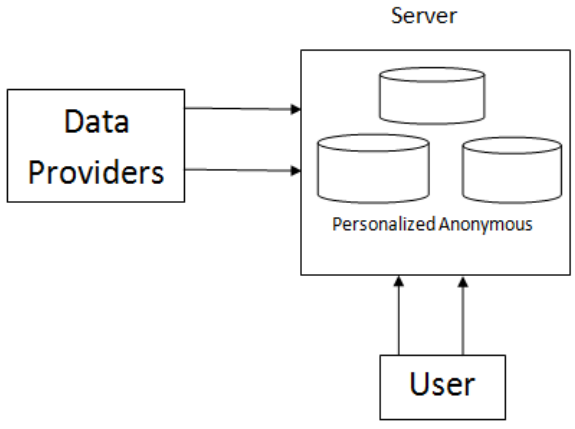

Figure 2: Anonymous Database

4. PROPOSED WORK

As k-anonymity has several drawbacks, the concept of personalized anonymity is used. 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 [5] .It organizes all diseases as leaves of a tree as shown in Figure 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 Figure 1; therefore it can be published directly.

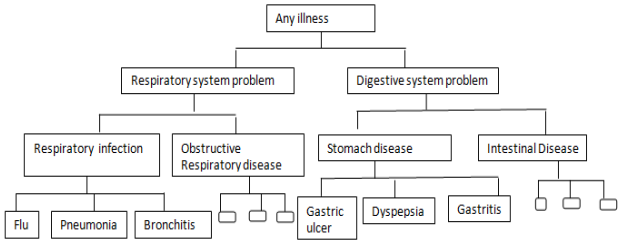

Figure 3: Taxonomy for Disease

Personalized privacy approach provides 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 [5] based on personalized anonymity concept that preserves a large amount of information in the microdata release without violating any privacy constraint.

To achieve personalized anonymity SA-generalization algorithm s 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.

4.1 Design View of Proposed System

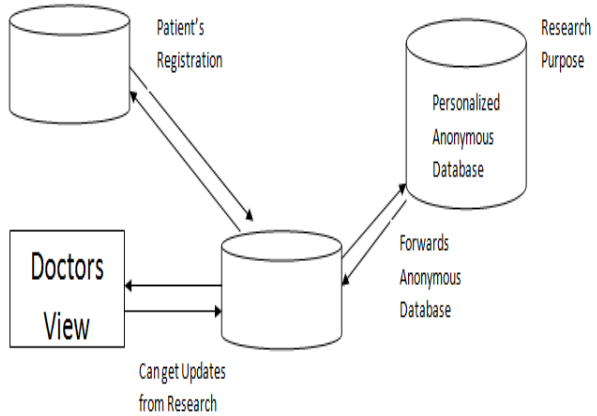

Figure 4: Design View of Proposed System

Figure 4 shows view of the proposed system for medical database for a patient, where the patient registers their details initially. The doctor can update record of patients and their treatments to the patient database. Also the doctor can retrieve information from other sources regarding the illnesses and their treatments.

The Personalized anonymous database can forward information to research center which has permission access the information. The research centre has its own access rights which will restrict use of the data. They can access only superficial data. They cannot access the patient details or the particular patient illness.

5. CONCLUSION

The 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 this paper, we work with the concept of personalized anonymity, and updates will be performed on these personally anonymized 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 which results mainly in less information loss.

REFERENCES

[1] 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.

[2] A.Machanavajjhala, J.Gehrke, et al., ℓ-diversity: Privacy beyond k-anonymity, In Proc. of ICDE, Apr.2006.B

[3] N. Li, T. Li, and S. Venkatasubramanian, t-Closeness: Privacy Beyond k-anonymity and l-Diversity, In Proc. Of ICDE, 2007, pp. 106-115.

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

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

[6] 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.

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

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

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