Neha Gosai, S.H.Patil / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 3, May-Jun 2012, pp.1596-1602

Generalization Based Approach to Confidential Database Updates

Neha Gosai*, S.H.Patil**
*(Department of Computer Science, Bharati Vidyapeeth University, Pune-43)
** (Department of Computer Science, Bharati Vidyapeeth University, Pune-43)

ABSTRACT

In many applications privacy becomes big issue as database size increases. For example, hospital or bank needs to share personal information of individuals in such a way that identities of the individuals cannot be revealed. In that case, personal identifier is removed to protect private or personal information is called anonymization.There are different data anonymization techniques. But in this paper k-anonymization approach is used. Suppose Alice having his own k-anonymous database and Bob wants to insert a tuple.So,the problem is to check after inserting a tuple whether database retains its k-anonymity or not. If allowing Alice to read content of tuple directly, it breaks the privacy of Bob and on the other hand database confidentiality violated once Bob has access to the contents of the database. In this paper, we propose a protocol to solve this problem on Generalization-based k-anonymous and confidential databases.

Keywords-**confidentiality, generalization, k-anonymity, privacy, SMC**

1. INTRODUCTION

In today's world database becomes valuable asset for many application, so security becomes critical. Bank information, medical research database – all these information can be dangerous if it fall into wrong hands. There is big concern of privacy. Privacy is the right of individual person to keep their information secret hiding from others. Privacy and confidentiality often used as synonyms but in reality there is a vast difference between these two. Privacy relates to person and Confidentiality relates to data. Data confidentiality is The nondisclosure of certain information except to authorized person. For example, In case of privacy, Health history or exam results are discussed in a private area. This may include asking an accompanying family member or friend to leave the room temporarily whereas in case of Confidentiality, Patient/participant information is not shared with other research team members in an elevator full of people.

To understand difference between Confidentiality and anonymity let's take an example of participants in research. In case of anonymity, identifying information is not used by researchers. Identifying information may be collected for regulatory or administrative purpose, but researches will not be able to use the data. In case of confidentiality identifying information collected, used in research but it will be removed after the research is complete. . Identifying information will not be available in any publication of the data.

There are lots of techniques developed for privacy. But one well-known technique is k-anonymization.such technique provide privacy by modifying data in such a way that it gives the same result for more than two tuples.So the problems of confidentiality and anonmization is different. The problem occurs when it comes to the updation of the database. When the tuple is to be inserted into the database, there are two problems: Is updated database still maintains privacy? And owner of the database really need to know the data to be inserted?

2. PRIVACY PRESERVING TECHNIQUES

A number of techniques have been developed to provide privacy to the databases such as, data modification techniques, query auditing techniques, randomization, and k- anonymity.

2.1 Randomization

The randomization method provides effective way of preventing the user from learning sensitive data which can be easily implemented because the noise added to the given record is independent from the other records. The amount of noise is large enough to smear original values, So individual record cannot be recovered. The randomization method is simple as compare to other methods because it does not require to knowledge of other records. That is why randomization can be used without the use of server that contains other records also. Large randomization increases the uncertainty and the personal privacy of the users. However, at the same time, larger randomizations can cause loss in the accuracy Kargupta [1] challenges perturbation and randomization –based approaches. They claim that approaches may lose information as well as not provide privacy by introducing random noise to the data by using random matrix properties, Kargupta successfully separates the data from the random noise and subsequently discloses the original data.

2.2 Secure multi-party computation

Goal of secure party computation is to compute function when each party has some input. It generally deals with problems of function computation with distributed inputs. In this protocol, parties have security properties e.g., privacy and correctness regarding privacy a secure protocol must not reveal any information other than output of the

function. Example of such a computation is the "millionaires' problem", in which two millionaires want to find out who is richer, without revealing their actual worth.Though there is difference between SMC and k-anonymity model. In k-anonymity model result can be out during the process but this is not the case in SMC.k-anonymity model protect actual value.

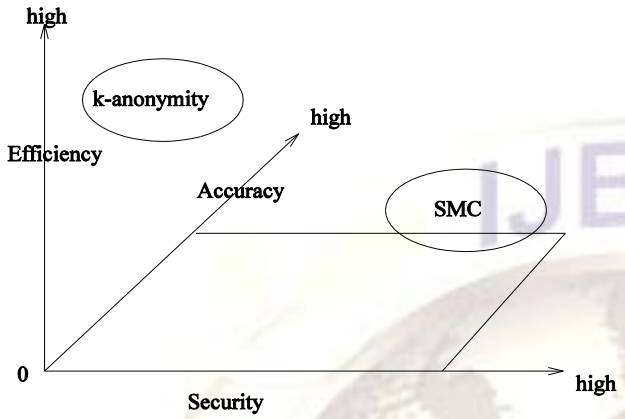


Figure 1: k-anonymity and SMC

K-anonymity and SMC are used in privacy-preserving data mining, but they are quite different in terms of efficiency, accuracy, security and privacy as shown in Figure 1[2].

2.3 K-anonymity Model

A large number of privacy models were developed most of which are based on the *k*-anonymity property. The K-anonymity model [3] was proposed to deal with 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. For example, Hospital contains large database in such a way that identity of individual cannot be revealed. It helps to reveal public databases without compromising privacy.Thus, it prevents database linkages. In k-anonymity the granularity of data representation is reduced by using techniques such as generalization and suppression. The granularity is reduced to such a level that any given record maps onto a least K other records in the dataset. A general method widely used for masking initial micro data to conform to the k-anonymity model is the generalization of the quasi identifier attributes [4].

3. PROBLEM STATEMENT

Consider an example of medical research database. All patients' records or data are to be stored in the research database under the condition that each patient's privacy is protected.Research database actually stored only anonymized version of each patient record. Suppose Certain data contains patient history related to negative effect of drug on patient. These information need to be confidential and accessible only by the few researchers. If data will be exposed then manufacturing company of drug gets affected.

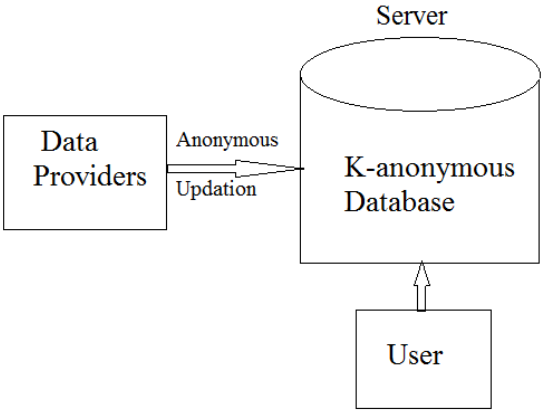


Figure 2 Anonymous Database System

Figure 2 shows anonymous database system [5]. Suppose Alice owns database and only he has having access to the database. Bob is data provider who inserts a tuple which contains patient's data. Database DB is confidential and anonymous. Database contains sensitive information so it is necessary to protect those data. It can be achieved by anonymization.So if the database is anonymous then it is not possible to reveal any information of patient. Now suppose bob inserts new tuple then obliviously database needs to be updated. The modification to the anonymous database can be achieved by two ways:

1) Bob or the server checks whether the updated database DB retains its anonyminity after inserting a tuple.2) Make entire database to Alice so he can insert a tuple by himself. But the problem occurs in both ways. In 1st problem if server inserts tuple then entire tuple needs to be revealed to the Bob. In 2nd problem Entire database needs to be available to the Alice which violates Bob or server confidentiality. To overcome this problem, in this paper we propose a protocol. Anonymous communication between Alice and Bob is carried out by communication protocol Crowds [6]. The main idea behind Crowds anonymity protocol is to hide each user's communications by routing them randomly within a group of similar users.

4. PROPOSED METHOD

The protocol rely on the fact that anonynimity of database does not affected by inserting tuple if the information contained in tuple ,properly anonymized,is already contained in database. Then the problem rose that privately checking whether there is a match between tuple to be inserted and tuple that already contained in database. To solve this problem the protocol Generalization based anonymous databases and it relies on Secure set intersection protocol to preserve privacy updates in generalization based method [7].

4.1 Prototype Architecture

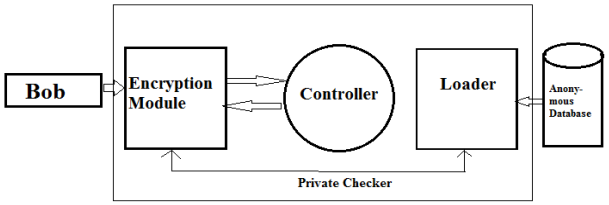


Figure 3: Proposed Block Diagram

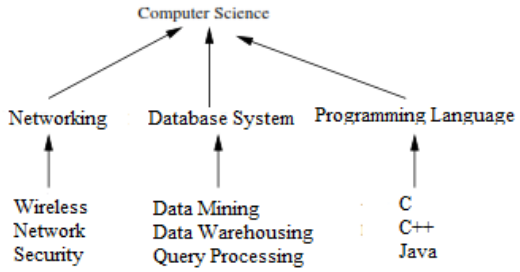
In Figure 3, Private checker module made of Encryption module, Controller, Loader module.
Bob enters data is stored into encryption module which is responsible for encrypting all tuples exchanged between Bob and the private Updater. Loader module reads data or tuple from anonymous database (Alice).Controller module that performs all the controls that is it checks whether the inserted data is matched with the data's in the anonymous database using generalization method. The main concept behind private checker is to check whether insertion is possible into the k anonymous database.

4.2 Generalization Based protocol

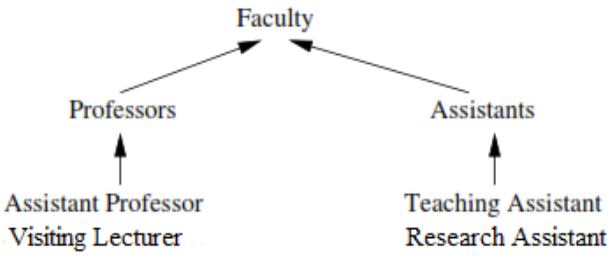
The generalization based protocol relies on well known cryptographic techniques. We consider table T= {t1… t2} over the attribute set A. The main idea of generalization based approach is original value is replaced by more general values according to value generalization hierarchies. So each attribute value replaced by more general value. For example, figure 4 shows value generalization hierarchy (VGH) which contains AREA, POSITION and SALARY. In VGH specific value is replace by more general value. For example, VGH of SALARY, $15,000 replaced by more general value [11k, 60k].Table 1 shows original dataset with column area, position, salary. Table 2 shows generalized data set with k (= 2) anonymity. It means there should be at least k tuple indistinguishable in table.

4.3 Fundamental Primitives

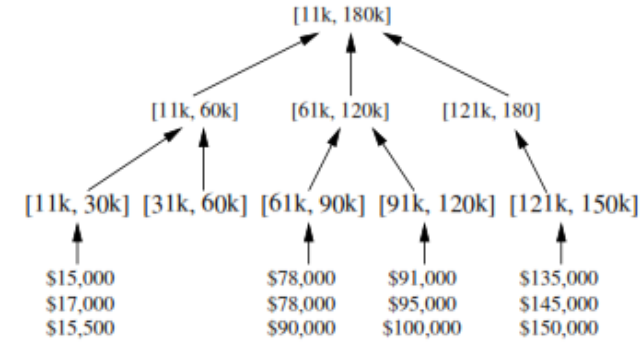
Generalization based k-anonymity protocol uses encryption scheme of commutative and product homomorphic. This encryption scheme allows to perform mathematical operation over encrypted data. We provide definition of Commutativity, product homomorphic and indistiguishability [8].



VGH of AREA



VGH of POSITION



VGH of SALARY

Figure 4: Value Generalization Hierarchy

TABLE 1 Original Data

AREA	POSITION	SALARY
Wireless	Assistant Professor	78000
Network Security	Assistant Professor	78000
C	Visiting Lecturer	15000
C++	Teaching Assistant	17000
Data Mining	Research Assistant	95000
Data Warehousing	Visiting lecturer	90000

TABLE 2 Generalized Data with k=2

AREA	POSITION	SALARY
Networking	Professor	[61k,120k]
Networking	Professor	[61k,120k]
Programming Language	Professor	[11k,60k]
Programming Language	Assistant	[11k,60k]
Database System	Assistant	[61k,120k]
Database System	Assistant	[61k,120k]

Given a finite set K of keys and a finite domain, a commutative, product homomorphic encryption scheme E is a polynomial time computable function $E: K \times D \rightarrow D$ satisfying the following properties:

1) *Commutativity:* For all key pairs $k1, k2 \in K$ and value $d \in D$,the following equality holds:
$$Ek1(Ek2(d)) = Ek2(Ek1(d))$$
2) *Product Homomorphism:* For every $k \in K$ and every value pairs $d1, d2 \in D$,the following equality holds:
$$Ek(d1) \cdot Ek(d2) = Ek(d1 \cdot d2)$$
3) *Indistinguishability:* It is infeasible to obtain data of plaintext from cipher text. The advantages are high privacy of data even after updation, 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 two users to establish shared secret key over insecure communication without having any prior knowledge. Here, Diffie Hellmen is used to agree on shared secret key to exchange data between two parties.

Here, we have assumed that database is k-anonymous. So it needs to check that after inserting properly anonymized tuple, by Bob, whether database (Alice) maintains its k-anonymity.If this is the case, tuple can be inserted otherwise tuple will be rejected.

4.4 Algorithm

Let t is Bob's private tuple from table T containing anonymous attributes ,so Bob can generate τ which holds corresponding values t[A1],…,t[Au];Let u (Size of anonymous tuple)be disjoint value Generalization hierarchies corresponding to anonymous attributes known to Alice. Let $\delta \in T$ and let *Getspec* (δ) be specific value that is bottom of VGH related to each anonymous attribute. Function f denotes to *GetSpec* (δ).Now,Bob generates a set τ containg corresponding values ot tuple t.We use Secure Set Intersection(SSI) protocol to compute cardinality of set.Here,we denote SSI(f,γ) as a secure protocol which computes cardinality of $f \cap \tau$ .On the receiving first request Alice chooses random tuple from table T. Alice computes function f= *GetSpec* (δ). Alice and Bob individually compute SSI (f,τ).next step to compare SSI (f, τ) with u.If both are equal then t in generalized form can insert tuple in database. Otherwise it again computes until we get both values same.

Algorithm:

1) Alice randomly choosesδ.
2) Alice computes f = *Getspec* (δ)
3) Alice and Bob privately compute s= SSI(f,τ)
4) If s= u, then generalized form of t can be inserted into T,
5) Otherwise, Compute until we get s=u.

Example:

Suppose for this example u=3 and Bob is having tuple t =[Wireless, Assistant Professor,$15000] So we can have τ={Wireless, Assistant Professor,$15000}.Now suppose δ=[Networking,assistant Professor,[11k,60k] then f={Wireless, Network Security, Assistant Professor,$15000,$17000,$15500}.Since $|f \cap \tau|$=3 =u, we can conclude that tuple t is properly anonymized so it can be inserted to T. Suppose t=[Database System, Research Assistant, $17,000] and δ = [Programming Language, Research Assistant, [11k, 30k]].then u=2<3 so t will be rejected.

5. IMPLEMETATION AND RESULT

Generalization based k-anonymity approach to provide privacy updates to confidential database is designed by using Java. The implementation setup Considered attributes AREA, POSITION and SALARY. Figure 5 shows home page of Generalization based approach. If the data entered by the user matches with the general value then this record will replaced by the general value and these general values being inserted into table. To carry out this task, we have made separate table for original values and general values. When user enters data it checks value in original table if it is valid then it matches with the values of generalized table. Based on this outcome data will get inserted or rejected as shown in snapshot below.

The implementation shows that the complexity of protocol depends on the number of message exchanged and their size. The complexity of protocol relies on the size of T and the complexity of SSI protocol. We implemented protocol using java and database created using My SQL. Experiment executed on Pentium 1GHz with 1 GB physical memory.

As a result if we enter all values correct then database will be updated successful otherwise tuple will not be inserted to the database. Thus we can say that database successfully updated while preserving privacy and k-anonymity.

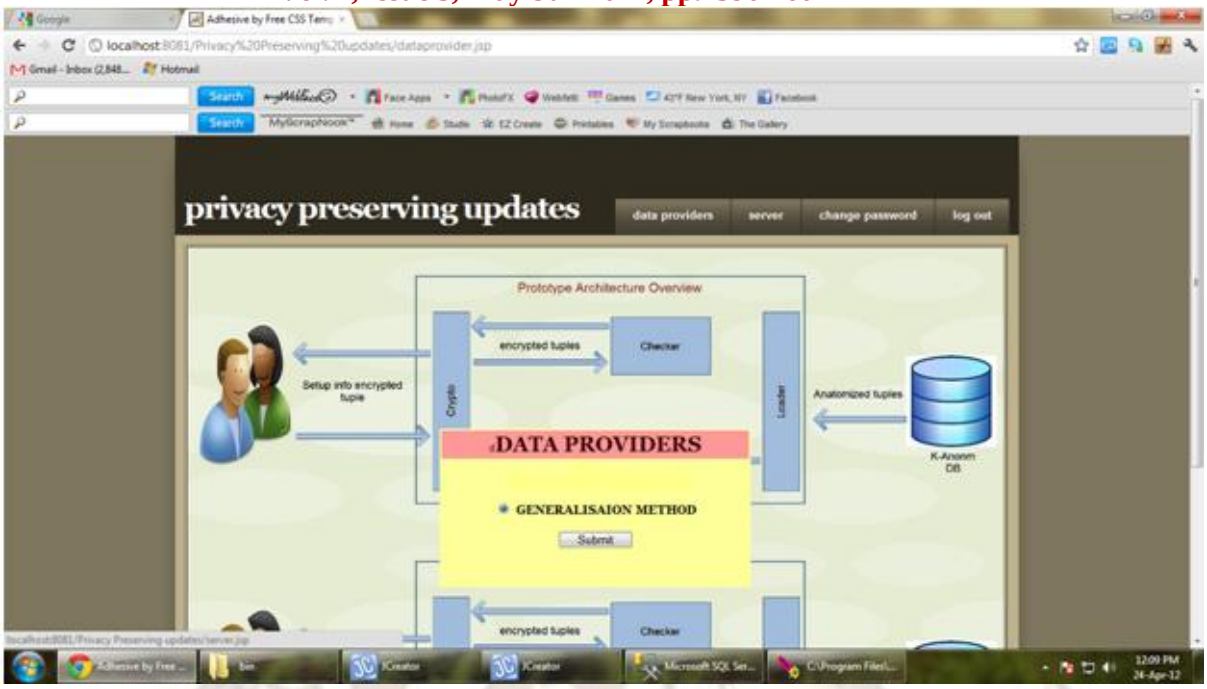


Figure 5: Generalization Based privacy preserving Updates

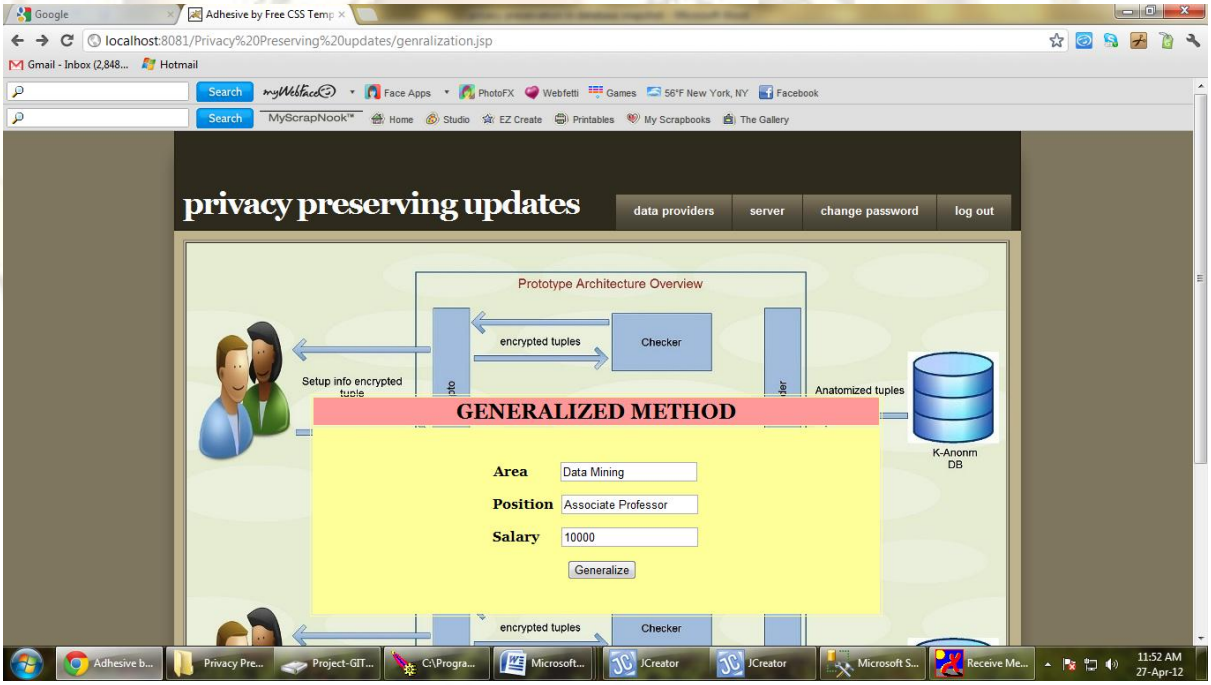


Figure 6: Attributes with correct data values

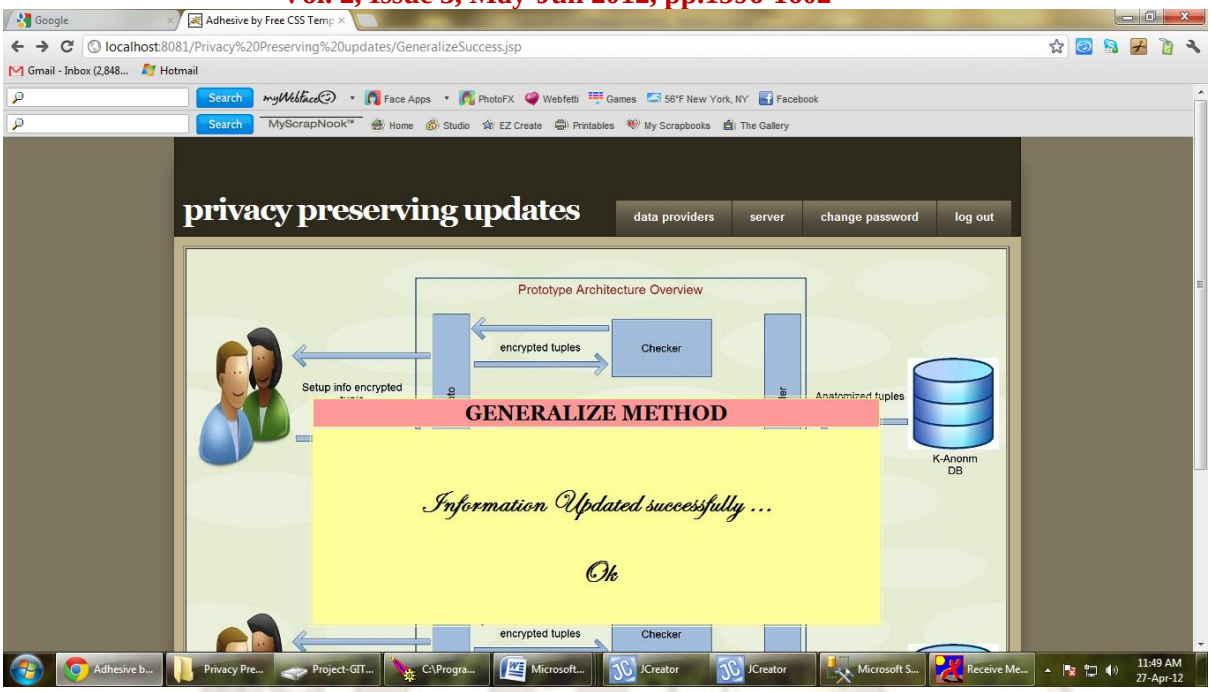


Figure 7: Output screen of correct data values

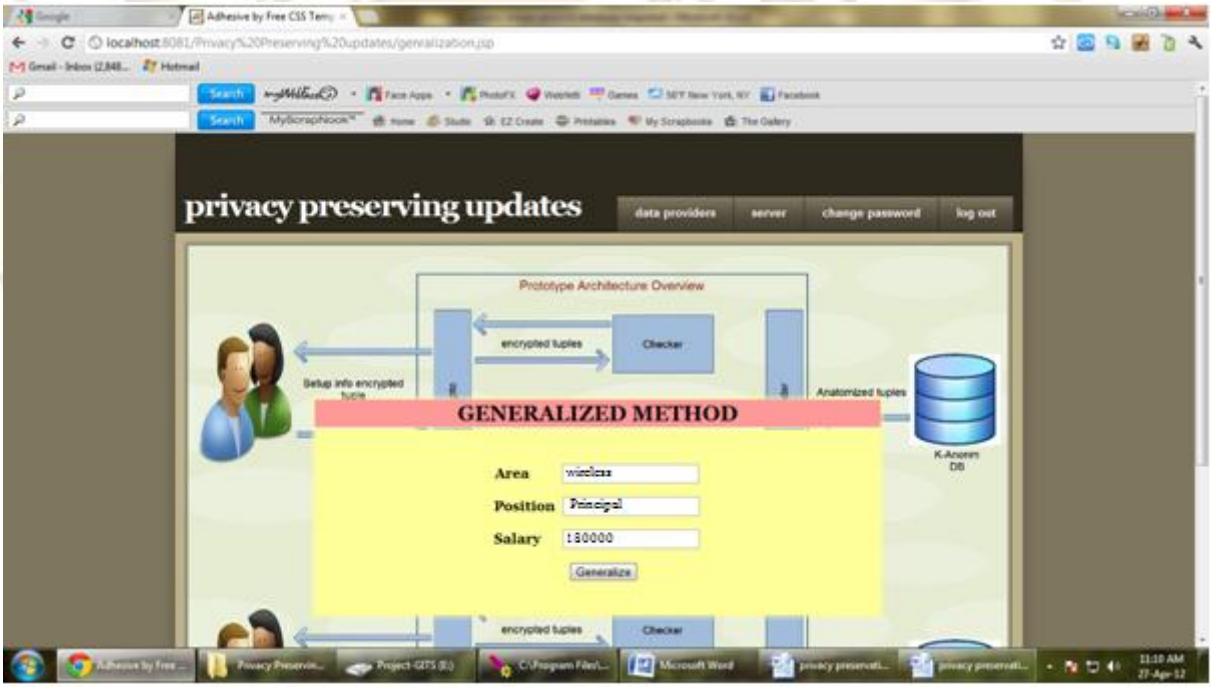


Figure 8: Attributes with wrong data values

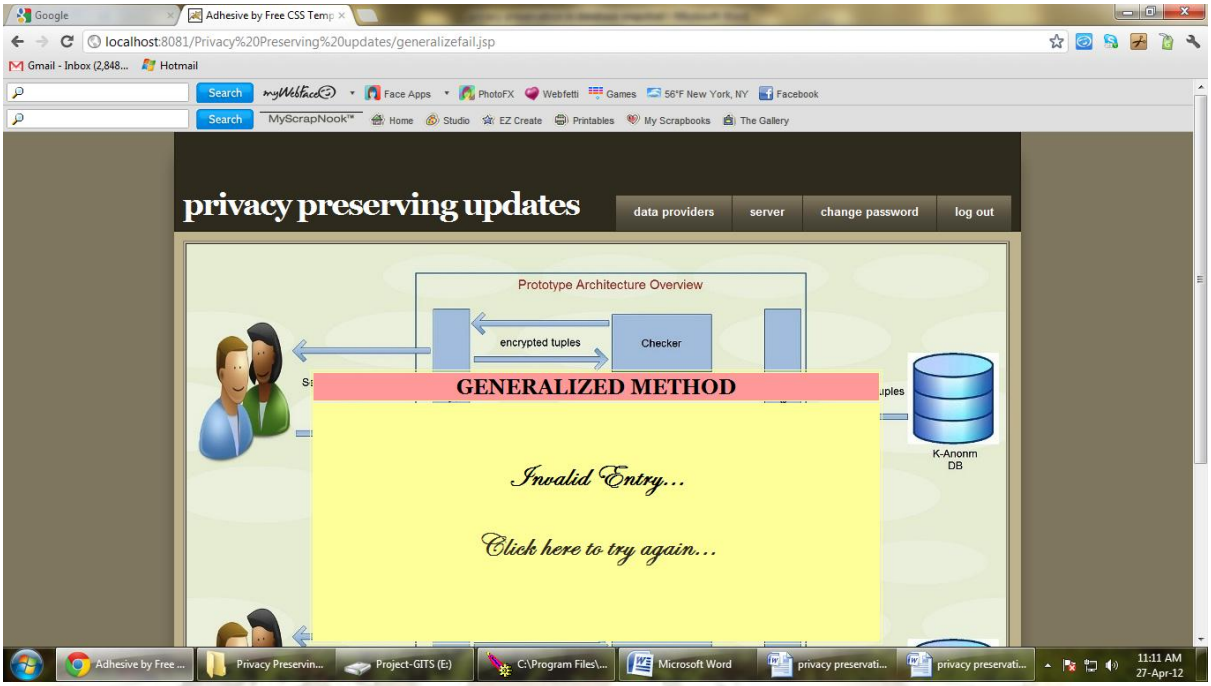


Figure 8: Output screen with wrong data values

6. ACKNOWLEDGEMENT

Prof. Dr. S. H. Patil provided the opportunity to write this paper and gave editorial suggestions on a very early draft. Finally, I thank to review my paper and his invaluable suggestion that make to improve quality of my paper.

7. CONCLUSION AND FUTURE WORK

In this paper, we have proposed secure protocol for privately checking whether a k-anonymous database retains anonymity once a new tuple is being inserted to it. In this protocol, we have shown that 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. But generalized based approach is not sufficient protocol as if a tuple fails to check, it does not insert to the database and wait until k-1.because of this much of long process waiting time also gets increase.

The important issues in future will be resolved:

1) Implement database for invalid entries.
2) Improving efficiency of protocol in terms of number of messages exchanged between user and database.
3) Implement real world database system.

8. REFRENCES

[1] Kargupta H. Datta, S. Q. Wang and K. Sivakumar ,"*On the privacy preserving properties of random perturbation techniques*"IEEEICDM,2003

[2] Current Developments of *k*-Anonymous Data Releasing, Jiuyong Li, Hua Wang, Huidong Jin, Jianming Yong' School of Computer and Information Science, University of South Australia, Mawson Lakes Adelaide, Australia, 5095

[3] Samarati P and Sweeney. L "*Protecting Privacy when Disclosing Information: k_anonymity and its Enforcement Through Generalization and Supression*"IEEE Symposium on Security and Privacy.(1998).

[4] Constrained k-Anonymity: Privacy with Generalization Boundaries, John Miller, Alina Campan, Traian Marius Truta.

[5] Privacy –Preserving Updates to Anonymous and Confidential Databases, Alberto Trombetta, Wei Jiang, Elisa Bertino and Lorenzo Bossi.

[6] M. K. Reiter, A. Rubin.crowds: anonymity with Web transctions.ACM Transactions on Information and System Security (TISSEC), 1(1), 1998; 66-92.

[7] Privacy Preserving Set Intersection Protocol Secure Against Malicious Behaviours Yingpeng Sang, Hong Shen School of Computer Science. The University of Adelaide, South Australia, 5005, Austrailia.

[8] S. Brands, Untraceable off-line cash in wallets with observers. In Proc. Of CRYPTO onf. Lecture Notes in Computer Science, 773, 1994; 302-318.

[9] www.wikipedia.com/wikifiles/.