International Journal of Technical Research and Applications e-ISSN: 2320-8163,
www.ijtra.com Special Issue 31(September, 2015), PP. 44-50

SECURE MINING FOR MEDICAL RESEARCH

Mr. Shetye A.N
PG Scholar, Department of Computer Engineering
Terna Engg.College ,Nerul , Mumbai University, India
akshay.shetye123@gmail.com

Prof. Dhumal R.A
Assistant Professor, Department of Computer Engineering,
Terna Engg.College, Nerul,Mumbai University, India.
rashmisalvi@gmail.com

***Abstract*—— Privacy preserving mining of distributed data has various applications. Privacy consideration may prevent mining approach to gather all data into central site. Data sharing among N different users is very common thing. In data sharing, privacy of corresponding users should be preserved. Many applications like hospital, bank need to share personal information of individuals that cannot be revealed. Data sharing means data is shared with other users but while doing so privacy of sensitive data should be kept confidential (Confidentiality).**

The proposed technique makes data to be shared with other members. This data sharing is possible, because the proposed mechanism will alter gathered data before delivering it to the data miner. The altered data is also available in different format depending upon the data miner. The system can preserve private information of any user, sensitive data about any company, and disease information about any patient. The proposed technique "Secure Mining in Medical Research" is mainly for preserving private information about patient, while sharing it with different researchers to define a policy, to find out drugs on some disease depending on the collected data from various patients.

***Keywords-Generalization,* Specialization, Privacy, K-Anonymous, Confidentiality**

I. INTRODUCTION

Data sharing among N different users is very common thing. Many applications like hospital, bank need to share personal information of individuals that cannot be revealed. Data sharing means data is data shared with other users but while doing so, privacy of sensitive data should be kept confidential. There are various techniques which are used for preservation of private data while sharing it with other users. Some of these techniques are randomization, secure multiparty computation, K-anonymity. In randomization, there is an addition of noise in a given record so that it is not possible to recover original data from it .In secure multiparty computation, each party will compute result depending on data from each party. There are two different approaches of making K-anonymous database, named as Generalization Approach and Suppression Approach.

In suppression approach, main aim is to form subsets of indistinguishable tuples by masking the values of some well chosen attributes. We mask with special value(*), then form a subset and then classify that subset by using Quasi Identifier(QI).Each record has number of attributes :some attributes are unique and personal(such as disease and salary) and some maybe repeated and general such as zip code, age, gender. By taking this we can easily identify someone. In generalization approach ,main idea is to replace original value by some general values depending upon the Value Generalization Hierarchy(V.G.H).For example value of salary 25000 is replaced by[11k ,30k].These general values are dependent on the value of k, used to maintain K-anonymous database.

II. LITERATURE REVIEW

In literature survey, we have studied how different previous techniques have been used for securely sharing data among different users/researchers/miners so that no sensitive information get revealed from it. Those techniques are listed below.

A. How to share secrete

This technique shows how the given data D can be reconstructed from any k pieces, but the complete knowledge of any k-1 pieces can reveal any information about D. It means divide number D into n pieces (D1, Di.. Dn) such way that knowledge of any k or more pieces of Di can reveal D. This (k, n) scheme is called as Threshold Scheme. This scheme is ideally suited for applications where group of mutually suspicious individuals with conflicting interests must cooperate with each other. [1]

This scheme is based on polynomial' interpolation:
Given k points in the 2-dimensional plane (x,, y,) . . . . .
(xk, Yk). with distinct xi's , there is one and only on epolynomial q(x) of degree k - 1 such that q(x) =yi for all i. Without loss of generality, we can assume that the data D is (or can be made) a number. To divide it into pieces Di, pick a random k-1 degree polynomial
$q(x)=a_o+ a_lx + . . . a_{k-1}x^{k-1}$ in which ao=D , and evaluate:
D1=q(1),Di=q(i),Dn=q(n)
given k points in the 2-dimensional plane (x,, y,) . . . . .
(xk, Yk). with distinct xi's , there is one and only one polynomial q(x) of degree k - 1 such that q(x) =yi for all i. Without loss of generality, we can assume that the data D is (or can be made) a number.

B. Secure Multiparty Computation

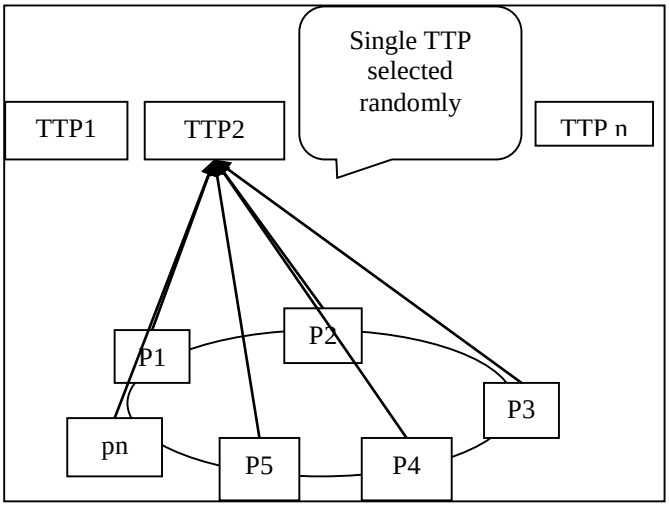

Figure 1: Secure Multi Party Computation Protocol

A secure multiparty Computation allows parties to compute results upon their private data, minimizing the threat of disclosure. This technique provide an encryption mechanism by which provide security in data sharing. Secure Multi-Party This work is to avoid ambiguities; at the same time ensuring the security of information by taking efficient measures. In this secure multiparty computation protocol, first data is distributed and then sent forward, so that no single party will become a victim of intercepting of data by other involved parties. Here the sole responsibility is not vested on single person/entity. This protocol is totally dynamic in nature. An encrypted nature of data provides security to the data. The identity of T.T.P. has been hidden until runtime, so the complexities are also reduced in it.[2]

C. SADS (Secure anonymous Database Search) Protocol

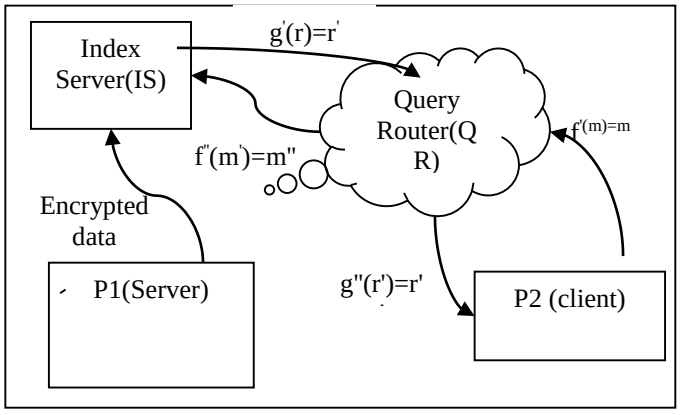

Figure2: Secure Anonymous Database Search

This protocol provides method for well defined and secure sharing of data between untrusting parties. Clients can search information residing on a server without revealing his/her identity as well as content of his query to server. This scheme is available only for authenticated clients. Means client's identity and query content is protected from server. This protocol is useful for those parties who do not want to share data any data until they do. The framework designed for this allows an authorized client to anonymously submit keyword query securely on documents/database stored on database.[5]

K-anonymity: a model for protecting privacy

It was carried in 2002 by L.Sweeney**,** The solution provided in this includes 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 atleast *k*-1 individuals whose information also appears in the release. This examines re-identification attacks that can be realized on releases that adhere to *a*nonymity unless accompanying policies are respected [6]. This paper has presented the K-anonymity protection model, explored related attacks and provides the way in which attacks can be thwarted.

III. Problem Statement

Various methods for anonymity of Database have been developed. These works are still subject to some drawbacks. 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. To achieve the objective to check whether the database inserted with the tuple is still k-anonymous, without letting admin and user know the contents of the tuple and the database. 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, and we provide theoretical analyses to proof their soundness and experimental results to illustrate their efficiency.

Research Objectives:

- To find the efficient method or techniques of updating database which ensure Privacy preserving and K-anonymity.
- Methods should be such that the privacy of Data provider and confidentiality of Database owner is maintained through Update

IV. Proposed System

A. K-Anonymous Property

Consider for a data holder, such as a hospital or a bank that has a privately held collection of person-specific, field structured data (Tabular form). Suppose the data holder wants to share a version of the data (i.e. some part of data) with researchers. The question is "How can a data holder release a version of its private data with scientific guarantees, that the individuals who are the subjects of the data cannot be reidentified, while the data remain practically useful?" One solution can be used named as *k*-anonymity. A release of data

provides *k*-anonymity protection. In this technique, 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. The re-identification attacks are also taking into consideration. The *k*-anonymity protection model provides guarantees of privacy protection. Privacy preserving data mining techniques clearly depend on the definition of privacy, which captures what information is sensitive in the original data and should be protected from either direct or indirect disclosure. K-anonymity states that each release of data must be such that every combination of values of released attributes that are also externally available and therefore exploitable for linking can be indistinctly matched to at least k respondents.[6][8].

B. Implementation & Result

In generalization-based anonymization consists of substituting general values. The main focus is on person-specific data, so the entities are people, and the property to be protected is the identity of the subjects whose information is contained in the data. However, other properties could also be protected. The values of a given attribute with more general values in the database, according to a priori established value generalization hierarchies (VGHs). Table 1 contains original information and after performing generalization based techniques original dataset is anonymized and table 2 shows generalized data with k=2. Generalization replaces a value with a "less-specific but semantically consistent" value. In a VGH, leaf nodes correspond to actual attribute values, and internal nodes represent less-specific values. For understanding Figure 3 contains VGHs for Quasi identifier or attributes DISEASE, GENDER and AGE. Generalization schemes can be defined based on the VGH that specify how the data will be generalized. According to the VGH of DISEASE, the value of disease is generalized according to the disease causes. Like "HIV" cause by virus so it can be generalized to "Diseases Caused by virus". The Gender hierarchy in the figure is generalized based on Male and Female category. The attribute Age is generalized to the interval (1-15) and (16-30), then to the interval (1-30).

DISEASE	GENDER	AGE
Typhoid	Girl	19
HIV	Man	50
Typhoid	Women	45
Exothrix	Man	68
HIV	Boy	20
Exothrix	Women	63

Table1: Original table

DESEASE	GENDER	AGE
Disease Caused by Bacteria	Female	[1-30]
Disease Caused by Virus	Male	[31-60]
Disease Caused by Bacteria	Female	[31-60]
Disease Caused by Fungi	Male	[61-100]
Disease Caused by Virus	Male	[1-30]

Table 1: Generalized Data with k=2

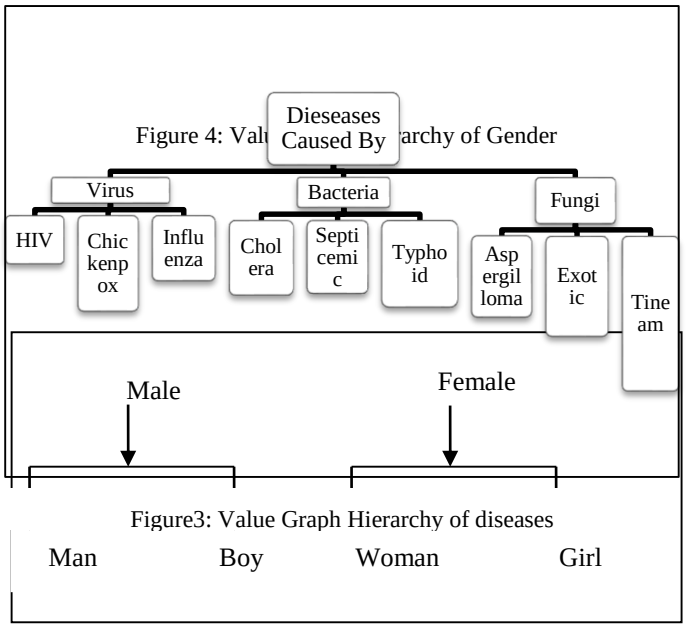

Figure4: Value Graph Hierarchy of Gender

Figure3: Value Graph Hierarchy of diseases

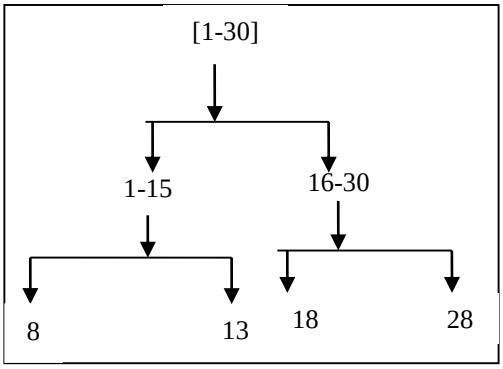

Figure 5: Value Graph Hierarchy of Age

C. Algorithms

Here we have used two algorithms to achieve K-anonymity namely generalization and suppression algorithm.

1) Suppression Algorithm

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.

2) Generalization Algorithm

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.

The protocol works as follows:

Step 1: User A randomly chooses a $\delta \varepsilon T_w$ (Witness Set).

Step 2: User A computes γ = GetSpec (δ).

Step 3: User A and User B collaboratively compute s= SSI (γ, τ).

Step 4: If s=u then t's generalized form can be safely inserted to T.

Step 5: Otherwise, User A repeats the above procedures until either s=u or witness set is empty.

Let t is User B's private tuple from table T containing anonymous attributes ,so User B 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 User A. Let $\delta \in T$ and let Getspec (δ) be specific value that is bottom of VGH (Value Graph Hierarchy) related to each anonymous attribute [14]. Function γ denotes to *GetSpec* (δ).Now, User B generates a set τ containing corresponding values to tuple t.We use Secure Set Intersection (SSI) protocol to compute cardinality of set. Here we denote SSI(γ, τ) as a secure protocol which computes cardinality of $\gamma \cap \tau$ .On the receiving first requests User A chooses random tuple from table T. User A computes function γ = GetSpec (δ). User A and User B individually compute SSI (γ,τ).next step to compare SSI (γ, τ) with u.If both are equal then t in generalized form can be inserted in database. Otherwise it again get computes until we get both values same.

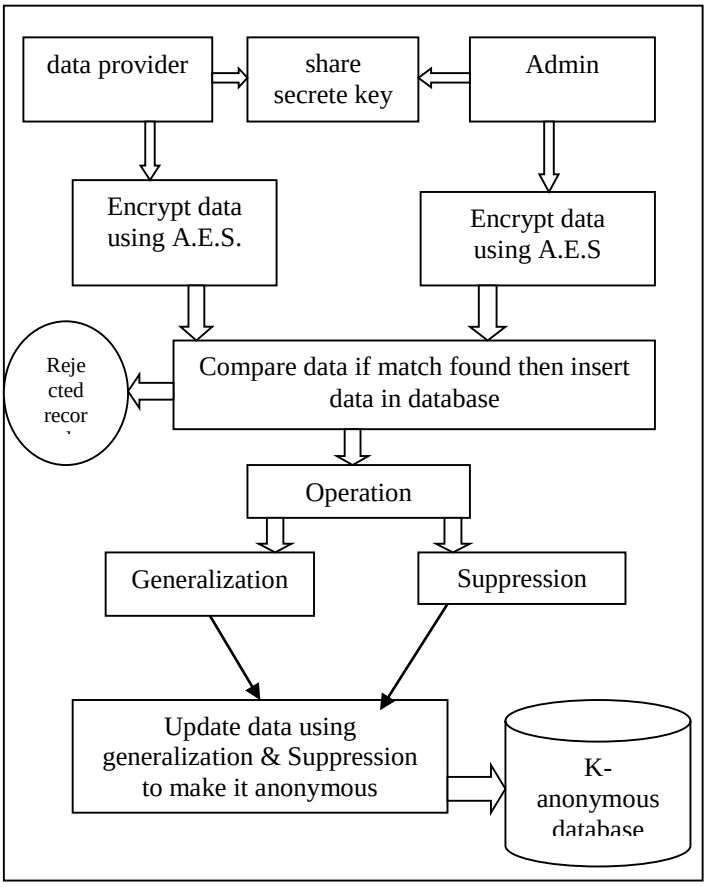

Figure 6:Overall architecture of Proposed system

D. Modules in Proposed system

1) Data-provider Module.

2) Researcher Module.

3) Cryptography Module.

4) Admin Module.

1) Data-provider Module:

This module is used to provide authorized access for data-provider. Authorization is achieved by using registration of data provider. If registration is successful then user should compulsory enter username and password if this information is correct then he/she will further proceed.

2) Researcher Module:

In this module, N number of researchers is validated by username and password after their registration process. A secrete key is shared among researcher and server by using define hellman key exchange algorithm. Total N number of anonymous key are generated for total N number of users.

3) Encryption Module:

In this module, if data-provider wants add some new tuples then encryption module convert this data into cyphertext .This precaution is to avoid any kind of attack on plain data due to shortfall of data transfer medium.

4) Admin Module:

Admin module checks converts original database into K-anonymous database depending on Quassi identifier and value of k. If new upcoming entry of tuple matches with k-anonymous property then and then only that entry will be allowed in database.. Here k-anonymous property is preserved by using generalization and suppression approach.

5) K-Anonymity Module:

The generated K-anonymous data is transferred to researcher depending on value of K .

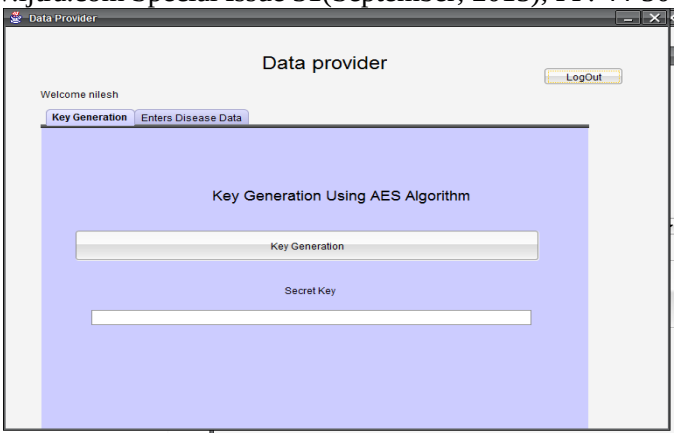

Figure 8: Secrete Key generation for provider and server

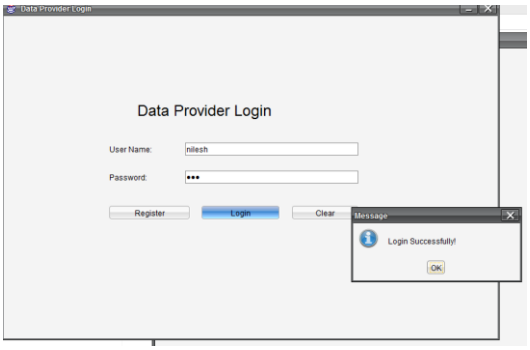

Figure 7: Output screen of Data Provider

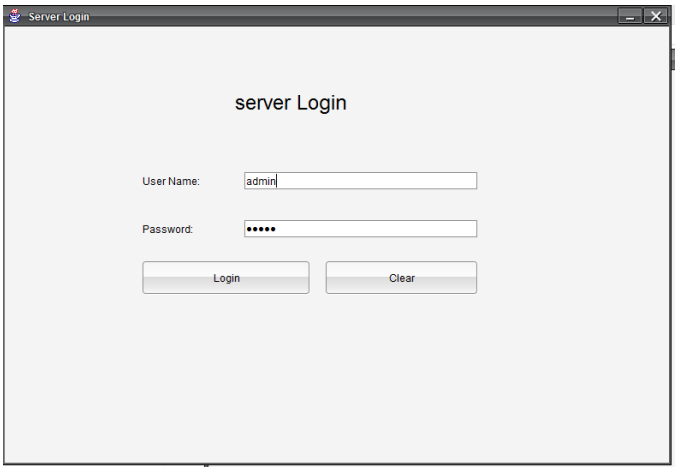

Figure 9: Data administrator login

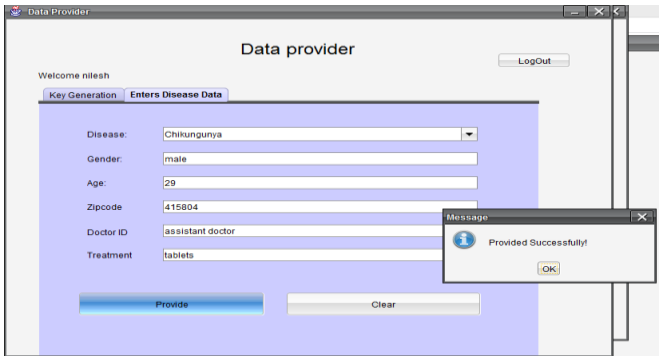

Figure 8: Data Provider inserts data

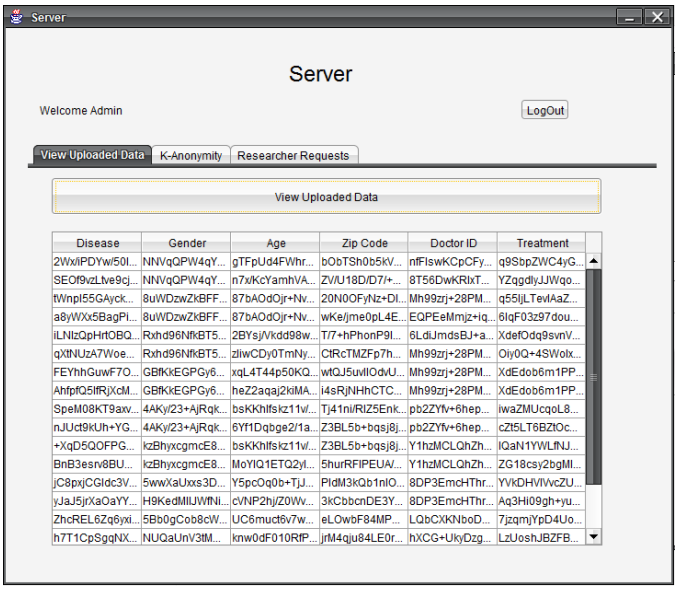

Figure 10: Encrypted form of data at server side

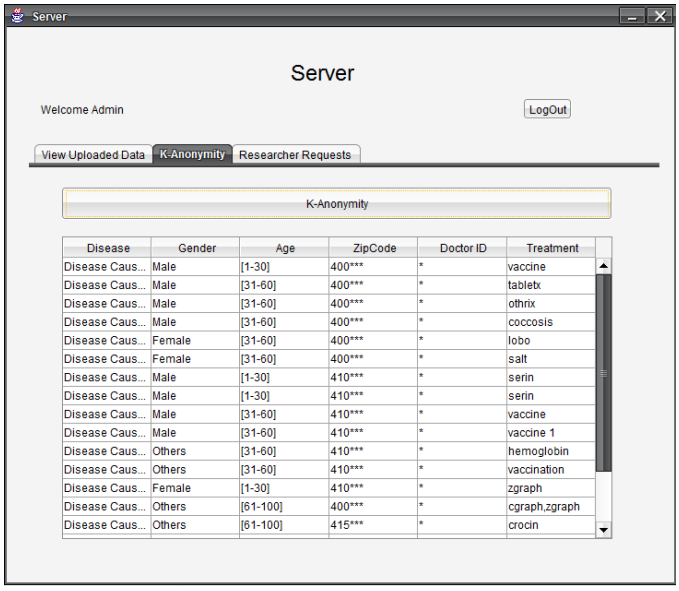

Figure11: K-anonymous form of data(By generalization, suppression)

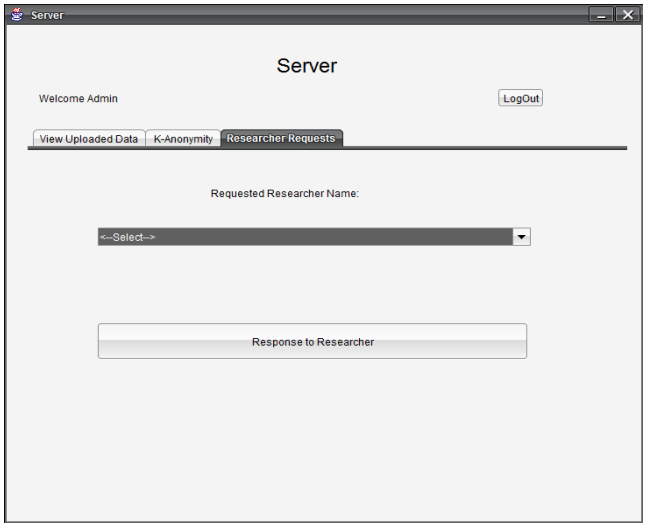

Figure 12: Pending requests from researchers

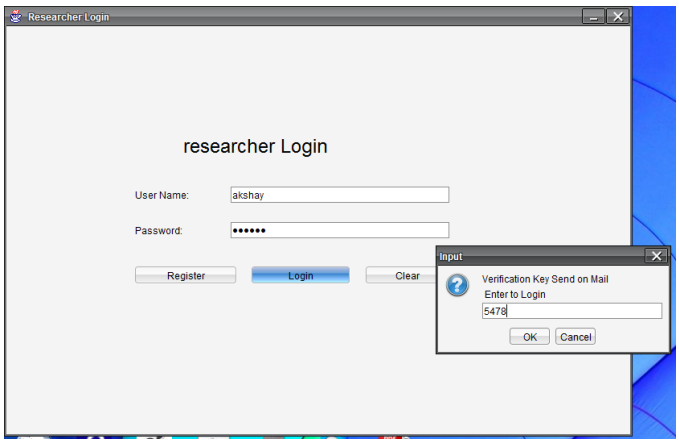

Figure13: Researcher Login

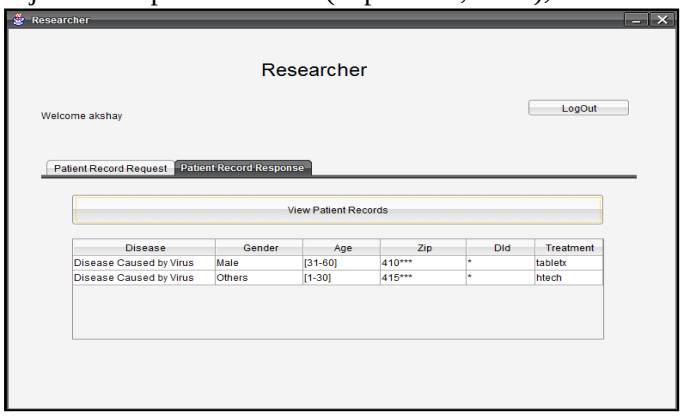

Figure 13: Response to researcher's request

E. *Performance Evaluation Factors*

- Information Loss: ILOS(vg)=$|v_g|$-1/$|D_A|$
- Information Loss for record ILOS(r)=$\sum w_i$.(ILOS(v_g))
  W_i=penalty weight of attribute
- Information Loss for table IOS(T)=$\sum_{r\in T}$ ILOS(r)
- PG= avg{A(QIDj)-As(QIDj)}.

Where A(QIDj) and As(QIDj) denote the anonymity of QIDj before and after specialization.

The Principle of information/privacy trade-off can also be used to select a generalization g, in the
which case it will minimize.

ILPG=IL(g)/PG(g)

Where IL(g) denotes the information loss and PG(g) denotes the privacy gain by performing g.

CONCLUSION

To preserve privacy in any research, we demonstrate that the proposed technique is better than existing system in achieving security. The proposed system uses anonymous id which do not require trusted centralized authority. The proposed system has better performance than multi party computation such as secure sum and power sum algorithm. Data is encoded in its general form (K-anonymous) depending on type of user accessing it. Also this K-anonymous property has been checked every time while entering any tuple in it. The solution on most probable attacks on K-anonymity has also provided in this system.

REFERENCES

[1] Shamir,"How to share a secret," Commun. ACM, vol. 22, no.11,pp. 612–613, 1979.

[2] Dr. Durgesh Kumar Mishra, Neha Koria, Nikhil Kapoor, Ravish Bahety,"A Secure Multi-Party Computation Protocol for Malicious Computation Prevention for preserving privacy during Data Mining", IJCSIS) International Journal of Computer Science and Information Security,Vol. 3, No. 1, 2009.

[3] M.Divya Meena, AR. Arunachalam, Dr T. Nalini "Confidential Data Sharing With Anonymous Id Assignment Using Central Authority", "International

Journal of Inventions in Computer Science and Engineering" ISSN ISSN (Print): 2348 – 3431 Volume I Issue 2 2014.

[4] J. Smith, "Distributing identity [symmetry breaking distributed access protocols]," IEEE Robot. Autom. Mag., vol. 6, no. 1, pp. 49–56, Mar 1999.

[5] R. Canetti, "Security and composition of multi-party cryptographic protocols," J. Cryptol., vol. 13, no. 1, pp. 143–202, 2000.

[6] Friedman, R. Wolff, and A. Schuster, "Providing k-anonymity in data mining," VLDB Journal, vol. 17, no. 4, pp. 789–804, Jul. 2008.

[7] Latanya Sweeney , "model for protecting privacy", International Journal on Uncertainty,Fuzziness and Knowledge-based Systems", 10 (5), 557-570,2002.

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

[9] Mahesh T.Dhande, Neeta A.Nemade," Performance Improvement of Privacy Preserving in K-anonymous Databases Using Advanced Encryption Standard Technique"," International Journal of Advanced Research in Computer Science and Software Engineering" Volume 3, Issue 6, ISSN: 2277 128X ,June 2013.

[10] Lambodar Jena1, Narendra Kumar Kamila," Distributed Data Mining Privacy by Decomposition (DDMPD) with Naive Bayes Classifier and Genetic Algorithm"," International Journal of Application or Innovation in Engineering & Management (IJAIEM)", Volume 2, Issue 7, ISSN:2319 – 4847,July 2013.

[11] Larry A. Dunning, Ray Kresman "Privacy preserving data sharing with anonymous id assignment", "IEEE Transaction On Information Forensics And Security", volume 8 & No2,February 2013.