ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

An Enhancement in Privacy Preserving Techniques for Collaborative Data Publishing

Aseema Jana

PG Scholar, Dept. of Computer Engineering, Dhole Patil College of Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

ABSTRACT: In recent years sharing of data becomes a significant activity and privacy takes a major role in securing the data from various attackers. Publication of shared data is either motivated by public benefit or by systems like NHIN i.e. nationwide health information network and Center for Disease Control and Prevention, who collects information from various health organizations and provides the same to health professionals and researchers to understand and address particular diseases more effectively. Preserving the privacy of individuals while publishing their linked data is a vital problem as original data contains sensitive information of individuals and publishing such data will violate individual privacy. There are chances of attacks when we are publishing collaborative data to various data providers. Attack can be done by people who are not part of data sharing, called as outsider attack and attack can also be done by those people who are part of this data sharing system, this is called as insider attack, where data providers themselves uses their own data to infer the data of other data providers. The paper discusses approaches to overcome these attacks by combining slicing and m-Privacy techniques. Implementation on Hive which is implemented on HDFS helps to reduce the computation time and Use of pattern matching algorithm will protect the system from sql injection attacks.

KEYWORDS: Anonymization, Collaborative publishing, Privacy, slicing, Hive, Sql injection attack.

I. INTRODUCTION

Generally health related data is gathered from different places like government, Health organizations and corporations. Further this data is used for giving health solutions, research and analysis purpose. As original data contains sensitive information of individual so directly releasing such information for above discussed purposes, may breach the privacy of individual. So anonymization techniques are used before releasing the data and this procedure is called as privacy preserving data publishing. Anonymization categorize the data as identifier, Quasi Identifier (QI) and Sensitive Attributes (SA) [8].Identifier is the key attribute which uniquely identifies a person such as SSN, name and this attribute is removed from data record before publishing. QIs are part of information which is well correlated with an entity and can create a unique identifier when combined with other QI, e.g. birth date, gender, zip code.SA includes sensitive information of an individual which may breach individual privacy if published, e.g. diseases and salary details. So objective is to secure individual's sensitive data from malicious users/attackers and preserving the privacy of individuals by using different techniques.

Section II presents the related work. In section III, the proposed approach and design is depicted. Section IV and Section V describes the Implementation details and results respectively. Section V covers conclusion and future scope of project.

II. RELATED WORK

In current years privacy preserving data study and shared data publishing has emerged as a promising approach which helps to preserve privacy of individuals. B.C.M. Fung et al [8] has given a survey on privacy preserving data publishing which gives different technique and tools for publishing the data while preserving the privacy of data. He has described different linking attacks where attacker is able to link owner of a record to a record in a published table, to a SA in published table, or to the published table itself. These are called record linkage, attribute linkage, and table

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

linkage, respectively. d-Presence protects the system from table linkage. K-anonymity prevents record linkage and it says that if any record in table has some value qid then at least k-1 record should also have the same value qid. L-diversity prevents attribute linkage and according to this concept each QI group should include at least l well represented SAs.

N. Mohammed et al [7] proposed a model for high dimensional relational data for healthcare system, called LKC privacy model. Model gives improved outcomes than conventional k anonymization model. This privacy model considers only relational data and data of healthcare is relatively complex, it can be the combination of relational data, transaction data and textual data. Privacy model works for centralized anonymization (anonymize and aggregate) and distributed anonymization (Aggregate and anonymize).

Alberto et al [6] has given the concept of privacy preserving updates to anonymous and confidential databases and developed a system to check whether the database inserted with record is still k-anonymous, without letting owner know, the contents of record and the database, respectively. To overcome this problem, two protocols have been proposed; first protocol is based on suppression and second is based on generalization and confidential databases.

Safe realization of the generalization privacy mechanism was given by Tristan Allard et al [5] .This paper focuses on the organization of the collection and anonymization phases at the data source (i.e., at each SPT) while compromising neither privacy nor data utility compared to a trusted central server approach. The problem is difficult due to three assumptions: (1) the data publisher and the data recipients are untrusted, (2) the SPTs are trusted but there is no direct communication between them and (3) there is no certainty about the connection frequency and duration of each SPT connection. This system focuses on this problem and proposes conventional Generalization privacy method which is composed of huge set of tamper-resistant smart portable tokens and this is connected to infrastructure that is well available. This conjunction of hypothesis makes the problem fundamentally different from any previously studied privacy-preserving data publishing problem we are aware of.

Tiancheng Li et al [4] has given a new approach for privacy preserving data publishing called Slicing which partitions the data both horizontally and vertically. This approach shows that slicing preserves better data utility than generalization and can be used for membership disclosure protection and it can handle high-dimensional data. The fundamental scheme of slicing is to break the association cross columns, and reserve the association within each column. This minimizes the dimensionality of the data and reserves better utility than generalization and bucketization. Slicing, groups highly correlated attributes together, that's why preserves utility and preserves the correlations between such attributes. Slicing protects privacy because it breaks the associations between uncorrelated attributes, which are infrequent and thus identifying.

When more number of similar attribute value and the sensitive value are present in the different tuples at that time slicing can give the original record while performing the random permutation. To overcome the drawback of slicing S. Kiruthika et al [2] has given enhanced slicing model. In this model suppression slicing is done by suppressing any one of the attribute value in the record and then perform the slicing. Thus utility is maintained with minimum loss by suppressing only very few values and privacy is maintained by random permutation. The next model is Mondrian slicing in this the random permutation is done with all the buckets not within the single bucket. Thus same utility of the original dataset is maintained.

M.KarthiKeyan et al [3] discusses Aho Corasick pattern matching algorithm for detecting and preventing the attacks related to sql injection. This is the pattern based technique used for static analysis. Here author has considered some standard attack pattern and shows that how this algorithm work against sql injection attack.

B. C. M. Fung et al. [1] has proposed an approach for handling 'insider attack'. In this attack data providers themselves uses their own data to infer the data of other data providers. Author has used m-privacy techniques and heuristic algorithms for overcoming the same.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

III. PROPOSED APPROACH AND DESIGN

A. Problem Definition

Attacks are the major problem in collaborative data publishing, additionally existing system has problems like single sensitive attribute, data loss/waiting, sql injection attack and more computation time. So goal is to publish an anonymized view of integrated data which is resistant to internal and external attack, reduce the computation time of system and take care of data loss.

B. Proposed Architecture and Design

The proposed system provides a competent approach to achieve enhanced privacy for collaborative data publishing and it overcomes the problems of existing system. Architecture follows Decentralized Anonymization approach i.e. aggregate and anonymizes approach and data publishing by using this approach is called as collaborative data publishing. In this approach data is first aggregated from different providers as P1,P2,P3..Pn and then data anonymization takes place. Left part of figure shows anonymization process. It takes the records from database; perform slicing on that, after that data is checked against privacy constraints. Resultant records are verified using Fscore algorithm and then final output is used for data publishing. Right part of figure shows search process, in this search process query is checked using pattern matching algorithm and if it's found malicious then user is considered as attacker and his information is sent to admin, else search information is provided to user/doctor as per specialization.

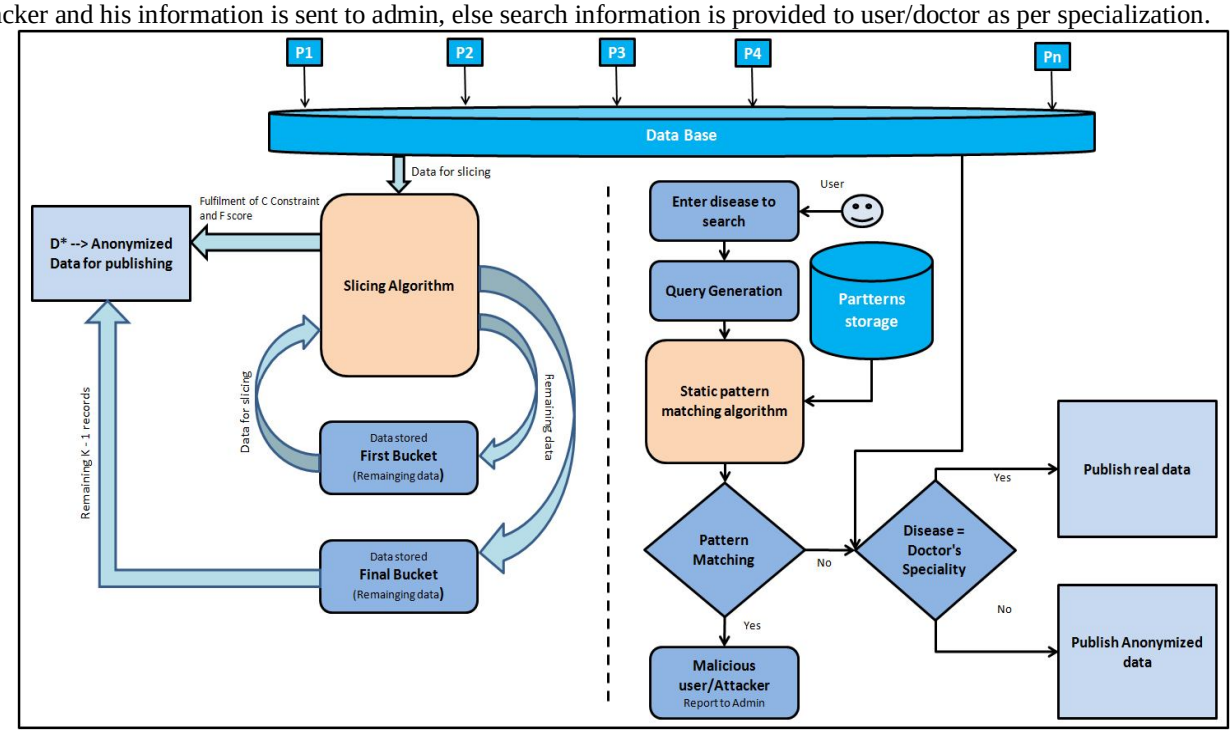


Figure 1: Proposed System Architecture

IV. IMPLEMENTATION DETAILS

System is implemented using 5 modules i.e. Main module, Provider module, Doctor Module, Admin module and Analysis module. Main module establishes the connection between application and database and initiates the process of other modules. In Provider module data providers are entering patient related details through patient registration form and this information is taken as an input to system for data publishing. Admin has access to original data records and can see the attack logs for detailed information. Doctor module can also be called as anonymization module. In this module slicing algorithm [4] is used which works in phases as Attribute Partition and Columns, Tuple partition and Buckets, Random permutation and Column Generalization. After partitioning the data record, privacy constraints are

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

applied for overcoming inference attacks which is next phase of this module and then data is getting verified using Fscore algorithm.

1. Attribute Partition and Columns: Attribute Partitioning is done by grouping highly correlated attributes into a column.
2. Tuple Partition and Buckets: A tuple partitioning is done by grouping the tuple in bucket. Each tuple belongs to exactly one subset and this tuples subset is called a bucket.
3. Random Permutation: Random permutation is done by randomly permuting or sorting each column values within each bucket. This permutation helps in breaking the linkage between columns. In project, random permutation is performed on sensitive attribute disease.
4. Column Generalization: Column generalization guarantees that one column satisfies the k-anonymity requirement. It is a multidimensional encoding and can be used as an additional step in slicing.
5. Privacy Constraints: It checks the privacy constraint of each record which follows k-anonymity and l-diversity of records. L diversity is the theory of sustaining uniqueness within data. In this system we used this concept on sensitive attributes. Our anonymized bucket size is 6 means K=6 and maintained L = 4 i.e. from 6 SA record 4 must be unique.
6. Fscore: After getting the anonymized view of data through above algorithm, it is verified through Fscore algorithm. Fscore algorithm is privacy fitness score which shows the level of fulfillment of privacy constraint C. If Fscore is greater than 1 than anonymized data is verified.

Step 5 is performed in iterative manner till all the records satisfy the privacy constraint and sent for data publishing but the last k-1, L<4 records which are not able to satisfy the constraints goes to data loss/waiting state for user. So to avoid this waiting situation in data publishing, one more additional bucket is used which contain mentioned records and published this in anonymized form. Additionally system uses Hive database, which is implemented on HDFS, it reduces the computation time of system while search and use of Aho Corasick pattern matching algorithm[3] protects the system from sql injection attack.

V. RESULTS

Below screenshots show the results. Figure 2 and figure 3 shows the anonymized view for data publishing which is resistant to attack and follow the methodology as proposed. Figure 4 shows the search result when asthma disease is searched and next figure 5. Shows the comparison how use of hive reduces the computation time as compared to sql in search, when large numbers of data records are there. Last two figures i.e. figure 6 and figure 7 show how system identifies the sql injection attack and generate detailed report for admin.

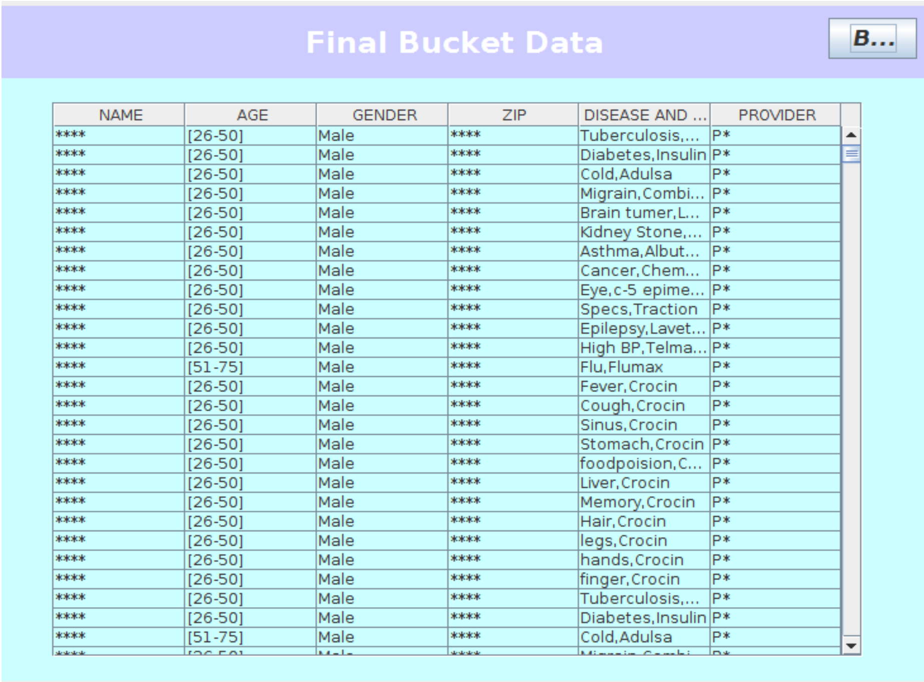
Figure2: Data view for data publishing

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

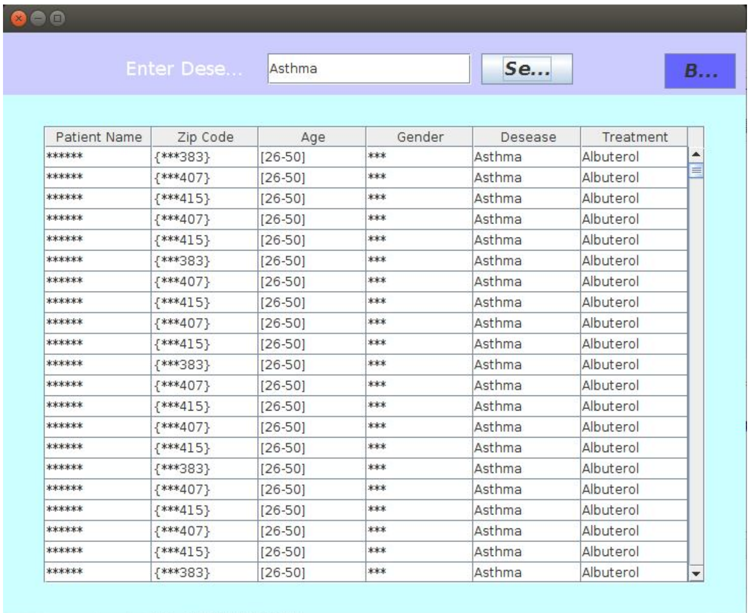
(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015



NAME	AGE	GENDER	ZIP	DISEASE AND ...	PROVIDER
****	[26-50]	Male	****	Tuberculosis,...	P*
****	[26-50]	Male	****	Diabetes,Insulin	P*
****	[26-50]	Male	****	Cold,Adulsa	P*
****	[26-50]	Male	****	Migrain,Combi...	P*
****	[26-50]	Male	****	Brain tumer,L...	P*
****	[26-50]	Male	****	Kidney Stone,...	P*
****	[26-50]	Male	****	Asthma,Albut...	P*
****	[26-50]	Male	****	Cancer,Chem...	P*
****	[26-50]	Male	****	Eye,c-5 epime...	P*
****	[26-50]	Male	****	Specs,Traction	P*
****	[26-50]	Male	****	Epilepsy,Lavet...	P*
****	[26-50]	Male	****	High BP,Telma...	P*
****	[51-75]	Male	****	Flu,Flumax	P*
****	[26-50]	Male	****	Fever,Crocin	P*
****	[26-50]	Male	****	Cough,Crocin	P*
****	[26-50]	Male	****	Sinus,Crocin	P*
****	[26-50]	Male	****	Stomach,Crocin	P*
****	[26-50]	Male	****	foodpoision,C...	P*
****	[26-50]	Male	****	Liver,Crocin	P*
****	[26-50]	Male	****	Memory,Crocin	P*
****	[26-50]	Male	****	Hair,Crocin	P*
****	[26-50]	Male	****	legs,Crocin	P*
****	[26-50]	Male	****	hands,Crocin	P*
****	[26-50]	Male	****	finger,Crocin	P*
****	[26-50]	Male	****	Tuberculosis,...	P*
****	[26-50]	Male	****	Diabetes,Insulin	P*
****	[51-75]	Male	****	Cold,Adulsa	P*

Figure3: Data view for data publishing(Final bucket)



Patient Name	Zip Code	Age	Gender	Desease	Treatment
******	{***383}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***383}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***383}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***383}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***407}	[26-50]	***	Asthma	Albuterol
******	{***415}	[26-50]	***	Asthma	Albuterol
******	{***383}	[26-50]	***	Asthma	Albuterol

Figure4: Search for disease Asthma

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

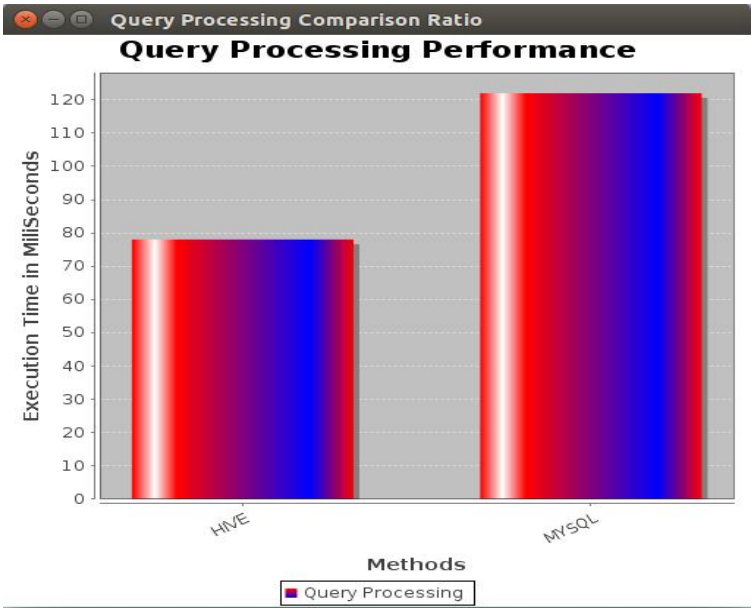


Figure5: Computation time comparison in search while searching Asthma

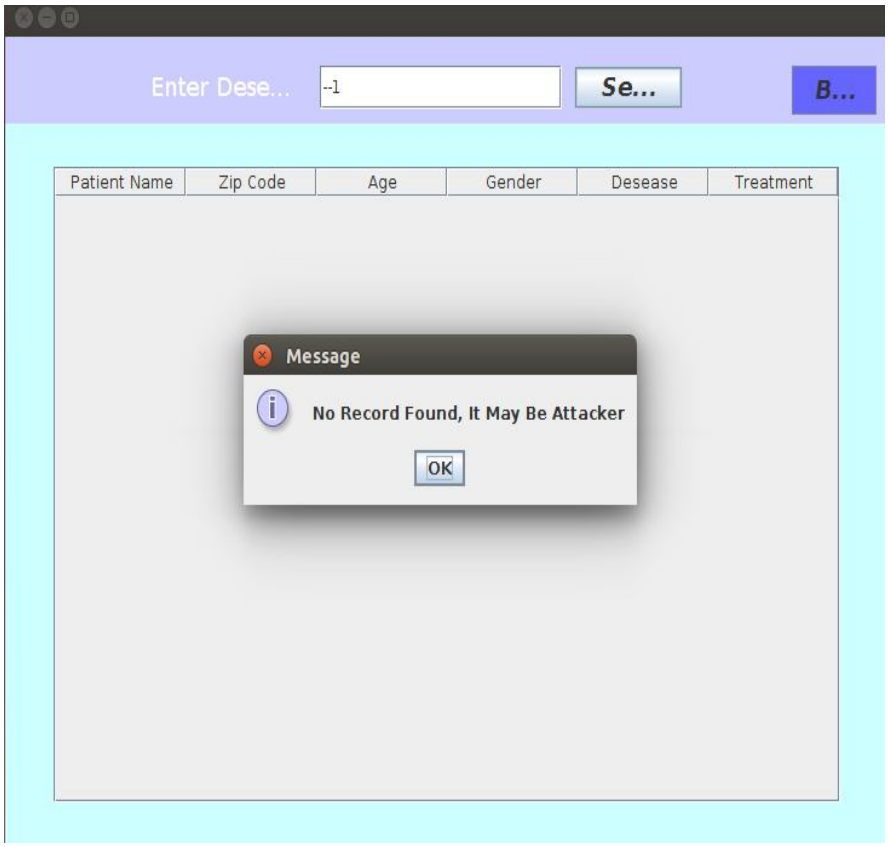


Figure6: Sql Injection attack

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

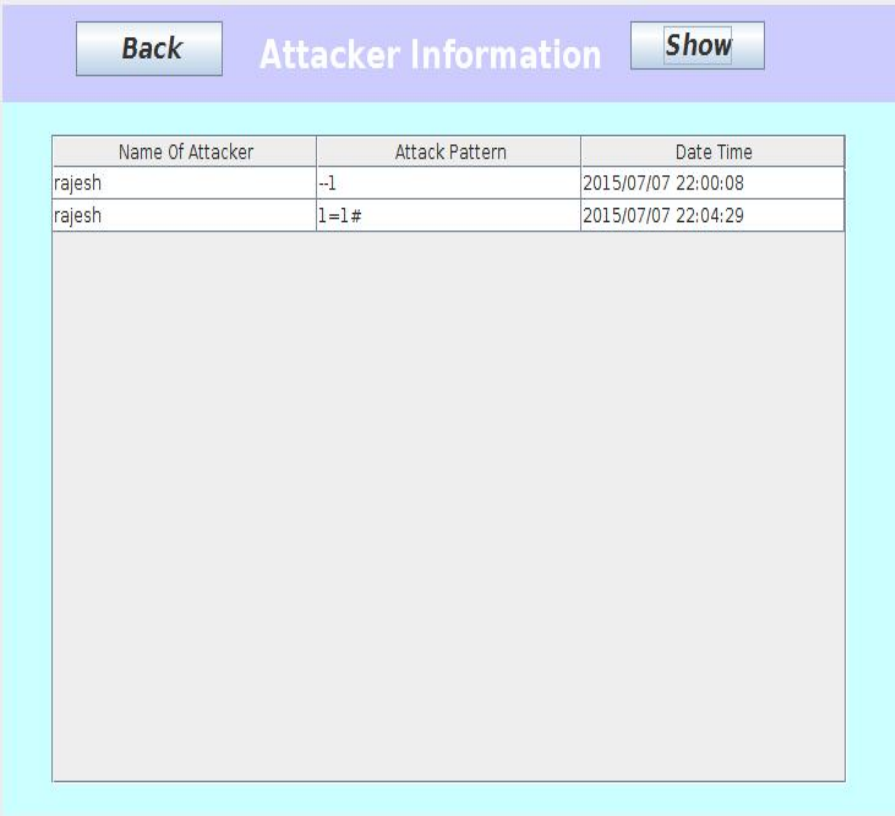


Figure7: Report generated for admin which contain Attacker information

VI. Conclusion and Future Scope

Here different categories of attacks on collaborative data publishing has been considered and for overcoming the same, combination of slicing technique with m-privacy techniques is used. Slicing provides the anonymized view of data and m-privacy assures that anonymized data satisfies a given privacy constraint against any collection of data providers and verifying the anonymization by using fitness score. Use of Aho Corasick pattern matching algorithm protects the system form sql injection attacks and enhanced search method reduces the computation time and increases the efficiency of system. In future, method can be generalize for different type of data, set valued data and can apply the techniques in scenario when data is distributed in ad-hoc manner.

VII. Acknowledgement

I would like to thanks my guide for the guidance and constant supervision. My sincere thank to Dhole Patil College of engineering for providing a strong platform to develop the skills and capabilities.

References

1. S. Goryczka, L. Xiong, and B. C. M. Fung, “m-Privacy for collaborative data publishing”, IEEE transactions on knowledge and data engineering,vol.26, no.10,oct 2014.
2. S.Kiruthika and Dr. M.Mohamed Raseen “Enhanced Slicing Models For Preserving Privacy In Data Publication”, ICCTET, 2013.
3. Dr.M.Amutha Prabakar, M.KarthiKeyan, Prof.K. Marimuthu, “An efficient technique for preventing Sql injection attack using pattern Matching algorithm” 2013 IEEE ICECCN 2013
4. Tiancheng Li, N inghui Li, Senior Member, IEEE, Jian Zhang, Member, IEEE, and IanMolloy “Slicing: A New Approach for Privacy Preserving Data Publishing” IEEE Transactions on knowledge and data engineering, vol. 24, no. march 2012.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 6, June 2015

5. Tristan Allard, Benjamin Nguyen, Philippe Pucheral, "Safe Realization of the Generalization Privacy Mechanism" Privacy, Security and Trust (PST), Ninth Annual International Conference July 2011.
6. Alberto Trombetta, Wei Jiang, Elisa Bertino, Lorenzo Bossi "Privacy-Preserving Updates to Anonymous and Confidential Databases" in IEEE TRANSACTIONS ON DEPENDABLE AND SECURE COMPUTING, VOL. 8, NO. 4, JULY/AUGUST 2011.
7. N. Mohammed, B. C. M. Fung, P. C. K. Hung, and C. Lee, "Centralized and distributed anonymization for high-dimensional healthcare data," ACM Trans. on Knowl. Discovery from Data, vol. 4 no. 4, pp. 18:1–18:33, October 2010.
8. M. Fung, K.Wang, R. Chen, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," ACM Comput.Surv., vol. 42, pp. 14:1–14:53, June 2010.