Privacy Preserving e-Health Data Sharing in Cloud
Extended utilization of anonymization with random effect

Privacy Preserving e-Health Data Sharing in Cloud

Extended utilization of anonymization with random effect

Mahmood Ahmad[1], Zahid Anwar[2], Sungyoung Lee[1]
[1]Dept. of Computer Engineering, Kyung Hee University, South Korea
e-mail : {rayemahmood, sylee}@oslab.khu.ac.kr
[2]National University of Sciences and Technology(NUST-SEECS), Pakistan
e-mail : zahid.anwar@seecs.edu.pk

Abstract

IT infrastructure has a verity of applications running on it. Due to voluminous growth of applications and associated data, usage of cloud services is inevitable. Depending upon the type of information that travels to and fro between consumer and service provider demands certain parameters on security and privacy. In recent years, applications in health domain have grabbed attention of its users and researchers in multiple directions. Amongst all these, concerns related to the outsourced data in cloud environment have serious challenge like secure data sharing and its effective utilization. Existing methodologies like encryption and anonymization are effective in certain circumstances. Preventing unauthorized access can be secured using these techniques, however to safeguard outsourced information these techniques sometimes cannot withstand against threats, like record linking and unsolicited disclosure of information in untrusted domain of public cloud. A reliable reach to secure private information is the essential goal. Most of the healthcare organizations either have their own custom solutions developed or third part CRM software tailored to their needs. Emergence of cloud computing has provided us the opportunity to take a step ahead and build applications available to users on a secure platform, with ability to scale the resources depending on the requirements. In our proposed model for outsourcing the health data in public cloud in integration with private cloud escalates data utility and alleviate burden for record linkage problem. Our solution protects against any eves dropper for identifying any patterns out of the communication that takes place between consumer and its owner.

1. Introduction

Smart CDSS is a cloud-based service that accepts user input as activity, social interaction, and clinical information in standard HL7 vMR format. This trove of information can be used as a knowledge repository for generating recommendations by domain experts. Besides patients, external EMR, EHR, computerized physician order entry system (CPOE) use smart CDSS according to their needs and demands. Due to contents secrecy in this information, in 1966 The Health Insurance Portability and Accountability Act (HIPPA) introduced regulations [5]on the privacy breach of this information. Keeping in view the enforcement policy of this act, information related to patients has to be protected not only from unauthorized access, but additional slipway of information has to be ceased. For this reason, mechanisms like encryption [6] and anonymity [1, 2] are widely adopted options to preserve the privacy aspect of this information. While preserving the privacy of this information, its utility may go down [4], therefore a tradeoff between privacy and security exist side by side. Usage frequency, massive volume, and dynamic access by a range of users, cloud environment is an inclined trend for housing applications like Smart CDSS.

Cloud computing is an epitome of on-demand computing. It provides virtualized computing resources (i.e., processing power, storage facility, application services) over the internet. However, as cloud is owned and managed by a third party, the risk of privacy infringement escalates when confidential data of the proposed Smart CDSS is outsourced to an un-trusted domain of cloud service provider. If encryption is applied to protect this outsourced data, it can minimize its utilization (i.e. searching and processing). To overcome this issue either data key is shared with cloud service provider (CSP) or it is downloaded by authorized user, decrypted and then processed according to one's wish. Both approaches have their own implications. First approach can reveal everything about data to the CSP and in second approach, even for a smaller segment of information complete data set has to be downloaded first. The situation becomes more complex when frequency of data updates is high. To avoid overly usage of network bandwidth and privacy issue, data can be processed through anonymization before publishing it publically. Anonymization ensures that at any individual record in total 'K' records is different from at least K-1 other records with respect to their quasi-identifiers [1, 2]. This approach is least expensive, however it is more prone for record linkage attacks. Identification of individuals through linking attack has been

explained in [3]. Out of two different sources of information where one is anonymized and other is not still 69% of individuals are not safe against linking attack and that too with just a combination of date of birth and zip code. Discouraging linking attacks and to preserve user privacy is prime focus of this paper.

Encryption is a good option to protect data against unauthorized access but at the cost of utility. Other than encryption, using anonymization can provide data utility but is prone to record linkage problem. Tweaking data utility and privacy issues while using anonymization, is another challenge. Keeping in view limitations associated with these two techniques we have come up with a solution that can be used where encryption is an expensive option and chances for record linking is inevitable if anonymization is used instead. Unlike anonymity, where data is suppressed or generalized, we further extend the existing methodology of anonmization by adding randomization. Randomization ensures that the overall privacy of outsourced data remains intact yet it is also very minimal on computation cost. Using proposed methodology on outsourced data in untrusted domain of cloud, additional slip-away of information can also be saved.

2. Working Methodology

After selecting the data, that has to be outsourced, it is processed with generalization and suppression

Table(a)

User SSN	Age	Disease	Gender	ZipCode
SS-1234	23	Cancer	Male	12563
SS-3452	25	Cancer	Female	12514
SS-2311	36	Asthma	Female	12401
Ss-1222	41	Diabetes	Male	12401

1

Table(b)

User SSN	Age	Disease	Gender	ZipCode
*	21-30	55	M	125*
*	21-30	55	F	125*
*	31-40	12	F	124*
*	41-50	32	F	124*

2

Table(c)

User SSN	Age	Disease	Gender	ZipCode
*	89	26	11	232
*	25	35	32	111
*	12	12	11	100
*	10	9	15	24

3

Table(d)

User SSN	Age	Disease	Gender	ZipCode
*	89	(x-26)(x-81)	11	232
*	25	(x-35)(x-90)	32	111
*	12	(x-12)(x-24)	11	100
*	10	(x-9)(x-41)	15	24

Figure 1: Table Transformation

wherever found necessary. The steps involved in this process are shown in Figure 1. For explanation, we have used a table that is transformed from Table (a) to Table (d). Table (a) represents the data in its initial stage. Table (b) is where transformation and generalization is applied on columns, age, disease, gender and Zip code. The user security number which uniquely identifies the patient is removed and is not shared with the outsourced data. In Table (c), value from each column is replaced with random number. To construct Table (d), mapper table is used which is shown in

Disease	Mapper value
Cancer	55
Asthma	12
Diabetes	32
Hair loss	65

Figure 2: Mapper Table

Figure2. The mapper table is a way to represent mapping that are used to construct Table (d) which is definitely constructed using standard codes. It is assumed that standardization of these codes has been done which is utilized here. All random values which are generated in Table (c) are added with these mappings and a composite polynomial table is constructed in such way that first valid root of this polynomial is random number. Second valid root of this polynomial is (random number + mapper value). Using this approach will further obfuscate same values with different polynomial expressions. Similarly all column values are treated the same way. The final version of transformation which is Table (d) is then outsourced on public cloud. Mapper table is preserved at private cloud. We have ignored the process of user authorization, however to access this information valid credentials are required. The CSP who is now hosting the data, cannot infer or deduce any knowledge from it.

2.1 Query Formulation- by user. Next step starts from user end and it is assumed that user is authorized and has valid credentials to access the data. For this reason, we will not discuss how this evaluation process will take place between the CSP and user. Schema of information which is meant to be accessed by that user is also known to him as it is shared during the user registration process. The user formulates the query and it is sent to the private cloud where it is modified with effect of randomization using the mapper table. The user query looks as shown below

q:1=Select count(disease) from table-name where disease="cancer"

This query is forwarded to the private cloud.

2.2 Query Formulation-by private cloud.

Private cloud is managed by the data custodian and will transform the user query such that it can be evaluated on public cloud. For this purpose parameter list from where clause is extracted. This parameter list is then compared with the mapper value and query will appear as

q-2 **:** *Select count(disease) from table-name where*

disease=”55”

Here, 55 is the code value of cancer which is taken from mapper table. A random ‘r’ value is generated again at private cloud such that where clause is now transformed completely and will look like as given in q:3

q-3 **:** *Select count (disease) from table-name where disease=”(x-r)(x-(r+55))”*

This query is then sent to public cloud. Information given in where clause of this query is altogether different what initially it had in q1.
For further protection on column names, name of columns can be replaced with random numbers just like the actual values to further hide additional details for CSP to know.

2.3 Query negotiation and response extraction.
The service of query evaluator will intercept the query and will prepare it for data retrieval. Services of third party are used for this query negotiation. In this negotiation the private cloud will share one valid root of all polynomials with trusted third party and likewise third party also shares one valid root of where clause from q:3. CSP will evaluate composite polynomials in disease column using this one root received by third party and forwards its result to third party. Meantime third party has also evaluated its results and will compare these results with those, that are shared by CSP. In case of exact match, row ids for these records are requested to be retrieved from CSP. These results are then shared with the user with exact count of cancer disease.

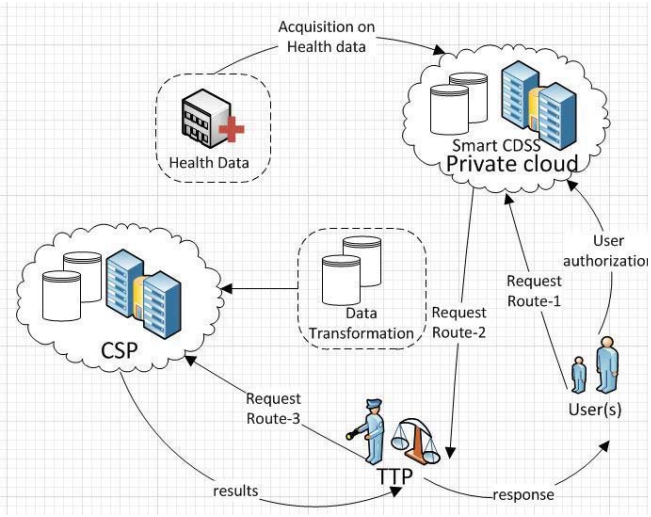

Figure 3: Proposed System Setup

Overall system architecture and involving entities are shown in Figure 3.

3. Conclusion
In this paper we have focused the problem related with health data outsourcing and possible disclosure of information with its usage. The proposed idea in synergy with anonymization further strengthens the overall communication between the stakeholder and consumer. Patterns associated with user query can reveal frequency of user request and associated output as a result, however incorporating randomization and using mapper tables can conceal the additional leakage of information. Maintaining multiple keys for different users or using encryption mechanism again and again for outsourced data is an expensive operation where frequency of update is very high. Affording a mapper table instead, is least demanding in terms of updates and is secure because of its localization in private cloud as well. Different channels of communication may slow down the overall process but it protects adequate secrecy of outsourced data.

4. Acknowledgement
This research was supported by the MSIP(Ministry of Science, ICT&Future Planning), Korea, under the ITRC(Information Technology Research Center) support program supervised by the NIPA(National IT Industry Promotion Agency)" (NIPA-2013-(H0301-13-2001)).

References

[1] P. Samarati, L. Sweeney, Generalizing data to provide anonymity when disclosing information, Proceedings of ACM Symposium on Principles of Database Systems (PODS), 1998, p. 188

[2] L. Sweeney, K-anonymity: a model for protecting privacy, International Journal on Uncertainty, Fuzziness and Knowledge-based Systems, 2002, pp. 557–570.

[3] Winkler, William E, Overview of record linkage and current research directions, Bureau of the Census, 2006

[4] N. Li, T. Li, t-Closeness: privacy beyond k-anonymity and l-diversity, Proceedings of the International Conference on Data Engineering (ICDE), 2007, pp. 106–115.

[5] Annas, George J HIPAA regulations-a new era of medical-record privacy, New England Journal of Medicine, medical publishing group-mass medic society, 2003

[6] Chow, Richard and Golle, Philippe and Jakobsson, Markus and Shi, Elaine and Staddon, Jessica and Masuoka, Ryusuke and Molina, Jesus, Controlling data in the cloud: outsourcing computation without outsourcing control, Proceedings of the 2009 ACM workshop on Cloud computing security, 2009