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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

An Medical Healthcare Information Sharing With Service Based Cloud Environment

S Santhiya[1], Dr M Karthikeyan[2]

Research Scholar, PG and Research Department of Computer Science, Government Arts College (Autonomous), Salem, India[1]

Assistant Professor, PG and Research Department of Computer Science, Government Arts College (Autonomous), Salem, India[2]

ABSTRACT: Electronic Medical Records (EMR) has been introduced in recent years. Medical Institutions and healthcare providers are required to store electronic records in a database and provide access for doctors and researchers. EMR provide health information on as a needed basis for diagnosis and treatment. EMR provide convenience, but such a system also introduces the new challenges of storing personal information securely. Based on personal information a specific person can be identified or quasi-identified. For preventing disclosure of person specific information, usually quasi-identified or anonymized data are published. *k*-Anonymity framework with generalization and suppression anonymizes the values of the quasi-identifiers which are provided in the EMR. But still *k*-Anonymity framework has some drawbacks which include re-identification attack. To address this issue, a framework combining *k*-Anonymity and *l*-diversity has been proposed. Each equivalence class of anonymized data contains at least *l* well represented values in the sensitive attributes. While using this approach re-identification attack can be reduced and information loss is also minimized.

KEYWORDS: privacy, Cloud, anonymization, EMI, EMR, Security

I. INTRODUCTION

Cloud computing is the use of computing resources (hardware and software) that are delivered as a service over a network (typically the Internet). The name comes from the common use of a cloud-shaped symbol as an abstraction for the complex infrastructure it contains in system diagrams. Cloud computing entrusts remote services with a user's data, software and computation. Cloud computing consists of hardware and software resources made available on the Internet as managed third-party services. These services typically provide access to advanced software applications and high-end networks of server computers.

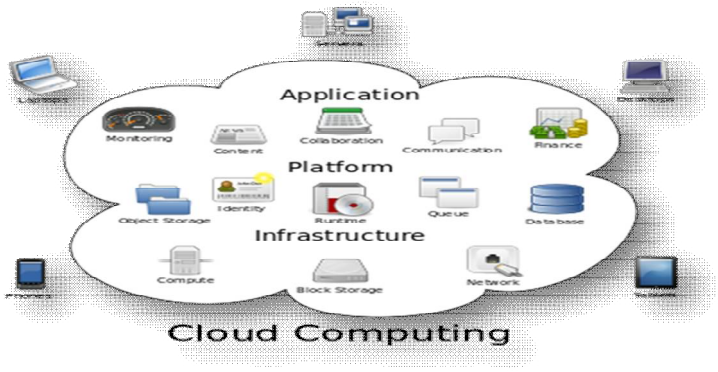


Figure 1.1 Structure of cloud computing

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Privacy is becoming an important issue when a researcher uses the data which contains the individual's sensitive information. Nowadays data collections have become easier and techniques to find the original data are becoming more efficient. In recent years, there have been many improvements to protect the privacy of an individual. The data which are published for research must be deidentified that is the identity of the individuals should be removed. The main aim of Privacy Preserving Data Publishing is to guarantee anonymity by means of controlled transformation of data.

The Public data which are published includes data's from Public Census Records, Electronic Medical Records, Salary Records, Web Search, Online Social Networks etc. The data publishing process includes various persons such as the individual from whom data is collected, the collection agent, the adversary, the end user who uses the data.

Re-identification problems create from linking attack, patients could be distinctively re-identified using the hospital summaries of demographics as date of birth, zip code and gender by linking publicly available voter registration lists.

An Electronic Medical Record (EMR) is a collection of patient record created in hospitals; it may include a range of data, demographics, medical history, medication and allergies, immunization status, laboratory test results, radiology images, vital signs, personal statistics like age and weight, and billing information. These Electronic Medical Records can be in the form of Relational Data, Transactional Data, sequential Data and Text Data.

The EMR dataset contains trajectory data, which is the combination of diagnosis code or disease with Age. The record with Age is said to be trajectory data because the values are recorded over a period of time. The diagnosis code here is referred as International Classification of Disease (ICD code). The ICD code is unique for each disease and the hierarchy of this ICD code is used for anonymization purpose.

Patient Privacy in publishing EMR is a growing concern for Healthcare Organizations. Safeguarding the confidentiality, integrity and availability of patient information is no longer a goal – it is a legal requirement [1].

To solve this concern Healthcare Providers must have secure access to clinical applications and protect the underlying IT infrastructure from misuse by insiders, hackers and identity thieves.

II. LITERATURE REVIEW

Anonymization is a process of removing or modifying the identifying variables contained in the dataset. Anonymization of data consists of two steps, first, the potential identifiers should be identified and removed, and it depends on one's personal judgement and then modifies the accuracy of the identifiers to reduce the risk of re-identification attack.

Sahai et al. proposed a scheme for ensuring data storage security in untrusted cloud [1]. Unlike most prior works for ensuring remote data integrity, the new scheme supports secure and efficient dynamic operations on data blocks, including: update, delete and append. Extensive security and performance analysis shows that the proposed scheme is highly efficient and resilient against Byzantine failure, malicious data modification attack, and even server colluding attacks. The straightforward and trivial way to support these operations is for user to download all the data from the cloud servers and re-compute the whole parity blocks as well as verification tokens. The user can always ask servers to send back blocks of the rows specified in the challenge and regenerate the correct blocks by erasure correction.

Lewko et al. proposed a scheme for enabling public verifiability and storage dynamics for cloud computing [5]. They have proposed a general formal PoR model with public verifiability for cloud data storage, in which both block less and stateless verification. The challenge-response protocol can both determine the data correctness and locate possible errors. They employ authenticated skip list data structure to authenticate the tag information of challenged or updated blocks first. It providing integrity verification under different data storage systems, the problem of supporting both public verifiability and data dynamics has not been fully addressed. The verification algorithm accepts when interacting with the valid prover (e.g., the server returns a valid response) and it is sound if any cheating server that convinces the client it is storing the data file is actually storing that file.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

V. Goyal et al. developed a new cryptosystem for fine-grained sharing of encrypted data that we call Key-Policy Attribute-Based Encryption (KP-ABE) [2]. In our cryptosystem, cipher texts are labeled with sets of attributes and private keys are associated with access structures that control which cipher texts a user is able to decrypt. It has limited applicability to access control of data, our primary motivation for this work. One drawback of encrypting data, is that it can be selectively shared only at a coarse-grained level. We demonstrate the applicability of our construction to sharing of audit-log information and broadcast encryption. Our construction supports delegation of private keys which subsumes Hierarchical Identity-Based Encryption (HIBE). As more sensitive data is shared and stored by third-party sites on the Internet, there will be a need to encrypt data stored at these sites (i.e., giving another party your private key). The cryptosystem of Sahai and Waters allowed for decryption when at least *k* attributes overlapped between a cipher text and a private key.

III. METHODOLOGY

3.1Anonymization Models

The Randomization Method

Randomization method is one of the privacy preserving technique in which a desired amount of noise is added to the current data in order to distort the original values of the records. $X = \{x_1, x_2,\ldots,x_n\}$ defines a set of records. For each record in X add a noise component $f_Y(y)$, which is drawn from the probability distribution .

The noise components are drawn independently which are denoted as $Y = \{y_1, y_2,\ldots y_n\}$, Tthe new set of distorted records are denoted as Z.

$Z = \{z_1,z_2,\ldots, z_n\}$ where $z_1 = \{x_1+y_1\},\ldots,z_n = \{x_n+y_n\}$.

The distorted data cannot be guessed because the large of variance added noise. So the original records can be recovered other than the distribution of the original record recovered. The diverse randomization methods are discussed below,

k-Anonymity

k-Anonymity is the property that each record have the same with atleast *k*-1 other records with respect to the quasi identifier [2]. In this method *k*-Anonymity states 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.

l-Diversity

A q* -block is *l*-diverse if contains atleast *l* “well-represented” values for the sensitive attribute S. A table is *l*-diverse if every q* -block is *l*-diverse [3]. Sensitive Attributes must be diverse within each quasi identifier equivalence class.

Domain Generalization Hierarchy (DGH)

A DGH of a attribute A, referred to as H_A, it is a partially ordered tree structure which defines valid mappings between specific and generalized values of A. The root of H_A is the most generalized value of A. The DGH for ICD and Age is given below in Figure 1 & 2.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

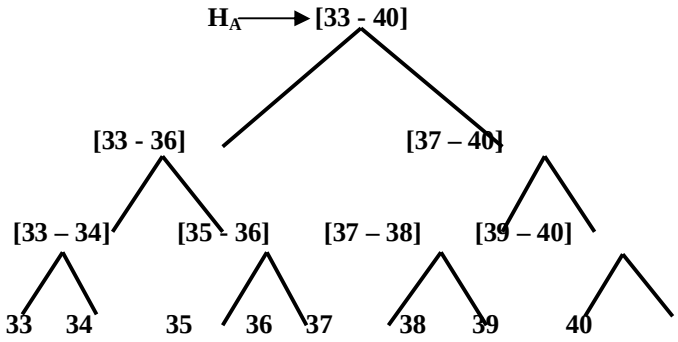


Figure 1 DGH of Age

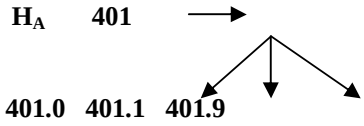


Figure 2 DGH of ICD code

K- Anonymity combined with l-diversity

There are many clustering algorithms like MDAV, k-means when applied to a dataset gives a *k*-anonymous data. This involves organizing records into clusters of size at least *k*, which are anonymized together.

The distance between two trajectories is defined as the cost of anonymizing them. Anonymization of two trajectories is achieved by finding a match between the pairs of trajectories that minimizes the cost of anonymization. After anonymization the dataset is verified to check whether it supports the principle of l-diversity.

k-Anonymity Model

In Associate in this earlier work, I introduced basic protection models termed null-map, k-map and wrong-map which offer protection by making certain that discharged data map or incorrect entities severally. To work out what percentage people every discharged tuble truly matches needs combining the discharged information with outwardly on the marker information and analyzing different attainable attacks. Making such a determination directly can be an extremely difficult task for the data holder who releases information. Although I can assume the data holder knows which data in PT also appear externally, and therefore what constitutes a quasi-identifier, the specific values contained in external data cannot be assumed. During this work by satisfying a rather totally different constraint on discharged data, termed the k-anonymity demand. This is a special case of kmap protection where k is enforced on the released data. A Data Anonymizer, a framework that incorporates alignment, clustering and l-diversity as separate components are modeled. Figure 3.1 gives the overall architecture of the data anonymizer.

Alignment

Alignment attempts to find a minimal cost pair matching between two trajectories. A simple heuristics called Baseline is developed for comparing the trajectories. For a given trajectories $X = \{x_1, x_2,\ldots,x_m\}$ and $Y = \{y_1, y_2,\ldots,y_n\}$ and DGHs of H_{ICD} and $H_{Age.}$ Baseline aligns X and Y by matching their pairs on the same index.

The steps involved in Baseline are, an empty trajectory T' is initialized to hold the anonymized trajectory, ILM(X) & ALM(X) is assigned to two variables. In shorter trajectories among those are determined and pair matching is performed based on their loss metric. It constructs a pair containing the LCA's of the ICD and Age and appends it to T'. Generalization is performed and the values for ILM & ALM are calculated. Then, the longer trajectories are determined and Suppression is performed for unmatched pairs of the ICD and Age. Finally, total loss metric is calculated and anonymized trajectory is formed.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

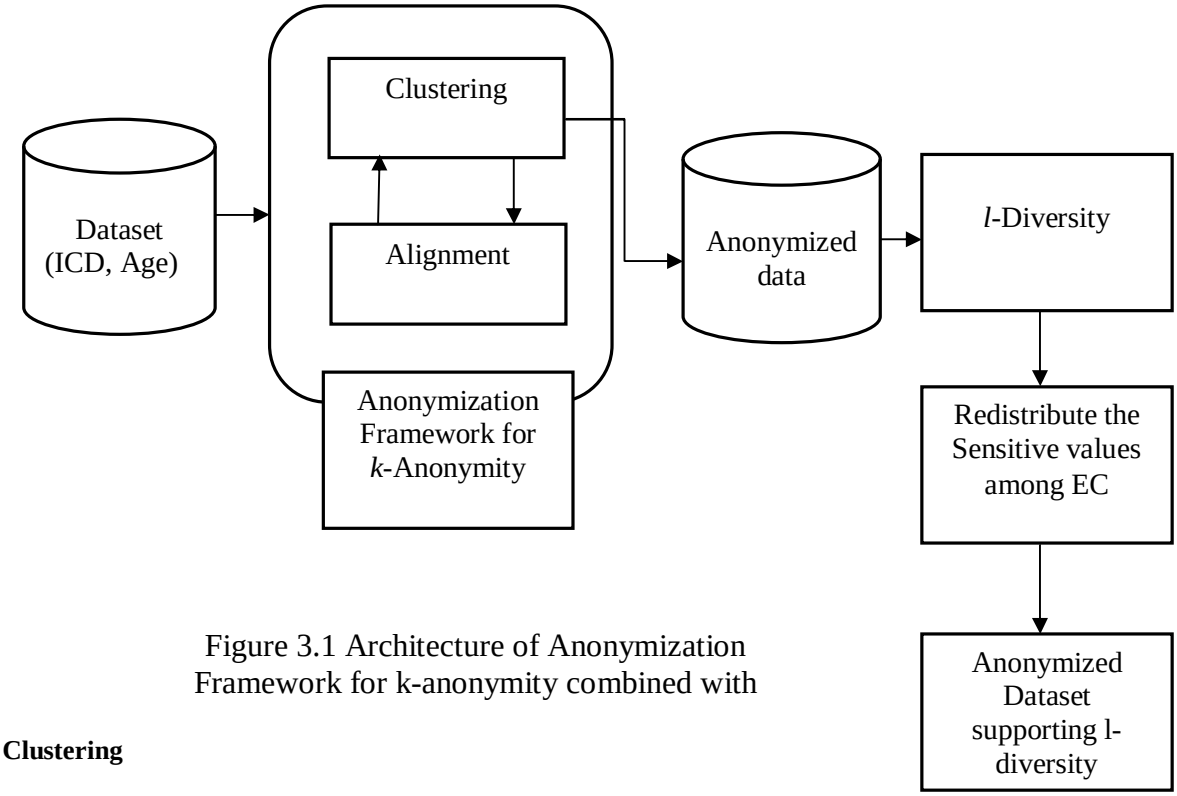


Figure 3.1 Architecture of Anonymization Framework for k-anonymity combined with

Clustering

Clustering interacts with the Alignment component to create clusters of at least *k* records. Clustering is based on the Maximum Distance to Average Vector algorithm which is a heuristic for k-anonymity named ad MDAV' with its helper function formCluster. MDAV' iteratively selects the most frequent trajectory in the dataset and find its distant trajectory and forms a cluster of k trajectories around the latter trajectory.

Cluster formation is performed by formCluster, a function that constructs a Cluster by aligning trajectories in a consecutive manner and returns the ILM and ALM.

***l*-diversity Model**

An anonymized block is *l*-diverse if it contains at least *l* "well-represented" values for the sensitive attribute. A table is *l*-diverse if every anonymized block is *l*-diverse. For a group of *k* different records that all share a particular QI, the attacker is interested on the sensitive values like medical diagnosis code, DNA sequence which are distributed evenly within the group, if the group contains distinct values, then the distribution of these values within a group is referred to as "*l*- Diversity".

The k-anonymized dataset is first checked for *l*-Diversity, sensitive attributes and set of quasi identifiers in the anonymized dataset is checked for the l-diversity principle. The information in every equivalence class is mapped to buckets. For every tuple, L(*t*) - list of statistics about the matching buckets are maintained. Each element in the list L(*t*) gets statistics about one matching bucket B. The matching probability p(*t*, B) and distribution of candidate sensitive value D(*t*, B).

p(*t*, B) is for a given tuple *t* and bucket B, the probability that *t* is in B depends on the fraction of *t*'s column values that match the column values in B.

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

$$p(t, B) = \frac{f(t, B)}{f(t)}$$

D(t, B) be the distribution of the candidate sensitive values in B.

Finally, p(t, s) is calculated which is the probability that t takes a sensitive values s. p(t, s) is calculated using the law of total probability.

$$p(t, s) = \Sigma_B \, p(t, B) \, p(s| \, t, B)$$

$$P(t, s) \leq 1 / l$$

An anonymized table is said to be l-diverse if and only if every tuple in it satisfies l-diversity.

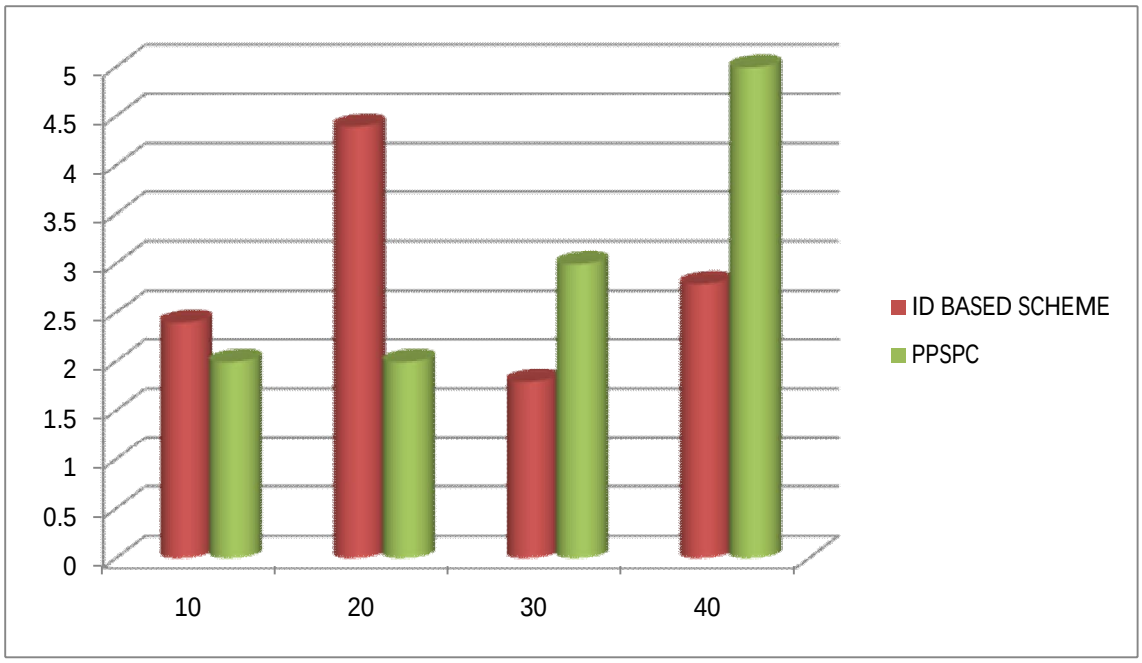
IV. RESULTS AND DISCUSSION

Performance Evaluation

Opportunistic computing:

In this framework consists opportunistic computing has gained the great interest from the research community recently, and we briefly review some of them related to our work. Introduce the opportunist computing k-anonymity and l-diversity combining cloud to unravel the matter of storing and execution an application that exceeds the memory resources. Especially, their resoulution relies on the thought of partitioning the appliance code into variety of opportunistically cooperating modules.

our proposed framework aims at the security and privacy problems, and develops a user-centric privacy access management of opportunistic computing in m-Healthcare emergency.



COMPARSION OF ID BASED SCHEME AND PPSPC

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

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Privacy-preserving scalar product computation: Research on privacy-preserving dot product computation (PPSPC) has been conducted for privacy-preserving data processing and likewise secure friend discovery in mobile social clouds quite recently at the start, PPSPC protocol was designed by involving a semi-trusted party. Later, to remove the semi-trusted party, several PPSPC protocols without a third party. In our planed SPOC framework, we tend to gift a replacement PPSPC protocol, which does not use any "homomorphic encryption", however is extremely economical in terms of computational and communication.

V. CONCLUSION

While applying k-anonymity the knowledge loss incurred by generalizing and suppressing the values of the attributes might cause less utility of data and sensitive values is also known. Once mistreatment *l*-diversity anonymization approach the data loss is also reduced and therefore the values within the sensitive attribute is distributed that is tough to link with alternative data.

REFERENCES

1. McAfee SIEM and FairWarning team(2012), ' Security and Privacy of Electronic Medical Records', Cost of Data Breach Study, United States, Benchmark research conducted by by Ponemon Institute LLC.
2. Latanya Sweeney(2002), '*k*-anonymity: A model for protecting privacy', International Journal on Uncertainity, Fuziness and Knowledge-based Systems, Vol.10(5), pp.557-570.
3. Machanavajjhala, A., Kifer, D., Gehrke, J., and Venkitasubramaniam(2007), '*l*-Diversity: Privacy beyond *k*-anonymity', ACM transactions on Knowledge Discovery from Data, Vol.1(1), Article No. 3.
4. Charu C. Aggarwal, Philip S. Yu(2008), 'A General Survey of Privacy Preserving Data Mining Models and Algorithms', Vol.34, pp.11 – 52.
5. Josep Domingo-Ferrer and Vicnec Torra(2005), 'Ordinal, Continuous and Heterogeneous *k*-Anonymity Through Microaggregation', Data Mining and Knowledge Discovery, Vol.11, pp.195 – 212.
6. Acar Tamersoy, Grigorios Loukides, Mehmet Ercan Nergiz, Yucel Saygin and Bradley Malin(2012), 'Anonymization of Longitudinal Electronic Medical Records', IEEE Transactions on Information Technology in Biomedicine, Vol.16(3), pp – 413 – 423.
7. Tiancheng Li, Ninghui Li, Jian Zhang and Ian Molloy, 'Slicing: A New Approach for Privacy Preserving Data Publishing', IEEE Transactions on Knowledge and Data Engineering, Vol.24(3), pp.561 – 574.
8. http://www.mathcs.emory.edu/~lxiong/cs573_s12/share/slides/0306_ppdm.pdf
9. http://missingdata.wordpress.com/2007/08/23/k-anonymity-and-l-diversity/
10. http://en.wikipedia.org/wiki/List_of_ICD9_codes_390%E2%80%93459:_diseases_of_the_circulatory_system