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 2, February 2015

Developing a Blueprint for Preserving Privacy of Electronic Health Records using Categorical Attributes

T.Kowshiga[1], T.Saranya[2], T.Jayasudha[3], Prof.M.Sowmiya[4] and Prof.S.Balamurugan[5]

Department of IT, Kalaignar Karunanidhi Institute of Technology, Coimbatore, TamilNadu, India[1,2,3,4,5]

ABSTRACT: Cloud computing offers unique opportunities for supporting long-term record preservation. MyPHRMachines , a patient owned health record system prototype based on remote virtual machines hosted in the cloud. MyPHRMachines is particularly promising for countries with a very heterogeneous architecture of systems across hospitals and other care institutions. In the view of developer PHRs should be portable. PHR systems typically offer functionality to share, visualize and analyze PHR data. Secure lifelong management of patient medical records since data are stored in the cloud and do not have to be carried around by patients. We also present method for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set.

KEYWORDS: Ontology, Micro aggregation, Differential Privacy, De-Identification, Biomedical Information System, Anonymous Authentication.

I. INTRODUCTION

The publishing of data to third parties is also as important assure as masking of data. Because the hackers can be traced with good amount of evidence, the leakage of data is detected. To tackle this problem an image is attached with the masked data and then it is distributed to the agents. Using steganography the masked data is shared to agent. The attached image contains the key which will give alert message to the distributor while agent distributed to any other third parties.

If the distributor sees enough evidence that an agent leaked data, he may stop doing business with him, or may initiate legal proceedings. In this paper we develop a model for assessing the guilt of agents. We also present method for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set. Such objects do not correspond to real entities but appear. If it turns out an agent was given one or more fake objects that were leaked, then the distributor can be more confident that agent was guilty.

II. LITERATURE SURVEY

k-Anonymity concept used for solve the tension between data utility and respondent privacy for individual data protection. The generalization and suppression approaches proposed in literature to achieve K-Anonymity is not equally suited for all types of attributes:

- Generalization/suppression is one of the few possibilities for nominal categorical attributes.
- It is one possibility for ordinal categorical attributes.
- It is completely unsuitable for continuous attributes, as it causes then to lose their numerical meaning.

The fundamental rights of patient to have their privacy protected by health care organizations. This information used to identify particular individual is not used to reveal sensitive patient such diagnoses,etc. If the degree of anonymity of a disseminated data set could be measured. Privacy protection in disseminated databases can be facilitated by use of special ambiguities algorithm.Generalization involves replacing a value with less specific but semantically consistent value. Suppression involves not replacing a value at all. In this paper authors plead on the need of knowledge intensive

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 2, February 2015

tools is data privacy more especially we discuss the role of knowledge related tools in data protection and in disclosure risk assessment. Statistical Disclosure Control(SDC) family of method for micro data masking the micro data so that they can be released while preserving the privacy. Aggregate original database records into small groups prior to publication. Each group should contain k-records.k->constant value. Recently micro aggregation achieve k-Anonymity in addition. Optimal micro aggregation computed in polynomial time for univariate data. This paper present new data oriented heuristics which improve on the trade off between computational complexity and information loss and are thus usable for large data sets. Microaggregation is the well known Microdata protection method, ensuring confidentiality. Authors propose and use for new approach like text documents. This method relies on word net framework that provide full semantic relationship taxonomy between words. Authors aim to ensure confidentiality of text document, at the same time preserve general meaning by applying some measures to evaluate the quality of the protection method relying on information loss.Inference control in data base also known as SDC. This is an important application in several areas such as official statistics, health statistics, e-commerce,etc. Hence it refers to data modification, challenge for SDC is to achieve protection with minimum loss of accuracy database we discuss several information loss and disclosure risk measures and analyse several ways of combining them to assess the performance of the various method. In US, the Health Insurance Portability and Accountability Act(HIPAA) protects the confidentiality of patient data and approval of internal review Board to use data for research but these requirements can be waived if data is de-identified. The De-identification of narrative text documents often realized and require significant resourses .In this method based performed better with PHI is rarely mentioned in clinical text but are more difficult to generalize.

Patient record data are highly sensitive so their secondary use raises both ethical and data protection issues. Disclosure of patient data could cause serious difficulties so individual damaging for patient and clinicians. In this paper grid based medical data repository accessing risk and suggest a new model for Statiscal Disclosure Control(SDC) of patient data. It provides enormous opportunities for data mining tasks. K-Anonymity which are vulnerable to privacy attacks based on background knowledge. Set-value data could be efficiently released under differential privacy with guaranteed utility help of taxonomy trees. Top down partitioning algorithm to generate a differentially privacy release scale with input data size. Protect of personal data in statistical data base has become major concern before they released to public use they applied statistical data bases. Microaggregation for SDC is to protect micro data that is record on individual compares. Micro data into groups at least K-records replace the record in each group.DBA(Density Based Algorithm) it form descending order of their densities in reverse order and compare with latest microaggregation methods.

III. BASIC PRIMITIVES AND TERMINOLOGIES

Micro Data

The customers or patient's data's are collected for this process. In this, we consider micro data such as census data and medical data. Typically, microdata is stored in a table, and each record corresponds to one individual. Each record has a number of attributes, which can be divided into the following three categories:

- **Identifier**. Identifiers are attributes that clearly identify individuals. Examples include *Social Security Number* and *Name*.
- **Quasi-Identifier**. Quasi-identifiers are attributes whose values when taken together can potentially identify an individual. Examples include *Zip-code*, *Birthdate*, and *Gender*.
- **Sensitive Attribute**. Sensitive attributes are attributes whose values should not be associated with an individual by the adversary. Examples include *Disease* and *Salary*.

Data Privacy

This effectively limits the amount of individual-specific information an observer can learn. However, an analysis on data utility shows that t-closeness substantially limits the amount of useful information that can be extracted from the released data.
This limits the amount of sensitive information about individuals while preserves features and patterns about large groups.

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 2, February 2015

View Agents List
The Admin can view all registered agents using this module. It contains the agent's full details for reference. It contains like Agent ID, Agent Name, Contact Number, Mail ID etc.,
Attach Fake Object: Fake objects are objects generated by the distributor that are not in set T. It contains the secret file and saved location and secret keys.
Stegnography(Secret File Sharing) Steganography is an alternative to encryption for keeping data or correspondence confidential. The Secret is embedding with in image. After that the key generated for secure sharing.
View Distribution List The Admin also known as distributed agent's data's. This module contains the details about already distributed data's in agent wise.
Data Leak Report
View Leaked Agent
The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. It displays the Agent ID, Agent Name etc.,
Image Extraction When the receiver gets the image, he will use the same random number generator.
AGENT
Agent Registration The registration module contains the agents personal details like Agent ID, Agent Name, Contact Number, Mail ID etc., and The Agent choose click points at the time of registration. It's very sensitive data's transaction.
Agent Login
Image Authentication Image authentication has been proposed as a user-friendly alternative to password generation and authentication.
Receive Data: Random allocation has also performance, since as the number of agents increases, the probability that at least two agents receive many common objects becomes higher. The every agent has successfully login into their application he can view the received sensitive file.

IV. WORKING METHODOLOGY

In this paper we describe understanding the textual data requires exploitation and integration of clinical resources. In past several approaches for assessing word similarity by exploiting different knowledge source have been proposed. These measures have been adapted to the biomedical field by incorporating domain information extracted from clinical data.

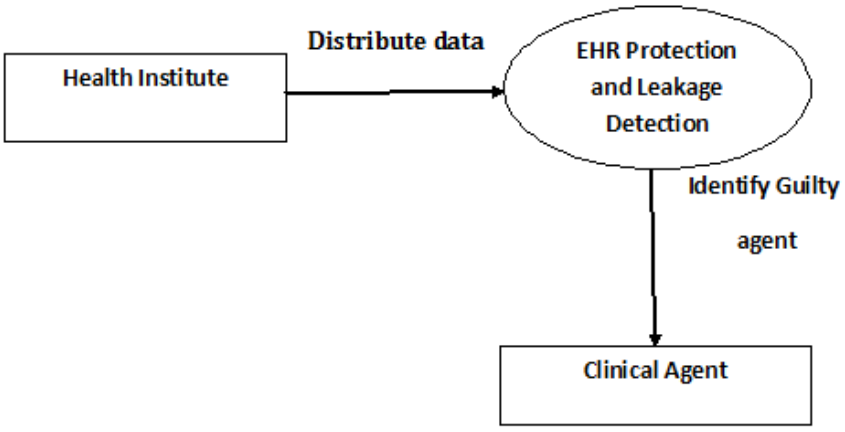

Fig.1 Distribution of Data through EHR mechanism

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 2, February 2015

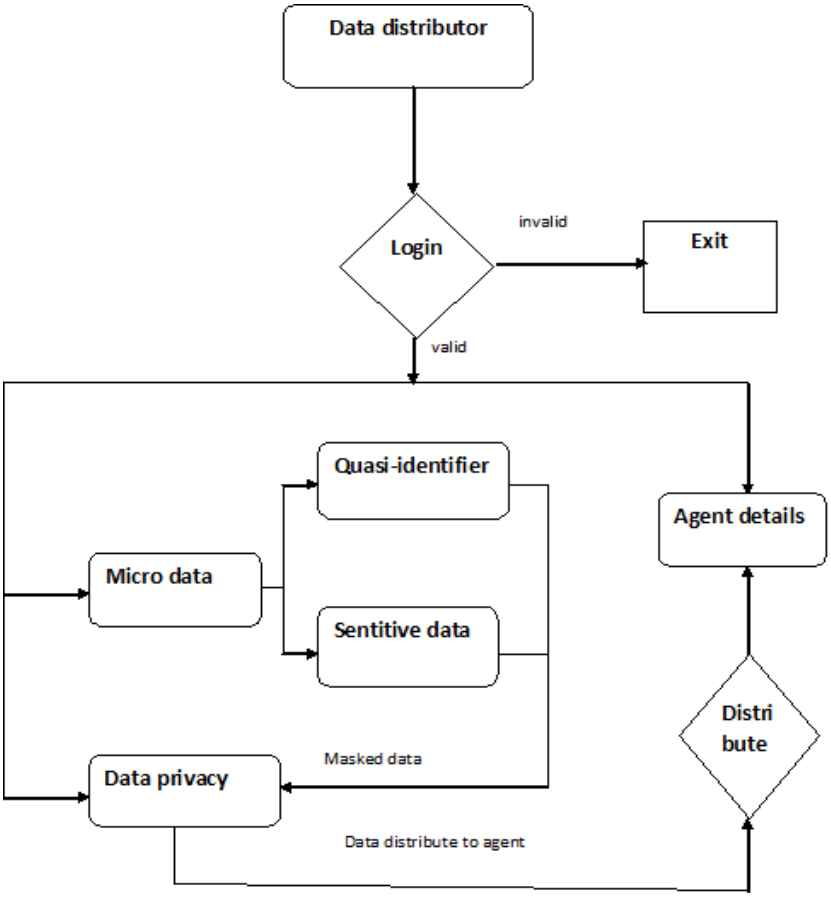

Fig.2 Flowchart for Agent-Based Leakage Detection

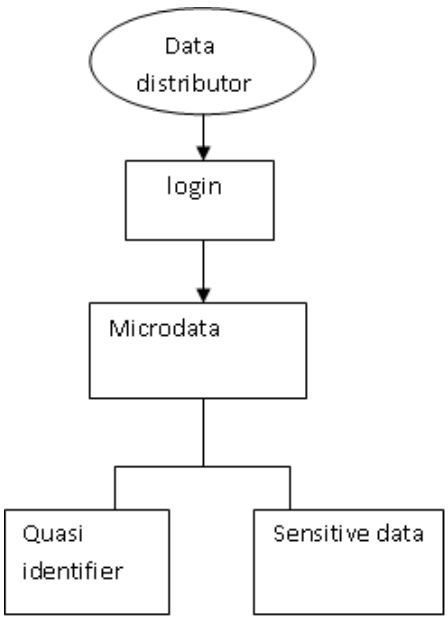

Fig.3 Classification of Data by Distributor

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 2, February 2015

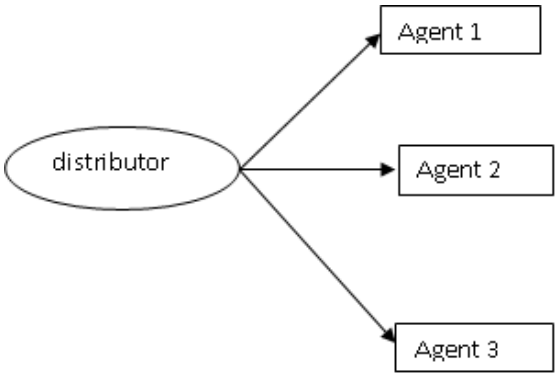

Fig.4 Role of Distributor in Distributor-Agent relationship

Distributor can send the data to different agent with hide of original data The distributor may be able to add fake objects to the distributed data in order to improve his effectiveness in detecting guilty agents. It displays the Agent ID, Agent Name etc., the objects are designed to look like real objects.

V. CONCLUSION AND FUTURE WORK

This paper detailed about various methods prevailing in literature for protecting privacy of anonymized medical data. Ontology Based measure to compute semantic similarity in Biomedicine is studied. Ordinal, continuous and heterogeneous K-Anonymity through Microaggregation are dealt in detail. Protecting patient privacy by quantifiable control of disclosure in disseminated databases and achieving k-Anonymity privacy protection using generalization and suppression are discussed in detail. Efficient Multivariate data-Oriented Micro aggregation of Categorical data for confidential documents is examined. Differential Privacy for Automatic De-Identification of textual documents in the electronic health records and Statistical Disclosure control for patient records in biomedical information System is considered. Density-based microaggregation for statistical disclosure control and anonymization of Set-Valued Data via Top-Down, Local Generalization are also aggregated in brief. In this paper we developed a model for assessing the guilt of agents. We also presented method for distributing objects to agents, in a way that improves our chances of identifying a leaker. Finally, we also consider the option of adding fake objects to the distributed set.

REFERENCES

[1] Sergio Martinez, David Sanchez,Aida Valls, "A Semantic Framework to Protect the Privacy of Electronic Health Records with Non-numerical Attributes", Journal of Biomedical Informatics 46 (2013) 294–303.
[2] GPO, US: 45 C.F.R. 164 Security and Privacy 2008. <http://www.access.gpo.gov/nara/cfr/waisidx_08/45cfr164_08.html>.
[3] Abril D, Navarro-Arribas G, Torra V. Towards semantic microaggregation of categorical data for confidential documents. In: Proceedings of the 7th international conference on Modeling decisions for artificial intelligence. Perpignan (France): Springer-Verlag; 2010. p. 266–76.
[4] Batet M, Sanchez D, Valls A. An ontology-based measure to compute semantic similarity in biomedicine. J Biomed Inform 2011;44:118–25.
[5] Chen R, Mohammed N, Fung BCM, Desai BC, Xiong L. Publishing set-valued data via differential privacy. PVLDB 2011;4:1087–98.
[6] Domingo-Ferrer J. A survey of inference control methods for privacy preserving data mining. In: Aggarwal CC, Yu PS, editors. Privacy-preserving data mining. Springer US; 2008. p. 53–80.
[7] Domingo-Ferrer J, Martínez-Ballesté A, Mateo-Sanz J, Sebé F. Efficient multivariate data-oriented microaggregation. VLDB J 2006;15:355–69.
[8] Domingo-Ferrer J, Mateo-Sanz JM. Resampling for statistical confidentiality incontingency tables. Comput Math Appl 1999;38:13–32.
[9] Domingo-Ferrer J, Mateo-Sanz JM. Practical data-oriented microaggregation for statistical disclosure control. IEEE Trans Knowl and Data Eng 2002;14:189–201.
[10] Domingo-Ferrer J, Torra V. Ordinal, continuous and heterogeneous kanonymity through microaggregation. Data Min Knowl Discov 2005; 11:195–212.
[11] Dwork C. Differential privacy. In: ICALP. Springer; 2006. p. 1–12.
[12] Elliot M, Purdam K, Smith D. Statistical disclosure control archite ctures for patient records in biomedical information systems. J Biomed Inform 2008;41:58–64.
[13] He Y, Naughton J. Anonymization of set-valued data via top-down, local generalization, VLDB '09: the thirtieth international conference on very large data bases. Lyon, France: VLDB Endowment; 2009.

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 2, February 2015

[14] Jones DH, Adam NR. Disclosure avoidance using the bootstrap and other resampling schemes. In: Proceedings of the fifth annual research conference. Washington (DC): U.S. Bureau of the Census; 1989. p. 446–55.
[15] Kullback S, Leibler R. On information and sufficiency. Ann Math Stat 1951;22:79–86.
[16] Lin J-L, Chang P-C, Liu JY-C, Wen T-H. Comparison of microaggregation approaches on anonymized data quality. Exp Syst Appl 2010;37:8161–5.
[17] Lin J-L, Wen T-H, Hsieh J-C, Chang P-C. Density-based microaggregation for statistical disclosure control. Exp Syst Appl 2010;37:3256–63.
[18] Machanavajjhala A, Kifer D, Gehrke J, Venkitasubramaniam M. L-diversity: privacy beyond k-anonymity. ACM Trans Knowl Discov Data 2007;1:3.
[19] Malin B, Sweeney L. How (not) to protect genomic data privacy in a distributed network: using trail re-identification to evaluate and design anonymity protection systems. J Biomed Inform 2004;37:179–92.
[20] Martínez S, Sánchez D, Valls A. Semantic adaptive microaggregation of categorical microdata. Comput Secur 2012;31:653–72.
[21] Martínez S, Sánchez D, Valls A. Towards k-anonymous non-numerical data via semantic resampling. In: Greco S, et al., editors. Information processing and management of uncertainty in knowledge-based systems. Catania, Italy; 2012. p. 519–28.
[22] Martinez S, Sanchez D, Valls A, Batet M. Privacy protection of textual attributes through a semantic-based masking method. Inf Fusion 2011;13:304–14.
[23] Martínez S, Sánchez D, Valls A, Batet M. The role of ontologies in theanonymization of textual variables. In: Proceeding of the 2010 conference on artificial intelligence research and development: proceedings of the 13th international conference of the Catalan association for Artificial intelligence. IOS Press; 2010. p. 153–62.
[24] Martínez S, Valls A, Sánchez D. Semantically-grounded construction of centroids for datasets with textual attributes. Knowl Based Syst 2012;35:160–72.
[25] Meystre S, Friedlin J, South B, Shen S, Samore M. Automatic de-identification of textual documents in the electronic health record: a review of recent research. BMC Med Res Methodol 2010;10:70.
[26] Mohammed N, Chen R, Fung BCM, Yu PS. Differentially private data release for data mining. In: Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining. San Diego, California (USA): ACM; 2011. p. 493–501.
[27] Nelson SJ, Johnston D, Humphreys BL. Relationships in medical subject headings. In: Publishers KA, editor. Relationships in the organization of knowledge. New York; 2001. p. 171–84.
[28] Nin J, Herranz J, Torra V. On the disclosure risk of multivariate microaggregation. Knowl Data Eng 2008;67:399–412.
[29] Ohno-Machado L, Silveira PSP, Vinterbo SA. Protecting patient privacy by quantifiable control of disclosures in disseminated databases. Int J Med Inform 2004;73:599–606.
[30] Pedersen T, Pakhomov SVS, Patwardhan S, Chute CG. Measures of semantic similarity and relatedness in the biomedical domain. J Biomed Inform 2007;40:288–99.
[31] Purdam K, Elliot M. A case study of the impact of statistical disclosure control on data quality in the individual UK Samples of Anonymised Records. Environ Plan A 2007;39:1101–18.
[32] Rogers J. Publically reported breaches in EPR confidentiality; 2005.
[33] Samarati P, Sweeney L. Protecting privacy when disclosing Information: kanonymity and its enforcement through generalization and suppression, technical report SRI-CSL-98-04, SRI Computer Science Laboratory; 1998.
[34] Sanchez D, Batet M. Semantic similarity estimation in the biomedical domain: an ontology-based information-theoretic perspective. J Biomed Inform 2011;44:749–59.
[35] Sánchez D, Batet M, Isern D, Valls A. Ontology-based semantic similarity: a new feature-based approach. Exp Syst Appl 2012;39:7718–28.
[36] Spackman KA, Campbell KE, Cote RA. SNOMED RT: a reference terminology forhealth care. Proc AMIA Annu Fall Symp 1997:640–4.
[37] Sweeney L. Achieving k-anonymity privacy protection using generalization and suppression. Int J Uncertain Fuzz Knowl Based Syst 2002;10:571–88.
[38] Sweeney L. K-anonymity: a model for protecting privacy. Int J Uncertain Fuzz Knowl Based Syst 2002;10:557–70.
[39] Torra V. Microaggregation for categorical variables: a median based approach.
[40] In: Domingo-Ferrer J, Torra V, editors. Privacy in statistical databases. Berlin (Heidelberg): Springer; 2004. p. 518.
[41] Torra V. Towards knowledge intensive data privacy. In: Proceedings of the 5th international workshop on data privacy management, and 3rd international conference on Autonomous spontaneous security. Athens (Greece): Springer- Verlag; 2011. p. 1–7.
[42] B.Powmeya , Nikita Mary Ablett ,V.Mohanapriya,S.Balamurugan,"An Object Oriented approach to Model the secure Health care Database systems,"In proceedings of International conference on computer , communication & signal processing(IC^3SP)in association with IETE students forum and the society of digital information and wireless communication,SDIWC,2011,pp.2-3
[43] Balamurugan Shanmugam, Visalakshi Palaniswami, "Modified Partitioning Algorithm for Privacy Preservation in Microdata Publishing with Full Functional Dependencies", Australian Journal of Basic and Applied Sciences, 7(8): pp.316-323, July 2013
[44] Balamurugan Shanmugam, Visalakshi Palaniswami, R.Santhya, R.S.Venkatesh "Strategies for Privacy Preserving Publishing of Functionally Dependent Sensitive Data: A State-of-the-Art-Survey", Australian Journal of Basic and Applied Sciences, 8(15) September 2014.
[45] S.Balamurugan, P.Visalakshi, V.M.Prabhakaran, S.Chranyaa, S.Sankaranarayanan, "Strategies for Solving the NP-Hard Workflow Scheduling Problems in Cloud Computing Environments", Australian Journal of Basic and Applied Sciences, 8(15) October 2014.
[46] Charanyaa, S., et. al., , A Survey on Attack Prevention and Handling Strategies in Graph Based Data Anonymization. International Journal of Advanced Research in Computer and Communication Engineering, 2(10): 5722-5728, 2013.
[47] Charanyaa, S., et. al., Certain Investigations on Approaches forProtecting Graph Privacy in Data Anonymization. International Journal of Advanced Research in Computer and Communication Engineering, 1(8): 5722-5728, 2013.
[48] Charanyaa, S., et. al., Proposing a Novel Synergized K-Degree L-Diversity T-Closeness Model for Graph Based Data Anonymization. International Journal of Innovative Research in Computer and Communication Engineering, 2(3): 3554-3561, 2014.
[49] Charanyaa, S., et. al., , Strategies for Knowledge Based Attack Detection in Graphical Data Anonymization. International Journal of Advanced Research in Computer and Communication Engineering, 3(2): 5722-5728, 2014.
[50] Charanyaa, S., et. al., Term Frequency Based Sequence Generation Algorithm for Graph Based Data Anonymization International Journal of Innovative Research in Computer and Communication Engineering, 2(2): 3033-3040, 2014.

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 2, February 2015

[51] V.M.Prabhakaran, Prof.S.Balamurugan, S.Charanyaa," Certain Investigations on Strategies for Protecting Medical Data in Cloud", International Journal of Innovative Research in Computer and Communication Engineering Vol 2, Issue 10, October 2014
[52] V.M.Prabhakaran, Prof.S.Balamurugan, S.Charanyaa," Investigations on Remote Virtual Machine to Secure Lifetime PHR in Cloud ", International Journal of Innovative Research in Computer and Communication Engineering Vol 2, Issue 10, October 2014
[53] V.M.Prabhakaran, Prof.S.Balamurugan, S.Charanyaa," Privacy Preserving Personal Health Care Data in Cloud" , International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 2, October 2014
[54] P.Andrew, J.Anish Kumar, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, "Investigations on Evolution of Strategies to Preserve Privacy of Moving Data Objects" International Journal of Innovative Research in Computer and Communication Engineering, 2(2): 3033-3040, 2014.
[55] P.Andrew, J.Anish Kumar, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, " Certain Investigations on Securing Moving Data Objects" International Journal of Innovative Research in Computer and Communication Engineering, 2(2): 3033-3040, 2014.
[56] P.Andrew, J.Anish Kumar, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, " Survey on Approaches Developed for Preserving Privacy of Data Objects" International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 2, October 2014
[57] S.Jeevitha, R.Santhya, Prof.S.Balamurugan, S.Charanyaa, " Privacy Preserving Personal Health Care Data in Cloud" International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 2, October 2014.
[58] K.Deepika, P.Andrew, R.Santhya, S.Balamurugan, S.Charanyaa, "Investigations on Methods Evolved for Protecting Sensitive Data", International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 4, Decermber 2014.
[59] K.Deepika, P.Andrew, R.Santhya, S.Balamurugan, S.Charanyaa, "A Survey on Approaches Developed for Data Anonymization", International Advanced Research Journal in Science, Engineering and Technology Vol 1, Issue 4, Decermber 2014.
[60] S.Balamurugan, S.Charanyaa, "Principles of Social Network Data Security" LAP Verlag, Germany, ISBN: 978-3-659-61207-7, 2014
[61] S.Balamurugan, ,M.Sowmiya and S.Charanyaa, "Principles of Scheduling in Cloud Computing" Scholars' Press, Germany,, ISBN: 978-3-639-66950-3, 2014
S.Balamurugan, S.Charanyaa, "Principles of Database Security" Scholars' Press, Germany, ISBN: 978-3-639-76030-9, 2014