ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (784-790), Month: April - June 2015, Available at: www.researchpublish.com

Efficiently Preserving Of Sensitive Resultant Cloud Datasets

Smt. Sandhya[1], Prof S.M Joshi[2]

[1]PG Student, GNDEC Bidar, Karnataka, India

[2]Professor, ECE Dept., GNDEC Bidar, Karnataka, India

Abstract: **Cloud computing is an internet based new age computer technology which can store one's applications and databases from distance through network and various devices in one location and provide authorised customers with on demand accessibility. Parallel dataflow programs generate enormous amounts of distributed data that are short lived, yet are critical for completion of job and for good run time performance. Preserving secrecy of these datasets has become risky as adversaries may recover confidential information by analyzing such multiple resultant datasets. In cloud storage service users upload their datasets together with authentication information to cloud storage server. Highly scalable computing resources are supplied as an outer service through Internet on pay for what you use basis. Encryption works well for preserving the privacy of resultant datasets, but encrypting all the datasets in cloud is widely adopted in existing approaches which is time consuming and cost in-effective. We propose a novel upper-bound privacy leakage constraint based approach that identifies all functionally encrypt able sensitive data, so that privacy preserving cost can be saved while simultaneously satisfying privacy requirements of data owners. Evaluation results demonstrate that this upper-bound approach is better than existing ones where all datasets are encrypted.**

Keywords: **sensitive-data leakage, resultant datasets, upper-limit constraint, privacy saving cost.**

1. INTRODUCTION

Cloud computing is future generation technology that uses a network's hardware and software, storage and retrieval services and interfaces which helps in providing computing as a service that has corporate management and end users and a third party to manage cloud.[1] these easily available resources without large scale investment is remarkable in history of IT. Since cloud users are abundant in number, volume of data stored is BIG creating an issue regarding privacy concerns of data owners which makes them to be hesitant to store their data in cloud though the service is free of infrastructure and software investment cost. The processing of BIG application data stored in cloud by users also prduce huge resultant data carrying short life span, but still considered essential for completion of work carried out by end users. Like all other utilities [2] preserving the sensitive-data of end users and resultant dataset management is viewed as 6th utility that fetches benefit to end users that is considered essential to meet everyday needs in life. Hence we need to store these data to save the cost of computing them again. But preserving the sensitive-data of resultant datasets arises a new problem of sensitive-data leakage by analysing multiple unpreserved data of resultant datasets. Thus though encryption technique works well to save the cost of preserving the sensitive information of ALL resultant datasets, it is cost ineffective to frequently encrypt/decrypt datasets. we use other techniques like aggregation and anonymization for datasets whose value is below upper-limit. In this paper we propose an upper-limit approach to preserve the sensitive-data information of resultant datasets using heuristic algorithm which is cost-effective and efficient compared to encryption of ALL datasets or only resultant datasets in cloud in existing approach.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (784-790), Month: April - June 2015, Available at: www.researchpublish.com

2. RELATED WORK

Let us have a brief overview of research work carried out in cloud on preserving of resultant datasets and their management.

Presently, cryptographic technique is exploited by most of existing applications in research work on maintaining private information of end users. Major part of the IT industry is transformed and shaped so that the hardware and software service becomes even more attractive because of the potential inherent in cloud computing technology [1].providing architecture for creation of cloud by virtual machine technology and allocation of resources in cloud and their management according to agreement oriented service level [2]. since many organisations rent for services provided by cloud service provider communities are benefited due to availability of resources on demand[3]. The virtual machine technology manages creation and allocation of images of operating system and applications to respective physical machines or a slice of server stack [4]. I.Gupta has proposed a resultant data storage system to reduce server failures and overhead of completion of work in time[5]. K.P.N puttaswamy and B.Y.zhao togetherly invented a set of tools that identifies encryptable data in cloud called silverline and have reduced the risk of managing key securely because of assigning keys only to specific data in cloud [6].D. Zissis and D. Lekkas have proposed a third party who is trust worthy and assures security in cloud, since cloud computing posses elements from various computing technologies like grid,autonomic and utility computing creating an issue for success of security [7].Roy et al. underwent an investigation regarding the data sensitivity problem caused by MapReduce and have put forward a technique named Airavat which involves compulsary access control with differential Sensitivity[8]. Zhang et al.have putforward a system called Sedic which partitions MapReduce calculation works in terms of the security which is required to carry out future work[9]. Ciriani et al.have putforward an approach that involves encryption and fragmentation of data to achieve sensitive-data protection for distributed type storage of data. We follow this but we integrate anonymization, aggregation and encryption techniques all together to satisfy the need of cost-effective sensitive-data protection [10]. Sensitive-data policies like k-anonymity and l-diversity concepts are put forth to overcome Sensitive-data leakage ,but most of them are applied only to single data set [11] [12]. Sensitive-data policies for N number of datasets are proposed. The research in [13],[14] exploits information theory to quantify the sensitive-data using maximum entropy policy[15].Many techniques related to anonymisation and aggregation have been put forth to save the sensitive data which works well for single datasets but they fail to solve the problem for N number of datasets. Our approach is to integrate all the techniques to fetch sensitivity preserving for N number of datasets.

3. EXPLANATION WITH REAL TIME EXAMPLE

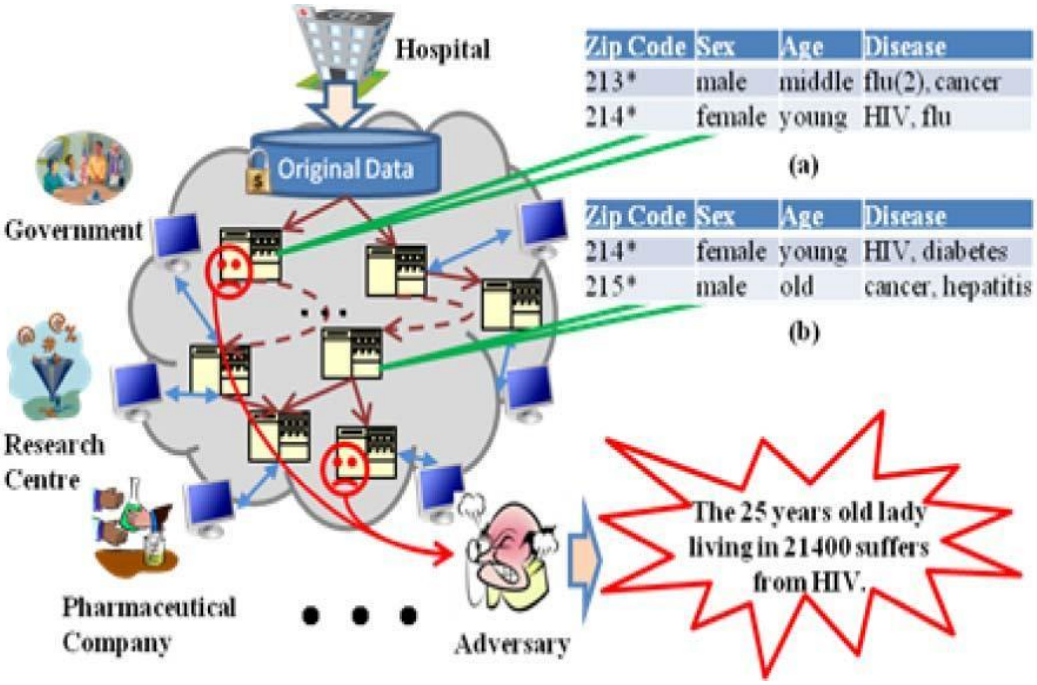

Figure 1 : Sensitive Data Leakage Problem due to resultant datasets

An example is illustrated in fig1 where hospital data records are moved into cloud servers as shown in fig from where the data can be accessed by remote computers for health benefits like all first aid treatment which can be given to diseased patients in emergency to avoid wastage of time in searching and running to hospital which can save a patient's life. Here we need to encrypt original data in cloud as there is an elevated risk of hacking for data stored in cloud by adversaries. As there are many users of cloud like government, research centres, pharmaceutical companies which may access or process original data after anonymization giving rise to huge resultant datasets as shown in fig.1(a) and fig.1(b) which are short lived but still required for completion of work carried out by users in future. Hence we need to save these resultant datasets to minimize the cost of recompiling them again as there are increasing chances of getting the sensitivity of original data leaked by analyzing these n number of resultant datasets by an unknown third party or hacker.

4. PROBLEM STATEMENT

In the existing system ALL datasets and ALL resultant datasets encryption approach was widely used in cloud to avoid leakage of sensitive-data which was costlier and inefficient hence now in this system we try to encrypt only part of resultant datasets in cloud but there is a high risk of loosing confidential data of original datasets by hackers as they analyse multiple resultant datasets. To overcome this kind of problem by following the approach proposed below.

5. PROPOSED METHODOLOGY

Heuristic algorithm is used for encryption of sensitive-data in cloud. First we attempt to calculate an upper-limit for sensitive-data leakage of N number of datasets in cloud. We use techniques like aggregation and anonymization for classified data to save the cost. Here we propose to encrypt only those datasets whose secrecy leakage is more than a threshold or heuristic value.

5.1 Upper-limit calculation:

We derive an upper-limit of PLm(Dune) that can be easily computed. Let du and dv be two data sets whose sensitive-data leakage are PLs(du) and PLs(dv), respectively. The combined secracy leakage obtained by them is PLm({du, dv}). As gain of information is never negative, PLm({du, dv}) is not less than neither PLs(du) nor PLs(dv).Further, PLm({du, dv}) will not exceed the sum of PLs(du) and PLs(dv) i.e. PLm({du, dv}) less than or equal to PLs(du) + PLs(dv), where the equality persists if and only if the information given by du and dv does not overlap.This property of combined secracy leakage can be extended to N number of unencrypted data sets in Dune: PLm(Dune).PLm (Dune) less than or equal to sum of PLs(di) where di belongs to Dune. so, the sum of sensitive-data leakage of unencrypted datasets can be deemed as an upper-limit of PLm(Dune).

5.2 Heuristic value calculation:

We obtain the heuristic value from a function called heuristic function denoted as f(SNi), is defined to calculate heuristic value of SNi i.e state node of SIT. It has two parts of information f(SNi)= g(SNi)+ h(SNi) i.e.where the information g(SNi) is obtained from the start state to the current state node SNi and the information h(SNi) is estimated from the present state node to the goal state. The heuristic function carry out the major role of the algorithm in selection of data sets with minimal cost but highest sensitivity leakage to encrypt. A directed acyclic graph represents the relationships of resultant data sets D from do the original dataset defined as a Sensitive Intermediate data set Graph, denoted as SIG which becomes a tree structure if each data set in D is obtained from only one parent data set called Sensitive intermediate data set tree (SIT) whose root is do.

Now we follow the following steps to encrypt datasets. We give SIT with root node do as input to the algorithm with all attribute values like size and frequency of accessing file, and sensitivity preserving cost etc and output of the algorithm is global sensitivity cost. Next we define priority for sensitive-data queue and first we construct the search node as the root of SIT and add it to queue, then pick the node with highest heuristic value from queue and we need to check for encrypted data, if it is not encrypted then it has solution and we repeat the process. Else we label the dataset as encrypted if its sensitivity leakage is more than threshold and list all solutions but select one out of that and calculate sensitive-data leakage upper-limit and encryption cost. Now calculate the remaining sensitivity leakage and its heuristic value. Next we design the newer search node from values obtained so far and add it to priority based queue.

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (784-790), Month: April - June 2015, Available at: www.researchpublish.com

6. SIMULATION RESULTS

The simulation results are as follows, Fig 2 depicts the starting of cloud server.

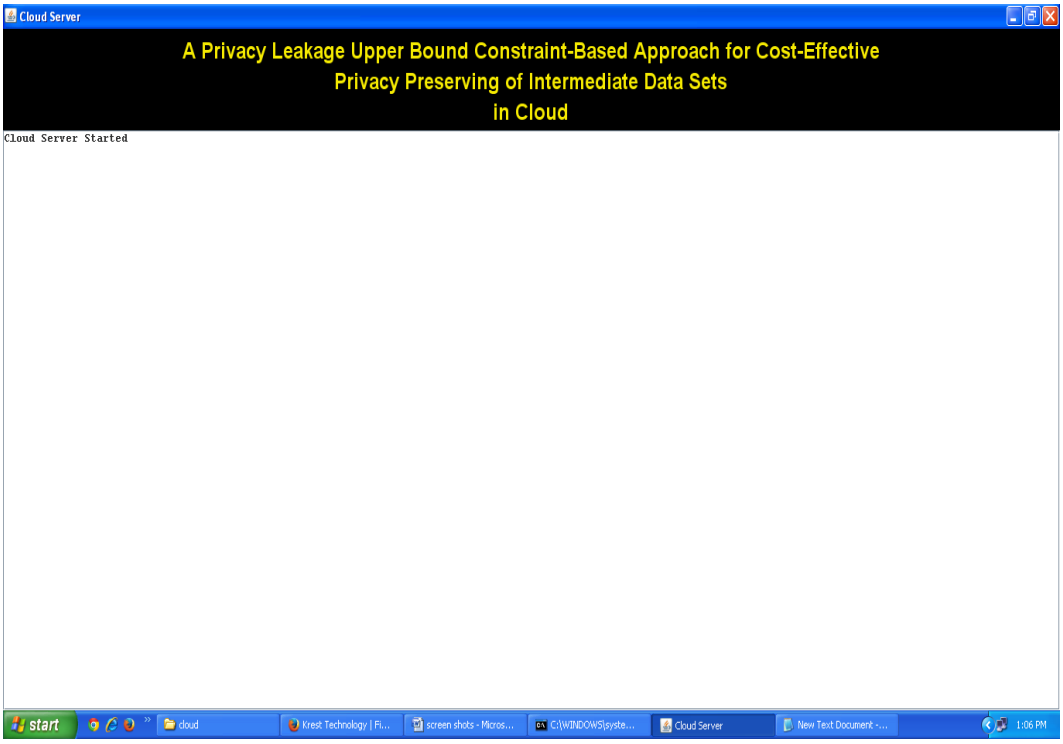

Figure 2: Cloud Server Started

Now the admin will login in his login screen and he has a home screen page as seen in figure 3:

Figure 3: Admin Home Screen

The admin will upload the dataset into the cloud by clicking on upload dataset button in admin home screen and aggregate the data such as age here by clicking on generalize algorithm button which is shown in next figure

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (784-790), Month: April - June 2015, Available at: www.researchpublish.com

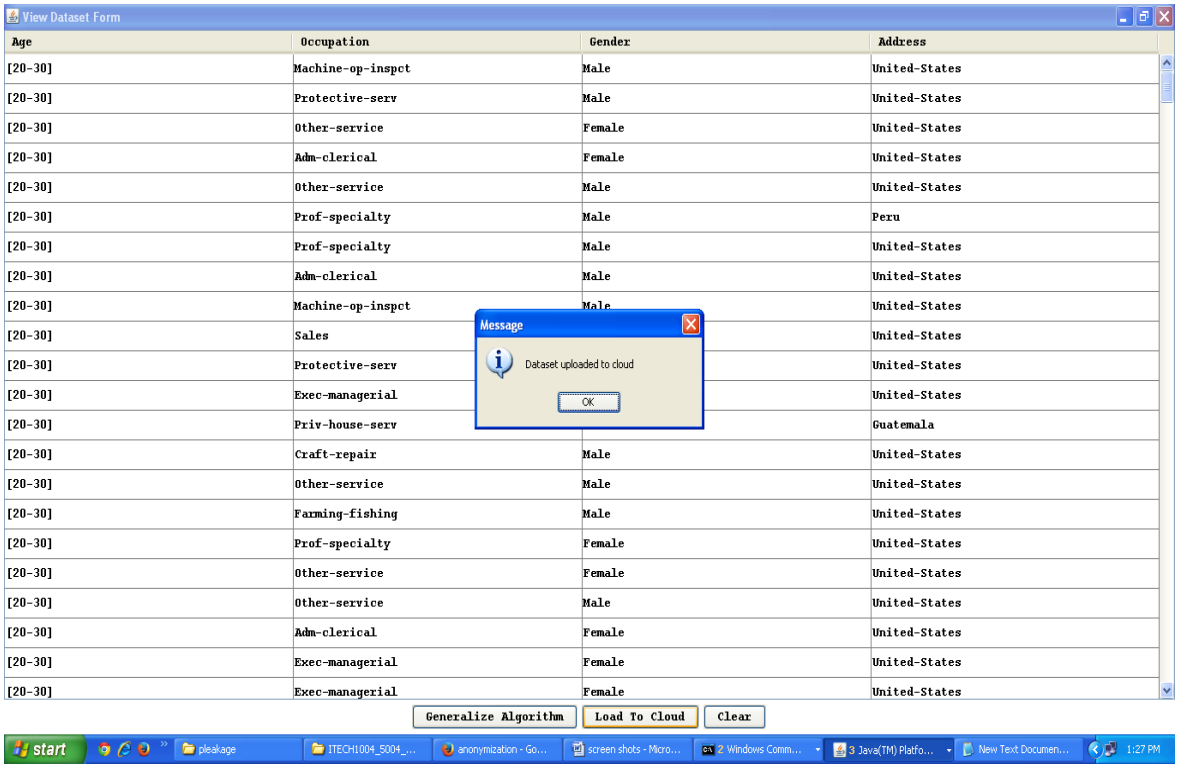

Figure 4: Aggregated Data Upload

Only the registered user can login into the cloud whose home page is as shown in fig5 below consisting of access dataset button and logout button

Figure 5: User Home Screen

For example if we try to search for united states candidates information like occupation here and if the access frequency is more than heuristic value then the details of occupation of united states people get encrypted .

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 3, Issue 2, pp: (784-790), Month: April - June 2015, Available at: www.researchpublish.com

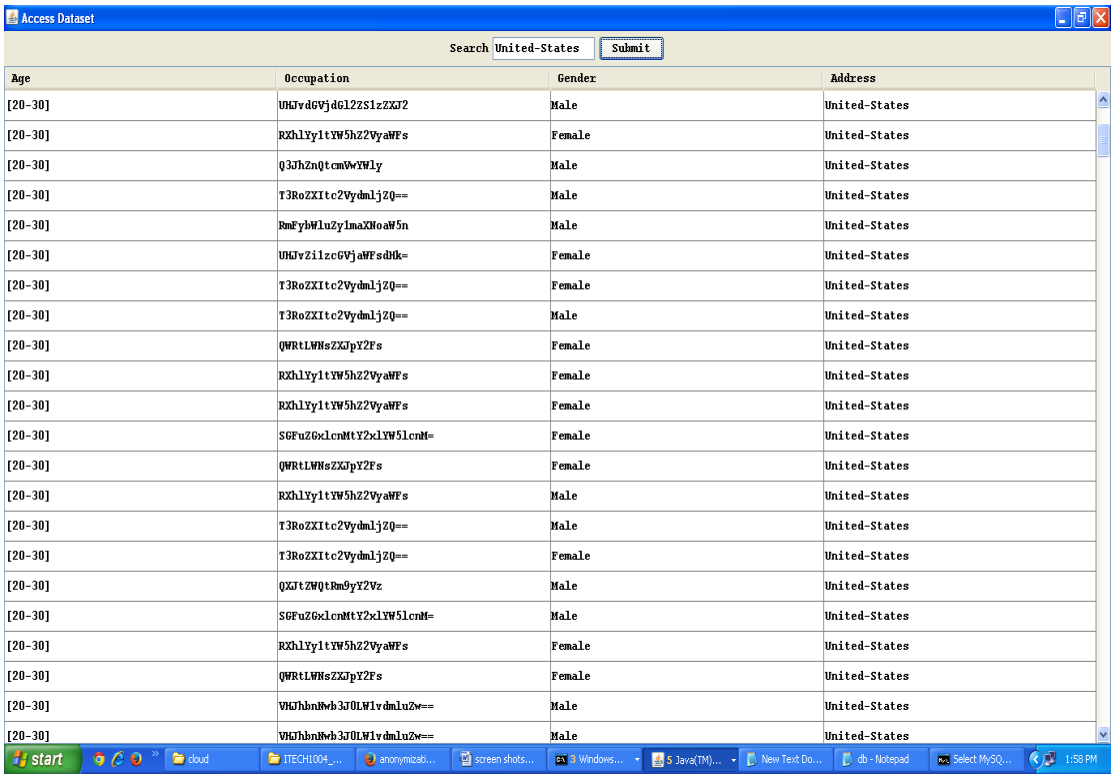

Figure 6: Encrypted Dataset

7. CONCLUSION

In this paper, we demonstrate heuristic algorithm for setting an upper-limit for encryption of datasets in cloud based on heuristic value instead of encrypting ALL datasets or ALL intermediate datasets as in existing system which was highly cost-ineffective and inefficient. Our approach minimizes the cost of sensitive-data preserving of resultant datasets in cloud as we integrate other techniques along with encryption like aggregation and anonymization for datasets below the heuristic value. In our approach we have considered the size and frequency of accessing resultant dataset files as shown in fig1 above as static for computation purposes, the dynamic nature of such files for computation purposes can be taken as future work.

ACKNOWLEDGEMENT

It's my immense pleasure to express my indebtedness to my guide Prof S.M. Joshi and Department of Electronics and Communication Engineering at Guru Nank Dev Engineering College Bidar for guiding me at various stages.

REFERENCES

[1] M. Armbrust, A. Fox, R. Griffith, A.D. Joseph, R. Katz, A.Konwinski, G. Lee, D. Patterson, A. Rabkin, I. Stoica, and M.Zaharia, "A View of Cloud Computing," Comm. ACM, vol. 53,no. 4, pp. 50-58, 2010.

[2] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic,"Cloud Computing and Emerging It Platforms: Vision, Hype, and Reality for Delivering Computing as the Fifth Utility," Future Generation Computer Systems, vol. 25, no. 6, pp. 599-616, 2009.

[3] L. Wang, J. Zhan, W. Shi, and Y. Liang, "In Cloud, Can Scientific Communities Benefit from the Economies of Scale?," IEEE Trans.Parallel and Distributed Systems, vol. 23, no. 2, pp. 296-303, Feb.2012.

[4] A Review of Cloud Computing Open Architecture and Its Security Issues Ashutosh Kumar Singh, Dr. Ramapati Mishra, Fuzail Ahmad, Raj Kumar Sagar, Anil Kumar Chaudhary

[5] S.Y. Ko, I. Hoque, B. Cho, and I. Gupta, "Making Cloud Intermediate Data Fault-Tolerant," Proc. First ACM Symp. Cloud Computing (SoCC '10), pp. 181-192, 2010.

[6] K.P.N. Puttaswamy, C. Kruegel, and B.Y. Zhao, "Silverline:Toward Data Confidentiality in Storage-Intensive Cloud Applications,"Proc. Second ACM Symp. Cloud Computing (SoCC '11), 2011.

[7] D. Zissis and D. Lekkas, "Addressing Cloud Computing Security Issues," Future Generation Computer Systems, vol. 28, no. 3, pp. 583-592, 2011.

[8] I. Roy, S.T.V. Setty, A. Kilzer, V. Shmatikov, and E. Witchel,"Airavat: Security and Privacy for Mapreduce," Proc. Seventh USENIX Conf. Networked Systems Design and Implementation (NSDI'10), p. 20, 2010.

[9] K. Zhang, X. Zhou, Y. Chen, X. Wang, and Y. Ruan, "Sedic: Privacy-Aware Data Intensive Computing on Hybrid Clouds,"Proc. 18th ACM Conf. Computer and Comm. Security (CCS '11),pp. 515-526, 2011.

[10] V. Ciriani, S.D.C.D. Vimercati, S. Foresti, S. Jajodia, S. Paraboschi,and P. Samarati, "Combining Fragmentation and Encryption to Protect Privacy in Data Storage," ACM Trans. Information and System Security, vol. 13, no. 3, pp. 1-33, 2010.

[11] P. Samarati, "Protecting Respondents' Identities in Microdata Release," IEEE Trans. Knowledge and Data Eng., vol. 13, no. 6,pp. 1010-1027, Nov. 2001.

[12] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,"L-Diversity: Privacy Beyond K-Anonymity," ACM

[13] G. Wang, Z. Zutao, D. Wenliang, and T. Zhouxuan, "Inference Analysis in Privacy-Preserving Data Re-Publishing," Proc. Eighth IEEE Int'l Conf. Data Mining (ICDM '08), pp. 1079-1084, 2008.

[14] W. Du, Z. Teng, and Z. Zhu, "Privacy-Maxent: Integrating Background Knowledge in Privacy Quantification," Proc. ACMSIGMOD Int'l Conf. Management of Data (SIGMOD '08), pp. 459-472, 2008.

[15] E.T. Jaynes, "Information Theory and Statistical Mechanics,"Physical Rev., vol. 106, no. 4, pp. 620-630, 1957.