International Journal of Applied Engineering Research
ISSN 0973-4562 Volume 10, Number 1 (2015) pp. 737-748
© Research India Publications
http://www.ripublication.com

A Resemblance Outsource Data Leakage Detection In Cloud Computing

Dr.R.Vasanthi[1]
M.E , Ph.D
Department of computer science And Engineering
Idhaya Engineering College for Women
Chinnasalem
E-mail:vasanti_me@yahoo.co.in
Dr.R.Prakash[2]
M.Tech., Ph.D
Department of Electrical and Electronics Engineering
Vivekananda Engineering College for Women
Trichengode
E-mail:prakashragu@yahoo.co.in

Abstract

Cloud computing is sought after for its reduced infrastructure and implementation costs. The cloud infrastructure as a Service paradigm and to create a flexible management framework. The rise in the usage of cloud computing also gave rise to privacy and security concerns. Concerns like data leakage are pretty common and can affect the original data owners drastically. Many traditional systems Upper Bound Privacy Leakage Constraint-Based approach to identify which intermediate data sets need to be encrypted and which do not .The method of encryption to hide the data and their respective data sets to ensure privacy and security. This method is not economically feasible as it involves encrypting every data over the cloud which is a huge collection of data. We propose Distributed System management and the Lattice compression Algorithm (LCA) to authenticate the data access from the cloud to ensure the security of the data over cloud and their data sets. Evaluation results demonstrate that the secrecy and privacy of the data stored over the network is achieved efficiently and effectively.

Keywords: Cloud computing, DSM, LCA

Introduction

Technically, cloud computing is regarded as an ingenious combination of a series of technologies, establishing a novel business model by offering IT services and using

economies of scale [1], [2]. Participants in the business chain of cloud computing can benefit from this novel model. Cloud customers can save huge capital investment of IT infrastructure, and concentrate on their own core business [3]. Therefore, many companies or organizations have been migrating or building their business into cloud. However, numerous potential customers are still hesitant to take advantage of cloud due to security and privacy concerns [4].

The privacy concerns caused by retaining intermediate data sets in cloud are important but they are paid little attention.

Storage and computation services in cloud are equivalent from an economical perspective because they are charged in proportion to their usage [1]. Thus, cloud users can store valuable intermediate data sets selectively when processing original data sets in data intensive applications like medical diagnosis, in order to curtail the overall expenses by avoiding frequent re-computation to obtain these data sets [6], [7]. Such scenarios are quite common because data users often reanalyze results, conduct new analysis on intermediate data sets, or share some intermediate results with others for collaboration. Without loss of generality, the notion of intermediate data set herein refers to intermediate and resultant data sets[5]. However, the storage of intermediate data enlarges attack surfaces so that privacy requirements of data holders are at risk of being violated. Usually, intermediate data sets in cloud are accessed and processed by multiple parties, but rarely controlled by original data set holders. This enables an adversary to collect intermediate data sets together and menace privacy-sensitive information from them, bringing considerable economic loss or severe social reputation impairment to data owners. But, little attention has been paid to such a cloud-specific privacy issue.

Existing technical approaches for preserving the privacy of data sets stored in cloud mainly include encryption and anonymization. On one hand, encrypting all data sets, a straightforward and effective approach, is widely adopted in current research [8], [9], [10]. However, processing on encrypted data sets efficiently is quite a challenging task, because most existing applications only run on unencrypted data sets. Although recent progress has been made in homomorphic encryption which theoretically allows performing computation on encrypted data sets, applying current algorithms are rather expensive due to their inefficiency [11]. On the other hand, partial information of data sets, e.g., aggregate information, is required to expose to data users in most cloud applications like data mining and analytics. In such cases, data sets are anonymized rather than encrypted to ensure both data utility and privacy preserving. Current privacy-preserving techniques like generalization [12] can withstand most privacy attacks on one single data set, while preserving privacy for multiple data sets is still a challenging problem [13] [14]. Thus, for preserving privacy of multiple data sets, it is promising to anonymize all data sets first and then encrypt them before storing or sharing them in cloud. Usually, the volume of intermediate data sets is huge [6]. Hence, argue that encrypting all intermediate data sets will lead to high overhead and low efficiency when they are frequently accessed or processed. As such, we propose to encrypt part of intermediate data sets rather than all for reducing privacy-preserving cost.

Related Work

In this, briefly review the research on privacy protection in cloud, intermediate data set privacy preserving and Privacy-Preserving Data Publishing (PPDP). Currently, encryption is exploited by most existing research to ensure the data privacy in cloud [8], [9], [10]. Although encryption works well for data privacy in these approaches, it is necessary to encrypt and decrypt data sets frequently in many applications. Encryption is usually integrated with other methods to achieve cost reduction, high data usability and privacy protection. Roy et al. [15] investigated the data privacy problem caused by MapReduce and presented a system named Airavat which incorporates mandatory access control with differential privacy. Puttaswamy et al. [16] described a set of tools called Silverline that identifies all functionally encryptable data and then encrypts them to protect privacy. Zhang et al. [17] proposed a system named Sedic which partitions MapReduce computing jobs in terms of the security labels of data they work on and then assigns the computation without sensitive data to a public cloud. The sensitivity of data is required to be labeled in advance to make the above approaches available. Ciriani et al. [18] proposed an approach that combines encryption and data fragmentation to achieve privacy protection for distributed data storage with encrypting only part of data sets. It follow this line, but integrate data anonymization and encryption together to fulfill cost-effective privacy preserving.The importance of retaining intermediate data sets in cloud has been widely recognized [6], [7], but the research on privacy issues incurred by such data sets just commences. Davidson et al. [19], [20], [21] studied the privacy issues in workflow provenance, and proposed to achieve module privacy preserving and high utility of provenance information via carefully hiding a subset of intermediate data. This general idea is similar to ours, yet our research mainly focuses on data privacy preserving from an economical cost perspective while theirs concentrates majorly on functionality privacy of workflow modules rather than data privacy. Our research also differs from theirs in several aspects such as data hiding techniques, privacy quantification and cost models. But, our approach can be complementarily used for selection of hidden data items in their research if economical cost is considered. The PPDP research community has investigated extensively on privacy-preserving issues and made fruitful progress with a variety of privacy models and preserving methods [13]. Privacy principles such as k-anonymity [22] and l-diversity [23] are put forth to model and quantify privacy, yet most of them are only applied to one single data set. Privacy principles for multiple data sets are also proposed, but they aim at specific scenarios such as continuous data publishing or sequential data releasing [13]. The research in [24], [25] exploits information theory to quantify the privacy via utilizing the maximum entropy principle [26]. The privacy quantification herein is based on the work in [24], [25]. Many anonymization techniques like generalization [12] have been proposed to preserve privacy, but these methods alone fail to solve the problem of preserving privacy for multiple data sets.

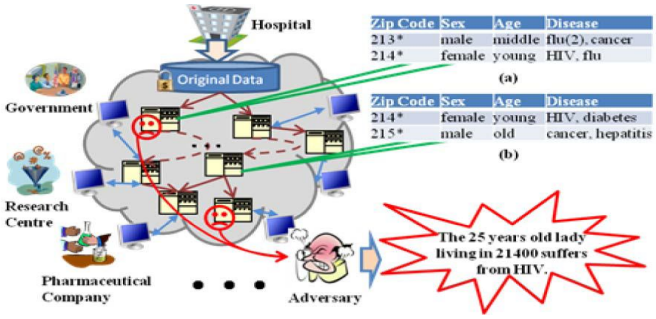


Fig. 1. A scenario showing privacy threats due to intermediate data sets

Our approach integrates anonymization with encryption to achieve privacy preserving of multiple data sets. Moreover, consider the economical aspect of privacy preserving, adhering to the pay-as-you-go feature of cloud computing.

Privacy Leakage Upper Bound Constraint-Based Approach For Privacy Preserving:

Here, an upper bound constraint-based approach to select the necessary subset of intermediate data sets that needs to be encrypted for minimizing privacy-preserving cost. In Section 3.1, specify relevant basic notations and elaborate two useful properties on an SIT. The privacy leakage upper bound constraint is decomposed layer by layer in Section 3.2. A constrained optimization problem with the PLC is then transformed into a recursive form in Section 3.3. In Section 3.4, a heuristic algorithm is designed for our approach and extend our approach to an SIG in Section 3.5.

Basic Notations and Properties on an SIT

Let $d_r \varepsilon D$ denote an intermediate data set in an SIT. Its directly generated data sets constitute a set $CD(d_r) \triangleq \{d_{r1}; \ldots ; d_{ri}\}$, where $d_{r1}, \ldots , d_{ri} \varepsilon D$. For any $d \varepsilon CD(d_r)$, $PLs(d) \leq PLs(dr)$ because all information in d is from d_r. Further, let $PD(d_r) \triangleq \{dr1; \ldots ; drj\}$ be all posterity data sets generated from d_r, where $dr_1, \ldots , d_{rj} \varepsilon D$. Similar to CD(dr), $PLs(d) \leq PLs(dr)$ for any $d \varepsilon P(dr)$. A subtree of an SIT with $dr \varepsilon D$ being its root is denoted as SDT(dr).

Recursive Privacy Leakage Constraint Decomposition

To satisfy the PLC, we decompose the PLC recursively into different layers in an SIT. Then, the problem stated can be addressed via tackling a series of small-scale optimization problems. Let the privacy leakage threshold required in the layer L_i be ε_i, $1 \leq i \leq H$. The privacy leakage incurred by UD_i in the solution π_i can never be larger than ε_i, i.e., $PL_m(Ud_i) \leq \varepsilon_i$. The threshold ε_i can be regarded as the privacy leakage threshold of the remainder part of an SIT after the layer L_{i-1}. In terms of the basic idea of our approach, the privacy leakage constraint $PLm(UDi) \leq \varepsilon_i$ is substituted by one of

its sufficient conditions. According to the PLC can be substituted by a set of privacy leakage constraints, named as PLC1:

$$\sum_{d\varepsilon UDi} PLs(d) \leq \varepsilon i, 1 \leq i \leq H \qquad (1)$$

The above threshold ε_i, $1 \leq i \leq H$, is calculated by

$$\begin{cases} \varepsilon i = \varepsilon i - 1 - \sum_{d\varepsilon UDi-1} PLs(d) \\ \varepsilon 1 = \varepsilon. \end{cases} \qquad (2)$$

A local encryption solution in the layer Li is feasible if it satisfies the PLC1 in (1). The set of feasible solutions in Li is denoted as $\Lambda_i^f \triangleq \{\pi ij | \pi ij \varepsilon \hat{} i\}$, where j is the number of feasible solutions. Similarly, a feasible global encryption solution can be denoted as $\prod_j^k \triangleq (\pi 1j1, \ldots, \pi Hjh)$,where

$$\prod ijl \in \bigwedge_i^f 1 \leq k \leq \prod_{i=1}^{\pi} |\hat{} i|$$

Given a feasible global solution π, for an SIT,we compress the SIT into a "compressed" tree layer by layer from L1 to LH, denoted as CT(π), where H is the height of the EDG in the SIT. The construction of CT(π) is achieved via three steps. First, the data sets in EDi are "compressed" into one encrypted node. According to the EDT property, these compressed nodes together with the original data set appear to be a string with the length being H. Second, all offspring data sets of the data sets in UDi are omitted. This will not affect the privacy preserving in terms of the RPC property. Third, the data sets in UDi are compressed into one node. Note that the action "compress" here is imaginary from a logical perspective for a demonstration purpose. Construction process of CT(π) is illustrated in Fig. 2 where the first layers have been built up according the above steps.

Minimum Privacy-Preserving Cost

Usually, more than one feasible global encryption solution exists under the PLC1 constraints, because there are many alternative local solutions in each layer. Further, each intermediate data set has various size and frequency of usage, leading to different overall cost with different solutions. Therefore, it is desired to find a feasible solution with the minimum privacy-preserving cost under privacy leakage constraints. Note that the minimum solution mentioned herein is somewhat pseudo minimum because an upper bound of joint privacy leakage is just an approximation of its exact value.

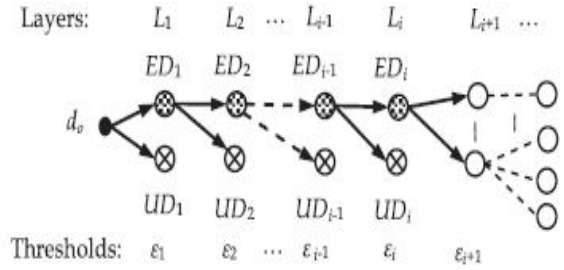


Fig. 2. Construction of a compressed tree.

Privacy-Preserving Cost Reducing Heuristic Algorithm

In this section, we design a heuristic algorithm to reduce privacy-preserving cost. In the state-search space for an SIT, a state node SNi in the layer Li herein refers to a vector of partial local solutions, i.e., SNi corresponds to ($\pi_{1j1,.....,\pi iji)}$, Note that the state-search tree generated according to an SIT is different from the SIT itself, but the height is the same. Appropriate heuristic information is quite vital to guide the search path to the goal state. The goal state in our algorithm is to find a near-optimal solution in a limited search space.

Extension to SIG

Although SITs can suit many applications, SIGs are also common, i.e., an intermediate data set can originate from more than one parent data set. An example of an SIG is illustrated in Fig. 3a. The data set d_5 is generated from d_2, d_3, and d_4, where d_4 is from d_2. Thus, it is possible that PLs($d_{5)}$ is larger than PLs(d2) or PLs(d3), resulting in the failure of Lemma 1 and RPC property. As a result, our approach cannot be directly applied to an SIG. However, we can adapt it to an SIG with minor modifications.

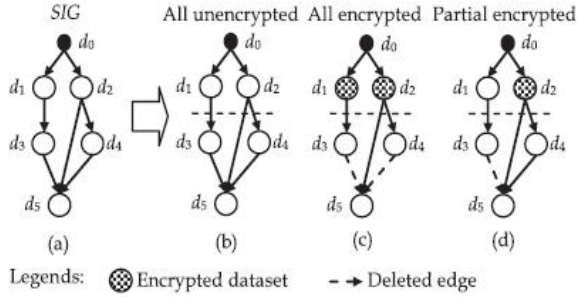


Fig. 3. Example for extending our approach to an *SIG*.

Let d_m denote a merging data set that inherits data from more than one predecessor, e.g., d_5 in Fig. 3. As only one root data set is assumed to exist in an SIG, all paths from the root data set d_o to d_m must converge at one point beside d_m itself. Let d_s denote this source data set, e.g., $d_s = d_o$ in Fig. 3. The inequality PLs(d_m) $\leq$ PLs(ds) holds because all privacy information of d_m comes from ds. Let PDi(ds) be

the set of the offspring data sets of ds in the layer Li, then PDi(ds) ≤ PD(ds). Data Sets in PDi(ds) are split into Edi and UDi when determining which data sets are encrypted.

To make the approach for an SIT available to an SIG as well, three minor modifications are required. The first one is to identify all merging data sets. The second one is to adjust the SIG according to the third case discussed above if $UD_{\pi} \neq \emptyset$; after we get a local solution $\pi=(ED_{\pi}, UD_{\pi}.)$. Fig. 4 shows the system overview of U-cloud

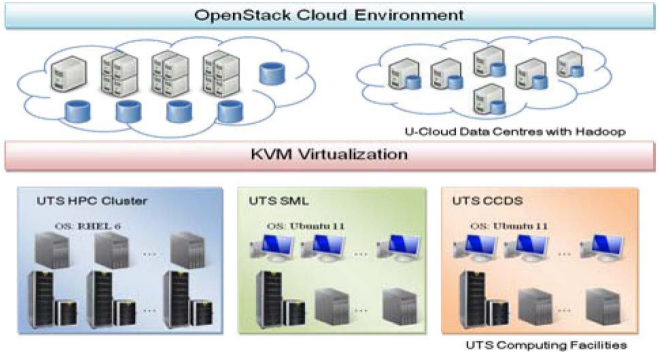


Figure 4: System overview of U-Cloud

The third one is to label the data sets that have been processed. In this way, it is unnecessary to explicitly delete edges discussed in the second case.

Proposed Method

A novel methodology to schedule the data oriented grid applications in an efficient way to reduce the overall execution time of the applications. It consists of the following steps:

Preprocessing

In this step read the semantic concepts and the meta data about the data sets available in the data grids, remove the noisy semantics and meta data and clean the meta data to ensure the quality of semantic indexing.

Distributed system management (DSM)

A **DSM** is a collection of applications designed to monitor & control the entire distribution network efficiently and reliably. It acts as a decision support system to assist the control room and field operating personnel with the monitoring and control of the electric distribution system. Improving the reliability and quality of service in terms of reducing outages, minimizing outage time, maintaining acceptable frequency and voltage levels are the key deliverables of a DSM. Fig. 5 shows the DSM in cloud.

DSMs access real-time data and provides all information on a single console at the control centre in an integrated manner. Their development varied across different

geographic territories. In the USA, for example, DMSs typically grew by taking Outage Management Systems to the next level, automating the complete sequences and providing an end to end, integrated view of the entire distribution spectrum. In the UK, by contrast, the much denser and more meshed network topologies, combined with stronger Health & Safety regulation, had led to early centralisation of high-voltage switching operations, initially using paper records and schematic diagrams printed onto large wallboards which were 'dressed' with magnetic symbols to show the current running states. There, DSMs grew initially from SCADA systems as these were expanded to allow these centralised control and safety management procedures to be managed electronically. These DSMs required even more detailed component/connectivity models and schematics than those needed by early OMSs as every possible isolation and earthing point on the networks had to be included. In territories such as the UK, therefore, the network component/connectivity models were usually developed in the DMS first, whereas in the USA these were generally built in the GIS.

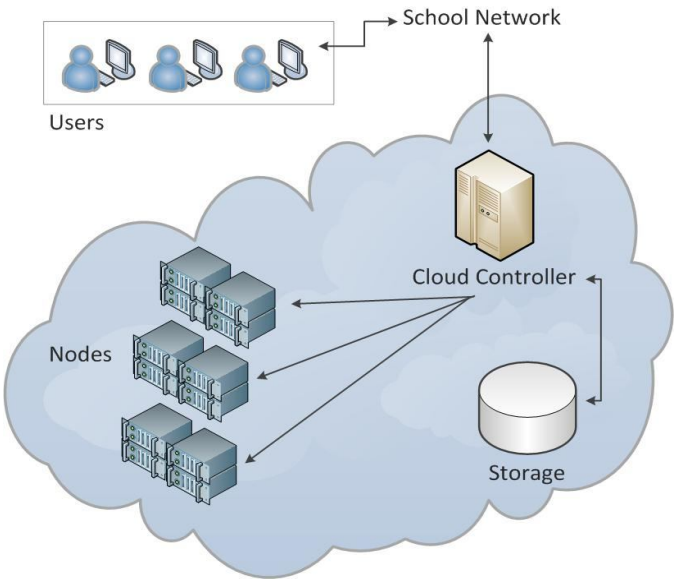


Figure 5: DSM in cloud

Distribution network usually covers over a large area and catering power to different customers at different voltage levels. So locating required sources and loads on a larger GIS/Operator interface is often very difficult. Panning & zooming provided with normal SCADA system GUI does not cover the exact operational requirement. Network connectivity analysis is an operator specific functionality which helps the operator to identify or locate the preferred network or component very easily. NCA does the required analyses and provides display of the feed point of various network loads.

Lattice Compression Algorithm

Lattice cryptography is arguably the most promising replacement for standard cryptography after the eventual coming of quantum computers. The most ubiquitous public-key cryptographic primitives, encryption schemes and digital signatures, already have somewhat practical lattice-based instantiations. In addition, researchers are rapidly discovering new lattice-based primitives, such as fully-homomorphic encryption multi-linear maps and attribute-based encryption that had no previous constructions based on classical number-theoretic techniques. Even though the above primitives are quite varied in their functionalities, many of them share the same basic building blocks. Thus an improvement in one of these fundamental building blocks, usually results in the simultaneous improvement throughout lattice cryptography. For example, the recent work on the lattice trapdoor generation algorithm resulted in immediate efficiency improvements in lattice-based hash-and-sign signatures, identity based encryption schemes, group signatures, and functional encryption schemes. In this work, propose an improvement of another such building block the rejection sampling procedure that is present in the most efficient constructions of lattice-based digital signatures authentication schemes blind signatures and zero-knowledge proofs used in multi-party computation As a concrete application, show that with our new algorithm, lattice-based digital signatures become completely practical.

Performance analysis

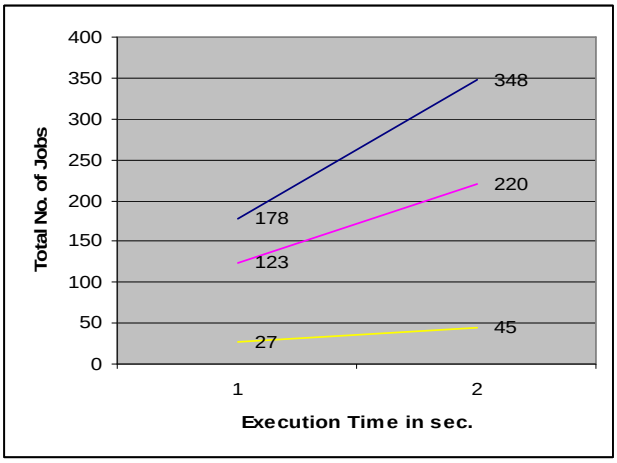


Figure 6: Performance analysis

Fig 6: shows the analysis of different no. of process and time taken with different algorithm

Conclusion and Future Work

In this paper, propose a Distributed System management and the Lattice compression Algorithm (LCA) to authenticate the data access from the cloud to ensure the security of the data over cloud and their data sets. Evaluation results demonstrate that the

secrecy and privacy of the data stored over the network is achieved efficiently and effectively.

In accordance with various data and computation intensive applications on cloud, intermediate data set management is becoming an important research area. Privacy preserving for intermediate data sets is one of important yet challenging research issues, and needs intensive investigation. With the contributions of this paper, we are planning to further investigate privacyaware efficient scheduling of intermediate data sets in cloud by taking privacy preserving as a metric together with other metrics such as storage and computation. Optimized balanced scheduling strategies are expected to be developed toward overall highly efficient privacy aware data set scheduling.

References

[1] Xuyun Zhang, Chang Liu, Surya Nepal, Suraj Pandey, and Jinjun Chen, "A Privacy Leakage Upper Bound Constraint-Based Approach for Cost-Effective Privacy Preserving of Intermediate Data Sets in Cloud", IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 24, NO. 6, JUNE 2013.

[2] 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.

[3] 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.

[4] 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.

[5] H. Takabi, J.B.D. Joshi, and G. Ahn, "Security and Privacy Challenges in Cloud Computing Environments," IEEE Security & Privacy, vol. 8, no. 6, pp. 24-31, Nov./Dec. 2010.

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

[7] D. Yuan, Y. Yang, X. Liu, and J. Chen, "On-Demand Minimum Cost Benchmarking for Intermediate Data Set Storage in Scientific Cloud Workflow Systems," J. Parallel Distributed Computing, vol. 71, no. 2, pp. 316-332, 2011.

[8] 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.

[9] H. Lin and W. Tzeng, "A Secure Erasure Code-Based Cloud Storage System with Secure Data Forwarding," IEEE Trans. Parallel and Distributed Systems, vol. 23, no. 6, pp. 995-1003, June 2012.

[10] N. Cao, C. Wang, M. Li, K. Ren, and W. Lou, "Privacy-Preserving Multi-Keyword Ranked Search over Encrypted Cloud Data," Proc.IEEE INFOCOM '11, pp. 829-837, 2011.

[11] M. Li, S. Yu, N. Cao, and W. Lou, "Authorized Private Keyword Search over Encrypted Data in Cloud Computing," Proc. 31st Int'l Conf. Distributed Computing Systems (ICDCS '11), pp. 383-392, 2011.

[12] C. Gentry, "Fully Homomorphic Encryption Using Ideal Lattices,"Proc. 41st Ann. ACM Symp. Theory of Computing (STOC '09), pp. 169-178, 2009.

[13] B.C.M. Fung, K. Wang, and P.S. Yu, "Anonymizing Classification Data for Privacy Preservation," IEEE Trans. Knowledge and Data Eng., vol. 19, no. 5, pp. 711-725, May 2007.

[14] B.C.M. Fung, K. Wang, R. Chen, and P.S. Yu, "Privacy-Preserving Data Publishing: A Survey of Recent Developments," ACM Computing Survey, vol. 42, no. 4, pp. 1-53, 2010.

[15] X. Zhang, C. Liu, J. Chen, and W. Dou, "An Upper-Bound Control Approach for Cost-Effective Privacy Protection of Intermediate Data Set Storage in Cloud," Proc. Ninth IEEE Int'l Conf. Dependable, Autonomic and Secure Computing (DASC '11), pp. 518-525, 2011.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] S.B. Davidson, S. Khanna, T. Milo, D. Panigrahi, and S. Roy, "Provenance Views for Module Privacy," Proc. 30th ACM SIGMOD-SIGACT-SIGART Symp. Principles of Database Systems (PODS '11), pp. 175-186, 2011.

[21] S.B. Davidson, S. Khanna, S. Roy, J. Stoyanovich, V. Tannen, and Y. Chen, "On Provenance and Privacy," Proc. 14th Int'l Conf. Database Theory, pp. 3-10, 2011.

[22] S.B. Davidson, S. Khanna, V. Tannen, S. Roy, Y. Chen, T. Milo, and J. Stoyanovich, "Enabling Privacy in Provenance-Aware Workflow Systems," Proc. Fifth Biennial Conf. Innovative Data Systems Research (CIDR '11), pp. 215-218, 2011.

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

[24] A. Machanavajjhala, D. Kifer, J. Gehrke, and M. Venkitasubramaniam,“L-Diversity: Privacy Beyond K-Anonymity,” ACM Trans.Knowledge Discovery from Data, vol. 1, no. 1, article 3, 2007.

[25] 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.

[26] W. Du, Z. Teng, and Z. Zhu, “Privacy-Maxent: Integrating Background Knowledge in Privacy Quantification,” Proc. ACM SIGMOD Int’l Conf. Management of Data (SIGMOD ’08), pp. 459-472, 2008.