

Authorized Public Auditing on Cloud Data

Mohammed Mujeer ulla

Assistant Professor,

Dept. of Information Science and Engineering,
HKBK College Of Engineering,
Bangalore -560045, India

VinayHegde,

Associate Professor

Dept. of Computer Science and Engineering
R.V College Of Engineering,
Bangalore -560059, India

Abstract— A new era in Information Technology is Cloud Computing, as it provides various scalable and elastic Information Technology services in pay-as-you-use basis, where the customers of cloud can reduce huge capital investments involved in IT infrastructure. In this aspect cloud users who use storage services no longer physically maintain direct control over the data that is stored in cloud, which makes data security as one of the major issues while using cloud. The earlier research work allows integrity data to be certified without possession of actual data file. When verification is done by a trusted third party, then such verification is known as Data Auditing and the person who does the auditing is called as Auditor.

Such schemes in reality suffer from several drawbacks:

- 1) A required process of Authentication/Authorization is unavailable between the cloud service provider and auditor. i.e anyone who is willing to challenge the cloud service provider to obtain the integrity of certain file, there by puts the quality of so called ‘auditing-as-a-service’ at risk.
- 2) The recent research work that was carried out on BLS signature can support updates full dynamic data on constant/fixed size of data blocks, this support is only towards fixed size blocks as basic unit which I call it as Coarse-grained updates. Due to which every small update would cause re-computation and updating of the authenticator for an entire file block, which results in over heads like higher usage of storage space and communication overheads. In this Project I would enable a formal analysis for all possible types of fine-grained updates and bring out a scheme that can fully support authorized auditing and fine grain update requests.

Keywords — Provable Data Possession, PDP, Homomorphic verifiable tags, storage security

I. INTRODUCTION

Cloud Computing in current Era is one of the influential innovations in computer science and technology in recent years. With recent advancements in Cloud research over last half century there is a perceived vision that cloud computing would be the 5th utility in Daily basis (After Water, electricity, telephone and gas). This computing utility would be able to provide basic essential needs of general community.

Defining Cloud Computing: “Cloud Computing is a technology to deliver applications as a services over the Internet and the hardware and system software in data centers that provide those services”.

The Cloud services are provided in form of

- **Software as a Service (SaaS):** Enables the consumer with the capability to use the provider’s application on a cloud infrastructure. Here the application is made available to various client devices through a thin client interface like web browser (web based e-mail). The cloud infrastructure management is not controlled by the consumer including server, network , operating system or even individual applications.
- **Platform as a Service (PaaS):** It enables the customer with the facility to deploy in cloud infrastructure , consumer created or acquired applications , generated using programming languages and tools supported by provider. Here there is no control from the consumer to manage or control over deployed applications.
- **Infrastructure as a Service (IaaS):** It enables the consumer with the capability of processing, storage, networks and some other fundamental computing resources and allows the consumer to deploy and run arbitrary software which can include operating systems and applications. The customer controls the operating systems, storage, deployed applications and possibly limited control of select network components.

Cloud computing is empowered by Virtualization technology, a technology which emerged in 1967 but for decades was made available on mainframe computers. In virtualization the host computer which runs an application called as Hypervisor; this creates one or more virtual machines which could simulate a physical computer so accurately that its simulations can run on any software from operating system. To end user applications. At hardware level cloud computing offers number of physical devices including processors , networking devices and hard drives.

The cloud identifies four deployment models described below:

- **Private Cloud:** the cloud infrastructure is operated for private organization. Management of it may be done by the organization or third party and may be available on premise or off premise.
- **Public Cloud:** The cloud infrastructure is made available to general public or a large industry group and the proprietor of such cloud is organization selling such cloud services.
- **Community Cloud:** here the cloud infrastructure is shared by several organization and is available for customers community of interest like mission critical, security requirements policy and compliance consideration). Community cloud is managed by third party or organizations and may be on premise or off remise.
- **Hybrid Cloud:** the hybrid cloud is a mixture or composition of two or more clouds like public, private or community cloud that remain unique entity that are bound together by proprietary or standardized technology that enables application and data probability like cloud bursting for load balancing between clouds.

Cloud computing is considered as one of the emerging technology in computing today to address number of issues. Some of the key characteristics of cloud computing are listed below:

- 1) **Infrastructure scalability:** New nodes could be added or removed from the network as can physical server with slight



modification to set up infrastructure and software. Depending on the demand cloud architecture can scale vertically or horizontally.

- 2) **Elasticity/Flexibility:** The provisioning of computing resources as and when needed avoiding human interaction. Provisioning of capabilities rapidly and elastically in some cases to quickly scale out or up.
- 3) **Broad network Access:** Capabilities could be made available over the network and are accessed through standard mechanism that promote usage of heterogeneous platform like laptops, mobile phones and Personal Digital assistance (PDA).
- 4) **Location Independence:** Here the customer has no knowledge or control over the perfect location of the resources provided, but he/she may be able to specify location at higher level of abstraction like country, state or data center.
- 5) **Cost effectiveness and Economy of scale:** regardless of deployment model cloud implementations tend to be as large as possible to En-cash the economies of scale. Huge cloud deployments are usually located close to cheap power stations and in low- priced real estate to reduce the costs.
- 6) **Sustainability:** comes through more efficient systems , improved resource utilization and carbon neutrality.

The different auditing schemes already have various characteristics potential risks, in efficiency's like inefficiency in processing small updates and inefficiency such as security risks in un authorised auditing requests. In my concept i will focus on higher support for small dynamic updates which benefits cloud storage server efficiency and scalability. To achieve this my scheme utilizes a Ranked Merkel hash tree (RMHT) and flexible data segmentation strategy . As well I will address potential security problem in supporting public verifiability to make the scheme more robust and secure which could be achieved by addition of three participating entities like client, CSS and third-party auditor (TPA).

Research contributions of the concept is summarized as follows:

- This is the first time that we formally analyze variety of different types of fine-grained dynamic update requests on variable-sized file blocks in single data set. To best of my knowledge this the first concept to propose a public auditing scheme on Merkle hash tree (MHT) and BLS signature that could support fine grained update requests. In contrast to peer schemes , this scheme supports variable size updates with size that is not restricted by the size of file blocks thereby allows scalability and extra flexibility in contrast to existing schemes.
- For enhanced security this scheme incorporates additional authorization process with intension of removing threats of pretending third party auditors or unauthorized audit challenge from malicious auditors.

This concept also investigates how improved efficiency can be achieved in frequent small updates which exist in many popular cloud and big data contexts.

Motivation: Cloud computing is a revolution today IT environment that enables IT services to be scalable and elastic in a pay-as-you-use fashion. The cloud users once they store the data in cloud they no longer have their control over the data , there by creates a security issues with respect to usage of cloud. Existing schemes suffer from several frequent disadvantages

- Primarily, a required authorization/authentication process is unavailable between the auditor and cloud service provider (CSP)

- Some of the earlier auditing schemes which we call coarse-grained support full dynamic data blocks only with respect to fixed size blocks as a unit. Due to which every small update would cause re-computation and updation of entire file block which results in higher communication and storage over heads. Based on this idea here is proposal for a technique that can dramatically reduce communication overheads for verifying small updates.

In future this Concept could be taken as Research step in domain of Green Computing to formally analyze Reduction in Millions of Kilo watts of Power consumed by Cloud Storage Server by just performing Dynamic Data updates.

Problem Statement:

The security and privacy of the data is one of the major issues in adoption of cloud computing [3],[4],[5]. In comparison with conventional system, the cloud users will be completely isolated with having direct control over their data. This concept will investigate the problem of integrity verification big data storage in cloud. This problem is known as data auditing [6][7] when verification is conducted by third party. From cloud users point of view it is known as 'Auditing as a service'.

In case of remote verification scheme, the cloud storage server (CSS) cannot certify the integrity proof of a given fraction of data to a verifier until whole data is intact.

To ensure stored data integrity this support is no less important than any data protection mechanism deployed by cloud service provider (CSP) [16], how much ever secure they pretend to be in that it will provide the verifier a piece of direct, trust worthy and real-timed intelligence of the integrity of cloud user's data through a challenge request. It is required that a improved and efficient data auditing scheme is to be conducted on regular intervals of time for users who ensure higher security demand for their data.

Objective:

- To formally analyze different types of fine grained dynamic data update requests on varying block sizes in a single data set, It can also be deployed in HPC (High Performance Computing) which has got a separate storage and computing nodes by developing Parallel processing Algorithms using OpenMP.
- To incorporate added authorization process with aim of eliminating bullying of unauthorized audit challenges from malicious or fake third party auditors.
- To investigate how to improve the Efficiency in verifying small updates as well .
- Compared to existing schemes, both hypothetical analysis & Experimental approach can significantly lower communication over heads. This can be assessed by making performance comparison through parallel programming across the clouds using HPC (High Performance Computing) by concept of either OpenMP which is a shared memory parallel programming / Java Multi threading concept.

Methodology:

- **Set Up:** The client will initially produce keying materials via KeyGen & FileProc , then upload the data to CSS (Cloud storage Server). The client will store a RMHT(Ranked Morkel Hash Tree) instead of MHT(Morkel Hash Tree) as meta data. More over the client will authorize the TPA by sharing a value.

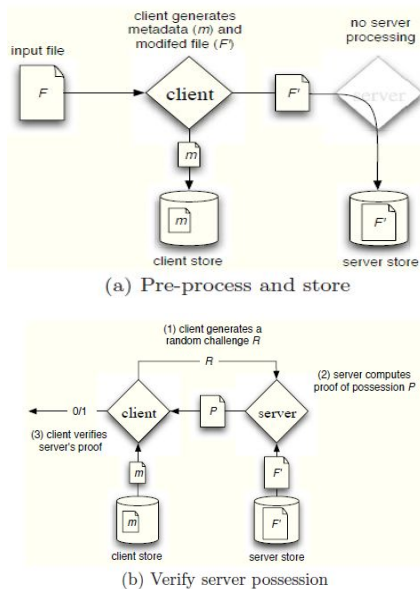


- **Verifiable Data Updating:** The CSS (Cloud storage Server) performs the client’s fine grained update request via perform update, then client runs Verifyupdate to check whether CSS has performed the updates on both the data blocks and their corresponding authenticators.
- **Challenge, Proof Generation & Verification:** Describes how the integrity of data stored on CSS is verified by TPA via Generate Challenge, GenProof and Verify.

II. RELATED WORK

Verification for the outsourced data storage for integrity has attracted extensive research interest. The first concept model for proof of retrievability (POR) was proposed by Jules [14]. Interestingly their scheme can only be applied to static data storages like library and achieve. In same year Ateniese came up with a similar model named ‘provable data possession’ [10]

Provable Data Possession [10] is a frame work is as shown in below figure



A PDP protocol verification of the outsourced storage site that retains file comprising of collection of n blocks. The Client C(owner of data) pre-processes the file , by produces a piece of data that is maintained and stored locally later it will transmit the file to the server S and finally deletes his own local file copy. The responsibility of server is to store the received file and respond to the challenges issued by the client. During pre-processing the client is allowed to altering of file that is to be stored at the server altering could be in form of expanding the file or include additional meta data to be stored at the server. Before deleting its local copy the client may execute data possession challenge to make sure that the server has successfully stored the file. And client is also allowed to perform file encryption prior to outsourcing the storage.

Homomorphic Verifiable Tags (HVTs): The Homomorphic Verifiable Tags (HVTs) is one of the building block for Provable Data Possession scheme. Which says that for a given message m(corresponding to a file block), we denote its Homomorphic verifiable tag as T_m . This tag would be stored on the server together with file F. Here the Homomorphic verifiable tag act as verification metadata for the file blocks and besides being unforgeble they also have following properties:

Homomorphic Tags: Given two values T_{m_i} and T_{m_j} one could combine them into a single value $T_{m_i+m_j}$ corresponding to sum of two messages m_i+m_j . In this construction HVT is a pair of values

(T_i, m, W_i) ,

W_i - is a random value obtained from a index i.

$T_{i,m}$ -is stored at server.

Since index i is never reused for computing tags & is used one time index. One common way to make sure that every tag uses a different index i is to use a global counter for i. The random value W_i is obtained by concatenating the index i to secret value which makes sure that W_i is unique and unpredictable each time a tag is computed. HVTs and their corresponding proofs have fixed constant size and are much smaller than actual file blocks.

3 Motivating Examples and Problem Analysis

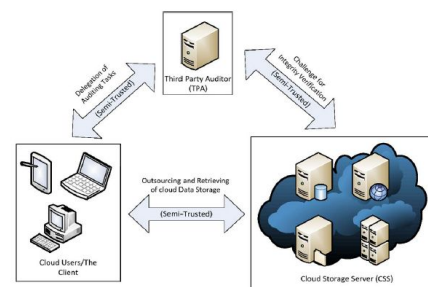
3.1 Motivating Examples

One of the reason that cloud is being extensively used is due to its elasticity feature. One such scenario is Vodafone Australia is using Amazon cloud to provide their users with mobile online video watching service. According to their statistics the number video requests per second (RPS) can be as much as 700 during public holiday or Friday nights in contrast with 70 in average in rest 90 percent of time and in this scenario if cloud computing is unavailable Vodafone has to purchase computing facilities to process 700 RPS but it is total waste of time and money. This is where cloud computing can be optimal for Vodafone in providing the extended service to its users on the fly. The other two companies that are using cloud is Amazon cloud is news.com.au and realestate.com

3.2 Problem analysis

3.2.1 Roles of participating entities:

Public data auditing can be supported by most of the PDP and POR schemes. Such schemes do contain 3 participating entities : client , CSS (Cloud storage server) and TPA (Third Party Auditor) The Relationship between three parties are shown in below diagram



In earlier model challenge message is very simple , every user of the cloud can send a challenge to Cloud storage server to obtain the proof of certain file blocks which could lead to following drawbacks:

A malicious party can launch a distributed denial-of-service attacks (DDOS) by overloading the cloud storage server by sending multiple challenge request and causing network congestion there by degenrating quality of service.

By challenging the CSS several times , the person trying to audit the cloud may get private sensitive information from proof returned by cloud storage server.

Verifiable Fine-Grained Dynamic Data Operations: The public auditing scheme that are used previous to this paper can also support full data dynamics [6],[7],[12]. These models can only perform insertions, deletions and modifications on fixed size blocks. In BLS-

signature based scheme [6],[7],[13],[15] with 80 bit security size of each data block is restricted by 160 bit prime group order p, each block is segmented into fixed number of 160- bit sectors. Such design is suitable to support variable-sized blocks, even though they have the advantage of shorter integrity proofs.

Support of coarse grained update is not recommendable though it provides integrity verification scheme with scalability, data updating operations could lead to complexity. For example in [6],[12] the verifiable update process introduced could not handle modifications or deletions in a size lesser than a block. CSS created a new block for every insertion. In such cases when there are many large number of small upgrades , the amount of space wasted is more. [Example the data block size that is recommended is 16k bytes [6],[12]. For every insertion of a 140-byte Twitter message , as much as 99% of newly allocated storage space is wasted- they cannot be reused until the block is deleted. These problems can be solved by use of fine grained data updates.

Assumption 1: All the data queries to the clients are honestly answered by Cloud Storage Server, i.e If a client asks his/her data on cloud then Cloud Storage Server will not try to cheat by giving in accurate answer.

The Proof of Retrieval and Provable Data Possession are two different model having different goals. The major difference is that the file is encoded with error correction code in the Proof of Retrieval model, this work is not restricted to either of the models.

4 The Proposed Scheme:

4.1 Preliminaries :

4.1.1 Bilinear Map:

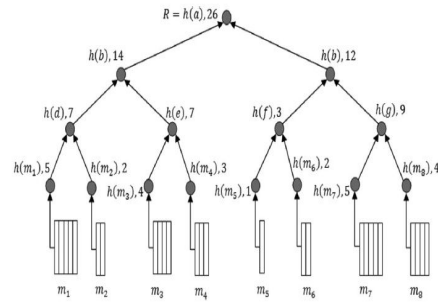
Let us assume that a group G is a gap Diffie-Hellman (GDH) group having prime order p. A Bilinear map is $e: G \times G \rightarrow GT$ is a multiplicative recurring group with prime order. A required e which is useful should have following properties:

- **Bilinearity:**
 $\forall m, n \in G \Rightarrow e(m^a, n^b) = e(m, n)^{ab}$
- **Non-degeneracy:**
 $\forall m \in G, m \neq 0 \Rightarrow e(m, m) \neq 1$; and
- **Computability:** e should be efficiently computable.

4.1.2 Ranked Merkle Hash Tree (RMHT): The Merkle Hash tree (MHT) is been used extensively by many cloud auditing researchers. This paper makes use of Merkle Hash Tree which is named here as RMHT. Just like our binary tree even our RMHT consists of a each node N having maximum of 2 child's nodes. According to update algorithm every non leaf node will constantly have one node N in RMHT T is represented as $\{H, r^N\}$ where H represents a Hash value and r^N represent rank of this node. Tree T is built as follows.

For a leaf node LN based on a message m_i , I have $H = h(m_i), r^LN = s_i$; A parent node of $N_1 = \{H_1, r^{N_1}\}$ and $N_2 = \{H_2, r^{N_2}\}$ is constructed as $N_p = \{h(H_1 || H_2), (r^{N_1} + r^{N_2})\}$

where k is a concatenation operator. A leaf node m_i 's is a set of hash values chosen from every of its upper level so that the root value R can be computed. For example, for the RMHT demonstrated in Fig. 2 $A \ A \ I \ \Omega_1 = \{h(m_2), h(e), h(d)\}$ According to the property of RMHT, I know that the number of hash values included in W_i equals the depth of m_i in T.



Example of Ranked Merkle Hash Tree

4.2 Frame Work and Definitions :

We would define the following block level fine grained update operations:

Varieties of Block level operation in Fine Grained Updates : consists of following types of operations:

Partial Modification PM - a successive fraction of certain block requires to be updated.

Complete Block Modification M - a complete block needs to be replaced by a set of data block

Block Deletion D - a whole block needs to be deleted from the tree structure

Block introduction I - a whole block needs to be created on the tree composition to contain newly inserted data and

Block Splitting SP - a portion of the data in a block needs to be removed out to form a new block to be inserted next to it.

The framework of public auditing scheme with data dynamic support is composed of series of algorithms. Keygen, FilePreProc, Challenge, Verify, Genproof , Perform Update and VerifyUpdate. KeyGen is a key generation algorithm that is run by the user to setup the scheme. SigGen is used by the user to generate verification metadata, GenProof is run by the cloud server to generate a proof of data storage correctness, while verify proof is run by TPA to audit the proof from the cloud server.

4.3 Our Scheme

Let us now describe our proposed scheme in the aim of supporting variable-sized data blocks, fine-grained dynamic data updates and third party auditing.

4.3.1 Overview

We will be describing our scheme in three parts:

Set Up: The client will initially produce keying materials via KeyGen & FileProc , then upload the data to CSS (Cloud storage Server). The client will store a RMHT (Ranked Morkel Hash Tree) instead of MHT (Morkel Hash Tree) as meta data. More over the client will authorize the TPA by sharing a value.

Verifiable Data Updating: The CSS (Cloud storage Server) performs the client's fine grained update request via perform update, then client runs Verifyupdate to check whether CSS has performed the updates on both the data blocks and their corresponding authenticators.

Challenge, Proof Generation & Verification : Describes how the integrity of data stored on CSS is verified by TPA via Generate Challenge, GenProof and Verify.



The below section describes our scheme in detail

4.3.2 Setup :

At first customer/client produces a secret value $r \in \mathbb{Z}_p$ and as well a generator g of G , then we compute $v = g^r$. A secret signing key pair $\{spk, ssk\}$ is chosen with respect to chosen provable secure signature scheme whose signing algorithm is denoted as $Sig()$. This algorithm outputs $\{ssk, r\}$ as a secret key sk and $\{spk, v, g\}$ as the public key pk .

FilePreProc(F,sk, SegReq) : as per the preemptively determined segmentation requirement SegReq (including s_{max} , a predefined upper-bound of the number of segments per block), segments the file F into $F = \{m_{ij}\}$, $i \in [1, l]$, $j \in [1, s]$, $s_i \in [1, s_{max}]$ i.e F is segmented into a total of l blocks, with the i^{th} block having s_i segments.

4.3.3 Prepare for Authorization

The client asks (his/her choice of) TPA for its ID VID (for security,

VID is used for authorization only). TPA will then return its ID, encrypted with the client's public key. The client will then compute $sig_{AUTH} = Sig_{ssk}(AUTH || t || VID)$ and sends sig_{AUTH} along with the auditing delegation request to TPA for it to create a challenge later on. Dissimilar from present schemes, after the execution of the above two algorithms, the client will keep the RMHT 'skeleton' with only ranks of each node and indices of each file block to reduce fine-grained update requests to block level operations. The client then sends $\{F, t, \Phi, sig, AUTH\}$ to CSS and deletes $\{F, F, t, \Phi, sig\}$ from its local storage. The CSS will construct an RMHT T based on m_i and keep T stored with $\{F, F, t, \Phi, sig\}$ for later verification, which should be identical to the tree spawned at client side which was done earlier.

References

- [1] R. Buyya, C.S. Yeo, S. Venugopal, J. Broberg, and I. Brandic, "Cloud Computing and Emerging IT Platforms: Vision, Hype, Reality for Delivering Computing as the 5th Utility," *Future Gen. Comput. Syst.*, vol. 25, no. 6, pp. 599-616, June 2009.
- [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," *Commun. ACM*, vol. 53, no. 4, pp. 50-58, Apr. 2010.
- [3] Customer Presentations on Amazon Summit Australia, Sydney, 2012, accessed on: March 25, 2013. [Online]. Available: <http://aws.amazon.com/apac/awssummit-au/>.
- [4] J. Yao, S. Chen, S. Nepal, D. Levy, and J. Zic, "TrustStore: Making Amazon S3 Trustworthy With Services Composition," in *Proc. 10th IEEE/ACM Int'l Symposium on Cluster, Cloud and Grid Computing (CCGRID)*, 2010, pp. 600-605.
- [5] D. Zisis and D. Lekkas, "Addressing Cloud Computing Security Issues," *Future Gen. Comput. Syst.*, vol. 28, no. 3, pp. 583-592, Mar. 2011.
- [6] Q. Wang, C. Wang, K. Ren, W. Lou, and J. Li, "Enabling Public Auditability and Data Dynamics for Storage Security in Cloud Computing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 22, no. 5, pp. 847-859, May 2011.
- [7] C. Wang, Q. Wang, K. Ren, and W. Lou, "Privacy-Preserving Public Auditing for Data Storage Security in Cloud Computing," in *Proc. 30th IEEE Conf. on Comput. and Commun. (INFOCOM)*, 2010, pp. 1-9.
- [8] G. Ateniese, R.D. Pietro, L.V. Mancini, and G. Tsudik, "Scalable and Efficient Provable Data Possession," in *Proc. 4th Int'l Conf. Security and Privacy in Commun. Netw. (SecureComm)*, 2008, pp. 1-10.
- [9] G. Ateniese, R. Burns, R. Curtmola, J. Herring, O. Khan, L. Kissner, Z. Peterson, and D. Song, "Remote Data Checking Using Provable Data Possession," *ACM Trans. Inf. Syst. Security*, vol. 14, no. 1, May 2011, Article 12.
- [10] G. Ateniese, R.B. Johns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song, "Provable Data Possession at Untrusted Stores," in *Proc. 14th ACM Conf. on Comput. and Commun. Security (CCS)*, 2007, pp. 598-609.
- [11] R. Curtmola, O. Khan, R.C. Burns, and G. Ateniese, "MR-PDP: Multiple-Replica Provable Data Possession," in *Proc. 28th IEEE Conf. on Distrib. Comput. Syst. (ICDCS)*, 2008, pp. 411-420.
- [12] C. Erway, A. Ku'pcu', C. Papamanthou, and R. Tamassia, "Dynamic Provable Data Possession," in *Proc. 16th ACM Conf. on Comput. and Commun. Security (CCS)*, 2009, pp. 213-222.
- [13] Y. Zhu, H. Hu, G.-J. Ahn, and M. Yu, "Cooperative Provable Data Possession for Integrity Verification in Multi-Cloud Storage," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 12, pp. 2231-2244, Dec. 2012.
- [14] A. Juels and B.S. Kaliski Jr., "PORs: Proofs of Retrievability for Large Files," in *Proc. 14th ACM Conf. on Comput. and Commun. Security (CCS)*, 2007, pp. 584-597.
- [15] H. Shacham and B. Waters, "Compact Proofs of Retrievability," in *Proc. 14th Int'l Conf. on Theory and Appl. of Cryptol. and Inf. Security (ASIACRYPT)*, 2008, pp. 90-107.
- [16] S. Nepal, S. Chen, J. Yao, and D. Thilakanathan, "DIAAS: Data Integrity as a Service in the Cloud," in *Proc. 4th Int'l Conf. on Cloud Computing (IEEE CLOUD)*, 2011, pp. 308-315.
- [17] Y. He, S. Barman, and J.F. Naughton, "Preventing Equivalence Attacks in Updated, Anonymized Data," in *Proc. 27th IEEE Int'l Conf. on Data Engineering (ICDE)*, 2011, pp. 529-540.
- [18] E. Naone, "What Twitter Learns From All Those Tweets," in *Technology Review*, Sept. 2010, accessed on: March 25, 2013. [Online]. Available: <http://www.technologyreview.com/view/420968/what-twitter-learns-from-all-those-tweets/> [19] X. Zhang, L.T. Yang, C. Liu, and J. Chen, "A Scalable Two-Phase Top-Down Specialization Approach for Data Anonymization Using MapReduce on Cloud," *IEEE Trans. Parallel Distrib. Syst.*, vol. 25, no. 2, pp. 363-373, Feb. 2014.
- [20] S.E. Schmidt, "Security and Privacy in the AWS Cloud," presented at the Presentation Amazon Summit Australia, Sydney, Australia, May 2012, accessed on: March 25, 2013. [Online].