Proofs of Retrievability with Public Verifiability and Constant Communication Cost in Cloud

Jiawei Yuan
University of Arkansas at Little Rock
Little Rock, USA
jxyuan@ualr.edu

Shucheng Yu
University of Arkansas at Little Rock
Little Rock, USA
sxyu1@ualr.edu

ABSTRACT

For data storage outsourcing services, it is important to allow data owners to efficiently and securely verify that the storage sever stores their data correctly. To address this issue, several proof-of-retrievability (POR) schemes have been proposed wherein a storage sever must prove to a verifier that all of a client's data are stored correctly. While existing POR schemes offer decent solutions addressing various practical issues, they either have a non-trivial (linear or quadratic) communication complexity, or only support private verification, i.e., only the data owner can verify the remotely stored data. It remains open to design a POR scheme that achieves both public verifiability and constant communication cost simultaneously.

In this paper, we solve this open problem and propose the first POR scheme with public verifiability and constant communication cost. In our proposed scheme, the message exchanged between the prover and verifier is composed of a constant number of group elements. Different from existing private POR constructions, our scheme allows public verification and releases the data owners from burden of staying online. Thorough analysis shows that our proposed scheme is efficient and practical. We prove the security of our scheme based on the Computational Diffie-Hellman Problem, the Strong Diffie-Hellman assumption and the Bilinear Strong Diffie-Hellman assumption.

Categories and Subject Descriptors

H.3.2 [**Information Storage and Retrieval**]: Information Storage; D.4.6 [**Security and Protection**]: Cryptographic controls

General Terms

Storage, Integrity, Security, Algorithm

Keywords

Proofs of Retrievability, Cloud Storage, Public Verification, Integrity Check, Constant Communication, Polynomial Commitment

1. INTRODUCTION

Due to a number of unprecedented advantages – resource elasticity, on-demand self-service, location independent resource polling, etc[22], cloud storage is increasingly attracting customers including both organizations and individuals. Currently, millions of users have been using cloud storage services including Amazon S3, Microsoft Skydrive, Google Cloud Storage, iCloud, and Dropbox. The users can access their cloud storage services from various devices such as laptops and mobile phones through wired or wireless networks.

Despite the proliferation of cloud storage service, there also raise security concerns since outsourcing data to the cloud make the data owner lose physical control over the storage sites. One significant concern, among the many, is data integrity, i.e., whether or not the cloud server indeed stores its clients' data correctly. Recently, a number of data loss events have been reported for best-known storage providers[2, 1, 3, 4], including Amazon S3 and Dropbox. To ensure the clients' confidence in the integrity of data remotely stored in the cloud, a reliable proof-of-retrievability (POR)[16] system is desirable. Specifically, in a POR system the data storage server must prove that it indeed stores the clients' data correctly. And the client shall be able to verify whether or not the proof generated by the server is valid. In practice the performance of a POR system is mainly determined by the following factors: 1) the communication cost between server and client during the verification process; 2) the way of verification, i.e., private verifiability or public verifiability; 3) the storage overhead on server side and client side; 4) the computation cost introduced to the server and the client for each verification. Toward designing a practical POR system, in past years several techniques[20, 11, 15] have been proposed.

Utilizing the idea of aggregating block integrity values, Shacham and Waters (SW)[20] proposed two POR schemes – one supports private POR verification and the other achieves public POR verification. While the two proposed schemes offer fast and flexible verification, their communication costs both grow linearly to the number of elements in the data block. For the integrity check of large data files, such a communication complexity represents a considerable cost to the system and can even make the POR system unaffordable to some bandwidth constrained users (e.g., mobile phone users with limited data plan). In Ref.[10], Dodis et al. enhanced the SW schemes by reducing the challenge message size, but

the response size is still linear to the number of elements in the data block. To overcome the limitation in Ref.[20], Xu et al.[15] constructed a POR scheme with constant communication cost by utilizing a recently proposed polynomial commitment technique. The main drawback of this scheme, however, is that it only supports private verification, which means that all the verifications must be performed by the data owner. In large-scale systems with many clients, this kind of verification will cause a heavy burden to the data owner in terms of both communication cost and computation cost because all the integrity check requests from other clients have to be performed by the owner. And the owner also has to stay online to provide the verification service. To our best knowledge, there is no existing POR solution that offers public verifiability with a constant communication cost [15].

In this paper, we fill this gap with our proposed secure and practical POR scheme, namely *PCPOR*. The main idea of our scheme can be summarized as follows: the data owner first breaks an erasure coded file into n blocks $\{m_i\}$,$1 \leq i \leq n$ and generates an authentication tag $\{\sigma_i\}$ for each block. Then, all the data blocks and tags are outsourced to the cloud storage server. When a client wants to retrieve the data from the server and check whether the data are stored correctly, he generates a challenge message and sends it to the cloud server. On receiving the challenge message, the cloud server generates a proof of the correctness of data storage based on the received message, the public key information and the previously stored tags, and then returns it to the client as the response. The client, having received the response, executes the verification algorithm to check data integrity using the public key information. In our *PCPOR* scheme, any client or auditing third-party can perform the verification process without contacting the data owner. Therefore, the data owner can stay off-line after having outsourced his data. To reduce the communication complexity, we first tailor a constant size polynomial commitment technique[17] and design a novel proof generation algorithm in which the proof information is aggregated into a polynomial. With our design, the size of the proof is made constant and independent to the number of elements in the data block. We further reduce the complexity of the challenge message to a constant level through uniquely integrating the similar techniques of Dotis et al.[10] with *PCPOR*. Despite the constant overall communication complexity, the storage overhead and the computational complexity introduced by *PCPOR* for integrity check are still comparable to existing schemes[20, 11, 15] thanks to our novel design. We conduct thorough analysis and show that our proposed scheme is efficient and practical. Based on the Computational Diffie-Hellman Problem (CDH), the Strong Diffie-Hellman (SDH) assumption and the Bilinear Strong Diffie-Hellman (BSDH) assumption, we prove the correctness and soundness of our proposed scheme.

Our main contributions can be summarized as below.

- We construct the first secure and efficient POR scheme with constant communication cost and public verifiability.
- Our proposed scheme omits the necessity for trade-off between communication cost and storage cost as in the public SW scheme[20].
- We formally prove the security of our proposed scheme under the standard model. Thorough analysis validates the advantages of our proposed scheme.

The rest of this paper is organized as follows. In Section 2, we review and discuss the related works; Section 3 describes the system model, security model and assumptions; We introduce the technique preliminaries of our work in Section 4, which is followed by the construction and security proof of our proposed scheme in Section 5. We evaluate the performance of our scheme in Section 6. Discussions about our paper are provided in Section 7. We conclude the paper in Section 8

2. RELATED WORK

In Ref.[16], Juels et al. first defined the POR formally, which allows a storage server to convince a client that it can correctly retrieve a file previously stored at the server. In their proposed POR scheme, disguised blocks hidden among regular file blocks are utilized to detect data modified by the server. However, the communication cost in this scheme is linear to the number of elements in each erasure coded file block. What is more, the number of challenges supported by this scheme is fixed a priori and thus limits its application. With similar purpose of Ref.[16], Ateniese et al.[5] proposed an efficient but weaker provable data possession model using homomorphic authentication tag. Nevertheless, an adversary was later introduced for this scheme by Shacham and Waters[20], which can answer a fraction of queries correctly with non-negligible probabilities.

To omit the limitation in Juels et al.'s POR scheme[16], Shacham and Waters (SW) [20] proposed two fast POR schemes based on the homomorphic linear authenticators[5], which enables the storage server to reduce the proof complexity by aggregating the authentication tags of individual file blocks. In their constructions, one proposed POR schemes supports private verification based on pseudorandom functions (PRFs) [12] and the other one achieves public verifiability by utilizing BLS signatures[7]. Compared with the scheme in Ref.[16], the communication cost for proof response in Ref.[20] is reduced to $\frac{1}{\lambda}$ of Ref.[16]'s and it can support unlimited number of challenges. At the same time, they first provide a security proof against arbitrary adversaries in the formal POR model. However, in SW schemes, the communication complexity for proof response is still linear to the number of elements in each erasure coded file block.

Following SW schemes[20], several POR schemes are proposed recently to enhance it in terms of communication cost. In Ref.[10], by using a $(\gamma, \delta) - hitter$ introduced by Goldreich[18], Dodis et.al. reduce the size of challenge message to $\frac{1}{\lambda}$ of Ref.[20]'s. Nevertheless, no change is made to the response size in this scheme, which is still linear to the number of elements in a data block. To further improve POR scheme and overcome the limitations in previous ones, Xu et.al[15] proposed a private POR scheme with constant communication cost with a polynomial based construction. In their scheme, the limitation of communication cost in SW's private POR scheme is omitted. However, their scheme requres the data owner to stay online and help the client to perform all the verification tasks. These onerous tasks will inevitably introduce heavy communication cost and computation cost to the data owner, especially when multiple clients submit verification requests at the same time. How to support POR

scheme with public verifiability and constant communication cost is pointed out as an open problem.

3. MODEL AND ASSUMPTION

3.1 System Model

In this work, we follow the POR model that is consistent with most existing POR schemes[16, 20, 8, 15]. We consider a POR system that has three participating entities: *Data owner*, *Client*, and *Cloud server*. Data owner has a collection of data and stores them on cloud server after erasure coding together with the corresponding authentication tags. The client who shares the stored data with the owner can access it with integrity check, which can also be performed as an independent procedure. To check the integrity of data files, the client generates a challenge message and sends it to the cloud server. The cloud server then responds the computed proof for the selected file blocks to the client. After receiving the proof, the client can verify the integrity of data files through the verification algorithm. W.l.o.g., we define the 1-round version of our POR model, which contains four algorithms, *KeyGen*, *Setup*, *Prove* and *Verify*, as below:

- **KeyGen:** Given a selected security parameter λ, the randomized *KeyGen* algorithm outputs the system public key and private key as (PK, SK).
- **Setup:** Given a data file $M \in \{0,1\}^*$ and the public-private key pair (PK, SK), the *Setup* algorithm generates the encoded file $\hat{M}$ as well as the corresponding authentication tag σ, which will be stored on the server.
- **Prove:** Given the public key PK, encoded file $\hat{M}$, authentication tag σ and a challenge message *Chall*, the *Prove* algorithm produces a proof response *Prf*.
- **Verify:** Given the public key PK and the *Prf*, the Verify algorithm checks the data integrity and outputs result as either accept or reject.

3.2 Security Model

We consider the storage server as untrusted and potentially malicious, which is consistent with existing POR schemes [16, 20, 15]. In our construction, we would like our POR scheme to be correct and sound. For the correctness of our scheme, we require that our *Verify* algorithm accepts a valid proof generated from all key pairs (PK, SK), all files $M \in \{0,1\}^*$, all encoded files $\hat{M}$ and authentication tags σ. For the soundness of our scheme, if any malicious cloud server can generate a proof and convince the *Verify* algorithm that it actually stores $\hat{M}$ correctly, it has to yield up the right $\hat{M}$ for the proof generation. By following Ref.[16, 20, 15], we define the security game for the soundness of our POR scheme as below.

DEFINITION 3.1. *Let $\nabla = (KeyGen, Setup, Prove, Veiry)$ be a POR scheme and A be a probabilistic polynomial-time adversary. Consider the following security game among a trust authority, a challenger and A.*

- The trust authority runs $KeyGen(1^\lambda) \rightarrow (PK, SK)$ and gives PK to the adversary A.
- The adversary A chooses a data file M and sends it to the trust authority. The authority then runs $Setup(M, SK, PK) \rightarrow (\sigma, \hat{M})$ and responds the encoded data file $\hat{M}$ together with the authentication tag σ back to A.
- With regard to the data file M chosen by adversary A, the challenger generates a random challenge message *Chall* and sends it to A.
- According to the received challenge *Chall*, the adversary A generates a file $\hat{M}', \hat{M}' \neq \hat{M}$ since he might have modified or lost some part of $\hat{M}$. A then produces a proof response *Prf* by running an arbitrary algorithm $Art(\hat{M}', \sigma, PK) \rightarrow Prf$ rather than the *Prove* algorithm. The proof response *Prf* is sent back to the challenger.
- The challenger verifies *Prf* by running *Verify* algorithm with *Prf* and PK. The output of $Verify(Prf, PK)$ is denoted as *Rst*.
- The adversary wins the game if and only if he can produce a Prf with data file $\hat{M}', \hat{M}' \neq \hat{M}$ and make the challenger generate *Rst* as *accept* through the *Verify* algorithm.

We say that ∇ is sound if any probabilistic polynomial-time adversary A can win the game with at most a negligible probability.

3.3 Assumption

DEFINITION 3.2. ***Computational Diffie-Hellman (CDH) Problem**[9]*

Let $a, b \overset{R}{\leftarrow} Z_p^*$. Given input as (g, g^a, g^b), where g is a generator of a cyclic group G of order p. It is computationally intractable to compute the value g^{ab}.

DEFINITION 3.3. ***t-Strong Diffie-Hellman (t-SDH) Assumption**[6]*

Let $\alpha \overset{R}{\leftarrow} Z_p^*$. Given input as a $(t+1)-tuple$ $(g, g^\alpha, \cdots, g^{\alpha^t}) \in G^{t+1}$, where G is a cyclic group of prime order p and g is the generator of G. For any probabilistic polynomial time adversary(PPT_{Adv}), the probability $Pr[PPT_{Adv}(g, g^\alpha, \cdots, g^{\alpha^t}) = (c, g^{\frac{1}{\alpha+c}})]$ is negligible for any value of $c \in Z_p^*/-\alpha$.

DEFINITION 3.4. ***t-Bilinear Strong Diffie-Hellman (t-BSDH) Assumption**[13]*

Let $\alpha \overset{R}{\leftarrow} Z_p^*$. Given input as a $(t+1)-tuple$ $(g, g^\alpha, \cdots, g^{\alpha^t}) \in G^{t+1}$, where G is a multiplicative cyclic group of prime order p and g is the generator of G. For any probabilistic polynomial time adversary (PPT_{Adv}), the probability $Pr[PPT_{Adv}(g, g^\alpha, \cdots, g^{\alpha^t}) = (c, e(g,g)^{\frac{1}{\alpha+c}})]$ is negligible for any value of $c \in Z_p^*/-\alpha$.

4. TECHNIQUE PRELIMINARIES

4.1 Bilinear Map

A bilinear map[7] is a map $e : G \times G \rightarrow G_1$, where G and G_1 are two multiplicative cyclic groups of the same prime order p. A bilinear map has the following properties:

- **Bilinear:** For all $g_1, g_2 \in G$ and $a, b \overset{R}{\leftarrow} Z_p^*$, $e(g_1^a, g_2^b) = e(g_1, g_2)^{ab}$.
- **Computable:** There exists a computable algorithm that can compute e efficiently.
- **Non-degenerate:** For $g \in G$, $e(g, g) \neq 1$

4.2 Constant Size Polynomial Commitment

A secure polynomial commitment scheme enables a committer to commit to a polynomial with a short string, which can be used by a verifier to confirm claimed evaluations of the committed polynomial. Utilizing an algebraic property of polynomials $f(x) \in Z[x]$: $(x - r)$ perfectly divides the polynomial $f(x) - f(r), r \overset{R}{\leftarrow} Z_p^*$, Kate et.al.[17] proposed a polynomial commitment scheme with constant communication size. In their scheme, a committer of polynomial $f(x)$ can generate a proof with constant size to verify the correctness of the polynomial evaluation $f(r)$, where $x = r$ is an index on the polynomial. Specifically, we summarize Kate et.al.'s scheme[17] as below.

- $Setup(1^\lambda, s)$**:** Given a security parameter λ and a fixed number s, a trust authority generates a public-private key pair (PK, SK):

  $$PK = (G, G_1, g, g^\alpha, \cdots, g^{\alpha^{s-1}})\ SK = \alpha \overset{R}{\leftarrow} Z_p^*$$

  where G and G_1 are two multiplicative cyclic groups with prime order $p(\lambda$ bits security), g is the generator of G and $e : G \times G \rightarrow G_1$.
- $Commit(PK, f_{\vec{m}}(x))$**:** For polynomial $f_{\vec{m}}(x) \in Z_p[x]$ with coefficient vector $\vec{m} = (m_0, m_1, \cdots, m_{s-1}) \overset{R}{\leftarrow} Z_p^*$, a committer computes the commitment $C = g^{f_{\vec{m}}(\alpha)} \in G$ and publishes C.
- $CreateWitness(PK, f_{\vec{m}}(x), r)$**:** For any index $r \overset{R}{\leftarrow} Z_p^*$, the committer divides polynomial $f_{\vec{m}}(x) - f_{\vec{m}}(r)$ with $(x - r)$ and outputs $\vec{w}$ as result, where $\vec{w} = (w_0, w_1, \cdots, w_{s-1})$ and $f_{\vec{w}}(x) \equiv \frac{f_{\vec{m}}(x) - f_{\vec{m}}(r)}{(x-r)}$. Then, the witness ψ is computed as $\psi = \prod_{j=0}^{s-1} (g^{\alpha^j})^{w_j} = g^{f_{\vec{w}}(\alpha)}$ based on PK.
- $VerifyEval(PK, C, r, f_{\vec{m}}(r), \psi)$**:** A verifier verifies that $f_{\vec{m}}(r)$ is the evaluation at the index r of the polynomial committed to by C as:

  $$e(C, g) \overset{?}{=} e(\psi, g^\alpha / g^r) \cdot e(g, g)^{f_{\vec{m}}(r)}$$

Due to the space limitation, please refer to Ref.[17] for the detail correctness and security proofs of this scheme.

5. CONSTRUCTION OF PCPOR

5.1 Scheme Description

Let $e : G \times G \rightarrow G_1$ and H be the one-way hash function[14], where G is a multiplicative cyclic group of prime order p and g, u be two random generators of G. We define $f_{\vec{c}(x)}$ as a polynomial with coefficient vector $\vec{c} = (c_0, c_1, \cdots, c_{s-1})$, where s is the number of elements in each data block. Our *PCPOR* scheme is described as follows.

- $KeyGen(1^\lambda) \rightarrow (PK, SK)$**:**

  Choose a random prime $p(\lambda$ bits security) and generate a random signing keypair $((spk, ssk) \overset{R}{\leftarrow} SKg)$ using BLS signature[6]. Choose two random numbers $\alpha, \epsilon \overset{R}{\leftarrow} Z_p^*$ and compute $v \leftarrow g^\epsilon$, $\kappa \leftarrow g^{\alpha\epsilon}$ as well as $\{g^{\alpha^j}\}_{j=0}^{s-1}$. Then, the public and private keys are

  $$PK = \{p, v, \kappa, spk, u, \{g^{\alpha^j}\}_{j=0}^{s-1}\}, SK = \{\epsilon, ssk, \alpha\}$$

- $Setup(PK, SK, M) \rightarrow (M^*, \sigma, \tau)$**:**

  Given a data file M, obtain M' by applying erasure code(e.g., Reed-Solomon code[19]). Split M' into n blocks, each of which is s elements long: $\{m_{ij}\}$, $1 \leq i \leq n$, $0 \leq j \leq s - 1$. Choose a random file name *name* from some sufficiently large domain ($e.g., Z_p^*$). Let τ_0 be "$name||n$"; the file tag τ is τ_0 together with a signature on τ_0 under ssk: $\tau \leftarrow \tau_0 || SSig_{ssk}(\tau_0)$. For each data $block_i, 1 \leq i \leq n$, an authentication tag is computed as:

  $$\sigma_i = (u^{H(name||i)} \cdot \prod_{j=0}^{s-1} g^{m_{ij}\alpha^j})^\epsilon \quad (1)$$
  $$= (u^{H(name||i)} \cdot g^{f_{\vec{\beta_i}}(\alpha)})^\epsilon$$

  where $\vec{\beta_i} = \{\beta_{i,0}, \beta_{i,1}, \cdots, \beta_{i,s-1}\}$ and $\beta_{i,j} = m_{i,j}$. The processed file M^* together with the authentication tags σ_i are outsourced for storage, where M^* is $\{m_{i,j}\}, 1 \leq i \leq n, 0 \leq j \leq s - 1$.
- $Verify(PK, \tau) \rightarrow Chall$**:**

  Stage 1: Verify the signature on τ: if the signature is not valid, reject and halt; otherwise, parse τ to recover *name* and n. Now, choose a random $l-$element subset I of the set $[1, n]$ and a random number $\rho \overset{R}{\leftarrow} Z_p^*$. Produce the challenge message as

  $$Chall = \{r, \rho, I\}$$

  where $r \overset{R}{\leftarrow} Z_p^*$. Challenge the data storage server with $Chall$.
- $Prove(PK, M^*, Chall, \tau) \rightarrow (\psi, y, \sigma)$**:**

  Parse the challenge message $Chall$ as $\{r, \rho, I\}$ and generate a l-element set $Q = \{(i, v_i)\}$, where $i \in I, v_i = \rho^i \ mod\ p$. Based on the processed file $M^* = \{m_{i,j}\}, 1 \leq i \leq n, 0 \leq j \leq s - 1$ and σ_i, compute

  $$\sigma = \prod_{(i,v_i) \in Q} \sigma_i^{v_i} \quad (2)$$

  We denote vector

  $$\vec{A} = (\sum_{(i,v_i) \in Q} v_i m_{i,0}, \cdots, \sum_{(i,v_i) \in Q} v_i m_{i,s-1})$$

  Compute

  $$y = f_{\vec{A}}(r) \quad (3)$$

  As we mentioned before, polynomials $f(x) \in Z[x]$ have the algebraic property that $(x-r)$ perfectly divides the polynomial $f(x)-f(r), r \overset{R}{\leftarrow} Z_p^*$. Now, divide the polynomial $f_{\vec{A}}(x)-f_{\vec{A}}(r)$ with $(x-r)$ using polynomial long

division, and denote the coefficients vector of the resulting quotient polynomial as $\vec{w} = (w_0, w_1, \cdots, w_{s-1})$, that is, $f_{\vec{w}}(x) \equiv \frac{f_{\vec{A}}(x) - f_{\vec{A}}(r)}{x-r}$. Produce

$$\psi = \prod_{j=0}^{s-1} (g^{\alpha^j})^{w_j} = g^{f_{\vec{w}}(\alpha)} \quad (4)$$

Respond $Prf = \{\psi, y, \sigma\}$.

- $Verify(PK, Prf) \rightarrow Rst$**:**

  Stage 2: After receiving the proof response Prf, compute

  $$\eta_i = u^{H(name||i)v_i}, (i, v_i) \in Q \quad (5)$$

  $$\eta = \prod_{(i,v_i)\in Q} \eta_i \quad (6)$$

  Parse Prf as $\{\psi, y, \sigma\}$ and check

  $$e(\eta, v) \cdot e(\psi, \kappa \cdot v^{-r}) \stackrel{?}{=} e(\sigma, g) \cdot e(g^{-y}, v) \quad (7)$$

  where $v = g^{\epsilon}, \kappa = g^{\epsilon\alpha}$ in PK. If Eq.7 holds, then output Rst as accept; otherwise, output Rst as reject.

- *Correctness:*

  For a storage server who honestly responds to the challenge with a $Prf = \{\psi, y, \sigma\}$, we can analyze the correctness of Eq.7 as:
  Left Part:

  $$\begin{aligned} & e(\eta, v) \cdot e(\psi, \kappa \cdot v^{-r}) \quad (8) \\ &= e(u, g)^{\epsilon(\sum_{(i,v_i)\in Q} t_i)} \cdot e(g^{f_{\vec{w}}(\alpha)}, g^{\epsilon(\alpha - r)}) \\ &= e(u, g)^{\epsilon(\sum_{(i,v_i)\in Q} t_i)} \cdot e(g, g)^{\frac{f_{\vec{A}}(\alpha) - f_{\vec{A}}(r)}{\alpha - r} \cdot \epsilon(\alpha - r)} \\ &= e(u, g)^{\epsilon(\sum_{(i,v_i)\in Q} t_i)} \cdot e(g, g)^{\epsilon(f_{\vec{A}}(\alpha) - f_{\vec{A}}(r))} \end{aligned}$$

  Right Part:

  $$\begin{aligned} & e(\sigma, g) \cdot e(g^{-y}, v) \quad (9) \\ &= e(u^{\epsilon(\sum_{(i,v_i)\in Q} t_i)} \cdot g^{\epsilon f_{\vec{A}}(\alpha)}, g) \cdot e(g^{-y}, v) \\ &= e(u^{\epsilon(\sum_{(i,v_i)\in Q} t_i)}, g) \cdot e(g^{\epsilon f_{\vec{A}}(\alpha)}, g) \cdot e(g, g)^{-\epsilon f_{\vec{A}}(r)} \\ &= e(u, g)^{\epsilon(\sum_{(i,v_i)\in Q} t_i)} \cdot e(g, g)^{\epsilon(f_{\vec{A}}(\alpha) - f_{\vec{A}}(r))} \end{aligned}$$

  where $t_i = H(name||i)v_i$. From the above Eq.8 and Eq.9, it is easy to see that the scheme is correct if the storage sever generate the Prf honestly.

5.2 Security Proof

In this section, we prove that the underlying authenticator is unforgeable.

THEOREM 5.1. *If $g^{f_{\vec{c}}(\alpha)}$ can be forged by an existed probabilistic polynomial time adversary A, we can construct an algorithm B that uses A to efficiently compute the solution to t-SDH problem.*

PROOF. Suppose there exists a probabilistic polynomial time adversity A that can forge $f_{\vec{c_1}}(\alpha)$ such that $g^{f_{\vec{c_1}}(\alpha)} = g^{f_{\vec{c}}(\alpha)}$, where $f_{\vec{c}}(x)$ and $f_{\vec{c_1}}(x)$ are known to A. A can construct another polynomial $f_{\vec{c_2}}(x) = f_{\vec{c}}(x) - f_{\vec{c_1}}(x)$ and obtain $g^{f_{\vec{c_2}}(\alpha)} = g^{f_{\vec{c}}(\alpha)} / g^{f_{\vec{c_1}}(\alpha)} = g^{f_{\vec{c}}(\alpha) - f_{\vec{c_1}}(\alpha)} \in Z_p[x]$. Since $f_{\vec{c_1}}(\alpha) = f_{\vec{c}}(\alpha)$ and $f_{\vec{c_2}}(\alpha) = 0$, i.e., α is a root of polynomial $f_{\vec{c_2}}(x)$. By factoring $f_{\vec{c_2}}(x)$[21], B can easily find $SK = \alpha$ and solve the instance of the t-SDH problem given by the system parameters. □

THEOREM 5.2. *If the signature scheme used for file tags is existentially unforgeable, the CDH problem is hard, the t-SDH assumption and the t-BSDH assumption hold. In our proposed scheme, the prover's response $Prf = (y, \psi, \sigma)$ is unforgeable.*

PROOF. Suppose a probabilistic polynomial time adversity can generate a forged proof response $(y', \psi', \sigma'), (y', \psi', \sigma') \neq (y, \psi, \sigma)$ after receiving a challenge message and make it to be accepted by the verification algorithm, we can get the following two equations:

$$e(\eta, g) \cdot e(\psi, \kappa \cdot v^{-r}) = e(\sigma, g) \cdot e(g^{-y}, v) \quad (10)$$

$$e(\eta, g) \cdot e(\psi', \kappa \cdot v^{-r}) = e(\sigma', g) \cdot e(g^{-y'}, v) \quad (11)$$

Dividing Eq.10 with Eq.11, we obtain:

$$\left(\frac{e(\psi, g)}{e(\psi', g)}\right)^{\epsilon(\alpha - r)} = \frac{e(\sigma, g)}{e(\sigma', g)} \cdot e(g, g)^{\epsilon(y' - y)} \quad (12)$$

Now we do a case analysis on whether $\sigma = \sigma'$.

Case 1: $\sigma \neq \sigma'$. As $\left(\frac{e(\psi,g)}{e(\psi',g)}\right)^{\epsilon(\alpha-r)}$, $e(g,g)^{\epsilon(y'-y)}$ and $e(\sigma, g)$ are known to the adversary, we rewrite Eq.12 as

$$\begin{aligned} & e(\sigma', g) = e(\sigma, g) \cdot \Upsilon \\ & e(\sigma', g) = e(u^{\epsilon(\sum_{(i,v_i)\in Q} t_i)}, g) \cdot e(g^{\epsilon f_{\vec{A}}(\alpha)}, g) \cdot \Upsilon \quad (13) \end{aligned}$$

where we denote $\Upsilon = e(g, g)^{\epsilon(y'-y)} / \left(\frac{e(\psi,g)}{e(\psi',g)}\right)$ as a known value to the adversary and $t_i = H(name||i)v_i$.

Recall that in this proof, the CDH problem is hard. If any probabilistic polynomial time adversity A can find σ' with non-negligible probability and make Eq.13 hold, we can construct an algorithm B that uses A to solve the instance of the CDH problem. Specifically, given $\sigma' \neq \sigma$ found by A, which makes Eq.13 hold, B can easily extract $g^{\epsilon f_A(\alpha)}$. With the given information, B can get $g^{\epsilon}, g^{f_A(\alpha)}$, where ϵ and $f_A(\alpha)$ are unknown, and thus solve the CDH problem. Therefore, no probabilistic polynomial time adversity can find a valid forged response $(y, \psi, \sigma) \neq (y', \psi', \sigma')$ and $\sigma \neq \sigma'$ with non-negligible probability.

Case 2: $\sigma = \sigma'$. In this case, we can rewrite Eq.12 as:

$$\left(\frac{e(\psi, g)}{e(\psi', g)}\right)^{\epsilon(\alpha - r)} = e(g, g)^{\epsilon(y' - y)} \quad (14)$$

We further do a case analysis on whether $y = y'$.

Case 2.1: $y = y'$. As $(y, \psi, \sigma) \neq (y', \psi', \sigma')$, $\sigma = \sigma'$ and $y = y'$, we can obtain that $\psi \neq \psi'$. In this case, since $y = y'$, we further rewrite the Eq.14 as:

$$\left(\frac{e(\psi, g)}{e(\psi', g)}\right)^{\epsilon(\alpha - r)} = 1 \quad (15)$$

We know that $\psi \neq \psi'$, i.e., $\frac{e(\psi,g)}{e(\psi',g)} \neq 1$, and $\epsilon \neq 0$, we can infer $\alpha = r$ from Eq.15. In our scheme, r is known to the adversary. Thus, the adversary can find $SK = \alpha$ and solve the instance of the t-SDH problem with solution $e(r, g^{\frac{1}{\alpha + r}})$ by given the system parameters. Therefore, no valid forged response $(y, \psi, \sigma) \neq (y', \psi', \sigma')$ and $y = y'$ can be found by probabilistic polynomial time adversity A with

non-negligible probability.

Case 2.2: $y \neq y'$. From Eq.14 and $y \neq y'$, we can infer that $\alpha \neq r$. In this case, we show how to construct an algorithm B, using the existed adversary, that can break the t-BSDH Assumption with a valid solution $(-r, \left(\frac{e(\psi,v)}{e(\psi',v)}\right)^{\frac{1}{y'-y}})$.

We denote ψ as g^{θ} and ψ' as $g^{\theta'}$, and then we can rewrite Eq.14 as :

$$\left(\frac{e(\psi,v)}{e(\psi',v)}\right)^{(\alpha-r)} = \frac{e(g,v)^{-y}}{e(g,v)^{-y'}}$$
$$\theta(\alpha-r)+y = \theta'(\alpha-r)+y'$$
$$\frac{(\theta-\theta')}{y'-y} = \frac{1}{\alpha-r} \quad (16)$$

Note that the operations in Eq.16 are modular operations with module p. Therefore, algorithm B can compute

$$\left(\frac{e(\psi,v)}{e(\psi',v)}\right)^{\frac{1}{y'-y}} = e(g,v)^{\frac{\theta-\theta'}{y'-y}} = e(g,g)^{\frac{1}{\alpha-r}} \quad (17)$$

and return $(-r, e(g,g)^{\frac{1}{\alpha-r}})$ as a solution for t-BSDH instance. It is easy to see that the success probability of solving the instance is the same as the success probability of the adversity, and the time required is a small constant larger than the time required by the adversary.

Therefore, *Theorem* 5.2 is proved. □

6. PERFORMANCE EVALUATION

In this section, we numerically evaluate the performance of our proposed *PCPOR* scheme in terms of communication cost, computation cost and storage cost. We compare our *PCPOR* scheme with existing POR techniques[20, 10, 15] and summarize the result in Table 1. For simplicity, in the following part of this paper, we denote the complexity of one multiplication operation on Group G as MUL and that of one exponentiation operation on Group G as EXP[1]. Furthermore, we use ZADD and ZMUL to represent the addition and multiplication operations on Z_p^* respectively. PRF is used to denote pseudorandom function and $|G|$ denotes the size (in number of bits) of a group element on G.

6.1 Communication

In our proposed *PCPOR* scheme, the communication cost comes from the challenge message *Chall* and the proof response Prf in each verification request. The challenge message consists of a $l-$element subset I and two random elements $\rho, r \in Z_p^*$. By utilizing Goldreich[12]'s $(\gamma,\delta)-hitter$[2], we can represent the subset I with $log^{|F|} + 3log^{(1/\delta)}$ bits, where $|F|$ is the size of the encoded data file, and δ is the error probability. In particular, given a data file less than 1024 TB (i.e. $|F| = 2^{43}$ bits, which is enough for most practical scenarios) and $\delta = 2^{-80}$ as in Ref.[15], the subset I can be represented with 283 bits. As a result, the total cost of *Chall* in our *PCPOR* scheme is $2\lambda + 283$ bits when the data file size is smaller than 1024 TB. The proof responses in our scheme are aggregated into 3 elements ψ, σ and y. Thus, the total size of the proof response is $2|G| + \lambda$ bits, where ψ and σ are two group elements and y the result of a polynomial. For a system with reasonable security parameter for our *PCPOR* scheme (e.g. set $\lambda = 160$ bits for ecliptic curves, $|G| = 1024$ bits and the data file is less than 1024 TB), the total communication cost becomes constant size. Therefore, the total complexity of communication cost in our *PCPOR* scheme is $O(1)$

Now, we compare existing POR schemes [20, 10, 15] with our *PCPOR* scheme and summarize the result in Table 1. In Ref.[20], the complexity of challenge message and proof response are $O(1)$ and $O(s)$ respectively, where s is the number of elements in each encoded block. Compared with our *PCPOR* scheme, which has constant communication cost in both challenge and response processes, this scheme have a communication size of proof response linear to s and a challenge message cost of λ times ours. The POR scheme proposed by Dodis et al.[10] achieves the challenge message size as same as our *PCPOR* scheme. However, the cost of proof response in their scheme is still $O(s)$. Considering only private verification, Xu et al.[15] reduce communication cost to constant size in each single verification. Nevertheless, their scheme only supports private verification, which centralizes all the verification tasks to the data owner. If the data owner wants to share his outsourced data with other individuals/organizations, it has to stay online and processes all verifications by himself. Differently, with the public verifiability of our *PCPOR* scheme, each challenger is able to conduct the verification independently with constant communication cost and the data owner can go off-line after outsourcing his data.

6.2 Computation

As shown in Section 5.1, our *PCPOR* scheme is composed of 4 algorithms: *KeyGen*, *Setup Prove* and *Verify*. Among these algorithms, *KeyGen* and *Setup* are performed off-line by the data owner as data preparation process. To generate the public key PK as well as the private key SK for the system, the data owner performs $(s+3)$EXP operations using the *KeyGen* algorithm. In the *Setup* procedure, to process an encoded data file with n blocks, each of which has s elements, $(2s+2)n$ EXP and $(s+1)$MUL operations are needed to generate the authentication tags. Note that the data preparation process in our *PCPOR* scheme is one-time cost for the data owner, it can go off-line after finishing this process. For each verification request, the server needs to perform $(l+s-1)$MUL, $(2l+s-1)$EXP, sl ZADD and $s(l+2)$ZMUL operations to generate the proof response, where l is the number of blocks chosen in each verification(Our discussion in Section.7 shows that the value of l can be small in practical). On receiving the proof information, the challenger first conducts l MUL and $2l$ EXP operations to generate η. Then, 2 more EXP and 5 *Pairing* operations are needed for the final verification. Therefore, as shown in Table 1, the total computation cost for the challenger in one verification is l MUL+$(2l+2)$EXP+$5Pairing$.

We now compare our *PCPOR* scheme with existing POR schemes[20, 10, 15] and show the result in Table 1. Compared with the public POR scheme in Ref.[20], our *PCPOR* will introduce $s-1$ more MUL, $l+s-1$ more EXP and $2s$ more ZMUL operations to the server side. But it achieves

[1]When the operation is on the elliptic curve, EXP means scalar multiplication operation and MUL means one point addition operation.

[2]In Goldreich[12]'s $(\gamma,\delta) - hitter$, it is guaranteed that any subset $Sub \subset [1,n]$ with size $|Sub| \geq |F|(1-gamma)$, $Pr[I \cap Sub \neq \varnothing]$. For more details, please refer to Ref.[12].

Scheme	Public verifiability	Comm. complexity (Challenge) (bits)	Comm. complexity (Response) (bits)	Comp. Cost (Server)	Comp. Cost (Challenger)	Storage Cost (Server) (bits)	Storage Cost (Challenger) (bits)	Data Preparation
[20]	Yes	$O(1)$	$O(s)$	l(MUL+EXP)+ sl(ZADD+ZMUL)	$(s+l)$MUL+ $(s+l)$EXP+ $2Pairing$	$(1+\frac{1}{s})\|F\|$	$\lambda+\|G\|$	$(s+1)$MUL+ $(s+2)n$ EXP
[10]	No	$O(1)$	$O(s)$	$(sl+l)$(ZMUL +ZADD)	$(s+l)$(ZMUL+ ZADD)+l PRF	$(1+\frac{1}{s})\|F\|$	2λ	l PRF+ n(ZMUL+ZADD)
[15]	No	$O(1)$	$O(1)$	$(s-1)$(EXP+MUL)+ $(s+l+sl)$(ZMUL +ZADD)	2EXP+l PRF l(ZMUL+ZADD)	$(1+\frac{1}{s})\|F\|$	$3\lambda+80$	l PRF+ n(ZMUL+ZADD)
PCPOR	Yes	$O(1)$	$O(1)$	$(l+s-1)$MUL+ $(2l+s-1)$EXP+ sl ZADD+$s(l+2)$ZMUL	l MUL+ $(2l+2)$EXP+ $5Pairing$	$(1+\frac{1}{s})\|F\|$	$\lambda+4\|G\|$	$(s+1)$MUL+ $(2s+2)n$ EXP

Table 1: Complexity Summary: in this table, MUL is one multiplication operation on Group G, EXP is one exponentiations operation on Group G, PRF denotes the pseudorandom function, ZADD and ZMUL represent the addition and multiplication operations on Z_p^* respectively; λ is the security parameter, $|G|$ is the size(in number of bits) of a group element on G, $|F|$ is size of data file, n is number of encoded blocks for the data file, s is the number of elements in each block and l is number of blocks selected for verification. Note: given a system security parameter, the values of λ and $|G|$ are fixed.

the same computational complexity at the challenger side (client). As the cloud sever is always much powerful than the challenger (e.g. Amazon EC2 vs Mobile devices), the additional computation cost brought to server side in our $PCPOR$ can be easily handled in practical scenarios and have little influence on the scheme's performance. Compared with the two private POR schemes[10, 15] as shown in Table 1, which have almost the same computation cost for the challenger except for 2 more EXP operations in Ref.[15], our $PCPOR$ scheme will cause a relative higher computation cost by replacing the operations on Z_p^* with operations on G in each verification. However, it is notable that the two private POR schemes[10, 15] have to centralize all the verification tasks to the data owner(i.e., the computation cost on the data owner is linear to the number of simultaneous verification requests), which requires the data owner to keep online. On the contrary, in our $PCPOR$, the computation tasks for each verification is distributed to the challengers without the participation of the data owner. This public verifiability makes our scheme easy to scale and have practical computation cost on each challenger. For the computation cost for data preparation, our $PCPOR$ scheme introduces sn more EXP operations compared with the public POR scheme in Ref.[20]. Since the data preparation process is one-time cost in our scheme and will not influence the realtime verification process, the additional operations in this process can be acceptable in practical scenarios. Compared with the private POR schemes in Ref.[10, 15], as shown in Table 1, our $PCPOR$ scheme and scheme in Ref.[20] need relative higher computation cost for the preparation due to the requirement of group operations for the public verifiability. Nevertheless, as mentioned above, these additional one-time computation cost is acceptable in practice.

6.3 Storage

In this section, we first analyze the storage cost of our $PCPOR$ scheme on both the challenger side and server side, and then compare it with existing POR schemes[20, 10, 15]. The results of our analysis are summarized in Table 1. At the challenger side, our $PCPOR$ scheme only requires the challenger to store partial public key $PK : \{p, v, k, spk, g\}$ in order to generate the challenge message $Chall$ and perform the verification algorithm $Verify$. Thus, the size of storage cost for each challenger is $4|G| + \lambda$ bits. Compared with existing POR schemes[20, 10, 15], which require $|G|+\lambda$ bits, 2λ bits and $3\lambda + 80$ bits respectively for the challenger side storage cost, our $PCPOR$ scheme achieves the same storage cost level as demonstrated in Table 1. The storage overhead on the server side mainly comes from the authentication tags for the encoded data blocks. In our $PCPOR$ scheme, each authentication tag is a group element with λ bits, thus the total size for tags is $n\lambda$ bits. As the total encoded data file size $|F| = ns\lambda$ bits, we represent the total storage overhead on the servers as $(1 + \frac{1}{s})|F|$ bits, which equals to Ref.[20, 10, 15]'s storage cost on servers when the values of s and λ are the same.

Note that the communication cost in our $PCPOR$ scheme is independent to s as we mentioned in Section 6.1. Therefore our scheme can reduce the storage cost by adjusting the value of s without affecting the communication performance. However, in Ref.[20, 10], as the communication cost for proof response is linear to the value of s, reducing the storage cost will lead to the sacrifice of communication performance.

7. DISCUSSION

In this section, we discuss the error detection probability of our $PCPOR$ scheme in Section.5.1. As we mentioned in the $Setup$ algorithm of $PCPOR$ scheme, Reed-Solomon code with rate ℓ is adopted for the data file encoding. For a ℓ Reed-Solomon encoded data file $(0 < \ell < 1)$, any ℓ fraction of encoded data blocks can recover the original file. If a data file encoded with ℓ Reed-Solomon code cannot be recovered from the erasure decoding, the probability of accessing a uncorrupted encoded data block will be less than ℓ. In this case, when we randomly choose l independent encoded data blocks and all these blocks are uncorrupted, the probability should be less than ℓ^l.

In our $PCPOR$ scheme, we can set $\ell = 0.98$ as previous POR scheme[15] does. In this case, by checking 200 encoded data blocks, the challenger can have at least 98.24% confidence that the stored data on the server is not corrupted if the $Verify$ algorithm outputs result as accpet. 99.99% confidence can be guaranteed if 1000 encoded data blocks are checked by the challenger.

8. CONCLUSIONS

Proofs of Retrievability (POR) technique enables individuals/organizations to verify the integrity of their outsourced data on a untrusted server (e.g., public cloud storage platform). While existing POR schemes have focused on various practical issues, they still have limitations either the com-

munication cost is linear to the number of elements in a data block, or the public verifiability is not supported. Such limitations cause these POR schemes to suffer from a severe scalability issue in terms of data file size or user number for practical use. In this work, we proposed the first public POR scheme with constant communication cost. By uniquely tailoring the polynomial commitment technique and designing a novel authentication tag, our *PCPOR* scheme achieves constant communication size, efficient computation performance as well as low storage overhead. What is more, by supporting the public verifiability, our scheme releases the data owner from onerous verification tasks, which need to be centralized to the data owner in previous private POR scheme with constant communication size. We prove the security of our scheme based on the CDH problem, the SDH assumption and the BSDH assumption. Our thorough analysis demonstrates the efficiency and scalability of our scheme.

9. REFERENCES

[1] Amazon forum. major outage for amazon s3 and ec2, https://forums.aws.amazon.com/thread.jspa?threadID=19714&start=15&tstart=0.

[2] Amazon web service. summary of the amazon ec2 and amazon rds service disruption in the us east region, http://aws.amazon.com/message/65648/.

[3] Business insider. amazon's cloud crash disaster permanently destroyed many customers' data, http://www.businessinsider.com/amazon-lost-data-2011-4.

[4] Dropbox. dropbox forums on data loss topic, http://forums.dropbox.com/tags.php?tag=data-loss.

[5] G. Ateniese, R. Burns, R. Curtmola, J. Herring, L. Kissner, Z. Peterson, and D. Song. Provable data possession at untrusted stores. In *Proceedings of the 14th ACM conference on Computer and communications security*, CCS '07, pages 598–609, New York, NY, USA, 2007. ACM.

[6] D. Boneh and X. Boyen. Short signatures without random oracles. pages 56–73. Springer-Verlag, 2004.

[7] D. Boneh, B. Lynn, and H. Shacham. Short signatures from the weil pairing. In *Proceedings of the 7th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology*, ASIACRYPT '01, pages 514–532, London, UK, UK, 2001. Springer-Verlag.

[8] K. D. Bowers, A. Juels, and A. Oprea. Proofs of retrievability: theory and implementation. In *Proceedings of the 2009 ACM workshop on Cloud computing security*, CCSW '09, pages 43–54, New York, NY, USA, 2009. ACM.

[9] W. Diffie and M. Hellman. New directions in cryptography. *IEEE Trans. Inf. Theor.*, 22(6):644–654, Sept. 1976.

[10] Y. Dodis, S. Vadhan, and D. Wichs. Proofs of retrievability via hardness amplification. In *Proceedings of the 6th Theory of Cryptography Conference on Theory of Cryptography*, TCC '09, pages 109–127, Berlin, Heidelberg, 2009.

[11] C. Erway, A. Küpçü, C. Papamanthou, and R. Tamassia. Dynamic provable data possession. In *Proceedings of the 16th ACM conference on Computer and communications security*, CCS '09, pages 213–222, New York, NY, USA, 2009. ACM.

[12] O. Goldreich, S. Goldwasser, and S. Micali. How to construct random functions. *J. ACM*, 33(4):792–807, Aug. 1986.

[13] V. Goyal. Reducing trust in the pkg in identity based cryptosystems. In *Proceedings of the 27th annual international cryptology conference on Advances in cryptology*, CRYPTO'07, pages 430–447, Berlin, Heidelberg, 2007. Springer-Verlag.

[14] P. Hawkes, M. Paddon, and G. G. Rose. On corrective patterns for the sha-2 family, 2004.

[15] X. Jia and C. Ee-Chien. Towards efficient provable data possession. In *Proceedings of the 7th ACM Symposium on Information, Computer and Communications Security*, ASIACCS '12, Seoul, Korea, 2012.

[16] A. Juels and B. S. Kaliski, Jr. Pors: proofs of retrievability for large files. In *Proceedings of the 14th ACM conference on Computer and communications security*, CCS '07, pages 584–597, New York, NY, USA, 2007. ACM.

[17] A. Kate, G. M. Zaverucha, and I. Goldberg. Constant-size commitments to polynomials and their applications. In *ASIACRYPT*, pages 177–194, 2010.

[18] G. Oded. A sample of samplers - a computational perspective on sampling (survey). *Electronic Colloquium on Computational Complexity (ECCC)*, 4(20), 1997.

[19] I. S. Reed and G. Solomon. Polynomial Codes Over Certain Finite Fields. *Journal of the Society for Industrial and Applied Mathematics*, 8(2):300–304, 1960.

[20] H. Shacham and B. Waters. Compact proofs of retrievability. In *Proceedings of the 14th International Conference on the Theory and Application of Cryptology and Information Security: Advances in Cryptology*, ASIACRYPT '08, pages 90–107, Berlin, Heidelberg, May 2008. Springer-Verlag.

[21] V. Shoup. *A computational introduction to number theory and algebra*. Cambridge University Press, New York, NY, USA, 2005.

[22] G. Timothy and M. M. Peter. The nist definition of cloud computing. NIST SP - 800-145, September 2011.