J. Parallel Distrib. Comput. 120 (2018) 62–76

Contents lists available at ScienceDirect

J. Parallel Distrib. Comput.

journal homepage: www.elsevier.com/locate/jpdc

Curve fitting based efficient parameter selection for robust provable data possession

Meng Liu [a,*], Xuyun Zhang [b], Chi Yang [c], Qiang He [d], Jianbing Zhang [e]

[a] School of Mechanical, Electrical and Information Engineering, Shandong University, Weihai, China
[b] Department of Electrical and Computer Engineering, University of Auckland, New Zealand
[c] School of Computing and Information Technology, University of Wollongong, Australia
[d] School of Software and Electrical Engineering, Swinburne University of Technology, Victoria, Australia
[e] State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, China

HIGHLIGHTS

- How to determine the number of sample blocks for PDP is formulated.
- A curve fitting based approach to determine the number of sample blocks is proposed.
- Extensive experiments demonstrate our approach is highly effective and efficient.

ARTICLE INFO

Article history:
Received 25 January 2018
Received in revised form 22 April 2018
Accepted 14 May 2018
Available online 22 May 2018

Keywords:
Provable data possession
Parameter selection
Integrity verification
Cloud computing
Cloud storage

ABSTRACT

Cyberspace faces a series of threats that can spread rapidly across distributed systems. Many such transmissible cyber threats aim to damage users' data. In recent years, the popularity of cloud computing has driven a lot of users to save their data in the cloud. The centralization of users' data in the cloud has created new opportunities and incentives for transmissible cyber threats targeting users' data. In this context, in addition to cloud vendor's security mechanisms, it is important to allow users to efficiently verify the integrity of their data saved in the cloud. The seminal sampling based PDP (Provable Data Possession) scheme can attain effective probabilistic verification with high computational efficiency for the integrity of users' data saved in the cloud by use of a set of randomly sampled data blocks. By integrating with the forward error correcting (FEC) technique, a recently-proposed robust PDP scheme can protect against arbitrarily small amounts of data corruption and has therefore been widely adopted in practice. It is a core task to determine the number of sample blocks in this scheme because this parameter plays a fundamental role in balancing the security of the scheme and the computational cost. A smaller value can comprise the security while a larger one can incur extra high computational cost. Existing work mainly leverages the Monte Carlo methods to estimate this parameter. However, these methods suffer from heavy computational cost. In this paper, we propose a method to determine this parameter based on the curve fitting technique. Specifically, we formally analyze the parameter selection process of the robust PDP scheme, and leverage the curve fitting technique to improve the efficiency of parameter selection while ensuring the optimality of the number of samples for the robustness of the scheme. Extensive experimental results demonstrate the effectiveness and efficiency of our approach. Specifically, it can be 25 times faster than the existing solution for 1, 000, 000 times simulated attacks.

1. Introduction

Nowadays, cyberspace faces a series of threats that can spread rapidly across distributed systems [12,13]. Every year, transmissible cyber threats bring huge economic losses and damage to users of various distributed systems. For example, malware, a typical transmissible cyber threat, has become the most concerned security issues in the cyberspace in recent years. CTB-Locker virus ran rampant in May 2015. Although we can use anti-virus software to kill the virus, it is often very difficult, sometimes impossible, to restore the data damaged by the virus. Transmissible virus can cause huge economic losses by threatening and damaging users' data. On the evening of May 12, 2017, the "WannaCry" infection of the ransomware virus broke out in the world. The virus has spread to

* Corresponding author.
E-mail addresses: liumeng@sdu.edu.cn (M. Liu), xuyun.zhang@auckland.ac.nz (X. Zhang), chiy@uow.edu.au (C. Yang), qhe@swin.edu.au (Q. He), zjb@nju.edu.cn (J. Zhang).

https://doi.org/10.1016/j.jpdc.2018.05.007
0743-7315/

hundreds of countries and regions in the world. The virus incident spread to education and scientific research units, business centers, medical units and so on. In recent years, cloud storage has become a popular choice for both enterprises and individuals to improve the security of their data because cloud storage vendors are capable of implementing different security mechanisms to protect their clients' data. As a result, an increasing number of companies and individuals have been moving their local data files to the cloud. However, due to the ubiquitous accessibility and multi-tenancy features, cloud computing faces a range of new security and privacy challenges (e.g., confidentiality, integrity, availability and privacy-preservability) that can spread rapidly across server farms or data centers [10,18,27,36]. In addition, as a type of transmissible cyber threat, malware could also compromise the integrity of the data in cloud platforms [11,23]. Such transmissible cyber attacks in cloud can jeopardize the hosted services and outsourced data storage severely [4,10,36]. The large amount of users' data centralized in the cloud has created new opportunities and incentives for cyber threats. Being an important service of cloud computing, cloud storage features unreliability and can lose the users' data, for instance some disks of the storage servers can be damaged and the maintainer might not be able to successfully restore the damaged data. But the cloud storage provider could not communicate this error to the affected users. Moreover, some cloud storage providers could be dishonest as well. To recycle the cloud storage space, they could discard the user data that have not been accessed or rarely accessed by the users.

Thus, the users must be capable of making periodic verifications of the integrity of the data storing in the untrusted cloud storage.

A variety of solutions have been proposed to solve these critical issues. It should be noted that we assume that the verifier cannot use the local copy data when the verifier executes the PDP protocol. An intuitive way is that the client stores a file to the server and computes a hash value of the file as a metadata stored in the client using a cryptographic hash function, for instance MD5, SHA, etc. Then the server sends the entire file to the client and the client recalculates a new hash value of the received file, followed by making a comparison of the two hash values to verify whether the file is stored correctly in cloud storage. But this way requires huge network commutation costs. In real cloud storage system, both of the number of files and the data size can be quite large, whereby an extensively large amount of bandwidth together with time is required to send these files from the servers to the clients across the network. To address these challenges, a range of approaches have been recently developed. Specifically, RSA-based hash functions [6,9] and homomorphic hash functions [20,38] are developed to reduce the communication cost to $O(1)$. But these solutions house the issue that the server will still be required to get access to the entire file and it will cost expensive disk I/O and a large amount of computation as well. Ateniese et al. [2] first adopted the sampling technique into the PDP scheme to reduce the computation cost. This scheme can significantly reduce the expensive disk I/O. A salient feature of this technique is that it only needs a constant number of data blocks to execute the cloud storage auditing protocol, independently of the total size of a file. Because of this, the sampling technique (also known as spot checking) has been widely adopted in the state-of-the-art cloud storage auditing protocols [17,28,41,45] to reduce the execution time and improve auditing performance. Furthermore, Ateniese et al. [1] proposed a more robust PDP scheme by integrating FEC codes with the classical PDP method [2] to enable the verifier to identify the corruption even if the corrupt data in a file are very small, while the latter fails to do so without a specific condition where a certain number of data blocks are corrupt. This means the robust PDP scheme can protect against arbitrary small amounts of data corruption.

However, the determination of the number of samples of the robust PDP scheme heavily depends on the total number of data blocks in cloud. The existing approach to determining the optimal number of samples is to use the Monte Carlo method, which leads to hefty computational costs. As it is in urgently required to enhance the computational efficiency to determine a suitable number of sample blocks while ensuring the precision of the robust PDP scheme, we propose a method herein based on the function fitting method to choose the sampling number of the robust PDP scheme while reducing the number of Monte Carlo simulation attacks. Besides theoretical analyses, the extensive empirical study shows that our approach can be both effective and efficient. Specifically, the computational cost can be significantly reduced when we choose the optimal number of samples for the robust PDP scheme.

Our contributions can be summarized as follows:

(1) By formal analyses, we formulate the issue of determining the optimality of number of sample blocks for the robust PDP scheme in order to protect outsourced data in the cloud that may be corrupted by malware.

(2) We are the first to propose the curve fitting based approach to determine the number of sample blocks in robust PDP scheme. Another vital finding is that the data damage probability can be captured well by decreasing the deviation of the experimental data.

(3) Extensive experiments demonstrate that our approach is highly effective and efficient with being able to 25 times faster than the existing methods for 1000,000 times simulated attacks.

The rest of the paper is organized as follows. The next section introduces the related work. Section 3 models and analyzes the research problem with an basic introduction to the PDP schemes as preliminaries. In Section 4, we formulate our method that leverage the curve fitting function to lower the deviation of the experimental data and improve the efficiency of parameter selection. We empirically evaluate our approach and report the experiment results in Section 5. Finally, we conclude the paper and discuss some future work in Section 6.

2. Related work

Ensuring the integrity of the data in cloud platforms is an effective way to avoid the damage from transmissible cyber attacks and their propagation [5,19,22,36,39]. In order to guarantee data integrity, integrity verification always appears as an essential problem on outsourced data. That is why a good amount of research has been carried out, together with achieving some exceptional results have been achieved [21,25,26,42].

In the context of some previous solutions, the data owners pose the responsibility for the verification of the data integrity of cloud storage. Furthermore, these solutions usually require huge network disk I/O, computation and commutation costs. Some solutions first reduced commutation costs by using some techniques, for example RSA-based hash functions [6,9] and homomorphic hash functions [20,38], the communication cost of which is $O(1)$. But all of them require expensive disk I/O and a large amount of computation. Accordingly, sampling technique was adopted by Ateniese et al. [2] in their PDP scheme. Their solution greatly reduced the expensive disk I/O and computation costs while still achieving a high probabilistic possession guarantee.

In order to guarantee the justice of data possession, introduction of the third-party auditing was provided [30,43]. Recently, several new third-party auditing protocols were proposed to allow the auditor to verify the data integrity on the cloud storage server [28,29,31,34,41,45]. Wang et al. [34] used homomorphic encryption to solve the data privacy problem. Yang et al. [41] proposed a new scheme based on paring-based cryptography to solve the data privacy issue. Both of these two auditing protocols