Liang, Xueqin; Yan, Zheng; Chen, Xiaofeng; Yang, Laurence T.; Lou, Wenjing; Hou, Thomas

Game Theoretical Analysis on Encrypted Cloud Data Deduplication

Published in:
IEEE Transactions on Industrial Informatics

DOI:
10.1109/TII.2019.2920402

Published: 01/10/2019

Document Version
Peer-reviewed accepted author manuscript, also known as Final accepted manuscript or Post-print

Please cite the original version:
Liang, X., Yan, Z., Chen, X., Yang, L. T., Lou, W., & Hou, T. (2019). Game Theoretical Analysis on Encrypted Cloud Data Deduplication. *IEEE Transactions on Industrial Informatics*, *15*(10), 5778-5789. https://doi.org/10.1109/TII.2019.2920402

Game Theoretical Analysis on Encrypted Cloud Data Deduplication

Xueqin Liang, Zheng Yan, *Senior Member, IEEE,* Xiaofeng Chen, *Senior Member, IEEE,*
Laurence T. Yang, *Senior Member, IEEE,* Wenjing Lou, *Fellow, IEEE,* and Y. Thomas Hou, *Fellow, IEEE*

***Abstract*—Duplicated data storage wastes memory resources and brings extra data-management load and cost to cloud service providers (CSPs). Various feasible schemes to deduplicate encrypted cloud data have been reported. However, their successful deployment in practice depends on whether all system players or stakeholders are willing to accept and execute them in a cooperative way, which was scarcely investigated in the previous literature. In this paper, we employ a non-cooperative game to model the interactions in a client-side server-controlled deduplication scheme (S-DEDU) [1] and construct an incentive mechanism based on payment discount to motivate its final acceptance. The experimental results based on a real-world dataset demonstrate the individual rationality, incentive compatibility, profitability and robustness of our incentive mechanism.**

***Index Terms*—cloud computing, encrypted data deduplication, game theory, incentive compatibility, incentive mechanism.**

I. Introduction

CLOUD computing [2] is a generic service platform architecture enabling ubiquitous, convenient, and on-demand access to various configurable resources in the form of X-as-a-Service. It is characterized by low energy consumption, resource sharing, scalability, elasticity, high reliability, and on-demand service. The popularity of cloud storage [3] along with the explosive growth of cloud data, especially big data, generates high demands for economical storage in the cloud. However, more than half of the data in standard file systems are duplicate and this proportion can be more than 90% in backup systems [4]. Duplicated data cause storage waste and burden data-management load and cost. In the era of big data, deduplication, which is a method to eliminate duplicated data storage, becomes essentially important in cloud computing.

Encrypted data are outsourced to the cloud to protect data security and privacy and various feasible schemes to deduplicate encrypted cloud data have been reported. Group data sharing [5] is regarded as deduplication over encrypted data. Searchable encryption [6] helps duplication check over encrypted data. Thus far, substantial encrypted cloud-data deduplication schemes were proposed based on convergent encryption [7]–[10], proofs-of-ownership (PoW) [1], [11]–[13], secret sharing [14], [15], password-authenticated key exchange (PAKE) [16], keywords search [10] and data ownership challenge with Proxy Re-Encryption (PRE) [1], [17]. Most explored encrypted data deduplication schemes support duplication check before uploading data, which significantly save network bandwidth. Classifying them based on which party controls data access, we obtain server-controlled deduplication (S-DEDU) and client-controlled deduplication (C-DEDU). S-DEDU holds a special advantage over C-DEDU because it releases data holders from any online service support, thus it outperforms C-DEDU and is widely preferred in the literature.

But it is desirable to analyze the practical deployment and acceptance of an encrypted cloud data deduplication scheme, which is scarcely investigated in the existing literature [1], [5]–[16] although some schemes have been commercially deployed [18], e.g., Bitcasa [19] and Wuala [20]. We found that there is a dearth of knowledge today on the acceptance of a deduplication scheme. Previous works [16], [21], [22] only mention the necessity of incentives for deduplication without providing any specific ones. Liu et al. [16] stated their scheme necessitates a direct incentive to attract data holders. Rabotka and Mannan [21] found cloud service providers (CSPs) need more incentive in C-DEDU after examining previous secure deduplication schemes. Youn and Chang [22] identified the dishonest actions of the first data holder and invoked an incentive system to suppress such actions. Miao et al. [23] proposed an incentive mechanism based on payment, which provides enough incentive to data holders. Yet, their mechanism seems not so comprehensive and practical because it is costly for CSPs (refer to detailed analysis in Section IV-B).

Analyzing whether all stakeholders have incentives to accept the deduplication scheme is practically important for ensuring the final adoption and wide usage of the cloud storage service. It is obvious that CSPs, as a direct beneficiary, are willing to accept deduplication. However, since data holders lose direct control over their data in S-DEDU, they may suffer from temporary data-unavailability due to the mismanagement of CSPs. More severely, data holders may confront permanent data-loss caused by illegal data erasure, because only one copy of data is stored. Thus, sensitive data holders may be reluctant to accept S-DEDU. On the other hand, S-DEDU also requests a proper incentive mechanism to motivate both CSPs and data

X. Q. Liang and Z. Yan are with the State Key Lab on Integrated Services Networks, School of Cyber Engineering, Xidian University, No.2 South Taibai Road, Xi'an, China, 710071; and the Department of Communications and Networking, Aalto University, Konemiehentie 2, P.O.Box 15400, Espoo 02150, Finland. E-mail: dearliangxq@126.com; zyan@xidian.edu.cn.

X. F. Chen is with the State Key Laboratory of Integrated Service Networks (ISN), Xidian University, No.2 South Taibai Road, Xi'an, China, 710071. E-mail: xfchen@xidian.edu.cn.

L. T. Yang is with the Department of Computer Science, St. Francis Xavier University, Antigonish, NS, B2G 2W5, Canada. E-mail: ltyang@gmail.com.

W. J. Lou is with the Department of Computer Science, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA. E-mail: wjlou@vt.edu.

Y. T. Hou is with the Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061 USA. E-mail: thou@vt.edu.

holders, primarily to encourage the cooperation of data holders with CSPs. Without any doubt, the success of S-DEDU relies on the acceptance of all involved system stakeholders (i.e. CSPs and data holders) and their cooperation.

However, we are still facing a number of challenges in the study of S-DEDU acceptance towards practical deployment. First, it is complicated to select a suitable analysis model to perform this study. Using game theory seems very helpful due to its advances in analyzing behavior strategies. Nevertheless, which game model is appropriate enough to conduct concrete analysis remains unsolved. Second, it is difficult to verify the analysis results due to the lack of real-world data. Using simulation to evaluate analysis results is not so convincing and hard to be recognized. It is also laborious to set up system parameters during analysis, evaluation, and proof. Third, for motivating all system stakeholders to accept and execute a deduplication scheme, i.e., to make the deduplication scheme deployable, we may need to figure out a novel incentive mechanism. But it is not easy to ensure each stakeholder can benefit with **individual rationality**, **incentive compatibility** among all system players, **profitability** for newly involved parties and system **robustness** to withstand unintentional system disturbances.

In this paper, we investigate the acceptance of S-DEDU [1] with game theory by analyzing the utilities of its stakeholders under different strategies with a non-cooperative game. Through theoretical analysis, we propose an incentive mechanism based on payment discount to motivate its acceptance. Experimental evaluations based on a real-world dataset further demonstrate its effectiveness in promoting the final acceptance of S-DEDU. Specifically, this paper has the following contributions:

1) We examine the detailed payoff structure of all types of system stakeholders in S-DEDU.
2) We propose an incentive mechanism based on payment discount to motivate the participation willingness of data holders while ensuring the profits of CSPs.
3) We consider the influence of data mismanagement to investigate the robustness of our incentive mechanism.
4) Real-world dataset based experiments show the effectiveness and correctness of our proposal.

The remainder of this paper is organized as follows. Section II presents a brief introduction of game theory and related work. S-DEDU is introduced in Section III along with problem statements, design goals, and our research assumptions. In Section IV, we give a game theory based economic model and propose an incentive mechanism in the cloud storage system with deduplication, which satisfies individual rationality, incentive compatibility, profitability, and system robustness. The experimental settings and results are detailed in Section V, followed by a conclusion section.

II. Background and Related Work

A. Game Theory

Game Theory is an effective mathematical model to analyze conflict and cooperation between rational decision-makers [24]. It has been widely deployed in many fields, such as economics, psychology, and even biology. Researchers in computer sciences also initiate applying game theory to analyze the interactions among system players. *Nash Equilibrium* (NE), a very important term in game theory, refers to a strategy profile that no player can obtain more profits by changing its current actions.

The essence of game theory in studying the interactions among rational players is its effectiveness and efficiency in handling various difficult problems. It models how a player will change its actions with regard to the others' actions and then the others act based on this player's actions and so forth. An analysis without considering the strategies of all stakeholders, like the price mechanism in [23], is not practical. In this paper, we intend to employ game theory to model how data holders and CSPs react when introducing S-DEDU and analyze how to motivate all players to accept S-DEDU.

B. Related Work

We found that there are already some researchers who noticed the incentive problem in encrypted cloud data deduplication although no concrete studies on it so far.

Rabotka and Mannan [21] analyzed some privacy-preserving based deduplication schemes in defending various attacks. Even though they did not mainly focus on analyzing the incentive in deduplication, they concluded that there is little incentive for CSPs to adopt C-DEDU.

Liu et al. [16] presented a C-DEDU scheme, based on a PAKE protocol without the presence of any additional independent servers. It can prevent the malicious behaviors of both CSPs and data holders. Notably, the authors commented that direct incentives should be investigated in order to motivate the data holders since the proposed scheme introduces extra burdens to them. Unfortunately, they did not elaborate a concrete incentive mechanism.

ClearBox [25] provides data holders a way to check the accurate number of data holders in the cloud. It motivates CSPs to honestly report the number of data holders of the same data in order to attract data storage. Moreover, the scheme can resist malicious data holders and CSPs. The authors stated that their model can promote the appearance of a novel economic price-model with fairness. However, they did not put forward a concrete proposal.

The deduplication scheme designed by Miao et al. [23] is also a server-controlled scheme. The authors proved that their scheme can ensure data confidentiality and signature unforgeability. Moreover, they formulated that the service fee of data holders and the deduplication rate are negatively correlated. Therefore, a rational CSP is unable to gain illegal benefits by reporting low deduplication rate to data holders. They also analyzed the economic essence of CSPs and data holders and provided an incentive mechanism to motivate the participation level of data holders. However, their mechanism cancnot provide incentive compatibility for CSPs (refer to Section IV-B for thorough analysis). A more comprehensive economic model is expected.

Economic factor has also been taken into consideration by Wang et al. [26] to address the side-channel attacks in

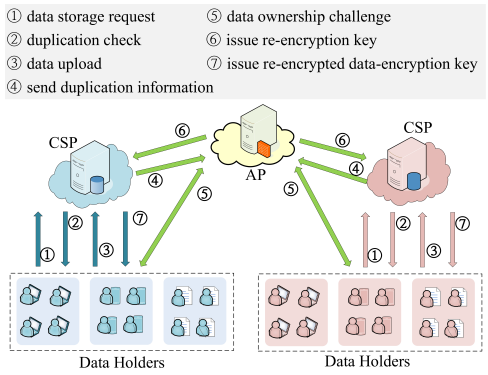

Fig. 1. The structure of a cloud storage system with S-DEDU.

deduplication schemes. The authors did not specify which entity holds the data access control but their solution is compatible with either server-controlled or client-controlled schemes. The authors employed game theory to form the interaction between an attacker and the CSP and designed a defence scheme.

Youn and Chang [22] analyzed the weakness of C-DEDU. They thought the privacy-disclosure risk of first data storage is higher than that of later duplicated data storage although they pay the same storage fee to CSPs. Therefore, no one has the incentive to be the first uploader. They pointed out the urgent need for an incentive mechanism in the C-DEDU, like giving a discount to the first uploader.

III. System Model and Problem Statement

In this paper, we focus on studying the deployment of S-DEDU and its acceptance. In another line of our study, we investigate the same issues with regard to C-DEDU.

In this section, we first detail the S-DEDU scheme that deduplicates encrypted data stored at CSPs with the help of PRE, in which data holders can get data access keys based on data ownership challenge. We also specify the economic model applied throughout the whole paper. We list some assumptions and numerous deployment problems of S-DEDU in practice along with the expected design goals.

A. System and Economic Model

Fig. 1 represents a cloud storage system $\mathbb{C} = \{\mathcal{K}, \mathcal{H}, \mathcal{A}\}$ with S-DEDU, where $\mathcal{K}$, $\mathcal{H}$ and $\mathcal{A}$ denote the sets of CSPs, data holders and Authorized Parties (AP), respectively.

CSPs: The cloud storage system consists of multiple CSPs (e.g., AWS S3, Dropbox, OneDrive, and Google drive) that are enabled to provide storage services to data holders. Once a data holder sends a data-storage request to a CSP, the CSP checks if this data has been uploaded in its storage space. If the check is negative, it asks this data holder to upload the encrypted data with access policy. Otherwise, the CSP cooperates with AP to verify the eligibility of this data holder to allow it access to pre-stored data by re-encrypting a data encryption key in a form that can be decrypted by the eligible data holder. Let $\mathcal{K} = \{k_1, k_2, \ldots, k_K\}$ denote the set of CSPs in the system.

Data holders: There exist many data holders, some of which may request to store the same data. If these holders store their data at a CSP with S-DEDU and they all accept S-DEDU, only one piece of data is stored in the cloud. $\mathcal{H} = \{h_1, h_2, \ldots, h_H\}$ denotes the set of data holders. The data set is denoted as $\mathcal{D} = \{d_1, d_2, \ldots, d_D\}$. Each data $d \in \mathcal{D}$ belongs to at least one holder. Let N_d represent the number of data holders of data d, then $N_d \geq 1$ and $\sum_{d\in\mathcal{D}} N_d = H$. The holder set of data d is denoted as $\mathcal{H}_d = \{h_{d,1}, h_{d,2}, \ldots, h_{d,N_d}\}$.

AP: AP is a third authorized party that does not collude with any parties in this system. Once a duplicated data storage is found by a CSP, it will be requested to challenge the data ownership. AP and the CSP cooperate as a proxy to provide deduplication service.

The following illustrates how S-DEDU works as shown in Fig. 1.

(1) Data holder $h_{d,1} \in \mathcal{H}_d$ sends its encrypted data d-storage request to CSP k_1. (2) k_1 checks its storage space and finds d is a unique data, then it asks $h_{d,1}$ to upload d. (3) Data holder $h_{d,2} \in \mathcal{H}_d$ sends its encrypted data d-storage request to CSP k_1. (4) k_1 finds the existence of data d, then it sends the information of $h_{d,2}$ to AP to verify its ownership. Once $h_{d,2}$ passes data ownership challenge, k_1 and AP cooperate to issue a re-encrypted data-encryption key to $h_{d,2}$. $h_{d,2}$ decrypts the key and accesses data d with this key. (5) S-DEDU supports data deletion and data update as well, which is beyond the analysis in this paper.

S-DEDU is secure since AP knows nothing about the encrypted data stored in CSP, and CSP cannot gain the data holder's plain data [1]. Therefore, no malicious behaviors happen in this model. The strategy of a CSP is whether to provide service with deduplication. The strategy of a data holder is whether to follow the procedure of S-DEDU or not.

We construct an economic model to analyze the acceptance of S-DEDU. The interactions among CSPs and data holders are modeled as a non-cooperative game since the objective of each stakeholder is to maximize its own profit. The utility functions of all system players are specified based on their interactions. The pricing mode applied in our analysis is pay-per-use, as often adopted in practice. For easy presentation and understanding, we summarize the main notations used in this paper in Table I with nomenclature. For the first three lines, the lowercase indicates an entity, the uppercase shows the total number of a class of entity, and the calligraphic ones refer to an entity set.

1) Utility of Data Holder: A data holder $h \in \mathcal{H}$, stores its data $d \in \mathcal{D}$ at CSP $k \in \mathcal{K}$. Let $U_h^0(t)$ denote its utility function when S-DEDU is not applied. The profit that h can gain from cloud storage at time t is denoted as $B_h(t)$. It should also pay storage fee $SF_h^k(t)$ to k at time t.

The loss for data mismanagement is $L_h(t)$. Assume data mismanagement happens at CSP k with the possibility of p_k. Let w_h denote the data-mismanagement influence on h, which could diverse from holder to holder. Then the loss suffered by a data holder without S-DEDU is

$$L_h(t) = p_k \times w_h \times B_h(t). \tag{1}$$

TABLE I
NOTATION WITH NOMENCLATURE

Notation	Description
$h, H, \mathcal{H}$	The parameters related to data holders;
$d, D, \mathcal{D}$	The parameters related to data;
$k, K, \mathcal{K}$	The parameters related to CSPs;
$U_a^{0,1,2}$	The utility functions of a, where a can be h, k, or AP;
SF_h^k	The storage-service fee h paid to k;
B_h	The benefit of h gained from cloud storage;
L_h	The loss suffered by h in terms of data mismanagement;
SC_k^h	The storage cost of k for storing the data of h;
RF_k	The service fee k paid to AP;
OC_{AP}	The operation cost of AP;
p_k	The possibility of data mismanagement in k;
r_k^d	The deduplication rate of d in k;
α_k^d	The storage-fee discount that k grants to the holder of d;
δ	The discount-adjustment parameter;
w_h	The influence of data-mismanagement on h;
φ_k	The unit storage fee to unit storage cost ratio of k;
$N_k^d(n_k^d)$	The data-holder number of d (that choose S-DEDU) in k;
AF_b^k	The parameters related to the access fee of k, where b can be in, out or $\emptyset$.

Therefore, we have

$$U_h^0(t) = B_h(t) - SF_h^k(t) - L_h(t). \tag{2}$$

When S-DEDU is applied, k grants discount $\alpha_k^d(t) (0 \leq \alpha_k^d(t) < 1)$ to the holders of d. Then, the storage fee h pays is $\left(1 - \alpha_k^d(t)\right) \times SF_h^k(t)$.

The data-mismanagement loss is exacerbated since all holders with the same deduplicated data are associated. Let $r_k^d(t)$ denote the deduplication rate of d, which is the proportion of the data holders that accept S-DEDU. The loss when S-DEDU is applied at time t becomes $\left(1 + r_k^d(t)\right) \times L_h(t)$.

To sum it up, for a data holder h that stores its data at k at time t, its utility with S-DEDU is

$$\begin{aligned} U_h^1(t) &= B_h(t) \\ &- \left(1 - \alpha_k^d(t)\right) \times SF_h^k(t) - \left(1 + r_k^d(t)\right) \times L_h(t). \end{aligned} \tag{3}$$

2) Utility of CSP: For each data $d \in \mathcal{D}$, there are $N_k^d(t)$ data holders. We employ $U_k^0(t)$ to present the utility function of CSP $k \in \mathcal{K}$ in the absence of S-DEDU at time t. In general, the utility of CSP k is the difference between the total storage fee it receives and the total storage cost. Then

$$U_k^0(t) = \sum_{d \in \mathcal{D}} N_k^d(t) \times \left(SF_h^k(t) - SC_k^h(t)\right). \tag{4}$$

When S-DEDU is applied and the deduplication rate of d is $r_k^d(t)$ at time t, there are $n_k^d(t) = r_k^d(t) \times N_k^d(t)$ holders of d that accept S-DEDU and $N_k^d(t) - n_k^d(t)$ holders refuse S-DEDU. k needs to store one copy of data for all the holders that accept S-DEDU and one copy for each data holder that refuses S-DEDU.

In this case, the storage cost for d is

$$SC_k^h(t) \times \left(N_k^d(t) - n_k^d(t) + 1\right).$$

Meanwhile, the storage service fees k obtains for d is

$$SF_h^k(t) \times \left(N_k^d(t) - \alpha_k^d(t) \times n_k^d(t)\right).$$

When k adopts S-DEDU, it should additionally pay service fee $RF_k(t)$ to AP for getting re-encryption keys, which is indispensable in S-DEDU. Then we conclude the utility function of k with S-DEDU as

$$\begin{aligned} U_k^1(t) &= \sum_{d \in \mathcal{D}} SF_h^k(t) \times \left(N_k^d(t) - \alpha_k^d(t) \times n_k^d(t)\right) \\ &- \sum_{d \in \mathcal{D}} SC_k^h(t) \times \left(N_k^d(t) - n_k^d(t) + 1\right) - RF_k(t). \end{aligned} \tag{5}$$

3) Utility of AP: The income of AP comes from the service fee $\sum_{k=1}^{K} RF_k(t)$ paid by all of its subscribed CSPs and the expenditure is mainly from its operation cost $OC_{AP}(t)$ for PRE. Hence, the utility of AP is:

$$U_{AP}(t) = \sum\nolimits_{k=1}^{K} RF_k(t) - OC_{AP}(t). \tag{6}$$

B. Assumptions

Our research holds a number of assumptions, as summarized below with justification.

Game assumption: All players are profit-driven in an economic environment. They rationally choose strategies from the perspective of maximizing their own utilities.

Data holder assumption: Data holders may not follow S-DEDU completely. When a data holder moves more data to the cloud, it could save more local storage costs thus benefit more. For simplifying our analysis, we assume all data holders can obtain the same stable benefits from cloud storage. According to the analysis in [27], cloud storage provides more convenience and benefits to data holders than local storages. Therefore, we assume all data holders are willing to store their data in the cloud for a long run.

Data assumption: For easy analysis, we simply assume that each data holder has and only has one data to store in the cloud. Different data have different data sizes.

CSP assumption: CSP cannot be fully trusted since it may not follow S-DEDU completely as promised in order to achieve higher profits. We assume the capacity of each CSP is infinite in case that all data holders choose to store at the same CSP. The unit storage fee of different CSPs may differ, but it is the same within one CSP. Moreover, CSP charges the same storage fee from all the holders of the same data and the discounts are only granted to the holders accept S-DEDU.

AP assumption: AP is a trusted party and does not collude with any other system stakeholders. It is possible that multiple APs exist and compete with each other. Since their business competition is not the focus of our research, we assume that there is only one AP for simplification. AP charges service fees from CSPs for providing the re-encryption service.

Mismanagement assumption: Mismanagement occurs in CSPs. S-DEDU makes the relationship between holders of the same data tight, data mismanagement on one data holder could influence the others. In a secure cloud storage system without deduplication, the impact of mismanagement is not so serious since user data are encrypted and one data loss will not affect another. Herein, we assume data encryption is sufficiently secure and the possible leakage caused by encryption vulnerability is trivial, thus data leakage is beyond our

consideration. Data mismanagement like service interruption and temporary data unavailability influences the cloud storage experiences of data holders; therefore, we assume the loss caused by mismanagement impacts the benefit of cloud storage and the loss obeys uniform distribution.

C. Problems Statements

When applying S-DEDU into a practical scenario, the following behaviors are considered in this paper.

First, a data holder could set its access policy as forbidding data sharing with others. It may also encrypt its data based on other schemes rather than the one specified in S-DEDU to hide its data characteristics (duplicated or unique). These actions result in cloud storage-resource waste and influence the quality of deduplication. The motivation for data holders to comply with the design of S-DEDU should be considered.

Second, a CSP can directly store all uploaded data without conducting duplication checks. Such uncooperative action may happen when the saved storage costs cannot make up the deduplication cost.

Third, the newly introduced party AP can only survive when its profits cover its operating costs. In other words, AP will only cooperate with CSP when it gains profits in reality.

Fourth, even though S-DEDU can resist deliberate security attacks [1] from outside, some unintentional system failures, e.g., service interruption and temporary data unavailability, may occur. For example, Dropbox suffered from a 10-hour outage in 2013. If S-DEDU is applied, one data unavailability will cause inconvenience to more than one holder. Such a bad experience will definitely hinder the acceptance of S-DEDU. Therefore, it is necessary to consider unintentional system failures, referred to as *mismanagement*.

D. Design Goals

Based on the problems stated above, we further specify our design goals as below: 1) **Individual Rationality:** No matter data holders or CSPs, they can earn more profits when adopting S-DEDU. 2) **Incentive Compatibility:** Each player cannot gain more profits by deviating the scheme design. 3) **Profitability:** The profits of a newly involved third party (AP in our paper) should be guaranteed. Otherwise, AP cannot survive to provide long-term services. 4) **Robustness:** When applying S-DEDU in practice, it should not only resist various attacks (as analyzed in [1]) but also be robust enough to withstand disturbances, i.e., mismanagement in its execution environment.

IV. Discount-Based Incentive Mechanism

In this section, we first we investigate the feasibility conditions of incentive mechanism, followed by an analysis of the problem of [23] before proposing our incentive mechanism.

A. Feasibility Conditions for Incentive Mechanism

CSPs need to carefully design $\alpha_k^d(t)$ in order to provide incentives to data holders without damaging their own profits. We first list the feasibility conditions as follows.

Definition 1. *Individual Rationality Constraint (IR-Constraint): An incentive mechanism in S-DEDU that a rational data holder/CSP accepts should guarantee a non-negative payoff, i.e., for all $h \in \mathcal{H}$ with S-DEDU, $U_h^1(t) \geq 0$; for all $k \in \mathcal{K}$ with S-DEDU, $U_k^1(t) \geq 0$.*

Definition 2. *Incentive Compatibility Constraint (IC-Constraint): The best strategy for a holder with duplicated data is to accept S-DEDU, i.e., for all $h \in \mathcal{H}$, $U_h^1(t) - U_h^0(t) \geq 0$. A CSP can obtain more profits by adopting S-DEDU, i.e., for all $k \in \mathcal{K}$, $U_k^1(t) - U_k^0(t) \geq 0$.*

Definition 3. *Profitability Constraint (P-Constraint): The incentive mechanism should guarantee the profit of AP, i.e., $U_{AP} \geq 0$.*

Before detailing the structure of the incentive mechanism with payment discount, we show some preconditions. A consensus throughout the literature is the acceptance of cloud storage service [27]. Namely, $U_h^0(t) \geq 0$ and $U_k^0(t) \geq 0$. Hence,

$$(1 - p_k \times w_h) \times B_h(t) - SF_h^k \geq 0, \tag{7}$$

$$SF_h^k(t) - SC_k^h(t) > 0. \tag{8}$$

B. Problem of An Existing Method

Herein, we consider such a case that for a data $d \in \mathcal{D}$, the number of its data holders in CSP k at time t is $N_k^d(t)$ and the deduplication rate is $r_k^d(t)$. The main idea in [23], which is the only one payment-based incentive mechanism in deduplication we found, is to let all holders of the same data to share the initial unit storage fee when S-DEDU is applied. That is,

$$\alpha_k^d(t) = 1 - \frac{1}{r_k^d(t) \times N_k^d(t)} = 1 - \frac{1}{n_k^d(t)}. \tag{9}$$

Proposition 1. *The IC-Constraint is not fulfilled when CSP k applies (9) to provide incentives.*

Proof. *(Counter-evidence.) If IC-Constraint for k is achieved, then $U_k^1(t) - U_k^0(t) \geq 0$. With (9), we obtain*

$$\left(n_k^d(t) - 1\right) \times \left(SC_k^h(t) - SF_h^k(t)\right) - RF_k(t) \geq 0.$$

Since $RF_k(t)$ is non-negative, then

$$\left(n_k^d(t) - 1\right) \times \left(SC_k^h(t) - SF_h^k(t)\right) \geq 0. \tag{10}$$

Therefore, $SC_k^h(t) - SF_h^k(t)$ must be greater than 0. However, $SF_h^k(t) - SC_k^h(t) > 0$ is a common consensus in our model (refer to (8)). Hence, the IC-Constraint is not fulfilled.

C. Our Proposed Mechanism

The above method [23] provides great benefits to data holders without considering CSPs, which is not applicable in our scenario. To solve this problem, we intend to set discount boundaries for data holders to ensure the profits of CSPs.

We design the discount for each data as follows:

$$\alpha_k^d(t) = r_k^d(t) \times \left(\alpha_{max}^{k,d} - \alpha_{min}^{k,d}\right), \tag{11}$$

where $\alpha_{max}^{k,d}$ and $\alpha_{min}^{k,d}$ are set to be the maximum and minimum discounts that k can give to the holders of data d.

Herein, let φ_k denote the constant $\frac{SF_h^k(t)}{SC_k^h(t)}$. Since the discount function is decided by k, it considers its IR-Constraint and IC-Constraint. As $U_k^0(t) > 0$ is a common consensus, then $\varphi_k > 1$. $U_k^1(t) - U_k^0(t) > 0$ (i.e., IC-Constraint) implies $U_k^1(t) > 0$ (i.e., IR-Constraint). Subtracting (4) from (5), we obtain

$$\sum_{d\in\mathcal{D}} \left(n_k^d(t) - \alpha_k^d(t) \times n_k^d(t) \times \varphi_k - 1\right) \times SC_k^h(t) - RF_k(t).$$

To make $n_k^d(t) - \alpha_k^d(t) \times n_k^d(t) \times \varphi_k - 1 > 0$, then

$$\alpha_k^d(t) < \frac{n_k^d(t) - 1}{n_k^d(t) \times \varphi_k}. \tag{12}$$

As $\frac{\partial\left(\frac{n_k^d(t)-1}{n_k^d(t)\times\varphi_k}\right)}{\partial r_k^d(t)} = \frac{N_k^d(t)\times\varphi_k}{\left(\varphi_k \times r_k^d(t) \times N_k^d(t)\right)^2} > 0$, then $\frac{n_k^d(t)-1}{n_k^d(t)\times\varphi_k}$ shares the same variation with $r_k^d(t)$. We can easily conclude that $max\frac{n_k^d-1}{n_k^d(t)\times\varphi_k} = \frac{N_k^d(t)-1}{\varphi_k\times N_k^d(t)}$ and $min\frac{n_k^d(t)-1}{n_k^d(t)\times\varphi_k} = 0$. We can set

$$\alpha_{max}^{k,d} = \frac{N_k^d(t) - 1}{(\varphi_k + \delta) \times N_k^d(t)}, \alpha_{min}^{k,d} = 0.$$

$\delta > 0$ is to adjust the discount in S-DEDU. It is uncomplicated to prove that $\alpha_{max}^{k,d}$ increases with the increase of $N_k^d(t)$ by the derivative method. Therefore, (11) is detailed as

$$\alpha_k^d(t) = r_k^d(t) \times \frac{N_k^d(t) - 1}{(\varphi_k + \delta) \times N_k^d(t)}. \tag{13}$$

Proposition 2. *With our proposed incentive mechanism, a data holder cannot obtain more profits by creating some fake identities to intentionally increase the deduplication rate for more discounts.*

Proof. *If there is a data holder who creates $N \geq 2$ accounts at CSP k to store the same data d in order to take the advantage of discounts, the storage fee it pays for all accounts is: (when all deduplicated, $r_k^d(t) = 1$)*

$$\begin{aligned}
&N \times \left(1 - \alpha_k^d(t)\right) \times SF_h^k(t) \\
&= N \times \left(1 - r_k^d(t) \times \frac{N-1}{(\varphi_k + \delta) \times N}\right) \times SF_h^k(t) \\
&\geq N \times \left(1 - \frac{N-1}{(\varphi_k + \delta) \times N}\right) \times SF_k^h(t) \\
&> N \times \left(1 - \frac{N-1}{N}\right) \times SF_h^k(t) = SF_h^k(t)
\end{aligned}$$

Therefore, $N \times \left(1 - \alpha_k^d(t)\right) \times SF_h^k(t) > SF_h^k(t)$. Namely, the storage fee it pays for all accounts is higher than that for one account. However, the benefits for storing several copies of data will not increase and the potential loss will not decrease. Hence, a data holder cannot obtain more profits by storing the same data in different accounts. Proposition 2 is proved.

D. Parameter-Setting and Strategy-Choosing Algorithm

In this section, we present an algorithm to instruct CSPs to choose system parameters and show how the data holders and CSPs select their strategies.

The first thing for all players to decide is whether to choose the cloud storage service. CSP k sets its default storage fee $SF_h^k(t)$ based on its storage cost $SC_k^h(t)$ to ensure (8). For a data holder h, the influence of data-mismanagement w_h and the cloud storage benefit $B_h(t)$ are known information and the mismanagement possibility p_k can be inferred through social networks. Data holder h calculates $(1 - p_k \times w_h) \times B_h(t) - SF_h^k$ and only chooses to store data at k when (7) is satisfied.

The next thing is to decide whether to accept S-DEDU with our payment discount based incentive mechanism (13). The value of φ_k is fixed when k has set the default value of $SF_h^k(t)$ and $SC_k^h(t)$ is a constant. k can calculate $r_k^d(t)$ easily from $r_k^d(t) = \frac{n_k^d(t)}{N_k^d(t)}$. Therefore, the only parameter in (7) that needs k to decide is δ. Since our incentive mechanism is designed from the point of individual rationality and incentive compatibility, as long as $\delta > 0$, our incentive can make sure the non-negative utility of a CSP and this utility is higher than that of a CSP without S-DEDU.

The best value of δ that ensures the highest benefits of CSP is difficult to decide. Theoretically, the larger δ is, the smaller $\alpha_k^d(t)$ is, the less discount k gives out, and the less benefit h obtains. The number of data holders in k could decrease with the drop of the utility of h, which poses a negative effect on the utility of k. An experienced CSP can infer the value of the minimum influence of data mismanagement on a data holder w_{min}, the maximum influence w_{max}, and the cloud storage benefit $B_h(t)$ from its empirical observation. According to the individual rationality and incentive compatibility of data holders, for the data holder with w_h, δ is should satisfy

$$\alpha_k^d(t) \times SF_h^k(t) - r_k^d(t) \times L_h(t) > 0. \tag{14}$$

Taking (1) and (13) into (14), we obtain

$$\delta < \frac{\left(N_k^d(t) - 1\right) \times SF_h^k(t)}{N_k^d(t) \times p_k \times w_h \times B_h(t)} - \varphi_k. \tag{15}$$

Let

$$\delta_1 = \frac{\left(N_k^d(t) - 1\right) \times SF_h^k(t)}{N_k^d(t) \times p_k \times w_{min} \times B_h(t)} - \varphi_k \tag{16}$$

$$\delta_2 = \frac{\left(N_k^d(t) - 1\right) \times SF_h^k(t)}{N_k^d(t) \times p_k \times w_{max} \times B_h(t)} - \varphi_k. \tag{17}$$

For each data, $N_k^d(t)$ is fixed and p_k is constant for any given k. δ_1 and δ_2 are fixed values if $SF_h^k(t)$ and $SC_k^h(t)$ are fixed. When k chooses $\delta_2 \leq \delta \leq \delta_1$ and $w_h \sim U\left[w_{min}, w_{max}\right]$, the number of data holders that accept S-DEDU is

$$n_k^d(t) = \frac{\delta - \delta_2}{\delta_1 - \delta_2} \times N_k^d(t). \tag{18}$$

Let $u(t) = U_k^1(t) - U_k^0(t)$ and with (13) and (18), we have

$$\begin{aligned}
u(t) = &\sum_{d\in\mathcal{D}} \frac{\delta - \delta_2}{\delta_1 - \delta_2} \times N_k^d(t) \times SC_k^h(t) - \sum_{d\in\mathcal{D}} SC_k^h(t) \\
&- \sum_{d\in\mathcal{D}} \left(\frac{\delta - \delta_2}{\delta_1 - \delta_2}\right)^2 \times \frac{N_k^d(t) - 1}{\varphi_k + \delta} \times \varphi_k \times SC_k^h(t) - RF_k(t).
\end{aligned}$$

Therefore, CSP k empirically chooses the value of δ, which satisfies $\delta > 0$ and $\delta_2 \leq \delta \leq \delta_1$, to achieve the maximized value of $u(t)$.

Parameter-Setting and Strategy-Selection Algorithm

Input: Parameters: $SC_k^h(t), B_h(t), L_h(t), \mathcal{H}, \alpha_k^d(t-1), T$
Output: Parameters: $SF_h^k(t), \alpha_k^d(t)$;
The strategies for each data holders;
Initialization: $\alpha_k^d(0) = 0,\ r_k^d(0) = 0$;
CSP k empirically chooses the value of $\delta > 0$ to set $\alpha_k^d(t)$;
For $t = 1$ to T **do**
  CSP k sets its required storage fee according to $SF_h^k(t) > SC_k^h(t)$;
  CSP k calculates $\alpha_k^d(t)$ according to $r_k^d(t-1)$;
  Foreach data holder $h \in \mathcal{H}$
    h calculates $U_h^0(t)$;
    If $U_h^0(t) > 0$
      h stores its data at k;
      h calculates $U_h^1(t)$;
      If $U_h^1(t) > U_h^0(t)$
        h follows S-DEDU (i.e., accepts S-DEDU);
      Else
        h refuses S-DEDU;
      End If
    Else
      h chooses local storage;
    End If
  End Foreach
  Get the strategy profile for all data holders at time t;
  CSP k calculates $r_k^d(t)$ based on the above strategy profile;
  $t \leftarrow t + 1$;
End For

Fig. 2. The parameter-setting and strategy-choosing algorithm.

After setting $\alpha_k^d(t)$, each data holder h calculates its utility $U_h^1(t)$ under this incentive provided by k and compares the utility with $U_h^0(t)$. If $U_h^1(t) > U_h^0(t)$, then h accepts S-DEDU. $n_k^d(t)$ increases by 1 and $r_k^d(t)$ also increases accordingly. After all data holders make up their decisions on the acceptance of S-DEDU, k re-calculates $r_k^d(t)$ and updates $\alpha_k^d(t)$. Fig. 2 provides an intuition on how CSPs set their parameters and data holders react to the settings.

E. Inter-CSP Deduplication Scenario

We further consider an inter-CSP scenario for S-DEDU, in which little adaptation is needed in the payoff function of data holders because the procedures in I-DEDU are almost the same as S-DEDU. We denote the utility of h with I-DEDU as $U_h^2(t)$. The structure of $U_h^2(t)$ is the same as $U_h^1(t)$. However, the deduplication rate $r_k^d(t)$, which influences the discount for h, may increase due to the cooperation of different CSPs.

When CSP k_1 notices a data to be unique throughout its storage space, it requests duplication check to other CSPs rather than directly storing the data in its local space. If CSP k_2 has stored the data, k_1 can cooperate with k_2 for deduplication and just pays some access fees $AF_{out}^k(t)$ to k_2, which is calculated based on the unit access fee AF^k and the number of related data holders. Similarly, k_1 can also obtain some access fee $AF_{in}^k(t)$ from other CSPs when it stores some unique data that the others want to store.

We list the detailed payoff functions of data holder h and CSP k with I-DEDU as follows.

$$U_h^2(t) = U_h^1(t).$$

$$U_k^2(t) = U_k^1(t) + AF_{in}^k(t) - AF_{out}^k(t).$$

In Section V, we further conduct an experiment to illustrate the acceptance of I-DEDU based on the above payoff structures.

TABLE II
PARAMETER SETTINGS

Symbols	Values	Symbols	Values	Symbols	Values
B_h	2.165	RF_k	40	OC_{AP}	10
SF_h^k	0.165	p_k	0.01	δ	3
SC_k^h	0.1	w_{min}	0	w_{max}	3
AF_k	0.1				

V. EVALUATION: EXPERIMENTAL RESULTS AND ANALYSIS

A. Experimental Data and Settings

The experimental data is a dataset consist of the information of Debian packages collected from the Debian Popularity Contest [28]. Each package represents a unique data in the cloud and the installation requests can simulate duplicated data storage requests. We select the packages in the section *contrib* (https://popcon.debian.org/contrib/index.html) to form our testing dataset. The reason for choosing this section is its data diversity. To be specific, its data sizes are diverse; different data have a different number of holders.

We took a snapshot on 19th June 2018 to record the number of package installations, the number of packages and the sizes of current-version packages. Our testing dataset consists of 309052 Debian package installations (i.e., data holders). The total number of packages (i.e., data) is 434. In our experiments, w_h obeys a uniform distribution on the interval $[w_{min}, w_{max}]$, denoted as $w_h \sim U\,[w_{min}, w_{max}]$.

After all players taking their actions, we state that a **time generation** passes. CSP announces its storage charge after each time generation based on its utility and deduplication situation. For the next generation, they take actions based on updated utilities and so forth. The game reaches its **NE** when all players have no incentive to change their strategies.

Deduplication percentage [16], [29] is a parameter employed to estimate the effectiveness of deduplication. We use the notation $\rho(t)$ to present the deduplication percentage in a cloud storage system at time t. If we apply $S'(t)$ and $S(t)$ to denote the total size of data that really stored and the total size of data that requested to be stored in the cloud at time t, the detailed expression of $\rho(t)$ is: $\rho(t) = \left(1 - \frac{S'(t)}{S(t)}\right) \times 100\%$.

Notably, the deduplication rate refers to the percentage of data holders that choose deduplication with regard to certain data. While the deduplication percentage refers to the percentage of saved data storage space within a CSP.

Table II provides the default experimental settings of system parameters. We set SF_k^h according to the price list in QI NIU [30]. B_h and OC_{AP} were set based on [27], in which PRE was applied. The value of SC_k^h was set based on (8). RF_k was chosen to ensure non-negative utility of AP (P-Constraint) and p_k was set as 0.01. We also chose different values of p_k to investigate its influence in Section V-C. In addtion, δ was set as 3 to ensure the utility of CSPs is non-dropping.

B. Experiment 1: The Acceptance of S-DEDU

Experiment 1 aims to evaluate whether the S-DEDU scheme is acceptable in practice with our incentive mechanism. We

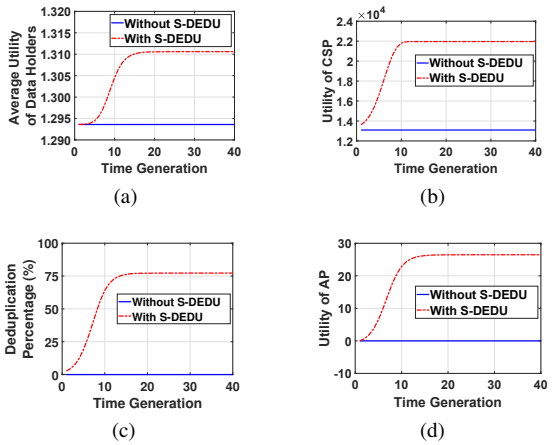

Fig. 3. The results of Experiment 1: (a) average utility of data holders; (b) utilities of two CSPs; (c) deduplication percentages of two CSPs; (d) utilities of AP without S-DEDU and with S-DEDU in different time generations.

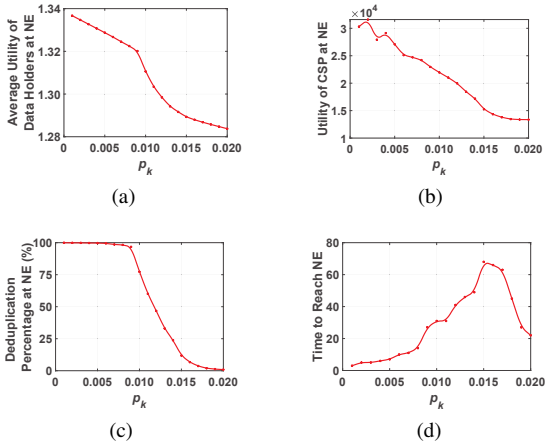

Fig. 4. The effect of p_k on: (a) average utility of data holders at NE; (b) utility of CSP at NE; (c) deduplication percentage at NE; (d) time to reach NE.

considered two CSPs, denoted as C1 and C2, in Experiment 1. C2 adopts S-DEDU and C1 does not. C1 and C2 have the same data holder set as specified in Section V-A. Namely, 309052 data holders with 434 unique data to be stored. C1 publishes it storage-service fee $SF_h^k(t)$ and the data holders decide whether to store at C1 according to their expected utilities (i.e., (2)). C1 stores the data of all the data holders directly. C2 publishes its storage-service fee $SF_h^k(t)$ and grants discount $\alpha_k^d(t)$ to the holders. The data holders compare their utilities without S-DEDU and with S-DEDU according to (2) and (3), and choose the one with more profits. C2 adjusts the discounts as time goes by since the deduplication rate is changing. We recorded the utilities of all stakeholders along with the deduplication percentage in each CSP.

To analyze the acceptance of data holders, we recorded the average utilities of data holders in C1 and C2. Fig. 3a shows that the average utilities of data holders in C1 and C2 are similar at the beginning (from the first to the third time generation). However, the difference between these two curves becomes larger and larger until about the 15th time generation, after which the average utility of data holders in C2 becomes stable. The average utility of data holders in C2 was non-negative and at least the same as that in C1; therefore, our incentive mechanism ensures the individual rationality and incentive compatibility of data holders. From this point of view, S-DEDU is acceptable to data holders.

Fig. 3b plots the utilities of C1 and C2. The utility of C1 was stable while that of C2 was gradually increased and reached stability from about the 10th time generation. The red dotted curve is always above the blue solid line that is larger than 0. Therefore, our incentive mechanism is individually rational and incentive compatible for C2, which illustrates the acceptance of CSPs to S-DEDU.

Fig. 3c depicts the deduplication percentage in C1 and C2. Since no deduplication schemes were adopted by C1, the deduplication percentage of C1 remained 0 all the time. The red dotted line, which shows the deduplication percentage in C2, increases as time goes by and becomes stable at about 75%. According to (13), the discount increases with the augment of the deduplication rate. Therefore, a data holder accepting S-DEDU can obtain more and more discount as time goes by, which is the reason behind the increase of red dotted curve in Fig. 3a. Fig. 3c shows that even if the data mismanagement with the possibility set in Table II is introduced, S-DEDU is also acceptable, which is reflected by its deduplication percentage. Therefore, our incentive mechanism makes S-DEDU robust enough to resist system disturbances.

The final feasibility condition for our incentive mechanism to achieve is profitability. Accordingly, we plotted the utility of AP in different time generations in Fig. 3d. C1 does not adopt S-DEDU so that it does not pay to AP. When S-DEDU is applied in C2, the utility of AP increases with the increasing number of data holders that select S-DEDU. The red dotted curve in Fig. 3d is above 0. Hence, our incentive mechanism is profitable for AP.We conducted additional experiments in the case of multiple CSPs and achieved similar results. Due to paper size limitation, we omit this part of the results.

C. Experiment 2: Effects of System Parameters

In Experiment 2, we investigated the effects of some system parameters on the acceptance of S-DEDU. To be precise, these system parameters include the possibility of mismanagement p_k, the maximum value of w_h (namely, w_{max}), and the parameter δ. We repeated the procedure of C2 in Experiment 1 by changing the above-mentioned parameters one by one while keeping other parameter settings as Experiment 1.

Fig. 4 shows the experimental results when p_k varies, $w_{max} = 3$ and $\delta = 3$. Fig. 4a and Fig. 4b indicate that the average utility of data holders and the utility of CSP at NE decrease with the increase of p_k. Fig. 4c shows that the deduplication percentage at NE does not decrease sharply before p_k reaches a threshold (0.01 in our settings). Fig. 4d displays the time to reach NE with different possibilities, and the approximate trend of the curve is presented as a bell curve. We can conclude that CSP should make efforts to improve its service robustness in order to gain more profits.

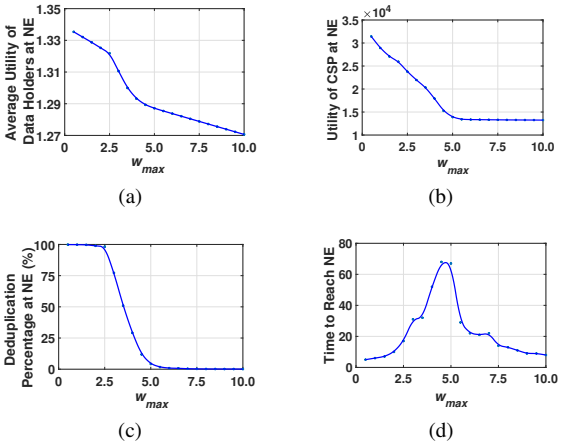

Fig. 5. The effect of w_{max} on: (a) average utility of data holders at NE; (b) utility of CSP at NE; (c) deduplication percentage at NE; (d) time to reach NE.

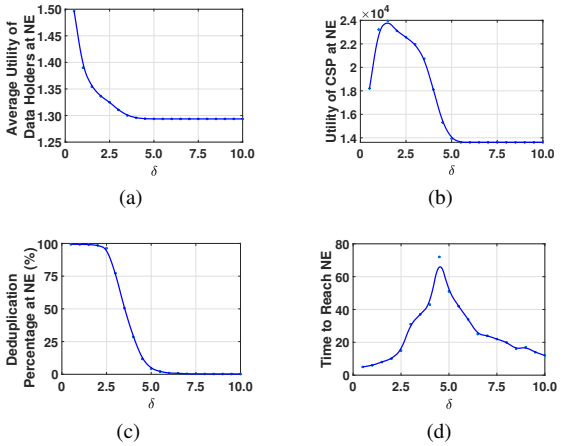

Fig. 6. The effect of δ on: (a) average utility of data holders at NE; (b) utility of CSP at NE; (c) deduplication percentage at NE; (d) time to reach NE.

Furthermore, we investigated the effect of w_{max}. The parameters p_k and δ were set as 0.01 and 3 in this test, respectively. We chose the value of w_{max} from 0 to 10. The average utility of data holders and the utility of CSP at NE decrease with the increase of w_{max}, as demonstrated in Fig. 5a and Fig. 5b, respectively. When w_{max} is bigger than 5, the utility of CSP at NE reaches its lowest value, which is almost the same as that when S-DEDU is not applied. In other words, almost all data holders refuse to accept S-DEDU when w_{max} is more than 5. Fig. 5c demonstrates that when w_{max} is larger than 5, the deduplication percentage drops to almost 0. Fig. 5d presents how the time to reach NE changes with w_{max}. The curve in Fig. 5d increases first and then decreases like that in Fig. 4d. An intuitive conclusion from this result is the acceptance of S-DEDU highly depends on the data-availability concerns of data holders.

Fig. 6 plots the effect of the parameter δ when $p_k = 0.01$ and $w_{max} = 3$. The discount a data holder obtains decreases with the increase of δ. Therefore, the average utility of data holders decreases when the value of δ increases, as shown in Fig. 6a. The utility of CSP increases when the discount is reduced at the beginning. However, when the discount is lower than a threshold (i.e., the value of δ is larger than a threshold), the number of data holders that accept S-DEDU reduces significantly because their expected utilities with S-DEDU are lower than those without it. Lacking enough data holders influences the utility of CSP with S-DEDU finally. Therefore, as illustrated in Fig. 6b, the utility of CSP increases first and then declines to a low value. Fig. 6c plots that the deduplication percentage starts to decline when δ is larger than 2. Fig. 6d implies the time to reach NE increases first and then decreases with the increase of δ. From this sub-experiment, we can conclude that CSP cannot increase the value of δ as large as possible for gaining more profits. There would be a different best δ under different parameter settings. Nevertheless, our incentive mechanism guarantees that the utilities of all players are no less than those when S-DEDU is not applied.

D. Experiment 3: The Acceptance of I-DEDU

We further conducted Experiment 3 to extend our analysis and exploration from intra-CSP deduplication to inter-CSP deduplication by comparing the experimental results of S-DEDU and I-DEDU.

We randomly labeled all data holders from number 1 to 309052 and classified them into two CSPs (C3 and C4). If a data holder was labeled with the number n and $n \mod 2 = 1$, the data holder chose C3. Otherwise, it stored data at C4. We performed two sub-experiments under the above settings. All players choose whether to accept S-DEDU in the first sub-experiment and to accept I-DEDU in the second one. We recorded the average utilities of data holders in C3 and C4, the average utilities of C3 and C4, the average deduplication percentages of C3 and C4, and the utilities of AP in both sub-experiments. The system parameters were set according to Table II. The blue solid curves in Fig. 7 express the results of the sub-experiment with S-DEDU and the red dotted curves represent the results of the sub-experiment with I-DEDU.

All the red dotted curves are above the blue solid ones Fig. 7 and they share the same variation trend in individualized sub-figures. Therefore, I-DEDU is more profitable than S-DEDU with our incentive mechanism. In addition, since we have showed the individual rationality, incentive compatibility, profitability and robustness of the proposed incentive mechanism in S-DEDU, we can easily conclude that it also satisfies the above properties in I-DEDU.

E. Comparison

We compare our paper work with [1], [16], [23], [25], [26] in terms of five aspects of a deduplication scheme: category, objective, incentive, features of incentive and security, as shown in Table III. We can see that the focus of our work is different from previous ones [16], [23], [25], [26]. It concentrates on studying a practical solution for technology deployment and acceptance, i.e., a novel incentive mechanism for S-DEDU adoption. This is missed in most of the previous related works [16], [23], [25], [26]. Compared with [23], our incentive mechanism is more advanced since it can support individual rationality, incentive compatibility and profitability,

TABLE III
COMPARISON WITH PREVIOUS WORKS

Scheme	Category	Objective	Incentive	Features of Incentive	Security
[1]	Server-controlled	Technical solution	×	-	Malicious data holders, Collusion-resistance; Data disclosure-resistance
[16]	Client-controlled	Technical solution	×	-	Collusion-resistance; Brute-force attack-resistance; Data disclosure-resistance
[23]	Server-controlled	Technical solution	√	IR	Information forgeability-resistance; Data disclosure-resistance
[25]	Server-controlled	Technical solution	×	-	Malicious data holders, curious/rational CSPs; Data disclosure-resistance
[26]	Not specified	Technical solution	×	-	Side-channel attack-resistance
This paper	Server-controlled	Practical acceptance study for technology adoption	√	IR, IC, and P	Malicious data holders, Collusion-resistance; Data disclosure-resistance; Sybil attack-resistance

IR: Incentive Rationality; IC: Individual Compatibility; P: Profitability.

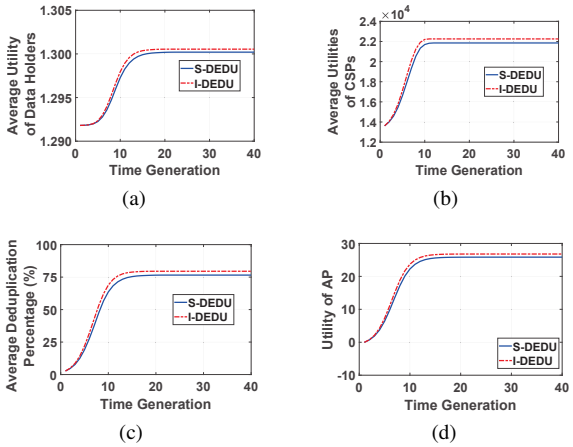

Fig. 7. The results of Experiment 3: (a) average utility of data holders; (b) average utilities of two CSPs; (c) average deduplication percentage of two CSPs; (d) utilities of AP without S-DEDU and with I-DEDU in different time generations.

while only individual rationality can be supported in [23]. In addition, we proved that a data holder cannot earn more by creating lots of identities and uploading the same data multiple times. Thus, our incentive mechanism can resist Sybil attack, which is not supported in other works.

VI. CONCLUSION

This paper analyzed the acceptance of S-DEDU based on a non-cooperative game. We detailed the payoff structure for S-DEDU and devised a new incentive mechanism based on payment discount for data holders. We also took data mismanagement into consideration in the evaluation of the robustness of our incentive mechanism. Through theoretical analysis and experimental evaluation over a real-world dataset, we proved the acceptance of S-DEDU in the condition that the proposed incentive mechanism should be applied. We also investigated the effects of three system parameters: the probability of CSP mismanagement, the influence of mismanagement and the parameter to adjust payment discount. Experimental results showed that the proposed incentive mechanism achieves the design goals of individual rationality, incentive compatibility, and profitability with the concern of system robustness.

ACKNOWLEDGMENT

Liang and Yan's research was sponsored by the NSFC (grants 61672410, 61802293 and U1536202), Academy of Finland (grants 308087 and 314203), the National Key Research and Development Program of China (grant 2016YFB0800704), National Postdoctoral Program for Innovative Talents (grant BX20180238), the Project funded by China Postdoctoral Science Foundation (grant 2018M633461), the Fundamental Research Funds for the Central Universities (grant JB191504), and the 111 project (grant B16037). The work of Lou and Hou was supported in part by US National Science Foundation under Grant CNS-1443889.

REFERENCES

[1] Z. Yan, W. Ding, X. Yu, H. Zhu, and R. H. Deng, "Deduplication on encrypted big data in cloud," *IEEE Trans. Big Data*, vol. 2, no. 2, pp. 138–150, Jun. 2016.
[2] P. Mell and T. Grance, "The NIST definition of cloud computing," *National institute of standards and technology*, vol. 53, no. 6, 2009.
[3] L. Wei, H. Zhu, Z. Cao, X. Dong, W. Jia, Y. Chen, and A. V. Vasilakos, "Security and privacy for storage and computation in cloud computing," *Inform. Sciences*, vol. 258, pp. 371–386, Feb. 2014.
[4] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," *ACM Trans. Storage*, vol. 7, no. 4, pp. 14:1–14:20, Feb. 2012.
[5] M. Ali, R. Dhamotharan, E. Khan, S. U. Khan, A. V. Vasilakos, K. Li, and A. Y. Zomaya, "SeDaSC: Secure data sharing in clouds," *IEEE Syst. J.*, vol. 11, no. 2, pp. 395–404, Jun. 2017.
[6] Z. Fu, F. Huang, K. Ren, W. Jian, and C. Wang, "Privacy-preserving smart semantic search based on conceptual graphs over encrypted outsourced data," *IEEE Trans. Inf. Forensics Security*, vol. 12, no. 8, pp. 1874–1884, Aug. 2017.
[7] J. Li, Y. K. Li, X. Chen, P. P. C. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication," *IEEE Trans. Parallel Distrib. Syst.*, vol. 26, no. 5, pp. 1206–1216, May 2015.
[8] M. Bellare, S. Keelveedhi, and T. Ristenpart, "Message-locked encryption and secure deduplication," in *Advances in Cryptology – EUROCRYPT 2013*, T. Johansson and P. Q. Nguyen, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2013, pp. 296–312.
[9] S. Keelveedhi, M. Bellare, and T. Ristenpart, "DupLESS: Server-aided encryption for deduplicated storage," in *USENIX Security 13*, Washington, D.C., 2013, pp. 179–194.
[10] M. Wen, K. Lu, J. Lei, F. Li, and J. Li, "BDO-SD: An efficient scheme for big data outsourcing with secure deduplication," in *INFOCOM WKSHPS 2015*, Hong Kong, China, 2015, pp. 214–219.
[11] J. Blasco, R. D. Pietro, A. Orfila, and A. Sorniotti, "A tunable proof of ownership scheme for deduplication using bloom filters," in *CNS'14*, San Francisco, CA, USA, 2014, pp. 481–489.
[12] L. González-Manzano and A. Orfila, "An efficient confidentiality-preserving proof of ownership for deduplication," *J. Netw. Comput. Appl.*, vol. 50, pp. 49–59, 2015.

[13] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in *CCS'11*, Chicago, Illinois, USA, 2011, pp. 491–500.

[14] J. Li, X. Chen, X. Huang, S. Tang, Y. Xiang, M. M. Hassan, and A. Alelaiwi, "Secure distributed deduplication systems with improved reliability," *IEEE Trans. Comput.*, vol. 64, no. 12, pp. 3569–3579, Dec. 2015.

[15] Y. Zheng, X. Yuan, X. Wang, J. Jiang, C. Wang, and X. Gui, "Enabling encrypted cloud media center with secure deduplication," in *ASIACCS'15*, Singapore, Republic of Singapore, 2015, pp. 63–72.

[16] J. Liu, N. Asokan, and B. Pinkas, "Secure deduplication of encrypted data without additional independent servers," in *CCS'15*, Denver, Colorado, USA, 2015, pp. 874–885.

[17] Z. Yan, L. Zhang, W. Ding, and Q. Zheng, "Heterogeneous data storage management with deduplication in cloud computing," *IEEE Transactions on Big Data*, 2017.

[18] Y. Shin, D. Koo, and J. Hur, "A survey of secure data deduplication schemes for cloud storage systems," *ACM Comput. Surv.*, vol. 49, no. 4, pp. 74:1–74:38, Jan. 2017.

[19] https://en.wikipedia.org/wiki/Bitcasa, retrieved December 7, 2018.

[20] https://en.wikipedia.org/wiki/Wuala, retrieved December 7, 2018.

[21] V. Rabotka and M. Mannan, "An evaluation of recent secure deduplication proposals," *J. Inf. Secur. Appl.*, vol. 27-28, pp. 3–18, 2016.

[22] T.-Y. Youn and K.-Y. Chang, "Necessity of incentive system for the first uploader in client-side deduplication," in *Advances in Computer Science and Ubiquitous Computing*, D.-S. Park, H.-C. Chao, Y.-S. Jeong, and J. J. J. H. Park, Eds. Singapore: Springer Singapore, 2015, pp. 397–402.

[23] M. Miao, T. Jiang, and I. You, "Payment-based incentive mechanism for secure cloud deduplication," *Int. J. Inform. Manage.*, vol. 35, no. 3, pp. 379–386, Jun. 2015.

[24] B. M. Roger, "Game theory: analysis of conflict," 1991.

[25] F. Armknecht, J.-M. Bohli, G. O. Karame, and F. Youssef, "Transparent data deduplication in the cloud," in *CCS'15*, Denver, Colorado, USA, 2015, pp. 886–900.

[26] B. Wang, W. Lou, and Y. T. Hou, "Modeling the side-channel attacks in data deduplication with game theory," in *2015 IEEE Conference on Communications and Network Security (CNS)*, 2015, pp. 200–208.

[27] L. Gao, Z. Yan, and L. T. Yang, "Game theoretical analysis on acceptance of a cloud data access control system based on reputation," *IEEE Trans. Cloud Comput.*, pp. 1–1, 2018.

[28] http://popcon.debian.org, home Page. Retrieved December 7, 2018.

[29] M. Dutch, "Understanding data deduplication ratios," in *SNIA Data Management Forum*, 2008, p. 7.

[30] https://www.qiniu.com/prices, retrieved December 7, 2018.

Xueqin Liang received the B.Sc. degree on Applied Mathematics from Anhui University, Anhui, China, 2015. She is currently working for her PhD degree at Xidian University, Xi'an, China, and Aalto University, Finland. Her research interests are in game theory based security solutions, cloud computing security and trust, and IoT security.

Zheng Yan is currently a professor at the Xidian University, China and a visiting professor and Finnish academy research fellow at the Aalto University, Finland. She received the Doctor of Science in Technology from the Helsinki University of Technology, Finland. Before joining academia in 2011, she was a senior researcher at the Nokia Research Center, Helsinki, Finland, since 2000. Her research interests are in trust, security, privacy, and security-related data analytics. She is an associate editor of IEEE Internet of Things Journal, Information Fusion, Information Sciences, IEEE Access, and JNCA. She served as a general chair or program chair for numerous international conferences including IEEE TrustCom 2015. She is a founder steering committee co-chair of IEEE Blockchain conference. She received several awards, including the 2017 Best Journal Paper Award issued by IEEE Communication Society Technical Committee on Big Data and the Outstanding Associate Editor of 2017/2018 for IEEE Access.

Xiaofeng Chen received his B.S. and M.S. on Mathematics from Northwest University, China in 1998 and 2000, respectively. He got his Ph.D degree in Cryptography from Xidian University in 2003. Currently, he works at Xidian University as a professor. His research interests include applied cryptography and cloud computing security. He has published over 200 research papers in refereed international conferences and journals. His work has been cited more than 7000 times at Google Scholar. He is in the Editorial Board of *IEEE Transactions on Dependable and Secure Computing*, *Security and Privacy*, and *Computing and Informatics (CAI)* etc. He has served as the program/general chair or program committee member in over 30 international conferences.

Laurence T. Yang received the B.E. degree in computer science and technology and the B.S. degree in applied physics both from Tsinghua University, Beijing, China, in 1992, and the Ph.D. degree in computer science from the University of Victoria, Victoria, BC, Canada, in 2006.

He is currently a Professor with St. Francis Xavier University, Antigonish, NS, Canada. His research interests include parallel and distributed computing, embedded and ubiquitous/pervasive computing, and big data. His research has been supported by the National Sciences and Engineering Research Council, and the Canada Foundation for Innovation.

Wenjing Lou is the W. C. English Professor of Computer Science at Virginia Tech and a Fellow of the IEEE. Her research interests cover many topics in the cybersecurity field, with her current research interest focusing on privacy protection in networked information systems and security and privacy problems in the Internet of Things (IoT) systems.

Prof. Lou is currently on the editorial boards of IEEE Transactions on Dependable and Secure Computing (TDSC), ACM/IEEE Transactions on Networking (ToN), IEEE Transactions on Mobile Computing (TMC), and Journal of Computer Security. She is the TPC chair for IEEE INFOCOM 2019 and SecureCom 2019. She is the Steering Committee Chair of IEEE Conference on Communications and Network Security (IEEE CNS), steering committee member of IEEE INFOCOM and IEEE Transactions on Mobile Computing. She served as a program director at US National Science Foundation (NSF) from 2014 to 2017.

Y. Thomas Hou is Bradley Distinguished Professor of Electrical and Computer Engineering at Virginia Tech, Blacksburg, VA, USA, which he joined in 2002. He received his Ph.D. degree from NYU Tandon School of Engineering (formerly Polytechnic Univ.) in 1998. During 1997 to 2002, he was a Member of Research Staff at Fujitsu Laboratories of America, Sunnyvale, CA, USA. Prof. Hou's current research focuses on developing innovative solutions to complex science and engineering problems arising from wireless and mobile networks. He is also interested in wireless security. He has over 250 papers published in IEEE/ACM journals and conferences. His papers were recognized by five best paper awards from the IEEE and two paper awards from the ACM. He holds five U.S. patents. He authored/co-authored two graduate textbooks: *Applied Optimization Methods for Wireless Networks* (Cambridge University Press, 2014) and *Cognitive Radio Communications and Networks: Principles and Practices* (Academic Press/Elsevier, 2009). Prof. Hou was named an IEEE Fellow for contributions to modeling and optimization of wireless networks. He was/is on the editorial boards of a number of IEEE and ACM transactions and journals. He is the Steering Committee Chair of IEEE INFOCOM conference and was a member of the IEEE Communications Society Board of Governors. He was also a Distinguished Lecturer of the IEEE Communications Society.