1-2020

# Game theoretical study on client-controlled cloud data deduplication

Xueqin LIANG

Zheng YAN

Robert H. DENG
*Singapore Management University*, robertdeng@smu.edu.sg

Published in Computers & Security, 2020 April, Volume 91, Article 101730.
https://doi.org/10.1016/j.cose.2020.101730

Accepted version

# Game theoretical study on client-controlled cloud data deduplication

Xueqin Liang [a,b], Zheng Yan [a,b,*], Robert H. Deng [c]

[a] School of Cyber Engineering, Xidian University, Xi'an, China
[b] Department of Communications and Networking, Aalto University, Espoo, Finland
[c] School of Information Systems, Singapore Management University, Singapore, Singapore

*Keywords:*
Cloud data deduplication
Free riding
Game theory
Incentive mechanism
Privacy

A B S T R A C T

Data deduplication eliminates redundant data and is receiving increasing attention in cloud storage services due to the proliferation of big data and the demand for efficient storage. Data deduplication not only requires a consummate technological designing, but also involves multiple parties with conflict interests. Thus, how to design incentive mechanisms and study their acceptance by all relevant stakeholders remain important open issues. In this paper, we detail the payoff structure of a client-controlled deduplication scheme and analyze the feasibilities of unified discount and individualized discount under this structure. Through game theoretical study, a privacy-preserving individualized discount-based incentive mechanism is further proposed with detailed implementation algorithms for choosing strategies, setting parameters and granting discounts. After theoretical analysis on the requirements of individual rationality, incentive compatibility, and profitability, we conduct extensive experiments based on a real-world dataset to demonstrate the effectiveness of the proposed incentive mechanism.

## 1. Introduction

Storing data in the cloud saves local storage spaces and reduces data management and operation costs. A data user can easily access its data in the cloud at any time and everywhere. Significant efforts have been made to securely and efficiently outsource data to the cloud in recent years, ranging from protecting data security and privacy (Chu et al., 2014; Wan et al., 2012; Wang et al., 2010a; 2011; Wei et al., 2014), reducing copyright risks (Hwang et al., 2009; Hwang and Li, 2010), controlling data access (Ruj et al., 2014; Wang et al., 2010b; Yan et al., 2017a; Yang and Jia, 2014; Zhou et al., 2013), to encrypted data deduplication (Harnik et al., 2010; Li et al., 2014; 2015; Liu et al., 2015; Xu et al., 2009; Yan et al., 2016a; 2016b; 2017b).

Cloud data deduplication greatly benefits Cloud Service Providers (CSPs). One data file might be uploaded by many users or by a single user multiple time either intentionally or unintentionally. A CSP with deduplication only stores one copy of every data file in either plaintext or encrypted form, and be able to provide all of its users a way to access the data based on certain access control policies. Hence, data deduplication can greatly reduce the storage overhead of CSPs and allow them to pass the cost savings to their users. There is no doubt that a CSP with less service fee can attract more data users, thus possibly gain more profits.

There are many schemes in the literature on deduplication over encrypted data in the cloud to achieve both data security and economic data storage. One prominent method is the client-side deduplication scheme, in which a data user only needs to upload the real data if the data has never been stored before. It saves more uplink bandwidth and CSP operation costs, and are widely studied by most researchers. Based on the eligibility verification and access control policies, we can classify the client-side deduplication scheme into client-controlled client-side deduplication (C-DEDU) and server-controlled client-side deduplication (S-DEDU).

However, almost all existing deduplication schemes (Harnik et al., 2010; Li et al., 2014; 2015; Liu et al., 2015; Xu et al., 2009; Yan et al., 2016a; 2016b; 2017b) are designed and analyzed only from technological perspectives. Few efforts in the literature were made to investigate the acceptability of deduplication schemes by all stakeholders (e.g., data users and CSPs). Only when a scheme brings tangible profits to its stakeholders can it be adopted in practice. In this paper, we focus on how to promote the acceptance of C-DEDU while the practical deployment problems in S-DEDU are studied in another line of our work.

Three kinds of stakeholders are involved in C-DEDU including data owners, data holders and CSPs. A data owner is the first one to upload a data and the later ones to upload are called data holders. Once a data holder requires to store this data, the data owner checks the eligibility of this holder and only grants the access right

* Corresponding author.
*E-mail addresses:* zyan@xidian.edu.cn, zheng.yan@aalto.fi (Z. Yan), robertdeng@edu.sum.sg (R.H. Deng).

to the eligible ones. CSP is the entity that provides cloud storage service.

Data owners have privileges due to the access control rights, however, need to keep online to perform this control. Various techniques have been proposed to mitigate the online requirement of data owners. In Yan et al. (2016a), a scheme was proposed to allow data owners to hand over the right of controlling data deduplication to a server. Harnik et al. (2010) introduced a simple mechanism that turns off deduplication artificially to ensure privacy preservation. A flexible deduplication scheme, which adaptively selects stakeholders control data deduplication according to the data protection policies of data owners, was introduced in Yan et al. (2017b). However, these methods either change C-DEDU into S-DEDU to avoid the online requirement or are not intelligent. Because the combination of client-side access control and deduplication introduces service-delay to data holders when the owner is temporarily offline, some time-sensitive data holders may be reluctant to adopt C-DEDU.

CSP is the direct beneficiary of deduplication schemes since they are primarily designed to save the storage cost of CSPs. Therefore, it is essential for a CSP to motivate the participation enthusiasm of data owners and data holders. Even though this necessity of incentives is mentioned by researchers, they either failed to propose a concrete mechanism (Liu et al., 2015), or the proposed mechanisms have privacy defect (Armknecht et al., 2015) or are proved to be not incentive compatible (Liang et al., 2019; Miao et al., 2015). Moreover, the complex interdependence among the various stakeholders in deduplication schemes increases the difficulty of weighting the profits from the stakeholders' perspective. Game theory, as a mathematical model of conflict and cooperation study between rational players, has natural advantages to address this problem. It helps to analyze how data owners and data holders choose strategies based on their utility functions. Unfortunately, to our knowledge, no systematically economic model for C-DEDU has been proposed until now.

In this paper, we first specify the employed economic model and introduce the detailed utility functions of data owners, data holders and CSPs. Then we apply game theory to analyze how data owners and data holders react according to different discount-charging models of CSPs and discuss the existence of Nash Equilibrium. To overcome the free-riding behaviors privacy issues, we propose a privacy-preserving incentive mechanism that can motivate rational players (i.e., data owners and data holders) to be honest. Furthermore, we conduct experiments to verify our theoretical analysis and illustrate the effectiveness of the incentive mechanism with a real-world dataset. Specifically, the contributions of this paper can be summarized as below:

1. We systematically propose an economic model for a cloud storage system with C-DEDU. The detailed utility function of each stakeholder is deeply discussed as well.
2. We analyze the advantages and disadvantages of two incentive mechanisms (i.e., unified discount and individualized discount) with a game model between a data owner and a data holder. We find that the individualized discount is more desirable due to the existence of Nash Equilibrium although it may intrude privacy.
3. We further present a new privacy-preserving incentive mechanism that is incentive compatible and motivates rational players (i.e., data owners and data holders) to be honest, thus eliminate the disadvantage of individualized discount.
4. We provide Parameter-Setting Algorithm, Discount-Granting Algorithm, and Strategy-Choosing Algorithms to instruct how our proposed incentive mechanism can be implemented in practice.
5. We discuss how the proposed incentive mechanism is compatible with existing encrypted data deduplication schemes and its scalability and robustness when being triggered by modification attacks.

The rest of the paper is organized as follows. Background and related works are briefly reviewed in Section 2. Section 3 overviews the cloud storage system with C-DEDU, and details its deployment problems, along with clearly specified game-model assumptions. An economic model for the cloud storage system with C-DEDU is proposed in Section 4 based on the assumptions. In Section 5, we perform game-theoretical analysis on two discount methods and propose a privacy-preserving individualized discount-based incentive mechanism, which is able to achieve individual rationality, incentive compatibility and profitability. In Section 6, we evaluate the effectiveness of our proposed incentive mechanism in promoting the acceptance of C-DEDU through a set of experiments based on a real-world dataset and further discuss its compatibility, scalability and robustness. Finally, concluding remarks are drawn in the last section.

## 2. Background and related work

### 2.1. Game theory

Game theory is a branch of applied mathematics but develops considerably in the field of economics. It has been widely deployed in many fields, such as economics, psychology, and even biology. It can flexibly and masterly capture the interactions between different participants. It studies how a rational entity will choose its strategy based on its preference and known information about the others at each step. Researchers in the field of security and privacy (Do et al., 2017; Manshaei et al., 2013) also initiate applying game theory to analyze the interactions among system players.

Player, action, information, strategy, payoff function and equilibrium are the essential elements to describe a game. Each participant as a player can make its own decision based on the obtained information that refers to all messages about the other players, like their payoff functions. After all players make decisions and take actions, they will obtain payoffs according to their interactions. When every player in a game has no incentive to change actions, we say this game reaches its equilibrium. The Nash Equilibrium (NE) in a non-cooperative game is a state where no player can gain more profits by deviating its current strategy.

### 2.2. Encrypted data deduplication

Current storage service faces the explosively growing digital data and the additional storage costs caused by the inadvertent multiple storages and backup considerations. A recent study (Meyer and Bolosky, 2012) performed by Microsoft shows that about 68% data are stored with duplication.

Deduplication is a popular technique for CSPs because it eliminates redundant copies of data stored in the cloud and substitutes them with pointers to a shared copy. Considering data privacy, data users prefer to upload encrypted data to the cloud. Before uploading the real data, the users calculate their data fingerprints based on hash functions and send to CSP for duplication check. Only the data whose fingerprint has not been stored before will be required to upload. Therefore, this kind of deduplication can save not only storage spaces but also network bandwidth consumption. A piece of data could be deduplicated at the file-level (Bolosky et al., 2000) or the chunk-level (Pooranian et al., 2018; Sun et al., 2018), while the latter one is more popular for advance compression performance (Xia et al., 2016).

**Deduplication percentage** is a parameter to indicate the effect of deduplication (Harnik et al., 2010; Liu et al., 2015). Let $\rho_k^i(t)$ be

the deduplication percentage of an owner $o_k^i$'s data at time $t$, then

$$\rho_k^i(t) = \frac{T_k^i(t) - S_k^i(t)}{T_k^i(t)} \times 100\%, \tag{1}$$

where $T_k^i(t)$ is the total data size requested to be stored by $c_k$ at time $t$ and $S_k^i(t)$ is the total size of data that really stored by $c_k$ at time $t$. If $\rho_k(t)$ denotes the deduplication percentage of $c_k$ at time $t$, then

$$\rho_k(t) = \sum_i \frac{T_k^i(t) - S_k^i(t)}{T_k^i(t)} \times 100\%. \tag{2}$$

The necessity of incentive in deduplication has attracted researchers' attention and interests. Liu et al. (2015) encouraged the presence of incentives in motivating data holders to participate in the deduplication scheme designed by them even though they did not solve this issue in their paper. Armknecht et al. (2015) developed ClearBox to enable the data holders to check the deduplication status in a CSP and guarantee a data holder to receive a correct discount with ClearBox. However, the authors did not discuss how to arrange the discounts and how to prevent privacy disclosure when granting discounts. The discount formulation proposed by Miao et al. (2015) is proved to be lack of incentive compatibility for CSPs by Liang et al. (2019).

For more details about encrypted cloud data deduplication schemes, which are not mentioned too much in this paper, the interested readers are recommended to refer to Yan et al. (2019).

### 2.3. Related work

Selfish behaviors are hard to be eliminated in the design of a secure scheme only from a technological perspective. The game theoretical analysis offers great help in solving this issue and ensuring the acceptance and long-term development of the scheme. It has been applied in the fields of cloud computing and networking. Yu et al. (2013) used a game theoretical method to analyze how vehicles optimally share resources to improve network performance when exploiting cloud computing in vehicular networks. In wireless multimedia social networks, Nan et al. (2014) proposed a distributed bandwidth allocation method based on game theory to effectively avoid selfish behaviors of players. Then, resource and reward fair allocation was addressed with a cooperative game theoretical approach in Niyato et al. (2011). Researchers in (Palmieri et al., 2013) proposed a game theory-based distributed task scheduling scheme that can eliminate all entities' selfish behaviors in the cloud and achieve social optimality.

Incentive mechanisms can help in addressing selfish behaviors. Xu et al. (2016) presented how an incentive mechanism can eliminate selfish behaviors in mobile social networks and promote players' active participation. Moreover, reputation mechanisms can induce players to cooperate with each other (Wang et al., 2015). Reward and punishment mechanisms can improve mutual trust between nodes in a cloud system (Wong et al., 2014), which ensures healthy system development. Wong et al. (2014) game-theoretically analyzed how a reputation-based cloud data access control system can be accepted by all system stakeholders by introducing a compensation mechanism and a punishment mechanism. Shen et al. (2014) formulated a trust-based punishment mechanism to incent network entities in a global trust management system to behave cooperatively.

However, the literature has not yet investigated the acceptance of cloud data deduplication and explored an effective mechanism to support its practical deployment and operation.

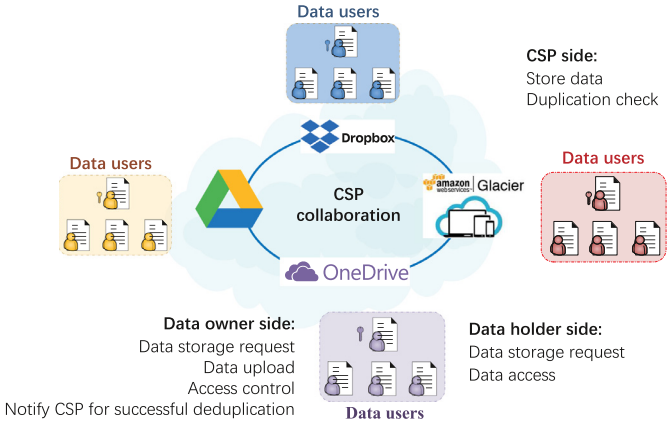


**Fig. 1.** A cloud storage system with C-DEDU.

## 3. System model and problem statement

### 3.1. System model

There is a CSP $c_k$ with $M_k$ unique data to be stored and $N_k$ data users. Let $\mathcal{D}_k$ and $\mathcal{U}_k$ represent the data set and user set. For each data $d_k^m \in D_k$, the number of its data users is denoted as $n_k^m$, then $\sum_{m=1}^{M_k} n_k^m = N_k$. Then $\mathcal{D}_k = \{d_k^m | m = 1, 2, \ldots, M_k\}$ and $\mathcal{U}_k = \left\{u_k^n \mid n = 1, 2, \ldots, N_k\right\} = \bigcup_m \mathcal{U}_{k,m}$, where $\mathcal{U}_{k,m} = \{u_{k,m}^s | s = 1, 2, \ldots, n_k^m\}$.

When $c_k$ adopts C-DEDU, $\mathcal{U}_k$ can be divided into a data owner set $\mathcal{O}_k$ and a data holder set $\mathcal{H}_k$. $\mathcal{O}_k = \left\{o_k^i \mid i = 1, 2, \ldots, M_k\right\} = \{u_{k,m}^1 | m = 1, 2, \ldots, M_k\}$ is composed of the first data uploader of all the unique data in $c_k$. The data holder set is composed of the rest of data users. $\mathcal{H}_k = \left\{h_k^j \mid j = 1, 2, \ldots, N_k - M_k\right\} = \bigcup_m \mathcal{U}_{k,m} \backslash \{u_{k,m}^1\} = \bigcup_m \{u_{k,m}^s | s = 2, \ldots, M_k\}$.

The data owners and data holders are the data users of CSPs. In C-DEDU, the data owner uploads its data to the cloud and controls the data access by verifying the eligibility of data holders. Only an eligible data holder can access the encrypted data stored in the cloud. Fig. 1 shows the structure of a cloud storage system with C-DEDU.

The procedure of C-DEDU is briefly described as below:

1. A CSP receives a data storage request, which contains the data fingerprint/hash value, from a data user.
2. The CSP checks if this data has been stored already by checking the existence of the fingerprint/hash value of this data (i.e., duplication check). If not, it requires the user to upload this data in an encrypted form and regards this user as the owner of this data. Otherwise, it contacts the real data owner for deduplication.
3. When a data owner receives a deduplication request, it controls the data access by checking the data user's eligibility and only issuing keys to the eligible ones. It also notifies the CSP for successful deduplication.
4. The eligible users, called data holders, can access and decrypt the cloud data with the keys from the data owner.

### 3.2. Problem Statement

Past work (Yan et al., 2016b) demonstrated the security and efficiency of C-DEDU. However, its deployment in a practical market depends on whether each system stakeholder can obtain expected profits. In particular, this scheme may confront the following problems in practical deployment.

C-DEDU requires the data owners to keep online so that they need to pay a relatively high cost to perform deduplication than data holders. If a data user cannot obtain enough incentive from this scheme, C-DEDU cannot be implemented smoothly. On the other hand, it is difficult to guarantee data owners to be online all the time, thus service delay is hard to avoid. This delay may cause data holders' serious economic losses in some cases.

CSPs adjust its storage fee according to the saved storage spaces to appear more subscribers; however, the adjustment is complicated. If a CSP sets the same unit storage fee for all data users, a data user can obtain this offer no matter it honestly follows C-DEDU or not. Therefore, they may modify their raw data somehow to avoid deduplication for fast storage service. Obviously, this selfish behavior is disliked by CSPs and damages the interests of the others that make contributions. Setting different unit storage fee for different subscribers may be too complicated and impractical to implement. If the holders of the same data pay the same adjusted storage fee, data privacy issue arises since a curious data holder can infer the existence of a data if its storage fee is adjusted. Therefore, the charging model of CSPs deserves deeply investigated.

Even though a CSP holds the absolute control of designing the charge model, it only survives when having massive users. The utilities of data users (i.e., data owners and data holders) and CSPs are directly linked. Data users pay for the cloud storage services offered by CSPs and CSPs spend some costs to provide such services. How to balance profits and interests among them is an interesting and practical problem.

By taking the above problems into consideration, we aim to propose an incentive mechanism that can promote the acceptance of C-DEDU to all system stakeholders (i.e., data owners, data holders and CSPs) and suppress selfish behaviors in practical implementation.

### 3.3. Assumptions

In what follows, we propose a number of assumptions based on Gao et al. (in press), Yan et al. (2016b). These assumptions play as the basis of our game theoretical analysis on the acceptance of C-DEDU with the proposed incentive mechanism.

*Game assumption*: All players are rational that they choose their own strategies that can bring the best profits. The acceptance of cloud storage system has been proved in Gao et al. (in press), thus we simply assume that all stakeholders would like to choose cloud storage. Local storage will not be a strategy for any players in our game model.

*Scheme assumption*: On the basis of the security analysis of C-DEDU (Yan et al., 2016b), the C-DEDU scheme is secure even when the CSPs cannot be fully trusted. Ineligible users and curious CSPs have no opportunities to access the raw data stored at the CSP. Therefore, data leakage is beyond our consideration in this paper. All CSPs can cooperate with each other to provide cloud storage services. Namely, inter-CSP deduplication is considered. We assume the interaction among CSPs is in a smooth and secure manner. Thus, there are no security concerns aroused during their interactions. A data owner should be online to send a deduplication-successful message. Therefore, it will not fake a deduplication-successful message since it will perform the C-DEDU and do not need to fake a message if it is online.

*User assumption*: The operation cost of a data owner to conduct C-DEDU is related to the number of data holders since it only performs C-DEDU when a user requests to store the same data. To simplify our analysis, we assume each data user requests to store one data at a time and obtains the same benefits for cloud storage. Note that there is no incentive for an owner to upload invalid data. First, if the data is modified before the duplication check, then the fingerprint of this data changes. The other data holders will not be detected as the same data users with this owner. Second, if the data is modified after the owner sent the fingerprint to CSP, it is easy to be detected by the CSP in the duplication check process since the modified data does not match with the previous data fingerprint. Each data upload request is in accordance with a data user index. A holder uploads its data several times will be granted with several indices. Refer to Proposition 2 in Liang et al. (2019), creating multiple copies of data intentionally will not bring a user more profits. Therefore, we keep our application scenario simplified by assuming that a data holder will not upload the same data more than one piece. When a user requires storing data at a CSP with deduplication while the data owner is offline, the user suffers from the storage-service delay. We assume data users are time-sensitive and different data users have different time sensitivities, which is a parameter to instantiate the influence of service delay on the cloud storage benefit of a data user. More sensitive to service delay means a user gains less benefits when the data owner is offline. Therefore, they suffer from some loss when storage-service delay happens. The loss is related to the cloud-storage benefit since the delay hinders the holders to enjoy cloud storage benefits. Due to the individualization of data users, we assume different data holders have different time sensitivities and different data owners have different possibilities to be offline when providing the C-DEDU service.

*Data assumption*: For the simplification of our model and simulation, we assume all data have the same size; therefore, the same storage fee and storage cost. Nevertheless, different data belong to different numbers of users. A data may be modified slightly by a holder to escape from being detected as duplicated. We assume this modification does not influence the data size. Furthermore, in privacy-preserving data deduplication schemes, the fingerprint/hash functions chosen in the duplication check procedure can normally make sure even if a slight change in the data content will result in a totally different output (or fingerprint/hash value). Therefore, a holder can slightly modify its data by adding a string of meaningless characters to avoid being detected as duplicated, but not burden with additional costs.

*CSP assumption*: We assume the storage and maintenance cost of a CSP is related to the number of data stored. Their C-DEDU operation cost is related to the total number of users. Note that there is a basic operation cost spent by CSP. But it exists no matter C-DEDU is applied or not. Thus, we ignore it in our economic model.

## 4. Economic model

Before starting to solve the above problems from the view of economics, we summarize the notations used in this paper in Table 1 for clear presentation and easy reference.

In this section, we build up an economic model to calculate the utilities of data owners, data holders and CSPs, respectively. Since CSPs hold absolute control over designing the charging model, we mainly consider the game between the data owner and data holders. We present the utilities of all stakeholders under different situations as follows.

### 4.1. Payoff structure in cloud storage without C-DEDU

We first present the utility functions of data users and CSPs in a cloud storage system without any deduplication schemes.

For a user $u_k^n$ with data $d_k^m$ stored at the CSP $c_k$, cloud storage provides it data-access convenience and local storage saving. As a profitable organization, $c_k$ also requires the data user to pay for this service. Let $bf_k^n(t)$ and $sf_k^n(t)$ denote the capitalization of the cloud storage benefits (including easy-access, cross-device-access, local-storage-saving, etc.) and the storage-service fee of $u_k^n$ at time

**Table 1**
Notations.

| Notations | Descriptions |
|---|---|
| $o_k^i, h_k^j$ | A specific data owner/holder; |
| $\mathcal{D}, d, m, M$ | The indications of parameters that are related to data; |
| $\mathcal{O}, o, i$ | The indications of parameters that are related to data owners; |
| $\mathcal{H}, h, j$ | The indications of parameters that are related to data holders; |
| $\mathcal{U}, u, n$ | The indications of parameters that are related to data users; |
| $c, k$ | The indications of parameters that are related to CSPs; |
| $f$ | The function mapping from the index of a data holder to that of its data owner; |
| $bf_k$ | The benefits for cloud storing at $k$; |
| $sf_k$ | The storage-service fee paid to $k$; |
| $\alpha_k$ | The discount of the storage-service fee that $k$ sets; |
| $rf_k/RF_k$ | The unit/total request fee paid by $k$; |
| $oc_g/OC_g$ | The unit/total C-DEDU operation cost of $g$, where $g = o, c$; |
| $sc$ | The storage cost; |
| $U_g$ | The utility function of $g$, where $g$=$o, h, u, c$; |
| $\varepsilon$ | The possibility of a data owner to be offline; |
| $l$ | The loss of a data holder suffering from service delay; |
| $\theta$ | The time sensitivity of a data holder; |
| $\eta$ | The data holder number threshold set by $k$ privately. |

$t$, respectively, the payoff structure of $u_k^n$ without C-DEDU at time $t$ is

$$U_u^n(t) = bf_k^n(t) - sf_k^n(t) \tag{3}$$

According to the above analysis, CSP $c_k$ receives $sf_k^n(t)$ from every data user $u_k^n$ for providing cloud storage service. However, it also needs to pay the storage and maintenance cost $sc_k^n(t)$ for each data user $u_k^n$. Therefore, the payoff structure of $c_k$ without C-DEDU at time $t$ is

$$U_c^k(t) = \sum_n sf_k^n(t) - \sum_n sc_k^n(t) \tag{4}$$

### 4.2. Payoff structure in cloud storage with C-DEDU

When C-DEDU is applied in the cloud storage system, data users are divided into a group of data owners and a group of data holders.

Data owner $o_k^i$ is the first one to upload data $d_k^m$ at CSP $c_k$ at time $t$. It obtains the cloud storage benefits $bf_k^i(t)$ and pays the storage-service fee $sf_k^i(t)$ with some discount $\alpha_k^i(t)$. The discount is discounted on the storage-service fee, which is the incentive that a CSP provides to its users. It can be a unified discount or an individualized discount. The more data are deduplicated, the more discount is granted. Besides this, C-DEDU requires it some operation costs $OC_o^i(t)$ for performing deduplication and keeping online. In order to motivate data owners, CSP $c_k$ pays the request fee $RF_k^i(t)$ to $o_k^i$ for the successful access of data holders at time $t$. We conclude the payoff structure of data owner $o_k^i$ at time $t$ when C-DEDU is applied as

$$U_o^i(t) = bf_k^i(t) + RF_k^i(t) - \left(1-\alpha_k^i(t)\right) \times sf_k^i(t) - OC_o^i(t). \tag{5}$$

The data holder $h_k^j$ shares the same payoff structure with data users and obtains the discount $\alpha_k^j(t)$ on the storage-service fee. However, it suffers from the service-delay loss $l_j(t)$ when it cannot receive a quick reply from $o_k^i$ with the possibility $\varepsilon_i(t)$ at time $t$. Therefore, when C-DEDU is applied, the payoff structure of $h_k^j$ at time $t$ is

$$U_h^j(t) = bf_k^j(t) - \left(1-\alpha_k^j(t)\right) \times sf_k^j(t) - \varepsilon_i(t) \times l_j(t). \tag{6}$$

According to the above analysis, CSP $c_k$ receives discounted storage-service fees from all data owners and data holders and stores only the data of data owners. $OC_c^k(t)$ represents the total operation cost of $c_k$ for performing C-DEDU at time $t$. $c_k$ also needs to pay some request fees to the data owners as stated above. Therefore, the payoff structure of $c_k$ with C-DEDU at time $t$ is concluded as

$$U_c^k(t) = \sum_i \left(1-\alpha_k^i(t)\right) \times sf_k^i(t) - RF_k^i(t) - sc_k^i(t) + \sum_i \sum_{j,f(j)=i} \left(1-\alpha_k^j(t)\right) \times sf_k^j(t) - OC_c^k(t). \tag{7}$$

### 4.3. Utility functions

The success of C-DEDU relies on its acceptance by data owners, data holders and CSPs. However, C-DEDU requires data owners to be online all the time, which will bring them high operation costs. Moreover, the loss caused by service delay hinders the acceptance of time-sensitive data holders. Introducing C-DEDU greatly relieves the storage costs of CSPs. CSP can play as a decisive party to grant its savings to data owners and data holders by giving them discounts and induce them to accept C-DEDU. In this section, we will further analyze the detailed composition of every stakeholder based on the game-model assumptions in Section 3.3.

We first specify the strategy spaces of all players. For a data owner $o_k^i$, it can choose the possibility to be offline, namely the value of $\varepsilon_i(t)$. If it keeps online (i.e., $\varepsilon_i(t) = 0$), we say it takes honest actions. If $o_k^i$ has a possibility to be offline (i.e., $0 < \varepsilon_i(t) \leq 1$), we say $o_k^i$ is a strategic player. For a data holder $h_k^j$, it also has two strategies. If it follows the design of C-DEDU, we say it is honest. If it modifies data to avoid being detected as duplicated, then it is taking the strategic action.

In this subsection, we analyze the detailed utility functions of data owners and data holders who are both rational, based on our assumptions in Section 3.3. A rational player (data owner or data holder) takes action from its strategy space {honest, strategic} according to which one can bring it more profits.

#### 4.3.1. Utility function of data holder

The proliferation of Internet technologies brings a variety of emerging services, like cloud computing. These services greatly facilitate people's life and become essential and irreplaceable now. In Gao et al. (in press), the acceptance of cloud storage services has been demonstrated. Therefore, we assume all data users are delighted to store data in the cloud, then,

$$bf_k^\delta(t) - sf_k^\delta(t) > 0 \ (\delta = n, i, j) \tag{8}$$

is true for each data user. CSPs charge data users the storage-service fees according to the data size in a pay-per-use scenario.

According to the data assumption and user assumption, each data has the same data size and all data users obtain the same benefit from cloud storage. Hence, for every $\delta = n, i, j$, $bf_k^{\delta}(t) = bf_k$, $sf_k^{\delta}(t) = sf_k$ and $bf_k > sf_k$. As stated in Section 3.2, the absence of data owner causes service delay to data holders and this delay impacts the benefit from cloud storage. Therefore, we assume the service-delay loss $l_j(t)$ of data holder $h_k^j$ is related to the cloud storage benefits $bf_k^j(t)$ as $l_j(t) = \theta_j \times bf_k^j(t) = \theta_j \times bf_k$. The coefficient $\theta_j$ diverses with regards to $j$ since different data holders have different time sensitivities. Hence, we can detail (6) as

$$U_h^j(t) = \left(1 - \varepsilon_i \times \theta_j\right) \times bf_k - \left(1 - \alpha_k^j(t)\right) \times sf_k. \tag{9}$$

*4.3.2. Utility function of data owner*

A data owner is a privileged data user that can control data access and manage data. For a data owner $o_k^i$, $bf_k^i(t) - sf_k^i(t) > 0$ holds as analyzed in Section 4.3.1. Cloud storage service often works as a backup service. A data user may upload the same data several times and many data users may upload the same data to the cloud simultaneously or sequentially. As specified in Section 3.1, a data owner needs to keep online to verify the eligibility of all data holders and issue keys to the eligible ones timely. Whether the data owner exploits its own resources or hires others' resources to complete this series of operations, the cost cannot be ignored. It is reasonable to assume $OC_o^i(t)$ is related to the number of data holders with deduplication at time $t$. Let the function $f(j,t) = i$ represent that $o_k^i$ is the owner of $h_k^j$ at time $t$, then, $OC_o^i(t) = \sum_{j, f(j,t)=i} oc_o^i$ when $oc_o^i$ represents the unit operation cost of $o_k^i$. CSP $c_k$ pays the request fee $RF_k^i(t)$ to $o_k^i$, which is related to how many times $o_k^i$ has performed C-DEDU at time $t$. Then, $RF_k^i(t) = \sum_{j, f(j,t)=i} rf_k$, where $rf_k$ is the unit request fee. If $o_k^i$ takes strategic behaviors that to be offline with the possibility $\varepsilon_i$, $0 < \varepsilon_i \leq 1$, its operation cost and request fee are modified as $(1 - \varepsilon_i) \times OC_o^i(t)$ and $(1 - \varepsilon_i) \times RF_k^i(t)$, respectively. Hence, we can extend (5) as

$$U_o^i(t) = bf_k - \left(1 - \alpha_k^i(t)\right) \times sf_k + (1 - \varepsilon_i) \times \sum_{j, f(j,t)=i} \left(rf_k - oc_o^i\right). \tag{10}$$

Notably, $\alpha_k^i(t) = \alpha_k^j(t)$ when $f(j,t) = i$.

*4.3.3. Utility function of CSP*

When C-DEDU is not applied in $c_k$, $U_c^k(t) = \sum_n sf_k^n(t) - \sum_n sc_k^n(t) = \sum_n sf_k - sc_k$. With the cloud-storage acceptance assumption, $sf_k > sc_k$ is a common condition throughout the whole paper. When $c_k$ adopts C-DEDU at time $t$, let $n_i(t)$ denote the total number of holders of $o_k^i$' data at time $t$. The storage cost $sc_k^i(t)$ for storing this data is exactly the unit storage cost $sc_k$ since only one copy stores. The C-DEDU operation cost for this data is related to the number of data owner and data holders. Therefore, $OC_c^k(t) = oc_c^k \times (1 + n_i(t))$. The detailed utility function of CSP $c_k$ is concluded as

$$\begin{aligned} U_c^k(t) = &\sum_i \left(1 - \alpha_k^i(t)\right) \times sf_k - rf_k \times n_i(t) - sc_k \\ &+ \sum_i \sum_{j, f(j,t)=i} \left(1 - \alpha_k^j(t)\right) \times sf_k - oc_c^k \times (1 + n_i(t)). \end{aligned} \tag{11}$$

## 5. Discount-based incentive mechanism

There are two kinds of discount-based incentive mechanism for CSPs to choose. The first one is a CSP grants discounts to all of its subscribers undifferentiatedly based on its saved storage space. The second one is to set the discount value for individual data based on how many times this data has been deduplicated. For simplification, we call these two kinds of mechanism as unified discount and individualized discount, respectively.

In this section, we first present the requirements for an incentive mechanism to be feasible and analyze these two mechanisms based on the game theoretical interactions between a data owner and a data holder. We further propose a privacy-preservation incentive mechanism and provide some algorithms to instruct its practical implementation.

### 5.1. Requirements

A feasible discount-based incentive mechanism should satisfy the following requirements (Liang and Yan, 2019).

*Individual rationality*: The introduction of an incentive mechanism ensures the non-negative profits for all players (i.e., data owners and data holders).

*Incentive compatibility*: Deviating the scheme design will not bring a player more profits. Specifically, a data holder will not gain more profits by modifying its raw data to avoid being detected as duplicated. A data owner gains more profits when keeping online.

*Profitability*: For the other party (i.e., CSP) except the players, the incentive mechanism should ensure its non-negative profit.

Combining with our economic model, we give the formal definitions of these requirements as follows.

**Definition 1.** A discount-based incentive mechanism is individually rational if for $\forall \delta = o, j$, $U_\delta(t) \geq 0$.

**Definition 2.** A discount-based incentive mechanism is incentive compatible if it ensures any player with duplicated data can obtain more profits by accepting C-DEDU and following the scheme procedure honestly. Namely, for $\forall j$, $U_h^j(t) \geq U_u^j(t)$ and for $\forall i$, $\frac{\partial U_o^i(t)}{\partial \varepsilon_i} < 0$.

**Definition 3.** A discount-based incentive mechanism is profitable for a CSP when the mechanism ensures its non-negative profits, i.e., $U_c^k(t) > 0$.

### 5.2. Unified discount

When CSP $c_k$ applies the unified discount, it calculates the value of discount according to the total storage spaces saved. All subscribers of $c_k$ are granted with the same discount no matter what strategies taken. The advantages of the unified discount are twofold. First, it can be easily conducted and reduce the computation cost of CSPs. The second and significant advantage is a data holder cannot speculate the deduplication information of its data, like whether is deduplicated and the number of data holders; therefore, it is resistant to side channel attacks.

There is a data owner $o_k^i$ and a data holder $h_k^j$. They are rational players and have the same data to store at $c_k$. Table 2 shows the utility functions under the game between $o_k^i$ and $h_k^j$ when the unified discount is applied. The first row represents the strategies of $h_k^j$ while the first column shows the strategies of $o_k^i$. The rests in this table are their utility arraies that the first element is the utility of $o_k^i$ and the second one is the utility of $h_k^j$.

**Proposition 1.** *With unified discount-based incentive mechanism, a rational data holder will always choose to be strategic. The best strategy for a data owner is to be strategic when the data holder is strategic.*

**Proof.** We first find the best strategy for $h_k^j$. $h_k^j$ can obtain the same benefit $bf_k - \left(1 - \alpha_k^i(t)\right) \times sf_k$ when $o_k^i$ is honest, no matter it is honest or strategic. When $o_k^i$ is a strategic player, $\left(1 - \varepsilon_i \times \theta_j\right) \times bf_k - \left(1 - \alpha_k^i(t)\right) \times sf_k -$

**Table 2**
Utility function matrix with unified discount.

| | Honest | Strategic |
|---|---|---|
| Honest | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k + rf_k - oc_o^i$,<br>$bf_k - \left(1-\alpha_k^j(t)\right) \times sf_k$ | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k - oc_o^i$,<br>$bf_k - \left(1-\alpha_k^j(t)\right) \times sf_k$ |
| Strategic | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k$<br>$+(1-\varepsilon_i) \times \left(rf_k - oc_o^i\right)$,<br>$\left(1-\varepsilon_i \times \theta_j\right) \times bf_k - \left(1-\alpha_k^j(t)\right) \times sf_k$ | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k$<br>$-(1-\varepsilon_i) \times oc_o^i$,<br>$bf_k - \left(1-\alpha_k^j(t)\right) \times sf_k$ |

**Table 3**
Utility function matrix with individualized discount.

| | Honest | Strategic |
|---|---|---|
| Honest | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k + rf_k - oc_o^i$,<br>$bf_k - \left(1-\alpha_k^j(t)\right) \times sf_k$ | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k - oc_o^i$,<br>$bf_k - sf_k$ |
| Strategic | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k$<br>$+(1-\varepsilon_i) \times \left(rf_k - oc_o^i\right)$,<br>$\left(1-\varepsilon_i \times \theta_j\right) \times bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k$ | $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k$<br>$-(1-\varepsilon_i) \times oc_o^i$, $bf_k - sf_k(t)$ |

$\left(bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k\right) = -\varepsilon_i \times \theta_j \times bf_k < 0$, the best strategy for $h_k^j$ is to be strategic. Therefore, to be strategic is the dominated strategy for $h_k^j$.

When $h_k^j$ is strategic, $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k - oc_o^i - bf_k + \left(1-\alpha_k^i(t)\right) \times sf_k + (1-\varepsilon_i) \times oc_o^i = -\varepsilon_i \times oc_o^i \leq 0$, the best strategy for $o_k^i$ is to be strategic. □

From Proposition 1, (honest, honest) will never be the only Nash Equilibrium of this game since (strategic, strategic) is already one NE according to the above analysis. The same discount is granted to all users. A rational data holder will always modify its data to avoid being detected as duplicated. It not only obtains the discount but also reduces the risk of storage-service delay. Therefore, the data holder benefits from C-DEDU without making any contribution. We call this behavior as free-riding behavior, which reduces the enthusiasm of other participants to contribute and makes the system collapse eventually. Therefore, applying the unified discount is not practical in C-DEDU.

### 5.3. Individualized discount

When CSP $c_k$ applies the individualized discount, it calculates the discount in terms of individual data. When data holder $h_k^j$ modifies its duplicated data as a unique one, $c_k$ will not give a discount to it. There is a data owner $o_k^i$ and a data holder $h_k^j$. They are rational players and have the same data to store at $c_k$. In Table 3, the first row represents the strategies of $h_k^j$ while the first column shows the strategies of $o_k^i$. The first element of the rest cells is the utility of $o_k^i$ and the second one is the utility of $h_k^j$.

**Proposition 2.** *With the individualized discount-based incentive mechanism, (honest, honest) is the only Nash Equilibrium when the following equations are satisfied simultaneously:*

$$\alpha_k^j(t) > \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}, \tag{12}$$

$$rf_k > oc_o^i. \tag{13}$$

**Proof.** When $h_k^j$ is strategic, $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k - oc_o^i - bf_k + \left(1-\alpha_k^i(t)\right) \times sf_k + (1-\varepsilon_i) \times oc_o^i = -\varepsilon_i \times oc_o^i < 0$ always holds, the best strategy for $o_k^i$ is to be strategic. When $o_k^i$ is honest, $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k - (bf_k - sf_k) = \alpha_k^i(t) \times sf_k > 0$ always holds, the best strategy for $h_k^j$ is to be honest.

If we want (honest, honest) to be the NE, the best strategy must be honest when $h_k^j$ is honest. Namely, $bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k + rf_k - oc_o^i - bf_k + \left(1-\alpha_k^i(t)\right) \times sf_k - (1-\varepsilon_i) \times \left(rf_k - oc_o^i\right)$ should be positive. Then, $rf_k > oc_o^i$.

Furthermore, if we want (honest, honest) to be the unique NE, the best strategy for $h_k^j$ must be honest when $o_k^i$ is strategic. Namely $\left(1-\varepsilon_i \times \theta_j\right) \times bf_k - \left(1-\alpha_k^i(t)\right) \times sf_k - (bf_k - sf_k)$ should be a positive value. Then we get $\alpha_k^j(t) > \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$.

From the above, if (honest, honest) is the only NE, $rf_k > oc_o^i$ *and* $\alpha_k^j(t) > \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ should be satisfied simultaneously. □

Analogously, we give the following conclusions without detailed proofs.

1. When $\alpha_k^j(t) > \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ and $rf_k < oc_o^i$, there is no Nash Equilibrium.
2. When $\alpha_k^j(t) < \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ and $rf_k < oc_o^i$, (strategic, strategic) is the only Nash Equilibrium.
3. When $\alpha_k^j(t) < \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ and $rf_k > oc_o^i$, there is two Nash Equilibria: (honest, honest) and (strategic, strategic).

Individualized discount suppresses the free-riding behavior since a strategic data holder cannot obtain more benefits than acting honestly. However, the privacy of the deduplication information may be violated since a data user can infer the existence of data by checking whether it can obtain a discount when uploading this data. To ensure data privacy, we set a random threshold for the discount on each data. Only when the number of holders of this data excesses the threshold, can they get the discounts. We will show how to implement this design in Section 5.4.

### 5.4. Individualized discount with privacy-preservation

#### 5.4.1. Feasibility analysis

The individualized discount-based incentive mechanism should satisfy the requirements of individual rationality and incentive compatibility to data owners and data holders without decreasing the profit of CSPs. Taking the utility functions into these requirements, we obtain the following conclusions.

**Proposition 3.** *The individualized discount-based incentive mechanism is individually rational for the data holder $h_k^j$ when*

$$\alpha_k^j(t) \geq 1 - \frac{bf_k}{sf_k} + \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}. \tag{14}$$

**Proof.** Taking $\alpha_k^j(t) \geq 1 - \frac{bf_k}{sf_k} + \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ into $U_h^j(t) = \left(1 - \varepsilon_i \times \theta_j\right) \times bf_k - \left(1 - \alpha_k^j(t)\right) \times sf_k$, then

$$U_h^j(t) \geq \left(1 - \varepsilon_i \times \theta_j\right) \times bf_k - \left(1 - \left(1 - \frac{bf_k}{sf_k} + \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}\right)\right) \times sf_k$$

$$= \left(1 - \varepsilon_i \times \theta_j\right) \times bf_k - \left(\frac{bf_k}{sf_k} - \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}\right) \times sf_k$$

$$= \left(1 - \varepsilon_i \times \theta_j\right) \times bf_k - bf_k - \varepsilon_i(t) \times \theta_j \times bf_k = 0.$$

Therefore, the individual rationality to $h_k^j$ is ensured since $U_h^j(t) \geq 0$. □

**Proposition 4.** *The individualized discount-based incentive mechanism is incentive compatible for the data holder $h_k^j$ when*

$$\alpha_k^j(t) \geq \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}. \tag{15}$$

**Proof.** To extend $U_h^j(t) - U_u^j(t)$, we have $bf_k^j(t) - \left(1 - \alpha_k^j(t)\right) \times sf_k^j(t) - \varepsilon_i \times l_j(t) - \left(bf_k^j(t) - sf_k^j(t)\right) = \alpha_k^j(t) \times sf_k - \varepsilon_i \times \theta_j \times bf_k$. We can easily obtain $U_h^j(t) - U_u^j(t) \geq 0$ when taking $\alpha_k^j(t) \geq \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ into the above equation. According to Definition 2, $U_h^j(t) \geq U_u^j(t)$ means incentive compatibility to the data holder. Therefore, we have proved Proposition 4. □

**Corollary 1.** *When a discount-based incentive mechanism satisfies $\alpha_k^j(t) \geq \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$, then this mechanism is individual rational and incentive compatible to the data holder $h_k^j$.*

To ensure the individual rationality and incentive compatibility simultaneously, $\alpha_k^j(t)$ should be no less than the bigger one between $1 - \frac{bf_k}{sf_k} + \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$ and $\frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$. Since $bf_k > sf_k$, $\frac{bf_k}{sf_k} > 1$. Then $1 - \frac{bf_k}{sf_k} + \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k} < \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$. Therefore, $\alpha_k^j(t) \geq \frac{\varepsilon_i \times \theta_j \times bf_k}{sf_k}$.

**Proposition 5.** *The individualized discount-based incentive mechanism is individually rational for the data owner $o_k^i$ when*

$$\alpha_k^i(t) \geq 1 - \frac{bf_k}{sf_k} - \frac{\varepsilon_i \times \sum_{j.f(j.t)=i} \left(rf_k(t) - oc_o^i\right)}{sf_k}. \tag{16}$$

Proposition 5 is easy to prove by taking (16) into $U_o^i(t) = bf_k - \left(1 - \alpha_k^i(t)\right) \times sf_k + (1 - \varepsilon_i) \times \sum_{j.f(j.t)=i} \left(rf_k(t) - oc_o^i\right)$ and evaluating if the result is non-negative.

**Proposition 6.** *The individualized discount-based incentive mechanism is incentive compatible for the data owner $o_k^i$ when*

$$rf_k > oc_o^i. \tag{17}$$

**Proof.** In our economic model, $bf_k$ and $oc_o^i$ are irrelevant to $\varepsilon_i$ and CSPs do not determine $\alpha_k^i(t)$, $sf_k$ and $rf_k$ according to the $\varepsilon_i$, therefore,

$$\frac{\partial U_o^i(t)}{\partial \varepsilon_i} = \sum_{j.f(j.t)=i} \left(oc_o^i(t) - rf_k\right), \tag{18}$$

which is negative when $rf_k > oc_o^i$. According to Definition 2, the individualized discount-based incentive mechanism is incentive compatible for $o_k^i$ when $rf_k > oc_o^i$. □

**Proposition 7.** *The individualized discount-based incentive mechanism is profitable for the CSP $c_k$ when*

$$\alpha_k^i(t) \leq 1 - \frac{sc_k + rf_k \times n_i(t) + oc_c^k \times n_i(t)}{sf_k \times (1 + n_i(t))}. \tag{19}$$

**Proof.** For data owner $o_k^i$, there are $n_i(t)$ data holders $h_k^j$, $j = \{1, 2, \ldots, n_i(t)\}$ that want to upload the same data as $o_k^i$. The benefits $c_k$ obtained from these users is

$$\left(1 - \alpha_k^i(t)\right) \times sf_k \times (1 + n_i(t)) - sc_k - rf_k \times n_i(t) - oc_c^k \times n_i(t). \tag{20}$$

The non-negative of (20) is guaranteed when

$$\alpha_k^i(t) \leq 1 - \frac{sc_k + rf_k \times n_i(t) + oc_c^k \times n_i(t)}{sf_k \times (1 + n_i(t))}.$$

□

*5.4.2. Algorithms*

In this subsection, we introduce the Parameter-Setting Algorithm (i.e., Algorithm 1) and Discount-Granting Algorithm (i.e., Algorithm 2) to show how CSPs implement the individualized discount-based incentive mechanism with privacy-preservation. The Strategy-Choosing Algorithms (i.e., Algorithm 3 and Algorithm 4) illustrate how rational data owners and data holders to choose their best strategies according to their known parameters.

We suggest CSP $c_k$ marking its duplication check result on $h_k^j$ at time $t$ as $x_i^j(t)$ and recording the deduplication report on $h_k^j$ at time $t$ from $o_k^i$ as $y_j^i(t)$. $x_i^j(t) = 1$ indicates $h_k^j$'s data has been stored by $c_k$ already and $x_k^j(t) = 0$ marks its data as unique. $y_j^i(t) = 1$ indicates CSP $c_k$ has received the deduplication-successful message from $o_k^i$ for $h_k^j$. Otherwise, $y_j^i(t)$ is marked as 0 and $h_k^j$ suffers from service delay. $c_k$ only pays $rf_k$ to $o_k^i$ when $y_j^i(t) = 1$.

Fig. 2 is the detailed parameter-setting algorithm, which is proposed based on (8) and (17) to ensure the existence of a cloud storage system and encourage data owners to accept C-DEDU, respectively. On the input of the state of all data holders $y_j^i(t-1)$, and $sc_k$, $oc_o^i$, CSP $c_k$ can calculate its $sf_k, rf_k$, and the discount $\alpha_k^i(t)$ for the next time by executing Algorithm 1.

Fig. 3 illustrates the discount-granting algorithm, which can be described as follows. If the data of a holder is detected as duplicated (i.e., $x_i^j(t) = 1$), CSP $c_k$ grants the discount calculating from Algorithm 1 to it. CSP $c_k$ will not pay the request fee to $o_k^i$ if it has not returned the deduplication-successful message of $h_k^j$. In detail, CSP $c_k$ only pays the request fee $rf_k$ to $o_k^i$ when $y_j^i(t) = 1$. According to our scheme assumption, no fake deduplication-successful message exists in this system. Therefore, $o_k^i$ will not gain any illegal $rf_k$.

Only when the incentive provided by the CSP satisfied the individual rationality and incentive compatibility of data owners and data holders, will they choose to be honest. Therefore, we propose Algorithm 3 (shown in Fig. 4) and Algorithm 4 (shown in Fig. 5) based on Propositions 4 and 7, respectively.

Fig. 4 illustrates what strategy a rational data owner will choose with the parameters $bf_k, rf_k, sf_k, oc_o^i$ $\alpha_k^i(t)$, and its private information $\varepsilon_i$.

It is difficult for a data holder to gain the value of $\varepsilon_i$, since $\varepsilon_i$ is the private information. However, they can infer the offline possibility $\varepsilon$ of the whole system through social networks. Therefore, we can take $\varepsilon$ as the public offline probability of the storage system, which is related to all $\varepsilon_i$.

## 6. Evaluation

We conducted a set of experiments to analyze the effectiveness of our incentive mechanism in promoting the acceptance of the C-DEDU scheme by all system players. In this section, we also discuss how to make the incentive mechanism compatible with existing deduplication schemes and its scalability and robustness with regard to modification attacks.

**Algorithm 1** Parameter-Setting Algorithm

**Input:** $y_j^i(t-1)$, $sc_k$, $oc_o^i$

**Output:** $sf_k$, $rf_k$, $\alpha_k^i(t)$

$c_k$ sets $sf_k$ according to $sf_k > sc_k$;
$c_k$ sets $rf_k$ according to $rf_k > \max_i\{oc_o^i\}$;
**Foreach** $o_k^i$
  $c_k$ sets a random threshold $\eta_i \geq 0$;
  **If** $\sum_{j,f(j,t-1)=i} y_j^i(t-1) \leq \eta_i$, **then**
    $\alpha_k^i(t) = 0$;
  **Else**
    $c_k$ calculates the discount based on an individualized discount function;
  **End If**
**End Foreach**

**Fig. 2.** The algorithm to set CSP parameters.

**Algorithm 2** Discount-Granting Algorithm

**Initialization**: $rf_k$, $\alpha_k^i(t)$;
**Repeat**
  **Foreach** $o_k^i$
    **Foreach** $h_k^j$
      **If** $f(j,t) = i$, **then**
        $c_k$ sets $x_i^j(t) = 1$;
        $c_k$ gives discount $\alpha_k^i(t)$ to $h_k^j$;
        **If** $o_k^i$ reports the deduplication-successful message on $h_k^j$ to $c_k$, **then**
          $c_k$ sets $y_j^i(t) = 1$;
          $c_k$ pays $rf_k$ to $o_k^i$;
        **Else**
          $c_k$ sets $y_j^i(t) = 0$;
        **End If**
      **End If**
    **End Foreach**
    $c_k$ gives discount $\alpha_k^i(t)$ to $o_k^i$;
  $t \leftarrow t + 1$;
**End Repeat**

**Fig. 3.** The algorithm to grant discounts for a CSP.

**Algorithm 3** Strategy-Choosing Algorithm for Data Owner $o_k^i$

**Initialization**: $bf_k, rf_k, sf_k, oc_o^i, \alpha_k^i(t)$
$o_k^i$ decides $\varepsilon_i$;
**If** $\alpha_k^i(t) \geq 1 - \frac{bf_k}{sf_k} - \frac{\varepsilon_i \times \sum_{j,f(j,t)=i} rf_k - oc_0^i}{sf_k}$, **then**
  $o_k^i$ chooses to be honest;
**Else**
  $o_k^i$ chooses to be strategic;
**End If**

**Fig. 4.** The algorithm to compute the optimal strategy for a rational data owner.

**Algorithm 4** Strategy-Choosing Algorithm for Data Holder $h_k^j$

**Initialization**: $bf_k, sf_k, \varepsilon, \alpha_k^j(t)$
$h_k^j$ decides $\theta_j$;
**If** $\alpha_k^j(t) \geq \frac{\varepsilon \times \theta_j \times bf_k}{sf_k}$, **then**
  $h_k^j$ chooses to be honest;
**Else**
  $h_k^j$ chooses to be strategic;
**End If**

**Fig. 5.** The algorithm to compute the optimal strategy for a rational data holder.

### 6.1. Experimental settings

The data used in our experiments were collected from section *contrib* (b23, 0000) in the Debian Popularity Contest (b24, 0000), which includes the usage of Debian packages. We recorded the number of installations for each package to simulate the number of data users. We took a snapshot on the 19th June 2018. The total number of package installations is 309052 and the total number of the package is 434. Therefore, in our dataset, the number of data files is 434 and the total number of data users is 309052. These dataset properties are very similar to those of cloud data storage. The users who access and download the same package can be treated as the data owner (first downloader) and its data holders (later downloaders). Thus, it is feasible to use this dataset to simulate cloud data storage with duplicated data and perform our experimental tests.

There is one CSP and 309052 data users to store 434 unique data in each experiment. At the beginning of each experiment, the CSP publishes its unit storage fee and the unit request fee to the public and calculates its discount for each data user. The data users

**Table 4**
Parameter settings.

| Parameter | Value | Parameter | Value | Parameter | Value |
|---|---|---|---|---|---|
| $bf_k$ | 1.5 | $sf_k$ | 1 | $rf_k$ | 0.1 |
| $sc_k$ | 0.8 | $oc_c^k$ | 0.05 | $oc_o^i$ | 0.05 |
| $\alpha_{min}$ | 0.001 | $\alpha_{max}$ | 0.85 | $k$ | 2 |
| $\varepsilon_i$ | [0,0.5] | $\eta_i$ | [0,5] | $\theta_j$ | [0,0.5] |

(i.e., data owners and data holders) independently choose an optimal strategy based on this public information and their private information. For a data owner, it chooses to keep online (i.e., honest) or to be offline with a probability (i.e., strategic). For a data holder, it chooses to follow the design of C-DEDU (i.e., honest) or modify its data to avoid being detected as duplicated for the sake of privacy (i.e., strategic). When they all made decisions, we say a time generation has passed. After each time generation, the CSP renews the discounts, and then all data users update their strategies accordingly. When all entities have no incentive to alter their strategies from the next time generation on, our game experiment reaches the NE state.

There are no specialized economic models on the discount function. In our experiments, we modelled the discount as a function of the deduplication percentage. Let $\alpha_{min}$ and $\alpha_{max}$ be the minimum and maximum discounts that $c_k$ can give. we set $\alpha_{min} = 0.001$ to initiate our storage system with C-DEDU. Due to the profitability of CSP, $\alpha_k^i(t) \leq 1 - \frac{sc_k + rf_k \times n_i(t) + oc_c^k \times n_i(t)}{sf_k \times (1+n_i(t))} = 1 - \frac{sc_k}{sf_k \times (1+n_i(t))} - \frac{rf_k + oc_c^k}{sf_k} \times \frac{n_i(t)}{1+n_i(t)}$. Therefore, $\alpha_{max} < \min\{\alpha_k^i(t)\}$. We set $\alpha_{max} = 1 - \frac{rf_k + oc_c^k}{sf_k}$ in our experiment. The discount function applied in our experiments is

$$\alpha_k^i(t) = \alpha_{min} + \frac{\rho_k^i(t) \times (\alpha_{max} - \alpha_{min})}{k}, \tag{21}$$

where $k \geq 1$.

If the individualized discount is applied, $\alpha_k^i(t) = \alpha_k^j(t) = \rho_k^i(t)$, where $f(j,t) = i$.We calculated the utilities of all stakeholders (i.e. data holders, data owners and CSPs) and recorded the deduplication percentage of the CSP for each experiment. We summarized the system parameter settings in Table 4 as follows. The value of $bf_k$, $sf_k$, $sc_k$, $rf_k$ and $oc_o^i$ were set according to $sc_k < sf_k < bf_k$ and $oc_o^i < rf_k$ for ensuring the long-term operation of the CSP. $oc_c^k$ was set to make sure $sc_k - rf_k - oc_c^k \geq 0$. We randomly chose a value between 0 and 0.5 as the possibility for a data owner to be offline; therefore, $\varepsilon_i \in [0, 0.5]$. We set the value of $\eta_i$ randomly from 0 to 5 for each data. Since time-sensitivity is different for different data holders, we regulated that $\theta_j$ obeys a uniform distribution between 0 and 0.5 (i.e., $\theta_j \sim U(0, 0.5)$). The public offline probability $\varepsilon$ of the system was modeled as the mean of all $\varepsilon_i$.

We conducted three experiments. The CSP is without C-DEDU in Experiment 1, which is worked as a benchmark. In Experiment 2, the CSP employs our proposed incentive mechanism with C-DEDU. We specify the evaluation indexes in the first two experiments as follows:

- The average utility of data holders at each time generation;
- The average utility of data owner at each time generation;
- The utility of CSP at each time generation;
- The deduplication percentage of CSP at each time generation.

To evaluate the influence of some system parameters ($\varepsilon_i$, $\eta_i$, $\theta_j$, to be specific), we further conducted Experiment 3. The CSP and data users are the same as those in Experiment 2. If denote $\varepsilon_i \in [0, a]$, $\eta_i \in [0, b]$ and $\theta_j \sim U(0, c)$, we set $a = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$, $b = 1, 2, 3, 4, 5, 6, 7, 8, 9, 10$ and $c = 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1$ in this experiment. The evaluation indexes for Experiment 3 are:

- The average utility of data holders at NE;
- The average utility of data owner at NE;
- The utility of CSP at NE;
- The deduplication percentage of CSP at NE;
- The time to reach NE.

### 6.2. Experimental results

There are 309052 data users requested to upload data to the cloud. If they are the users in a CSP without C-DEDU (i.e., Experiment 1), the CSP stores their data directly. If they are the users in a CSP with C-DEDU and the CSP adopts our incentive mechanism (i.e., Experiment 2), the data holders choose to duplicate its data or not according to (15) and the data owners choose whether to keep online based on (19). After all users taking actions, one time generation passed and the CSP adjusts the discount for the next time generation. We plotted the results of Experiment 1 and Experiment 2 together in Fig. 6 for easy comparison. The solid and dotted lines show the results in Experiment 1 and Experiment 2, respectively.

Fig. 6 (a)to –(c) show that the utilities of all system stakeholders (i.e., data owners, data holders and CSP) in the C-DEDU are increasing with an S-shape as the time generation goes by. All these utilities reach the maximum value eventually and stay stable from then on. Therefore, there is an NE state of our game model. This state is also a social optimal state that our stakeholders obtain the largest benefits. Fig. 6(d) illustrates the deduplication percentage variation of the cloud storage system. Under our parameter settings, the proposed incentive mechanism motivates most of the data users to adopt C-DEDU due to obvious profits and the system saves more than 90% (92.46% to be specific) storage spaces at last. The reason for the rest users that do not select deduplication is the incentive provided by our mechanism is still not enough to compensate the potential loss of some extreme time-sensitive users under our experimental parameter settings. If we adjust $a$ and/or $b$ to a smaller value, the deduplication percentage in the equilibrium state will be higher than 92.46%. Taking the experimental results in Experiment 1 as the benchmark, all the dotted lines are above the solid ones. Therefore, the acceptance of C-DEDU is guaranteed by our incentive mechanism.

Experiment 3 was to evaluate the effect of some system parameters: the offline possibility $\varepsilon_i$ of a data owner, the time sensitivity $\theta_j$ of a data holder, and the holder number threshold $\eta_i$ set by the CSP. The detailed evaluation method was to adjust the upper limits of the distribution functions to regulate $\varepsilon_i$, $\eta_i$, and $\theta_j$, namely the value of $a$, $b$, and $c$, and re-execute the procedure in Experiment 2.

Fig. 7 shows the influence of $\varepsilon_i$ by setting $a$ from 0.1 to 1 with $b = 5$ and $c = 0.5$. At the NE state, the utility of the CSP, data owners and data holders decrease with the increase of $a$. Only the time to reach this state increases while $a$ is increasing. Fig. 7(d) represents the deduplication percentage of the system with different $a$. Furthermore, even when $a = 1$ (i.e., $\varepsilon_i \in [0, 1]$), the deduplication percentage can still be above 85% (87.4739% precisely), which means the C-DEDU is accepted by most users.

The influence of $\eta_i$ was tested by setting $b$ from 1 to 10 with $a = 0.5$ and $c = 0.5$. As illustrated in Fig. 8, the utilities of all stakeholders and the deduplication percentage at the NE state share the same decrease trend with the increase of $b$. The reason for this downtrend is that there are some data belongs to less than $b$ users that cannot obtain discounts thus be hesitated to accept C-DEDU. The hesitation further influences the deduplication percentage, which is directly linked to the discount. Nevertheless, the influence is not so serious since the deduplication percentage

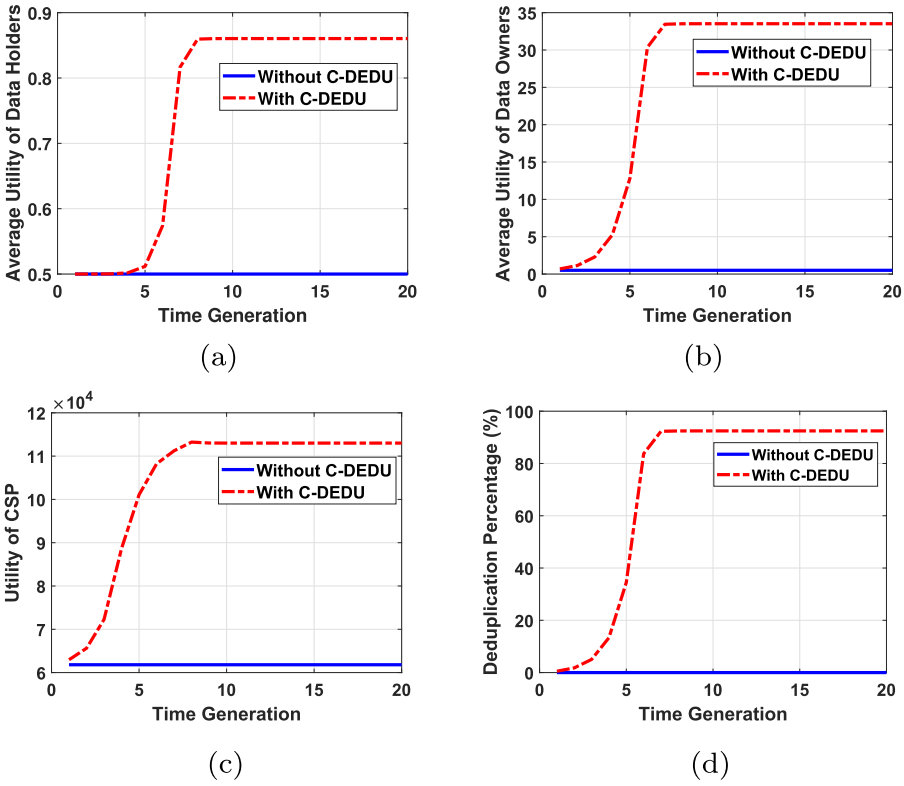


**Fig. 6.** The results of Experiment 1 and Experiment 2: (a) the average utility of data owners; (b) the average utility of data holders; (c) the utility of the CSP; (d) the system deduplication percentage in different time generations.

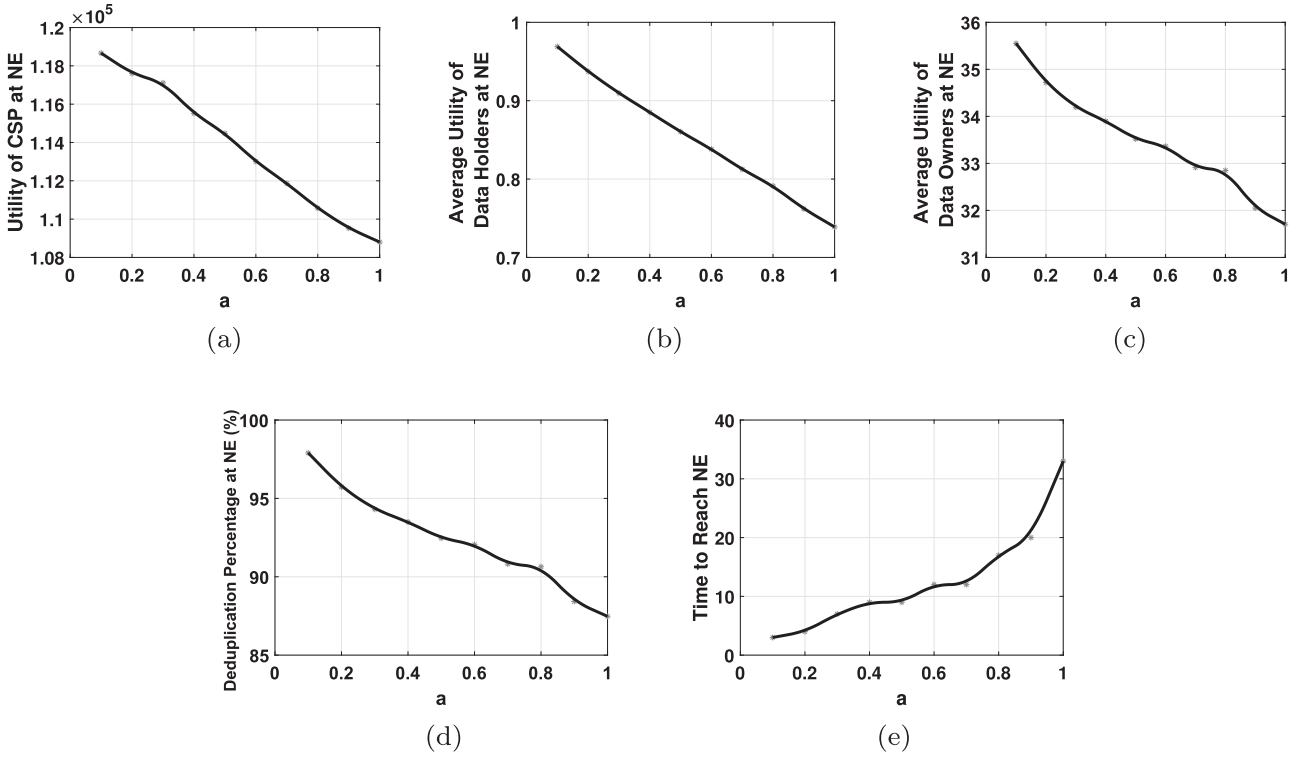


**Fig. 7.** The effect of $a$ on: (a) the utility of CSP at NE; (b) the average utility of data owners at NE; (c) the average utility of data holders at NE; (d) the deduplication percentage of a storage system at NE; (e) time to reach NE.

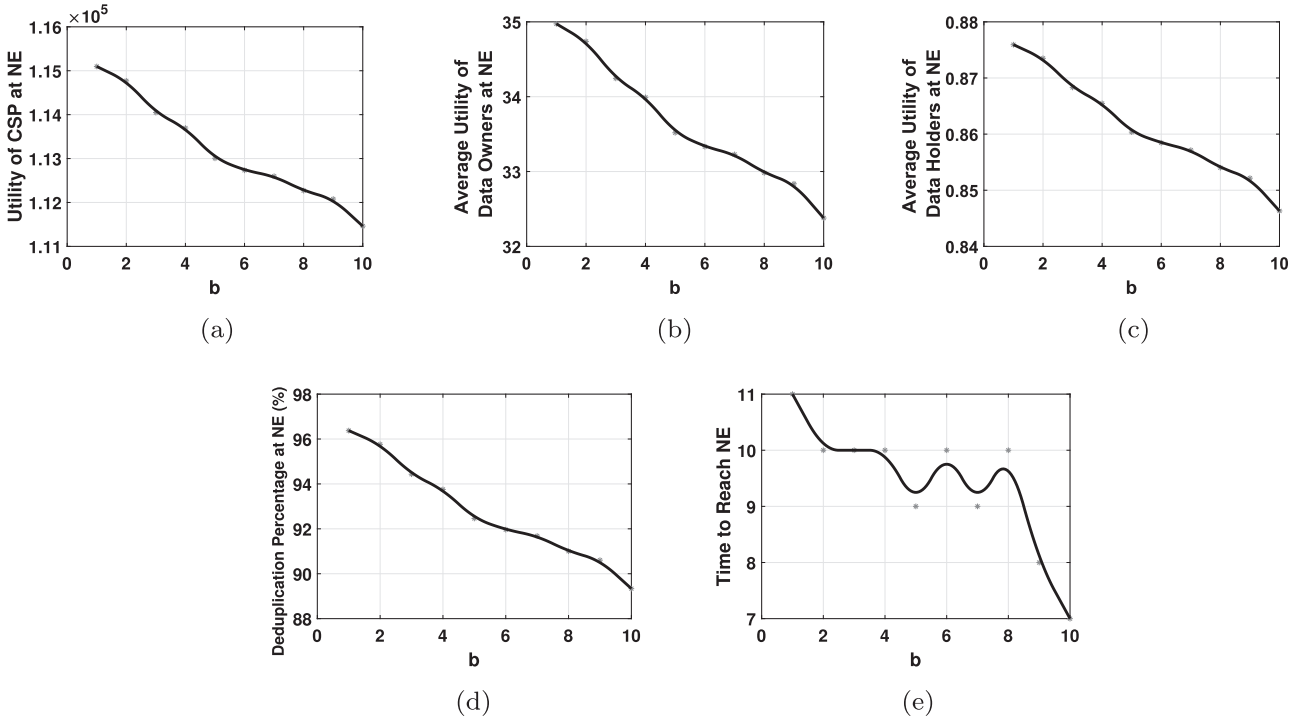


**Fig. 8.** The effect of $b$ on: (a) the utility of CSP at NE; (b) the average utility of data owners at NE; (c) the average utility of data holders at NE; (d) the system deduplication percentage at NE; (e) time to reach NE.

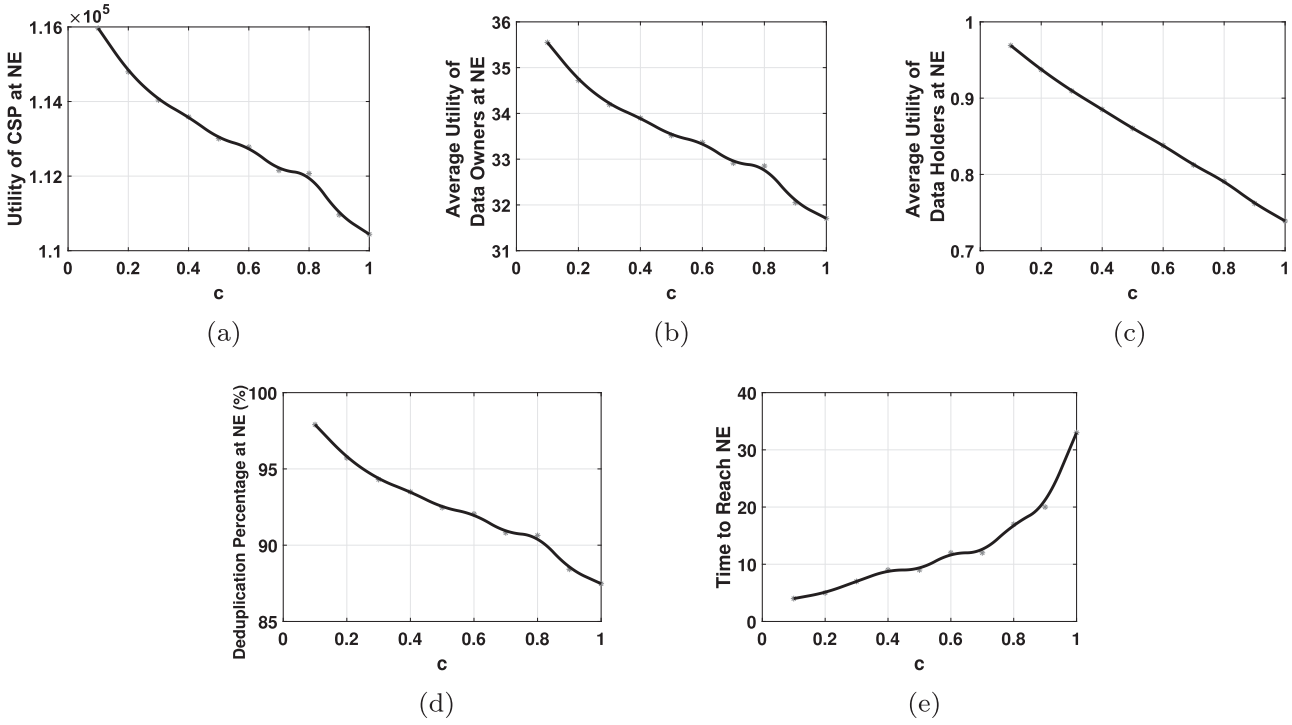


**Fig. 9.** The effect of $c$ on: (a) the utility of CSP at NE; (b) the average utility of data owners at NE; (c) the average utility of data holders at NE; (d) the system deduplication percentage at NE; (e) time to reach NE.

still remains at a high level (above 89% in Fig. 8)(d). The time to reach NE fall between the 9th generation and the 10th generation mostly.

To evaluate the influence of $\theta_j$, we fixed $a = 0.5$ and $b = 5$ while varying $c$ from 0.1 to 1 and plotted the experimental results in Fig. 9. All the evaluation indexes for Experiment 3, except the time to reach NE, decline with the increase of $c$. Notably, even though the indexes are decreased, they are still higher than those in Experiment 1.

In a nutshell, taking the results in Experiment 1 as the benchmark, we evaluated the acceptance of C-DEDU with our proposed incentive mechanism in Experiment 2. The increase of deduplication percentage shows C-DEDU is gradually accepted. Even though the percentage would decrease when adjusting some parameters,

the influence is not significant. Therefore, our incentive mechanism guarantees the widely adoption of C-DEDU.

### 6.3. Further discussions

In this section, we further discuss the compatibility of our incentive mechanism when being applied into existing cloud data deduplication schemes and its scalability and robustness when being triggered with modification attacks, where the cloud users attempt to modify their data stored in the cloud.

Herein, we first illustrate how our incentive mechanism works based on the procedure of the deduplication scheme in Yan et al. (2016b). The parameters, $x_i^j(t)$ and $y_j^i(t)$, needed to be collected in our incentive mechanism are compatible with the deduplication scheme in Yan et al. (2016b). Specifically, the value of $x_i^j(t)$ indicates the result of duplication check, which is the inevitable process in deduplication. Furthermore, once a data owner performs deduplication successfully, it will send the deduplication-successful message to the CSP so that it can obtain its request fee as compensation. This message is recorded as $y_j^i(t)$ in our mechanism.

To apply our incentive mechanism in ClearBox (Armknecht et al., 2015), which is also a client-controlled deduplication scheme, the value of $x_i^j(t)$ is easy to determine since the owner will check if any other clients have already uploaded the same data. When the data has already been stored and the user passes the possession verification, the owner will append this user to the data structure. This operation can be approximately regarded as sending the deduplication-successful messages in Yan et al. (2016b), thus the value of $y_j^i(t)$ is also decided.

Heen et al. (2012) proposed a client-controlled deduplication scheme that can resist the side-channel attack. In this scheme, a new gateway server is introduced as the proxy of a data owner to perform deduplication. The gateway client checks the existence of a piece of data at the storage space of the gateway server (that can be considered as the CSP), thus duplication check happens and $x_i^j(t)$ is determined. Furthermore, the gateway server handles the other users' upload requests of the same data, therefore, it can calculate $y_j^i(t)$ as well.

In a nutshell, our incentive mechanism is compatible with not only the deduplication scheme in Yan et al. (2016b) but also ClearBox (Liu et al., 2015) and the one that resists to side-channel attacks (Heen et al., 2012). Thanks to the compatibility, introducing this incentive mechanism to existing deduplication schemes burdens no extra communication costs. The computational complexity is only linearly related to the number of data users.

Our incentive mechanism can also suppress the dishonest behaviors of data owners. If an owner conducts a modification attack then its data will be regarded as unique, it will not gain the compensation from deduplication. Moreover, the owner can hardly obtain the discount of storage-service fee when applying our individualized discount-based incentive. Without concrete proof, we are safe to conclude that our incentive mechanism is scalable to the application scenario with the modification attack and is robust to this attack.

## 7. Conclusion

In this paper, we detailed an economic model for cloud storage systems with C-DEDU. A game theoretical approach is employed to analyze the feasibilities of two discount-based incentive mechanisms: unified discount and individualized discount. The unified discount ensures data privacy but introduces free-riding behaviors, which is difficult to eliminate without changing the design of C-DEDU. The individualized discount can suppress free-riding behaviors in some cases; however, data holders can infer some private information from the discount value. To address this privacy issue, we further proposed an adapted privacy-preserving individualized discount-based incentive mechanism with the concern of individual rationality, incentive compatibility and profitability. Corresponding algorithms were proposed as well to show the practical implementation of our mechanisms. Comprehensive experiments based on a real dataset further illustrated the effectiveness of our incentive mechanisms for the final acceptance of C-DEDU.

## Declaration of Competing Interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgments

The work is supported in part by the National Natural Science Foundation of China under Grants 61672410 and 61802293, the National Postdoctoral Program for Innovative Talents under grant BX20180238, the Project funded by China Postdoctoral Science Foundation under grant 2018M633461, the Academy of Finland under Grants 308087 and 314203, the Key Lab of Information Network Security, Ministry of Public Security under grant No. C18614, the open grant of the Tactical Data Link Lab of the 20th Research Institute of China Electronics Technology Group Corporation, P.R. China under grant CLDL-20182119, the Shaanxi Innovation Team project under grant 2018TD-007, and the 111 project under grant B16037.

## References

http://popcon.debian.org/contrib/index.html.
http://popcon.debian.org.
Armknecht, F., Bohli, J., Karame, G.O., Youssef, F., 2015. "transparent data deduplication in the cloud,". In: CCS'15, pp. 886–900. New York, USA.
Bolosky, W.J., Corbin, S., Goebel, D., Douceur, J.R., 2000. "single instance storage in windows 2000,". In: Proceedings of the 4th USENIX Windows Systems Symposium, pp. 13–24. Seattle, WA.
Chu, C.K., Chow, S.S., Tzeng, W.G., Zhou, J., Deng, R.H., 2014. "key-aggregate cryptosystem for scalable data sharing in cloud storage,". IEEE Trans. Parallel Distrib. Syst. 25 (2), 468–477.
Do, C.Y., Tran, N.H., Hong, C., Kamhoua, C.A., Kwiat, K.A., Blasch, E., Ren, S., Pissinou, N., Iyengar, S.S., 2017. "game theory for cyber security and privacy,". ACM Computing Surveys (CSUR) 50 (2). Article no. 30.
Gao L., Yan Z., Yang L.Y. "game theoretical analysis on acceptance of a cloud data access control system based on reputation,". In: IEEE Transactions on Cloud Computing, in press.
Harnik, D., Pinkas, B., Shulman-Peleg, A., 2010. "side channels in cloud services: deduplication in cloud storage,". IEEE S&P 8 (6), 40–47.
Heen, O., Neumann, C., Montalvo, L., Defrance, S., 2012. "improving the resistance to side-channel attacks on cloud storage services,". In: NTMS'12, pp. 1–5.
Hwang, K., Kulkareni, S., Hu, Y., 2009. "cloud security with virtualized defense and reputation-based trust management,". In: IEEE DASC'09, pp. 717–722. Chendu, China.
Hwang, K., Li, D., 2010. "trusted cloud computing with secure resources and data coloring,". IEEE Internet Comput 14 (5), 14–22.
Li, J., Chen, X., Li, M., Li, J., Lee, P.P., Lou, W., 2014. "secure deduplication with efficient and reliable convergent key management,". IEEE Trans Parallel Distrib Syst 25 (6), 1615–1625.
Li, J., Li, Y.K., Chen, X., Lee, P.P., Lou, W., 2015. "a hybrid cloud approach for secure authorized deduplication,". IEEE Trans. Parallel Distrib. Syst. 26 (5), 1206–1216.
Liang, X., Yan, Z., 2019. "a survey on game theoretical methods in human-machine networks,". Future Gener. Comput. Syst. 92, 674–693.
Liang, X., Yan, Z., Chen, X., Yang, L.T., Lou, W., Hou, Y.T., 2019. "game theoretical analysis on encrypted cloud data deduplication,". IEEE Trans. Ind. Inform. 15 (10), 5778–5789.
Liu, J., Asokan, N., Pinkas, B., 2015. "secure deduplication of encrypted data without additional independent servers,". In: CCS'15, pp. 874–885. New York, NY.
Manshaei, M.H., Zhu, Q., Alpcan, T., Bacşar, T., Hubaux, J.P., 2013. "game theory meets network security and privacy,". ACM Computing Surveys (CSUR) 45 (3). Article no. 25.
Meyer, D.T., Bolosky, W.J., 2012. "a study of practical deduplication,". ACM Trans. Storage (TOS) 7 (4), 14.
Miao, M., Jiang, T., You, I., 2015. "payment-based incentive mechanism for secure cloud deduplication,". Int. J. Inf. Manage. 35 (3), 379–386.

Nan, G., Mao, Z., Li, M., Zhang, Y., Gjessing, S., Wang, H., Guizani, M., 2014. "distributed resource allocation in cloud-based wireless multimedia social networks,". IEEE Network 28 (4), 74–80.
Niyato, D., Vasilakos, A.V., Kun, Z., 2011. "resource and revenue sharing with coalition formation of cloud providers: Game theoretic approach,". In: IEEE CCGrid'11, pp. 215–224. Newport Beach, CA.
Palmieri, F., Buonanno, L., Venticinque, S., Aversa, R., Martino, B.D., 2013. "a distributed scheduling framework based on selfish autonomous agents for federated cloud environments,". Future Gener. Comput. Syst. 29 (6), 1461–1472.
Pooranian, Z., Chen, K., Yu, C., Conti, M., 2018. "RARE: defeating side channels based on data-deduplication in cloud storage,". In: IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), pp. 444–449.
Ruj, S., Stojmenovicn, M., Nayak, A., 2014. "decentralized access control with anonymous authentication of data stored in clouds,". IEEE Trans. Parallel Distrib. Syst. 25 (2), 384–394.
Shen, Y., Yan, Z., Kantola, R., 2014. "analysis on the acceptance of global trust management for unwanted traffic control based on game theory,". Comput. Secur. 47, 3–25.
Sun, W., Zhang, N., Lou, W., Hou, Y.T., 2018. "tapping the potential: Secure chunk-based deduplication of encrypted data for cloud backup,". IEEE Conference on Communications and Network Security (CNS) 1–9. 2018.
Wan, Z., Liu, J.E., Deng, R.H., 2012. "HASBE: a hierarchical attribute-based solution for flexible and scalable access control in cloud computing,". IEEE Trans. Inf. Forensics Secur. 7 (2), 743–754.
Wang, C., Wang, Q., Ren, K., Lou, W., 2010a. "privacy-preserving public auditing for data storage security in cloud computing,". In: Proc. IEEE INFOCOM, pp. 1–9. San Diego, CA.
Wang, G., Liu, Q., Wu, J., 2010b. "hierarchical attribute-based encryption for fine–grained access control in cloud storage services,". In: CCS'10, pp. 735–737. Chicago, Illinois.
Wang, Q., Wang, C., Ren, K., Lou, W., Li, J., 2011. "enabling public auditability and data dynamics for storage security in cloud computing,". IEEE Trans. Parallel Distrib. Syst. 22 (5), 847–859.
Wang, Y., Guo, C., Li, T., Xu, Q., 2015. "secure two-party computation in social cloud based on reputation,". In: IEEE WAINA'15, pp. 242–245. Gwangiu, South Korea.
Wei, L., Zhu, H., Cao, Z., Dong, X., Jia, W., Chen, Y., Vasilakos, A.V., 2014. "security and privacy for storage and computation in cloud computing,". Inf. Sci. 258, 371–386.
Wong, W., Lee, W., Wei, H., 2014. "base on game theory model to improve trust access control in cloud file-sharing system,". In: IIHMSP'14, pp. 702–705. Kitakyushu, Japan.
Xia, W., Jiang, H., Feng, D., Douglis, F., Shilane, P., Hua, Y., Fu, M., Zhang, Y., Zhou, Y., 2016. "a comprehensive study of the past, present, and future of data deduplication,". Proc. IEEE 104 (9), 1681–1710.
Xu, J., Chang, E.C., Zhou, J., 2009. "weak leakage-resilient client-side deduplication of encrypted data in cloud storage,". In: ASIACCS'09, pp. 195–206. Hangzhou, China.
Xu, Q., Su, Z., Guo, S., 2016. "a game theoretical incentive scheme for relay selection services in mobile social networks,". IEEE Trans. Veh. Technol. 65 (8), 6692–6702.
Yan, Z., Ding, W., Yu, X., Zhu, H., Deng, R.H., 2016a. "deduplication on encrypted big data in cloud,". IEEE Trans. Big Data 2 (2), 138–150.
Yan, Z., Li, X.Y., Wang, M.J., Vasilakos, A.V., 2017a. "flexible data access control based on trust and reputation in cloud computing,". IEEE Trans. Cloud Comput. 5 (3), 485–498.
Yan, Z., Liang, X., Ding, W., Yu, X., Wang, M., Deng, R., 2019. "encrypted big data deduplication in cloud storage,". Book Smart Data: State-of-the-Art Perspectives in Computing and Applications. published by Taylor and Francis doi:10.1201/9780429507670-4.
Yan, Z., Wang, M., Li, Y., Vasilakos, A.V., 2016b. "encrypted data management with deduplication in cloud computing,". IEEE Cloud Comput. 3 (2), 28–35.
Yan, Z., Zhang, L., Ding, W., Zheng, Q., 2017b. "heterogeneous data storage management with deduplication in cloud computing,". IEEE Trans Big Data 99. 1–1.
Yang, K., Jia, X., 2014. "expressive, efficient, and revocable data access control for multi-authority cloud storage,". IEEE Trans. Parallel Distrib. Syst. 25 (7), 1735–1744.
Yu, R., Zhang, Y., Gjessing, S., Xia, W., Yang, K., 2013. "toward cloud-based vehicular networks with efficient resource management,". IEEE Netw. 27 (5), 48–55.
Zhou, L., Varadharajan, V., Hitchens, M., 2013. "achieving secure role-based access control on encrypted data in cloud storage,". IEEE Trans. Inf. Forensics Secur. 8 (12), 1947–1960.

**Xueqin Liang** received the B.Sc. degree on Applied Mathematics from Anhui University, Anhui, China, 2015. She is currently working for her Ph.D. 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 Blockchain.

**Zheng Yan** received the BEng degree in electrical engineering and the M.Eng. degree in computer science and engineering from the Xin Jiaotong University, Xin, China in 1994 and 1997, respectively, the second M.Eng. degree in information security from the National University of Singapore, Singapore in 2000, and the licentiate of science and the doctor of science in technology in electrical engineering from Helsinki University of Technology, Helsinki, Finland. She is currently a professor at the Xidian University, China and a visiting professor and Finnish academy research fellow at the Aalto University, 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 inventor of 24 patents, all of them having been adopted in industry. 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 a number of 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.

**Robert H. Deng** is AXA Chair Professor of Cybersecurity and Professor of Information Systems in the School of Information Systems, Singapore Management University since 2004. His research interests include data security and privacy, multimedia security, network and system security. He served/is serving on the editorial boards of many international journals, including TFIS, TDSC. He has received the Distinguished Paper Award (NDSS 2012), Best Paper Award (CMS 2012), Best Journal Paper Award (IEEE Communications Society 2017). He is a fellow of the IEEE.