Liang, Xueqin; Yan, Zheng; Deng, Robert; Zheng, Qinghua

Investigating the Adoption of Hybrid Encrypted Cloud Data Deduplication with Game Theory

Published in:
IEEE Transactions on Parallel and Distributed Systems

DOI:
10.1109/TPDS.2020.3028685

Published: 01/03/2021

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., Deng, R., & Zheng, Q. (2021). Investigating the Adoption of Hybrid Encrypted Cloud Data Deduplication with Game Theory. *IEEE Transactions on Parallel and Distributed Systems* , *32*(3), 587-600. https://doi.org/10.1109/TPDS.2020.3028685

Investigating the Adoption of Hybrid Encrypted Cloud Data Deduplication with Game Theory

Xueqin Liang, Zheng Yan, *Senior Member, IEEE,* Robert H. Deng, *Fellow, IEEE,* and Qinghua Zheng, *Member, IEEE*

Abstract—Encrypted data deduplication, along with different preferences in data access control, brings the birth of hybrid encrypted cloud data deduplication (H-DEDU for short). However, whether H-DEDU can be successfully deployed in practice has not been seriously investigated. Obviously, the adoption of H-DEDU depends on whether it can bring economic benefits to all stakeholders. But existing economic models of cloud storage fail to support H-DEDU due to complicated interactions among stakeholders. In this paper, we establish a formal economic model of H-DEDU by formulating the utilities of all involved stakeholders, i.e., data holders, data owners, and Cloud Storage Providers (CSPs). Then, we construct a multi-stage Stackelberg game, which consists of Holder Participation Game, Owner Online Game and CSP Pricing Game, to capture the interactions among all system stakeholders. We further analyze the conditions of the existence of a sub-game perfect Nash Equilibrium and propose a gradient-based algorithm to help the stakeholders choose near-optimal strategies. Extensive experiments show the feasibility of the proposed algorithm in achieving the Nash Equilibrium of the Stackelberg game. Additionally, we investigate the effects of parameters related to CSP, data owners and data holders on H-DEDU adoption. Our study advises all stakeholders the best strategies to adopt H-DEDU.

Index Terms—Cloud Computing, Deduplication, Gradient-Based Algorithm, Multi-Stage Stackelberg Game.

1 Introduction

Deduplication, as an efficient way to eliminate redundant data storage, has become a popular research topic in the field of economic cloud computing. Current storage service faces explosive growth of data volume and additional storage costs caused by inadvertent multiple storage and backup demands. A recent study [1] performed by Microsoft shows that about 68% of data are duplicately stored. Deduplication is a technology to find the existence of a duplicate and substitute it with a pointer to a single shared copy. A piece of data could be deduplicated at a file-level [2] or a chunk-level [3], [4], and the latter one is popular due to better compression performance [5].

The benefits for a Cloud Service Provider (CSP) to adopt deduplication is noticeable. First, the CSP can greatly reduce its storage cost by only storing one copy for each unique data. Second, network bandwidth is conserved by avoiding the transmission of redundant data. Third, the data management cost sharply drops and the CSP can provide more storage spaces at the same price. Security and privacy concerns require data to be stored in an encrypted form in the cloud. However, outsourcing encrypted data to the cloud greatly increases the difficulty of deduplication. Without secure access control over deduplication, data disclosure could happen and impose significant loss to the users.

Efforts [6] in removing duplicated encrypted data while ensuring data security mainly fall into the following directions: message-dependent encryption [7], [8], [9], proof of ownership [10], [11], traffic obfuscation [12], [13], and deterministic information dispersal [14]. Existing deduplication schemes can support data owners to control deduplication [15] or the CSP works as a proxy of the owners to perform deduplication [11], [16], [17]. Unfortunately, these two solutions either require the owners to keep online or force data users to lose direct data control. Taking the advantages of existing deduplication schemes, a hybrid encrypted cloud data deduplication scheme [10] (H-DEDU) can flexibly control deduplication at either the user side or the CSP side, depending on the preference of data users. In our presentation, a data owner is the first data user to upload data to CSP and has responsibility to control deduplication later on. Data holders refer to the users that upload duplicated data subsequently. We classify data users into these two groups because they play different roles in H-DEDU.

H-DEDU is theoretically feasible and safe; however, whether it can motivate all stakeholders to adopt it remains unstudied. For example, Google Drive provides storage services with an optional deduplication feature. Therefore, whether deduplication can be selected by the users of Google Drive depends on if it can provide enough incentives to them. Liu et al. [7] articulated the necessity of an incentive mechanism in promoting the adoption of deduplication. Youn and Chang [18] recognized the absence of incentives in motivating the participation of data owners and indicated a potential solution by granting discounts. Miao et al. [19] integrated a payment-based incentive scheme into deduplication and provided payment-based incentives

- *X. 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, 710071, China; 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.
- *R. H. Deng is with the School of Information Systems, Singapore Management University, Singapore 188065, Singapore.*
  Email: robertdeng@smu.edu.sg.
- *Q. Zheng is with the School of Computer Science and Technology, Xi'an Jiaotong University, Xi'an, 710049, China.*
  Email:qhzheng@mail.xjtu.edu.cn.

to data users. However, this model is too simple to be extended for applying into other scenarios [20]. As a market-oriented and profit-driven entity, the CSP evaluates the effectiveness of deduplication and makes an optimal decision by considering economic factors. But in order to make deduplication applicable, its optimal decision should also provide economic profits to other stakeholders. Therefore, it becomes essential to investigate the adoption of H-DEDU from the perspective of all stakeholders.

However, exploring the adoption of H-DEDU faces two challenges. First, the literature lacks an economic model of a cloud storage system with H-DEDU. Previous works either only mentioned the need to consider economic factors or were infeasible to be applied into H-DEDU. Second, it is difficult to model the complicated interactions among all stakeholders since their interests are interdependent.

In this article, we attempt to investigate the adoption of H-DEDU with game theory. We first formulate the utility function of each stakeholder as its total gains minus its total costs and analyze their compositions. We further detail a storage discount function for data users and specify the influences of other entities' strategies on a data holder's benefit. Under this economic structure, we model the interactions in H-DEDU as a multi-stage Stackelberg game, where the CSP plays as an absolute leader and the data owner, as a follower of the CSP, is a leader of data holders. The multi-stage Stackelberg game consists of three sub-games: Holder Participation Game, Owner Online Game and CSP Pricing Game. We solve the Stackelberg game with a backward induction method and prove the existence of a perfect Nash Equilibrium in each sub-game. We further propose a gradient-based algorithm in order to help the stakeholders choose near-optimal strategies. Extensive experiments show the feasibility of the proposed algorithm in reaching the Nash Equilibrium of the Stackelberg game (in short Stackelberg Equilibrium). Additionally, we investigate the effects of parameters related to CSP, data owners and data holders and summarize meaningful and interesting insights on H-DEDU deployment. Specifically,

1) We establish an economic model of H-DEDU by specifying the utilities of all stakeholders.
2) We apply a multi-stage Stackelberg game to model the interactions among all stakeholders in H-DEDU and analyze the existence of a Nash Equilibrium by adopting a backward induction method.
3) We design a gradient-based algorithm to search the near-optimal strategies for all stakeholders.
4) We conduct extensive experiments to illustrate that the results of the gradient-based searching algorithm converge to reach the Stackelberg Equilibrium. we also test the effects of a number of parameters related to CSP, data owners and data holders on the adoption of H-DEDU.
5) We discover some interesting insights from our experimental results. Concretely, a wise strategy for CSP is to set an access fee and make it cover the H-DEDU operation costs of data owners; the data holders with popular data are likely to accept H-DEDU and the CSP tends to control popular data deduplication; CSP should grant additional discounts to sensitive data in order to promote H-DEDU adoption.

The rest of this article is structured as follows. Section 2 provides the basic of game theory and deduplication. It gives a brief review on deduplication incentives and the applications of game theory. We detail the procedure of H-DEDU in Section 3, along with its practical deployment problems. We also describe our research assumptions in this section, followed by the proposed economic model in Section 4. In Section 5, we formulate a multi-stage Stackelberg game to model the interactions among all stakeholders and analyze the existence of sub-game perfect equilibrium. The experimental results are presented in Section 6 with essential discussions on our discovery. Finally, we conclude this article in the last section.

2 Background and Related Work

In this section, we first present the basic of game theory and deduplication. Then, we briefly review the state-of-art economic incentives in deduplication and game-theoretical approaches applied in computer sciences.

2.1 Game Theory

Game Theory [21] is a cross-discipline subject to study the interactions and competitions among individuals. It considers the predictive behavior and actual behavior of individuals and studies their optimization strategies. To describe a game, we need to know who is participating in and making decisions (i.e., players), what decisions can they make (i.e., strategies), and what are the possible results for these decisions (i.e., outcomes).

A Nash Equilibrium (NE) of a strategic game is a strategy profile with the property that no player can increase its payoff by deviating to a different action, provided the other players' actions. Therefore, no player has the incentive to change its action unilaterally in a NE state.

Stackelberg game [22], [23] is a sequential strategic game that one player moves first and all the others react to the first mover's decision subsequently to minimize their costs or maximize their utilities, after which the first mover updates its strategy to optimize its utility. The first mover is called the leader while the others are the followers. Obviously, the leader has an overwhelming advantage in such a game.

The NE in a Stackelberg game model is called Stackelberg Equilibrium, which is achieved by finding the sub-game perfect Nash Equilibrium (SPNE). A prevalent solution to gain the SPNE is the backward induction, which first seeks the best responses for the followers and afterwards solves an optimization problem for the leader.

2.2 Deduplication

Based on deduplication is controlled by CSP, data owners, or both of them, encrypted cloud data deduplication schemes can be classified into three categories: server-controlled deduplication (S-DEDU), client-controlled deduplication (C-DEDU) and hybrid deduplication (H-DEDU). C-DEDU requires the data owner to keep online for providing deduplicated storage services to data users. Otherwise, data users

will suffer from service delays. S-DEDU relieves the online requirement on data owners by allowing a CSP to work as a proxy to control deduplication. However, the participation of a third party increases the risk of malicious behaviors and collusion. H-DEDU enables flexible data deduplication. A data owner controls deduplication when it is online and grants the control right to CSP when it is offline. Therefore, H-DEDU holds the advantages of both S-DEDU and C-DEDU. For the details of the above three types of deduplication schemes, please refer to [24].

Deduplication rate refers to a parameter to estimate the effectiveness of deduplication. Let n and N be the number of data holders that accept deduplication and the total number of users of this data. The deduplication rate with notation r can be represented as:

$$r = \frac{n}{N}. \tag{1}$$

2.3 Related Work

2.3.1 Incentives in Deduplication

Researchers have noticed the importance of incentives in deduplication. Some mentioned the necessity of incentives to promote the acceptance of deduplication [7], [18], some applied technical approaches to guarantee the performance of CSPs [16], and some provided incentives from an economic perspective [19], [20], [25], [26].

Liu et al. [7] encouraged the CSP, the direct beneficiary of deduplication, to rise storage quotas for a user whose data is deduplicated. Armknecht et al. [16] proposed ClearBox to restrict CSPs by allowing the users to attest to the behaviors of CSP. ClearBox provides strong incentives to the users of popular data, however, it keeps unpopular data out.

Youn and Chang [18] identified the selfish actions of data holders caused by the lack of incentives in a deduplication scheme. They analyzed the disadvantages of being the first data user to adopt deduplication and listed some potential compensation strategies. They stated that discounting on the first data user's storage-service fee could help but they did not provide a real mechanism.

Jin et al. [25] figured out the existence of selfish data holders that only enjoy the benefits of deduplication but refuse to take part in data ownership verification. They proposed a solution that all the holders share one piece of data storage cost. However, this incentive mechanism fails to benefit the CSPs. They believed that market competition among CSPs would decide the best pricing strategy.

Miao et al. [19] monitored the selfish behavior of CSP in publishing an unfair pricing strategy to data holders. They applied a payment-based incentive into S-DEDU and charged the users according to a deduplication rate. The payment structure is similar to that in [25], where all the holders with the same deduplicated data share the storage fee paid to CSPs. However, Liang et al. [20] formally proved that this incentive cannot guarantee the profits of CSPs. To solve this problem, they set an upper limit to the discount, which is related to the total number of data holders and other system parameters. They proved their discount-based incentive mechanism to be individual-rational, incentive-compatible, profitable, and robust. Liang et al. [26] further analyzed the feasibility of unified discount and individualized discount based on their proposed economic structure in C-DEDU. They also considered how to preserve the privacy of data holders when designing an incentive mechanism. However, the economic models in [20], [26] cannot be directly applied into the scenario of H-DEDU due to the difference of deduplication scheme design.

2.3.2 Applications of Game Theory

Interdisciplinary cooperation increases the application of game theory in solving difficult problems and making optimal decisions. Numerous survey papers [22], [27], [28], [29], [30], [31] have shown the great potential of game theory in addressing security and privacy issues in computer science.

Game-theoretical analysis offers great help in eliminating selfish behaviors and promoting scheme acceptance, thus ensuring the long-term development of a scheme. Yu et al. [32] employed a game-theoretical method to analyze how vehicles optimally share resources to improve network performance. In wireless multimedia social networks, Nan et al. [33] 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 [34]. Researchers in [35] proposed a game theory-based distributed task scheduling scheme that can eliminate all entities' selfish behaviors and achieve social optimal.

The two-stage Stackelberg game is widely applied to model the interactions among services or resource suppliers and buyers [36], or among buyers [37]. Yu and Hong [36] considered a smart grid scenario with one service provider (i.e., energy management center) and multiple buyers (i.e., devices) and applied the Stackelberg game to capture their interactions. An existing and unique Stackelberg equilibrium was proved to be the optimal strategy profile for all players. Wei et al. [37] investigated how to allocate virtual cloud computing resources by adopting the Stackelberg game with imperfect information to capture the competition among two buyers. Through dynamic bid prediction and strategy update, the game finally reaches its NE, where all players gain increased profits.

Xiong et al. [38] adopted a three-stage Stackelberg game to model the interactions among a service provider, a content provider, and end-users. The game is solved through backward induction and converges to a unique Stackelberg Equilibrium, at which all players obtain their optimal benefits. The interplay in a supply chain scenario with duopolies was also formulated as a three-stage Stackelberg game in [39]. The players are a manufacturer with a pricing strategy, a distributor with a pricing strategy, and two retailers with their demand strategy.

However, the Stackelberg game has never been applied into the study of encrypted data deduplication although it shows specific advantages in analyzing the complicated interactions among multiple players.

3 System Model

This section describes a cloud storage system with H-DEDU. We briefly present the work flow of this system and discuss potential deployment problems of H-DEDU. We also summarize our research assumptions in this section.

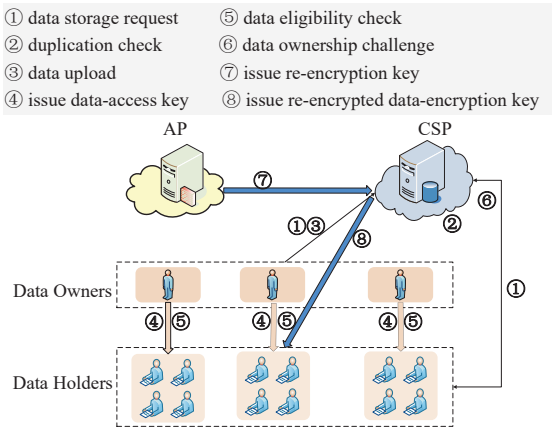

Fig. 1. System model

3.1 H-DEDU

Fig. 1 drafts the structure of a cloud storage system with H-DEDU. There are four kinds of system entities: data holders, data owners, CSP and Authorized Party (AP). A data owner refers to the first user that uploads data and controls its deduplication with responsibility, while data holders are the users that upload duplicated data subsequently. CSP provides cloud storage services to data owners and data holders. To ensure data security, CSP cooperates with AP to control deduplication when the owner grants deduplication right to it. AP is a fully trusted party that is introduced for verifying data ownership and controlling data access, as well as eliminating malicious behaviors of CSPs [11]. A simple description of H-DEDU is provided below:

- A user sends its data storage request to a CSP;
- The CSP performs duplication check on this data and asks the user to upload this data if it has not been stored before. If the user would like to be online to control its data deduplication, it is regarded as a data owner. It encrypts the data with a data-encryption key that is encrypted with the public key of this user. Otherwise, the user encrypts the data-encryption key with the public key of AP to grant the right of controlling deduplication to the CSP and AP. If the data exists already, the CSP first challenges the ownership of this user by executing an ownership challenge protocol [11], [15], [40], e.g., asking the user to response with the hash value of a randomly selected part of this data. H-DEDU is performed only when the user passes the ownership challenge;
- If deduplication is directly controlled by the data owner, the CSP contacts the owner for performing H-DEDU, e.g., by applying Attribute-Based Encryption (ABE)-based deduplication [10]. The attribute used for deduplication access control could be user identity. After receiving the deduplication request from the CSP, the data owner checks the eligibility (or attribute) of this user and only issues a data access key to an eligible one. With the data-access key, the user can decrypt the data-encryption key and then access the data;
- If deduplication is controlled by the CSP, the CSP works as a proxy to provide deduplication by applying Proxy Re-Encryption (PRE). Based on the public and secret keys of the AP and the public key of this user, AP generates a re-encryption key and sends it to the CSP. With the re-encryption key, the CSP transfers the encrypted data-encryption key to a new one, which can be decrypted by the user's secret key. Then the CSP issues the newly re-encrypted data-encryption key to this user;
- After receiving the key issued by either the data owner or the CSP, this user can obtain raw data through decryption. The data deduplication procedure ends.

The data ownership challenge and the attribute employed in the above illustration are based on [10]. Other kinds of proof-of-ownership methods [11], [15], [40], and more complex attribute structures are also applicable, which do not impact the economic analysis performed in this article.

3.2 Practical Deployment Problems

Google Drive has provided a cloud storage service with optional deduplication up to data users' choice. When a data user uploads a duplicated file, the user can choose whether to activate deduplication. From this wide-use cloud storage service, we can see that deduplication's adoption relies on data user's willingness. Similarly, the success of H-DEDU deployment requires the acceptance of all system entities (i.e., stakeholders) including data users. However, whether they are willing to adopt H-DEDU has been scarcely investigated.

The adoption of H-DEDU reduces the storage cost of a CSP. If the CSP transfers saved costs as a data storage discount shared among its users, it will attract more subscribers and gain more profits. However, the CSP with H-DEDU needs to choose an appropriate discount policy for the first time when it enters into a cloud storage market. The art of discount selection lies in the balance between the saved costs and the given discount. With a large discount, the CSP could attract a large number of data holders, but the discount could exceed its saved storage costs, thus lower its expected utility. The CSP should also consider the influence of uncooperative data users.

A data owner can utterly determine whether to control deduplication according to its preference. Since keeping online is costly and impractical, the data owners may control deduplication when they are online and grant the right to CSPs when being offline. A trade-off between online operation cost and offline probability needs to be considered when investigating H-DEDU deployment in practice.

Data holders are the most passive participants in the cloud storage system. The pricing policy and data access policy are determined by the CSP and the data owner, respectively. The data holders have right to select a CSP with or without H-DEDU. Therefore, it is reasonable for the CSP to understand that not all data holders participate in H-DEDU during its decision-making.

In addition, all system entities' behaviors are impacted with each other. The pricing model decided by CSPs in-

fluences the participation willingness of data users. The offline probability of a data owner implies the possibility of deduplication to be controlled by CSPs, which impedes the acceptance of the data users that suspect the credibility of CSPs. Furthermore, the acceptance of the data users on H-DEDU has a direct impact on the pricing model.

In practice, the CSP first determines whether to adopt H-DEDU and publishes its pricing strategies, then the data owner determines its online probability, and subsequently, the data holders take actions based on the online probability and their individual characteristics accordingly.

3.3 Assumptions

In this subsection, we summarize our research assumptions with justification.

We assume all stakeholders are profit-driven, which is a common phenomenon in a practical market. Data users are usually individuals or organizations that have data to be stored. CSPs pursue profits from cloud storage service provision. Therefore, we assume all stakeholders are rational to take actions for maximizing their individual utilities. We regard the acceptance of cloud storage as a consensus based on previous study [41], which means all data users are willing to choose cloud storage.

Due to the data security concern, data users encrypt data no matter a deduplication scheme is applied or not. Since the operation cost of data encryption cannot be avoided and unrelated to H-DEDU deployment, we ignore it in the utility functions of data users for simplifying our analysis.

Since the data user chooses whether to activate deduplication on its own in practice, it has no incentive to modify data fingerprint for avoiding deduplication. Thus, we assume the data user honestly provides data fingerprints to CSP for allowing it to calculate a data deduplication rate.

Based on [41], collusion will worsen the reputation of CSPs. Therefore, AP and CSP do not collude due to different business incentives and interests.

We assume a CSP will back up its data timely to prevent irreversible data loss. And cloud data backup is beyond the focus of this article work.

We assume CSP regularly pays a service fee to AP. AP always charges the CSP a reasonable fee that can be afforded by the CSP; otherwise, AP will lose subscribers and incomes. The utility of AP is only directly related to the CSP and will not be affected by other factors. Therefore, we consider a simple game model with three types of players, which are the data holders, the data owners, and the CSPs.

4 Economic Model

This section constructs the utility functions of all players. Our analysis is based on a simple scenario: a piece of data d_i held by a number of data users stores at CSP k. To avoid any misinterpretation, o_i denotes the data owner that is the first data user to upload data d_i and control its deduplication. $\mathcal{H} = \{h_j^i | j = 1, 2, \ldots, N\}$ denotes the data holders of d_i. N is the total number of data holders.

Gao et al. [41] designed the utility functions of a cloud storage system with reputation and trust, which cannot be directly applied in this article. Miao et al. [19] presented the utility functions of a cloud storage system with deduplication. However, they failed to provide incentive compatibility to CSPs [20]. Therefore, we proposed new utility functions for a cloud storage system with H-DEDU. The utility of each entity amounts to its total gains minus its total costs. For easy presentation, Table 1 describes all the notations used in the rest of this article.

TABLE 1
Notations

Notations	Descriptions
uH_j^i	The utility of data holder h_j^i;
uO_i	The utility of data owner o_i;
uC_k	The utility of CSP k;
b	The cloud storage benefit for a data user;
sf	The storage-service fee paid by a data user;
sc	The storage cost of CSP;
α	The discount of storage-service fee;
p_i	The online probability of data owner o_i;
RF	The data request fee;
oc	The H-DEDU operation cost of the data owner;
OC	The H-DEDU operation cost of CSP;
r	The deduplication rate;
N	The total number of data holders;
c	The confidence of a data holder on CSP security.

4.1 Utility of Data Holder

Cloud storage enables data users to access their data at any time and any where, which greatly saves their local storage spaces. We quantify the benefit that a data holder h_j^i benefits from the cloud storage as b. Data holder h_j^i should also pay a storage-service fee to CSPs, which is denoted as sf. Therefore, the utility of data holder h_j^i is presented as:

$$uH_j^i = b - sf. \tag{2}$$

It is essential to state that it is difficult to quantify and calculate the benefit of cloud storage. Nevertheless, we define it with b through the inspiration of [41] and default it to be larger than sf, which has been already proved by existing CSPs (like Google Drive, iCloud, etc).

If H-DEDU is applied, only when o_i is online and controls deduplication, can h_j^i obtain the total cloud storage benefit. Otherwise, deduplication is controlled at the server-side and something unexpected (like collusion between malicious data users and CSP) would happen because of the loss of direct control. The increase of deduplication rate subsequently introduces malicious behavior of data users. Their malicious behavior decreases the cloud storage benefits of honest data users. Hence, the cloud storage benefit of a data holder when H-DEDU is applied is related to the data owner's online probability and the deduplication rate. We express it as $f(p_i, r)b$. To motivate data users to accept H-DEDU, CSP k gives a discount on storage-service fee to its users. Based on our previous work [20], [26], the discount should be determined based on its upper limit α and current data deduplication rate r, represented as $g(\alpha, r)$. Hence, the utility of h_j^i that adopts cloud storage with H-DEDU is:

$$uH_j^i = f(p_i, r)b - sf + g(\alpha, r)sf. \tag{3}$$

Where, $f(p_i, r)$ and $g(\alpha, r)$ are defined below:

$$f(p_i, r) = p_i + (1 - p_i)(1 - r^c) = 1 + (p_i - 1)r^c, c > 2 \tag{4}$$

$$g(\alpha, r) = \alpha(1 - e^{-r}). \quad (5)$$

4.2 Utility of Data Owner

Since being a data owner has some disadvantage, as analyzed in [18], the data owner should be motivated to firstly upload data to the cloud. Apart from the discount on storage-service fee, we also propose to let the data owner charge an access fee (denoted as AF) from the CSP. Therefore, the data owner can obtain more benefits than data holders to incent first data uploading. On the other hand, the operation cost of the data owner to perform deduplication for data holders can also be compensated by the access fee. Obviously, AF should be discounted by the deduplication rate r. The storage-service fee of o_i is the same as its holders. However, keeping online takes o_i a cost oc and online probability p_i directly impacts this cost. Hence, we formulate the utility of o_i with online probability p_i as:

$$uO_i = b - sf + g(\alpha, r)sf + rAF - p_i oc \quad (6)$$

4.3 Utility of CSP

In a cloud storage system without H-DEDU, CSP k needs to store one copy of data at a cost sc for every user $u \in \mathcal{H} \cup \{o_i\}$. Therefore, the storage-service fee charged from its users should rationally compensate for this cost. Namely,

$$sf > sc. \quad (7)$$

The utility function of a CSP without H-DEDU is

$$uC_k = \sum_{u \in \mathcal{H} \cup \{o_i\}} (sf - sc). \quad (8)$$

According to the previous discussion in Section 4.1 and Section 4.2, we can summarize the utility of CSP k for providing cloud storage services with H-DEDU. If the deduplication rate is r, the total storage-service fee that CSP k obtains from the data users that accept H-DEDU is $rNsf - rNg(\alpha, r)sf$, which is at a cost of storing one copy of data. The access fee that k pays to the data owner is rAF. The operation cost for conducting H-DEDU is OC, which contains the service fee paid to AP. Hence, we conclude the utility of CSP k for performing H-DEDU as

$$uC_k = rNsf - rNg(\alpha, r)sf - sc - rAF - OC. \quad (9)$$

5 Game Formulation and Analysis

In this article, we model the interactions among the CSP, the data owner and the data holders as a multi-stage Stackelberg game. As the direct beneficiary of H-DEDU, CSP takes an unquestionably leading role in deciding whether to adopt H-DEDU and all data users play as its followers. Among all the users, the data owner is the leader since it is the first user to upload data. In a nutshell, CSP selects its pricing strategy in Stage I, based on which the owner decides its online probability in Stage II. In Stage III, the data holders together determine the deduplication rate. In this section, we mathematically formulate the problems needed to be solved in the above three stages by constructing three sub-games (namely Holder Participation Game, Owner Online Game and CSP Pricing Game). Then, we discover the sub-game perfect Nash Equilibrium by applying backward induction.

5.1 Game Formulation

5.1.1 Holder Participation Game (HPG)

Given the discount α set by the CSP and the online probability p_i of data owner o_i, data holders cooperate with each other and choose the strategies to maximize their utilities. We denote the game as $\mathcal{G}_H = \{\mathcal{H}, \{r_j\}_{h_j^i \in \mathcal{H}}, \{uH_j^i\}_{h_j^i \in \mathcal{H}}\}$, where $\mathcal{H}$ is the data holder set, $\{r_j\}_{h_j^i \in \mathcal{H}}$ is the strategy set and $\{uH_j^i\}_{h_j^i \in \mathcal{H}}$ is the utility set. Each data holder $h_j^i \in \mathcal{H}$ chooses the best strategy r_j to maximize its utility. Under the same cloud storage environment, the best strategy of all holders should be quite similar, thus we simply denote it as r. So, the holder participation sub-game is to find the solution for the following optimization problem:

$$\max_{0 \le r \le 1} uH_j^i(r; \alpha, p_i). \quad (10)$$

5.1.2 Owner Online Game (OOG)

Based on the NE in $\mathcal{G}_H$ and the given CSP pricing strategy α, the data owner o_i decides its best strategy $p_i \in [0, 1]$ to maximize its utility. Therefore, the data owner tries to find the solution of the following problem:

$$\max_{0 \le p_i \le 1} uO_i(p_i; \alpha). \quad (11)$$

5.1.3 CSP Pricing Game (CPG)

Knowing the best strategies of all data holders and the owner, CSP k decides its best strategy $\alpha \in [0, 1]$ to optimize its pricing strategy. The best strategy of CSP is the solution of the following problem:

$$\max_{0 \le \alpha \le 1} uC_k(\alpha). \quad (12)$$

5.2 Equilibrium Analysis

All the sub-games specified above form a multi-stage Stackelberg game with complete information. We employ the backward induction method to find its equilibrium, which is the state that CSP achieves its maximum payoff when both the data owner and the data holders play the best-response strategies.

5.2.1 Equilibrium Analysis of HPG

We first provide the Nash Equilibrium definition of HPG as follows.

Definition 1. The Nash Equilibrium of $\mathcal{G}_H$ is r^*, if for any $h_j^i \in \mathcal{H}$,

$$uH_j^i(r^*; \alpha, p_i) \ge uH_j^i(r; \alpha, p_i) \quad (13)$$

is satisfied for all $r \in [0, 1]$.

Theorem 1. The Nash Equilibrium of HPG $\mathcal{G}_H = \{\mathcal{H}, \{r_j\}_{h_j^i \in \mathcal{H}}, \{uH_j^i\}_{h_j^i \in \mathcal{H}}\}$ exists.

Proof 1. We first calculate the first-order and second-order partial derivatives of (3) with (4) and (5):

$$\frac{\partial uH_j^i}{\partial r} = c(p_i - 1)r^{c-1}b + \alpha sfe^{-r}, \quad (14)$$

$$\frac{\partial^2 uH_j^i}{\partial r^2} = c(c-1)(p_i - 1)r^{c-2}b - \alpha sfe^{-r}. \quad (15)$$

Since $c > 2$ and $p_i \in [0,1]$, then $c(c-1)(p_i-1) \leq 0$.

$$\frac{\partial^2 uH_j^i}{\partial r^2} < 0. \tag{16}$$

Therefore, uH_j^i is strictly concave with respect to r. Considering the strategy space of r is a non-empty convex and a compact subset of the Euclidean space, the existence of NE is proved.

Let $\frac{\partial uH_j^i}{\partial r} = 0$, we have

$$c(1-p_i)r^{c-1}b = \alpha sfe^{-r}. \tag{17}$$

$$r = \ln c(1-p_i)b + \ln r^{c-1} - \ln \alpha sf. \tag{18}$$

Let r^* be the best response of data holders, then r^* is the solution of (18) and

$$r^* - \ln {r^*}^{c-1} = \ln c(1-p_i)b - \ln \alpha sf. \tag{19}$$

5.2.2 Equilibrium Analysis of OOG

In Stage II, data owner o_i chooses the optimal strategy based on the sub-game perfect equilibrium achieved in HPG. The utility function of o_i is formulated as (20) when taking (5) and r^* into consideration.

$$uO_i = b - sf + \alpha(1 - e^{-r^*})sf + r^*AF - p_i oc. \tag{20}$$

Below is the definition of Nash Equilibrium of OOG.

Definition 2. The Nash Equilibrium of OOG is p_i^*, if for data owner o_i,

$$uO_i(p_i^*; \alpha) \geq uO_i(p_i; \alpha) \tag{21}$$

is satisfied for all $p_i \in [0,1]$.

Theorem 2. The NE in the OOG exists when

$$(c - r^* - 1)^2 + 1 - c > 0 \tag{22}$$

Proof 2. The first-order and second-order partial derivatives of (20) are listed as follows:

$$\frac{\partial uO_i}{\partial p_i} = (\alpha e^{-r^*} sf + AF)\frac{\partial r^*}{\partial p_i} - oc, \tag{23}$$

$$\frac{\partial^2 uO_i}{\partial {p_i}^2} = -\alpha e^{-r^*}(\frac{\partial r^*}{\partial p_i})^2 sf + (\alpha e^{-r^*} sf + AF)\frac{\partial^2 r^*}{\partial {p_i}^2}. \tag{24}$$

According to (19), we can calculate $\frac{\partial r^*}{\partial p_i}$ and $\frac{\partial^2 r^*}{\partial {p_i}^2}$.

$$\frac{\partial r^*}{\partial p_i} = \frac{r^*}{(p_i-1)(r^* - c + 1)} > 0. \tag{25}$$

$$\frac{\partial^2 r^*}{\partial {p_i}^2} = r^*(p_i-1)^{-2}\frac{(c - r^* - 1)^{-2} + 1 - c}{(r^* - c + 1)^3}. \tag{26}$$

When (22) holds, as $c > 1$ and $r^* > 0$, $r^* - c + 1 < 0$, then we can easily conclude that

$$\frac{\partial^2 r^*}{\partial {p_i}^2} < 0, \tag{27}$$

$$\frac{\partial^2 uO_i}{\partial p_i^2} < 0. \tag{28}$$

Furthermore, with the property that the strategy space of data owner, $p_i \in [0,1]$ is a non-empty convex and a compact subset of the Euclidean space, we complete the proof of the existence of NE.

Let $\frac{\partial uO_i}{\partial p_i} = 0$, we have

$$(\alpha e^{-r^*} sf + AF)\frac{r^*}{(p_i-1)(r^* - c + 1)} - oc = 0. \tag{29}$$

If p_i^* denotes the best response of the data owner, then

$$p_i^* = 1 + \frac{r^*(\alpha e^{-r^*} sf + AF)}{(r^* - c + 1)oc} \tag{30}$$

5.2.3 Equilibrium Analysis of CPG

With the optimal deduplication rate r^* of data holders and the optimal online probability p_i^* of the data owner, the CSP determines its best strategy by solving the optimization problem (12). The strategy of CSP is the pricing strategy, or discount decision $\alpha \in [0,1]$, to be precise. Considering the expressions of r^* and p_i^* and Theorem 2, we reformulate (12) as follows:

$$\begin{aligned}
&\underset{\alpha}{\text{maximize}}\ uC_k(\alpha) \\
&\text{subject to } \alpha \in [0,1] && (31)\\
&\qquad p_i^* = \arg\min_{p_i} uO_i && (32)\\
&\qquad (c - r^* - 1)^2 + 1 - c > 0 && (33)\\
&\qquad r^* - \ln {r^*}^{c-1} = \ln c(1-p_i)b - \ln \alpha sf && (34)
\end{aligned}$$

The constraint (31) represents the strategy space of CSP. The constraint (32) imposes that p_i^* is the best response of the data owner when the pricing strategy of CSP is fixed. The constraint (33) is added according to Theorem 2. The constraint (34) is to indicate that r^* is the best response of data holders when p_i and α are determined.

Theorem 3. The NE of the CPG exists when the following equation is satisfied:

$$Nsf - \alpha Nsf - AF > 0. \tag{35}$$

Proof 3. Taking (5) into (9), we obtain

$$uC_k(\alpha) = (1 - \alpha + \alpha e^{-r^*})r^* Nsf - sc - r^* AF - OC. \tag{36}$$

Likewise, we calculate the first-order and second-order partial derivatives of (36) and present them in (37) and (38), respectively.

$$\frac{\partial uC_k}{\partial \alpha} = \frac{\partial r^*}{\partial \alpha}(Nsf - \alpha Nsf - AF) - r^* Nsf, \tag{37}$$

$$\frac{\partial^2 uC_k}{\partial \alpha^2} = \frac{\partial^2 r^*}{\partial \alpha^2}(Nsf - \alpha Nsf - AF) - 2\frac{\partial r^*}{\partial \alpha} Nsf. \tag{38}$$

By calculating the first-order partial derivative and the second-order partial derivative of (19) with respect to α, we can conclude $\frac{\partial r^*}{\partial \alpha}$ and $\frac{\partial^2 r^*}{\partial \alpha^2}$.

$$\frac{\partial r^*}{\partial \alpha} = \frac{r^*}{\alpha(c - 1 - r^*)}. \tag{39}$$

$$\frac{\partial^2 r^*}{\partial \alpha^2} = r^* \alpha^{-2}\frac{(c - 1 - r^*)^2 + r(c-1)}{(r^* - c + 1)^3}. \tag{40}$$

According to $c > 2$ and $r^* \in [0,1]$, we can easily obtain

$$\frac{\partial r^*}{\partial \alpha} > 0, \tag{41}$$

$$\frac{\partial^2 r^*}{\partial \alpha^2} < 0. \tag{42}$$

Based on (35), (41) and (42),

$$\frac{\partial^2 u C_k}{\partial \alpha^2} < 0. \tag{43}$$

Therefore, the existence of NE in CPG is proved.

5.3 Algorithm to Decide Optimal Strategies

We apply a low-complexity gradient-based searching algorithm to find the optimal strategy of all players, which is shown in Algorithm 1. The algorithm first inputs the initial values of discount $\alpha \in [0,1]$ and online probability $p_i \in [0,1]$. In each iteration, according to the given pricing strategy of CSP and the online probability of the data owner, the data holders calculate their best response based on (19). The owner's best response to the sub-game in Stage II, which is calculated according to (30), also needs to renew when the optimal strategy of data holders is updated. Taking the optimal strategies of all followers (i.e., the data holders and the data owner) into Stage I, the CSP derives its optimal discount by applying the gradient-based algorithm. Then the game goes to another iteration and terminates until $\frac{\left\|\alpha^{[t]}-\alpha^{[t-1]}\right\|_1}{\left\|\alpha^{[t-1]}\right\|_1} < \epsilon$. The optimal strategies of all players are the strategies gained when the game ends.

Algorithm 1 The algorithm to decide optimal strategies

Input:

$\alpha \in [0,1]$, $p_i \in [0,1]$, iteration $t \leftarrow 1$, CSP step size μ, accuracy threshold ϵ;

Output:

Optimal strategies of data holders, data owner and CSP: $r^{*[t]}$, $p_i^{*[t]}$ and $\alpha^{*[t]}$;

1: **repeat**

2: For data holders, they all together decide the deduplication rate $r^{[t]}$ according to

$$r^{[t]} - \ln {r^{[t]}}^{c-1} \leftarrow \ln c(1-p_i^{[t-1]})b - \ln \alpha^{[t-1]} sf;$$

3: Data owner o_i updates its online probability based on

$$p_i^{[t]} \leftarrow 1 + \frac{r^{[t]}(\alpha^{[t-1]} e^{-r^{[t]}} sf + AF)}{(r^{[t]} - c + 1)oc};$$

4: CSP k updates its pricing strategy $\alpha^{[t]}$ according to the gradient assisted searching algorithm:

$$\alpha^{[t]} \leftarrow \alpha^{[t-1]} + \mu \frac{\partial u C_k(r^{[t]}, p_i^{[t]})}{\partial \alpha};$$

5: $t \leftarrow t+1$;

6: **until** $\frac{\left\|\alpha^{[t]}-\alpha^{[t-1]}\right\|_1}{\left\|\alpha^{[t-1]}\right\|_1} < \epsilon$;

7: $r^{*[t]} \leftarrow r^{[t]}$;

8: $p_i^{*[t]} \leftarrow p_i^{[t]}$;

9: $\alpha^{*[t]} \leftarrow \alpha^{[t]}$.

Furthermore, when the initialization state is constant, the output of Algorithm 1 is determined. Therefore, the execution of Algorithm 1 helps the multi-stage Stackelberg game converge to a unique stable state.

TABLE 2
Parameter Settings

$b = 2$	$OC = 2$	$sf = 1$
$sc = 0.5$	$c = 3$	$AF = 3$
$\mu = 0.0001$	$oc = 2$	$\epsilon = 0.01$

6 Experimental Evaluation

We implemented the multi-stage Stackelberg game and evaluated the adoption of H-DEDU in a cloud storage system with five experiments. Experiment 1 evaluated the effectiveness of Algorithm 1 in helping the proposed game model converge to the Stackelberg Equilibrium. Then, we analyzed whether the step size in Algorithm 1 has significant influences on the experimental results in Experiment 2. We investigated the impact of the system parameters related to CSPs, data owners and data holders in Experiment 3 and Experiment 4, respectively. Experiment 5 was conducted based on a real-world dataset by employing the same evaluation metrics as Experiment 1. In this section, we report the above experimental results. In addition, we summarize our findings from the experiments and suggest future work.

6.1 Experimental Settings

This subsection first introduces the real-world dataset used in our experimental test. Next, it specifies the experimental settings about players and system parameters. Finally, it presents our evaluation metrics.

Since the information of data holders and stored data is confidential, CSPs in the real world rarely disclose such information to the public. In this article, we applied relevant real-world data to simulate duplicated cloud data storage. We employed Debian packages in the section *contrib* of Debian Popularity Contest [42] to construct a data storage system. Debian Popularity Contest is a project that tracks the usage of the Debian packages, including a list of packages and the installation times of each package. Different packages are installed by different numbers of users. Some packages are installed by several users and some popular packages are used by tens of thousands of users. This makes the dataset gained from the Debian packages hold very similar properties to a practical cloud storage dataset. Thus, it has been applied by many researchers to simulate cloud data storage status [7], [20], [26]. Specifically, each Debian package represents one piece of data and the number of installation requests can represent the number of data users. We recorded the status of Debian packages on June 19th, 2018 and formulated a dataset with 434 unique data and 309052 data holders to perform our experiments.

In our experiments, we set one CSP to provide cloud storage services with H-DEDU for simplification and analyze the acceptance of H-DEDU without considering the competition among CSPs.

Table 2 shows the default parameter settings in our experiments. We also varied them individually to evaluate their effects on the adoption of H-DEDU. Based on the profitability requirement of players in a cloud storage system, (2) and (8) should be positive. That is,

$$b > sf > sc \tag{44}$$

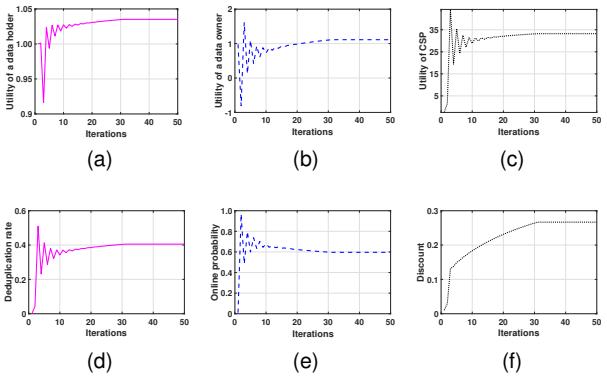

Fig. 2. The results of Experiment 1

should be satisfied. The access fee AF is set to be larger than the operation cost oc of the data owner to ensure that the owner will not decrease its profit by adopting H-DEDU. Note that the utility of the owner would be negative without charging the access fee. Hence,

$$b - sf - oc < 0. \tag{45}$$

We applied the following evaluation metrics to demonstrate the performance of our research result: the utility and strategies of all stakeholders; the number of iterations to reach the NE state for evaluating the performance of Algorithm 1. In the following figures, the experimental results of the utility and deduplication rate related to data holders are shown with (magenta) solid lines (with triangles). The (blue) dashed lines (with circles) and the (black) dotted lines (with five-pointed stars) illustrate the utility and online probability of the data owner as well as the utility and discount of CSP, respectively. The iteration numbers are drawn with the (green) stem.

6.2 Experimental Results

6.2.1 Experiment 1

We conducted Experiment 1 to illustrate how Algorithm 1 helps the multi-stage Stackelberg game converge to a stable state. Experiment 1 evaluates our game model in a simple scenario: one CSP, one data owner and 100 data holders. The reason to set 100 data holders is because the holder numbers of most data in our real-world dataset are within 100. The CSP publishes its initial discount and the data owner initializes its online probability and then all the data holders update their strategies based on Algorithm 1. Specifically, in order to respond to the strategies of the CSP and the data owner at iteration $t-1$, the data holders calculate their best response at iteration t according to Line 2 of Algorithm 1. With the new response $r^{[t]}$ from the data holders, the owner updates its online probability in accordance with $\alpha^{[t-1]}$ (i.e., Line 3 of Algorithm 1). Then, the CSP renews its pricing strategy $\alpha^{[t]}$ accordingly by applying the gradient-based algorithm with $r^{[t]}$ and $p_i^{[t]}$ as input. After that, the deduplication rate is updated and the CSP and the data owner renew their countermeasures. The experimental results are plotted in Fig. 2.

Fig. 2d to Fig. 2f show the strategy variations of all players. The strategies of the owner and the holders fluctuate as the iteration goes by. However, the fluctuation ranges shrink and the curves converge to relatively stable states. The online probability is increased due to the compensation made by discounts and access fees. It decreases mainly because a previous discount is not enough to make up its operation cost, which meanwhile causes the decline of deduplication rate in the next iteration. The discount value witnesses an increasing trend in Fig. 2f. It increases quickly in the first three iterations, then the increase rate gradually reduces to 0 and an equilibrium strategy is reached.

Fig. 2a to Fig. 2c show the utilities of all players. Once the utility of a data user decreases, no matter it is a holder or an owner, the deduplication rate or online probability becomes lower in the next iteration. Before reaching the NE, even if a player obtains a high utility at an iteration, the utilities of other players are at a lower level compared with those at the NE. Such a state is not stable since the players with low utilities have incentives to change their strategies. By applying the parameter settings in Table 2, the game finally reaches the NE at the 31st iteration.

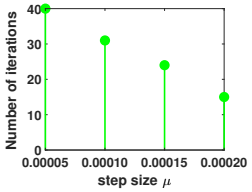

Fig. 3. The influence of step size on the number of iterations

6.2.2 Experiment 2

We changed the value of the step size μ and varied it from 0.00005 to 0.0002 in Experiment 2. All the other parameters in Table 2 except μ were kept the same. The experimental execution goes exactly the same as that in Experiment 1. We found that the step size affects the time to reach convergence as shown in Fig. 3. CSP iterates for a long time in Algorithm 1 to find the best response with a small step size. Nevertheless, the outputs of the fine-grained search algorithm based on a gradient with a smaller step size are theoretically closer to the optimal results.

6.2.3 Experiment 3

We investigated the impact of CSP-related parameters on the evaluation metrics in this experiment. Besides the variation of a specified parameter in each sub-experiment, the other parameters remain the same as in Table 2. The game procedure has no difference from that in Experiment 1.

We first evaluated the influence of OC by varying it from 0 to 10 with a step 1. We found that OC only has an influence on the utility of CSP, which decreases with OC increase. The calculation of deduplication rate in Algorithm 1 is built on (19), which has no relationship with OC. Likewise, the online probability calculation is built on (30) that is also not influenced by OC. Furthermore, OC does not exist in (3) and (6); therefore, the utilities of all data users remain stable when OC changes. In addition, even though the utility function of CSP is directly related to OC, the computation of α only depends on the first-order partial derivative of uC_k with respect to α, which is irrelevant

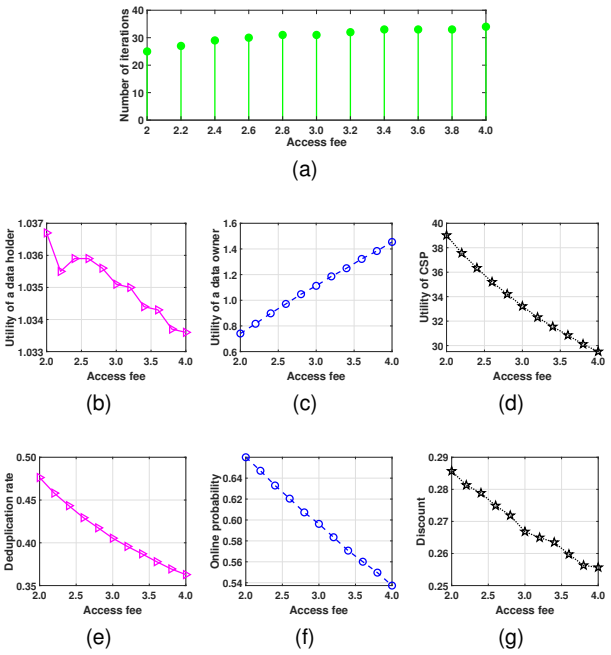

Fig. 4. The influence of access fee AF

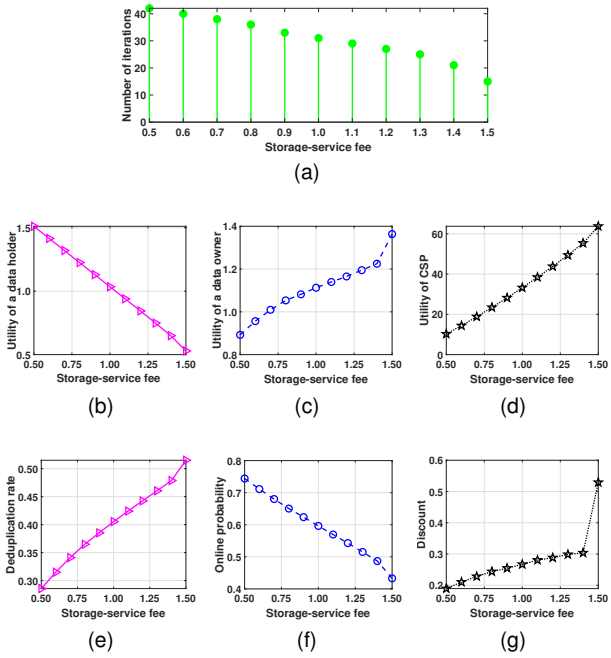

Fig. 5. The influence of storage-service fee sf

to OC according to (37). All the steps in Algorithm 1 are indifferent to the value of OC, therefore, OC does not affect the convergence speed and the number of iterations.

Fig. 4 plots the influence of access fee AF on the evaluation metrics. From Fig. 4a, we observe that the number of iterations needed to reach NE is around 30 with a slightly increasing trend when the access fee increases. The reason of this trend lies in that AF has a negative correlation to $\frac{\partial uC_k}{\partial \alpha}$ (as shown in (37)), which positively affects the time to reach the optimal strategy of CSP. Fig. 4d and Fig. 4g show that when the CSP pays more access fees to the data owner, the CSP will decrease its discount. However, this discount reduction shrinks the utility of data holders, therefore, makes more and more data holders reluctant to adopt H-DEDU, so that the curves in Fig. 4b and Fig. 4e are decaying with the increase of AF. A low deduplication rate reduces data storage frequency; hence, the data owner degrades its online probability to save its operation cost. In spite of the proportion of access fee that the owner can obtain from CSP decreases, the extra access fee and saved operation cost ensures the non-falling profit of the data owner, as shown in Fig. 4c.

Additionally, we investigated how the storage-service fee sf influences the experimental results. As a self-determined parameter, the CSP can set sf as any value between its storage cost and the cloud storage benefit of data users according to (44). In our experiment, we chose the value of sf from 0.5 to 1.5 with a step 0.1, kept all the other parameters as the same as those in Table 2. The experimental results are shown in Fig. 5. Fig. 5a illustrates that the convergence speed of Algorithm 1 decreases with the rise of sf. Fig. 5b shows that the increased storage-service fee reversely influences the utility of data holders. The increase of the deduplication rate in Fig. 5e relies on an increasing discount, as plotted in Fig. 5g. It's worth noting that the sharp increase of the discount in Fig. 5g is caused by the shrink of iteration times. The curve in Fig. 5g has an increasing fashion, which means the CSP can improve its revenue by raising sf when the market competition among CSPs is not considered. The response of the data owner to the increased storage-service fee is to reduce its online probability to save the operation cost. Fortunately, the increasing deduplication rate provides it with more access fees. Therefore, even if the owner needs to pay more storage-service fees, the reduced operation cost and the increased access fee guarantee its profitable utility as shown in Fig. 5c. As the followers, the data holders are the only ones to gain less profits, shown in Fig. 5b.

6.2.4 *Experiment 4*

The fourth experiment evaluates the impact of the parameters related to data users on evaluation metrics, as shown in Fig. 6 to Fig. 9.

We set the number of holders N from 50 to 150 with a step 10 and remained the other parameter settings in Experiment 1. Fig. 6a shows the reduction in the number of iterations needed to reach NE when N increases. Fig. 6b to Fig. 6d illustrate that the number of data holders poses a positive impact on the utilities of all players. With more and more data holders flooding into the CSP, the CSP can save more and more storage spaces even if with the same deduplication rate. This allows the CSP to offer a higher discount to attract more users, as shown in Fig. 6e and Fig. 6g. Fig. 6f illustrates that the willingness of an owner to be online decreases with the rise of parameter N.

Let $\bar{r}$ be the upper limit of the best response of data holders. According to (33) and $c > 2$, we have

$$r^* < c - 1 - \sqrt{c-1}. \tag{46}$$

Therefore,

$$\bar{r} = \begin{cases} c - 1 - \sqrt{c-1}, & c < \frac{5}{2} + \frac{\sqrt{5}}{2} \\ 1. & \text{otherwise} \end{cases} \tag{47}$$

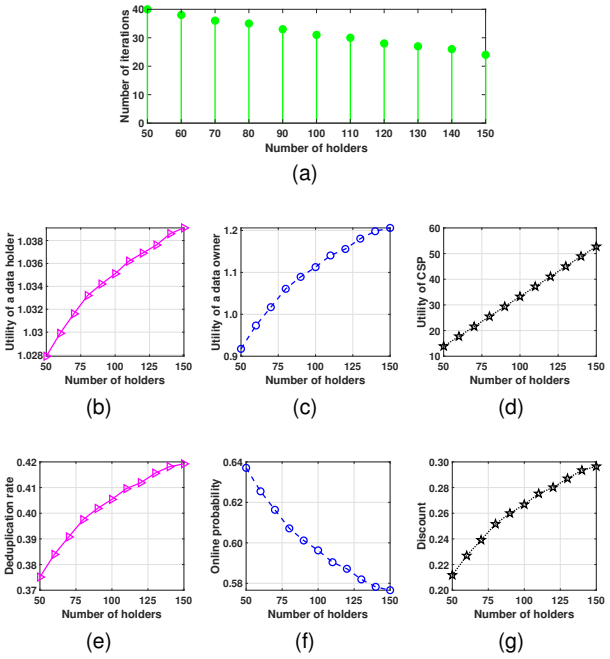

Fig. 6. The influence of data holder number N

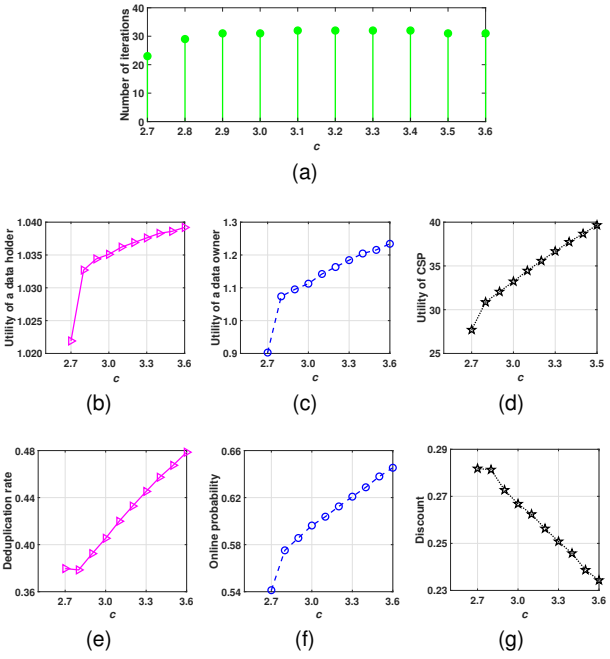

Fig. 7. The influence of parameter c

Fig. 7 depicts the evaluation metrics when the value of parameter c is changed from 2.7 to 3.6, while others were kept the same as in Experiment 1, where 3.6 is no larger than $\frac{5}{2}+\frac{\sqrt{5}}{2}$. The number of iterations to reach NE has little relationship with c and it is around 30, as illustrated in Fig. 7a. The reason is that the algorithm to calculate the optimal strategy of CSP is irrelevant to the value of c. The first-order derivative of (47) is larger than 0; hence, $\bar{r}$ and c are positively correlated. The corresponding r^* increases when changing c from 2.7 to 3.6. The reason is: the larger the parameter c, the more confidence the holders have on the CSP security and they are more likely to accept H-DEDU,

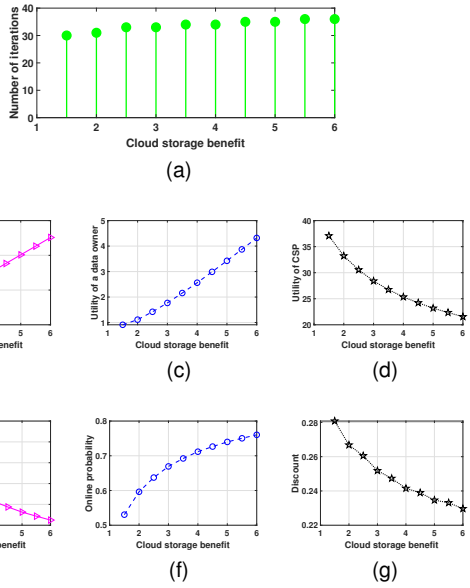

Fig. 8. The influence of cloud storage benefit b

which is experimentally proved in Fig. 7e. The inherent incentive in c lowers the required discount to reach NE. Therefore, a rational CSP cuts down its discount as shown in Fig. 7g. With a high deduplication rate and a low discount, the only possible way for the owner to increase its utility is to maintain its online probability as high as possible. Fig. 7c and Fig. 7f together demonstrate this fact. Fig. 7b to Fig. 7d illustrate that all players' utilities at the NE are raised when increasing the value of c.

Furthermore, we investigated the impact of cloud storage benefit b by increasing it from 1.5 to 6 and summarized the results in Fig. 8. The number of iterations to reach NE fluctuates around 30, as plotted in Fig. 8a. The utilities of all data users are directly related to b. The data owner obtains more revenues to compensate its operation costs with the rise of b; therefore, its online probability is increased with the increase of b. The expression of $f(p_i, r)b$ in (4) shows that the deduplication rate adversely influences the cloud storage benefit b of the data holders. Therefore, a rational data holder decreases the deduplication rate to relieve the influence for a high profit. Fig. 8b and Fig. 8c verify the above statements. The value of b has no direct impact on CSP, but the drop of deduplication rate causes the decrease of the saved storage costs. Therefore, a rational CSP will degrade its discount on the storage-service fee to maintain its profits, as shown in Fig. 8g. Fig. 8b, Fig. 8c together with Fig. 8d demonstrate that the CSP is the only player whose utility goes down when the value of b goes up.

At last, we tested how the evaluation metrics are influenced by the operation cost oc of the data owner and summarized the results in Fig. 9. Fig. 9a illustrates that the number of iterations linearly drops with the augment of oc. Intuitively, a higher operation cost of data owner means a lower utility, which is proved by Fig. 9c. A direct strategy to increase the utility of a data owner is to arise its online probability as shown in Fig. 9f; however, the expensive operation cost still cannot be totally compensated. The rise of the online probability motivates the participation

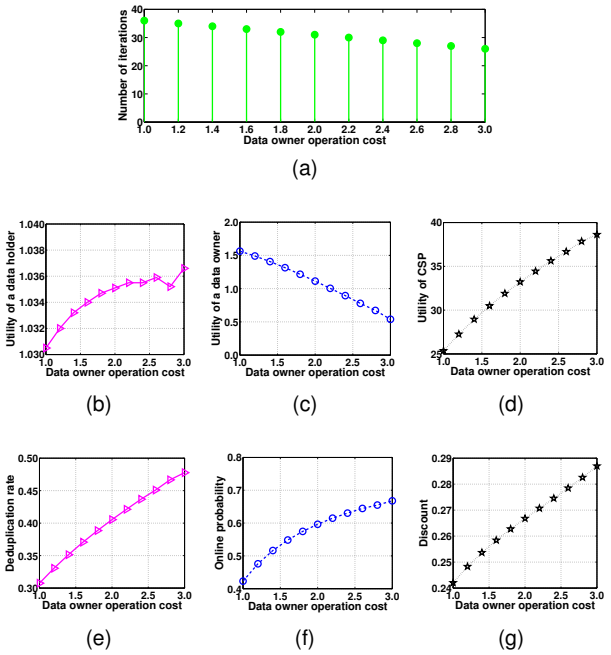

Fig. 9. The influence of owner operation cost oc

willingness of data holders; therefore, the deduplication rate grows and the CSP can save more storage spaces as well as grant more discounts, as shown in Fig. 9e and Fig. 9g. The utility curves of the data holders and CSP in Fig. 9b and Fig. 9d demonstrate that the increase of oc has no negative impact on them and the data owner whose utility is directly related to the value of the operation cost is the only player being evidently impacted.

6.2.5 Experiment 5

In Experiment 5, we considered a complicated scenario by employing the real-world dataset where multiple pieces of data exist. But, we found a fast convergence problem if applying the parameter settings in Table 2. The reason is that the numbers of some data's holders are much more than 100, which decreases the number of iterations to reach NE (refer to Fig. 6a) and negatively affects the accuracy of optimal strategy calculation based on the result of Experiment 4. To overcome this problem, we linked the step size μ to the number of data holders N in Algorithm 1. Concretely, we modified μ in Algorithm 1 as $0.01/N$. The experimental results were plotted in Fig. 10 by employing the same evaluation metrics as Experiment 1. We can observe that the variation trends of all evaluation metrics in Experiment 5 are quite similar to those in Experiment 1 (shown in Fig. 2). This also provides a strong support on the validity of our simulation results achieved in Experiment 1-4.

6.3 Insights and Future Work

In this subsection, we summarize insights derived from the above experimental results and indicate some future work.

The increase of access fee drives a drop of the discounts, which suppresses the willingness of the data owner to be online and reduces the participation enthusiasm of data holders in H-DEDU. Without the competition among CSPs,

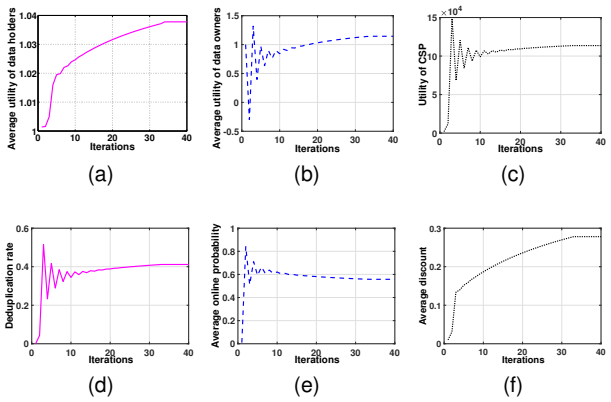

Fig. 10. The results of Experiment 5

a high storage-service fee encourages the rise of the deduplication rate, which implies an increasing acceptance of H-DEDU. On the other hand, the increase of storage-service fee encourages the data owner to grant the deduplication control right to CSP.

When a data belongs to numerous data holders, the CSP is likely to control deduplication. Data holders prefer to store popular data at CSPs with H-DEDU and the CSPs can grant large discount on the popular data. The parameter c reflects the security confidence of data holders in CSPs. The larger c is, the more positive influence the deduplication rate has on a data holder's utility. A large c also means that the data holders care little about its data sharing with others. An insight from this finding is that a CSP can provide a large discount to privacy-related data for encouraging the adoption of H-DEDU. When a data user can obtain a large benefit from cloud storage, no matter it is a holder or a owner would try to prevent any adverse behavior. Specifically, an owner will keep online as long as possible to control its data and the data holders reduce the deduplication rate to prevent possible malicious behaviors caused by H-DEDU.

Overall, H-DEDU is more likely to be accepted by three kinds of data users: the ones that hold popular data, the ones that have high confidence on CSP security, and the ones that benefit a lot from the convenience of cloud storage services. We have some suggestions for all players based on the experimental results. We advise a CSP to set the access fee that just covers the H-DEDU operation cost of data owners and build a good reputation to increase the users' confidence on its security. We advise data users to choose a CSP on which they have more confidence and store popular data at the CSP. A wise choice for the data owner is to grant the deduplication control right to CSPs when the storage-service fee is high.

Although our analysis is based on the H-DEDU in [10], our methodology is applicable to any kinds of hybrid deduplication schemes, where deduplication is controlled by a data owner when the owner is online or the CSP when the data owner is offline.

Our analysis is based on the existence of one CSP that takes absolute control over the economic market. When multiple CSPs exist in the cloud storage system, they will determine a low storage-service fee to attract users. According to (7), the storage-service fee must be able to compensate

for its storage cost. The competition capability of CSPs is highly related to the storage-service fees set by them. We indeed investigated the impact of storage-service fee change on CSP's profit and subsequently its competition capability. In the future, we will further consider the scenario with multiple CSPs that compete with each other in a global market.

Our discussion is based on homogeneous data holders. However, a practical market is more complicated than this. One potential research extension is to take the internal difference among data holders into consideration.

7 Conclusion

In this article, we established the economic model of the cloud storage system with H-DEDU. Based on the economic model, we formulated the multi-stage Stackelberg game to capture the interactions among data holders, data owners, and CSPs for the purpose of investigating H-DEDU adoption and practical deployment conditions. We applied the backward induction method to discover the sub-game perfect Nash Equilibrium in each stage of the Stackelberg game. We further proposed a gradient-based searching algorithm to calculate the near-optimal strategies for all players. Extensive experimental results illustrated the convergence of the formulated game to the Stackelberg Equilibrium. Meanwhile, we also investigated the effects of a number of system parameters on the utilities and strategies of system players at the NE state. Through game-theoretical investigation, we found that H-DEDU is intended to be accepted by the users that hold popular data and have confidence on CSP security.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 61672410, Grant 62072351 and Grant 61802293; in part by the Academy of Finland under Grant 308087, Grant 314203, and Grant 335262; in part by the Shaanxi Innovation Team Project under Grant 2018TD-007; and in part by the 111 Project under Grant B16037.

References

[1] D. T. Meyer and W. J. Bolosky, "A study of practical deduplication," *ACM Transactions on Storage (TOS)*, vol. 7, no. 4, p. 14, 2012.

[2] W. J. Bolosky, S. Corbin, D. Goebel, and J. R. Douceur, "Single instance storage in windows 2000," in *Proceedings of the 4th USENIX Windows Systems Symposium*. Seattle, WA, 2000, pp. 13–24.

[3] D. Bhagwat, K. Eshghi, D. D. Long, and M. Lillibridge, "Extreme binning: Scalable, parallel deduplication for chunk-based file backup," in *2009 IEEE International Symposium on Modeling, Analysis & Simulation of Computer and Telecommunication Systems*. IEEE, 2009, pp. 1–9.

[4] W. Sun, N. Zhang, W. Lou, and Y. T. Hou, "Tapping the potential: Secure chunk-based deduplication of encrypted data for cloud backup," in *2018 IEEE Conference on Communications and Network Security (CNS)*. IEEE, 2018, pp. 1–9.

[5] W. Xia, H. Jiang, D. Feng, F. Douglis, P. Shilane, Y. Hua, M. Fu, Y. Zhang, and Y. Zhou, "A comprehensive study of the past, present, and future of data deduplication," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1681–1710, 2016.

[6] Y. Shin, D. Koo, and J. Hur, "A survey of secure data deduplication schemes for cloud storage systems," *ACM computing surveys (CSUR)*, vol. 49, no. 4, p. 74, 2017.

[7] J. Liu, N. Asokan, and B. Pinkas, "Secure deduplication of encrypted data without additional independent servers," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2015, pp. 874–885.

[8] M. Bellare and S. Keelveedhi, "Interactive message-locked encryption and secure deduplication," in *IACR International Workshop on Public Key Cryptography*. Springer, 2015, pp. 516–538.

[9] J. Li, Y. K. Li, X. Chen, P. P. Lee, and W. Lou, "A hybrid cloud approach for secure authorized deduplication," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 5, pp. 1206–1216, 2014.

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

[11] Z. Yan, W. Ding, X. Yu, H. Zhu, and R. H. Deng, "Deduplication on encrypted big data in cloud," *IEEE Transactions on Big Data*, vol. 2, no. 2, pp. 138–150, Jun. 2016.

[12] D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Side channels in cloud services: Deduplication in cloud storage," *IEEE Security & Privacy*, vol. 8, no. 6, pp. 40–47, 2010.

[13] Y. Shin and K. Kim, "Differentially private client-side data deduplication protocol for cloud storage services," *Security and Communication Networks*, vol. 8, no. 12, pp. 2114–2123, 2015.

[14] M. Li, C. Qin, and P. P. C. Lee, "CDstore: Toward reliable, secure, and cost-efficient cloud storage via convergent dispersal," in *2015 USENIX Annual Technical Conference (USENIX ATC 15)*. Santa Clara, CA: USENIX Association, jul 2015, pp. 111–124.

[15] Z. Yan, M. Wang, Y. Li, and A. V. Vasilakos, "Encrypted data management with deduplication in cloud computing," *IEEE Cloud Computing*, vol. 3, no. 2, pp. 28–35, 2016.

[16] F. Armknecht, J.-M. Bohli, G. O. Karame, and F. Youssef, "Transparent data deduplication in the cloud," in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2015, pp. 886–900.

[17] D. Koo and J. Hur, "Privacy-preserving deduplication of encrypted data with dynamic ownership management in fog computing," *Future Generation Computer Systems*, vol. 78, pp. 739–752, 2018.

[18] 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*. Springer, 2015, pp. 397–402.

[19] M. Miao, T. Jiang, and I. You, "Payment-based incentive mechanism for secure cloud deduplication," *International Journal of Information Management*, vol. 35, no. 3, pp. 379–386, 2015.

[20] X. Liang, Z. Yan, X. Chen, L. T. Yang, W. Lou, and T. Y. Hou, "Game theoretical analysis on encrypted cloud data deduplication," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 10, pp. 5778–5789, October 2019.

[21] R. B. Myerson, *Game theory*. Harvard university press, 2013.

[22] X. Liang and Z. Yan, "A survey on game theoretical methods in human–machine networks," *Future Generation Computer Systems*, vol. 92, pp. 674–693, 2019.

[23] A. A. Kulkarni and U. V. Shanbhag, "An existence result for hierarchical stackelberg v/s stackelberg games," *IEEE Transactions on Automatic Control*, vol. 60, no. 12, pp. 3379–3384, 2015.

[24] Z. Yan, X. Liang, W. Ding, X. Yu, M. Wang, and R. H. Deng, "Encrypted big data deduplication in cloud storage," in *Smart Data: State-of-the-Art Perspectives in Computing and Applications*, K.-C. Li, B. D. Martino, L. T. Yang, and Z. Qingchen, Eds. CRC Press, 2019, ch. 4, pp. 63–92.

[25] X. Jin, L. Wei, M. Yu, N. Yu, and J. Sun, "Anonymous deduplication of encrypted data with proof of ownership in cloud storage," in *2013 IEEE/CIC International Conference on Communications in China (ICCC)*. IEEE, 2013, pp. 224–229.

[26] X. Liang, Z. Yan, and R. H. Deng, "Game theoretical study on client-controlled cloud data deduplication," *Computers & Security*, vol. 91, p. 101730, 2020.

[27] S. Roy, C. Ellis, S. Shiva, D. Dasgupta, V. Shandilya, and Q. Wu, "A survey of game theory as applied to network security," in *2010 43rd Hawaii International Conference on System Sciences*. IEEE, 2010, pp. 1–10.

[28] M. H. Manshaei, Q. Zhu, T. Alpcan, T. Bacşar, and J.-P. Hubaux, "Game theory meets network security and privacy," *ACM Computing Surveys (CSUR)*, vol. 45, no. 3, p. 25, 2013.

[29] I. Butun, S. D. Morgera, and R. Sankar, "A survey of intrusion detection systems in wireless sensor networks," *IEEE communications surveys & tutorials*, vol. 16, no. 1, pp. 266–282, 2013.
[30] X. Liang and Y. Xiao, "Game theory for network security," *IEEE Communications Surveys & Tutorials*, vol. 15, no. 1, pp. 472–486, 2012.
[31] C. T. Do, N. H. Tran, C. Hong, C. A. Kamhoua, K. A. Kwiat, E. Blasch, S. Ren, N. Pissinou, and S. S. Iyengar, "Game theory for cyber security and privacy," *ACM Computing Surveys (CSUR)*, vol. 50, no. 2, p. 30, 2017.
[32] R. Yu, Y. Zhang, S. Gjessing, W. Xia, and K. Yang, "Toward cloud-based vehicular networks with efficient resource management," *IEEE Network*, vol. 27, no. 5, pp. 48–55, Sep. 2013.
[33] G. Nan, Z. Mao, M. Li, Y. Zhang, S. Gjessing, H. Wang, and M. Guizani, "Distributed resource allocation in cloud-based wireless multimedia social networks," *IEEE Network*, vol. 28, no. 4, pp. 74–80, 2014.
[34] D. Niyato, A. V. Vasilakos, and Z. Kun, "Resource and revenue sharing with coalition formation of cloud providers: Game theoretic approach," in *2011 11th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing*. IEEE, 2011, pp. 215–224.
[35] F. Palmieri, L. Buonanno, S. Venticinque, R. Aversa, and B. Di Martino, "A distributed scheduling framework based on selfish autonomous agents for federated cloud environments," *Future Generation Computer Systems*, vol. 29, no. 6, pp. 1461–1472, 2013.
[36] M. Yu and S. H. Hong, "A real-time demand-response algorithm for smart grids: A stackelberg game approach," *IEEE Transactions on Smart Grid*, vol. 7, no. 2, pp. 879–888, 2015.
[37] W. Wei, X. Fan, H. Song, X. Fan, and J. Yang, "Imperfect information dynamic stackelberg game based resource allocation using hidden markov for cloud computing," *IEEE Transactions on Services Computing*, vol. 11, no. 1, pp. 78–89, 2016.
[38] Z. Xiong, S. Feng, D. Niyato, P. Wang, and Y. Zhang, "Economic analysis of network effects on sponsored content: a hierarchical game theoretic approach," in *GLOBECOM 2017-2017 IEEE Global Communications Conference*. IEEE, 2017, pp. 1–6.
[39] N. M. Modak, S. Panda, and S. S. Sana, "Three-echelon supply chain coordination considering duopolistic retailers with perfect quality products," *International Journal of Production Economics*, vol. 182, pp. 564–578, 2016.
[40] S. Halevi, D. Harnik, B. Pinkas, and A. Shulman-Peleg, "Proofs of ownership in remote storage systems," in *Proceedings of the 18th ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM, 2011, p. 491–500.
[41] L. Gao, Z. Yan, and L. T. Yang, "Game theoretical analysis on acceptance of a cloud data access control system based on reputation," *IEEE Transactions on Cloud Computing*, pp. 1–1, 2018.
[42] https://popcon.debian.org/contrib/index.html, accessed June 19, 2018.

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

Zheng Yan received the D.Sc. degree in technology from the Helsinki University of Technology, Espoo, Finland, in 2007. She is currently a Professor in the School of Cyber Engineering, Xidian University, Xi'an, China and a Visiting Professor and Finnish Academy Research Fellow at the Aalto University, Helsinki, Finland. Her research interests are in trust, security, privacy, and security-related data analytics. Dr. Yan is an area editor or an associate Editor of IEEE INTERNET OF THINGS JOURNAL, Information Fusion, Information Sciences, IEEE ACCESS, and Journal of Network and Computer Applications. She served as a General Chair or Program Chair for numerous international conferences, including IEEE TrustCom 2015, IFIP Networking 2021. She is a Founder Steering Committee Co-Chair of IEEE Blockchain conference. She received several awards, including the 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, Director of the Secure Mobile Centre, and Deputy Dean for Faculty & Research, School of Information Systems, Singapore Management University. He has been Professor of Information Systems at SMU since 2004. Prior to that, he was Principal Scientist and Manager of Infocomm Security Department, Institute for Infocomm Research, Singapore. Since April 2019, he has been Programme Director of the National Satellite of Excellence in Mobile Systems Security and Cloud Security, a five-year research initiative sponsored by NRF's National Cybersecurity R&D programme. His research interests are in the areas of data security and privacy, cloud security, network and distributed systems security. He serves/served on many editorial boards and conference committees, including the editorial boards of ACM Transactions on Privacy and Security, IEEE Security and Privacy, IEEE Transactions on Dependable and Secure Computing, IEEE Transactions on Information Forensics and Security, Journal of Computer Science and Technology, and Steering Committee Chair of the ACM Asia Conference on Computer and Communications Security. He received the Outstanding University Researcher Award from National University of Singapore, Lee Kuan Yew Fellowship for Research Excellence from Singapore Management University, Asia-Pacific Information Security Leadership Achievements Community Service Star from International Information Systems Security Certification Consortium in 2010. He is a Fellow of IEEE and Fellow of Academy of Engineering Singapore.

Qinghua Zheng received the Ph.D. degree in System Engineering from the Xi'an Jiaotong University, Xi'an, China. He is currently a distinguished professor in this university. He achieved the National Funds for Distinguished Young Scientists, and is among the first batch of leading scientists of the "Ten-Thousand Talents Project", a candidate of the "New Century National BaiQianWan Talents Project" in China. He is leading the Innovation Team of National Natural Science Foundation of China, the Innovative Team of Ministry of Education, and the Shaanxi Key Scientific and Technological Innovation Team. His main research interests are in intelligent e-learning, big data mining and application, as well as software reliability.