Three-Level Storage and Nested MDS Codes for Perfect Secrecy in Multiple Clouds

Ping Hu and Chi Wan Sung
Department of Electronic Engineering
City University of Hong Kong
Email: hping2@my.cityu.edu.hk, albert.sung@cityu.edu.hk

Siu-Wai Ho and Terence H. Chan
Institute for Telecommunications Research
University of South Australia
Email: {siuwai.ho,terence.chan}@unisa.edu.au

***Abstract*—The problem of storing data reliably and securely in multiple cloud storage providers (CSPs) with minimum cost is investigated. A jointly optimal coding and storage allocation scheme, which achieves perfect secrecy with minimum cost, is derived. The optimal coding scheme is shown to be the nested maximum-distance-separable code and the optimal amounts of data to be stored in the CSPs is proven to exhibit a three-level structure. The exact parameters of the code and the exact storage amount to each CSP can be determined numerically by simple one-dimensional search.**

I. Introduction

Cloud computing enables users to utilize software and hardware resources over the clouds without keeping them on their own computing devices. In particular, cloud data storage is now becoming increasingly popular. Cloud storage providers (CSPs) such as Amazon, Apple, Cisco, IBM, Google and Microsoft offer storage space for their customers. When storing data over the clouds, users are free from the costs of setting up their own servers, electric power charges, space expenses and maintenance costs. Furthermore, they can download their stored data anywhere, provided that the CSP is accessible. In reality, however, CSPs may be unavailable temporarily or even permanently due to various reasons including disk failure, hacker attack, network disconnection, natural disaster, or even political influence. Furthermore, from the users' perspective, their data stored in a CSP is not confidential, since the CSP has full access to its customers' data. To address these issues, users may subscribe services from multiple CSPs and encode their confidential files into several pieces and distribute them to the CSPs.

CSPs generally have different charging schemes. Naturally, a user would like to minimize the total cost of subscription to several CSPs. How to keep confidentiality, ensure data availability and minimize cost at the same time is a challenge. This problem has only be partially investigated in the literature. The work in [1] addressed the first two issues, ensuring computational security in the sense that an eavesdropper has negligibly small probability to access the data within polynomial time. In [2], the authors tackled the same problem subject to the weak security requirement [3], which ensures that an eavesdropper cannot decode any of the original symbols. To the best of our knowledge, the only work that addresses all the three issues simultaneously is [4]. Their approach decouples the problem into two sub-problems. In the first sub-problem, an existing code is adopted and the file is encoded into a pre-determined number of pieces. In the second sub-problem, allocation of the pieces are performed. Since the two sub-problems are solved separately, the solution is in general sub-optimal. In our formulation, the system-level security requirement is that the data to be stored should be confidential under potential collusion of a certain number of CSPs. Besides, instead of adopting existing coding schemes and focusing only on the optimization aspects, we consider the problem as a whole and determine the jointly optimal storage allocation and coding scheme.

Both computational security and weak security allow some information leakage to an eavesdropper. In this paper, we consider perfect secrecy, which requires that no information should be leaked. This notion of security provides the strongest form of privacy and also serves as a theoretic bound for all other weaker forms. In [5], it was shown that coset coding based on linear codes can achieve the secrecy capacity of *wiretap channel II*. Utilizing the similar idea, the authors of [6] showed that nested MDS codes can achieve the secrecy capacity of erasure-erasure wiretap channel, which is a generalization of wiretap channel II. In [7], the authors developed some schemes that can achieve the secure capacity of distributed storage systems under dynamic repair.

Our work is also related to the well-known problem called *secret sharing* [8], in which there is a secret to be shared among a group of people. When the number of gathered people is larger than or equal to a threshold t, they can reveal the secret. No information is disclosed when there are fewer than t people. This corresponds to a special case of our problem.

To summarize, we consider storage in multiple CSPs with availability, perfect secrecy and minimum cost from the user's viewpoint. Our main results are as follows:

- *Optimal Coding Scheme*: We obtain a lower bound on the cost for keeping availability and perfect security. A coding scheme which achieves the bound is identified.
- *Three-level Storage*: The optimal storage amounts of the coded data on the CSPs are proven to have three levels at most, regardless of the differences in charging schemes.
- *Low Complexity*: The problem was simplified from an optimization problem with exponential number

This work was partially supported by a grant from the University Grants Committee of the Hong Kong Special Administrative Region, China (Project No. AoE/E-02/08).

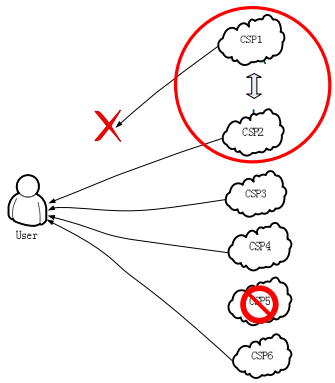

Fig. 1. An example when $N = 6$, $K = 4$ and $T = 2$.

of constraints to a problem solvable by fast one-dimensional search.

II. System Model

Our model includes one customer and N cloud storage providers (CSPs). The customer stores its file into the N CSPs. When the customer wants to get the file, it should be able to retrieve it from these CSPs. However, it cannot be guaranteed that all the CSPs are available all the time. We require that the customer can get the file back by connecting to any $K (K \leq N)$ CSPs. Besides, the customer faces a privacy threat if some CSPs collude together to access the file. The customer does not want any CSP to get any information about its file. Assuming the potential number of colluding CSPs is $T < K$, it should be required that no information of the stored file is exposed against the T colluding CSPs. An example of the model is shown in Fig. 1.

Suppose the file consists of B blocks of data, denoted by a vector $\boldsymbol{x} \triangleq (x_1, x_2, \ldots, x_B)$. Each block x_i is a symbol drawn uniformly at random from the finite field $\mathbb{F}$ of size 2^L. Therefore, the file size, or equivalently, the entropy of $\boldsymbol{x}$, $H(\boldsymbol{x})$, is equal to BL bits.

An encoding function $f : \mathbb{F}^B \to \mathbb{F}^n$ maps $\boldsymbol{x}$ into $\boldsymbol{y} \triangleq (\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_N)$, where $\boldsymbol{y}_i \triangleq (y_{i,1}, y_{i,2}, \ldots, y_{i,n_i})$ for $i = 1, 2, \ldots, N$, and $\sum_{i=1}^{N} n_i = n$. Note that $\boldsymbol{y}_i$ is the vector of coded blocks stored in CSP i, and the number of coded blocks stored in CSP i is n_i. In the special case where $n_i = 0$, the vector $\boldsymbol{y}_i$ degenerates into an empty vector, which means that CSP i does not store any data. The cost of storing a unit block of data in CSP i is denoted by c_i. The total storage cost is given by

$$C = \sum_{i=1}^{N} c_i n_i. \tag{1}$$

For any $\mathcal{S} \subseteq \mathcal{N} \triangleq \{1, 2, ..., N\}$, define $\boldsymbol{y}(\mathcal{S})$ as the sub-vector of $\boldsymbol{y}$ obtained by retaining only $\boldsymbol{y}_i$ for $i \in \mathcal{S}$. To fulfill the *reconstruction* requirement that the customer can reconstruct its file from "any K out of N" CSPs, for any $\mathcal{S} \subset \mathcal{N}$ of cardinality K, there should exist a decoding function $g_{\mathcal{S}}$, such that $g_{\mathcal{S}}(\boldsymbol{y}(\mathcal{S})) = \boldsymbol{x}$. This requirement can also be expressed as

$$H(\boldsymbol{x}|\boldsymbol{y}(\mathcal{S})) = 0, \ \forall \mathcal{S} \subset_K \mathcal{N}, \tag{2}$$

where we have used the shorthand notation $\mathcal{A} \subset_k \mathcal{B}$ to mean that $\mathcal{A}$ is a k-subset of $\mathcal{B}$.

To ensure that the file is perfectly secure against any T colluding CSPs, we must have

$$H(\boldsymbol{x}|\boldsymbol{y}(\mathcal{T})) = H(\boldsymbol{x}), \ \forall \mathcal{T} \subset_T \mathcal{N}. \tag{3}$$

We call the above requirement the *perfect secrecy* requirement.

The problem is to determine f and the vector $(n_1, n_2, \ldots, n_N)$ so as to minimize the total storage cost, C, in (1) with respect to the reconstruction requirement (2) and perfect secrecy requirement (3). Note that the code length, n, is a variable depending on the values of n_i's.

III. Lower Bound on the Minimum Cost

In this section, we derive a lower bound on the minimum cost by elementary manipulations of Shannon's information measures.

Theorem 1. *The cost C is bounded below by*

$$C_{LB} \triangleq \min \sum_{i=1}^{N} c_i n_i, \tag{4}$$

where n_i's are subject to

$$\min_{\mathcal{S} \subset_K \mathcal{N}, \mathcal{T} \subset_T \mathcal{S} \subset_K \mathcal{N}} \left(\sum_{i \in \mathcal{S}} n_i - \sum_{i \in \mathcal{T}} n_i \right) \geq B. \tag{5}$$

Proof: Consider an arbitrary $\mathcal{S} \subset_K \mathcal{N}$ and an arbitrary $\mathcal{T} \subset_T \mathcal{N}$. From the perfect secrecy requirement (3) and the reconstruction requirement (2), we have

$$H(\boldsymbol{x}) = H(\boldsymbol{x}|\boldsymbol{y}(\mathcal{T})) - H(\boldsymbol{x}|\boldsymbol{y}(\mathcal{S})).$$

Denote $\mathcal{I} \triangleq \mathcal{T} \cap \mathcal{S}$. We can then rewrite the above equation as

$$H(\boldsymbol{x}) = H(\boldsymbol{x}|\boldsymbol{y}(\mathcal{I}), \boldsymbol{y}(\mathcal{T} \backslash \mathcal{I})) - H(\boldsymbol{x}|\boldsymbol{y}(\mathcal{I}), \boldsymbol{y}(\mathcal{S} \backslash \mathcal{I})).$$

Take elementary manipulations of Shannon's information measures, we have

$$\begin{aligned} H(\boldsymbol{x}) &= I(\boldsymbol{x}; \boldsymbol{y}(\mathcal{S} \backslash \mathcal{I}) | \boldsymbol{y}(\mathcal{I})) - I(\boldsymbol{x}; \boldsymbol{y}(\mathcal{T} \backslash \mathcal{I}) | \boldsymbol{y}(\mathcal{I})) \\ &\leq I(\boldsymbol{x}; \boldsymbol{y}(\mathcal{S} \backslash \mathcal{I}) | \boldsymbol{y}(\mathcal{I})) \\ &\leq H(\boldsymbol{y}(\mathcal{S} \backslash \mathcal{I}) | \boldsymbol{y}(\mathcal{I})) \\ &\leq H(\boldsymbol{y}(\mathcal{S} \backslash \mathcal{I})) \\ &\leq \sum_{i \in \mathcal{S} \backslash \mathcal{I}} H(\boldsymbol{y}_i) \\ &\leq L \sum_{i \in \mathcal{S} \backslash \mathcal{I}} n_i \\ &= L \left(\sum_{i \in \mathcal{S}} n_i - \sum_{i \in \mathcal{I}} n_i \right) \end{aligned} \tag{6}$$

The above bound on $H(\boldsymbol{x})$ holds for any $\mathcal{S} \subset_K \mathcal{N}$ and any $\mathcal{T} \subset_T \mathcal{N}$. The tightest ones are the cases where $\mathcal{T}$ happens to be a subset of $\mathcal{S}$. Therefore, (6) can be re-written as

$$H(\boldsymbol{x}) \leq L \min_{\mathcal{S}\subset_K\mathcal{N}, \mathcal{T}\subset_T\mathcal{S}\subset_K\mathcal{N}} (\sum_{i\in\mathcal{S}} n_i - \sum_{i\in\mathcal{T}} n_i).$$

As $H(\boldsymbol{x}) = BL$, we have shown that (5) is a necessary condition to be met. Therefore, if we minimize the total cost subject to (5), the result so obtained becomes a lower bound for our problem. ■

IV. A Coding Scheme with Perfect Secrecy

In this section, we construct a coding scheme based on the *nested maximum-distance-separable (MDS) code*. Recall that an (n, k) MDS code is a class of linear codes matching the Singleton bound, i.e. the minimum Hamming distance of the code, $d_{\min}$, equals $n-k+1$. The generator matrix of an (n, k) MDS code has the property that all its $k \times k$ submatrices are full rank. A coding scheme based on the cosets of MDS codes was shown to achieve the secrecy capacity of *wiretap channel II* [5]. In wiretap channel II, k data bits are encoded into $n > k$ bits and transmitted to the legitimate receiver without loss. An eavesdropper can observe an arbitrary subset of μ bits. A modified version of wiretap channel II is the *erasure-erasure wiretap channel* introduced in [6], in which the bits transmitted to legitimate receiver also experience erasures. In an erasure-erasure wiretap channel with parameters (θ, M, μ), the transmitter sends out θ symbols and the legitimate receiver and the eavesdropper receive M and μ symbols respectively. The maximum amount of data that can be sent without any leakage of information to the eavesdropper is called the *secrecy capacity* of the channel. The secrecy capacity of the erasure-erasure wiretap channel can be achieved by *nested MDS code* [6][7]. We present the definition of nested MDS code following the lines in [7]:

Definition 1. *If the generator matrix of an (n, k) MDS code G can be written as $G = \begin{bmatrix} G_1 \\ G_2 \end{bmatrix}$, where G_1 itself is a generator matrix of an (n, k_0) MDS code, then we refer to the code defined by G as nested MDS code with parameters (n, k, k_0).*

According to [6] and [7], the secrecy capacity of the erasure-erasure wiretap channel with parameters (θ, M, μ) is equal to $M-\mu$, which can be achieved by a nested MDS code with parameters (θ, M, μ). Based on the above observation, we obtain the following achievability result.

Theorem 2. *C^* as defined below is achievable:*

$$C^* \triangleq \min \sum_{i=1}^{N} c_i n_i, \tag{7}$$

where n_i's are subject to

$$\min_{\mathcal{S}\subset_K\mathcal{N}, \mathcal{T}\subset_T\mathcal{N}} (\sum_{i\in\mathcal{S}} n_i - \sum_{i\in\mathcal{T}} n_i) \geq B. \tag{8}$$

Proof: Let $(n_1^*, n_2^*, \ldots, n_N^*)$ be the solution of the minimization problem (7) and $n^* = \sum_{i=1}^{N} n_i^*$. Define

$$\nu_{\min}^* \triangleq \min_{\mathcal{S}\subset_K\mathcal{N}} \sum_{\mathcal{S}} n_i^*, \quad \text{and} \quad \mu_{\max}^* \triangleq \max_{\mathcal{T}\subset_T\mathcal{N}} \sum_{\mathcal{T}} n_i^*.$$

The user then generates a random key $\boldsymbol{k} = (k_1, k_2, \ldots, k_Z)$ of length $Z = \nu_{\min}^* - B$, in which the k_i's are uniformly and independently distributed over the finite field $\mathbb{F}_{2^L}$. The encoder output is defined by

$$\boldsymbol{y} = [\ \boldsymbol{k} \quad \boldsymbol{x}\ ] \begin{bmatrix} G_k \\ G_x \end{bmatrix}, \tag{9}$$

where $G \triangleq \begin{bmatrix} G_k \\ G_x \end{bmatrix}$ is the generation matrix of a nested MDS code with parameter $(n^*, \nu_{\min}^*, \mu_{\max}^*)$.

Since the user is supposed to retrieve the file through any K CSPs, the channel can be seen as a collection of erasure-erasure wiretap channels with parameters (n^*, ν^*, μ^*), where

$$\nu^* \in \Big\{ \sum_{i\in\mathcal{S}} n_i^* : \mathcal{S} \subset_K \mathcal{N} \Big\} \quad \text{and} \quad \mu^* \in \Big\{ \sum_{i\in\mathcal{T}} n_i^* : \mathcal{T} \subset_T \mathcal{N} \Big\}.$$

Among all these channels, the worst one is the case where the user receives the least number of blocks while the eavesdropper observes the most number of blocks, that is, the channel with parameters $(n^*, \nu_{\min}^*, \mu_{\max}^*)$. If the secrecy capacity of this channel is larger than B, i.e.,

$$\nu_{\min}^* - \mu_{\max}^* \geq B, \tag{10}$$

then the code generated by (9) can store B blocks of data with perfect secrecy. It is obvious to see that constraints (8) and (10) are equivalent. ■

Note that nested MDS code can be obtained from a $\nu_{\min}^* \times n^*$ Vandermonde matrix $V_{i,j} = \alpha_i^{j-1}$ where $i = 1, \ldots, \nu_{\min}^*$, $j = 1, \ldots, n^*$, and all α_i's shall be distinct. Now we can see that the file which has B blocks can be securely stored if (8) is satisfied and the minimum cost for this coding method is exactly the value of C^* in Theorem 2.

V. Optimal coding scheme and three-level storage

In this section, we will show that the coding scheme in Section IV is optimal and prove that the optimal solution to the minimum cost problem has a three-level structure. From now on, we assume that the CSPs are labeled in a way such that $c_1 \leq c_2 \leq \ldots \leq c_N$. We denote an optimal solution to the storage cost bound in Theorem 1 by $(n_1^{LB}, n_2^{LB}, \ldots, n_N^{LB})$ and the optimal solution to the storage cost minimization problem in Theorem 2 by $(n_1^*, n_2^*, \ldots, n_N^*)$.

Definition 2. *Let $m \triangleq N - K + T + 1$.*

Note that m satisfies $1 \leq m \leq N$ and thus represents the index of one of the N CSPs. As we are going to prove, the optimal solution to the minimum cost problem has a three-level structure, and m plays the role of a lower bound on the number of CSPs that have to store the amount that attains the top level.

Lemma 3. *The storage cost minimization problem in Theorem 2 can be simplified into*

$$C^* = \min \sum_{i=1}^{N} c_i n_i,$$

where n_i's are subject to

$$n_1 \geq n_2 \geq ... \geq n_N \geq 0, \tag{11}$$

and

$$(n_{N-K+1} + ... + n_N) - (n_1 + ... + n_T) \geq B. \tag{12}$$

Proof: We claim that $(n_1^*, n_2^*, \ldots, n_N^*)$ must satisfy (11). Assume that $n_i^* < n_j^*$ for some $i < j$. If we swap the values of n_i^* and n_j^*, the feasible region remains unchanged because (8) is symmetric in n_i's. However, the cost in (7) will increase, violating that n_i^*'s are optimal. Therefore, (11) must hold, meaning that adding (11) as an additional constraint does not change the optimal solution. But with (11) added, the family of constraints in (8) can be simplified to the most stringent one, that is, the one in (12). ■

Lemma 4. *The storage cost minimization problem in Theorem 2 can be simplified into*

$$C^* = \min \sum_{i=1}^{N} c_i n_i,$$

where n_i's are subject to

$$n_1 = n_2 = \cdots = n_m. \tag{13}$$

Furthermore, $n_m + n_{m+1} \cdots + n_N = B$.

Proof: First of all, note that $m > \max\{N-K+1, T\} \triangleq j_{\max}$. We must have $n_1^* = n_2^* = \cdots = n_T^*$. To see it, suppose on the contrary that $n_i > n_{i+1}$ for some $i < T$. We can then reduce the value of n_i without violating (12), thus yielding a smaller storage cost. Now suppose there exists r, where $T \leq r < m$, such that $n_r^* > n_{r+1}^*$. Let $\delta \triangleq n_r^* - n_{r+1}^*$, $j_{\min} \triangleq \min\{N-K+1, T\}$, and $g \triangleq r - j_{\max}$. Note that $n_{j_{\max}}, n_{j_{\max}+1}, \ldots, n_r$ always fall in the first bracket of (12) and $n_{j_{\min}-g}, n_{j_{\min}-g+1}, \ldots, n_{j_{\min}}$ always fall in the second bracket of (12). We can subtract δ from each of $n_{j_{\min}-g}, n_{j_{\min}-g+1}, \ldots, n_r$ without changing the value of the expression in the left hand side of (12). Since this subtraction can reduce the storage cost, we conclude that such a r does not exist and (13) must hold. With (13), it can be seen that (12) is equivalent to

$$n_m^* + n_{m+1}^* \cdots + n_N^* \geq B.$$

Equality must hold for otherwise we can reduce the value of n_j^* to reduce the storage cost for some $j \geq m$. ■

Theorem 5. *The coding scheme in Section IV is optimal and C^* is the minimum cost.*

Proof: Following similar steps in the proof of Lemma 3, we can show that $n_1^{LB} \geq n_2^{LB} \geq ... \geq n_N^{LB} \geq 0$. With (5) and following similar steps in the proof of Lemma 4, we can show that $n_m^{LB} + n_{m+1}^{LB} \cdots + n_N^{LB} = B$. The bound in Theorem 1 can thus be simplified into

$$C_{LB} = \min \sum_{i=1}^{N} c_i n_i,$$

where n_i's are subject to $n_1 = n_2 = \cdots = n_m$, and $n_m + n_{m+1} \cdots + n_N = B$. Lemma 4 implies $C^* = C_{LB}$, thus completing the proof. ■

Now we proceed to show the three-level structure of the optimal solution and further simplify the optimization problem.

Lemma 6. *An optimal solution must be in the form*

$$n_m^* = n_{m+1}^* = \cdots = n_p^* > n_{p+1}^* = B - n_m^*(p-m+1) \geq 0 \tag{14}$$

and

$$n_{p+2}^* = \cdots = n_N^* = 0, \tag{15}$$

where $m \leq p \leq N$.

Proof: With Lemma 4, the storage cost minimization can be re-written as

$$C^* = \min \left[d_m n_m + \sum_{j=m+1}^{N} c_j n_j \right]$$

subject to

$$n_m \geq n_{m+1} \geq \cdots \geq n_N \geq 0, \tag{16}$$

$$n_m + n_{m+1} + \cdots n_N = B, \tag{17}$$

where $d_m \triangleq \sum_{j=1}^{m} c_j$.

Suppose n_m^* is known. Then the constraints (16)(17) become

$$n_m^* \geq n_{m+1} \geq \cdots \geq n_N \geq 0,$$
$$n_{m+1} + \cdots n_N = B - n_m^*.$$

Since the storage costs are sorted in increasing order, it is clear that one would put more coded blocks of data to storage nodes of smaller indices. As a result, we must have

$$n_m^* = n_{m+1}^* = n_{m+2}^* = \cdots = n_p^*$$

and

$$n_{p+1}^* = B - n_m^*(p-m+1) \geq 0.$$

■

From the above lemma, we can see that the allocation of the blocks has at most three levels, n_m^*, n_{p+1}^* and 0. If the storage amount of a particular CSP is equal to 0, it means that the user does not need to subscribe to the service provided by that CSP. Suppose $\phi \triangleq N - p - 1$ of the CSPs are not in use. The file can be retrieved from any $K' \triangleq K - \phi$ out of $N' \triangleq N - \phi$ CSPs. If each CSP fails independently with the same probability, it can be proved that the reliability of the solution yielded is actually better than the original any-K-out-of-N requirement.

Now it remains to determine the values of p and n_m^*. We are going to show that it can be done by a one-dimensional search over all possible values of p. Before presenting the result, we define the following: For $m \leq p < N$, define $n_{p,\max} \triangleq \lfloor \frac{B}{p-m+1} \rfloor$, $n_{p,\min} \triangleq \lceil \frac{B}{p-m+2} \rceil$, $d_p \triangleq \sum_{j=1}^{p} c_j$ and $d'_p \triangleq c_{p+1}(p-m+1)$. Furthermore, define $I(u) \triangleq u$ if u is an integer. Otherwise, define $I(u)$ to be ∞, assuming that we are now working in the extended real number system.

Theorem 7. *An optimal solution to the storage cost minimization problem is given by the following:*

$$n_1^* = n_2^* = \cdots = n_m^* = \cdots = n_p^*,$$
$$n_{p+1}^* = B - n_m^*(p-m+1),$$
$$n_{p+2}^* = \cdots = n_N^* = 0,$$

where

$$n_m^* = \begin{cases} I(\frac{B}{K-T}), & p = N \\ n_{p,\max}, & m \le p < N,\ d_p \ge d'_p \\ n_{p,\min}, & m \le p < N,\ d_p < d'_p \end{cases} \quad (18)$$

and p is an integer satisfying $m \le p \le N$ that minimizes

$$C(p) = \begin{cases} d_p n_m^*, & p = N \\ c_{p+1}B + n_m^*(d_p - d'_p), & m \le p < N \end{cases} \quad (19)$$

Proof: First, note that $m \le p \le N$. According to (4), p can possibly be equal to N only if B is an integer multiple of $(K-T)$. In that case, if $p = N$, then $n_m^* = B/(K-T)$ and the corresponding storage cost equals $d_p n_m^*$.

Next, consider the case where $m \le p < N$. From (14), we obtain

$$n_m^* > B - n_m^*(p - m + 1) \ge 0,$$

which implies that $n_{p,\min} \le n_m^* \le n_{p,\max}$. Using the results in Lemmas 4 and 6, the storage cost minimization problem is equivalent to finding p so as to minimize

$$C(p) = d_p n_m^* + c_{p+1}(B - n_m^*(p - m + 1)),$$

which can be rewritten as (19). To minimize the value of $C(p)$ for $m \le p < N$, it is easy to see that the value of n_m^* should be the one given in (18). ■

To solve the storage cost minimization problem, it suffices to determine the value of p. We can find the minimum value of $C(p)$ through a linear search through $N - m + 1 = K - T$ values, with time complexity of $O((K-T))$. To apply the above result, we need to first sort c_i's in increasing order, with time complexity of $O(N \log N)$. Therefore, the whole problem can be solved with time complexity of $O(N \log N)$.

To minimize the cost, the encoded block length is optimized subject to the reconstruction requirement and the secrecy requirement. After the optimal encoded block length is determined, we should allocate the encoded blocks into the N CSPs. The solution has at most three levels. The highest level is the amount of data to be stored on the p cheapest CSPs. The second level, which is the amount of data stored on the $(p+1)$-th cheapest CSP, is actually an artifact due to the indivisibility of the encoded block length by p. The third level is 0, meaning that data should not be stored in expensive CSPs. Because of the "any K out of N" reconstruction requirement and the "any T out of N" security requirement, equal allocation of coded data to cheaper CSPs, as shown in Theorem 7, is intuitively reasonable. The problem is then reduced to a one-dimensional search for the value of p.

VI. Numerical Examples

In this section, we give examples on the distribution of storage. We assume there are $N = 10$ CSPs and the charge of each CSP is listed in the second column of Table I. The customer wants to store $B = 50$ blocks of data. The user require that the data retrieval can be done through any $K = 7$ CSPs, which means that the data is reliable even $N - K = 3$ CSPs are disconnected. Table I lists the data storage on each CSP when the number of colluding CSPs increases from $T = 1$ to $T = 4$. When $T = 1$, we use nested MDS code with parameters $(118, 67, 17)$ to encode. CSP 8 to 10 are not used and this code and its distribution is sufficient for file reconstruction through any $K' = 4$ CSPs out of the $N' = 7$ CSPs that are in use. When $T = 2$, we adopt nested MDS code with parameters $(130, 82, 32)$ to encode. When $T = 3$, we adopt nested MDS code with parameters $(128, 89, 39)$ to encode. The code length is shorter than the case $T = 2$ but the total cost is higher. We can see that the code length varies with different scenarios. Normally, to minimize the storage cost, CSPs with higher price will store nothing. However, when T becomes large, all the CSPs should be used. This is because when we want to keep privacy against more colluding CSPs, we should disperse the message into more blocks and store them in a more scattering way.

TABLE I. Storage Distribution

CSP	c_i	$n_i(T=1)$	$n_i(T=2)$	$n_i(T=3)$	$n_i(T=4)$
1	10	17	16	13	17
2	23	17	16	13	17
3	44	17	16	13	17
4	85	17	16	13	17
5	100	17	16	13	17
6	140	17	16	13	17
7	160	16	16	13	17
8	210	0	16	13	17
9	260	0	2	13	17
10	300	0	0	11	16

VII. Conclusion

In this paper, we investigate the minimization of storage cost when the user stores its data in multiple untrustful and unreliable clouds. We give a lower bound on the cost and present a coding scheme that can achieve this bound. This optimal scheme can be solved through one-dimensional search, which has very low computational complexity. Given $c_1 < c_2 < \cdots < c_N$ and large block size B, one may incline to think that the optimal storage allocation would be $n_1^* > n_2^* > \cdots > n_N^*$. Somewhat surprisingly, we show the optimal storage amounts exhibit a three-level structure.

References

[1] H. Lin and W. Tzeng, "A secure erasure code-based cloud storage system with secure data forwarding," *IEEE Trans. Parallel Distrib. Syst.*, vol. 23, no. 6, pp. 995–1003, Jun. 2012.

[2] P. Oliveira, L. Lima, T. Vinhoza, J. Barros, and M. Médard, "Coding for trusted storage in untrusted networks," *IEEE Trans. on Information Forensics and Security*, vol. 7, no. 6, pp. 1890–1898, Dec. 2012.

[3] K. Bhattad and K. Narayanan, "Weakly secure network coding," in *Proc. of the First Workshop on Network Coding, Theory, and Applications (NetCod)*, Apr. 2005.

[4] Y. Singh, F. Kandah, and W. Zhang, "A secured cost-effective multi-cloud storage in cloud computing," in *2011 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Apr. 2011.

[5] L. H. Ozarow and A. D. Wyner, "Wire-tap channel-II," *AT&T Bell Lab Tech. J.*, vol. 63, no. 10, pp. 2135–2157, 1984.

[6] A. Subramanian and S. McLaughlin, "MDS codes on erasure-erasure wire-tap channel," *Available:arXiv:0902.3286v1*, 2009.

[7] S. Pawer, S. E. Rouayheb, and K. Ramchandran, "Securing dynamic distributed storage systems against eavesdropping and adversarial attacks," *IEEE Trans. Inf. Theory*, vol. 57, no. 10, pp. 6734–6752, Oct. 2011.

[8] A. Shamir, "How to share a secret," *Commun. ACM*, vol. 22, no. 11, pp. 612–613, Nov. 1979.