

Secure Convertible Codes for Passive Eavesdroppers

Justin Zhang*

Department of Computer Science
Purdue University, West Lafayette, IN
Email: zhan3554@purdue.edu

K.V. Rashmi

Computer Science Department
Carnegie Mellon University, Pittsburgh, PA
Email: rvinayak@cs.cmu.edu

Abstract—Large-scale distributed storage systems rely on erasure codes to ensure fault tolerance against node failures. Due to the observed changing failure rates within these systems, code redundancy tuning, or *code conversion* has been shown to reduce storage cost. Previous work has developed theoretical bounds and constructions for *convertible codes*, a specialized class of erasure codes optimizing either access or bandwidth costs during conversion.

In this paper, we address the challenge of securing convertible codes in the presence of an eavesdropper. We introduce an eavesdropper-secrecy model for convertible codes wherein an eavesdropper gaining read access to a subset of the codeword symbols learns nothing (information-theoretically) about the underlying message. We then focus on access-cost optimal convertible codes, and we then derive the information-theoretic upper bound on the number of message symbols that can be stored securely. Finally, we provide an explicit construction that simultaneously reaches this secrecy bound while admitting access-cost optimal conversion using concatenation of nested codes with traditional convertible codes. Since our construction works with all traditional access-optimal convertible codes, we show that access-optimal secure convertible codes exist for all message and codeword length parameters.

I. INTRODUCTION

Erasure codes provide a low-storage overhead solution to ensure fault tolerance against node failures in large-scale distributed storage systems [1], [2]. In this approach, the data is divided into k message symbols, which are then encoded using an $[n, k]$ erasure code into n coded symbols, forming a *codeword* using an $[n, k]$ erasure code. These codewords are distributed across n different nodes in the storage system. To achieve optimal storage efficiency and fault tolerance, Maximum Distance Separable (MDS) codes are typically employed. Informally, the MDS property ensures data integrity by allowing recovery of the original data even if up to $(n - k)$ nodes fail. In other words, any k out of the n codeword symbols are sufficient to decode the original data.

The parameters n and k are selected based on the observed node failure rates, which, as shown by Kadekodi et al., can vary over time [3]. During periods of high failure rates, n and k are configured to achieve a high redundancy ratio $\frac{n}{k}$, ensuring greater fault tolerance at the expense of increased storage overhead. Conversely, during periods of low failure rates, a lower redundancy ratio is preferred, reducing storage overhead.

*This work was completed while the author was a student at Carnegie Mellon University.

This work was supported in part by the NSF CAREER Award under Grant 19434090 and in part by a Sloan Faculty Fellowship.

However, changing the parameters n and k on already encoded data using the conventional approach—decoding the data from the initial code and re-encoding it with a new code—incur significant costs in terms of I/O, and network bandwidth [4].

This problem has been formalized under the theoretical framework of *code conversion* [4], which defines the conversion of data from an initial code \mathbb{C}^I with parameters $[n^I, k^I]$ to a final code \mathbb{C}^F with parameters $[n^F, k^F]$. *Convertible codes* [4] are a class of codes that by design minimize the costs of code conversion, while maintaining certain decodability guarantees (such as the MDS property) in both the initial and final codes. Convertible codes have been studied primarily in terms of minimizing conversion costs, with two key cost metrics: access cost [4], [5], which measures the number of symbols accessed during conversion, and bandwidth cost [6], [7], which measures the amount of information downloaded. Access-optimal convertible codes are known for all parameter settings, while bandwidth-optimal convertible codes have been developed for certain parameter regimes.

In this paper, we consider the problem of information-theoretic security of convertible codes. Specifically, we investigate security against passive *eavesdroppers* who gain read access to some of codeword symbols stored in the system and try to learn information about the message symbols. This problem setting has been inspired by several prior works on information-theoretic security in distributed storage codes under various models, such as secure regenerating codes [8]. We first introduce a secrecy model for convertible codes, incorporating requirements for data decoding, code conversion, and eavesdropper secrecy. For a specified security parameter ℓ , the objective is to ensure that an eavesdropper who reads any ℓ code symbols of a convertible code learns no information about the message symbols.

We then focus on access-optimal convertible codes and establish an upper bound on the number of message symbols that can be securely stored using convertible codes using an information-theoretic approach. Finally, we present an explicit construction of an *access-optimal* secure convertible code that achieves this upper bound for all parameter settings. The proposed construction uses code concatenation of nested codes [9] with traditional convertible codes [4], [5].

The outline of the paper is as follows. Section II presents the secrecy model for convertible codes. Section III proves the secrecy capacity of any secure convertible code. Section IV presents a construction of access-optimal secure convertible

codes that reach secrecy capacity for all parameters. Lastly, the paper concludes with a discussion in Section V.

A. Notations Used

This section introduces notation used throughout the paper. Calligraphic, uppercase letters \mathbb{T} denote sets. Bold lowercase letters will denote vectors, e.g. a n -length vector, $\mathbf{x} \in \mathbb{F}^n$, where \mathbb{F} is a finite field. When relevant, we denote \mathbb{F}_q as the finite field of size q . The i 'th symbol of a vector \mathbf{x} is written (non-bold) as x_i . A vector subscripted with a set, e.g. \mathbf{x}_S , denote the projection of the vector to each coordinate in the set \mathbf{x} e.g. $\mathbf{x}_S = [x_i : i \in S]$. Uppercase letters denote matrices, e.g. a matrix of size $k \times n$, $G \in \mathbb{F}^{k \times n}$, while Calligraphic letters \mathcal{C} will denote codes. For any vector \mathbf{x} , its corresponding random variable is denoted as \mathcal{X} (uppercase, calligraphic, and bold). Let $[i] = \{1, 2, \dots, i\}$. Let $\Pi(i)$ be the set of all partitions of $[i]$. Lastly, let H be the entropy function (in base $|\mathbb{F}|$).

II. SECRECY MODEL FOR CODE CONVERSIONS

This section presents the eavesdropper threat model for distributed storage systems that employ convertible codes. The formal definition of a convertible code is first provided to establish intuition and motivation for the necessary modifications to accommodate eavesdroppers. The definition of a secure convertible code is then introduced, encompassing both the original properties of convertible codes and the added requirement of eavesdropper security.

The traditional convertible codes framework captures the conversion between an initial and a final *configuration* of stored data [4]. In the initial configuration, data is encoded using an (n^I, k^I) code \mathcal{C}^I , while in the final configuration, the same data is encoded using an (n^F, k^F) code \mathcal{C}^F . Non-trivial conversion occurs when $k^I \neq k^F$, allowing multiple codewords in both configurations. Let $\mathbf{m} \in \mathbb{F}^k$ be the message symbols to be stored, where $k = \text{lcm}(k^I, k^F)$. The initial configuration contains $\lambda^I = k/k^I$ codewords, and the final configuration contains $\lambda^F = k/k^F$ codewords. The message symbols in each codeword is determined by the initial and final partitions \mathbb{P}^I and \mathbb{P}^F of $[k]$. A conversion procedure is then defined to transform the initial configuration into the final one. The access cost of conversion is measured by the number of codeword symbols used by the conversion procedure.

More formally,

Definition 1 (Convertible Code [4]): A $[n^I, k^I; n^F, k^F]$ convertible code is defined by:

- 1) A pair of codes $(\mathcal{C}^I, \mathcal{C}^F)$ where \mathcal{C}^I is a (n^I, k^I) code over \mathbb{F} and \mathcal{C}^F is a (n^F, k^F) code.
- 2) A pair of partitions $(\mathbb{P}^I, \mathbb{P}^F) \in \Pi(k)$ where each subset $\mathbb{P}_i^I \in \mathbb{P}^I$ has size k^I , and each $\mathbb{P}_j^F \in \mathbb{P}^F$ has size k^F .
- 3) A conversion procedure that takes initial codewords $\{\mathcal{C}^I(\mathbf{m}_{\mathbb{P}_i^I}) : \mathbb{P}_i^I \in \mathbb{P}^I\}$ to final codewords $\{\mathcal{C}^F(\mathbf{m}_{\mathbb{P}_j^F}) : \mathbb{P}_j^F \in \mathbb{P}^F\}$.

A convertible code is MDS if the initial and final code are both MDS. Similarly, a convertible code is linear if the initial and final code are both linear.

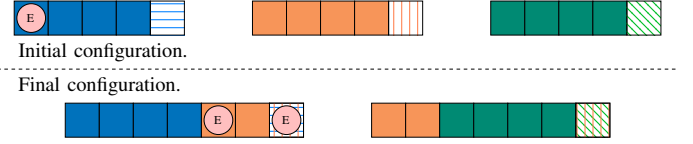


Fig. 1. A $[5, 4; 7, 6]$ convertible code with 3 symbols read by the eavesdropper (pink circles). Initial codewords are on the top of the diagram while the final codewords are on the bottom of the diagram. The initial/final codewords make up the initial/final configuration. Note that an eavesdropper can choose to read symbols across different initial and final configurations, or they can read all the symbols from a single codeword.

We define secure convertible codes by enhancing the existing convertible code framework with protection against eavesdroppers. A message $\mathbf{m} \in \mathbb{F}^k$ is stored on the convertible code, resulting in an initial/final configuration comprising of initial/final codewords. Now, suppose that an eavesdropper gains read-access to any $\ell < \min\{k^I, k^F\}$ codeword symbols, spanning across initial and final configuration, as illustrated in Figure 1.

Note that the eavesdropper may choose to read some out of the ℓ compromised symbols from the initial configuration and/or wait for the conversion to occur to choose the remaining compromised symbols from the final configuration. Also, note that this secrecy model captures the special case when the codeword symbols downloaded during the conversion process are compromised, in the access-cost setting. This scenario is identical to the case of eavesdropper reading the corresponding codeword symbols in the initial configuration.

We introduce additional notation to formally define the desired properties of secure convertible codes under passive eavesdroppers. Let $\mathbf{s} \in \mathbb{F}^{k_S}$ be the message symbols to be securely stored for some $k_S \in \mathbb{N}$. Let \mathcal{S} be the corresponding random variable, which is assumed to be uniformly distributed over \mathbb{F}^{k_S} representing (incompressible) data. Hence $H(\mathcal{S}) = k_S$.

Next, we introduce notation to specify the partition of the secure message symbols into the initial and final codewords. Let $\mathcal{S}^I, \mathcal{S}^F \in \Pi(k_S)$, where $|\mathcal{S}^I| = \lambda^I$ and $|\mathcal{S}^F| = \lambda^F$ denote *secure symbol partitions* that specify the mapping of secure message symbols into the initial and final codewords. Likewise, for $i \in [\lambda^I]$ let \mathcal{S}_i^I be the random variable corresponding to the secure message symbols of i 'th initial codeword, and for $j \in [\lambda^F]$, let \mathcal{S}_j^F be the random variables corresponding to the secure message symbols of the j 'th final codeword.

In traditional convertible codes, initial and final configurations are implicitly defined as the collection of their corresponding codewords. However, to analyze an eavesdropper who may read symbols across multiple codewords during the conversion (see figure 1), it is more convenient to define these configurations as vectors. Let $\mathbf{x}^I \in \mathbb{F}^{\lambda^I n^I}$ represent the vector consisting of all the codewords in the initial configuration (in the implicit ordering specified by the Convertible code), and $\mathbf{x}^F \in \mathbb{F}^{\lambda^F n^F}$ represent the same for the final configuration. Then, for each $i \in [\lambda^I]$, the vector $\mathbf{x}_i^I \in \mathbb{F}^{n^I}$ is the i 'th initial

codeword and for each $j \in [\lambda^F]$ the vector $\mathbf{x}_j^F \in \mathbb{F}^{n^F}$ is the j 'th final codeword.

Definition 2: A (ℓ, k_S) -Secure $[n^I, k^I; n^F, k^F]$ convertible code is a $[n^I, k^I; n^F, k^F]$ convertible code that can store a message $\mathbf{s} \in \mathbb{F}^{k_S}$ satisfying following decoding and secrecy properties:

- 1) **Decoding (MDS property).** For each $i \in [\lambda^I]$ and any subset $\mathbb{B} \subset [n^I]$ of size k^I ,

$$H(\mathcal{S}_i^I | \mathcal{X}_{i, \mathbb{B}}^I) = 0,$$

and for each $j \in [\lambda^F]$ and any subset $\mathbb{B} \subset [n^F]$ of size k^F ,

$$H(\mathcal{S}_j^F | \mathcal{X}_{j, \mathbb{B}}^F) = 0.$$

- 2) **ℓ -Secrecy.** For any $\mathbb{E}^I \subset [\lambda^I n^I]$, $\mathbb{E}^F \subset [\lambda^F n^F]$ of combined size $|\mathbb{E}^I| + |\mathbb{E}^F| \leq \ell$,

$$H(\mathcal{S} | \mathcal{X}_{\mathbb{E}^I}^I, \mathcal{X}_{\mathbb{E}^F}^F) = H(\mathcal{S}).$$

As in traditional convertible codes, the access cost is measured by the number of initial symbols accessed in the conversion procedure. We are interested in secure convertible codes that maximize k_S and minimize access cost simultaneously. In the following section, we prove the information-theoretic upper bound on the number of secure message symbols that be stored using a convertible code. For (ℓ, k_S) -secure convertible codes that reach the upper bound secrecy capacity, we drop the k_S from the notation, simply denoting them as optimal ℓ -secure convertible codes.

III. UPPER BOUND ON THE SECRECY CAPACITY OF CONVERTIBLE CODES

In order to derive an upper bound on the secrecy capacity, we first address a necessary nuance of ℓ -secure convertible codes. In this model, an eavesdropper is given the highest level of flexibility, where she can choose any symbol within the initial or final configuration to access. In particular, she may choose to read only the symbols of an individual codeword. Thus, in order for ℓ -secrecy to hold for the overall convertible code, *each codeword* must be secure to ℓ eavesdroppers. This intuition is captured in the following lemma.

Lemma 3: For any (ℓ, k_S) -secure $[n^I, k^I; n^F, k^F]$ convertible code with secure message symbols \mathbf{s} and symbol partitions $(\mathbb{S}^I, \mathbb{S}^F)$ with initial and final configuration codewords $\mathbf{x}^I \in \mathbb{F}^{\lambda^I n^I}$, $\mathbf{x}^F \in \mathbb{F}^{\lambda^F n^F}$, the following hold:

- 1) **Initial codeword Secrecy:** For any $i \in [\lambda^I]$ and subset $\mathbb{E}_i^I \subset [n^I]$ of size ℓ , we have $H(\mathcal{S}_{\mathbb{S}_i^I} | \mathcal{X}_{\mathbb{E}_i^I}^I) = H(\mathcal{S}_{\mathbb{S}_i^I})$.
- 2) **Final codeword Secrecy:** For any $j \in [\lambda^F]$ and subset $\mathbb{E}_j^F \subset [n^F]$ of size ℓ , $H(\mathcal{S}_{\mathbb{S}_j^F} | \mathcal{X}_{\mathbb{E}_j^F}^F) = H(\mathcal{S}_{\mathbb{S}_j^F})$.

Proof: This follows from ℓ -secrecy of definition 2. ■

Lemma 3 is used to derive the upper bound on the number of secure message symbols k_S for a ℓ -secure convertible codes.

Theorem 4: For positive integers $k^I, n^I, k^F, n^F, \ell, k_S$ such that $k^I \leq n^I, k^F \leq n^F, \ell < \min\{k^I, k^F\}$, any (ℓ, k_S) -secure $[n^I, k^I; n^F, k^F]$ convertible code satisfies

$$k_S \leq \min\{\lambda^I(k^I - \ell), \lambda^F(k^F - \ell)\}.$$

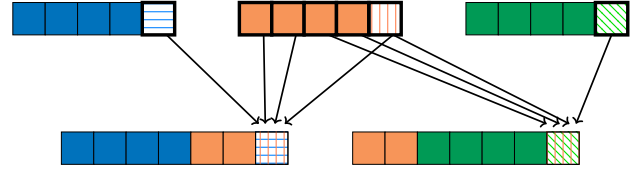


Fig. 2. A systematic access-optimal $[5, 4; 7, 6]$ convertible code, where only the non-systematic symbols are changed. The bold-edge symbols were accessed by the conversion procedure. The arrows represent which accessed symbol was used in the computation of each new non-systematic symbol in the final codewords.

Proof: Suppose $k^I \leq k^F$. Fix $i \in [\lambda^I]$ and suppose $\mathbb{E} \subset \mathbb{B} \subset [n^I]$ such that $|\mathbb{E}| = \ell$ and $|\mathbb{B}| = k^I$. Then,

$$k_S = H(\mathcal{S}) = \sum_{i=1}^{\lambda^I} H(\mathcal{S}_{\mathbb{S}_i^I}) \leq \lambda^I(k^I - \ell).$$

where the last inequality follows from

$$\begin{aligned} H(\mathcal{S}_{\mathbb{S}_i^I}) &= H(\mathcal{S}_{\mathbb{S}_i^I} | \mathcal{X}_{\mathbb{E}}) - H(\mathcal{S}_{\mathbb{S}_i^I} | \mathcal{X}_{\mathbb{B}}) \quad (\text{Lemma 3}) \\ &= H(\mathcal{S}_{\mathbb{S}_i^I} | \mathcal{X}_{\mathbb{E}}) - H(\mathcal{S}_{\mathbb{S}_i^I} | \mathcal{X}_{\mathbb{E}}, \mathcal{X}_{\mathbb{B} \setminus \mathbb{E}}) \\ &= I(\mathcal{S}_{\mathbb{S}_i^I}; \mathcal{X}_{\mathbb{B} \setminus \mathbb{E}} | \mathcal{X}_{\mathbb{E}}) \\ &\leq H(\mathcal{X}_{\mathbb{B} \setminus \mathbb{E}} | \mathcal{X}_{\mathbb{E}}) \leq H(\mathcal{X}_{\mathbb{B} \setminus \mathbb{E}}) \leq (k^I - \ell), \end{aligned}$$

Suppose $k^F < k^I$. Fix $j \in [\lambda^F]$ and suppose $\mathbb{E} \subset \mathbb{B} \subset [n^F]$ such that $|\mathbb{E}| = \ell$ and $|\mathbb{B}| = k^F$. Then, symmetric to the previous case, we have $H(\mathcal{S}_{\mathbb{S}_j^F}) \leq (k^F - \ell)$ and $k_S \leq \lambda^F(k^F - \ell)$. Putting the cases together, we have our desired bound. ■ Note that $\lambda^I(k^I - \ell) \leq \lambda^F(k^F - \ell)$ if and only if $\lambda^I \geq \lambda^F$ i.e., the upperbound on the secrecy capacity is determined by whether there are more initial codewords ($\lambda^I \geq \lambda^F$) or there are more final codewords ($\lambda^I < \lambda^F$).

The intuition for the secrecy capacity upperbound is the tension between the information needed for decoding and the information hidden by ℓ -secrecy. First, the MDS property of the initial code implies that any k^I initial codeword symbols is sufficient for decoding the underlying k^I message symbols. An eavesdropper reading ℓ codeword symbols can get at most ℓ symbols worth of information that, in the worst case, directly overlaps with the k^I message symbols, so at most $k^I - \ell$ of these symbols may be meaningful. Since the same holds for the final codewords, we have our secrecy capacity upperbound.

IV. ACCESS-OPTIMAL SECURE CONVERTIBLE CODE CONSTRUCTIONS FOR ALL PARAMETERS

In this section, access-optimal secure convertible codes that reach the secrecy capacity derived in Theorem 4 are constructed for all parameters k^I, n^I, k^F, n^F , and ℓ . As a starting point, traditional access-optimal convertible code constructions for all parameters from Maturana et al. [5] are used. For example, Figure 2 depicts an access-optimal $[5, 4; 7, 6]$ convertible code. The proposed construction concatenates existing access-optimal convertible codes with another code, known as the *nested code* [9].

A. Nested codes

Nested codes were constructed by Subramanian and McLaughlin in the context of securing messages through a wiretap channel with erasures [9]. The wiretap channel with erasures considers the eavesdropper-security and decoding of a single codeword: any ℓ codeword symbols reveal nothing of the secured symbols, and any k codeword symbols suffices to decode the message. Subramanian and McLaughlin show that the secrecy capacity is $(k - \ell)$. Given this, they constructed nested codes to satisfy the requirements of the wiretap channel with erasures while reaching secrecy capacity.

Definition 5 (Nested Code [9]): An MDS $[n, k]$ code \mathcal{C} is a ℓ -nested code if it has a generator matrix $G = \begin{bmatrix} G_s \\ G_\kappa \end{bmatrix} \in \mathbb{F}^{k \times n}$, where $G_\kappa \in \mathbb{F}^{\ell \times n}$ is a generator matrix of a MDS code.

First, a new message vector of length k is constructed comprising of the message symbols to be encoded, $\mathbf{s} \in \mathbb{F}^{k-\ell}$, and some *masking* symbols, $\boldsymbol{\kappa} \in \mathbb{F}^\ell$, where each masking symbol is chosen uniformly at random over \mathbb{F} . Let $\mathbf{m} = [\mathbf{s} \ \boldsymbol{\kappa}]$ be this message vector. Then, its encoding is $\mathbf{m}G = \mathbf{s}G_s + \boldsymbol{\kappa}G_\kappa$. One can verify that no information about the secure message symbols \mathbf{s} is leaked from any $j < \ell$ codeword symbols due to the addition of the encoding of the masked symbols $\boldsymbol{\kappa}G_\kappa$. The example below illustrates this:

1) *Example:* Consider a nested coding for $n = k = 4, \ell = 2$. Suppose \mathcal{D} is the MDS 2-nested code over \mathbb{F}_5

with generator matrix G defined as $G = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}$. The

secrecy capacity is $4 - 2 = 2$. The secure message symbols are $s_1, s_2 \in \mathbb{F}_4$ and the (uniformly random) masking symbols are $\kappa_1, \kappa_2 \in \mathbb{F}_5$. Let the message be $\mathbf{m} = [s_1 \ s_2 \ \kappa_1 \ \kappa_2]$.

Then $\mathcal{D}(\mathbf{m}) = [s_1 + \kappa_{1,1} \ s_2 + \kappa_{1,2} \ \kappa_1 \ \kappa_2]$, where $\kappa_{i,j} = i\kappa_1 + j\kappa_2$. Any eavesdropper reading any 2 symbols learns nothing about the secure message symbols s_1 and s_2 .

B. Constructing Access-Optimal Secure Convertible Codes

In the context of securing convertible codes, we apply a concatenated code, where the outer code is a nested code, and the inner code is the initial code of the convertible code. Intuitively, the nested code applies secrecy onto the message before it is stored on a convertible code. After conversion, the applied secrecy from the nested code will be present in the final codewords.

1) *Example:* Consider a 1-secure $[5, 4; 7, 6]$ convertible code over \mathbb{F}_7 . Here, $\lambda^I = 3$ and $\lambda^F = 2$, so the upperbound on the secrecy capacity is $\min\{3(4 - 1), 2(6 - 1)\} = 9$. Let $\mathbf{s} = s_1 \dots s_9$ be the secure message symbols and let $\boldsymbol{\kappa} \in \mathbb{F}$ be a redundant symbol. Consider the MDS 1-nested $[4, 4]$

code \mathcal{D}^I with generator $G = \begin{bmatrix} G_s \\ G_\kappa \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$.

Then, set the message vectors in the initial configuration $\mathbf{m}_1^I, \mathbf{m}_2^I, \mathbf{m}_3^I \in \mathbb{F}_7^4$ as

$$\begin{aligned} \mathbf{m}_1^I &= \mathcal{D}^I([s_1 \ s_2 \ s_3 \ \boldsymbol{\kappa}]) = [\hat{s}_1 \ \hat{s}_2 \ \hat{s}_3 \ \boldsymbol{\kappa}], \\ \mathbf{m}_2^I &= \mathcal{D}^I([s_4 \ s_5 \ s_6 \ \boldsymbol{\kappa}]) = [\hat{s}_4 \ \hat{s}_5 \ \hat{s}_6 \ \boldsymbol{\kappa}], \\ \mathbf{m}_3^I &= \mathcal{D}^I([s_7 \ s_8 \ s_9 \ \boldsymbol{\kappa}]) = [\hat{s}_7 \ \hat{s}_8 \ \hat{s}_9 \ \boldsymbol{\kappa}], \end{aligned}$$

where $\hat{s}_i = s_i + \boldsymbol{\kappa}$. An eavesdropper reading any 1 symbol learns nothing about the secure message symbols; either they read $\boldsymbol{\kappa}$ or an obfuscated secure symbol $s_i + \boldsymbol{\kappa}$. Note that the secure message symbols in any initial configuration message vector \mathbf{m}_i^I can be decoded by reading all 4 of its symbols, following from the MDS property of \mathcal{D}^I .

Let $(\mathcal{C}^I, \mathcal{C}^F)$ be an access-optimal $[5, 4; 7, 6]$ convertible code. The initial configuration codeword is set to

$$\mathbf{x}^I = (\mathcal{C}^I(\mathbf{m}_1^I), \mathcal{C}^I(\mathbf{m}_2^I), \mathcal{C}^I(\mathbf{m}_3^I)),$$

Using the conversion procedure of the convertible code $(\mathcal{C}^I, \mathcal{C}^F)$ on the initial codewords results in final codewords $\mathcal{C}^F(\mathbf{m}_1^F), \mathcal{C}^F(\mathbf{m}_2^F)$, where the message vectors in the final configuration $\mathbf{m}_1^F, \mathbf{m}_2^F$ are defined as

$$\begin{aligned} \mathbf{m}_1^F &= [\hat{s}_1 \ \hat{s}_2 \ \hat{s}_3 \ \hat{s}_4 \ \hat{s}_5 \ \boldsymbol{\kappa}], \\ \mathbf{m}_2^F &= [\hat{s}_6 \ \hat{s}_7 \ \hat{s}_8 \ \hat{s}_9 \ \boldsymbol{\kappa} \ \boldsymbol{\kappa}]. \end{aligned}$$

Again, in the final configuration, any 1 symbol that an eavesdropper reads is either a masking symbol or a masked secure symbol. Lastly, the secure message symbols can be decoded from any final codeword $j = 1, 2$.

The following construction shows how to use the approach in the previous example for general parameters.

2) *General construction:* By Theorem 4, the secrecy capacity is $k_S \leq \min\{\lambda^I(k^I - \ell), \lambda^F(k^F - \ell)\}$. Without loss of generality, suppose that $\lambda^I(k^I - \ell) \leq \lambda^F(k^F - \ell)$, which is equivalent to $\lambda^I \geq \lambda^F$.

Construction 6:

Preliminaries. Suppose the secure message symbols are $s_1, \dots, s_{\lambda^I(k^I - \ell)} \in \mathbb{F}$ and the redundant symbols are $\kappa_1, \dots, \kappa_\ell \in \mathbb{F}$. Further, let $\mathbf{s}_i^I = [s_{i(k^I - \ell) + 1} \dots s_{(i+1)(k^I - \ell)}]$ for $i \in [\lambda^I]$ and $\boldsymbol{\kappa} = [\kappa_1 \dots \kappa_\ell]$. By [5], there exists an access optimal $[n^I, k^I; n^F, k^F]$ convertible code $(\mathcal{C}^I, \mathcal{C}^F)$, which will be used in the construction. Next, let \mathcal{D}^I be an MDS ℓ -nested $[k^I, k^I]$ code with generator $G^I = \begin{bmatrix} G_s^I \\ G_\kappa^I \end{bmatrix}$

where $G_\kappa^I \in \mathbb{F}^{\ell \times k^I}$ is the generator of an MDS $[k^I, \ell]$ code and G_s^I is defined as $G_s^I = \begin{bmatrix} \mathbf{I}_{k^I - \ell} & \mathbf{0} \end{bmatrix}$, where

$\mathbf{I}_{k^I - \ell}$ is the identity matrix of size $k^I - \ell$ and $\mathbf{0} \in \mathbb{F}^{k^I \times \ell}$ is the all-zeros matrix. It is not hard to confirm that \mathcal{D}^I is MDS i.e., G^I is invertible.

Encoding in the initial configuration. Form the i 'th message vector in the initial configuration \mathbf{m}_i^I for each $i \in [\lambda^I]$ as $\mathbf{m}_i^I = \mathcal{D}^I([\mathbf{s}_i^I \ \boldsymbol{\kappa}]) = \mathbf{s}_i^I G_s^I + \boldsymbol{\kappa} G_\kappa^I$. Next, form the initial configuration codeword as $\mathbf{x}^I = (\mathcal{C}^I(\mathbf{m}_i^I))_{i \in [\lambda^I]}$.

Decoding in the initial configuration. To decode any initial codeword $\mathcal{C}^I(\mathbf{m}_i^I)$, use the decoding algorithm for \mathcal{C}^I , then

apply the decoding algorithm for \mathcal{D}^I (apply the inverse of its generator matrix G^{-1}).

Code conversion The final configuration (and final codewords) is constructed by running the conversion procedure of the underlying convertible code as-is to obtain $\mathbf{x}^F = (\mathcal{C}^F(\mathbf{m}_j^F))_{j \in [\lambda^F]}$, where \mathbf{m}_j^F are the message vectors in the final configuration.

Decoding in the final configuration. To decode all secure message symbols of a given final codeword $j \in [\lambda^F]$: 1) apply the decoder of \mathcal{C}^F to recover \mathbf{m}_j^F , 2) recover κ from \mathbf{m}_j^F , 3) Each message symbol in the final configuration has at most one secure symbol s_{iq} to decode, where $i \in [\lambda^I]$, and $q \in [k^I - \ell]$. For each message symbol $(\mathbf{m}_j^F)_p$, where $p \in [k^F]$, corresponding to a unique secure symbol s_{iq} (as will be shown below), output $(m_j^F)_p - (\kappa G_\kappa)_q$. In Theorem 7, we prove that each assertion is valid and this procedure always correctly decodes s_{iq} .

When $\lambda^I < \lambda^F$, the construction follows along similar lines. In this case, we start the construction by defining the final configuration and then work backwards. Form \mathbf{s}_j^F for $j \in [\lambda^F]$, message vectors in the final configuration codeword \mathbf{m}_j^F , and final configuration codeword $\mathbf{x}^F = (\mathcal{C}^F(\mathbf{m}_j^F))_{j \in [\lambda^F]}$ similarly. Then, define \mathbf{m}_i^I for each $i \in [\lambda^I]$ such that running the conversion procedure on the initial configuration codeword $\mathbf{x}^I = (\mathcal{C}^I(\mathbf{m}_i^I))_{i \in [\lambda^I]}$ results in \mathbf{x}^F .

We prove that our construction is an optimal secure convertible code with optimal access cost.

Theorem 7: For any integers n^I, n^F, k^I, k^F such that $0 \leq k^I \leq n^I, 0 \leq k^F \leq n^F$, and $\ell < \min\{k^I, k^F\}$, construction 6 is an optimal ℓ -secure $[n^I, k^I; n^F, k^F]$ convertible code with optimal access cost.

Proof: Without loss of generality, consider the construction when $\lambda^I \geq \lambda^F$. First, the initial codewords are decodable and ℓ -secure by construction of the initial configuration. Decodability follows from the MDS property of codes \mathcal{C}^I and \mathcal{D}^I . Each initial codeword $\mathcal{C}^I(\mathbf{m}_i^I)$ is ℓ -secure since each message vector in the initial configuration \mathbf{m}_i^I are codewords of code \mathcal{D}^I , which is an ℓ -secure code. Next, the final codewords $\mathcal{C}^F(\mathbf{m}_j^F)$ retain ℓ -secrecy. If not, this implies that the message vectors in the final configuration \mathbf{m}_j^F do not have ℓ -secrecy because the contrapositive, that messages with ℓ -secrecy imply their encodings have ℓ -secrecy, is true. Then, there is some subset of message symbols in the final configuration of size less than ℓ that reveal nonzero information about the secure message symbols. However, these message symbols in the final configuration were originally message symbols in the initial configuration and since initial messages are codewords of \mathcal{D}^I , there exists a subset of less than ℓ codeword symbols that reveal information about the secure message symbols, contradicting the ℓ -secrecy of code \mathcal{D}^I .

It is left to show that each final codeword $\mathcal{C}^F(\mathbf{m}_j^F)$ is decoded correctly by our specified algorithm. In step 1, \mathbf{m}_j^F is recovered by the decoder of \mathcal{C}^F . Next, step 2 is always possible i.e., every message vector in the final configuration contains a copy of κ . Each final codeword will contain all

symbols of some message vector in the initial configuration \mathbf{m}_i^I due to properties of the access-optimal convertible codes constructed by Maturana et al. [5]. In their construction, for all $\mathbb{P}_j^F \in \mathbb{P}^F$, there is a $\mathbb{P}_i^I \in \mathbb{P}^I$ such that $\mathbb{P}_i^I \subset \mathbb{P}_j^F$. Thus, each final codeword has all symbols of some message vector in the initial configuration $\mathbf{m}_i^I = \mathcal{D}^I([\mathbf{s}_i^I \ \kappa])$. Thus, since \mathcal{D}^I is MDS, we can recover κ .

For step 3, we first show that each message symbol in the final configuration symbol $(m_j^F)_p$ corresponds to at most one secure symbol s_{iq} . We show this for *initial* message symbols, since message symbols in the final configuration are just a repartitioning of the symbols of message vectors in the initial configuration. This is true by the construction of G_s^I , which maps each secure symbol to a unique symbol of a message vector in the initial configuration. Also, this implies that since $(m_j^F)_p$ has the unique secure symbol s_{iq} , $(m_j^F)_p = (m_i^I)_q$. Thus, the decoding procedure correctly decodes s_{iq} since

$$(m_j^F)_p - (\kappa G_\kappa)_q = (m_i^I)_q - (\kappa G_\kappa)_q = (\mathbf{s}_i^I G_s^I)_q = s_{iq}$$

Since the constructed code uses the same conversion procedure as the underlying access optimal convertible code $(\mathcal{C}^I, \mathcal{C}^F)$, our construction also achieves access optimal conversion. Lastly, the proof for the construction when $\lambda^I < \lambda^F$ follows a symmetric argument. ■

Remark: The field size requirement for the construction is the same as that of the access-optimal convertible code used in [5]. More specifically, the construction utilizes an ℓ -nested code with field size at most linear in $\min\{k^I, k^F\}$. Thus, construction 6 has the same field size requirement as the utilized access-optimal convertible codes, and benefit from recent works improving the field size requirement of access-optimal convertible codes [10]. In terms of computational overhead, in addition to the decoding procedure of the underlying convertible code, there is an additional decoding step for the MDS nested code in our secure convertible code construction.

V. CONCLUSION

In this paper, we introduce an information-theoretic secrecy model for convertible codes in the presence of eavesdroppers. We derived fundamental upper bounds on the number of message symbols that can be stored securely using convertible codes that provide security against eavesdroppers while maintaining access cost optimality of code conversions. We also presented explicit construction of optimal secure convertible codes meeting these bounds.

These results establish a foundation for designing secure and efficient convertible codes. This work opens up several avenues for future work. For example, the notion of secrecy can be expanded for convertible codes in the bandwidth cost model. Since bandwidth-optimal convertible codes are constructed using access-optimal convertible codes, a natural follow-up to this work would be to see if our construction retains secrecy and reaches bandwidth secrecy capacity.

REFERENCES

- [1] C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin, "Erasure coding in windows azure storage," in *2012 USENIX Annual Technical Conference (USENIX ATC 12)*. Boston, MA: USENIX Association, Jun. 2012, pp. 15–26. [Online]. Available: <https://www.usenix.org/conference/atc12/technical-sessions/presentation/huang>
- [2] S. Ghemawat, H. Gobioff, and S.-T. Leung, "The google file system," *SIGOPS Oper. Syst. Rev.*, vol. 37, no. 5, p. 29–43, oct 2003. [Online]. Available: <https://doi.org/10.1145/1165389.945450>
- [3] S. Kadekodi, K. V. Rashmi, and G. R. Ganger, "Cluster storage systems gotta have heart: improving storage efficiency by exploiting disk-reliability heterogeneity," in *Proceedings of the 17th USENIX Conference on File and Storage Technologies*, ser. FAST'19. USA: USENIX Association, 2019, p. 345–358.
- [4] F. Maturana and K. V. Rashmi, "Convertible codes: enabling efficient conversion of coded data in distributed storage," *IEEE Transactions on Information Theory*, vol. 68, pp. 4392–4407, 2022.
- [5] F. Maturana, V. S. C. Mukka, and K. V. Rashmi, "Access-optimal linear mds convertible codes for all parameters," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 577–582.
- [6] F. Maturana and K. V. Rashmi, "Bandwidth cost of code conversions in distributed storage: Fundamental limits and optimal constructions," *IEEE Transactions on Information Theory*, vol. 69, no. 8, pp. 4993–5008, 2023.
- [7] —, "Bandwidth cost of code conversions in the split regime," in *2022 IEEE International Symposium on Information Theory (ISIT)*, 2022, pp. 3262–3267.
- [8] S. Pawar, S. El Rouayheb, and K. Ramchandran, "Securing dynamic distributed storage systems against eavesdropping and adversarial attacks," *IEEE Transactions on Information Theory*, vol. 57, no. 10, pp. 6734–6753, 2011.
- [9] A. Subramanian and S. W. McLaughlin, "MDS codes on the erasure-erasure wiretap channel," *arXiv preprint arXiv:0902.3286*, 2009.
- [10] S. Chopra, F. Maturana, and K. V. Rashmi, "On low field size constructions of access-optimal convertible codes," in *2024 IEEE International Symposium on Information Theory (ISIT)*, 2024, pp. 1456–1461.