Secure Convertible Codes for Eavesdropper Attacks

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*—THIS PAPER IS ELIGIBLE FOR THE STUDENT PAPER AWARD. 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. We then derive the information-theoretic upper bound on the number of message symbols that can be stored securely on an access-optimal convertible code. 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 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 chunks, which are then encoded into n chunks, 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 chunks 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. 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—incurs significant costs in terms of I/O, CPU, 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 $\mathcal{C}^I$ with parameters $[n^I, k^I]$ to a final code $\mathcal{C}^F$ 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 *eavesdroppers* who gain read access to some of codeword symbols stored in the system and try to learn information about the original data. 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 symbols of a convertible code learns no information about the underlying data.

We then focus on access-optimal convertible codes and establish an upper bound on the number of secure message symbols that can be 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. Bold lowercase letters will denote vectors, e.g. a n-length vector, $\boldsymbol{x} \in \mathbb{F}^n$. The i'th symbol of a vector $\boldsymbol{x}$ is written (non-bold) as x_i. A vector subscripted with a set, e.g. $\boldsymbol{x}_S$, denote the projection of the vector to each coordinate in the set $\boldsymbol{x}$ e.g. $\boldsymbol{x}_S = [x_i : i \in S]$. Bold uppercase letters denote matrices, e.g. a matrix of size $k \times n$, $\boldsymbol{G} \in \mathbb{F}^{k \times n}$. Let $[i] = \{1, 2, \ldots, i\}$. 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. 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 $\boldsymbol{m} \in \mathbb{F}^{\mathcal{M}}$ be the message symbols to be stored, where $\mathcal{M} = \text{lcm}(k^I, k^F)$. The initial configuration contains $\lambda^I = \mathcal{M}/k^I$ codewords, and the final configuration contains $\lambda^F = \mathcal{M}/k^F$ codewords. The message symbols in each codeword is determined by the initial and final partitions $\mathcal{P}^I$ and $\mathcal{P}^F$ of $[\mathcal{M}]$. A conversion procedure is then defined to transform the initial configuration into the final one. In the access model, the access cost is measured by the number of symbols used by the conversion procedure.

More formally, let $[i] = \{1, 2, \ldots, i\}$ and for vector $\boldsymbol{x}$, let $\boldsymbol{x}_S$ be the vector formed by projecting each index $i \in S$ of $\boldsymbol{x}$ in order.

Definition 1 (Convertible Code [5]): 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 $\mathcal{P}^I, \mathcal{P}^F \in \Pi(\mathcal{M})$, such that each subset $P_i^I \in \mathcal{P}^I$ has size k^I, and each subset $P_j^F \in \mathcal{P}^F$ has size k^F.
3) A conversion procedure mapping initial codewords $\{\mathcal{C}^I(\boldsymbol{m}_{P_i^I}) : P_i^I \in \mathcal{P}^I\}$ to final codewords $\{\mathcal{C}^F(\boldsymbol{m}_{P_j^F}) : P_j^F \in \mathcal{P}^F\}$.

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. Secure convertible codes are defined by enhancing the existing convertible code framework with protection against eavesdroppers. A message $\boldsymbol{m} \in \mathbb{F}^{\mathcal{M}}$ is stored on the convertible code, resulting in an initial/final configuration comprising of initial/final codewords.

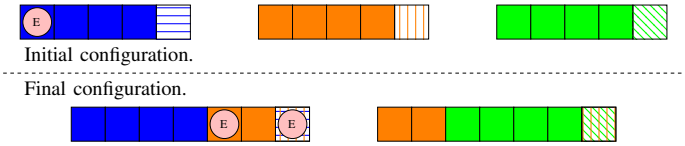

Fig. 1. A $[5, 4; 7, 6]$ convertible code with 3 symbols read by the eavesdropper (pink circles). Initial codewords are on the top of the diagram while the final codewords are on the bottom of the diagram. The initial/final codewords make up the initial/final configuration. Note that an eavesdropper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

Now, suppose that an eavesdropper gains read-access to any $\ell < \min\{k^I, k^F\}$ symbols at any point during the conversion from the initial configuration to the final configuration, as illustrated in Figure 1. Consequently, the eavesdropper may read symbols across multiple initial and final codewords.

Note that this secrecy model captures the special case when the eavesdropper reads the symbols accessed (or downloaded) during the conversion process.

An additional parameter $\mathcal{M}_S \leq \mathcal{M}$ is introduced, which represents the number of symbols stored securely. The eavesdropper is required to learn nothing about these $\mathcal{M}_S$ *secure symbols* of the message $\boldsymbol{m}$, which are denoted by $\boldsymbol{s} \in \mathbb{F}^{\mathcal{M}_S}$. In contrast, the remaining symbols – referred to as *redundant symbols* and denoted $\boldsymbol{\kappa}$– may be exposed to the eavesdropper, and their decoding is not a requirement (although our construction does decode the redundant nodes, and in fact, it is necessary).

To formalize methods used to secure and decode secure symbols in a convertible code, we first define how message symbols are structured into initial and final configurations. Toward this, we first need to partition the secure message symbols. Specifically, we partition secure symbols into *secure symbol partitions*, denoted as $\mathcal{S}^I, \mathcal{S}^F \in \Pi(\mathcal{M}_S)$, where $|\mathcal{S}^I| = \lambda^I$ and $|\mathcal{S}^F| = \lambda^F$.

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: $\boldsymbol{X}^I \in \mathbb{F}^{\lambda^I n^I}$ for the initial configuration and $\boldsymbol{X}^F \in \mathbb{F}^{\lambda^F n^F}$ for the final configuration. The natural *codeword partitions* are defined as

$$\mathcal{W}^I = \{W_i^I := \{in^I + 1, \ldots, (i+1)n^I\} : i \in [\lambda^I]\}$$
$$\mathcal{W}^F = \{W_j^F := \{jn^F + 1, \ldots, (j+1)n^F\} : j \in [\lambda^F]\}.$$

In other words, the i'th initial codeword is represented by $\boldsymbol{X}^I_{W_i^I}$ (compared to traditional convertible codes, where the i'th initial codeword is given by $\mathcal{C}^I(\boldsymbol{m}_{P_i^I})$). For notational simplicity, these codeword partitions will be assumed for the remainder of the paper and be omitted from explicit mention.

Formally, the decoding (MDS) property — which guarantees that any secure symbol can be decoded using only symbols of its corresponding codeword — and the (ℓ-)secrecy

property — which ensures that any ℓ read symbols provide no information about the secure symbols — can both be expressed in terms of the secure symbol partitions, codeword partitions, and configurations.

Definition 2: A $(\ell, \mathcal{M}_S)-$Secure $[n^I, k^I; n^F, k^F]$ convertible code is specified by

- a message to be securely stored $\boldsymbol{s} \in \mathbb{F}^{\mathcal{M}_S}$,
- a pair of configurations $\boldsymbol{X}^I \in \mathbb{F}^{\lambda^I n^I}, \boldsymbol{X}^F \in \mathbb{F}^{\lambda^F n^F}$,
- a pair of secure symbol partitions $\mathcal{S}^I = \{S_1^I, \ldots, S_{\lambda^I}^I\}$ and $\mathcal{S}^F = \{S_1^F, \ldots, S_{\lambda^F}^F\}$, where $\mathcal{S}^I, \mathcal{S}^F \in \Pi(\mathcal{M}_S)$,
- and a conversion procedure mapping $\{\boldsymbol{X}_{W_i^I}^I : i \in [\lambda^I]\}$ to $\{\boldsymbol{X}_{W_j^F}^F : j \in [\lambda^F]\}$.

satisfying:

1) **Decoding (MDS property).** For each $i \in [\lambda^I]$ and any subset $B \subset W_i^I$ of size k^I,
$$H(\boldsymbol{s}_{S_i^I} | \boldsymbol{X}_B^I) = 0,$$
and for each $j \in [\lambda^F]$ and any subset $B \subset W_j^F$ of size k^F,
$$H(\boldsymbol{s}_{S_j^F} | \boldsymbol{X}_B^F) = 0.$$

2) **ℓ-Secrecy.** For any $E^I \subset [\lambda^I n^I]$, $E^F \subset [\lambda^F n^F]$ of combined size $\left|E^I\right| + \left|E^F\right| \leq \ell$,
$$H(\boldsymbol{s} \mid \boldsymbol{X}_{E^I}^I, \boldsymbol{X}_{E^F}^F) = \mathcal{M}_S.$$

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 $\mathcal{M}_S$ and minimize access cost simultaneously. In the following section, we prove the information-theoretic upper bound on the number of secure symbols on a convertible code to derive the secrecy capacity. For $(\ell, \mathcal{M}_S)$-secure convertible codes that reach the secrecy capacity, we drop the $\mathcal{M}_S$ from the notation, simply denoting them as optimal *ℓ-secure convertible codes.*

III. Secrecy Capacity of Convertible Codes

In order to derive 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 $(\ell, \mathcal{M}_S)$-secure $[n^I, k^I; n^F, k^F]$ convertible code with secure symbols $\boldsymbol{s}$ and secure symbol partitions $(\mathcal{S}^I, \mathcal{S}^F)$ with initial and final encodings $\boldsymbol{X}^I \in \mathbb{F}^{\lambda^I n^I}, \boldsymbol{X}^F \in \mathbb{F}^{\lambda^F n^F}$, the following must hold:

1) **Initial codeword Secrecy**: For each $i \in [\lambda^I]$ and any subset $E_i^I \subset W_i^I$ of size ℓ,
$$H(\boldsymbol{s}_{S_i^I} | \boldsymbol{X}_E) = H(\boldsymbol{s}_{S_i^I}).$$

2) **Final codeword Secrecy**: For each $j \in [\lambda^F]$ and any subset $E_j^F \subset W_j^F$ of size ℓ,
$$H(\boldsymbol{s}_{S_j^F} | \boldsymbol{X}_E) = H(\boldsymbol{s}_{S_j^F}).$$

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

Lemma 3 is used to derive the upper bound on $\mathcal{M}_S$ to get the secrecy capacity for ℓ-secure convertible codes.

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

Proof: Suppose $k^I \leq k^F$. Fix $i \in [\lambda^I]$ and suppose $E \subset B \subset W_i^I$ such that $|E| = \ell$ and $|B| = k^I$.
Then,
$$\begin{aligned} H(\boldsymbol{s}_{S_i^I}) &= H(\boldsymbol{s}_{S_i^I} | \boldsymbol{X}_E) - H(\boldsymbol{s}_{S_i^I} | \boldsymbol{X}_B) && \text{(Lemma 3)} \\ &= H(\boldsymbol{s}_{S_i^I} | \boldsymbol{X}_E) - H(\boldsymbol{s}_{S_i^I} | \boldsymbol{X}_E, \boldsymbol{X}_{B \setminus E}\} \\ &= I(\boldsymbol{s}_{S_i^I}; \boldsymbol{X}_{B \setminus E} | \boldsymbol{X}_{E_i}) \\ &\leq H(\boldsymbol{X}_{B \setminus E} | \boldsymbol{X}_E) \\ &\leq H(\boldsymbol{X}_{B \setminus E}) \\ &\leq k^I - \ell, \end{aligned}$$
where
$$\mathcal{M}_S = H(\boldsymbol{s}) = \sum_{i=1}^{\lambda^I} H(\boldsymbol{s}_{S_i^I}) \leq \lambda^I(k^I - \ell).$$
Now, suppose $k^F < k^I$. Fix $j \in [\lambda^F]$ and suppose $E \subset B \subset W_j^F$ such that $|E| = \ell$ and $|B| = k^F$. Then, symmetric to the previous case, we have $H(\boldsymbol{s}_{S_j^F}) \leq k^F - \ell$ and $\mathcal{M}_S \leq \lambda^F(k^F - \ell)$. Putting the two cases together, we have our desired bound. ∎

Note that $\lambda^I(k^I - \ell) \leq \lambda^F(k^F - \ell)$ iff $\lambda^I \geq \lambda^F$ i.e. 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 of convertible codes is the interplay between initial decoding, final decoding, and ℓ-secrecy. In the initial configuration, any k^I nodes of an initial codeword must give full information of its underlying secure message symbols by the decoding property. 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 symbols used for the initial decoding, so at most $k^I - \ell$ of these symbols may be meaningful. Since the same holds when considering final codeword decoding, we have our secrecy capacity.

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

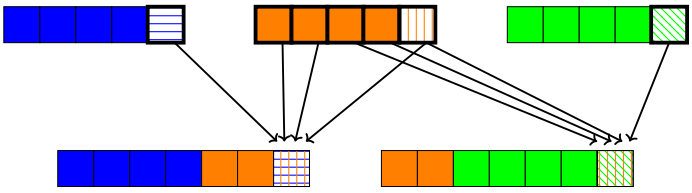

Fig. 2. A systematic access-optimal $[5, 4; 7, 6]$ convertible code, where only the non-systematic symbols are changed. The bold symbols were accessed by the conversion procedure. The arrows represent which accessed symbol was used in the computation of each new non-systematic symbol in the final codewords.

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 construction concatenates existing access-optimal convertible codes with another code, known as the *nested code*.

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 decode the message. Subramanian and McLaughlin show that the secrecy capacity is $k - \ell$, a similar bound to the conversion setting. 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 generator $\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G_s} \\ \boldsymbol{G_\kappa} \end{bmatrix} \in \mathbb{F}^{k \times n}$, where $\boldsymbol{G_\kappa} \in \mathbb{F}^{\ell \times n}$ is the generator of a MDS code.

A codeword of a nested code can be interpreted as applying secrecy to the $k - \ell$ secure symbols by adding ℓ redundancy symbols. Then, to read any secure symbol, ℓ redundancy symbols must first be read to recover any secure symbol. 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 with generator $\boldsymbol{G}$ defined as

$$\boldsymbol{G} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 1 & 1 & 0 \\ 1 & 2 & 0 & 1 \end{bmatrix}.$$

The secrecy capacity is $4 - 2 = 2$. Let the secure symbols be $s_1, s_2 \in \mathbb{F}$ and the redundant symbols be $\kappa_1, \kappa_2 \in \mathbb{F}$. Let the message be $\boldsymbol{m} = \begin{bmatrix} s_1 & s_2 & \kappa_1 & \kappa_2 \end{bmatrix}$.

Then $\mathcal{D}(\boldsymbol{m}) = \begin{bmatrix} s_1 + \kappa_{1,1} & s_2 + \kappa_{1,2} & \kappa_1 & \kappa_2 \end{bmatrix}$, where $\kappa_{i,j} = i\kappa_1 + j\kappa_2$. Any eavesdropper reading any 2 symbols learns nothing about the secure symbols s_1 and s_2; reading $s_1 + \kappa_{1,1}$ or $s_2 + \kappa_{1,2}$ gives no information on s_1 or s_2. The same holds for reading at least one redundant symbol. In other words, $H(s_1 s_2 | \mathcal{D}(\boldsymbol{m})_i) = 2$ for all $i \in [4]$.

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. Here, $\lambda^I = 3$ and $\lambda^F = 2$, so the secrecy capacity is $\min\{3(4-1), 2(6-1)\} = 9$. Let $\boldsymbol{s} = s_1 \dots s_9$ be the secure message symbols and let $\kappa \in \mathbb{F}$ be a redundant symbol. Consider the MDS 1-nested $[4, 4]$ code $\mathcal{D}^I$ with generator

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_s \\ \boldsymbol{G}_\kappa \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

Then, set the initial messages $\boldsymbol{m}_1^I, \boldsymbol{m}_2^I, \boldsymbol{m}_3^I$ as

$$\begin{aligned} \boldsymbol{m}_1^I &= \mathcal{D}^I\left(\begin{bmatrix} s_1 & s_2 & s_3 & \kappa \end{bmatrix}\right) = \begin{bmatrix} \hat{s_1} & \hat{s_2} & \hat{s_3} & \kappa \end{bmatrix}, \\ \boldsymbol{m}_2^I &= \mathcal{D}^I\left(\begin{bmatrix} s_4 & s_5 & s_6 & \kappa \end{bmatrix}\right) = \begin{bmatrix} \hat{s_4} & \hat{s_5} & \hat{s_6} & \kappa \end{bmatrix}, \\ \boldsymbol{m}_3^I &= \mathcal{D}^I\left(\begin{bmatrix} s_7 & s_8 & s_9 & \kappa \end{bmatrix}\right) = \begin{bmatrix} \hat{s_7} & \hat{s_8} & \hat{s_9} & \kappa \end{bmatrix}, \end{aligned}$$

where $\hat{s_i} = s_i + \kappa$. An eavesdropper reading any 1 symbol learns nothing about the secure symbols; either they read κ or an obfuscated secure symbol $s_i + \kappa$.. Moreover, the decoding of the secure symbol is possible by 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 is set to

$$\boldsymbol{X}^I = \left(\mathcal{C}^I\left(\boldsymbol{m}_1^I\right), \mathcal{C}^I\left(\boldsymbol{m}_2^I\right), \mathcal{C}^I\left(\boldsymbol{m}_3^I\right)\right),$$

with the (natural) secure symbol partition

$$\mathcal{S}^I = \{\{1, 2, 3\}, \{4, 5, 6\}, \{7, 8, 9\}\}.$$

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

$$\begin{aligned} \boldsymbol{m}_1^F &= \begin{bmatrix} \hat{s_1} & \hat{s_2} & \hat{s_3} & \hat{s_4} & \hat{s_5} & \kappa \end{bmatrix}, \\ \boldsymbol{m}_2^F &= \begin{bmatrix} \hat{s_6} & \hat{s_7} & \hat{s_8} & \hat{s_9} & \kappa & \kappa \end{bmatrix}. \end{aligned}$$

Again, for the final messages, any 1 symbol that an eavesdropper reads is either a redundancy symbol or an obfuscated secure symbol. Lastly, the decoding of any secure symbol from a final codeword $\boldsymbol{m}_j^F$ is possible in two steps: $\boldsymbol{m}_j^F$ is decoded by the MDS property of $\mathcal{C}^F$, then κ is subtracted from each $\hat{s_i}$.

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 $\mathcal{M}_S \leq \min\{\lambda^I(k^I - \ell), \lambda^F(k^F - \ell)\}$. Without loss of generality, suppose that $\lambda^I(k^I - \ell) \geq \lambda^F(k^F - \ell)$, which is equivalent to $\lambda^I \geq \lambda^F$.

Construction 6:

Preliminaries. Suppose the secure symbols are $s_1, \ldots, s_{\lambda^I(k^I-\ell)} \in \mathbb{F}$ and the redundant symbols are $\kappa_1, \ldots, \kappa_\ell \in \mathbb{F}$. Further, let $\boldsymbol{s}_i^I = [s_{i(k-\ell)+1} \cdots s_{(i+1)(k-\ell)}]$ for $i \in [\lambda^I]$ and $\boldsymbol{\kappa} = [\kappa_1 \ldots \kappa_\ell]$. By [5], there exists an access optimal convertible code $(\mathcal{C}^I, \mathcal{C}^F)$ under the desired parameters. Next, let $\mathcal{D}^I$ be an MDS ℓ-nested $[k^I, k^I]$ code with generator $\boldsymbol{G}^I = \begin{bmatrix} \boldsymbol{G}_s^I \\ \boldsymbol{G}_\kappa^I \end{bmatrix}$ where $\boldsymbol{G}_\kappa^I \in \mathbb{F}^{\ell \times k^I}$ is the generator of an MDS $[k^I, \ell]$ code and $\boldsymbol{G}_s^I$ is defined as

$$\boldsymbol{G}_s^I = \left[\begin{array}{c|c} \boldsymbol{I}_{k^I-\ell} & \boldsymbol{0} \end{array} \right],$$

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

Encoding in the initial configuration. Form the i'th initial message $\boldsymbol{m}_i^I$ for each $i \in [\lambda^I]$ as

$$\boldsymbol{m}_i^I = \mathcal{D}^I\left(\begin{bmatrix} \boldsymbol{s}_i^I & \boldsymbol{\kappa} \end{bmatrix}\right) = \boldsymbol{s}_i^I \boldsymbol{G}_s^I + \boldsymbol{\kappa} \boldsymbol{G}_\kappa^I$$

Next, form each the initial configuration (and each initial codeword) as $\boldsymbol{X}^I = (\mathcal{C}^I(\boldsymbol{m}_i^I))_{i \in [\lambda^I]}$.

Decoding in the initial configuration. To decode any initial codeword, use the decoding algorithm for $\mathcal{C}^I$, then apply the decoding algorithm for $\mathcal{D}^I$ (apply the inverse of the generator matrix $\boldsymbol{G}^{I-1}$).

Code conversion The final configuration (and final codewords) is constructed by running the conversion procedure as-is to obtain $\boldsymbol{X}^F = (\mathcal{C}^F(\boldsymbol{m}_j^F))_{j \in [\lambda^F]}$, for some λ^F final messages $\boldsymbol{m}_j^F$.

Decoding in the final configuration. To decode all secure symbols of a given final codeword $j \in [\lambda^F]$: 1) apply the decoder of $\mathcal{C}^F$ to recover $\boldsymbol{m}_j^F$, 2) recover $\boldsymbol{\kappa}$ from $\boldsymbol{m}_j^F$, 3) Each final message symbol has at most one secure symbol s_{iq} to decode, where $i \in [\lambda^I], q \in [k^I - \ell], p \in [k^F]$. For each final message symbol $(m_j^F)_p$ that does, output $(m_j^F)_p - (\boldsymbol{\kappa}\boldsymbol{G}_\kappa)_q$.

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 is symmetric; swap initial/final parameters and "undo" the conversion procedure.

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: First, the initial codewords are decodable and ℓ-secure by construction of the initial configuration. Next, the final codewords retain ℓ-secrecy. Suppose not, then this implies that the final *messages* do not have ℓ-secrecy (the contrapositive, that messages with ℓ-secrecy imply their encodings have ℓ-secrecy, is true). Then, there is some subset of final message symbols of size less than ℓ that reveal nonzero information about the secure symbols. However, these final message symbols were originally initial message symbols and since initial messages are codewords of $\mathcal{D}^I$, there exists a subset of less than ℓ $\mathcal{D}^I$ codeword symbols that reveal information about the secure symbols, contradicting the ℓ-secrecy of $\mathcal{D}^I$.

It is left to show that each final codeword is decoded correctly by our specified algorithm. We start by show that step 2 is always possible i.e. every final message contains a copy of $\boldsymbol{\kappa}$. Indeed, more generally, each final codeword will contain all symbols of some initial message $\boldsymbol{m}_i^I$ due to properties of the access-optimal convertible codes constructed by Maturana et al. [5]. In their construction, for all $P_j^F \in \mathcal{P}^F$, there is a $P_i^I \in \mathcal{P}^I$ such that $P_i^I \subset P_j^F$. Thus, each final codeword has all symbols of some initial message. $\boldsymbol{m}_i^I = \mathcal{D}^I\left(\begin{bmatrix} \boldsymbol{s}_i^I & \boldsymbol{\kappa} \end{bmatrix}\right)$. Thus, since $\mathcal{D}^I$ is MDS, we can recover $\boldsymbol{\kappa}$.

For step 3, we first show that each final message symbols corresponds to at most one secure symbol. We show this for *initial* message symbols, since final message symbols are just a repartitioning of initial message symbols. Indeed, this is true by the construction of $\boldsymbol{G}_s^I$, which maps each secure symbol to a unique initial message symbol. 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

$$(m_j^F)_p - (\kappa \boldsymbol{G}_\kappa)_q = (m_i^I)_q - (\kappa \boldsymbol{G}_\kappa)_q = (\boldsymbol{s}_i^I \boldsymbol{G}_s^I)_q = s_{iq}$$

Lastly, our construction achieves access-optimal conversion since the access-optimal convertible code conversion procedure is used as-is. ∎

Remark: The field size requirement for the construction is the same as that of the access-optimal convertible code used. 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].

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.