# Secure Locally Repairable Codes for Distributed Storage Systems

Ankit Singh Rawat, O. Ozan Koyluoglu, Natalia Silberstein, and Sriram Vishwanath
Dept. of ECE, The University of Texas at Austin, Austin, TX 78751 USA
Email: ankitsr@utexas.edu, {ozan, natalys, sriram}@austin.utexas.edu

***Abstract*—This paper presents coding schemes for distributed storage systems (DSS) that are secure against eavesdroppers, while simultaneously enabling efficient node repair (regeneration). Towards this, novel upper bounds on secrecy capacity for minimum storage regenerating (MSR) codes and locally repairable codes (LRCs) are derived. The eavesdropper model considered in this paper incorporates the ability to listen in on data downloaded during $\ell_2$ node repairs in addition to content stored on $\ell_1$ nodes. Finally, this paper presents coding schemes, based on precoding using Gabidulin codes, that achieve the upper bounds on secrecy capacity and characterize the secrecy capacity of DSS for various settings of system parameters.**

***Index Terms*—Coding for distributed storage systems, locally repairable codes, minimum storage regenerating codes, security.**

## I. Introduction

Data intensive and data generative applications are increasingly pervasive, ranging from social networking to multimedia uploads. Thus, storage in the “cloud” is gaining prominence, where individuals and institutions use distributed storage services whose physical structure and location is mostly unknown to the user. This decentralized nature of cloud storage systems makes them susceptible to a variety of issues, including failure and passive/active attacks. Of particular interest to us in this paper is an eavesdropping attack, where an unauthorized party wishes to gain access to information stored on a distributed storage system (DSS) through a set of compromized nodes within the DSS.

In this paper, we focus on designing strategies for DSS that are eavesdropper-resistant while simultaneously being resilient to node failures. Resilience to node failures is an essential and well studied feature of DSS, and it is desirable that such resilience be possible with minimum repair bandwidth [1] and (or) small locality [2]. In our work, we desire to build on this literature, by developing coding schemes that counter eavesdropping while being efficient in node repair.

The primary contribution of this paper is the design of secure locally repairable codes (LRCs) for DSS. Towards this, we first study the secrecy against eavesdropping attacks in DSS that employ minimum storage regenerating (MSR) codes. We adopt the eavesdropper model presented in [3], where, during the entire life span of the DSS, the eavesdropper can access data stored on $\ell_1$ nodes, and, in addition, it observes data downloaded during node repair of an additional $\ell_2$ nodes. We derive an upper bound on secrecy capacity, the amount of data that can be stored on the system without leaking information to an eavesdropper, for DSS that employ MSR codes to facilitate bandwidth efficient node repair. Our bound is novel in that it can take into account the additional downloaded data observed by the eavesdropper during node repairs, and is tighter than the available bounds in the literature. We then present a secure, exact repairable coding scheme with higher code rate compared to that of [3]. Utilizing a special case of the obtained bound, we show that the proposed coding scheme achieves the optimal secure file size for any $(\ell_1, \ell_2)$ with $\ell_2 \leq 2$ at the MSR point.

Further, we study eavesdropper secrecy for LRCs that are minimum distance optimal, i.e., codes with maximum worst-case resilience for a given locality constraint. To the best of our knowledge, there is limited understanding of secrecy for LRCs, and our work is aimed at bringing these two concepts together in a meaningful fashion. We derive an upper bound on the size of data that can be stored in a locally repairable minimum distance optimal DSS that is secure against an $(\ell_1, \ell_2)$-eavesdropper. We consider two cases: (i) single parity node per local group and (ii) multiple parity nodes per local group. Multiple parity nodes per local group allow regenerating codes to be used inside local groups [4], [5]. This enables bandwidth efficient local repair of a failed node, which decreases the amount of data that is revealed to an eavesdropper during node repairs. We then provide coding schemes that achieve the respective upper bounds in both cases.

## II. System Model and Preliminaries

Consider a DSS with $n$ nodes at a given time storing a file $\mathbf{f}$ of size $\mathcal{M}$ over $\mathbb{F}_{q^m}$ without any secrecy constraint. The file $\mathbf{f} = (f_1, \ldots, f_{\mathcal{M}})$ is first encoded into $n$ data blocks, $(\mathbf{x}_1, \ldots, \mathbf{x}_n)$, each of length $\alpha$ over $\mathbb{F}_{q^m}$. In this paper, we focus on linear coding where the encoding process can be summarized as $\mathbf{x}_i = \{\mathbf{f}^T\mathbf{g}_i^1, \ldots, \mathbf{f}^T\mathbf{g}_i^\alpha\}$, for $i \in [n]$. ($[n]$ denotes the set $\{1, 2, \ldots, n\}$.)

In the event of a node failure, a newcomer node contacts $d$ surviving nodes and downloads $\beta$ symbols from each of these $d$ nodes. We use $\mathbf{d}_{i,j} = \mathbf{f}^T(\mathbf{g}_i^1, \ldots, \mathbf{g}_i^\alpha)V_{i,j}$ to denote $\beta$ symbols downloaded from node $i$ for repair of node $j$. Here $V_{i,j}$ represents the $\alpha \times \beta$ repair matrix for node $j$ associated with node $i$. We refer to $\mathcal{D}_{i,j}$ as the subspace spanned by rows of $(\mathbf{g}_i^1, \ldots, \mathbf{g}_i^\alpha)V_{i,j}$. $\mathcal{D}_j$ denotes the subspace downloaded to node $j$. For a given set of nodes $\mathcal{A}$, we use the notation $\mathbf{s}_{\mathcal{A}} \triangleq \{\mathbf{s}_i, i \in \mathcal{A}\}$, where $\mathbf{s}_i$ denotes data stored on node $i$. (Note that $\mathbf{s}_i = \mathbf{x}_i$ for $i \in [n]$.) A similar notation is adopted for the downloaded symbols, and the subspace representation.

### A. Regenerating codes

In their seminal work [1], Dimakis et al. characterize an information theoretic trade off between repair bandwidth ($d\beta$) and per node storage ($\alpha$) for DSS satisfying the *maximum distance separable* (MDS) or "any $k$ out of $n$" property. Two classes of codes that achieve two extreme points of this trade off are known as *minimum storage regenerating (MSR)* codes and *minimum bandwidth regenerating (MBR)* codes, corresponding to minimum storage per node (i.e., $\alpha = \mathcal{M}/k$) and minimum possible repair bandwidth ($\gamma = d\beta = \alpha$) respectively. For MSR codes, we have $(\alpha_{\rm msr}, \beta_{\rm msr}) = \left(\frac{\mathcal{M}}{k}, \frac{\mathcal{M}}{k(d-k+1)}\right)$. On the other hand, MBR codes are characterized by $(\alpha_{\rm mbr}, \beta_{\rm mbr}) = \left(\frac{2\mathcal{M}d}{k(2d-k+1)}, \frac{2\mathcal{M}}{k(2d-k+1)}\right)$.

In [6]–[8] and references therein, regenerating codes that allow *exact node repair* (data on the regenerated node is the same as that stored on the failed node) are presented. In what follows, we use the term *exact-MSR* to denote the MSR codes that allow exact node repair.

### B. Gabidulin codes

Gabidulin codes are an essential component of various coding schemes presented in this paper. Gabidulin codes are an example of maximum rank distance (MRD) codes [9]. Encoding a message $(f_1, f_2, \ldots, f_K)$ to a codeword of an $[N, K, D]$ Gabidulin code over $\mathbb{F}_{q^m}$ consists of two steps:

**Step 1:** Construct a data polynomial $f(y) = \sum_{i=1}^{K} f_i y^{q^{i-1}}$ over $\mathbb{F}_{q^m}$.

**Step 2:** Evaluate $f(y)$ at $\{y_1, y_2, \ldots, y_N\} \subset \mathbb{F}_{q^m}$, $N$-linearly independent (over $\mathbb{F}_q$) points, to obtain a codeword $\mathbf{c} = (f(y_1), \ldots, f(y_N))$.

**Remark 1.** *The data polynomial ($f(\cdot)$) constructed in the first step of encoding for Gabidulin codes is called linearized polynomial as it satisfies $f(ay_1 + by_2) = af(y_1) + bf(y_2)$, where $y_1, y_2 \in \mathbb{F}_{q^m}$ and $a, b \in \mathbb{F}_q$ [10].*

**Remark 2.** *Given evaluations of $f(\cdot)$ at any $K$ linearly independent (over $\mathbb{F}_q$) points in $\mathbb{F}_{q^m}$, say $(z_1, \ldots, z_K)$, one can get evaluations of $f(\cdot)$ at $q^K$ points spanned by $\mathbb{F}_q$-linear combinations of $(z_1, \ldots, z_K)$ using linearized property of $f(\cdot)$ (Remark 1). This allows one to recover $q^{K-1}$-degree polynomial $f(\cdot)$, and therefore to reconstruct data vector $(f_1, \cdots, f_K)$, by performing polynomial interpolation. This also establishes that Gabidulin codes are MDS codes.*

### C. Eavesdropper model and proof of secrecy

In [11], Pawar et al. consider a passive eavesdropper with access to the data stored on $\ell$ ($< k$) storage nodes. However, at the MSR point, an eavesdropper with access to data downloaded during node repairs as well may gain more information as repair bandwidth is strictly greater than $\alpha_{\rm msr}$. Keeping this in mind, we adopt the eavesdropper model defined in [3]. We consider an $(\ell_1, \ell_2)$-eavesdropper, which can access the stored data of nodes in the set $\mathcal{E}_1$, and additionally can access both the stored and downloaded data at the nodes in the set $\mathcal{E}_2$ with $|\mathcal{E}_1| = \ell_1$ and $|\mathcal{E}_2| = \ell_2$. The eavesdropper is assumed to know the coding scheme employed by the DSS. We present the definition of achievability of a secure file size in the following.

**Definition 3** (Security against an $(\ell_1, \ell_2)$-eavesdropper)**.** *A DSS is said to achieve a secure file size of $\mathcal{M}^s$ against an $(\ell_1, \ell_2)$-eavesdropper, if, for any sets $\mathcal{E}_1$ and $\mathcal{E}_2$ of size $\ell_1$ and $\ell_2$, respectively, mutual information $I(\mathbf{f}^s; \mathbf{e}) = 0$. Here $\mathbf{f}^s$ denotes the secure file of size $\mathcal{M}^s$, which is first encoded to file $\mathbf{f}$ of size $\mathcal{M}$, and $\mathbf{e} = (\mathbf{s}_{\mathcal{E}_1}, \mathbf{d}_{\mathcal{E}_2})$ represents data observed by the eavesdropper.*

Throughout the paper, we use the following lemma to prove that a coding scheme is secure.

**Lemma 4** (Secrecy Lemma [3], [12])**.** *Consider a system with information symbols $\mathbf{f}^s$, random symbols $\mathbf{r}$ (independent of $\mathbf{f}^s$), and an eavesdropper with observations given by $\mathbf{e}$. If $H(\mathbf{e}) \leq H(\mathbf{r})$ and $H(\mathbf{r}|\mathbf{f}^s, \mathbf{e}) = 0$, then $I(\mathbf{f}^s; \mathbf{e}) = 0$. Here $H$ denotes entropy.*

*Proof.* We have $I(\mathbf{f}^s; \mathbf{e}) = H(\mathbf{e}) - H(\mathbf{e}|\mathbf{f}^s) \overset{(a)}{\leq} H(\mathbf{e}) - H(\mathbf{e}|\mathbf{f}^s) + H(\mathbf{e}|\mathbf{f}^s, \mathbf{r}) \overset{(b)}{\leq} H(\mathbf{r}) - I(\mathbf{e}; \mathbf{r}|\mathbf{f}^s) \overset{(c)}{=} H(\mathbf{r}|\mathbf{f}^s, \mathbf{e}) \overset{(d)}{=} 0$, where (a) follows by non-negativity of $H(\mathbf{e}|\mathbf{f}^s, \mathbf{r})$, (b) is due to $H(\mathbf{e}) \leq H(\mathbf{r})$, (c) follows as $\mathbf{r}$ and $\mathbf{f}^s$ are independent, (d) is due to $H(\mathbf{r}|\mathbf{f}^s, \mathbf{e}) = 0$. □

### D. Locally repairable codes

An $(n, k)$-DSS is said to be an $(r, \delta, \alpha)$ LRC when for each stored block $\mathbf{s}_i$ (of size $\alpha$), there exists a set of nodes $\Gamma(i)$ such that (i) $i \in \Gamma(i)$, (ii) $|\Gamma(i)| \leq r + \delta - 1$, and (iii) minimum distance of $\mathcal{C}|_{\Gamma(i)}$, the code obtained by puncturing $\mathcal{C}$ over $\Gamma(i)$, is at least $\delta$.

**Remark 5.** *Property (iii) implies that each element $j \in \Gamma(i)$ can be written as a function of any set of $r$ elements in $\Gamma(i)$ (not containing $j$). Whereas, properties (ii) and (iii) imply that $H(\Gamma(i)) \leq r\alpha$.*

Distinct sets in $\{\Gamma(i)\}_{i \in [n]}$ are called local groups. We have following general bound on the minimum distance ($d_{\min}$) of an $(n, k, d, r, \delta, \alpha)$ DSS [4], i.e., $(n, k, d)$-DSS with $(r, \delta, \alpha)$ locality,

$$d_{\min}(\mathcal{C}) \leq n - \left\lceil \frac{\mathcal{M}}{\alpha} \right\rceil + 1 - \left( \left\lceil \frac{\mathcal{M}}{r\alpha} \right\rceil - 1 \right)(\delta - 1).$$

Note that $d = r$ is possible as $r$ nodes in $\Gamma(i) \backslash \{i\}$ can be used to repair node $i$. [4], [5] present codes that achieve this bound for general $\delta$ and have $g$ disjoint local groups with $\{\mathcal{G}_i\}_{i \in [g]}$ denoting the set of indices of nodes in $g$ local groups. [2], [13], [14], and [15] present $d_{\min}$-optimal scalar LRCs ($\alpha = 1$). In [16], Papailiopoulos et al. present $d_{\min}$-optimal vector LRCs with a single local parity, i.e, $\delta$ = 2.

## III. Secure Minimum Storage Regenerating Codes

In [11], Pawar et al. establish the following upper bound on the secure file size when an eavesdropper observes the content of $\ell$ nodes.

$$\mathcal{M}^s \leq \sum_{i=\ell+1}^{k} \min\{(d - i + 1)\beta, \alpha\}. \qquad (1)$$

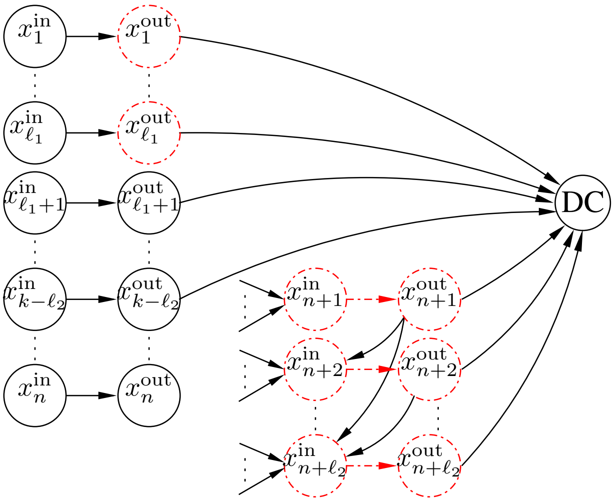
$x_1^{\text{in}}$
$x_1^{\text{out}}$
$x_{\ell_1}^{\text{in}}$
$x_{\ell_1}^{\text{out}}$
$x_{\ell_1+1}^{\text{in}}$
$x_{\ell_1+1}^{\text{out}}$
DC
$x_{k-\ell_2}^{\text{in}}$
$x_{k-\ell_2}^{\text{out}}$
$x_{n+1}^{\text{in}}$
$x_{n+1}^{\text{out}}$
$x_{n+2}^{\text{in}}$
$x_{n+2}^{\text{out}}$
$x_n^{\text{in}}$
$x_n^{\text{out}}$
$x_{n+\ell_2}^{\text{in}}$
$x_{n+\ell_2}^{\text{out}}$

Fig. 1: An information flow graph associated with an $(n, k)$ DSS in the presence of an $(\ell_1, \ell_2)$-eavesdropper. For the eavesdropper, we have $\mathcal{E}_1 = \{\mathbf{x}_1, \ldots, \mathbf{x}_{\ell_1}\}$ and $\mathcal{E}_2 = \{\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+\ell_2}\}$. Here, we assume that $\mathbf{x}_{k-\ell_2+1}, \ldots, \mathbf{x}_k$ fail subsequently in the order specified by their indices and are repaired by introducing nodes $\mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+\ell_2}$ respectively. Data collector (DC) contacts $\mathbf{x}_1, \ldots, \mathbf{x}_{k-\ell_2}, \mathbf{x}_{n+1}, \ldots, \mathbf{x}_{n+\ell_2}$ to reconstruct the original data stored on the DSS.

At the MBR point, when $d = n-1$, Pawar et al. [11] show the tightness of this bound. [3] proposes product matrix based secure coding schemes achieving this bound for any $\ell < k$ at the MBR point with general $d$. However, the product matrix based coding scheme proposed in [3] can only store a secure file size of $(k-\ell_1-\ell_2)(\alpha-\ell_2\beta)$ at the MSR point, where the bound in (1) reduces to $\mathcal{M}^s \leq (k-\ell_1-\ell_2)\alpha$, which concludes that the coding scheme from [3] characterizes secrecy capacity only when $\ell_2 = 0$. In this paper, we improve the upper bound on secrecy capacity against $(\ell_1, \ell_2)$-eavesdropper at the MSR point. We subsequently present a secure coding scheme against $(\ell_1, \ell_2)$-eavesdropper by combining the classical secret sharing scheme [17] with an existing class of exact-MSR codes. The proposed coding scheme has higher rate compared to that proposed in [3] and characterizes the secrecy capacity when $\ell_2 \leq 2$ for any $\ell_1$.

### *A. Improved bound on secrecy capacity at the MSR point*

In this subsection, we utilize the standard approach of computing a cut in information flow graph [1], [11] associated with a DSS in order to get the following (improved) bound on secrecy capacity at the MSR point:

**Theorem 6.** *For an $(n, k)$ MSR code, we have*

$$\mathcal{M}^s \leq \sum_{i=\ell_1+1}^{k-\ell_2} \left( \alpha - \dim \left( \sum_{j=1}^{\ell_2} \mathcal{D}_{i,n+j} \right) \right). \tag{2}$$

*Proof.* We consider the information flow graph and the eavesdropper shown in Fig. 1. The proof follows, compare to that of (1) in [11], by considering leakage from nodes $\ell_1+1, \ldots, k-\ell_2$ to $\ell_2$ node repairs corresponding to $\mathcal{E}_2$. See [4] for details. □

In Theorem 6, $\dim \left(\sum_{j=1}^{\ell_2} \mathcal{D}_{n+j}^{i}\right)$ can be trivially lower bounded by $\beta$ to obtain the following corollary.

**Corollary 7.** *For a DSS employing an $(n, k, d, \alpha, \beta)$ MSR code, we have:*

$$\mathcal{M}^s \leq (k-\ell_1-\ell_2)(\alpha-\beta). \tag{3}$$

This shows that the secure code construction proposed in [3] is optimal for $\ell_2 \leq 1$. Next, we restrict ourselves to exact-MSR codes with $d = n-1$. These codes necessarily require interference alignment for node repairs [18]. For $(n, k, d = n-1)$-DSS, employing systematic exact-MSR code, it follows from Lemma 7 in [4] that, for such codes, we have the following result,

$$\dim \left( \bigcap_{j \in \mathcal{A}} \text{rowspace}(V_{i,j}) \right) \leq \frac{\alpha}{(n-k)^{|\mathcal{A}|}}, \tag{4}$$

where $\mathcal{A} \subseteq [k] \backslash \{i\}$ and $\text{rowspace}(V_{i,j})$ denotes row space of repair matrix $V_{i,j}$. Noting that $\text{rowspace}(V_{i,j}) = \mathcal{D}_{i,j}$, we use (4) to conclude that

$$\dim \left(\mathcal{D}_{i,n+1} + \mathcal{D}_{i,n+2}\right) \geq 2\beta - \frac{\alpha}{(n-k)^2}, \tag{5}$$

Combining (5) with Theorem 6, we get:

**Corollary 8.** *Given an $(n, k, d, \alpha, \beta)$ exact-MSR code with $d = n-1$, for $\ell_2 \leq 2$, we have*

$$\mathcal{M}^s \leq \begin{cases} (k-\ell_1-\ell_2)(\alpha-\beta) & \text{if } \ell_2 = 1, \\ (k-\ell_1-\ell_2)\left(\alpha - 2\beta + \frac{\alpha}{(n-k)^2}\right) & \text{if } \ell_2 = 2. \end{cases} \tag{6}$$

### *B. Construction of secure MSR codes for $d = n-1$*

In this subsection, we present a construction which is based on concatenation of Gabidulin codes [9] and zigzag codes [7] (over $\mathbb{F}_q$). For zigzag codes, $\alpha = p^k$, where $p = n-k$. The node repair for a systematic node (say $j$) is performed by accessing the symbols associated with set $Y_j = \{x \in [0, p^k - 1] : x \cdot e_j = 0\}$ from each surviving node [7], where $e_j$ is an element of the standard basis for $\mathbb{Z}_p^k$, and $x$ is represented with an element of $\mathbb{Z}_p^k$. We first state the following relevant property of zigzag codes.

**Lemma 9.** *For a DSS employing an $(n = k+p, k)$ zigzag code, an $(\ell_1, \ell_2)$-eavesdropper with $\mathcal{E}_2 \subseteq [k]$ can observe only $kp^k - (k-\ell_1-\ell_2)p^k \left(1 - \frac{1}{p}\right)^{\ell_2}$ independent symbols.*

*Proof.* Refer to [4]. □

We now detail the achievability scheme of this section: Take a file $\mathbf{f}^s$ of size $(k-\ell_1-\ell_2)p^k(1-\frac{1}{p})^{\ell_2}$ symbols over $\mathbb{F}_{q^m}$. Append this file with $kp^k - (k-\ell_1-\ell_2)p^k(1-\frac{1}{p})^{\ell_2}$ random symbols uniformly and independently drawn from $\mathbb{F}_{q^m}$. Encode these $kp^k$ symbols using an $[N = kp^k, K = kp^k, D = 1]$ Gabidulin code using a linearized polynomial as specified in

Section II-B. Encode the output of the previous step (codeword from Gabidulin code) to a codeword of an $(n = p + k, k)$ zigzag code (over $\mathbb{F}_q$) with $\alpha = p^k$.

Note that the $k\alpha$ symbols from any $k$ nodes are enough to reconstruct the original data (see Remark 2). Next, we establish the secrecy guarantee of the coding scheme.

**Theorem 10.** *The proposed coding scheme, obtained by Gabidulin precoding of a zigzag code, securely stores a file of size $\mathcal{M}^s = (k - \ell_1 - \ell_2)p^k \left(1 - \frac{1}{p}\right)^{\ell_2}$ against an $(\ell_1, \ell_2)$-eavesdropper with $\mathcal{E}_2 \subseteq [k]$[1]. In addition, when $\ell_2 \leq 2$, the proposed coding scheme attains the upper bound on the secure file size given in Corollary 8, and therefore characterizes the secrecy capacity at the MSR point with $d = n - 1$.*

*Proof.* The proof of security follows by Lemma 9, Lemma 4, and the linearized property of Gabidulin codes. Note that we can invoke linearized property here as the construction uses zigzag codes over $\mathbb{F}_q$ with Gabidulin codes over $\mathbb{F}_{q^m}$. (A similar proof of security when utilizing polynomials for encoding is provided in the seminal paper of A. Shamir on secret sharing [17].) See [4] for a detailed proof.

Substituting $\ell_2 = 1$ (or 2), $\alpha = p^k$, $\beta = \frac{p^k}{p} = p^{k-1}$ and $p = n - k$ in (6) shows that the proposed code construction achieves the upper bound on secure file size, specified in Corollary 8, for $\mathcal{E}_2 \subseteq [k]$ with $\ell_2 \leq 2$. □

## IV. Secrecy in Locally Repairable DSS

In this section, we address the issue of secrecy in $d_{\min}$-optimal locally repairable DSS under the eavesdropper model defined in Section II-C. Before describing our results, we present a short note on the notation, which is specific to the present section. Let $\mathcal{E}_1 = \cup_{i=1}^{g} \mathcal{E}_1^i$ and $\mathcal{E}_2 = \cup_{i=1}^{g} \mathcal{E}_2^i$ be two sets of indices of the nodes observed by an $(\ell_1, \ell_2)$ eavesdropper. Here, $\mathcal{E}_1^i$ ($|\mathcal{E}_1^i| = l_1^i$) and $\mathcal{E}_2^i$ ($|\mathcal{E}_2^i| = l_2^i$) denote the sets of indices of storage-eavesdropped and download-eavesdropped nodes in local group $i$ respectively. Note that we have $\sum_{i=1}^{g} l_1^i = \ell_1$, and $\sum_{i=1}^{g} l_2^i = \ell_2$. A DC is associated with the indices of nodes, $\mathcal{K} = \cup_{i=1}^{g} \mathcal{K}_i$ with $|\mathcal{K}| \leq n - d_{\min} + 1$, it contacts to reconstruct the original file. Here $\mathcal{K}_i$ denotes the set of indices of nodes that the DC contacts in local group $i$.

We first derive a generic upper bound on the secrecy capacity of an $(r, \delta, \alpha)$ LRC, which we later specialize for specific cases of system parameters. While addressing specific cases, we also present secure code constructions that achieve the respective upper bound for certain set of system parameters.

**Theorem 11.** *For a DSS employing an $(r, \delta, \alpha)$ LRC that is secure against an $(\ell_1, \ell_2)$-eavesdropper, we have*

$$\mathcal{M}^s \leq \sum_{i=1}^{g} H(\mathbf{s}_{\mathcal{K}_i} | \mathbf{s}_{\mathcal{E}_1^i}, \mathbf{d}_{\mathcal{E}_2^i}) \;\; \forall \left(\{\mathcal{E}_1^i, \mathcal{E}_2^i, \mathcal{K}_i\}_{i=1}^{g}\right) \in \mathcal{X}, \quad (7)$$

*where $\mathcal{X}$ denotes the set of tuples $(\{\mathcal{E}_1^i, \mathcal{E}_2^i, \mathcal{K}_i\}_{i=1}^{g})$ that are allowed under our model.*

*Proof.* For a tuple $(\{\mathcal{E}_1^i, \mathcal{E}_2^i, \mathcal{K}_i\}_{i=1}^{g})$, the upper bound follows by subtracting the leakage associated with the eavesdropper $(I(\mathbf{s}_{\mathcal{K}_i}; \mathbf{s}_{\mathcal{E}_1^i}, \mathbf{d}_{\mathcal{E}_2^i}))$ from data observed at DC from local group $i$ $(H(\mathbf{s}_{\mathcal{K}_i}))$. See [4] for details. □

Next, we consider two cases depending on the number of local parities per local group: (i) single parity node per local group ($\delta = 2$) and (ii) multiple parity nodes per local group ($\delta > 2$). In both cases, vectors $\mathbf{l}_1 = (l_1^1, \ldots, l_1^g)$ and $\mathbf{l}_2 = (l_2^1, \ldots, l_2^g)$ denote patterns of eavesdropped nodes. In what follows, we use $\tau$ and $h$ as short-hand notation for $\left\lfloor \frac{n - d_{\min} + 1}{r + \delta - 1} \right\rfloor$ and $n - d_{\min} + 1 - (r + \delta - 1) \left\lfloor \frac{n - d_{min} + 1}{r + \delta - 1} \right\rfloor$, respectively.

### A. Case 1: Single local parity per local group ($\delta = 2$)

For an LRC with single local parity per group, all the information stored in a local group is revealed to an eavesdropper that observes a node repair in the local group as a newcomer node downloads all the data stored on other $(r)$ nodes in the local group it belongs to. Therefore, we have $H(\mathbf{s}_{\mathcal{G}_i} | \mathbf{d}_{\mathcal{E}_2^i}) = 0 \Rightarrow H(\mathbf{s}_{\mathcal{K}_i} | \mathbf{d}_{\mathcal{E}_2^i}) = 0$, when $\mathcal{E}_2^i \neq \emptyset$. We use this fact to present the following result on secrecy capacity of an LRC with $\delta = 2$.

**Theorem 12.** *Secrecy capacity of an $(r, \delta = 2, d = r, \alpha)$ LRC, against an $(\ell_1, \ell_2)$-eavesdropper, is*

$$\mathcal{M}^s = [\tau r + h - (\ell_2 r + \ell_1)]^+ \alpha. \quad (8)$$

*Proof.* In order to get the upper bound (stated as RHS of (8)) on secrecy capacity, consider a DC ($\mathcal{K}$) with $\mathcal{K}_1 = \mathcal{G}_1, \mathcal{K}_2 = \mathcal{G}_2, \ldots, \mathcal{K}_\tau = \mathcal{G}_\tau, \mathcal{K}_{\tau+2} = \ldots = \mathcal{K}_g = \emptyset, \mathcal{K}_{\tau+1} \subset \mathcal{G}_{\tau+1}$ s.t. $|\mathcal{K}_{\tau+1}| = h$; and an eavesdropper with eavesdropping pattern $\mathbf{l}_2 = (1, 1, \ldots, 1, 0, \ldots, 0)$ with ones at first $\ell_2$ positions and $\mathbf{l}_1 = (0, \ldots, 0, l_1^{\ell_2+1}, \ldots, l_1^g)$ with zeros in first $\ell_2$ positions. This eavesdropping pattern and DC along with (7) give the upper bound on $\mathcal{M}^s$, RHS of (8), for an $(r, \delta = 2, \alpha)$ LRC.

Next, we establish Theorem 12 by presenting a secure coding scheme that shows tightness of the upper bound stated above. Append $(\ell_2 r + \ell_1)\alpha$ random symbols (independent of file $\mathbf{f}^s$) over $\mathbb{F}_{q^m}$, $\mathbf{r} = (r_1, \ldots, r_{(\ell_2 r + \ell_1)\alpha})$, with $(\tau r + h - (\ell_2 r + \ell_1))\,\alpha$ symbols of file $\mathbf{f}^s$. Encode these $\mathcal{M} = (\tau r + h)\,\alpha$ symbols (including both $\mathbf{r}$ and $\mathbf{f}^s$) using an $[\mathcal{M}, \mathcal{M}, 1]$ Gabidulin code following encoding process specified in Section II-B. Encode $\mathcal{M}$ symbols of the Gabidulin codeword with a $d_{\min}$-optimal $(r, \delta = 2, \alpha)$ LRC, e.g., coding scheme proposed in [16].

To prove secrecy of the proposed scheme against an $(\ell_1, \ell_2)$-eavesdropper, we use Lemma 4. In particular, the coding scheme meets two sufficient conditions (i) $H(\mathbf{e}) \leq H(\mathbf{r})$ (eavesdropper observes at most $(\ell_2 r + \ell_1)\alpha$ independent symbols) and (ii) $H(\mathbf{r} | \mathbf{f}^s, \mathbf{e}) = 0$, the eavesdropper can decode $\mathbf{r}$

[1] A high rate MSR coding schemes has a small number of parity-check nodes. Therefore the assumption that $\mathcal{E}_2$ lies in the set of systematic nodes is not detrimental to our contributions as small number of parity nodes may have additional mechanisms in place for secure node repairs. Here, we point out that, for $\ell_2 = 1$ case, the proposed scheme is optimal even when repair of a parity node is eavesdropped.

given its observed data and $\mathbf{f}^s$ (using Remark 1 and 2). See [4] for a detailed proof. □

Theorem 12 shows that the performance of an LRC with $\delta = 2$ degrade substantially in the presence of an eavesdropper that can observe node repairs, i.e., $\ell_2 > 0$.

### *B. Case 2: Multiple local parities per local group ($\delta > 2$)*

For LRCs with $\delta > 2$ that allow only naïve node repair, i.e., a newcomer downloads all the information from $r$ out of $r + \delta - 2$ surviving nodes, the characterization of secrecy capacity is similar to the previous case and therefore omitted due to lack of space. The main aim in the present section is to show that using regenerating codes within local groups (when $\delta > 2$) can improve the secrecy capacity of DSS against $(\ell_1, \ell_2)$-eavesdropper. Here, we focus only on those LRCs that, when restricted to a local group, give an MSR code. (The analysis for MBR codes is similar and will be presented in a future work.)

In the following, we assume that node repairs are performed with a newcomer downloading $\beta^{\mathrm{loc}}_{\mathrm{msr}}$ symbols from each of $d$ surviving node of its own local group.

**Theorem 13.** *For an $(r, \delta > 2, \alpha)$ LRC with MSR codes as local codes, secrecy capacity against $(\ell_1, \ell_2)$-eavesdropper satisfies*

$$\begin{aligned}\mathcal{M}^s \leq \sum_{i=1}^{\rho}&(r - (l_1^i + s + 1))\left(\alpha - \theta(\alpha, \beta^{\mathrm{loc}}_{\mathrm{msr}}, s + 1)\right)\\ &+ \sum_{i=\rho+1}^{\tau} (r - (l_1^i + s))\left(\alpha - \theta(\alpha, \beta^{\mathrm{loc}}_{\mathrm{msr}}, s)\right)\\ &+ (\min\{r, h\} - (l_1^i + \nu))\left(\alpha - \theta(\alpha, \beta^{\mathrm{loc}}_{\mathrm{msr}}, \nu)\right). \qquad (9)\end{aligned}$$

*where $s, \rho,$ and $\nu$ are positive integers such that $0 \leq \rho + \nu \leq s$, $\nu \leq h$, and $\ell_2 = s\tau + \rho + \nu$. Here $\theta(\alpha, \beta^{\mathrm{loc}}_{\mathrm{msr}}, t)$ denotes the amount of information that an eavesdropper receives from one intact node (a node not eavesdropped) during the repair of $|\mathcal{E}_2^i| = t$ nodes in the $i$th local group.*

*Moreover, the above bound is tight and characterizes secrecy capacity for the LRC, when $d = r + \delta - 2$ and $\ell_2 \leq 2\left\lfloor \frac{n - d_{min} + 1}{r + \delta - 1} \right\rfloor + \min\{2, h\}$.*

*Proof.* For an LRC with local MSR codes, we apply Theorem 6 to obtain

$$H(\mathbf{s}_{\mathcal{K}_i} | \mathbf{s}_{\mathcal{E}_1^i}, \mathbf{d}_{\mathcal{E}_2^i}) \leq \sum_{j=1}^{\min(|\mathcal{K}_i|, r) - l^i} (\alpha - \theta(\alpha, \beta^{\mathrm{loc}}_{\mathrm{msr}}, l_2^i)), \qquad (10)$$

where $l^i = l_1^i + l_2^i$. Next, we consider the DC associated with the pattern $(\mathcal{K}_1, \ldots, \mathcal{K}_g)$ used in Section IV-A, and the eavesdropping pattern $\mathbf{l}_2$ such that $l_2^1 = \ldots = l_2^\rho = s + 1$, $l_2^{\rho+1} = \ldots = l_2^\tau = s$, $l_2^{\tau+2} = \ldots = l_2^g = 0$, and $l_2^{\tau+1} = \nu$. Given this particular choice of a DC and an eavesdropper, using Theorem 11 and (10), we get the upper bound in (9).

Now, we restrict ourselves to LRCs with $d = r + \delta - 2$. For such codes, it follows from Lemma 7 in [4] (similar to Corollary 8) that, for $l_2^i \leq 2$,

$$\theta(\alpha, \beta^{\mathrm{loc}}_{\mathrm{msr}}, l_2^i) \geq \begin{cases} \beta^{\mathrm{loc}}_{\mathrm{msr}}, & \text{if } l_2^i = 1 \\ 2\beta^{\mathrm{loc}}_{\mathrm{msr}} - \frac{\alpha}{(\delta-1)^2}, & \text{if } l_2^i = 2 \end{cases} \qquad (11)$$

Next, we sketch a secure coding scheme against an eavesdropper when $\ell_2 \leq 2\tau + \min\{2, h\}$ and $l_2^i \leq 2$. Take a file $\mathbf{f}^s$ of size $\mathcal{M}^s$ (over $\mathbb{F}_{q^m}$) equal to the RHS of (9) and append this to $\mathcal{M} - \mathcal{M}^s = \tau r\alpha + \min\{h, r\}\alpha - \mathcal{M}^s$ i.i.d. uniform random symbols (independent of $\mathbf{f}^s$) from $\mathbb{F}_{q^m}$. Encode these $\mathcal{M}$ symbols (secure data symbols and random symbols) using two stage encoding scheme for vector LRCs presented in [4]: (i) Encode $\mathcal{M}$ symbols to a $gr\alpha$-long Gabidulin codeword over $\mathbb{F}_{q^m}$. (ii) Partition $gr\alpha$ symbols of the Gabidulin codeword into $g$ disjoint groups of size $r\alpha$ each. Then, apply an $(r + \delta - 1, r, \alpha)$ zigzag code (over $\mathbb{F}_q$) to each group of $r\alpha$ symbols. Note that $g(r + \delta - 1) = n$. The proof of secrecy of the proposed scheme is similar to that in Theorem 12. See [4] for a detailed proof. □

## REFERENCES

[1] A. Dimakis, P. Godfrey, Y. Wu, M. Wainwright, and K. Ramchandran, "Network coding for distributed storage system," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4539-4551, Sep. 2010.

[2] P. Gopalan, C. Huang, H. Simitchi, and S. Yekhanin, "On the locality of codeword symbols," *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6925-6934, Nov. 2012.

[3] N. B. Shah, K. V. Rashmi, and P. V. Kumar, "Information-theoretically secure regenerating codes for distributed storage," in *Proc. of IEEE Globecom*, Dec. 2011.

[4] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath, "Optimal locally repairable and secure codes for distributed storage systems," *CoRR*, vol. abs/1210.6954, Oct. 2012.

[5] G. M. Kamath, N. Prakash, V. Lalitha, and P. V. Kumar, "Codes with local regeneration," *CoRR*, vol. abs/1211.1932, Nov. 2012.

[6] K. V. Rashmi, N. B. Shah, and P. V. Kumar, "Optimal exact-regenerating codes for distributed storage at the MSR and MBR point via a product-matrix construction," *IEEE Trans. Inf. Theory*, vol. 57, no. 57, pp. 5227-5239, Aug. 2011.

[7] I. Tamo, Z. Wang, and J. Bruck, "Zigzag codes: MDS array codes with optimal rebuilding," *IEEE Trans. Inf. Theory*, vol. 59, no. 3, pp. 1597-1616, Mar. 2013.

[8] Z. Wang, I. Tamo, and J. Bruck, "Long MDS codes for optimal repair bandwidth," in *Proc. of IEEE ISIT*, Jul. 2012.

[9] E. M. Gabidulin, "Theory of codes with maximum rank distance," *Problems of Information Transmission*, vol. 21, pp. 1-12, July 1985.

[10] F. J. MacWilliams and N. J. A. Sloane, *The theory of error-correcting codes*, North-Holland, 1978.

[11] S. Pawar, S. El Rouayheb, and K. Ramchandran, "Securing dynamic distributed storage systems against eavesdropping and adversarial attacks," *IEEE Trans. Inf. Theory*, vol. 57, no. 9, pp. 6734-6753, Sep. 2011.

[12] A. Wyner, "The wire-tap channel," *The Bell System Technical Journal*, vol. 54, no. 8, pp. 1355-1387, Oct. 1975.

[13] C. Huang, M. Chen, and J. Li, "Pyramid code: flexible schemes to trade space for access efficiency in reliable data storage systems," in *NCA*, 2007.

[14] N. Prakash, G. M. Kamath, V. Lalitha, and P. V. Kumar, "Optimal linear codes with a local-error-correction property," in *Proc. of IEEE ISIT*, Jul. 2012.

[15] N. Silberstein, A. S. Rawat, and S. Vishwanath, "Error resilience in distributed storage via rank-metric codes," in *Proc. of 50th Allerton*, Oct. 2012.

[16] D. S. Papailiopoulos and A. G. Dimakis, "Locally repairable codes," in *Proc. of IEEE ISIT*, Jul. 2012.

[17] A. Shamir, "How to share a secret," *Communications of the ACM*, vol. 22, no. 11, pp. 612-613, Nov. 1979.

[18] N. B. Shah, K. Rashmi, P. V. Kumar, and K. Ramchandran, "Interference alignment in regenerating codes for distributed storage: Necessity and code constructions," *IEEE Trans. Inf. Theory*, vol. 58, no. 4, pp. 2134-2158, Apr. 2012.