



# CHALMERS

## Chalmers Publication Library

### Secure Repairable Fountain Codes

This document has been downloaded from Chalmers Publication Library (CPL). It is the author's version of a work that was accepted for publication in:

**IEEE Communications Letters (ISSN: 1089-7798)**

Citation for the published paper:

Kumar, S. ; Rosnes, E. ; Graell i Amat, A. (2016) "Secure Repairable Fountain Codes". IEEE Communications Letters

Downloaded from: <http://publications.lib.chalmers.se/publication/236930>

Notice: Changes introduced as a result of publishing processes such as copy-editing and formatting may not be reflected in this document. For a definitive version of this work, please refer to the published source. Please note that access to the published version might require a subscription.

Chalmers Publication Library (CPL) offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all types of publications: articles, dissertations, licentiate theses, masters theses, conference papers, reports etc. Since 2006 it is the official tool for Chalmers official publication statistics. To ensure that Chalmers research results are disseminated as widely as possible, an Open Access Policy has been adopted. The CPL service is administrated and maintained by Chalmers Library.

(article starts on next page)

# Secure Repairable Fountain Codes

Siddhartha Kumar, *Student Member, IEEE*, Eirik Rosnes, *Senior Member, IEEE*,  
and Alexandre Graell i Amat, *Senior Member, IEEE*

**Abstract**—In this letter, we provide the construction of repairable fountain codes (RFCs) for distributed storage systems that are information-theoretically secure against an eavesdropper that has access to the data stored in a subset of the storage nodes and the data downloaded to repair an additional subset of storage nodes. The security is achieved by adding random symbols to the message, which is then encoded by the concatenation of a Gabidulin code and an RFC. We compare the achievable code rates of the proposed codes with those of secure minimum storage regenerating codes and secure locally repairable codes.

## I. INTRODUCTION

**T**he design of information-theoretically secure distributed storage systems (DSSs) has attracted a significant interest in the last few years [1], [2]. DSSs use erasure correcting codes (ECCs) to yield fault tolerance against storage node failures. The resiliency of the DSS against passive attacks is a good measure of its security. Passive attacks are those where the attacker (referred to as the eavesdropper) gains access to a subset of storage nodes and thereby to partial information on the data stored on the DSS. Information-theoretic security against such attacks involves mixing of information symbols (called the *message*) with random symbols, prior to encoding by an ECC, in a manner such that the eavesdropper does not gain any information about the original message even if he has access to some code symbols.

Using this idea, [1], [2] provided explicit constructions of minimum storage regenerating (MSR) codes that achieve security for an  $(\ell_1, \ell_2)$  eavesdropper model where the eavesdropper has access to the content of  $\ell_1$  storage nodes and the data that needs to be downloaded to repair  $\ell_2$  additional storage nodes. The design of secure locally repairable codes (LRCs) was also addressed in [2]. In particular, to achieve security, random symbols are appended to the message and the resulting vector of symbols is precoded by a Gabidulin code [3] prior to encoding by an LRC (or MSR code) in [2]. Achieving security comes at the expense of a lower code rate with respect to the original LRC (or MSR code), due to appending random symbols to the message [2]. For the LRC- and MSR-based secure codes, the authors in [2] derived the maximum message size (equivalently, the maximum code rate) that allows to achieve security. Moreover, the code constructions in [2] achieve this maximum. A sufficient condition for the

information leakage to the eavesdropper to be zero was also given in [1], [2].

LRCs [4] and MSR codes [5] are appealing code families because they are repair efficient. Repairable fountain codes (RFCs) are another class of repair-efficient ECCs [6]. Like LRCs, they yield a good locality, which implies that few storage nodes are involved in the repair of a failed node.

In this letter, we present the construction of RFCs that are information-theoretically secure for the  $(\ell_1, \ell_2)$  eavesdropper model. As in [2], we achieve security by appending random symbols to the message and precoding by a Gabidulin code. We prove that the proposed code construction is completely secure for the  $(\ell_1, \ell_2)$  eavesdropper model. To prove security, we give a necessary condition for the information leakage to the eavesdropper to be zero, thus extending the sufficient condition in [1], [2]. Our proof differs from the one in [1], [2] and is based on simple information theory equalities. We compare the achievable code rates (the maximum code rate that allows to achieve security) of the proposed codes with those of secure MSR codes and LRCs in [2]. We show that, for a given rate of the underlying code (RFC, LRC, or MSR code), secure RFCs yield the same achievable code rates as those of secure LRCs and better than those of secure MSR codes when the rate of the underlying code is high enough.

## II. SYSTEM MODEL

We consider a DSS with  $n$  storage nodes, each storing one symbol. A message  $\mathbf{m} = (m_1, m_2, \dots, m_k)$ , of length  $k$  symbols  $m_i \in \text{GF}(q^p)$ ,  $i = 1, \dots, k$ , where  $q$  is a prime and  $p$  is a positive integer, is first encoded using an  $(n, k)$  ECC of rate  $R = k/n$  into a codeword  $\mathbf{c} = (c_1, c_2, \dots, c_n)$  of length  $n$ . Each of the  $n$  code symbols  $c_i$ ,  $i = 1, \dots, n$ , is then stored into a different storage node. We assume that code symbol  $c_i$  is stored in the  $i$ th storage node and, with a slight abuse of notation, we will refer to both code symbol and storage node  $i$  by  $c_i$ .

**Example 1.** *The bipartite graph shown in Fig. 1(a) represents a message stored on a DSS with  $n = 6$  storage nodes using a  $(6, 4)$  ECC. Each code symbol  $c_i$ ,  $i = 1, \dots, 6$ , is a linear combination of its neighboring message symbols  $m_i$ ,  $i = 1, \dots, 4$  (circles). Each code symbol (squares) is stored on a different storage node.*

### A. Security Model

We consider an  $(\ell_1, \ell_2)$  eavesdropper model [2], where the eavesdropper can passively observe, but not modify, the content of  $\ell = \ell_1 + \ell_2 < k$  storage nodes. Out of the  $\ell$  nodes, the eavesdropper can observe the symbols stored

The work of S. Kumar and E. Rosnes was partially funded by the Research Council of Norway (grant 240985/F20) and by Simula@UiB. A. Graell i Amat was supported by the Swedish Research Council under grant #2011-5961.

S. Kumar and E. Rosnes are with the Department of Informatics, University of Bergen, N-5020 Bergen, Norway, and the Simula Research Lab (e-mail: kumarsi@simula.no; eirik@ii.uib.no).

A. Graell i Amat is with the Department of Signals and Systems, Chalmers University of Technology, SE-41296 Gothenburg, Sweden (e-mail: alexandre.graell@chalmers.se).

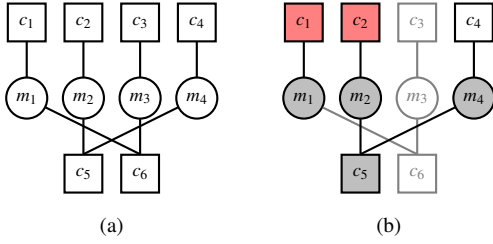


Fig. 1: A DSS with 6 storage nodes employing a  $(6, 4)$  ECC.  $\mathbf{m} = (m_1, m_2, m_3, m_4)$  is encoded into the codeword  $\mathbf{c} = (c_1, c_2, c_3, c_4, c_5, c_6)$ . Each code symbol is stored in a storage node. (a) Bipartite graph of the  $(6, 4)$  ECC; squares and circles represent code symbols (storage nodes) and message symbols, respectively. (b) Example of a  $(1, 1)$  eavesdropper, where  $\mathcal{S}_1 = \{c_1\}$  and  $\mathcal{S}_2 = \{c_2\}$  (in red). Gray symbols are the symbols that the eavesdropper obtains.

in a subset of  $\ell_1$  storage nodes, which we denote by  $\mathcal{S}_1$  ( $|\mathcal{S}_1| = \ell_1$ ). Furthermore, it can observe the data downloaded during the repair of a subset of  $\ell_2$  storage nodes, denoted by  $\mathcal{S}_2$  ( $|\mathcal{S}_2| = \ell_2$ ), where  $\mathcal{S}_1 \cap \mathcal{S}_2 = \emptyset$ . This model is relevant in the scenario where nodes are located at different geographical locations. Peer-to-peer storage systems are examples of such DSSs [1]. We denote the subset of storage nodes from which data is downloaded to repair storage nodes in  $\mathcal{S}_2$  by  $\mathcal{S}_d$ . We will refer to the symbols the eavesdropper obtains as the *eavesdropped symbols*. We also assume that the eavesdropper has perfect knowledge of the ECC used for encoding.

**Definition 1** ([1], [2]). *Let  $\mathbf{e}$  be the vector of eavesdropped symbols that the eavesdropper obtains from the storage nodes in  $\mathcal{S}_1 \cup \mathcal{S}_d$ . A DSS storing a message  $\mathbf{m}$  (possibly encoded by an ECC) is said to be completely secure against an  $(\ell_1, \ell_2)$  eavesdropper if the mutual information between the message and the eavesdropped symbols is zero, i.e.,  $I(\mathbf{m}; \mathbf{e}) = 0$ .*

**Example 2.** *Fig. 1(b) shows an example of a  $(1, 1)$  eavesdropper where  $\mathcal{S}_1 = \{c_1\}$  and  $\mathcal{S}_2 = \{c_2\}$ . Thus, the eavesdropper obtains  $c_1 = m_1$  and the downloaded data  $c_5 = m_2 + m_4$  and  $c_4 = m_4$ , and thereby  $m_2$ , during the repair of  $\mathcal{S}_2 = \{c_2\}$ . In all, the eavesdropper obtains the symbols  $m_1, m_2, m_4$ , and  $c_5 = m_2 + m_4$ , colored in gray in the figure.*

### III. GABIDULIN AND REPAIRABLE FOUNTAIN CODES

We summarize Gabidulin codes and RFCs, which are the building blocks of the secure RFCs presented in Section IV.

#### A. Gabidulin Codes

Gabidulin codes are a class of rank codes [3]. An  $(N, K)$  Gabidulin code (over  $\text{GF}(q^p)$ ) of length  $N$ , dimension  $K$ , and minimum rank distance  $D_{\min}$ , can correct up to  $D_{\min} - 1$  rank erasures. Gabidulin codes are maximum rank distance codes, i.e., they achieve the Singleton bound,  $D_{\min} \leq N - K + 1$ , and are obtained by evaluations of polynomials. More specifically, Gabidulin codes use linearized polynomials.

**Definition 2.** *A linearized polynomial  $f(y)$  of degree  $t > 0$  over  $\text{GF}(q^p)$  has the form*

$$f(y) = \sum_{i=0}^t a_i y^{q^i},$$

where  $a_i \in \text{GF}(q^p)$  and  $a_t \neq 0$ .

A message  $\mathbf{m} = (m_1, \dots, m_K)$  is encoded using an  $(N, K)$  Gabidulin code as follows.

- 1) Construct a polynomial  $f(y) = \sum_{i=1}^K m_i y^{q^{i-1}}$ .
- 2) Evaluate  $f(y)$  at  $N$  linearly independent (over  $\text{GF}(q)$ ) points  $\{y_1, \dots, y_N\} \subset \text{GF}(q^p)$  to obtain a codeword  $(f(y_1), \dots, f(y_N))$ .

Decoding proceeds as follows.

- 1) Obtain any  $K$  evaluations at  $K$  linearly independent (over  $\text{GF}(q)$ ) points. Otherwise, decoding fails.
- 2) Perform polynomial interpolation on the  $K$  evaluations and recover the original message  $\mathbf{m}$  by solving a system of linear equations.

#### B. Repairable Fountain Codes

An  $(n, k)$  systematic RFC encodes a message  $\mathbf{m} = (m_1, \dots, m_k) \in \text{GF}(q^p)$ ,  $q > k$ , into a codeword  $\mathbf{c} = (c_1, \dots, c_n)$ , where  $c_i = m_i$  for  $i = 1, \dots, k$ . The parity symbols  $c_i$ ,  $i = k + 1, \dots, n$ , are constructed according to the following three-step procedure.

- 1) Successively select  $\xi = O(\log k)$  message symbols independently and uniformly at random with replacement.
- 2) For each of the  $\xi$  message symbols, a coefficient is drawn uniformly at random from  $\text{GF}(q) \subset \text{GF}(q^p)$ .
- 3) The parity symbol is then obtained as the linear combination of the  $\xi$  chosen message symbols, weighted by the corresponding coefficients.

Each of the  $n$  code symbols is stored in a different storage node. From the code construction, each parity symbol is a weighted sum of at most  $\xi$  message symbols. A parity symbol and the corresponding (at most)  $\xi$  message symbols is referred to as a local group. The existence of local groups is a hallmark of any ECC having low locality. Unlike LRCs, which have only disjoint local groups, RFCs also have overlapping local groups [6]. Furthermore, for each systematic symbol there exist a number of disjoint local groups from which it can be reconstructed. This allows multiple parallel reads of the systematic symbol, accessing the disjoint local groups. When a storage node fails, it is repaired from one of its local groups. This requires the download of at most  $\xi$  symbols (from the other at most  $\xi$  nodes of the local group). Thus, RFCs have low locality,  $\xi$ , and their repair bandwidth is  $\xi p \log q$ . Also, RFCs are near maximum distance separable codes.

### IV. SECURE REPAIRABLE FOUNTAIN CODES

In this section, we present the construction of RFCs that are secure against the  $(\ell_1, \ell_2)$  eavesdropper model. The proposed secure RFCs are obtained by concatenating a Gabidulin code and an RFC. More precisely, consider an  $(n, \tilde{k})$  RFC such that each parity symbol is a random linear combination of up to  $\xi$  randomly chosen input symbols. Let  $\mathbf{m}$  denote the message of length  $k = \tilde{k} - \ell_1 - \xi \ell_2$  symbols. A codeword of the proposed secure RFC is constructed as follows.

- 1) Append to  $\mathbf{m}$  a random vector  $\mathbf{r} = (r_1, \dots, r_u)$  of length  $u = \ell_1 + \xi \ell_2$  symbols, drawn independently and

uniformly at random from  $\text{GF}(q^p)$ , thus obtaining the vector  $\tilde{\mathbf{m}} = (\mathbf{m}, \mathbf{r})$ .

- 2) *Outer code.* Encode  $\tilde{\mathbf{m}}$  using a  $(\tilde{k}, \tilde{k})$  Gabidulin code to obtain the intermediate codeword  $\tilde{\mathbf{c}} = (\tilde{c}_1, \dots, \tilde{c}_{\tilde{k}}) = (f(y_1), \dots, f(y_{\tilde{k}}))$ .
- 3) *Inner code.* Encode  $\tilde{\mathbf{c}}$  using an  $(n, \tilde{k})$  RFC into the codeword  $\mathbf{c} = (c_1, \dots, c_n)$ . The  $n$  code symbols are then stored in  $n$  storage nodes.

**Remark 1.** A  $\text{GF}(q)$ -linear combination of evaluations of a linearized polynomial  $f(y) = \sum_{i=0}^t a_i y^{q^i}$  of some degree  $t$  over  $\text{GF}(q^p)$  (see Definition 2) is itself an evaluation of the same linearized polynomial. In particular,  $\sum_{j=1}^{\kappa} \gamma_j f(\beta_j) = f(\sum_{j=1}^{\kappa} \gamma_j \beta_j)$ , where  $\kappa$  is a positive integer,  $\gamma_j \in \text{GF}(q)$ , and  $\beta_j \in \text{GF}(q^p)$ , i.e.,  $f(\cdot)$  is a linear map over  $\text{GF}(q)$  [2, Remark 8]. Thus, each code symbol  $c_i$ ,  $i = 1, \dots, n$ , is an evaluation of a linearized polynomial  $f(\cdot)$  of degree at most  $\tilde{k} - 1$  and with coefficients from  $\tilde{\mathbf{m}}$  at some point  $y_i \in \text{GF}(q^p)$ , i.e.,  $c_i = f(y_i) = \sum_{j=1}^{\tilde{k}} \tilde{m}_j y_i^{q^{j-1}}$ .

**Example 3.** Fig. 2 depicts a toy example of a  $(20, 10)$  secure RFC for a  $(1, 1)$  eavesdropper. Here,  $\tilde{\mathbf{m}}$  comprises  $k = \tilde{k} - \ell_1 - \xi \ell_2 = 6$  message symbols and  $u = \ell_1 + \xi \ell_2 = 4$  random symbols.  $\tilde{\mathbf{m}}$  is encoded using the concatenation of a  $(10, 10)$  Gabidulin code and a  $(20, 10)$  RFC. Due to the outer encoding by the Gabidulin code, each code symbol  $c_i$ ,  $i = 1, \dots, 20$ , is an evaluation of a linearized polynomial (see Remark 1). Another consequence is that the final code retains the repair properties of the inner code (the RFC). For this example, the code locality is  $\xi = 3$ .

In the following, we show that the proposed secure RFCs achieve complete security for the  $(\ell_1, \ell_2)$  eavesdropper model. We first prove a sufficient and necessary condition for  $I(\mathbf{m}; \mathbf{e}) = 0$  using an alternative proof to the one in [1], [2].

**Theorem 1.** Let  $\mathbf{m}$  be a message which is stored in a DSS by first appending to it a vector  $\mathbf{r}$  of random symbols and then encoding  $(\mathbf{m}, \mathbf{r})$  by an ECC. Also, let  $\mathbf{e}$  be the vector of code symbols the eavesdropper has access to. Then, the information leakage to the eavesdropper is zero, i.e.,  $I(\mathbf{m}; \mathbf{e}) = 0$ , if and only if  $H(\mathbf{r}|\mathbf{m}, \mathbf{e}) = H(\mathbf{r}) - H(\mathbf{e})$ .

*Proof:* We prove the theorem using simple information theory equalities,

$$\begin{aligned}
 I(\mathbf{m}; \mathbf{e}) &= H(\mathbf{m}) - H(\mathbf{m}|\mathbf{e}) \\
 &\stackrel{(a)}{=} H(\mathbf{m}) - H(\mathbf{m}|\mathbf{e}) + H(\mathbf{e}|\mathbf{m}, \mathbf{r}) \\
 &= H(\mathbf{m}) - H(\mathbf{m}|\mathbf{e}) + H(\mathbf{e}|\mathbf{m}) - I(\mathbf{r}; \mathbf{e}|\mathbf{m}) \\
 &\stackrel{(b)}{=} H(\mathbf{e}) - I(\mathbf{r}; \mathbf{e}|\mathbf{m}) \\
 &= H(\mathbf{e}) - H(\mathbf{r}|\mathbf{m}) + H(\mathbf{r}|\mathbf{e}, \mathbf{m}) \\
 &\stackrel{(c)}{=} H(\mathbf{e}) - H(\mathbf{r}) + H(\mathbf{r}|\mathbf{e}, \mathbf{m}),
 \end{aligned}$$

where (a) follows from the fact that  $H(\mathbf{e}|\mathbf{m}, \mathbf{r}) = 0$ , since eavesdropped symbols are a function of  $\mathbf{m}$  and  $\mathbf{r}$ , (b) follows from  $H(\mathbf{e}) - H(\mathbf{e}|\mathbf{m}) = H(\mathbf{m}) - H(\mathbf{m}|\mathbf{e})$ , and (c) follows from the fact that  $\mathbf{r}$  and  $\mathbf{m}$  are stochastically independent of each other, i.e.,  $H(\mathbf{r}|\mathbf{m}) = H(\mathbf{r})$ . Thus,

$$I(\mathbf{m}; \mathbf{e}) = 0 \Leftrightarrow H(\mathbf{r}|\mathbf{m}, \mathbf{e}) = H(\mathbf{r}) - H(\mathbf{e}). \quad (1)$$

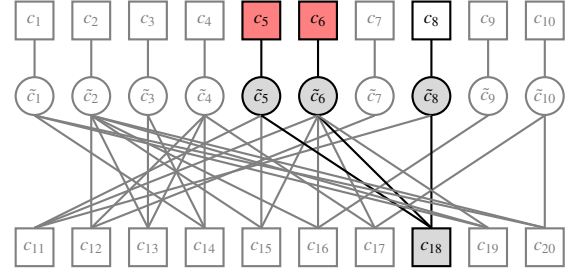


Fig. 2: A  $(20, 10)$  secure RFC. Storage nodes  $\mathcal{S}_1 = \{c_6\}$  and  $\mathcal{S}_2 = \{c_5\}$  are eavesdropped. Gray symbols are the symbols that the eavesdropper obtains. ■

We remark that in [1] and [2, Lem. 4] a sufficient condition on  $I(\mathbf{m}; \mathbf{e}) = 0$  was proved, whereas Theorem 1 gives a sufficient and necessary condition. ECCs for which Theorem 1 is satisfied do not leak any information, i.e., they are completely secure. In Theorem 2 below, we use the following lemma to prove that our proposed code construction is completely secure for the  $(\ell_1, \ell_2)$  eavesdropper model.

**Lemma 1.** Consider the  $(\ell_1, \ell_2)$  eavesdropper model. For the proposed code construction (with  $u = \ell_1 + \xi \ell_2$  random symbols  $\mathbf{r}$ ),  $H(\mathbf{e}) \leq H(\mathbf{r}) = (\ell_1 + \xi \ell_2)p \log q$ , where  $\mathbf{e}$  is the vector of code symbols the eavesdropper has access to.

*Proof:* Consider the repair of a single storage node  $c_i$  in  $\mathcal{S}_2$ , and let  $\Gamma^{(i)}$  denote the local group (there are many) used for the repair of storage node  $c_i$ . Each local group contains one inner code parity symbol and at most  $\xi$  inner code message symbols to which it is connected. Thus,  $|\Gamma^{(i)}| \leq \xi + 1$ . Since the inner code parity symbol is a  $\text{GF}(q)$ -weighted linear combination of the (at most)  $\xi$  inner code message symbols from the local group,  $\Gamma^{(i)}$  contains at most  $\xi$  stochastically independent symbols. Considering the repair of all storage nodes in  $\mathcal{S}_2$ , it follows by the argument above that at most  $\xi \ell_2$  stochastically independent inner code symbols are eavesdropped during the repair process. Also, since each storage node stores a single symbol, the eavesdropper has access to an additional  $\ell_1$  inner code symbols from the storage nodes in  $\mathcal{S}_1$ . Hence, in total, the eavesdropper has access to at most  $\ell_1 + \xi \ell_2$  stochastically independent symbols from  $\mathbf{c}$ . Thus,  $H(\mathbf{e}) \leq (\ell_1 + \xi \ell_2)p \log q$ . Furthermore, since  $\mathbf{r}$  contains  $u = \ell_1 + \xi \ell_2$  uniform independent random symbols,  $H(\mathbf{r}) = (\ell_1 + \xi \ell_2)p \log q$ , and the result follows. ■

**Theorem 2.** The code comprising of a Gabidulin code as its outer code and an RFC as its inner code, which encodes a vector  $\tilde{\mathbf{m}} = (\mathbf{m}, \mathbf{r})$  that consists of a message  $\mathbf{m}$  of length  $k$  and a random vector  $\mathbf{r}$  of length  $u = \ell_1 + \xi \ell_2$  is completely secure for the  $(\ell_1, \ell_2)$  eavesdropper model.

*Proof:* To prove security, we show that  $H(\mathbf{r}|\mathbf{e}, \mathbf{m}) = H(\mathbf{r}) - H(\mathbf{e})$ , which is equivalent to  $I(\mathbf{m}; \mathbf{e}) = 0$  according to Theorem 1. Each eavesdropped symbol  $e_i$ ,  $i = 1, \dots, w$ , where  $w = |\mathbf{e}|$ , corresponds to a code symbol and therefore is an evaluation of  $f(\cdot)$  at some point  $z_i \in \{y_1, \dots, y_n\} \subset \text{GF}(q^p)$ ,

where  $c_i = f(y_i)$  (see Remark 1). Thus, for  $i = 1, \dots, w$ ,

$$e_i = f(z_i) = \sum_{j=1}^{\tilde{k}} \tilde{m}_j z_i^{q^{j-1}} = \sum_{j=1}^k m_j z_i^{q^{j-1}} + \sum_{j=1}^u r_j z_i^{q^{k+j-1}} \quad (2)$$

since  $\tilde{\mathbf{m}} = (\mathbf{m}, \mathbf{r})$ . In the following,  $\mathbf{r} = (r_1, \dots, r_u)$  is assumed to be the unknowns (the message  $\mathbf{m}$  and the eavesdropper vector  $\mathbf{e}$  are assumed to be known) in the linear system of equations defined in (2).

Let  $1 \leq \nu \leq w$  (by definition) be the number of GF( $q$ )-linear independent symbols of  $\{e_1, \dots, e_w\}$ , denoted by  $\tilde{\mathbf{e}} = (\tilde{e}_1, \dots, \tilde{e}_\nu)$ . The corresponding vector of points from  $\{z_1, \dots, z_w\}$  is denoted by  $\tilde{\mathbf{z}} = (\tilde{z}_1, \dots, \tilde{z}_\nu)$ . From (2),  $\tilde{\mathbf{e}} = \mathbf{b}(\tilde{\mathbf{z}}, \mathbf{m}) + \mathbf{r} \cdot \mathbf{A}(\tilde{\mathbf{z}})$ , where  $\mathbf{b}(\tilde{\mathbf{z}}, \mathbf{m})$  is a length- $\nu$  row vector and  $\mathbf{A}(\tilde{\mathbf{z}})$  is a  $u \times \nu$  matrix. Since  $\{\tilde{e}_1, \dots, \tilde{e}_\nu\}$  are GF( $q$ )-linear independent, the matrix  $\mathbf{A}(\tilde{\mathbf{z}})$  is of full column-rank (i.e., its column space is a vector space over GF( $q$ ) of dimension  $\nu$ ), and since the  $u$  random symbols in  $\mathbf{r}$  are chosen independently and uniformly at random from GF( $p^q$ ),  $\{\tilde{e}_1, \dots, \tilde{e}_\nu\}$  are also *stochastically* independent uniformly distributed random variables over GF( $q^p$ ) ( $\tilde{e}_i$  is uniformly distributed over GF( $q^p$ ) for all  $i$  and  $\tilde{\mathbf{e}}$  is uniformly distributed over GF( $q^p$ ) $^\nu$ ). Finally, since  $e \in \{e_1, \dots, e_w\} \setminus \{\tilde{e}_1, \dots, \tilde{e}_\nu\}$ , it follows that  $H(\mathbf{e}) = H(\tilde{\mathbf{e}}) = \nu \cdot p \log q$ . From Lemma 1,  $H(\mathbf{e}) \leq H(\mathbf{r})$ . Thus,  $u \geq \nu$  since  $H(\mathbf{e}) = \nu \cdot p \log q$  and  $H(\mathbf{r}) = u \cdot p \log q$ . The conditional entropy  $H(\mathbf{r}|\mathbf{e}, \mathbf{m})$  is equal to the logarithm (base-2) of the number of solutions of (2) when the number of unknowns  $u$  is larger than or equal to the number of independent equations  $\nu$ , i.e., when  $u \geq \nu$ . Hence,  $H(\mathbf{r}|\mathbf{m}, \mathbf{e}) = (u - \nu)p \log q = H(\mathbf{r}) - H(\mathbf{e})$  from which the result follows from (1) (see Theorem 1). ■

**Example 4.** Consider the (20, 10) secure RFC over GF( $q^p$ ) in Fig. 2 that encodes the message  $\mathbf{m} = (m_1, \dots, m_6)$  of 6 symbols and a vector  $\mathbf{r} = (r_1, \dots, r_4)$  of 4 random symbols. Each  $c_i, i = 1, \dots, 20$ , is an evaluation of a linearized polynomial  $f(\cdot)$  at  $y_i$ . For the (1, 1) eavesdropper model, the scenario where  $\mathcal{S}_1 = \{c_6\}$  and  $\mathcal{S}_2 = \{c_5\}$ , i.e., the eavesdropper gains access to the symbols  $\mathbf{e} = (c_5, c_6, c_8, c_{18} = c_5 + c_6 + c_8)$ , is depicted. It can easily be seen that  $H(\mathbf{r}) = 4p \log q$ ,  $H(\mathbf{e}) = 3p \log q$ , and  $H(\mathbf{r}|\mathbf{m}, \mathbf{e}) = (4 - 3)p \log q$ . Therefore, there is no information leakage to the eavesdropper.

## V. NUMERICAL RESULTS

We compare the proposed secure RFCs with the secure MSR codes and secure LRCs in [2] in terms of the maximum code rate  $k/n$  that allows to achieve security. In particular, we consider  $(r, \delta)$   $d_{\min}$ -optimal LRCs [2], where  $r$  is the code locality (and thus has an analogous meaning to  $\xi$  for RFCs) and  $d_{\min}$  is the minimum distance of the code. Each local group of such a code can be seen as a subcode (punctured from the LRC) of minimum distance at least  $\delta$ .

In Fig. 3, we fix the code rate of the inner code (RFC, LRC, or MSR code),  $\tilde{k}/n$ , to 0.5 and 0.8, and plot the achievable code rates (the maximum  $k/n$  that allows to achieve security) for the (2, 2) eavesdropper model as a function of  $\tilde{k}$ . Note that  $\tilde{k}/n$  is an upper bound on the achievable code rate  $k/n$ , since

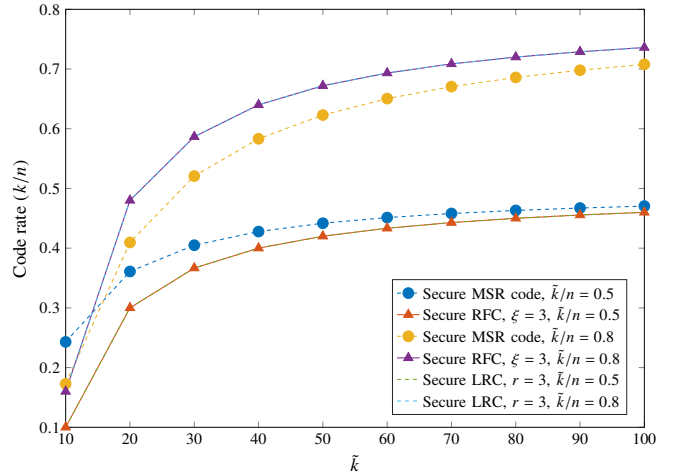


Fig. 3: Comparison of code rates for different classes of secure ECCs for the (2, 2) eavesdropper model.

to achieve security a number of random symbols needs to be appended to the message of length  $k$ . Note also that  $n$  is the total number of storage nodes. We remark that, unlike LRC- and RFC-based DSSs, where each code symbol is stored in a different storage node, for MSR codes each storage node stores  $\alpha = (n - \tilde{k})^{\tilde{k}-1}$  code symbols. For a fair comparison between RFCs and LRCs, we set  $r = \xi$ . It can be seen that the achievable code rates for secure RFCs and secure LRCs are identical. On the other hand, secure RFCs yield higher achievable code rates compared to secure MSR codes for  $\tilde{k}/n = 0.8$ , while the opposite is observed for  $\tilde{k}/n = 0.5$ .

## VI. CONCLUSION

We proposed a code construction based on RFCs that is secure against the  $(\ell_1, \ell_2)$  eavesdropper model. We gave a necessary and sufficient condition for the information leakage to the eavesdropper to be zero, and subsequently proved that the proposed construction is completely secure. The proposed secure RFCs yield the same achievable code rates as LRCs, and higher than MSR codes (when the code rate of the underlying code is high enough). An interesting extension of this work would be the design of secure and repair-efficient vector RFCs, i.e., code symbols are distributed over the storage nodes, each containing  $\alpha > 1$  code symbols.

## REFERENCES

- [1] N. B. Shah, K. V. Rashmi, and P. V. Kumar, "Information-theoretically secure regenerating codes for distributed storage," in *Proc. IEEE Global Telecommun. Conf. (GLOBECOM)*, Houston, TX, Dec. 2011.
- [2] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath, "Optimal locally repairable and secure codes for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 212–236, Jan. 2014.
- [3] E. M. Gabidulin, "Theory of codes with maximum rank distance," *Problems Inf. Transmiss.*, vol. 21, pp. 1–12, Jul. 1985.
- [4] D. S. Papailiopoulos and A. G. Dimakis, "Locally repairable codes," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Cambridge, MA, Jul. 2012.
- [5] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4539–4551, Sep. 2010.
- [6] M. Asteris and A. G. Dimakis, "Repairable fountain codes," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 5, pp. 1037–1047, May 2014.