Optimal Fractional Repetition Codes

Natalia Silberstein and Tuvi Etzion

January 21, 2014

Abstract

Fractional repetition (FR) codes is a family of codes for distributed storage systems that allow for uncoded repair having the minimum repair bandwidth. However, in contrast to minimum bandwidth regenerating codes, where a random set of certain size of available nodes is used for a node repair, the repairs with FR codes are table based.

In this work we consider bounds on the fractional repetition capacity. Optimal FR codes which attain these bounds are presented. The constructions of optimal FR codes are based on combinatorial designs and on different families of regular and biregular graphs. Finding optimal codes raises some interesting questions in graph theory. We discuss these questions and their solutions.

In addition, we analyze other properties of the constructed codes, allowing parallel independent reads of many subsets of the stored symbols, by showing a connection to combinatorial batch codes. We also define for each code a rate hierarchy which resembles to the well known generalized Hamming weight hierarchy.

1 Introduction

In distributed storage systems (DSS), data is stored across a network of nodes, which can unexpectedly fail. To provide a reliability, data redundancy based on coding techniques is introduced in such systems. Moreover, existing erasure codes allow to minimize the storage overhead. In [10] Dimakis et al. introduce a new family of erasure codes, called regenerating codes, which allow for efficient single node repairs by minimizing repair bandwidth. In particular, they present two families of regenerating codes, called minimum storage regenerating (MSR) and minimum bandwidth regenerating (MBR) codes, which correspond to the two extreme points on the storage-bandwidth trade-off [10]. The constructions for these two families of codes can be found in [10,11,23,24,27,29,30] and references therein.

An \((n, k, d, \alpha, \beta)_q\) regenerating code \(C\), for \(k \leq d \leq n - 1, \beta \leq \alpha\), is used to store a file across a network of \(n\) nodes, where each node stores \(\alpha\) symbols from \(\mathbb{F}_q\), a finite field with \(q\) elements, such that the stored file can be recovered by downloading the data from any set of \(k\) nodes. Note, that this means that any \(n - k\) node failures (i.e., erasures) can be corrected by this code. When a single node fails, a newcomer node which substitutes the failed node contacts with any set of \(d\) nodes and downloads \(\beta\) symbols of each node in this set to reconstruct the failed data. This process is called a node repair process.

In [23] Rashmi et al. presented a construction for MBR codes which have the additional property of exact repair by transfer, or uncoded repair. In other words, the code proposed in [23] allows for efficient node repairs where no decoding is needed. Every node participating in a node repair process just passes one symbol (\(\beta = 1\)) which will be directly stored in the newcomer node. This construction is based on a concatenation of an MDS code with a repetition code based on a complete graph as follows. Let \(k\alpha - \binom{k}{2}\) be the size of a file over \(\mathbb{F}_q\). This file is first encoded by using an \(\left(\binom{n}{2}, k\alpha - \binom{k}{2}\right)\) MDS code \(C\). The symbols of the corresponding codeword of \(C\) are placed on \(n\) different nodes of size \(\alpha = n - 1\) in a following way. Every node is associated with a vertex in \(K_n\), the complete graph with \(n\) vertices. Every symbol of the codeword of \(C\) is associated with an edge in \(K_n\). Every node \(i\) of the DSS stores the symbols of the codeword which are associated with the edges incident to vertex \(i\) in \(K_n\). The authors in [23] proved the uniqueness of this construction for the given parameters \(\alpha = d = n - 1\).
Rouayheb and Ramchandran [26] generalized the construction of [23] and defined a new family of codes for DSS, called fraction repetition (FR) codes. This family of codes was proposed for efficient uncoded repairs for a wide range of parameters. However, the definition of FR codes relaxes the requirement of a random $d$-set for a repair of a failed node and instead of this the repairs become table based. This modified model requires a modification for the bounds on the maximum amount of data that can be stored on a DSS based on a FR code.

An $(n, k, \alpha, \rho)$ fractional repetition code $C$ with repetition degree $\rho$ is a collection of $n$ subsets $N_1, \ldots, N_n$ of $[\theta] = \{1, 2, \ldots, \theta\}$, each one of size $\alpha$. Node $i$ of the DSS stores $\alpha$ symbols from a codeword of an MDS code.

These symbols are located in the positions of the codeword indexed by the elements of the subset $N_i$. Each element of $[\theta]$ belongs to exactly $\rho$ sets of $C$ (this implies that $n\alpha = \rho \theta$). More precisely, let $f \in \mathbb{F}_q^M$ be a file of size $M$. First, $f$ is encoded by using a $(\theta, M)$ MDS code $C$, where $\theta = n\alpha / \rho$. Second, the $\theta$ symbols of a codeword $c_f \in C$ which encodes the file $f$ are placed on $n$ nodes as explained above. The FR code should satisfy the requirement that from any set of $k$ nodes it is possible to reconstruct the stored file $f$. When some node $N_i$ fails, it can be repaired by using a set of $\alpha$ other nodes $\{N_{i_1}, N_{i_2}, \ldots, N_{i_{\alpha}}\}$, such that $|N_j \cap N_{i_t}| > 0$ and $N_j \cap N_{i_s} \neq N_j \cap N_{i_t}$, $s, t \in [\alpha]$.

The encoding scheme based on a FR code is shown in Fig. 1.

For a FR code we have that $\alpha = d$ and $\beta = 1$. The repair bandwidth of a FR code is the same as the repair bandwidth of an MBR code. Constructions of FR codes based on regular graphs and different types of combinatorial designs like Steiner systems, affine resolvable designs and Latin squares were proposed in [4, 17, 19, 21, 31].

The rate $R_C(k)$ of an $(n, k, \alpha, \rho)$ FR code $C = \{N_1, \ldots, N_n\}$ is defined by

$$R_C(k) = \min_{|I| = k} |\bigcup_{i \in I} N_i|.$$  

An $(n, k, \alpha, \rho)$ FR code is called universally good if its rate is no less than the capacity of MBR codes, given by $k\alpha - \binom{k}{2}$, for any $k \leq \alpha$. In particular, it is of interest to consider codes whose rate exceeds the capacity of MBR codes. Two upper bounds on the maximum rate of an $(n, k, \alpha, \rho)$ FR code, called FR capacity and denoted by $C_{FR}(n, k, \alpha, \rho)$, were presented in [26]:

$$C_{FR}(n, k, \alpha, \rho) \leq \left\lfloor \frac{n\alpha}{\rho} \left( 1 - \frac{\binom{n-\rho}{k}}{\binom{n}{k}} \right) \right\rfloor; \quad (1)$$

Figure 1: The encoding scheme based on a FR code
\[ C_{FR}(n, k, \alpha, \rho) \leq \varphi(k), \]

where

\[ \varphi(1) = \alpha, \quad \varphi(k + 1) = \varphi(k) + \alpha - \left\lceil \frac{\rho \varphi(k) - k \alpha}{n - k} \right\rceil. \]  \hspace{1cm} (2)

Note, that bound (2) is tighter than bound (1).

In this paper, we address the problem of constructing of optimal FR codes. We present a family of codes whose rate attains the upper bound in (2). In addition, we consider some parameters where this bound can be improved and present explicit constructions for FR codes that have the maximum rate for these parameters.

First, we propose constructions for FR codes with \( \rho = 2 \) which attain the bound in (2). Note, that the case \( \rho = 2 \) corresponds to the case of the highest data/storage ratio, since the repetition degree is the lowest one. All these constructions are based on different families of regular graphs. One construction is based on Turán graphs. A special case of a Turán graph is a complete regular bipartite graph. Another construction is based on different regular graphs with a given girth and in particular, cage graphs. Next, we consider FR codes with \( \rho > 2 \). Note that in contrast to the case with \( \rho = 2 \), here a failed node can be repaired from several sets of other nodes. One construction is based on a family of combinatorial designs, called transversal designs. This construction generalizes the construction based on complete regular bipartite graphs for \( \rho = 2 \). Another construction is based on biregular bipartite graphs with a given girth. One important family of such graphs are the generalized polygons. We analyze the parameters of the constructed codes and find the conditions for which the bound in (2) is attained.

We analyze additional properties of the constructed FR codes which allow parallel independent reads of the stored data by showing a connection of these codes to combinatorial batch codes. We consider a scenario when it is possible to reconstruct a \( t \)-subset of data symbols from the stored file, by reading at most one element from each node. In other words, we want to provide the load balancing in partial data reconstruction, which can be performed by several users independently and in parallel.

The rest of the paper is organized as follows. In Section 2 we provide the main definitions of the structures which will be used in our constructions. In particular, we provide definitions for some families of regular and biregular graphs, graphs with a given girth, transversal designs, projective planes, generalized polygons and their incidence matrices. In Sections 3 and 4 we consider FR codes with \( \rho = 2 \) and \( \rho > 2 \), respectively. In Section 5 we establish a connection between the rates of FR codes and generalized Hamming weights. In Section 6 we analyze the constructed FR codes as combinatorial batch codes. Conclusions and some problems for future research are given in Section 7.

2 Preliminaries

In this section we provide the definitions of all the combinatorial objects used for the constructions of FR codes presented in this paper.

2.1 Regular and Biregular Graphs

A graph \( G = (V, E) \) consists of a vertex set \( V \) and an edge set \( E \), where an edge is an unordered pair of vertices of \( V \). For an edge \( e = \{x, y\} \in E \) we say that \( x \) and \( y \) are adjacent and that \( x \) and \( e \) are incident. The degree of a vertex \( x \) is the number of edges incident with it. We say that a graph \( G \) is regular if all its vertices have the same degree and \( G \) is \( d \)-regular if each vertex has degree \( d \). A graph is called connected if there is path between any pair of vertices. A graph is called complete if every pair of vertices are adjacent. A complete graph on \( n \) vertices is denoted by \( K_n \). A subgraph \( G_2 = (V_2, E_2) \) of a graph \( G_1 = (V_1, E_1) \) is a graph such that \( V_1 \subseteq V_2 \) and \( E_1 \subseteq E_2 \). A \( k \)-clique in a graph \( G \) is a subgraph with \( k \) vertices of \( G \) which is complete.

The incidence matrix \( I(G) \) of a graph \( G = (V, E) \) is a binary \( |V| \times |E| \) matrix with rows and columns indexed by the vertices and edges of \( G \), respectively, such that \((I(G))_{i,j} = 1\) if and only if vertex \( i \) and edge \( j \) are incident.
A graph $G$ is called bipartite ($r$-partite) if its vertex set can be partitioned into two ($r$) parts such that every two adjacent vertices belong to two different parts. The complete bipartite graph with left part of size $n$ and right part of size $m$ is denoted by $K_{n,m}$. A bipartite graph $G$ is called biregular if the degree of the vertices in one part is $d_1$ and the degree of the vertices in the other part is $d_2$. Note that in $K_{n,m}$ the degree of a vertex in the left part is $m$ and the degree of a vertex in the right part is $n$.

The following theorem, known as Turán theorem, shows the conditions when a graph does not contain a clique of a given size [15].

**Theorem 1.** If a graph $G = (V, E)$ on $n$ vertices has no $(r + 1)$-clique, $r \geq 2$, then

$$|E| \leq (1 - \frac{1}{r}) \frac{n^2}{2}. \quad (3)$$

**Remark 1.** Note that if $G$ is an $\alpha$-regular graph then from Theorem [1] we have $n \geq \frac{r}{r-\alpha}$.

We consider a family of regular graphs, called Turán graphs, which attain the bound of Theorem [1], or equivalently, have the smallest number of vertices. Let $r, n$ be two integers such that $r$ divides $n$. An $(n, r)$-Turán graph is defined as a complete $r$-partite graph, i.e., a graph formed by partitioning a set of $n$ vertices into $r$ parts of size $\frac{n}{r}$ and connecting each two vertices of different parts by an edge. Clearly, an $(n, r)$-Turán graph does not contain a clique of size $r + 1$ and it is an $(r - 1)\frac{n}{r}$-regular graph.

A cycle in a graph $G$ is a connected subgraph of $G$ in which each vertex has degree two. The girth of a graph is the length of the shortest cycle in it. A $(d, g)$-cage is a $d$-regular graph with girth $g$ and minimum number of vertices. For example, a $(d, 4)$-cage is a complete bipartite graph $K_{d,d}$. Constructions for cages are known for $g \leq 12$ [12]. A lower bound on the number of vertices in a $(d, g)$-cage is given in the following theorem, known as Moore bound [15].

**Theorem 2.** The number of vertices in a $(d, g)$-cage is at least

$$N_0(d, g) = \begin{cases} 1 + d \sum_{i=0}^{\frac{g-3}{2}} (d - 1)^i & \text{if } g \text{ is odd} \\ 2 \sum_{i=0}^{\frac{g-2}{2}} (d - 1)^i & \text{if } g \text{ is even} \end{cases} \quad (4)$$

Similar result to the Moore bound for biregular bipartite graphs can be found in [5][14].

### 2.2 Combinatorial Designs

A set system is a pair $(\mathcal{P}, \mathcal{B})$, where $\mathcal{P} = \{p_i\}$ is a finite nonempty set of points and $\mathcal{B} = \{B_i\}$ is a finite nonempty set of subsets of $\mathcal{P}$ called blocks. A design $D$ is a set system with a constant number of points per block and no repeated blocks. A design $D$ can be described by an incidence matrix $I(D)$, which is a $|\mathcal{P}| \times |\mathcal{B}|$ matrix, with rows indexed by the points, columns indexed by the blocks, where

$$(I(D))_{i,j} = \begin{cases} 1 & \text{if } p_i \in B_j \\ 0 & \text{if } p_i \notin B_j \end{cases}$$

The incidence graph $G_I(D) = (V, E)$ of $D$ is the bipartite graph with the vertex set $V = \mathcal{P} \cup \mathcal{B}$, where $\{p, B\} \in E$ if and only if $p \in B$, for $p \in \mathcal{P}, B \in \mathcal{B}$.

A transversal design of group size $h$ and block size $\ell$, denoted by TD($\ell, h$) is a triple $(\mathcal{P}, G, \mathcal{B})$, where

1. $\mathcal{P}$ is a set of $\ell h$ points;
2. $G$ is a partition of $\mathcal{P}$ into $\ell$ sets (groups), each one of size $h$;
3. $\mathcal{B}$ is a collection of $\ell$-subsets of $\mathcal{P}$ (blocks);
4. each block meets each group in exactly one point;
5. any pair of points from different groups is contained in exactly one block.

The properties of a transversal design $TD(\ell, h)$ which will be useful for our constructions are summarized in the following lemma [3].

**Lemma 3.** Let $(\mathcal{P}, \mathcal{G}, \mathcal{B})$ be a transversal design $TD(\ell, h)$. Then

- The number of points is given by $|\mathcal{P}| = \ell h$;
- The number of groups is given by $|\mathcal{G}| = \ell$;
- The number of blocks is given by $|\mathcal{B}| = h^2$;
- The number of blocks that contain a given point is equal to $h$.
- The girth of the incidence graph of a transversal design is equal to 6.

A TD$(\ell, h)$ is called resolvable if the set $\mathcal{B}$ can be partitioned into sets $\mathcal{B}_1, \ldots, \mathcal{B}_h$, each one contains $h$ blocks, such that each element of $\mathcal{P}$ is contained in exactly one block of each $\mathcal{B}_i$, i.e., the blocks of $\mathcal{B}_i$ partition the set $\mathcal{P}$. Resolvable transversal design $TD(\ell, q)$ is known to exist for any $\ell \leq q$ and prime power $q$ [3].

**Remark 2.** A TD$(2, h)$, for any integer $h \geq 2$ is equivalent to the complete bipartite graph $K_{h,h}$.

Next, we consider two families of designs whose incidence graphs attain the Moore bound [4].

A projective plane of order $n$ denoted by PG$(2,n)$, is a set system $(\mathcal{P}, \mathcal{B})$, such that $|\mathcal{P}| = |\mathcal{B}| = n^2 + n + 1$, each block of $\mathcal{B}$ is of size $n + 1$, and any two points are contained in exactly one block. Note that any two blocks in $\mathcal{B}$ have exactly one common point. It is well known (see [15]) that the incidence graph of a projective plane has girth 6.

A generalized quadrangle of order $(s, t)$, denoted by GQ$(s, t)$ is a set system $(\mathcal{P}, \mathcal{B})$, where

- Each point $p \in \mathcal{P}$ is incident with $t + 1$ blocks, and each block $B \in \mathcal{B}$ is incident with $s + 1$ points.
- Any two blocks have at most one common point.
- For any pair $(p, B) \in \mathcal{P} \times \mathcal{B}$, such that $p \notin B$, there is exactly one block $B'$ incident with $p$, such that $|B' \cap B| = 1$.

In a generalized quadrangle GQ$(s, t)$, the number of points $|\mathcal{P}| = (s + 1)(st + 1)$, the number of blocks $|\mathcal{B}| = (t + 1)(st + 1)$ and the girth of the incidence graph is 8 [15].

We note that transversal designs, projective planes, and generalized quadrangles belong to a class of designs called partial geometries. In addition, projective planes and generalized quadrangles are examples of designs called generalized polygons (or $n$-gons). Their incidence graphs have girth $2n$ and they attain the Moore bound. Such structures are known to exist only for $n \in \{3, 4, 6, 8\}$ [15].

### 2.3 FR Codes based on Graphs and Designs

Let $C$ be an $(n, k, \alpha, \rho)$ FR code. $C$ can be described by an incidence matrix $I(C)$, which is an $n \times \theta$ binary matrix, $\theta = \frac{n\alpha}{\rho}$, with rows indexed by the nodes of the code and columns indexed by the symbols of the corresponding MDS codeword, such that $(I(C))_{i,j} = 1$ if and only if node $i$ contains symbol $j$.

Let $G$ be an $\alpha$-regular graph with $n$ vertices. We say that an $(n, k, \alpha, \rho = 2)$ FR code $C$ is based on $G$ if $I(C) = I(G)$. Such a code will be denoted by $C_G$.

Let $D = (\mathcal{P}, \mathcal{B})$ be a design with $|\mathcal{P}| = n$ points such that each block $B \in \mathcal{B}$ contains $\rho$ points and each point $p \in \mathcal{P}$ is contained in $\alpha$ blocks. We say that an $(n, k, \alpha, \rho)$ FR code $C$ is based on $D$ if $I(C) = I(D)$. Such a code will be denoted by $C_D$.  

5
3 Fractional Repetition Codes with $\rho = 2$

In this section we present constructions of FR codes with $\rho = 2$ and analyze their rate. These constructions are based on different types of regular graphs. Note that FR codes based on regular graphs were also considered in [17,26].

Let $G = (V, E)$, $|V| = n$, be an $\alpha$-regular graph. When an $(n, k, \alpha, \rho = 2)$ FR code $C_G$ based on $G$ is used to store a file of size $M$, the file is first encoded by using an $(\frac{n\alpha}{2}, M)$ MDS code $C$ and the symbols of the codeword of $C$ are stored on $n$ different nodes which correspond to the vertices of $G$. Node $i$, $1 \leq i \leq n$, stores $\alpha$ symbols indexed by the edges which are incident with vertex $i$. When node $j$ fails, it can be repaired by using $\alpha$ other nodes which correspond to the vertices of $G$ incident to all the edges (symbols) of the failed node $j$. To ensure the file recoverability from any set of $k$ nodes, the rate of the FR code $C_G$ should satisfy $R_{C_G}(k) \geq M$. In this section, we analyze the connection between the rate $R_{C_G}(k)$ of $C_G$ and the properties of the regular graph $G$. Based on graphs with these properties, we construct FR codes which attain the upper bound on the FR capacity.

First we consider FR codes based on the complete bipartite graphs. These codes are analog to the construction for MBR codes based on the complete graphs, presented by Rashmi et al. [23]. We prove that our codes attain the upper bound in (2) on the rate for all $k \leq \alpha$.

Theorem 4. The rate of the $(2\alpha, k, \alpha, 2)$ FR code $C_{K\alpha,\alpha}$, for $\alpha \geq 2$, is given by

$$R_{C_{K\alpha,\alpha}}(k) = \begin{cases} \frac{\alpha k - k^2}{\alpha k - \frac{k^2 - 1}{4}} & \text{if } k \text{ is even} \\ \frac{\alpha k - \frac{k^2 - 1}{4}}{\alpha k - \frac{k^2 - 1}{4}} & \text{if } k \text{ is odd} \end{cases}$$

(5)

which attains the upper bound in (2) for all $k \leq \alpha$.

Proof. For an even $k = 2t$, there are $t + i$ nodes which correspond to the vertices form one part of the graph and $t - i$ nodes that correspond to the second part, for some integer $i$. Hence, $R_{C_{K\alpha,\alpha}}(k) = \min_i \{k\alpha - (t + i)(t - i)\} = k\alpha - t^2 = k\alpha - \frac{k^2}{4}$. For an odd $k = 2t + 1$, there are $t + i$ nodes that correspond to the vertices form one part and $t - i$ from another one, hence $R_{C_{K\alpha,\alpha}}(k) = \min_i \{k\alpha - (t + 1 + i)(t - i)\} = k\alpha - (t + 1)t = k\alpha - \frac{k^2 - 1}{4}$.

To prove that the rate $R_{C_{K\alpha,\alpha}}(k)$ attains the upper bound in (2), we note the rate of $C_{K\alpha,\alpha}$ satisfies the following recursion

$$R_{C_{K\alpha,\alpha}}(k + 1) = R_{C_{K\alpha,\alpha}}(k) + \alpha - \left\lfloor \frac{k}{2} \right\rfloor.$$  

(6)

It is easy to prove (by induction) that for the parameters of the constructed FR code $C_{K\alpha,\alpha}$, the recursive formula in (6) is the same as the recursive formula in (6).

Remark 3. Note, that for any $k \geq 3$, the rate of the code $C_{K\alpha,\alpha}$ is strictly larger than the rate of an MBR code, i.e.,

$$R_{C_{K\alpha,\alpha}}(k) > k\alpha - \left(\frac{k}{2}\right).$$

Example 1. The $(6, k, 3, 2)$ FR code based on $K_{3,3}$ and its rate is shown in Fig. 2

Note that $K_{\alpha,\alpha}$ is a $(2\alpha, 2)$-Turán graph and $(\alpha, 4)$-cage graph. Now, we will consider the rate of FR codes based on Turán graphs and cage graphs. The following lemma follows directly from the definition of a clique.

Lemma 5. An $\alpha$-regular graph $G$ with $n$ vertices contains a clique of size $k$ if and only if $R_{C_G}(k) = k\alpha - \left(\frac{k}{2}\right)$.

Corollary 6. The rate $R_{C_G}(k)$ of a FR code $C_G$, where $G$ is a graph which does not contain a clique of size $k$, is strictly larger than the MBR capacity.

Since it is desirable to have FR codes with a rate which exceeds the MBR bound, we consider different families of regular graphs which do not contain a clique of a certain size. We start with the Turán graphs. The following theorem shows that FR codes obtained from Turán graphs are optimal for all $k \leq \alpha$. This is a generalization of Theorem 4.
Theorem 7. Let \( T \) be an \((n, r)\)-Turán graph for \( r \) which divides \( n \) and let \( k = br + t \) for \( b, t \geq 0 \) such that \( t \leq r - 1 \). The \((n, k, (r - 1)\frac{n}{r}, 2)\) FR code \( C_T \) based on \( T \) has rate

\[
R_{C_T}(k) = k\alpha - \left(\frac{k}{2}\right) + r\left(\frac{b}{2}\right) + bt.
\]  

which attains the upper bound in (2) for all \( k \leq \alpha \).

Proof. We consider a subgraph \( T' \subseteq T \) of \( k \) vertices with the minimum number of edges. The number of edges in such a graph is equal to the rate \( R_{C_T}(k) \). One can verify that since \( T \) is a complete \( r \)-partite graph, it follows that its minimum subgraph \( T' \) is a complete \( r \)-partite graph with exactly \( t \) parts of size \( b + 1 \) and \( r - t \) parts of size \( b \). Hence the number of edges in \( T' \) is given by

\[
\alpha k - \left[\left(\frac{t}{2}\right)(b + 1)^2 + \left(\frac{r - t}{2}\right)b^2 + t(r - t)(b + 1)b\right].
\]  

It is easy to verify that (8) equals to (7). One can check (by induction) that for the parameters of the constructed code \( C_T \) the bound in (2) equals to (7). \( \square \)

Next, we consider the girth of a graph and show that FR codes obtained from a graph with a large girth is optimal.

Lemma 8. The girth of an \( \alpha \)-regular graph \( G \) with \( n \) vertices is at least \( k + 1 \) if and only if \( R_{C_{G}}(k) = k\alpha - (k - 1) \).

Proof. Let \( G \) be a graph with girth \( g \).

Assume that \( g \geq k + 1 \). If \( R_{C_G}(k) < k\alpha - (k - 1) \) then there exists a set of \( k \) vertices of \( G \) which are incident to at most \( k\alpha - k \) edges. Therefore, there exists a cycle of length less than \( k + 1 \) in \( G \), and hence \( R_{C_G}(k) \geq k\alpha - (k - 1) \). \( R_{C_G}(k) \leq k\alpha - (k - 1) \) trivially follows from the fact that \( \rho > 1 \).

Assume that \( R_{C_G}(k) = k\alpha - (k - 1) \). If there exists a cycle of length \( \ell \leq k \) in \( G \) then any set of \( k \) vertices that contain this cycle are adjacent to at most \( k\alpha - k \) edges, which contradicts the given rate. \( \square \)

Corollary 9. A FR code \( C_G \) based on a graph \( G \) with girth \( g \) attains the bound in (2) for all \( k \leq g - 1 \).
Corollary 10. Let $TD$ be a $TD(q, q)$ transversal design. The $(2q^2, k, q, 2)$ FR code $C_{G_{TD}}$, based on the incidence graph $G_1(TD)$, attains the bound in (2) for all $k \leq 5$.

Corollary 11. Let $PG$ be a $PG(2, q)$ projective plane. The $(2q^2 + 2q + 2, k, q + 1, 2)$ FR code $C_{G_{PG}}$, based on the incidence graph $G_1(PG)$, attains the bound in (2) for all $k \leq 5$.

The incidence graph of a projective plane $PG(2, q)$ is a $(q + 1, 6)$-cage. It attains the Moore bound (see Theorem 2). The graphs that attain the Moore bound are called Moore graphs (for odd $g$) and generalized polygons (for even $g$). The known regular Moore graphs and generalized polygons and the parameters of FR codes corresponding to these graphs can be found in the following table.

| name of a graph            | degree | girth | parameters of a FR code               |
|----------------------------|--------|-------|---------------------------------------|
| Complete graph $K_n$       | $n - 1$| 3     | $(n, k, n - 1, 2)$                     |
| Complete bipartite graph $K_{r,r}$ | $r$   | 4     | $(2r, k, r, 2)$                       |
| Petersen graph             | 3      | 5     | $(10, k, 3, 2)$                       |
| Hoffman-Singleton graph    | 7      | 5     | $(50, k, 7, 2)$                       |
| Projective plane           | $q + 1$| 6     | $(2q^2 + 2q + 2, k, q + 1, 2)$        |
| Generalized quadrangle     | $q + 1$| 8     | $(2q^3 + 2q^2 + 2q + 2, k, q + 1, 2)$ |
| Generalized hexagon        | $q + 1$| 12    | $(2q^4 + 2q^3 + 2q^2 + 2q + 2, k, q + 1, 2)$ |

Now we use the Moore bound to show that the bound in (2) can be improved in many cases. Let $G$ be an $\alpha$-regular graph with girth $g$ and $n$ vertices. By (4) we have that $n \geq N_0(\alpha, g)$. By Lemma 8 and Corollary 9 we have that $R_{CG}(k) = \alpha k - (k - 1) = \varphi(k)$ for any $k \leq g - 1$. However, $R_{CG}(g) = \alpha g - g$, while $\varphi(g) = \alpha g - g + 2 - \left\lfloor\frac{\alpha(g-1)-2g+4}{n-g+1}\right\rfloor$. Note that $\varphi(g) = \alpha g - g + 1$ if and only if $n \geq \alpha g - \alpha - g + 3$. However, if $\varphi(g) = \alpha g - (g - 1)$ and this bound is tight, then the graph on $n$ vertices related to the code which attains this bound has no cycle of length $g$. Hence, its girth is at least $g + 1$ and therefore $n \geq N_0(\alpha, g + 1)$. Thus, the bound in (3) is not tight if $\max\{\alpha g - \alpha - g + 3, N_0(\alpha, g)\} \leq n \leq N_0(\alpha, g + 1)$.

Corollary 12. An FR code $C_G$ based on a graph $G$ with girth $g$ has optimal rate $R_{CG}(k)$ for all $k \leq g$.

We summarize the results on an FR code from a graph with a given girth in the following theorem.

Theorem 13. Let $G$ be a graph with girth $g$. Then the rate of a FR code $C_G$ based on $G$ satisfies

$$R_{CG}(k) = \begin{cases} k\alpha - k + 1 & \text{if } k \leq g - 1 \\ k\alpha - k & \text{if } g \leq k \leq g + \left\lceil \frac{g}{2} \right\rceil - 2. \end{cases}$$

Proof. For $k \leq g - 1$ the result directly follows from Lemma 8. Since the graph $G$ has a cycle of length $g$, then $R_{CG} = g\alpha - g$. It is easy to verify that in a graph $G$ with girth $g$, two cycles have at most $\left\lceil \frac{g}{2} \right\rceil + 1$ common vertices. Hence, the maximum number of vertices in a subgraph of $G$ with at most one cycle of length $g$ is $g + \left\lceil \frac{g}{2} \right\rceil - 2$. $lacksquare$

3.1 Rate of a FR Code with $\rho = 2$

First, we observe from Lemma 5 and Lemma 8 that for any $1 \leq k \leq \alpha$, the rate $R_C(k)$ of an $(n, k, \alpha, 2)$ FR code $C$ satisfies

$$k\alpha - \left(\frac{k}{2}\right) \leq R_C(k) \leq k\alpha - (k - 1), \quad (9)$$

and the value of the rate depends on the structure of the underlying regular graph $G$. If the graph contains a clique $K_k$ then the rate attains the lower bound in (9). If the graph does not contain a cycle of length $k$ then the rate attains the upper bound in (9). The intermediate values for the rate can be obtained by excluding certain subgraphs of $K_k$. 

8
from the graph $G$. For example, to have a rate equals to $k\alpha - \binom{k}{2} + 2$, the graph $G$ should not contain $K_k - e$, i.e., a $k$-clique without an edge.

For $k = 3$, there are only two possible values for $R_C(3)$, $3\alpha - 3$ and $3\alpha - 2$. To have a code $C$ with rate $3\alpha - 2$, one should exclude a clique $K_3$, which is also a cycle of length 3. From (4) it follows that for a given $\alpha$, if $n < 2\alpha$ then $R_C(3) = 3\alpha - 3$. Equivalently, the necessary condition for $R_C(3) = 3\alpha - 2$ is that $n \geq 2\alpha$.

Constructions for codes with rate $R_C(3) = 3\alpha - 2$ are provided in the previous subsection, based on optimal (Turán, Moore) graphs, for specific choices for the parameters $\alpha, n$, where $n$ is even. In addition, we provide another two constructions of FR codes with rate $3\alpha - 2$, the first one for even $n$, and the second one for odd $n$.

Let $I(K_{\alpha+i, \alpha+i})$, $i \geq 0$, be the $2(\alpha+i) \times (\alpha+i)^2$ incidence matrix of the complete bipartite graph $K_{\alpha+i, \alpha+i}$. Note that it is also the incidence matrix of a resolvable transversal design $TD(2, \alpha+i)$. Based on the resolvability of the design, $I(K_{\alpha+i, \alpha+i})$ can be written in a blocks form, i.e., $I(K_{\alpha+i, \alpha+i})$ is a $2 \times (\alpha+i) \times (\alpha+i)$ blocks matrix, where every block is a permutation matrix of size $(\alpha+i)^2$. Each such permutation matrix block will be called a $p$-block. Let $I_{\alpha,i}^{even}$ be a matrix obtained from $I(K_{\alpha+i, \alpha+i})$ by removing $i(\alpha+i)$ columns which correspond to $2i$ $p$-blocks. Note that there are exactly $\alpha$ ones in each row of $I_{\alpha,i}^{even}$. Let $C_{\alpha,i}^{even}$ be a FR code obtained from the graph $I_{\alpha,i}^{even}$, whose incidence matrix is $I_{\alpha,i}^{even}$. It is easy to verify that $C_{\alpha,i}^{even}$ is a $(2\alpha + 2i, k, \alpha, 2)$ code with rate $R_{C_{\alpha,i}^{even}}(3) = 3\alpha - 2$. Note that the data/storage ratio $\frac{R_{C_{\alpha,i}^{even}}(3)}{\alpha i} = \frac{3\alpha - 2}{(2\alpha + 2i)\alpha}$ decreases when $i$ increases.

For even $\alpha$ and odd $n = 3\alpha - 1$ we construct FR codes as follows. Let $I_{\alpha}^{odd}$ be the $(3\alpha - 1) \times 3\alpha^2 / 2$ matrix of the form

$$I_{\alpha}^{odd} = \begin{pmatrix} A & 0 \\ B & \frac{I_{\alpha,i}^{even}}{\alpha^2} - 1 \end{pmatrix},$$

where $0$ is the $(\alpha + 1) \times \frac{\alpha(\alpha-1)}{2}$ zero matrix, $I_{\alpha,i}^{even}$ is the $2\alpha - 2 \times \frac{\alpha(\alpha-2)}{2}$ matrix defined above, and $\begin{pmatrix} A \\ B \end{pmatrix}$, $A \in \mathbb{F}_2^{(\alpha+1)\times\alpha^2}$, $B \in \mathbb{F}_2^{(2\alpha-2)\times\alpha^2}$, is the $(3\alpha - 1) \times \alpha^2$ incidence matrix of the following graph. We take two copies of $K_{\alpha,\alpha-1}$, denoted by $K^1 = (L^1 \cup R^1, E^1)$ and $K^2 = (L^2 \cup R^2, E^2)$, and an additional vertex $u$ which is adjacent to all the vertices in the left part $L^i$ of both graphs $K^i$, $i = 1, 2$. The rows of matrix $A$ correspond to $u \cup L^1 \cup L^2$ and the rows of $B$ correspond to $R^1 \cup R^2$. Let $G_{\alpha,i}^{odd}$ be the graph with the incidence matrix $I_{\alpha,i}^{odd}$. One can verify that the FR code $C_{G_{\alpha,i}^{odd}}$ is a $(3\alpha - 1 - k, \alpha, 2)$ code with rate $3\alpha - 2$ for $k = 3$.

We illustrate this construction in the following example.

**Example 2.** For $\alpha = 4$ the incidence matrix $I_{\alpha}^{odd}$ is given by

$$I_{\alpha}^{odd} = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{pmatrix}.$$

where an empty entry in the matrix is 0.

For any odd $n \geq 5\alpha - 1$, let $G_{\alpha,i}^{odd}$, $i \geq 0$, be the graph whose incidence matrix $I_{\alpha,i}^{odd}$ has the form

$$I_{\alpha,i}^{odd} = \begin{pmatrix} I_{\alpha}^{odd} & 0 \\ 0 & I_{\alpha,i}^{even} \end{pmatrix}.$$

One can verify that the FR code $C_{G_{\alpha,i}^{odd}}$ is a $(5\alpha + 2i - 1, k, \alpha, 2)$ code with rate $3\alpha - 2$ for $k = 3$.

The following theorem follows immediately from the discussion above.
Theorem 14. The FR capacity for $k = 3$ satisfies

- For even $n$
  \[ C_{FR}(n,3,\alpha,2) = \begin{cases} 3\alpha - 3 & \text{if } n < 2\alpha \\ 3\alpha - 2 & \text{if } n \geq 2\alpha \end{cases} \]

- For odd $n$
  \[ C_{FR}(n,3,\alpha,2) = \begin{cases} 3\alpha - 3 & \text{if } n < 3\alpha - 1 \\ 3\alpha - 2 & \text{if } n \geq 3\alpha - 1 + 2(\alpha + i)j, i \geq 0, j \in \{0,1\} \end{cases} \]

For $k = 4$, we have that $R_C(4) \in \{4\alpha - 3, 4\alpha - 4, 4\alpha - 5, 4\alpha - 6\}$.

1. If $R_C(4) = 4\alpha - 3$ then the corresponding graph $G$ has girth at least 5. Codes with rate $4\alpha - 3$ can be given for any known graph $G$ with girth $\geq 5$, e.g., Hoffman-Singleton graph and its generalizations [1][18].

2. If $R_C(4) = 4\alpha - 4$ then the corresponding graph $G$ contains a subgraph of $K_4$ with 4 edges, but does not contain $K_4 - e$, a 4-clique without an edge. Codes with rate $4\alpha - 4$ and minimum number of nodes are constructed from $K_{\alpha,\alpha}$ by Theorem[4]

3. If $R_C(4) = 4\alpha - 5$ then the corresponding graph $G$ contains $K_4 - e$, but does not contain $K_4$. Codes with rate $4\alpha - 5$ and minimum number of nodes are constructed from $(n,3)$-Turán graphs.

4. If $R_C(4) = 4\alpha - 6$ then the corresponding graph $G$ contains $K_4$. Codes with rate $4\alpha - 6$ and minimum number of nodes are given by complete graphs.

The different rates for a general $k$ can be obtained by graphs with a given girth and by $(n, r)$-Turán graphs with different $r$’s (see Theorem[7] and Theorem[13]). However, constructions of FR codes with any given rate in the range between $k\alpha - \left(\frac{k}{2}\right)$ and $k\alpha - k + 1$ will be discussed in Section[7].

As we already saw, the problem of constructions for FR codes and estimation of their rates can be formulated in terms of graph theory. Generally, an answer for the following three problems provides solutions to these problems.

**Problem 1.** Find the value of $N(k, \alpha, \delta) = \frac{n}{\alpha} \geq 0$, which is the maximum number of vertices such that any $\alpha$-regular graph $G$ with less vertices contains a subgraph of $K_{\alpha,\alpha}$ with size $(\frac{k}{2}) - \delta$.

**Problem 2.** Find the value of $N'(k, \alpha, \delta) = \frac{n}{\alpha} \geq 0$, which is the maximum number of vertices such that for any $n \geq N'(k, \alpha, \delta)$, there exists an $\alpha$-regular graph $G$ with $n$ vertices which does not contain a subgraph of $K_{\alpha,\alpha}$ with size $(\frac{k}{2}) - \delta$.

**Problem 3.** Let $n, k, \alpha, \delta$ be positive integers such that $3 \leq k \leq \alpha$ and $0 \leq \delta \leq \left(\frac{k}{2}\right) - k$. Do there exist an $\alpha$-regular graph $G$ with $n$ vertices which does not contain a subgraph of $K_{\alpha,\alpha}$ whose size is $(\frac{k}{2}) - \delta$?

Clearly, an answer to Problem 1 provides a solution to the existence question of FR codes with any rate for $\rho = 2$. Based on the solution for Problem 1 and Problem 2, one can show that the FR capacity satisfies

\[ C_{FR}(n,k,\alpha,2) \leq k\alpha - \left(\frac{k}{2}\right) + \delta \quad \text{if } n < N(k,\alpha,\delta) \]
\[ C_{FR}(n,k,\alpha,2) \geq k\alpha - \left(\frac{k}{2}\right) + \delta + 1 \quad \text{if } n \geq N'(k,\alpha,\delta). \]

By our discussion above we have

**Corollary 15.** $N(3, \alpha, 0) = 2\alpha; N'(3, \alpha, 0) = 2\alpha$ if $\alpha$ is odd, and $3\alpha - 1 \leq N'(3, \alpha, 0) \leq 5\alpha - 2$ if $\alpha$ is even.

From the Moore bound and from the Turán bound we have the following corollary.

**Corollary 16.**

- $N(k, \alpha, \left(\frac{k}{2}\right) - k) \geq N_0(\alpha, k + 1)$, where $N_0(\alpha, k)$ is defined in Theorem[2]

- $N(k, \alpha, 0) \geq \left\lceil \frac{k-1}{k-2}\alpha \right\rceil$. 

10
4 Fractional Repetition Codes with $\rho > 2$

In this section, we consider FR codes with a general $\rho > 2$. Note, that while codes with $\rho = 2$ have the maximum data/storage ratio, codes with $\rho > 2$ provide multiple choices for node repairs. In other words, when a node fails, it can be repaired from different sets of available nodes.

We present generalizations of the constructions from the previous section, which were based on Turán graphs and graphs with a given girth. This generalizations employ transversal designs and generalized polygons, respectively.

We start with a construction of FR codes from transversal designs. Let $TD$ be a $TD(\rho, \alpha)$, $\rho \leq \alpha + 1$, be a transversal design with block size $\rho$ and group size $\alpha$. By Lemma 3, there are $\rho \alpha$ points in $TD$. Let $n \overset{\text{def}}{=} \rho \alpha$ and let $C_{TD}$ be an $(n, k, \alpha, \rho)$ FR code such that every node $i$, $1 \leq i \leq n$, corresponds to point $i \in [n]$ of $TD$, and all the symbols stored in node $i$ correspond to the set $N_i$ of blocks from $TD$ that contain the point $i$. Note that by Lemma 3, there are $\alpha$ blocks that contain a given point, hence each node stores $\alpha$ symbols.

Similarly to Theorem 7 we can prove the following theorem.

**Theorem 17.** Let $k = b\rho + t$, for $b, t \geq 0$ such that $t \leq \rho - 1$. For an $(n = \rho \alpha, k, \alpha, \rho)$ FR code $C_{TD}$ based on $TD(\rho, \alpha)$ we have

$$R_{C_{TD}}(k) \geq k\alpha - \binom{k}{2} + \rho \binom{b}{2} + bt.$$  

**Remark 4.** Note that for all $k \geq \rho + 1$, the rate of the FR code $C_{TD}$ is strictly larger than the MBR bound.

**Corollary 18.** Let $C_{TD}$ be an $(r\alpha, k, \alpha, r)$ FR code based on $TD(r, \alpha)$ and $C_T$ be an $(r - 1, \alpha, k, \alpha, 2)$ FR code based on $(n, r)$-Turán graph. Then

1. $C_{TD} = C_T$ for $r = 2$;
2. $R_{C_{TD}}(k) \geq R_{C_T}(k)$ for all $r \geq 2$.

**Example 3.** Let $TD$ be a transversal design $TD(3, 4)$ defined as follows: $\mathcal{P} = \{1, 2, \ldots, 12\}; \mathcal{G} = \{G_1, G_2, G_3\}$, where $G_1 = \{1, 2, 3, 4\}$, $G_2 = \{5, 6, 7, 8\}$, and $G_3 = \{9, 10, 11, 12\}; \mathcal{B} = \{B_1, B_2, \ldots, B_{16}\}$, with incidence matrix given by

$$I(TD) = \begin{pmatrix}
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1
\end{pmatrix}$$

The placement of symbols from a codeword of the corresponding MDS code of length 16 is shown in Fig.3. The rate values for different $k$’s are given in the following table.

| $k$ | $R_{C_{TD}}(k)$ |
|-----|-----------------|
| 1   | 4               |
| 2   | 7               |
| 3   | 9               |
| 4   | 11              |
Remark 5. The incidence matrix of a resolvable transversal design can be always written in a block form, were each block is a permutation matrix. In particular, by applying column permutations on the incidence matrix from Example 3 one can obtain a block matrix.

In the following theorem we find the conditions on the parameters such that the rate of a FR code $C_{TD}$ attains the recursive bound in (2).

Theorem 19. Let $k = b\rho + t \leq \alpha$, $0 \leq t \leq \rho - 1$, and $\alpha > \max\{\alpha_0(k'), \alpha_0(k)\}$, where $k' = (b - 1)\rho + \rho - 1$ and

\[
\alpha_0(k) = \begin{cases} 
\frac{1}{2}b^2\rho^2(b\rho - 1)(\rho - 2) + b(\rho(\rho - 1) + (4 - 3t) + (t - 1) + \frac{\rho - 2}{2}(t - 1)(t - 2)}{\rho + 1 - t} & \text{if } \rho \nmid k \\
\frac{b^2\rho^2}{2}(\rho - 1)(\rho - 2) + b(\rho^2 - 3\rho + 1) - 1 & \text{if } \rho | k
\end{cases}
\]

The rate $R_C$ of the $(\rho\alpha, k, \alpha, \rho)$ FR code $C_{TD}$ is given by

\[
R_C(k) = k\alpha - \binom{k}{2} + \rho \binom{b}{2} + bt
\]

and attains the bound in (2) for all $k \leq \alpha$.

Proof. We observe that if the lower bound on the rate of Theorem 17 is equal to the upper bound in (2) then we have that

\[
R_C(k) = k\alpha - \binom{k}{2} + \rho \binom{b}{2} + bt.
\]

Therefore, to satisfy this condition it follows from the recursion in (2) that

- If $\rho \nmid k$ then
  \[
k - b - 2 < \frac{(\rho - 1)(k - 1)\alpha - \rho \binom{k-1}{2} + \rho^2 \binom{b}{2} + \rho(t - 1)b}{\rho\alpha - k + 1}
  \]

- If $\rho | k$ then
  \[
k - b - 1 < \frac{(\rho - 1)(k - 1)\alpha - \rho \binom{k-1}{2} + \rho^2 \binom{b-1}{2} + \rho(\rho - 1)(b - 1)}{\rho\alpha - k + 1}
  \]
Hence, we obtain that

\[
\alpha > \begin{cases} 
\frac{b^2 \rho (\rho -1)(\rho -2) + b(\rho^2(t-1)+\rho(4-3t)+t-1)+\frac{2}{\rho^2} (t-1)(t-2)}{\rho+1-t} & \text{if } \rho \nmid k \\
\frac{b^2 \rho (\rho -1)(\rho -2) + b(\rho^2 - 3\rho + 1) - 1}{\rho} & \text{if } \rho | k
\end{cases}
\]  

(10)

In addition we note the function \(\alpha_0(k)\) increases in the interval \(i\rho \leq k < i\rho + \rho - 1\), however it might hold that \(\alpha_0(k_1) > \alpha_0(k_2)\) for \(k_1 < k_2\) such that \(k_1 = i\rho + \rho - 1\) and \(k_2 = (i + 1)\rho + j, j < \rho - 1\). Thus, we have \(\alpha > \max\{\alpha_0(k), \alpha_0(k')\}\) for \(k' = k - (t + 1)\).

**Example 4.** We illustrate the minimum values of \(\alpha\) for which the FR code obtained from a TD(\(\rho, \alpha\)) is optimal as a consequence of Theorem 19.

| \(k\) | \(\rho\) | 2 | 3 | 4 | 5 | 6 |
|------|------|---|---|---|---|---|
| 3    | 1    | 1 | 1 | 1 | 1 |
| 4    | 1    | 2 | 6 | 4 | 4 |
| 5    | 2    | 3 | 6 | 18 | 12 |
| 6    | 2    | 9 | 7 | 18 | 40 |
| 7    | 3    | 9 | 14 | 18 | 40 |
| 8    | 3    | 10 | 37 | 20 | 40 |

We continue similarly to the case with \(\rho = 2\), to find the conditions when there exists a FR code \(C\) with rate \(R_C(3) = 3\alpha - 2\). To have a rate greater than \(3\alpha - 3\), we should avoid the existence of a \(3 \times 3\) submatrix \(I'\) of \(I(C)\) such that each row of \(I'\) has exactly two ones. Such a matrix \(I'\) will be called a triangle.

**Lemma 20.** If \(n < \rho(\rho -1)\alpha - \rho(\rho - 2)\) then there exists a triangle in the incidence matrix of a FR code \(C\). Equivalently, the necessary condition for \(R_C(3) = 3\alpha - 2\) is that \(n \geq \rho(\rho - 1)\alpha - \rho(\rho - 2)\).

**Proof.** Given \(\theta = \frac{\alpha}{\rho}\), we need to prove that if \(\theta < (\rho - 1)\alpha - (\rho - 2)\alpha\) then there exists a triangle. We consider \(1 + (\rho - 1)\alpha\) rows which have common ones with a given row of the \(n \times \theta\) incidence matrix \(I(C)\) of an FR code. To avoid a triangle in the matrix, we must have that \(\theta \geq \alpha + (\alpha - 1)(\rho - 1)\alpha = (\rho - 1)\alpha^2 - (\rho - 2)\alpha\).

By Lemma 20 it follows that for \(n < \rho(\rho -1)\alpha - \rho(\rho - 2)\) and \(k = 3\) the rate of a FR code \(C\) equals \(R_C(3) = 3\alpha - 3\). However, the bound in (2) satisfies \(\varphi(3) = 3\alpha - 1 - \left\lfloor \frac{2(\rho - 1)\alpha - \rho}{\rho - 2} \right\rfloor = 3\alpha - 2\) if and only if \(n \geq 2(\rho - 1)\alpha - (\rho - 2)\). Thus, this bound is not tight in the interval \(n \in [2(\rho - 1)\alpha - (\rho - 2), \rho(\rho - 1)\alpha - \rho(\rho - 2)]\).

Next, we present a construction of a FR code \(C\) based on a generalized quadrangle, for which \(R_C(3) = 3\alpha - 2\). This code attains the bound on \(n\) presented in Lemma 20. The following lemma follows directly from the definition of a generalized quadrangle.

**Lemma 21.** Let \(GQ\) be a generalized quadrangle \(GQ(s, t)\), where \(t \geq s\), and let \(C_{GQ}\) be the FR code based on \(GQ\). \(C_{GQ}\) is an \((n = (s + 1)(st + 1), k, \alpha = t + 1, \rho = s + 1)\) FR code for which \(R_{C_{GQ}}(3) = 3\alpha - 2\) and \(R_{C_{GQ}}(4) = 4\alpha - 4\). Moreover, this code attains the bound on \(n\) of Lemma 20.

**Remark 6.** Similarly to a FR code \(C_G\) with \(\rho = 2\) based on a graph \(G\) with girth \(g\), we can consider a FR code \(C_{GP}\) based on a generalized \(g\)-gon (generalized polygon \(GP\)) for \(\rho > 2\). One can prove that the rate of \(C_{GP}\) is identical to the rate of \(C_G\) for \(k \leq g + \lfloor \frac{g}{2} \rfloor - 2\) given in Theorem 13. However, a generalized \(g\)-gon is known to exist only for \(g \in \{3, 4, 6, 8\}\). This observation also holds for the biregular bipartite graph of girth 2\(g\). The existence of such graphs was considered in [12][13][14].

**Remark 7.** Note that both generalized quadrangles and transversal designs are examples of a partial geometry. Codes for distributed storage systems based on partial geometries were also considered in [20].
5 Rate of FR Codes and Generalized Hamming Weights

Let \( C \) be a \([\theta, k]\) linear code and \( D \) be a subcode of \( C \). The support of \( D \), denoted by \( \chi(D) \), is defined as follows:

\[
\chi(D) = \{ i : \exists (c_1, c_2, \ldots, c_{\theta}) = c \in D, c_i \neq 0 \}.
\]

The \( r \)-th generalized Hamming weight of a linear code \( C \), denoted by \( d_r(C) \), is the minimum support of any \( r \)-dimensional subcode of \( C \), \( 1 \leq r \leq k \), namely,

\[
d_r(C) = \min_D \{ |\chi(D)| : D \subseteq C, \dim(D) = r \}.
\]

The set \( \{d_1, d_2, \ldots, d_k\} \) is called the generalized Hamming weight hierarchy of \( C \) \([32]\).

There are a few definitions of generalized Hamming weights for nonlinear codes \([9][13][25]\). We propose now a straightforward analog definition for generalized Hamming weight hierarchy for nonlinear codes. This generalization is strongly connected to the different rates of a given FR code \( C \).

Let \( C \) be a code of length \( \theta \) with \( n \) codewords. Assume further that the all-zero vector is not a codeword of \( C \) (if the all-zero vector is a codeword of \( C \) we omit it from the code). The \( r \)-th generalized Hamming weight of \( C \), \( d_r(C) \), is the minimum support of any subcode of \( C \) with \( r \) codewords. Note that an \((n, k, \alpha, \rho)\) FR code \( C \) can be represented as a binary constant weight nonlinear code \( C \) of length \( \theta \) and weight \( \alpha \). Note further that the minimum Hamming distance of \( C \) is \( 2\alpha - 2 \). Finally note that with these definitions we have that \( R_{C}(k) = d_k(C) \).

Therefore, by our previous discussion and the definition of the generalized Hamming weight hierarchy it is natural to define the rate hierarchy of a FR code \( C \) to be the same as the generalized Hamming weight hierarchy of the related binary constant weight code \( C \).

In addition to the questions discussed in the previous sections, the definition of the rate hierarchy raises some natural questions.

1. Do there exist two FR codes \( C_1 \) and \( C_2 \), with the same parameters \( n, \alpha \) and \( \rho \), and two integers \( k_1 \) and \( k_2 \), such that \( R_{C_1}(k_1) < R_{C_2}(k_1) \) and \( R_{C_1}(k_2) > R_{C_2}(k_2) \)?

2. Given \( n, \alpha \) and \( \rho \), do there exist a FR code \( C \) with these parameters, which satisfies for each \( k \leq \alpha \) that \( R_{C}(k) = C_{FR}(n, k, \alpha, \rho) \)?

6 Fractional Repetition Codes and Batch Codes

In this section we analyze additional properties of FR codes which allow parallel independent reads of the stored data by establishing a connection to combinatorial batch codes. We consider a scenario when in addition to the uncoded repairs of failed nodes and to the recoverability of the stored file from any set of \( k \) nodes, it is possible to reconstruct a \( t \)-subset of data symbols by reading at most one element from each node. In other words, we are interested in balancing the load of partial data reconstruction, performed potentially by several users independently and in parallel.

We note that a family of codes called batch codes has exactly this property. Batch codes, introduced in \([16]\), represent the distributed storage of a data set with \( \theta \) items on \( n \) servers in such a way that any batch of \( t \) data items can be decoded by reading at most one item from each server, while keeping the total storage in the \( n \) servers equal to \( N \). In \( \rho \)-uniform combinatorial batch code, proposed in \([22]\), each server stores a subset of data items and the decoding is permormed only by reading items from the servers. Each item is stored in exactly \( \rho \) servers and hence it is also called a replication based batch code. A \( \rho \)-uniform combinatorial batch code is denoted by \( \rho - (\theta, N, t, n)-C\text{CBC} \) and it has total storage \( N = \rho \theta \). These codes were studied in \([6][8][22][28]\).

We identify the servers of a batch code with the nodes of a code for DSS and the items of a batch code with the symbols stored in the DSS. The retrieval of \( t \) data items in a batch code is identified with the independent parallel
reads of a \(t\)-subset from the stored data of the DSS. Recall that the whole data should be recovered from any set of \(k\) nodes of the DSS. In particular, we analyze the FR codes presented in the previous sections as \(\rho\)-uniform combinatorial batch codes. In this context, we need to analyze the parameter \(t\) of the code, as other parameters are obvious. Note that while the batch property of a FR code allows to retrieve a \(t\)-subset of the \(\theta\) symbols of a codeword of the \((\theta, M)\)-MDS code, used in the first step of encoding a file of size \(M\), by choosing a systematic MDS code, any \(t\)-subset of the file symbols could be retrieved. First, we need the following results on batch codes.

**Theorem 22.** \([22]\) Let \(G\) be a graph with \(n\) vertices, \(\theta\) edges and girth \(g\). Then the batch code \(C^B_G\) with servers indexed by the vertices of \(G\) and with items indexed by the edges of \(G\), is a \(2-(\theta, 2\theta, t, n)\)-CBC with \(t = 2g - \lfloor g/2 \rfloor - 1\).

**Theorem 23.** \([28]\) Let \(TD\) be a resolvable transversal design \(TD(q - 1, q)\), for a prime power \(q\). Then the batch code \(C^B_TD\) with servers indexed by points and items indexed by blocks of \(TD\), is a \((q - 1) - (q^2, q^2 - q^2, q^2 - q - 1, q^2 - q)\)-CBC.

By applying Theorems 22 and 23 we obtain the following result for some FR codes constructed in the previous sections.

**Corollary 24.** Consider the following three families of FR codes.

- The FR code \(C_{K,\alpha,\alpha}\) allows 5 independent parallel reads for any \(\alpha\).
- The FR code \(C_G\) constructed from a regular graph \(G\) with girth \(g\) allows \(2g - \lfloor g/2 \rfloor - 1\) independent parallel reads.
- The FR code \(C_{TD}\) constructed from a resolvable transversal design \(TD(\rho, \alpha)\) with \(\alpha = \rho + 1 = q\), for prime power \(q\), allows \(q^2 - q - 1 = n - 1\) independent parallel reads.

**Example 5.**

- Consider the \((6, 3, 3, 2)\) FR code \(C_{K,3,3}\) from Example 1. By Corollary 24 any 5 out of 9 symbols of an MDS codeword can be read independently, and in particular any 5 out of the 7 information symbols, if a systematic \((9, 7)\) MDS code is used.
- Consider the \((12, 4, 4, 2)\) FR code \(C_{TD}\) from Example 3. By Corollary 24 any 11 out of 16 symbols of the MDS codeword can be read independently, and in particular all the 11 data symbols.

### 7 Conclusions and Problems for Future Research

We considered the problem of constructing \((n, k, \alpha, \rho, M)\) FR codes which attain the FR capacity for these parameters. We presented several constructions and schemes based on regular and biregular graphs, graphs with a given girth, transversal designs, projective planes, generalized polygons and their incidence matrices. The rate of some FR codes obtained from these schemes attain a known upper bound. The rate of some other codes is proved to be optimal, and as a consequence the bound on the FR capacity is improved for their parameters.

Finding the value of the FR capacity can be formulated in terms of graph theory. We have discussed this problem and some related ones. For large range of parameters solutions for these problems were given.

In general, given four of the five parameters of FR codes, namely, the number \(n\) of nodes, the number \(k\) of nodes needed to reconstruct the whole stored file, the number \(\alpha\) of stored symbols in a node, the number \(\rho\) of repetitions of a symbol in the code, and the size \(M\) of the stored file \(f\), one can ask what are the possible values of the fifth parameter. For this we define the following five functions.

1. Let \(N(k, \alpha, \rho, M)\) be the minimum number \(n\) of nodes in an \((n, k, \alpha, \rho)\) FR code which stores a file of size \(M\).
2. Let $K(n, \alpha, \rho, M)$ be the minimum number $k$ of nodes from which the whole stored file of size $M$, of an $(n, k, \alpha, \rho)$ FR code, can be reconstructed.

3. Let $A(n, k, \rho, M)$ be the minimum degree $\alpha$ of a node in an $(n, k, \alpha, \rho)$ FR code which stores a file of size $M$.

4. Let $P(n, k, \alpha, M)$ be the minimum number $\rho$ of repetitions of a symbol in an $(n, k, \alpha, \rho)$ FR code which stores a file of size $M$.

5. Let $C_{FR}(n, k, \alpha, \rho)$ be the fractional repetition capacity, i.e., the maximum size of a stored file in an $(n, k, \alpha, \rho)$ FR code, also known as the rate of the code.

In this paper, we mainly considered the values of $C_{FR}(n, k, \alpha, \rho)$, but we also considered the values of $N(k, \alpha, \rho, M)$. Some values of the other functions are either simple to compute or can be also deduced from our discussion. We note that the choice of a parameter to optimize, i.e., the choice of one of the functions defined above, depends on the requirements to the specific DSS.

The existence problem of FR codes, with $\rho = 2$, for any given rate in the range between $k\alpha - \left(\frac{k}{2}\right)$ and $k\alpha - k + 1$ was considered in Subsection 3.1. Generally, one can prove that any rate in this range can be obtained. To prove this claim one can start with a known graph $G = (V, E)$ in which there is no subgraph of $K_t$ with $t$ edges, $t \leq \left(\frac{k}{2}\right)$. Let $H$ be a subgraph of $G$ with $k$ vertices and $t - 1$ edges; let $v_1, v_2$ two vertices in $H$ and $v_3, v_4$ two vertices in $G \setminus H$ such that $e_1 = \{v_1, v_2\}, e_2 = \{v_3, v_4\} \notin E$ and $e_3 = \{v_1, v_3\}, e_4 = \{v_2, v_4\} \in E$. One can easily verify that the graph $G \setminus \{e_3, e_4\} \cup \{e_1, e_2\}$ contains a subgraph of $K_t$ with $t$ edges but does not contain a subgraph of $K_t$ with $t + 1$ edges.

We considered the existence problem of FR codes with minimum number of nodes, with $\rho = 2$, for any given rate in the range between $k\alpha - \left(\frac{k}{2}\right)$ and $k\alpha - k + 1$, only for $k = 3$ and $k = 4$. The problem is getting more complicated as $k$ increases. For example, we consider now constructions of codes, based only on our previous discussion, for $k = 5$. For a given FR code $C$ the possible values of $R_C(5)$ belong to the set $\{5\alpha - i : 4 \leq i \leq 10\}$.

- A code $C$ for which $R_C(5) = 5\alpha - 9$ is obtained from an $(n, 4)$-Turán graph.
- A code $C$ for which $R_C(5) = 5\alpha - 8$ is obtained from an $(n, 3)$-Turán graphs.
- A code $C$ for which $R_C(5) = 5\alpha - 6$ is obtained from an $(n, 2)$-Turán graph, which is also a graph with girth 4.
- A code $C$ for which $R_C(5) = 5\alpha - 5$ is obtained from a graph with girth 5.
- A code $C$ for which $R_C(5) = 5\alpha - 4$ is obtained from a graph with girth 6.

Next, we examine the storage overhead of DSS, by using a specific example. Suppose we want to store a file of size 36 by using a FR code with $n$ nodes, where each node stores 8 symbols and the repetition degree $\rho = 2$. We can use a $(114, 5, 8, 2)$ FR code based on the incidence graph of a projective plane of order 7, according to Corollary 1. The length of a corresponding MDS codeword is 456 and hence the storage overhead is much more than 1000\%. To have smaller overhead one can use a $(10, 5)$-Turán graph which by Theorem 7 yields a $(10, 7, 8, 2)$ FR code for which we can take a file of size 37 and encode it with a $(40, 37)$ MDS code. Hence, the total overhead is only about 10\%. This example illustrates the fact that given the file size $M$ and the number of symbols $\alpha$ stored in a node, decreasing the number of nodes $k$ used to recover the whole file increases the storage overhead significantly. However, even so the overhead for small $k$ is much higher, it is desirable for various reasons to access as less nodes as possible when one needs to reconstruct the file. It is also desirable that the data/storage ratio will be small and hence $\rho$ should be as small as possible. Therefore, this trade-off should be taken into account by the designer of a DSS.

Finally, we considered the rate hierarchy of FR codes and connections between FR codes and batch codes in the context of parallel independent reads by several users. These concepts raise some more interesting problems for future research.
References

[1] M. Abreu, G. Araujo-Pardo, C. Balbuena, and D. Labbate, “Families of small regular graphs of girth 5,” *Discrete Mathematics* 312 (2012), 2832-2842.

[2] M. Abreu, G. Araujo-Pardo, C. Balbuena, D. Labbate, and G. López-Chávez, “Constructions of biregular cages of girth five,” *Electronic Notes in Discrete Mathematics* 40 (2013), 9-14.

[3] I. Anderson, *Combinatorial designs and tournaments*, Clarendon Press, Oxford, 1997.

[4] S. Anil, M. K. Gupta, and T. A. Gulliver, “Enumerating some fractional repetition codes,” *arXiv:1303.6801*, Mar. 2013.

[5] G. Araujo-Pardo, C. Balbuena, and J.C. Valenzuela, “Constructions of bi-regular cages,” *Discrete Mathematics* 309 (2009), 1409-1416.

[6] N. Balachandran and S. Bhattacharya, “On an extremal hypergraph problem related to combinatorial batch codes”, *arXiv:1206.1996v4*, Oct. 2012.

[7] S. Bhattacharya, S. Ruj, and B. Roy, “Combinatorial batch codes: A lower bound and optimal constructions,” *Advances in Mathematics of Communications* 6 (2012), 165-174.

[8] Cs. Bujtás and Zs. Tuza. “Turán numbers and batch codes”, *arXiv:1309.6506v1*, Sep. 2013.

[9] G. Cohen, S. Litsyn, and G. Zemor, “Upper bounds on generalized distances,” *IEEE Trans. Inform. Theory*, vol. 40, pp. 2090-2092, Nov. 1994.

[10] A. G. Dimakis, P. Godfrey, M. Wainwright and K. Ramachandran, “Network coding for distributed storage system,” *IEEE Trans. on Inform. Theory*, vol. 56, no. 9, pp. 4539-4551, Sep. 2010.

[11] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, “A survey on network codes for distributed storage,” in *Proc. of the IEEE*, pp. 476–489, Mar. 2011.

[12] G. Exoo and R. Jajcay, “Dynamic Cage Survey,” *Electron. J. Combin.*, Dynamic survey: DS16, (2008).

[13] T. Etzion and A. Vardy, “On perfect codes and tilings: Problems and solutions,” *SIAM J. Discrete Math.*, vol. 11, pp. 205-223, 1998.

[14] Z. Furedi, F. Lazebnik, A. Seress, V. A. Ustimenko, ans A. J. Woldar, “Graphs of prescribed girth and bi-degree,” *J. Combin. Theory*, Ser. B 64 (2) (1995), 228–239.

[15] C. Godsil, R. Royle, *Algebraic Graph Theory*, Springer, Graduate Texts in Mathematics, vol. 207, 2001.

[16] Y. Ishai, E. Kushilevitz, R. Ostrovsky, and A. Sahai, “Batch codes and their applications,” in *Proc. 36th annual ACM symposium on Theory of computing*, vol. 36, pp. 262–271, 2004.

[17] J. C. Koo and J. T. Gill. III, “Scalable constructions of fractional repetition codes in distributed storage systems,” in *Proc. 49th Annual Allerton Conf. on Communication, Control, and Computing (Allerton)*, pp. 1366–1373, 2011.

[18] U. S. R. Murty, “A generalization of the Hoffman - Singleton graph,” *Ars Combin.* 7 (1979), 191-193.

[19] O. Olmez and A. Ramamoorthy, “Repairable replication-based storage systems using resolvable designs,” in *Proc. 50th Annual Allerton Conf. on Communication, Control, and Computing (Allerton)*, pp. 1174 – 1181, 2012.
[20] L. Pamies-Juarez, H. D. L. Hollmann, and F. E. Oggier, “Locally repairable codes with multiple repair alternatives,” in Proc. IEEE ISIT, pp. 2338-2342, Aug. 2011.

[21] S. Pawar, N. Noorshams, S. El Rouayheb, and K. Ramchandran, “Dress codes for the storage cloud: Simple randomized constructions,” in Proc. IEEE ISIT, pp. 892 – 896, Jul. 2013.

[22] M. B. Patterson, D. R. Stinson, and R. Wei, “Combinatorial batch codes,” Advances in Mathematics of Communications, vol. 3, pp. 13–27, 2009.

[23] K. Rashmi, N. B. Shah, P. V. Kumar, and K. Ramchandran, “Explicit construction of optimal exact regenerating codes for distributed storage,” in Proc. 47th Annual Allerton Conf. on Communication, Control, and Computing (Allerton), pp. 1243 - 1249, 2009.

[24] 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, pp. 5227-5239, Aug. 2011.

[25] I. Reuven and Y. Beery, “Generalized Hamming weights of nonlinear codes and the relation to the Z4-linear representation,” IEEE Trans. Inform. Theory, vol. 45, pp. 713-720, Mar. 1999.

[26] S. El Rouayheb and K. Ramchandran, “Fractional repetition codes for repair in distributed storage systems,” in Proc. 48th Annual Allerton Conf. on Communication, Control, and Computing (Allerton), pp. 1510 –1517, 2010.

[27] N. B. Shah, K. V. Rashmi, P. V. Kumar, and K. Ramchandran, “Explicit codes minimizing repair bandwidth for distributed storage,” in Proc. IEEE ITW, pp. 1–5, Jan. 2010.

[28] N. Silberstein and A. Gál, “Optimal combinatorial batch codes based on block designs”, arXiv:1312.5505, Dec. 2013.

[29] C. Suh and K. Ramchandran, “Exact-repair MDS codes for distributed storage using interference alignment,” in Proc. IEEE ISIT, pp. 161–165, Jul. 2010.

[30] 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.

[31] C. Tian, V. Aggarwal, and V. Vaishampayan, “Exact-repair regenerating codes via layered erasure correction and block designs,” arXiv:1302.4670, Feb. 2013.

[32] V. K. Wei, “Generalized Hamming weights for linear codes,” IEEE Trans. Inform. Theory, vol. 37, pp. 1412-1418, Sep. 1991.