Minimizing the alphabet size of erasure codes with restricted decoding sets

Mira Gonen, Ishay Haviv, Michael Langberg, Alex Sprintson

***Abstract*—A Maximum Distance Separable code over an alphabet F is defined via an encoding function $C : F^k \to F^n$ that allows to retrieve a message $m \in F^k$ from the codeword $C(m)$ even after erasing any $n-k$ of its symbols. The minimum possible alphabet size of general (non-linear) MDS codes for given parameters n and k is unknown and forms one of the central open problems in coding theory. The paper initiates the study of the alphabet size of codes in a *generalized* setting where the coding scheme is required to handle a pre-specified subset of all possible erasure patterns, naturally represented by an n-vertex k-uniform hypergraph. We relate the minimum possible alphabet size of such codes to the strong chromatic number of the hypergraph and analyze the tightness of the obtained bounds for both the linear and non-linear settings. We further consider variations of the problem which allow a small probability of decoding error.**

I. Introduction

Maximum Distance Separable codes are known to play an important and influential role in the area of coding theory. An MDS code over an alphabet F is defined via an encoding function $C : F^k \to F^n$ that allows to retrieve a message $m \in F^k$ from the codeword $C(m)$ even after erasing any $n-k$ of its symbols. Equivalently, the Hamming distance between any two distinct codewords is at least $n-k+1$. The well-known Singleton bound implies that MDS codes are optimal with respect to the number of erasures that they can handle. However, the minimum possible alphabet size of such codes for given parameters n and k is unknown and forms a central open question in coding theory (see Conjectures 1 and 2).

The present paper initiates the study of the alphabet size of codes in a *generalized* setting where the coding scheme is required to handle a pre-specified subset of all possible erasure patterns. Such scenarios arise naturally in the distributed storage settings in which there is a need to rebuild the entire data set by contacting one of the pre-specified sets of storage nodes, referred to as a *recovery group*. The desired set of recovery groups is determined based on the network configuration, the reliability of storage nodes, as well as the network access patterns. Similar constraints appear in *availability codes* [1], which have recently attracted interest from the research community. However, in availability codes each recovery group is used for retrieving a single symbol, whereas we are focusing on retrieving the entire set of k symbols.

Mira Gonen is with the Department of Computer Science, Ariel University, Ariel 40700, Israel (e-mail: mirag@ariel.ac.il).

Ishay Haviv is with the School of Computer Science, The Academic College of Tel Aviv-Yaffo, Tel Aviv 61083, Israel.

Michael Langberg is with the Department of Electrical Engineering, State University of New-York at Buffalo, Buffalo, NY 14260, USA (e-mail: mikel@buffalo.edu). Work supported in part by NSF grant 1909451.

Alex Sprintson is with the Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA (e-mail: spalex@tamu.edu). Work supported in part by NSF grants 1642983 and 1718658.

The set of erasure patterns is naturally represented by an n-vertex k-uniform hypergraph in which the vertices represent the n coordinates of the codewords and the (hyper)edges[1] correspond to the possible sets of locations of unerased symbols (i.e., *decoding sets*). For a given uniform hypergraph G we are interested in minimizing the size of an alphabet over which there exists a coding scheme with respect to the erasure patterns defined by G.

Definition 1 (The q parameter) *Let $G = ([n], E)$ be a k-uniform hypergraph on the vertex set $[n] = \{1, \ldots, n\}$. Let $q(G)$ denote the smallest size q of an alphabet F for which there exist an encoding function*

$$C : F^k \to F^n$$

and a decoding function

$$D : (F \cup \{\perp\})^n \to F^k$$

such that for every edge $e \in E$ and every message $m \in F^k$ it holds that

$$D(C_e(m)) = m.$$

Here, $C_e(m)$ stands for the word obtained from the codeword $C(m)$ by replacing the symbols in the locations of $[n] \setminus e$ by the erasure symbol $\perp$.

Similarly, let $q_{lin}(G)$ denote the smallest prime power q for which there exist linear *encoding and decoding functions as above when F is a field of size q.*

Observe that for the complete n-vertex k-uniform hypergraph, denoted by $\kappa_{n,k}$, the values of $q(\kappa_{n,k})$ and $q_{lin}(\kappa_{n,k})$ are equal, respectively, to the minimum alphabet sizes of general and linear MDS codes of length n and dimension k. We state below the MDS conjectures for general and for linear codes (see, e.g., [2]–[5]).

Conjecture 1 (MDS Conjecture for general codes) *For given integers $k < q \neq 6$, let $n(q,k)$ be the largest integer n such that $q(\kappa_{n,k}) \leq q$. Then,*

$$n(q,k) \leq \begin{cases} q+2 & \text{if } 4|q \text{ and } k \in \{3, q-1\} \\ q+1 & \text{otherwise.} \end{cases} \quad (1)$$

[1]For clarity, in the rest of the paper we refer to the hyperedges of a hypergraph as *edges*.

Conjecture 2 (MDS Conjecture for linear codes) *For given integers $k < q$ where q is a prime power, let $n(q,k)$ be the largest integer n such that $q_{lin}(\kappa_{n,k}) \leq q$. Then,*

$$n(q,k) \leq \begin{cases} q+2 & \text{if } q \text{ is even and } k \in \{3, q-1\} \\ q+1 & \text{otherwise.} \end{cases} \quad (2)$$

Note that in Conjecture 2 the right-hand side is known to form a lower bound on the left-hand side, and that the special case of the above conjectures for $k = 2$ is known to hold (see, e.g., [5]). The MDS Conjecture for linear codes over prime fields has been proven by S. Ball in his seminal paper [6].

In this work we aim to study the behavior of the q parameter for general sub-hypergraphs of $\kappa_{n,k}$. Our results imply strong relations between the q parameter of uniform hypergraphs and their chromatic number. A *valid* coloring of a hypergraph G is an assignment of colors to its vertices so that the vertices of each edge are assigned to distinct colors. This is at times referred to as a *strong* coloring and is consistent with the notion of graph coloring (i.e., the coloring of hypergraphs with edges of size two). The chromatic number $\chi(G)$ of G is the minimum number of colors that allows a valid coloring of G.

II. Our results

In what follows we give an overview of our results. Some of the proofs can be found in Section III. All proofs appear in the full version of this work [7].

A. Relation between q and χ

We start with the following upper bounds.

Theorem 1 *For every k-uniform hypergraph G,*

$$q(G) \leq q(\kappa_{\chi(G),k}) \quad \text{and} \quad q_{lin}(G) \leq q_{lin}(\kappa_{\chi(G),k}).$$

In particular,

$$q(G) \leq q_{lin}(G) \leq [\chi(G)-1]_{pp}.$$

Here, for an integer x, $[x]_{pp}$ represents the smallest prime power that is greater or equal to x.

Theorem 1 formalizes the natural intuition that for *simple* collections of erasure patterns G, i.e., the setting in which $\chi(G)$ is small, a *small* alphabet size q suffices for a suitable erasure code. The question we next consider is whether the upper bound provided by Theorem 1 is tight.

B. Tightness of Theorem 1, the case $k \geq 3$

The following result shows that Theorem 1 is not tight in general.

Proposition 1 *There exists a 3-uniform hypergraph G with $q_{lin}(G) = q(G) = 2$ and yet $q(\kappa_{\chi(G),3}) \geq 5$.*

We further show that for every $k \geq 3$ the chromatic number of k-uniform hypergraphs can be significantly larger than their q parameter (even while restricted to the linear setting). This implies a large gap between the q parameter and its upper bound provided by Theorem 1.

Proposition 2 *For every $k \geq 3$ and every prime power q, there exists a k-uniform hypergraph G with $q_{lin}(G) \leq q$ and yet $\chi(G) \geq \frac{q^k-1}{q-1}$.*

For $k = 2$, the question at hand seems to be more challenging. Here we ask whether there exists a graph G for which $q(G)$ is significantly smaller than $q_{lin}(G)$ which in turn is known to be at most $[\chi(G)-1]_{pp}$. We thus study the relationship between $q(G)$ and $\chi(G)$. Our results for this case are outlined below.

C. Tightness of Theorem 1, the case $k = 2$

To study the relationship between q and χ for the case $k = 2$, we define the following graph family.

Definition 2 (The graph family G_q) *For an integer q, let G_q be the graph whose vertex set consists of all the balanced vectors of length q^2 over $[q]$, that is, the vectors $u \in [q]^{q^2}$ such that $|\{i \in [q^2] \mid u_i = j\}| = q$ for every $j \in [q]$, where two vertices $u = (u_1, \ldots, u_{q^2})$ and $v = (v_1, \ldots, v_{q^2})$ are adjacent if the collection of pairs $\{(u_i, v_i)\}_{i \in [q^2]}$ is equal to $[q] \times [q]$.*

The graph family G_q is extremal with respect to the q parameter in the sense given by the following lemma.

Lemma 1 *For every integer q,*

1) *$q(G_q) \leq q$ and*
2) *$\chi(G) \leq \chi(G_q)$ for every graph G with $q(G) = q$.*

By Lemma 1, the challenge of obtaining graphs with chromatic number much larger than the q parameter reduces to the study of the chromatic number $\chi(G_q)$ of G_q. We first show that if we require the coding scheme of G_q to be linear then the size of the used alphabet cannot be smaller than $[\chi(G_q)-1]_{pp}$ (implying that Theorem 1 is tight in this case). Recall that this is in contrast to the situation of $k \geq 3$ (see Proposition 2).

Proposition 3 *For every integer q,*

$$q_{lin}(G_q) = [\chi(G_q)-1]_{pp}.$$

And what about non-linear codes? Is $q(G_q)$ significantly smaller than $\chi(G_q)$? For $q = 2$, it is not difficult to see that $\chi(G_q) = 3 = q(G_q)+1$. However, an exhaustive analysis due to [8] shows that for $q = 3$ it holds that $\chi(G_q) = 6 > q(G_q)+1$. For general values of q, we provide the bounds stated below. We use here the notation $n(q,2)$ which stands for the largest integer n for which there exists an MDS code of length n and dimension 2 over an alphabet of size q. Note that if q is a prime power then $n(q,2) = q+1$.

Proposition 4 *For every integer q,*

$$n(q,2) \leq \chi(G_q) \leq \binom{q+1}{2}.$$

In particular, if q is a prime power then

$$q+1 \leq \chi(G_q) \leq \binom{q+1}{2}.$$

It is interesting to understand the asymptotic behavior of $\chi(G_q)$ as a function of q. In an attempt to shed some light on this question, we provide a couple of related results, described next.

D. *Understanding $\chi(G_q)$*

Firstly, we consider a natural family of independent sets of G_q which we refer to as *canonical* independent sets. The canonical independent set $A_{i,j}$ associated with two distinct indices $i, j \in [q^2]$ is the set of all vertices $u \in [q]^{q^2}$ of G_q that satisfy $u_i = u_j$. For a prime power q, it can be seen that $A_{i,j}$ is an independent set of maximum size in G_q. In fact, this type of independent set is used to obtain the upper bound in Proposition 4. However, we show that if we restrict ourselves to colorings of G_q whose color classes are all contained in canonical independent sets then the number of used colors has to be quadratic in q (see Proposition 5) and is thus close to the upper bound of Proposition 4.

Secondly, we focus on the subgraph of G_q induced by the vertices whose vectors in $[q]^{q^2}$ form concatenations of q permutations of $[q]$. It is shown that the chromatic number of this subgraph is only linear in q, corresponding now to the lower bound of Proposition 4. Intuitively speaking, this might hint that the difficulty in coloring the graph G_q using few colors comes from the 'less-structured areas' of the graph (see Proposition 6).

We conclude our work with a further extension of the q parameter to an error-tolerant model.

E. *Allowing an error in decoding*

In this last study, for a given uniform hypergraph G and an error parameter $\varepsilon > 0$, we are interested in minimizing the size of the alphabet over which there exists a coding scheme with respect to the erasure patterns defined by G that guarantees success probability at least $1 - \varepsilon$. We first study the setting in which the success probability is taken over the uniform distribution on the messages. This is given formally in the following definition.

Definition 3 (The q_ε parameter) *Let $G = ([n], E)$ be a k-uniform hypergraph on the vertex set $[n]$ and let $\varepsilon > 0$. Let $q_\varepsilon(G)$ denote the smallest size q of an alphabet F for which there exist an encoding function $C : F^k \to F^n$ and a decoding function $D : (F \cup \{\perp\})^n \to F^k$ such that for every edge $e \in E$ it holds that*

$$\Pr_m[D(C_e(m)) = m] \geq 1 - \varepsilon,$$

where m is uniformly chosen from F^k.

Notice that the q parameter given in Definition 1 coincides with Definition 3 when $\varepsilon = 0$. Similar to Theorem 1, the following holds.

Theorem 2 *For every $\varepsilon \geq 0$ and a k-uniform hypergraph G,*

$$q_\varepsilon(G) \leq q_\varepsilon(\kappa_{\chi(G),k}) \leq q(\kappa_{\chi(G),k}).$$

We focus our study on the case $k = 2$. To demonstrate Definition 3, we consider the complete graph $\kappa_{n,2}$. For $\varepsilon = 0$, this takes us to 2-dimensional MDS codes. However, if the probability of success is slightly relaxed it turns out that the required alphabet size can be reduced. For example, in Proposition 7 of Section IV we show that

- for $\varepsilon = \frac{1}{3}$, $q_\varepsilon(\kappa_{20,2}) \leq 3$ whereas $q(\kappa_{20,2}) = 19$,
- for $\varepsilon = \frac{1}{4}$, $q_\varepsilon(\kappa_{7,2}) \leq 4$ whereas $q(\kappa_{7,2}) = 6$, and
- for $\varepsilon = \frac{1}{6}$, $q_\varepsilon(\kappa_{6,2}) \leq 6$ whereas $q(\kappa_{6,2}) = 7$.

Notice that, for $k = 2$, $\varepsilon > 0$ implies $\varepsilon \geq \frac{1}{q^2}$ (as our error is measured over a sample space of size q^2). Thus, as a first step in understanding q_ε, in the analysis above and those that follow, we study ε for the intermediate value of $\frac{1}{q}$. A full study addressing q_ε for general ε is left for future work.

In Section IV, we study the q_ε parameter for general graphs G (with $k = 2$ and $\varepsilon = 1/q$). As in our study of the q parameter, we employ the tool of universal graphs (i.e., an analog to the graph family of Definition 2) and apply it for the error-tolerant setting. See Propositions 8, 9, and 10 in Section IV.

Finally, in Section IV-C, we further extend the notion of error to allow an average error $\tilde{\varepsilon}$ over both the messages and the edges in E. Here, the decoding error is computed assuming a uniform set of k messages and a uniform decoding edge in E (see Definition 6 in Section IV-C). Roughly speaking, we show in Theorem 3 that this last notion of error allows significant flexibility in the sense that for a given $\tilde{\varepsilon}$ the relaxed $q_{\tilde{\varepsilon}}$ parameter is bounded by approximately $1/\tilde{\varepsilon}$ on complete graphs.

III. The q Parameter

A. *Proof of Theorem 1*

Let G be a k-uniform hypergraph on the vertex set $[n]$ and let $\chi = \chi(G)$. Denoting $q = q(\kappa_{\chi,k})$, it follows that for an alphabet F of size q there exist an encoding function $C : F^k \to F^\chi$ and a decoding function $D : (F \cup \{\perp\})^\chi \to F^k$ such that for every k-subset e of $[\chi]$ and every message $m \in F^k$ it holds that $D(C_e(m)) = m$.

To prove that $q(G) \leq q$ we define a coding scheme over the alphabet F as follows. Fix a valid coloring $g : [n] \to [\chi]$ of G. Consider the encoding function $\widetilde{C} : F^k \to F^n$ that given a message $m \in F^k$ outputs the vector in F^n whose ith entry $\widetilde{C}_i(m)$ is $C_{g(i)}(m)$, i.e., the symbol in the codeword $C(m)$

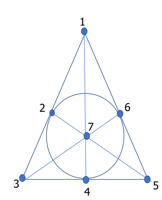

Fig. 1. Illustration of the Fano matroid.

which corresponds to the color of the ith vertex. Here, and throughout, we use the notation $C_i(m)$ to denote the ith entry in the codeword $C(m)$. It remains to show that given a word $\widetilde{C}_e(m) \in (F \cup \{\perp\})^n$ for an edge e of G and a message $m \in F^k$, it is possible to retrieve m. Indeed, consider the word $\widetilde{C}_e(m)$ restricted to the entries corresponding to an edge e (recall that the value of these entries is not the erasure symbol $\perp$). As the vertices of e are colored by g using distinct colors, the values of the corresponding entries in $\widetilde{C}_e(m)$ equal to k distinct entries in $C(m)$. As $C(m)$ is decodable from any set of k distinct entries via D, we conclude that given $\widetilde{C}_e(m)$ we can retrieve m as well. An identical proof shows that $q_{lin}(G) \leq q_{lin}(\kappa_{\chi(G),k}) \leq [\chi(G)-1]_{pp}$ (where the rightmost inequality follows from the known upper bounds on q_{lin} as discussed after Conjecture 2).

B. *Proof of Proposition* 1

The proof is based on the Fano plane illustrated in Figure 1. Recall that the Fano plane is defined over a set of 7 points, denoted here by the integers in [7], and consists of 7 lines with 3 points on every line and 3 lines on every point. Let G be the 3-uniform hypergraph on the vertex set [7] whose edges are all the 3-subsets of [7] that do *not* form lines in the Fano plane. We claim that G satisfies the assertion of the proposition. To this end, we turn to show that

1) $\chi(G) = 7$,
2) $q_{lin}(G) = q(G) = 2$, and
3) $q(\kappa_{7,3}) \geq 5$.

For Item 1, observe that every two vertices of G are included in some edge of G, hence every valid coloring of G assigns every vertex to a distinct color, implying that $\chi(G) = 7$. For Item 2, consider the three vectors $v_1 = (0,0,0,1,1,1,1)$, $v_2 = (0,1,1,1,0,0,1)$, and $v_3 = (1,1,0,1,1,0,0)$ over the binary field F_2, and let $C : F_2^3 \to F_2^7$ be a linear encoding function whose image is the linear span of $\{v_1, v_2, v_3\}$. Namely, $C(m_1, m_2, m_3) = \sum_{i=1}^{3} m_i v_i$. It is straightforward to verify that if we restrict the images of the function C to the coordinates of any edge of G we get an invertible function from F_2^3 to F_2^3. For example, restricting C to coordinates $\{1,2,4\}$ gives a function whose image is spanned by the restrictions of $\{v_1, v_2, v_3\}$ to these coordinates, i.e., a function with image spanned by $\{(0,0,1),(0,1,1),(1,1,1)\}$. As the latter vectors are linearly independent over F_2 it follows that the restriction of C at hand is invertible. We conclude that for every edge e of G and for every message $m \in F_2^3$, one can decode m from the codeword $C(m)$ even if the symbols in the locations of $[7] \setminus e$ are erased, hence $q_{lin}(G) \leq 2$. Since it clearly holds that $q_{lin}(G) \geq q(G) \geq 2$, Item 2 follows. Finally, Item 3 follows from the fact that $q(\kappa_{n,k}) \geq n-k+1$ for every $n \geq k \geq 2$ (see [9]).

C. *Proof of Proposition* 2

Let F be a field of size q. Consider a k-uniform hypergraph G whose vertices correspond to *normalized* vectors in F^k. Here, a normalized vector $(x_1, \ldots, x_k) \in F^k$ is one in which the leading nonzero entry equals 1. The number of vertices n in G is thus $\sum_{i=1}^{k} q^{k-i} = (q^k-1)/(q-1)$. The edge set of G consists of all k-collections of vertices in G that correspond to linearly independent vectors. As any two vertices in G have corresponding vectors that can be completed to a linearly independent set of size k, any two vertices in G are included in at least one edge in G. Thus $\chi(G) = n = (q^k-1)/(q-1)$. In what follows, we use a natural relation between vertices in G and linear functions $F^k \to F$. Specifically, if vertex i in G is defined by the vector $(x_1, \ldots, x_k)$, then the function f_i corresponding to i maps $m = (m_1, \ldots, m_k) \in F^k$ to $f_i(m) = \sum_{\ell=1}^{k} x_\ell m_\ell$. Now, let $C : F^k \to F^n$ be a linear code which maps messages $m = (m_1, \ldots, m_k)$ to the codeword $C(m)$ using the relation above. Namely, the ith entry $C_i(m)$ of $C(m)$ is defined to be $f_i(m)$. As edges in G consist of vertices that correspond to linearly independent vectors, it follows that for every edge in G, the message m can be recovered from $C_e(m)$. This implies that $q_{lin}(G) \leq q$.

D. *Proof of Lemma* 1

For an integer q, let n_q denote the number of vertices in the graph G_q. Recall that the vertices of G_q are the balanced vectors of $[q]^{q^2}$, and note that each of them can be realized as a function from F^2 to F, where F is an alphabet of size q. Namely, with each vertex v in G_q we associate a function $f_v(m)$ that takes $m \in F^2$ and returns an element in F. We turn to show a coding scheme over the alphabet F with respect to the erasure patterns defined by the graph G_q. To this end, consider the encoding function $C : F^2 \to F^{n_q}$ for which for any m, $C_v(m) = f_v(m)$. Here, for any vertex v in G_q, $C_v(m)$ denotes the entry of C corresponding to v. For the decoding, consider a word $C(m)$ and assume that all of its symbols but the two that correspond to some adjacent vertices u and v are erased. We claim that given $C_u(m)$ and $C_v(m)$ it is possible to retrieve m. Indeed, by the definition of G_q, the possible pairs $(C_u(m), C_v(m))$ over all messages $m \in F^2$ are all the distinct pairs in F^2, hence the pair $(C_u(m), C_v(m))$ fully determines the message m. This implies that $q(G_q) \leq q$.

For the second item, let $G = (V,E)$ be a graph with $q(G) = q$. Then there exists a coding scheme over an alphabet F of size q with respect to the erasure patterns defined by G. Let $C : F^2 \to F^{|V|}$ be the encoding function of such a coding scheme. Observe that the existence of a corresponding decoding function implies that for every adjacent vertices u and v in G it holds that the possible pairs $(C_u(m), C_v(m))$ over all messages $m \in F^2$ are all the pairs in F^2. We assign to every non-isolated vertex $v \in V$ the vertex of G_q that represents the function that assigns every $m \in F^2$ to $C_v(m)$. We further assign isolated vertices of G to arbitrary vertices of G_q. This mapping forms a homomorphism from G to G_q, so in particular, it holds that $\chi(G) \leq \chi(G_q)$.

E. *Proof of Proposition* 3

For an integer q, let n_q denote the number of vertices in the graph G_q. Assume in contradiction that for some prime power q', $q_{lin}(G_q) = q' < [\chi(G_q)-1]_{pp}$, and let C be a linear encoding function for the graph G_q over a field F of size q'. Note that for every message $m \in F^2$, $C(m) = (C_1(m), \ldots, C_{n_q}(m))$, where each $C_i(m)$ is a linear function of m. One may represent each such function C_i by a vector $(x_1, x_2) \in F^2$ such that $C_i(m) = x_1 m_1 + x_2 m_2$ for every $m = (m_1, m_2)$. By the decodability of C, it follows that the vectors associated with the endpoints of any edge in G_q are linearly independent (in particular, there are no entries in C which correspond to the zero linear function). Thus, similar to the proof of Proposition 2, we may assume that the vectors corresponding to entries of C are normalized, i.e., their leading nonzero coefficient is 1, since such a normalized C is still an erasure code for G_q. A simple counting argument shows that the number of distinct normalized vectors in F^2 is precisely $q' + 1$. This implies that $\chi(G_q) \leq q' + 1$, since using the distinct normalized vectors of C to represent color classes one could color G_q. This in turn implies that $[\chi(G_q) - 1]_{qq} \leq q'$, in contradiction to our assumption.

F. Proposition 4 and Canonical Independent Sets of G_q

For an integer q consider the graph G_q given in Definition 2. For two distinct indices $i, j \in [q^2]$ let $A_{i,j}$ be the set of all vertices $u \in [q]^{q^2}$ of G_q that satisfy $u_i = u_j$. Every set $A_{i,j}$ forms an independent set in G_q since for every two distinct vertices $u, v \in A_{i,j}$ we have $u_i = u_j$ and $v_i = v_j$, and thus $(u_i, v_i) = (u_j, v_j)$, which implies that the collection of pairs $\{(u_i, v_i)\}_{i \in [q^2]}$ is not equal to $[q] \times [q]$. We refer to such independent sets of G_q as *canonical*. The canonical independent sets of G_q are used in the proof of Proposition 4.

Proposition 5 *For every sufficiently large integer q, the number of canonical independent sets required to cover the vertex set of G_q is $\Omega(q^2)$.*

G. Subgraphs of G_q

Definition 4 (The graphs H_q and H_q^{cyclic}) *Let q be an integer. Let the vector representation of a permutation $\sigma \in S_q$ be $(\sigma(1), \ldots, \sigma(q))$. A vector $(v_1, \ldots, v_{q^2})$ of $[q]^{q^2}$ is said to be a concatenation of q permutations if for every $i \in [q]$ the vector $(v_{q(i-1)+1}, \ldots, v_{qi})$ is a permutation of $[q]$. Let H_q be the subgraph of G_q induced by the vectors in $[q]^{q^2}$ that are concatenations of q permutations. Let H_q^{cyclic} be the subgraph of G_q induced by the vectors in $[q]^{q^2}$ that are concatenations of q cyclic permutations.*

Proposition 6 *For every integer q, $\chi(H_q^{cyclic}) \leq \chi(H_q) \leq q$.*

IV. THE q_ε PARAMETER

A. The study of κ.

Proposition 7

- *For $\varepsilon = \frac{1}{3}$, $q_\varepsilon(\kappa_{20,2}) \leq 3$ whereas $q(\kappa_{20,2}) = 19$, and*
- *For $\varepsilon = \frac{1}{4}$, $q_\varepsilon(\kappa_{7,2}) \leq 4$ whereas $q(\kappa_{7,2}) = 6$, and*
- *For $\varepsilon = \frac{1}{6}$, $q_\varepsilon(\kappa_{6,2}) \leq 6$ whereas $q(\kappa_{6,2}) = 7$.*

B. The study of general graphs G ($k = 2$).

Definition 5 (The graphs $G_{q,\varepsilon}$, $H_{q,\varepsilon}$, and $H_{q,\varepsilon}^{cyclic}$) *For an integer q and an error parameter ε, let $G_{q,\varepsilon}$ be the graph whose vertex set consists of all the vectors of length q^2 over $[q]$, that is, the vectors $u \in [q]^{q^2}$, where two vertices $u = (u_1, \ldots, u_{q^2})$ and $v = (v_1, \ldots, v_{q^2})$ are adjacent if the collection of pairs $\{(u_i, v_i)\}_{i \in [q^2]}$ is at least of size $(1-\varepsilon)q^2$. $H_{q,\varepsilon}$ and $H_{q,\varepsilon}^{cyclic}$ are subgraphs of $G_{q,\varepsilon}$ induced on the vertices of H_q and H_q^{cyclic} respectively.*

The graph family $G_{q,\varepsilon}$ is extremal with respect to the q_ε parameter in the sense given by the following lemma.

Lemma 2 *For every integer q,*
1) *$q_\varepsilon(G_{q,\varepsilon}) \leq q$ and*
2) *$\chi(G) \leq \chi(G_{q,\varepsilon})$ for every graph G with $q_\varepsilon(G) = q$.*

As with the graphs G_q, by Lemma 2, the challenge of obtaining graphs with chromatic number much larger than the q_ε parameter reduces to the study of the chromatic number χ of $G_{q,\varepsilon}$ and its subgraphs.

Proposition 8 *For any q and $\varepsilon = \frac{1}{q}$, $\chi(G_{q,\varepsilon}) \leq O\left(5^q q^{2q}\right)$.*

Proposition 9 *For any q and $\varepsilon = \frac{1}{q}$, $\chi(H_{q,\varepsilon}^{cyclic}) \leq q^2$.*

Proposition 10 *For any q and $\varepsilon = \frac{1}{q}$, $\chi(H_{q,\varepsilon}) \leq q! \cdot q$.*

C. Average error ϵ

In this subsection we allow average error ϵ over both the messages and the edges in E. Here, the decoding error is computed assuming a uniform set of k messages and a uniform decoding edge in E.

Definition 6 (The $q_{\tilde{\varepsilon}}$ parameter) *Let $G = ([n], E)$ be a k-uniform hypergraph on the vertex set $[n]$ and let $\tilde{\varepsilon} > 0$. Let $q_{\tilde{\varepsilon}}(G)$ denote the smallest size q of an alphabet F for which there exist an encoding function $C : F^k \to F^n$ and a decoding function $D : (F \cup \{\perp\})^n \to F^k$ such that*

$$\Pr_{e,m}[D(C_e(m)) = m] \geq 1 - \tilde{\varepsilon},$$

where m is uniformly chosen from F^k, and e is uniformly chosen from E.

Theorem 3 *Let n be any integer. For any prime power p, $q_{\tilde{\varepsilon}}(\kappa_{n,2}) \leq p$ for $\tilde{\varepsilon} = \frac{1}{p+1}$.*

References

[1] S. B. Balaji and P. V. Kumar. Bounds on the rate and minimum distance of codes with availability. In *2017 IEEE International Symposium on Information Theory (ISIT)*, pages 3155–3159, June 2017.

[2] K. A. Bush. Orthogonal arrays of index unity. *Ann. Math. Statistics*, 23:426 – 434, 1952.

[3] B. Segre. Curve razionali normali e k-archi negli spazi finiti. *Ann. Mat. Pura Appl.*, 39(4):357–379, 1955.

[4] F. J. MacWilliams and N. J. A. Sloane. *The Theory of Error-CorrectingCodes*. Amsterdam: North-Holland, 1977.

[5] S. Huntemann. The upper bound of general maximum distance separable codes, 2012. Honours Project.

[6] S. Ball. On sets of vectors of a finite vector space in which every subset of basis size is a basis. *Journal of the European Mathematical Society*, 14(3):733–748, 2012.

[7] M. Gonen, I. Haviv, M. Langberg, and A Sprintson. Minimizing the alphabet size of erasure codes with restricted decoding sets. *Manuscript, available on arXiv.org*, 2020.

[8] A. R. Lehman and E. Lehman. Network coding: does the model need tuning? In *SODA'05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 499–504, 2005.

[9] R. Silverman. A metrization for power-sets with applications to combinatorial analysis. *Canad. J. Math.*, 12:158–176, 1960.