Electronic Colloquium on Computational Complexity, Report No. 338 (2026)

Unshardable Codes*

Prashanth Amireddy† Yevgeniy Dodis‡ Rohan Goyal§ Ethan Mook¶

Madhu Sudan‖ Daniel Wichs**

November 23, 2026

Abstract

We propose a new coding-theoretic problem motivated by censorship resistance. Is it possible to intertwine many pieces of data by encoding them jointly in a way that makes it impossible to remove any single piece without destroying many others as well? In particular, we want to encode k data components $(m_1, \ldots, m_k)$ into a codeword $(c_1, \ldots, c_n)$ with the following properties:

1. *Local recovery*: every data component can be recovered (in the absence of erasures or errors) by reading at most ℓ correct codeword symbols.
2. *Unshardability*: if an adversary erases some arbitrary subset of the codeword symbols, the remaining codeword symbols either uniquely determine all k data components, or there are at least s data components that cannot be determined.

We call such codes (ℓ, s)-*unshardable*. Intuitively, such codes prevent an adversary from selectively censoring individual pieces of data by erasing codeword symbols, without destroying large amounts of other data as collateral damage.

We show that codes whose generator matrix has certain *superregularity* properties are unshardable. Using this connection, we construct linear codes that are (ℓ, s)-unshardable with $s = (1 - \varepsilon)k$ and locality $\ell = O_\varepsilon(1)$. We also show that linear unshardable codes are subject to the lower bound $s \leqslant k - k/\ell + 1$. Finally, we consider a strengthening of local recovery to also require local decodability from some constant fraction of erasures or even errors. With the help of AI tools, we achieve the above notion by showing that any *Locally Decodable Code (LDC)* can be converted into one that is unshardable while preserving locality via a certain careful change of basis over an appropriate extension field.

*To appear in 18th Innovations in Theoretical Computer Science (ITCS) 2027.

†School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported in part by Madhu Sudan's Simons Investigator Award and NSF Award CCF 2152413 and Salil Vadhan's Simons Investigator Award. Email: pamireddy@g.harvard.edu

‡Department of Computer Science, New York University, New York, NY, USA. Supported in part by a gift from the Stellar Foundation. Email: dodis@cs.nyu.edu

§Department of Electrical Engineering & Computer Science, MIT, Cambridge, Massachusetts, USA. Supported by Yael Tauman Kalai's NSF Award 6956054. Email: rohan_g@mit.edu

¶Department of Computer Science, University of Toronto, Toronto, ON, Canada. Email: ethan@ethanmook.com

‖School of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts, USA. Supported in part by a Simons Investigator Award, NSF Award CCF 2152413 and AFOSR award FA9550-25-1-0112. Email: madhu@cs.harvard.edu

**Northeastern University and NTT Research. Research supported by NSF CNS-2349972. Email: wichs@ccs.neu.edu

ISSN 1433-8092

1 Introduction

We introduce a new coding-theoretic property which we call *unshardability*. This property seeks to address the following central question: Can we intertwine many pieces of data by jointly encoding them such that it is impossible to delete any single piece without also deleting many others? This problem is intuitively motivated by censorship resistance: an adversary playing the role of a public data store holds the encoded data and should not be able to selectively censor any individual piece of data without causing significant collateral damage deleting much more.

To formalize this goal we first need a more precise notion of what it means for a piece of data to be deleted. Let $C : \Sigma^k \to \Sigma^n$ be any encoding map We say a subset $T \subseteq [n]$ *determines* $i \in [k]$ (with respect to C) if every pair of messages $\mathbf{m}, \mathbf{m}'$ that disagree at the ith coordinate also disagree on their codewords restricted to T: $C(\mathbf{m})|_T \neq C(\mathbf{m}')|_T$We think of a component m_i of the message as having been deleted if i is not determined by the set of remaining codeword symbols after some erasures. We are now ready to present our main definition.

Definition 1.1 (Unshardable Codes)**.** *Let $C : \Sigma^k \to \Sigma^n$. We say C is* (ℓ, s)-unshardable *if it satisfies the following two properties:*

1. *Local decoding under no errors: For every $i \in [k]$ there is a subset $D \subseteq [n]$ of size at most ℓ that determines i.*
2. *Unshardability: Take any arbitrary pattern of codeword symbol erasures and let $T \subseteq [n]$ denote the set of non-erased symbols. If there is some index $i \in [k]$ that is not determined by T, then the same is true of at least s such indices. Equivalently, T either determines all of $[k]$, or fails to determine all j in a set $S \subseteq [k]$ of size at least s.*

We propose this as a natural coding-theoretic property that has a distinct flavor relative to other notions considered in the literature. For instance, whereas error correction seeks to upper bound the impact of a "small enough" number of errors, unshardability lower bounds the damage of a "large enough" number of erasures. Local decoding says that each data location can be recovered efficiently by only reading a few codeword locations, while unshardability requires the data to be intertwined. Therefore, the two properties are seemingly contradictory and it is apriori unclear whether they can be satisfied simultaneously with good parameters e.g., $\ell = o(k)$ and $s = \Omega(k)$.[1]

Unshardability with Erasures vs. Errors. The unshardability property, as we've defined it, gives a guarantee on the behavior of destroying data in the presence of an arbitrary unbounded set of *erasures*. It is natural to ask what happens if consider an arbitrary set of *errors*. However this is too strong to be achievable. For instance, for any target $i \in [k]$, an adversary could simply overwrite the codeword with the values $(m_j)_{j \neq i}$ (and erase/zero out any remaining symbols). Clearly under these errors, the modified codeword determines all components of $\mathbf{m}$ except the ith. Therefore, we naturally restrict our definition to erasures. Furthermore, referring back to the motivating example of censorship resistance, one also could potentially justify the erasure model using cryptographic authentication techniques (e.g., based on Merkle hashing) to enforce that the adversary can only erase and not introduce new errors.[2]

[1] We note that it is much easier to achieve unshardability if we do not require local decodability. For instance, using our Proposition 3.1, any code whose generating matrix is *superregular* has unshardability k.

[2] We leave this idea as an informal approach to handling errors. It is arguable whether this achieves the intuitive goal, since the trivial attack above still allows the server to censor a single piece of data while enabling all other pieces

Local Decoding with Bounded Errors. It is also natural to ask whether the errorless local decoding property can be strengthened to support local decoding in the presence of a bounded fraction of errors. Here the answer is yes, and we also consider that definition in this work. Concretely we say C is a (δ, ℓ, s)-*unshardable code* if it meets the above definition where we replace the first condition with the following: There exists a randomized local decoding algorithm $\mathcal{L}^{\mathbf{w}}(i)$ that reads at most ℓ locations from $\mathbf{w}$ and outputs m_i with probability at least $3/4$ whenever $\mathbf{w}$ is at most δ-far in relative Hamming distance from $C(\mathbf{m})$. Effectively such a code is a (δ, ℓ)-locally decodable code with an additional unshardability property added on.

General vs. Linear Codes. Our notion of unshardability is well defined for general codes. However, unshardability in the case of nonlinear codes is quite subtle. For example, if C is a nonlinear code and $\mathbf{m}$ is uniformly random, $C(\mathbf{m})|_T$ may information theoretically determine partial information about m_i (such as its most significant bit) without (fully) determining it in the above sense. On the other hand, in the setting of linear codes, if T determines m_i in the sense defined above, then $C(\mathbf{m})|_T$ uniquely determines m_i in the information-theoretic sense and otherwise contains no information about m_i. This makes unshardability much more effective and intuitively appealing for linear codes. Coupled with the efficiency and simplicity of linear codes, in this work all our constructions will be restricted to linear codes.

Limitation of Unshardable Codes. We observe that any unshardable code cannot be systematic. Indeed if $C(\mathbf{m}) = (\mathbf{m}, \mathbf{c}')$ then setting $T = [k] \setminus \{i\}$ for any $i \in [k]$ and erasing the rest of the symbols in the codeword clearly deletes i, but T still determines all other indices. This argument can be generalized in a straightforward manner to yield the following lower bound that demonstrates a trade-off between unshardability and locality that we will prove formally later in Section 3.

Theorem 1.2 (Limits of unshardable codes)**.** *If a linear code $C : \mathbb{F}^k \to \mathbb{F}^n$ is (ℓ, s)-unshardable, then we must have*

$$s \leqslant k - k/\ell + 1.$$

Simple Constructions of Unshardability without Local Decoding. Next we note that achieving unshardability without local decoding is easy. For an illustrative example, consider a Reed-Solomon code $C : \mathbb{F}^k \to \mathbb{F}^n$ under the evaluation encoding map: interpret $\mathbf{m}$ as the evaluations of a degree $k-1$ polynomial $p_{\mathbf{m}}$ on some fixed set of distinct $\alpha_1, \ldots, \alpha_k \in \mathbb{F} \setminus \{0\}$, then $C(\mathbf{m}) = (p_{\mathbf{m}}(\beta_1), \ldots, p_{\mathbf{m}}(\beta_n))$ for another set of distinct $\beta_1, \ldots, \beta_n \in \mathbb{F} \setminus \{0\}$ disjoint from the α_i's.[3] For any $T \subseteq [n]$, the data of $C(\mathbf{m})|_T$ is a set of $|T|$ evaluations of a polynomial of degree $k-1$. If $|T| \geqslant k$ then T determines all $i \in [k]$ by interpolation, but if $|T| < k$ then T doesn't determine any i because $k-1$ evaluations of a degree $k-1$ polynomial cannot determine its value at any other different kth point. Thus this Reed-Solomon code has unshardability $s = k$. This argument works even if $k = n$ which demonstrates that the unshardability of a code is independent of its distance. The fact

to be recovered by clients who ignore the lack of proper authentication. However, it is reasonable in setting where client don't fully trust the server and are only willing to accept authenticated codeword symbols.

[3] This is a change of basis relative to the more standard Reed-Solomon encoding map that treats $\mathbf{m}$ as the coefficients of a polynomial. We use the evaluation encoding here for expository reasons because it makes unshardability easier to see, although unshardability can also be shown for the standard encoding map.

that Reed-Solomon codes have high unshardability is related to the fact that they are maximum distance separable (MDS).[4]

Challenges. Here, we try to point to some of the difficulties in constructing unshardable codes (with local decoding), and show why certain natural extensions do not work, before moving on to our final constructions. First, we note that we can modify the trivial construction from the above paragraph to get some simple tradeoff between unshardability and locality, though with poor parameters. Simply "batch" the message into $\sqrt{k}$-size chunks and apply a Reed-Solomon code block-wise. To decode i, one must only look at the codeword for its batch, and to delete i one must delete the whole batch, hence this is a $(\sqrt{k}, \sqrt{k})$-unshardable code. This trivial argument works for any parameters not just $\sqrt{k}$, but never gives unshardability s better than locality ℓ. Thus, in our actual results below, we will aim at achieving $(o(k), \Omega(k))$-unshardable codes. In fact, our main constructions will achieve optimal or almost optimal trade-offs, nearly matching the limits set by Theorem 1.2.

Another natural attempt may be to try and extend the Reed-Solomon construction by taking more variables (denoted $m > 1$), i.e., a tensor-product of Reed-Solomon codes. For such a design to work, one has to come up with a suitable choice of the message evaluation points (denoted $I \subseteq \mathbb{F}_q^m$)[5] and the codeword evaluation points (denoted $E \subseteq \mathbb{F}_q^m$). While such codes can be shown to have local decodability (e.g. $O(d) = O(\sqrt{k})$ queries when $m = 2$), in fact, even in the presence of a constant fraction of errors, obtaining interesting bounds on unshardability remains unclear (for any choice of I and E). In fact, we establish a negative result for some natural evaluation sets (in particular, when $I = S^2$, where $S \subseteq \mathbb{F}_q$ is a subfield and $E \supseteq \overline{S}^2$). We defer the details to Appendix A.

1.1 Our Results

As our main technical contribution we give a generic transformation from any linear code to one with almost optimal unshardability. The transformation preserves the rate and distance of the underlying code but does require a larger (but still constant) field size. Instantiating our transformation with an asymptotically good code yields the following theorem.

Theorem 1.3 (Explicit unshardable codes)**.** *For every $R < 1$ and $\varepsilon > 0$, there exists $q \leqslant 2^{O_R(1/\varepsilon)}$ and an explicit infinite family of linear codes $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ of rate at least R that are $(O_R(1/\varepsilon), (1-\varepsilon)k)$-unshardable.*

We complement the above results with another set of incomparable results as we discuss in the below remark:

Remark 1.4. *We note that we can also obtain an optimal trade-off between locality and unshardability (i.e., match Theorem 1.2 exactly without the $O(\cdot)$ factor) using simpler direct constructions that work over large fields (see Theorem 6.4). Along the same lines we get unshardable codes with minimal locality 2: we prove the existence of a $(2, \lceil (k+1)/2 \rceil)$-unshardable code and give an explicit construction of a $(2, \lceil k/3 \rceil)$-unshardable code. Furthermore, over sufficiently large fields, we can also*

[4] We later generalize this to show that any code with a certain superregularity condition on its generator matrix has high unshardability. See Proposition 3.1 for more details.

[5] Recall that unlike many other natural coding-theoretic properties, the message space crucially affects unshardability; this is also clear from our earlier discussion about why systematic codes cannot be unshardable.

achieve Maximum-Distance-Separability (MDS) in addition to unshardability (see Theorem 6.7). We refer to Section 6 for these results.

Next we shift our focus to the setting of (δ, ℓ, s)-unshardable codes that have local decoding in the presence of errors. First we show that if we instantiate our generic transformation with a locally correctable code (LCC) we can achieve local decoding in the presence of errors, at the cost of increasing the locality parameter, degrading the tradeoff between locality and unshardability. Plugging in the state of the art LCCs due to Kopparty, Meir, Ron-Zewi, and Saraf [KMRS17] to our transformation yields the following theorem.

Theorem 1.5 (High-rate unshardable LDCs)**.** *For every $R < 1$ and $\varepsilon > 0$, there exist $\delta = \delta(R) > 0$, $q \leqslant 2^{O_R(1/\varepsilon)}$, and an explicit infinite family of linear codes $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ of rate at least R that are $(\delta, \ell, (1-\varepsilon)k)$-unshardable, where $\ell = O_R(\exp(\sqrt{\log n \log\log n}) \cdot \varepsilon^{-2} \cdot \log(1/\varepsilon))$.*

In particular, the query complexity is subpolynomial in the blocklength of the code n. As with Theorem 1.3, we prove a more general transformation of any LCC into an unshardable one (see Theorem 5.1) – in particular, on the other end of the spectrum, we can also achieve relative distance approaching $1/2$ (and error parameter $\delta \to 1/4$) with constant rate.

Remark 1.6. *We note that our construction above also guarantees that the final code is also locally* correctable *with the same query complexity and error parameter, in addition to being locally decodable. We can further make the code family approach the Singleton bound (i.e., any rate $R \in (0,1)$ and relative distance $1-R-o(1)$) by using Theorem 1.3 of [KMRS17] instead, although this increases the required alphabet size. Furthermore, we can obtain a better query complexity of $(\log n)^{O(1)}$ if we allow the local decoder to return "I don't know" (while requiring completeness when there are no errors), by using the recent progress on* Relaxed LDCs *[KM24; CY24].*

Finally, we show that we can slightly modify our generic transformation into a similar transformation that converts a given LDC directly into an unshardable LDC. We instantiate this with the constant-query LDCs of Efremenko [Efr09] yielding the following theorem.

Theorem 1.7 (Constant-query unshardable LDCs)**.** *For every $r \in \mathbb{N}$ and $\varepsilon > 0$, there exist $\delta = \delta(r) > 0$, $q \leqslant 2^{O(1/\varepsilon)}$, and an explicit infinite family of linear codes $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that are $(\delta, \ell, (1-\varepsilon)k)$-unshardable, where $\ell = O(2^r \cdot \varepsilon^{-2} \cdot \log(1/\varepsilon))$ and $n = \exp(\exp(\sqrt[r]{\log k \log\log^{r-1} k}))$.*

In particular, we obtain subexponential length codes with near-maximal unshardability and constant-query local decoding.

Local-Local Unshardability. In a separate direction, we also briefly consider an incomparable variant of unshardability and prove a strong lower bound for that variant. Our main definition (Definition 1.1) considers a global notion of deletion: if i is not determined by reading all of $\mathbf{c}|_T$ then so are many other indices. We can also consider a "local-local" notion of unshardability: if i is not determined by reading any small number of locations in $\mathbf{c}|_T$ then the same is true of many other indices. This local-local unshardability is incomparable to (standard) unshardability because both the premise and conclusion are weaker. Nonetheless we prove a strong lower bound that any (ℓ, s)-local-local unshardable code must have $s \leqslant \ell$, effectively ruling out any non-trivial constructions. See Section 7 for details.

1.2 Technical Overview

The key steps involved in arriving at our constructions of unshardable codes are sketched below:

- **Idea 1: A useful criterion for unshardability.** Let us start with the goal of constructing the generating matrix $G \in \mathbb{F}_q^{k \times n}$ of a linear code with some non-trivial unshardability $s \geqslant 2$. A preliminary observation highlighted in the above subsection is that G cannot contain the $k \times k$ identity matrix (denoted I_k) as a submatrix. On the contrary, if G *does* contain I_k as a submatrix, there exists a subset of columns $T \subseteq [n]$ of size $k-1$ such that the corresponding submatrix G_T has a vector of Hamming weight 1 in its (left) kernel. This suggests that we want to have vectors of *large* Hamming weight in the kernel of the submatrix G_T (whenever it is non-empty). Indeed if we have $\mathbf{m}^\top G_T = \mathbf{0}^\top$ for some message $\mathbf{m} \in \mathbb{F}_q^k$, then it is easy to see that every coordinate $i \in [k]$ such that $m_i \neq 0$ is *not* determined by T. This is exactly the condition we will exhibit in our constructions to show that a given code has high unshardability (see Proposition 3.1).

- **Idea 2: Sparse rebasing.** Now, it remains to design the code G with the above property while also being locally decodable (under no erasures or errors, to begin with). Equivalently, we want to ensure that each standard basis vector $\mathbf{1}_i$ can be spanned by a small number of columns of G. One natural way to achieve this (while avoiding I_k as a submatrix) is by taking

$$G = \begin{bmatrix} A & A + I_k \end{bmatrix},$$

  for some $A \in \mathbb{F}_q^{k \times k}$. Viewed differently, the above matrix G is obtained by applying some simple elementary column operations to manufacture locality (hence the name *sparse rebasing*).

  However, we have to choose A such that the linear-algebraic criterion from Idea 1 applies to G. It turns out that one can further show that if all the "large" submatrices of G are full rank (this property is known as *superregularity* (see Definition 2.13)), then the linear-algebraic criterion indeed applies, giving the desired bounds on unshardability. Using known (relatively simple) constructions of superregular matrices for the matrix A, and a careful argument to "transfer" the superregularity from A to G, we thus get explicit constructions of codes with locality 2 and unshardability $\Omega(k)$ over sufficiently large fields (see Section 6).

- **Idea 3: Sparse rebasing w.r.t. extension fields.** The challenge remains to obtain unshardable codes over a constant-alphabet (i.e., prove Theorem 1.3). For this we abstract out the sparse rebasing notion above more generally such that we need not have to analyze the superregularity of a base code anymore but instead only take advantage of its relative distance. Let $A \in \mathbb{F}_q^{k \times n}$ be a code and $B \in \mathbb{F}_q^{k \times k}$ be an invertible matrix such that each column of B is a "sparse" linear combination of columns of A. Then, we essentially have access to a locally decodable code (under no errors) given by the map

$$\mathbf{m}^\top B \mapsto \mathbf{m}^\top A.$$

  The linear-algebraic criterion from Item 1 (under this basis change) translates to saying that we want to exhibit a message $\mathbf{m} \in \mathbb{F}_q^k$ such that $\langle \mathbf{m}, \mathbf{b}_i \rangle \neq 0$ for many $i \in [k]$, while at the same time having $\langle \mathbf{m}, \mathbf{a}_i \rangle = 0$ for all $i \in T$, where $\mathbf{b}_1, \ldots, \mathbf{b}_n$ and $\mathbf{a}_1, \ldots, \mathbf{a}_n$ denote the columns of B and A respectively. Taking $\mathbf{m} \neq \mathbf{0}$ to be an arbitrary vector in the kernel of A_T (if

one exists), we observe that $\langle \mathbf{m}, \mathbf{a}_i \rangle \neq 0$ for a constant fraction of the coordinates if A has a constant relative distance.

The goal now is to design the sparse column combinations of A so that these non-zeroes translate to many of the inner products $\langle \mathbf{m}, \mathbf{a}_i \rangle$ being non-zero. In turn, we want an edge-labeled (with field elements, denoting the sparse linear combinations) $n \times k$ bipartite graph H such that the many non-zeroes on the left are "well-propagated" to the right (the precise notion we need here is *vertex-expansion* – see Claim 4.5). This is because $\langle \mathbf{m}, \mathbf{b}_i \rangle$ (for some $i \in [k]$) is a corresponding sparse linear combination of the inner products corresponding to the set of left-neighbors of i. Now, without loss of generality, suppose we have

$$\mathbf{b}_i = \sum_{i=1}^{\ell} \lambda_i \cdot \mathbf{a}_i,$$

for some small sparsity $\ell \in \mathbb{N}$ and $\lambda_i \in \mathbb{F}_q$. In turn, we get

$$\langle \mathbf{m}, \mathbf{b}_i \rangle = \sum_{i=1}^{\ell} \lambda_i \cdot \langle \mathbf{m}, \mathbf{a}_i \rangle \,. \tag{1}$$

Had it turned out that each inner product term of (1) lies in a small subfield of $\mathbb{F}_q$, and λ_i's were say "generic" field elements, then one can guarantee that if any of the inner products on the RHS of (1) is non-zero (which happens for many $i \in [k]$ due to high distance of the code A and good vertex expansion of H), then the LHS of (1) is also non-zero (since some λ_i "survives"). Stated slightly differently and more explicitly, it suffices to take the coefficients λ_i to be $\mathbb{F}_q$-linearly independent elements from a field extension $\mathbb{F}_{q^b}$ for a sufficiently large $b \in \mathbb{N}$. Thus, the basis change matrix B and our final code (i.e., $G = B^{-1}A$) will be over an extension field and have the desired sparsity/locality (governed by the maximum degree of H) and unshardability (governed by the expansion of H and distance of the code A) – see Theorem 4.4 for the final parameters (for $q = 2$).

We note that one has to prove that matrix B we get at the end of the above rebasing is invertible. This is done by carefully setting the values of some of the coefficients $\lambda_i \in \mathbb{F}_{q^b}$ of (1) in a special manner (see Claim 4.6).

We point out here that the above construction closely resembles constructions of graph-based codes of Gallager [Gal62] (and many more later developments in coding theory; see e.g. [Tan81; ABNNR92; AEL95; SS96]) which use similar "belief-propagation" arguments. However, it is worth noting that in our context, the graph (i.e., H) interacts with the generator matrix of a code (i.e., A), whereas in expander code constructions, (the parity-check matrix of) a code is defined based on a graph. Another key innovation is in using a field extension to reason about the number of right-vertices that evaluate to a non-zero value given a non-zero assignment (from the base field) to the left-vertices.

- **Idea 4: Local decodability under errors.** Our final result is for constructing unshardable codes that are locally decodable in the presence of constant fraction of erasures or errors. The key ideas from the above construction (Idea 3) are sufficiently robust to almost immediately give this result. More specifically, we take the base code A above to be one that is locally correctable from errors. Then, since the columns of B are sparse combinations of columns of

A, the local correctability of A is enough to get local decodability for G up to the same error parameter. We instantiate this reduction with existing LCCs and LDCs to obtain unshardable codes with this stronger local decodability feature (see Theorem 5.1).

We remark that the source of Ideas 3 and 4 was AI. In particular, while we can construct explicit unshardable codes using connection to superregular matrices (and other results from Section 6), it was not clear to us how to extend those constructions to the setting of local decodability under errors (i.e., Theorem 1.5), until the notion of a sparse rebasing over an extension field was brought into the picture.

Organization. We present the required coding-theoretic preliminaries in Section 2. Then in Section 3 we provide a linear-algebraic criterion for high unshardability that will be used throughout the paper. In Section 4, we set up the main frameworks to construct unshardable codes and prove Theorem 1.3 (i.e., unshardable codes that only locally decodable under no errors). In Section 5, we then give constructions that also work under a constant fraction of errors (i.e., Theorem 1.5 and Theorem 1.7). We then give multiple improvements and incomparable constructions (for the no-error version) in Section 6 based on superregularity criteria discussed above. Finally in Section 7, we discuss close variants of our main notion of unshardability.

2 Preliminaries

Let $\mathbb{F} = \mathbb{F}_q$ denote the finite field of size q (where q is a prime power) and $\mathbb{F}^\times = \mathbb{F} \setminus \{0\}$ denote its multiplicative group. For positive integers k and $i \in [k]$, let $\mathbf{1}_i \in \mathbb{F}^k$ denote the i-th standard basis vector. We use $\mathrm{span}(\cdot)$ (or $\mathrm{span}_{\mathbb{F}}(\cdot)$ to be more explicit when dealing with subfields/extension fields) and $\dim(\cdot)$ to denote the $\mathbb{F}$-linear span and dimension of a subset of vectors in $\mathbb{F}^k$. We use $\mathrm{rank}(\cdot)$ to denote the rank of a matrix, and $\det(\cdot)$ the determinant of a square matrix.

For a finite alphabet Σ, $D \subseteq [n]$, and $\mathbf{u} \in \Sigma^n$, let $\mathbf{u}|_D \in \Sigma^D$ denote the vector obtained by restricting $\mathbf{u}$ to the coordinates of D. For $\mathbf{a}, \mathbf{b} \in \Sigma^n$, let $\Delta(\mathbf{a}, \mathbf{b})$ denote their Hamming distance and $\delta(\mathbf{a}, \mathbf{b}) := \Delta(\mathbf{a}, \mathbf{b})/n$ their relative Hamming distance. Let $|\mathbf{a}| := \Delta(\mathbf{a}, \mathbf{0})$ denote the Hamming weight of $\mathbf{a}$. We say that a code $C : \Sigma^k \to \Sigma^n$ has *rate* k/n and *distance* $\min_{\mathbf{m}_1 \neq \mathbf{m}_2 \in \Sigma^k} \Delta(C(\mathbf{m}_1), C(\mathbf{m}_2))$. We refer to the distance divided by the blocklength n as the *relative distance* of the code. It is desirable to have both rate and relative distance be constants bounded away from zero – we call such families of codes *asymptotically good*. For a matrix $M \in \Sigma^{m \times n}$ and $A \subseteq [m]$ and $B \subseteq [n]$, we let $M_{A,B}$ denote the submatrix of M with rows indexed by A and columns indexed by B. We let $M_{\cdot,B} := M_{[m],B}$ and $M_{A,\cdot} := M_{A,[n]}$. We say that C is a *linear code* with *generating matrix* $G \in \mathbb{F}^{k \times n}$ if $\Sigma = \mathbb{F}$, $\mathrm{rank}(G) = k$, and $C(\mathbf{m}) := \mathbf{m}^\top G$. We say that G (and C) is *systematic* if $G_{\cdot,k} = I$ (i.e., the identity matrix).

2.1 Main Notions: Locality and Unshardability

Now, we define the notions of local decodability and unshardability of codes more formally.

Definition 2.1 (Locality under no errors)**.** *A code $C : \Sigma^k \to \Sigma^n$ is said to be* ℓ-locally decodable *(or "has* locality ℓ*") if for all indices $i \in [k]$, there exists a set $D \subseteq [n]$ of size at most ℓ such that*

$$\forall \mathbf{m}, \mathbf{m}' \in \Sigma^k : C(\mathbf{m})|_D = C(\mathbf{m}')|_D \implies m_i = m'_i. \tag{2}$$

We refer to the condition (2) *as saying* D determines i *w.r.t.* C.

The following standard observation gives a linear-algebraic characterization for the notion of determining message symbols for linear codes.

Observation 2.2. *Let the columns of* $G \in \mathbb{F}^{k \times n}$ *be denoted by* $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n \in \mathbb{F}^k$. *For* $i \in [k]$ *and* $D \subseteq [n]$, D *determines* i *w.r.t* G *if and only if* $\mathbf{1}_i \in \operatorname{span}\{\mathbf{v}_j : j \in D\}$.

Definition 2.3 (Locality under errors)**.** *A code* $C : \Sigma^k \to \Sigma^n$ *is said to be* (δ, ℓ)-locally decodable *(or "has* locality ℓ up to error δ"*) if there exists a randomized algorithm* $\mathcal{L}$, *which when given as input an index* $i \in [k]$ *and oracle access to a word* $\mathbf{w} \in \mathbb{F}^n$ *such that* $\delta(\mathbf{w}, C(\mathbf{m})) \leqslant \delta$ *for some message* $\mathbf{m} \in \mathbb{F}^k$, *makes at most* ℓ *oracle queries and outputs* m_i *with probability at least* $3/4$. *That is*[6]*:*

$$\Pr[\mathcal{L}^{\mathbf{w}}(i) = m_i] \geqslant 3/4.$$

Remark 2.4. *A similar notion of* (δ, ℓ)-local correctability *corresponds to a* ℓ*-query algorithm* $\mathcal{L}$ *such that for all* codeword indices $i \in [n]$ *and received words* $\mathbf{w}$ *such that* $\delta(\mathbf{w}, \mathbf{c}) \leqslant \delta$ *for some codeword* $\mathbf{c} \in \operatorname{Im}(C)$, *it holds that*

$$\Pr[\mathcal{L}^{\mathbf{w}}(i) = c_i] \geqslant 3/4.$$

For systematic codes, local correctability implies local decodability.

Unless mentioned otherwise, we reserve the phrases *Locally Decodable Codes (LDCs)* and *Locally Correctable Codes (LCCs)* for when the error parameter $\delta = \Omega(1)$ and $\ell = o(k)$. Although our primary interest is local decodable codes (LDCs) in this paper, some of our constructions make use of locally correctable codes (LCCs). In particular, we will use the following constructions of LDCs and LCCs in a black-box way for some of our main results.

Theorem 2.5 ([KMRS17, Theorem 1.1])**.** *The following statements hold true for infinitely many* k. *For every* $\varepsilon > 0$, *there exist* $\delta > 0$ *and explicit linear codes* $C : \mathbb{F}_2^k \to \mathbb{F}_2^n$ *of rate at least* $1 - \varepsilon$ *and relative distance at least* δ *that are* $(\delta, \exp(\sqrt{\log n \log\log n}))$*-locally correctable.*[7] *Moreover, the above codes are systematic linear codes, so they are also locally* decodable *with the same parameters.*

Theorem 2.6 ([Efr09, Theorem 3.6])**.** *For every* $r \in \mathbb{N}$, *there exists* $\delta = \delta(r) > 0$ *and infinitely many* $k \in \mathbb{N}$ *such that there are explicit linear codes* $C : \mathbb{F}_2^k \to \mathbb{F}_2^n$ *of relative distance at least* δ *that are* $(\delta, O(2^r))$*-locally decodable with* $n = \exp(\exp(\sqrt[r]{\log k \log\log^{r-1} k}))$.

We now formally define the notion of unshardability that we introduce in this paper.

Definition 2.7 (Unshardability)**.** *A code* $C : \Sigma^k \to \Sigma^n$ *is said to be* s-unshardable *(or "has* unshardability s*") if the following holds for all sets* $T \subseteq [n]$*: If there exists an index* $i \in [k]$ *that is* not *determined by* T, *then there are at least* s *such indices.*

We note that unshardability is most interesting in the context of linear codes (as also indicated by our results) since Observation 2.2 implies that if some message index $i \in [k]$ is not determined, then it is *completely undetermined*, as in each alphabet symbol is equally likely to be the actual message symbol m_i for a random message.

Locality and unshardability are opposing features of a code. We combine these two definitions to arrive at the main focus of this paper.

[6] Note that the success probability can be amplified to $1 - 1/2^r$ by blowing up the query complexity by factor of $O(r)$.

[7] $\exp(x)$ stands for $2^{O(x)}$.

Definition 2.8 ((δ, ℓ, s)-unshardable)**.** *A code $C : \Sigma^k \to \Sigma^n$ is said to be* (δ, ℓ, s)-unshardable *if it is (δ, ℓ)-locally decodable and s-unshardable.*

Abusing notation slightly, we refer to a weaker notion of (ℓ, s)-unshardable to mean that the code is ℓ-locally decodable and s-unshardable.

2.2 MDS Codes and Superregularity

We now define *Maximum Distance Separable (MDS)* codes, which are codes that achieve the best rate-distance trade-off.

Definition 2.9 (MDS code)**.** *We say that $G : \mathbb{F}^k \to \mathbb{F}^n$ is a (linear) MDS code if $d = n - k + 1$, where d is the distance of the code G. Equivalently, every collection of k columns of G are linearly independent.*

It is easy to see that for linear codes, MDS property is independent under a basis change, i.e., for all invertible matrices $A \in \mathbb{F}^{k \times k}$, G is MDS if and only if AG is MDS. Moreover, MDS codes are easy to construct explicitly, where throughout the paper, we say that a code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ is *explicit* if its generating matrix can be deterministically constructed in time $\mathrm{poly}(k, n, \log q)$ given as input k, n, and q. In fact, it is well-known that the *Reed-Solomon code* is MDS:

Definition 2.10 (Reed-Solomon code)**.** *Let $k, n \geqslant 1$ be integers and $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$ be pairwise distinct elements. Then the corresponding Reed-Solomon code $G : \mathbb{F}^k \to \mathbb{F}^n$ is defined by*

$$G(m_1, \ldots, m_k)_i = \sum_{j=1}^{k} m_{j-1} \cdot \alpha_i^{j-1},$$

for $i \in [n]$. In words, the messages correspond to polynomials of degree at most $k-1$ (in particular, G is said to be a degree-$(k-1)$ *RS code) and the codewords correspond to their evaluations at n field elements.*

Fact 2.11 (RS code is MDS)**.** *For every choice of integers $n \geqslant k \geqslant 1$, field $\mathbb{F}$ with pairwise distinct elements $\alpha_1, \ldots, \alpha_n \in \mathbb{F}$, the corresponding degree-$(k-1)$ Reed-Solomon code is MDS.*

The following (closely related) fact (often called the Schwartz-Zippel lemma) bounds the fraction of zeroes of low-degree polynomials, and will be useful for probabilistic constructions.

Fact 2.12 (Polynomial identity lemma, [DL78; Sch80; Zip79])**.** *Let $\mathbb{F}$ be an arbitrary field, $d \geqslant 0$, and $S \subseteq \mathbb{F}$ be non-empty. If $P \in \mathbb{F}[X_1, \ldots, X_n]$ is a non-zero polynomial of degree at most d, then*

$$\Pr_{\mathbf{a} \in S^n}[P(\mathbf{a}) = 0] \leqslant \frac{d}{|S|}.$$

We now define the notion of superregular matrices, which will be crucial for some of our constructions of unshardable codes.

Definition 2.13 (Superregular matrices)**.** *For integers $a \leqslant k$ and $b \leqslant n$, we say that a matrix $A \in \mathbb{F}^{k \times n}$ is* (a, b)-superregular *if every $a \times b$ submatrix of A is full rank.*

We say that A is (full) superregular *if it is (a, b)-superregular for all $a \leqslant k$ and $b \leqslant n$.*

It is easy to give an existential construction of superregular matrices using the probabilistic method and the polynomial identity lemma (Fact 2.12). The following lemma shows how to construct superregular matrices *explicitly* using MDS codes (i.e., using Fact 2.11).

Lemma 2.14 (Explicit superregular matrices, e.g. [RL89])**.** *Let $n \geqslant k$ be positive integers, $q \geqslant n$ be a prime power, and*

$$G = \begin{bmatrix} I & A \end{bmatrix},$$

where $I \in \mathbb{F}^{k\times k}$ is the identity matrix and $A \in \mathbb{F}^{k\times(n-k)}$. Then G is MDS if and only if A is superregular.

Proof. Let $S \subseteq [n]$ be a set of k columns of G such that $S_1 \subseteq [k]$ and $S_2 \subseteq [k]$ are the indices of S corresponding to the first half and the second half of the columns respectively (i.e., $S = S_1 \cup (k+S_2)$). By rearranging rows and columns, we notice that $G_{\cdot,S} = \begin{bmatrix} I_{\cdot,S_1} & A_{\cdot,S_2} \end{bmatrix}$ can be transformed into a matrix of the form

$$\begin{bmatrix} I & A_{S_1,S_2} \\ 0 & A_{\overline{S_1},S_2} \end{bmatrix},$$

where $\overline{S_1} = [k] \setminus S_1$. Thus, we have $\det(G_{\cdot,S}) = \det(A_{\overline{S_1},S_2})$. Observing that $A_{\overline{S_1},S_2}$ is a submatrix of A, we get the desired equivalence that G is MDS if and only if A is superregular. □

We need the notion of subfields/extension fields and linear independence.

Definition 2.15. *For positive integers $\ell \leqslant b$ and $q = 2^b$, we say that $\lambda_1, \ldots, \lambda_\ell \in \mathbb{F}_q$ are* $\mathbb{F}_2$-linearly independent *if for all non-zero Boolean vectors $\mathbf{b} \in \mathbb{F}_2^\ell$, it holds that*

$$\sum_{i=1}^{\ell} b_i \cdot \lambda_i \neq 0.$$

Equivalently, the vectors $(\rho(i))_{i\in[\ell]} \in \mathbb{F}_2^b$ are linearly independent, where $\rho : \mathbb{F}_q \to \mathbb{F}_2^b$ is an arbitrary $\mathbb{F}_2$-linear bijection[8]*. In particular, the field elements $(\rho^{-1}(\mathbf{1}_i))_{i\in[b]}$ are $\mathbb{F}_2$-linearly independent.*

2.3 Bipartite Expanders

We will need explicit bipartite graphs with the following expansion properties.

Lemma 2.16 (Bipartite expanders)**.** *For every $R < 1$ and $\delta, \varepsilon > 0$, there exist infinitely many $k \in N$ and $k \geqslant Rn$ such that there are explicit bipartite graphs $H : [n] \times [k] \to \{0,1\}$ of degree at most $d = O(1/(R\delta\varepsilon))$ such that for all $L \subseteq [n]$ of size at least δn, the size of the neighborhood of L is at least $(1-\varepsilon)k$. Moreover, there exists a matching of size k in H.*

*We refer to H satisfying the above expansion condition as a $(\geqslant \delta n, (1-\varepsilon)k)$-*vertex expander.

Proof. We take H to be a subgraph of a suitable bipartite *spectral expander*. In particular, we let $d = O(1/(R\delta\varepsilon))$ be a sufficiently large integers and choose $G : [n] \times [n] \to \mathbb{N}$ be a d-regular bipartite *Ramanujan* multigraph: that is, the second eigenvalue of its adjacency matrix is

$$\lambda \leqslant 2\sqrt{d-1}. \tag{3}$$

[8] i.e., $\rho(\lambda + \lambda') = \rho(\lambda) + \rho(\lambda')$ for all $\lambda, \lambda' \in \mathbb{F}_q$

Explicit constructions of such graphs are known to exist from the works of [MSS18; Coh16].

Now, we define the graph H to be the subgraph of G obtained by removing some arbitrary $n-k$ right vertices and their incident edges in G. Furthermore, we also replace all repeated edges with a single edge so that H is a simple graph.

Let $S \subseteq [n]$ be an arbitrary subset of size at least δn. Suppose, towards a contradiction, that the set of neighbors of S in H is of size less than $(1-\varepsilon)k$. Then, we have a subset $T \subseteq [n]$ of size greater than εk such that there are no edges between S and T in H (and hence in G). In other words, $e_G(S,T) = 0$. Applying the expander-mixing lemma (see e.g. [Vad12, Lemma 4.15]) for bipartite graphs, we obtain

$$\left| \frac{e_G(S,T)}{n \cdot d} - \delta \cdot \frac{|T|}{n} \right| \leqslant \frac{\lambda}{d} \sqrt{\delta \cdot \frac{|T|}{n}},$$

which in turns implies that $\lambda \geqslant d\sqrt{R\delta\varepsilon}$. Combined with (3), we obtain

$$d \leqslant 4/(R\delta\varepsilon),$$

which is a contradiction by taking $d := \lceil 10/(R\delta\varepsilon) \rceil$. Hence, we conclude that H is a $(\geqslant \delta n, (1-\varepsilon)k)$-vertex expander.

Finally, we show that there is a matching of size k in H. For this, it suffices to show that there is a perfect matching in G. We prove this using Hall's theorem. Let $S \subseteq [n]$ be arbitrary and $T \subseteq [n]$ be its neighborhood. The number of edges between S and T is equal to $d \cdot |S|$ since the left-degree is d. On the other hand, since the right-degree is also d, the number of edges between S and T is at most $d \cdot |T|$. Put together, this gives $|T| \geqslant |S|$. Hence, H contains a matching of size k . □

3 Characterizing Unshardability

We prove the following proposition, which establishes a clean linear-algebraic proxy for unshardability that will be used for our constructions.

Proposition 3.1 (Characterizing unshardability)**.** *Let $G : \mathbb{F}^k \to \mathbb{F}^n$ be a linear code and $s \leqslant k$ be a positive integer. Consider the following statements.*

1. *G is s-unshardable.*
2. *For all subsets $T \subseteq [n]$, if $\mathrm{rank}(G_{\cdot,T}) < k$, then there exists $\mathbf{m} \in \mathbb{F}^k$ such that $\mathbf{m}^\top G_{\cdot,T} = \mathbf{0}^\top$ and $|\mathbf{m}| \geqslant s$.*
3. *G is $(s-1, k-1)$-superregular.*
4. *For every two messages $\mathbf{m}_1 \neq \mathbf{m}_2 \in \mathbb{F}^k$, it holds that*

$$\Delta(\mathbf{m}_1, \mathbf{m}_2) < s \implies \Delta(G(\mathbf{m}_1), G(\mathbf{m}_2)) > n - k + 1.$$

We have that Item 4 $\implies$ Item 3 $\implies$ Item 2 $\implies$ Item 1.

Remark 3.2. *We note that while we never use Item 4 in our main proofs, it has an interesting interpretation that close-by messages are encoded to far-away codewords. It also suggests that unshardable codes are "close" to being MDS in a sense*[9].

Proof of Proposition 3.1. (*Item 2* $\implies$ *Item 1*) Let $T \subseteq [n]$ be arbitrary. We have two cases:

- **Case 1:** $\operatorname{rank}(G_{\cdot,T}) = k$**:** Equivalently, the columns of G indexed by T span $\mathbb{F}^k$. In particular, we have for all $i \in [k]$, $\mathbf{1}_i \in \operatorname{span}\{\mathbf{g}_j : j \in T\}$ so T determines i (by Observation 2.2), and we are done.
- **Case 2:** $\operatorname{rank}(G_{\cdot,T}) < k$**:** Let $\mathbf{m} \in \mathbb{F}^k$ be the vector guaranteed by Item 2: i.e., we have $|\mathbf{m}| \geqslant s$ and

$$\mathbf{m}^\top G_{\cdot,T} = \mathbf{0}^\top. \tag{4}$$

  We will now argue for all $i \in [k]$, $m_i \neq 0$ implies that T doesn't determine i. To prove the contrapositive, suppose T determines i. Denoting the j-th column of G by $\mathbf{g}_j$, we then have

$$\mathbf{1}_i \in \operatorname{span}\{\mathbf{g}_j : j \in T\}.$$

  Combining the above fact with (4), we get

$$m_i = \mathbf{m}^\top \mathbf{1}_i = 0.$$

  Therefore, we conclude that the number of indices not determined by T is at least $|\mathbf{m}| \geqslant s$.

(*Item 3* $\implies$ *Item 2*) Suppose G is $(s-1, k-1)$-superregular and $T \subseteq [n]$ be an arbitrary subset of columns of G such that $\operatorname{rank}(G_{\cdot,T}) < k$. We first note that we can assume that the columns indexed by T are linearly independent since the condition $\mathbf{m}^\top G_{\cdot,T} = \mathbf{0}^\top$ only depends on the column span of $G_{\cdot,T}$. That is, $\operatorname{rank}(G_{\cdot,T}) = |T| \leqslant k-1$. We can further assume that $|T| = k-1$ since adding additional indices to T only makes it harder to satisfy $\mathbf{m}^\top G_{\cdot,T} = \mathbf{0}^\top$.

Now, since the matrix $G_{\cdot,T}$ has rank less than its number of rows k, by the rank-nullity theorem, there exists a non-zero message $\mathbf{m} \in \mathbb{F}^k$ such that

$$\mathbf{m}^\top G_{\cdot,T} = \mathbf{0}^\top. \tag{5}$$

Since every non-empty subset of $s-1$ rows of $G_{\cdot,T}$ are linearly independent (using $(s-1, k-1)$-superregularity of G), we must have that the Hamming weight of $\mathbf{m}$ be least s, proving Item 2.

(*Item 4* $\implies$ *Item 3*) We first note that because G is linear, Item 4 is equivalent to the following:

$$\forall \mathbf{m} \in \mathbb{F}^k : 0 < |\mathbf{m}| < s \implies |G(\mathbf{m})| > n - k + 1. \tag{6}$$

We will show for every subset of rows $R \subseteq [k]$ of size $s-1$ and subset of columns $C \subseteq [n]$ of size $k-1$, that $G_{R,C}$ has full row rank i.e., $s-1$. Let $\boldsymbol{\alpha} \in \mathbb{F}^R$ be an arbitrary non-zero vector. Applying (6) for $\mathbf{m}$ given by

$$m_i = \begin{cases} \alpha_i, & \text{if } i \in R, \\ 0, & \text{if } i \in [k] \setminus R, \end{cases}$$

[9] This is only a qualitative statement since unshardability (Item 1) does not imply Item 4 in general due to the following simple counter example: Let $G : \mathbb{F}^k \to \mathbb{F}^n$ be an s-unshardable code. We observe that the code $G' : \mathbb{F}^k \to \mathbb{F}^{n+k-1}$ obtained by appending $k-1$ columns of all zeroes to G continues to be s-unshardable. However, G' is clearly not $(s-1, k-1)$-superregular for any $s \geqslant 2$.

we have

$$|G(\mathbf{m})| = |\boldsymbol{\alpha}^\top G_{R,\cdot}| > n - k + 1.$$

Note that $\boldsymbol{\alpha}^\top G_{R,C}$ is a subvector of $\boldsymbol{\alpha}^\top G_{R,\cdot}$ obtained by restricting to the indices of C. Thus, we have the following lower bound for the number of non-zero entries of $\boldsymbol{\alpha}^\top G_{R,C}$:

$$|\boldsymbol{\alpha}^\top G_{R,C}| \geqslant |\boldsymbol{\alpha}^\top G_{R,\cdot}| - (n - |C|) > (n - k + 1) - (n - (k - 1)) = 0.$$

Hence, every non-zero linear combination of the rows of $G_{R,C}$ produces a non-zero vector. In other words, $G_{R,C}$ has full row rank. This finishes the proof of the key proposition needed to construct unshardable codes (Proposition 3.1). □

We now give a formal justification for the opposing nature of locality and unshardability alluded to in the introduction:

Theorem 1.2 (Limits of unshardable codes)**.** *If a linear code $C : \mathbb{F}^k \to \mathbb{F}^n$ is (ℓ, s)-unshardable, then we must have*

$$s \leqslant k - k/\ell + 1.$$

Proof. For each $i \in [k]$, let $D_i \subseteq [n]$ be the subset of size at most ℓ guaranteed by local decoding that determines i. Let m be the largest integer such that $|\cup_{i=1}^m D_i| < k$. Clearly, we have $m + 1 \geqslant k/\ell$. Then $T := \cup_{i=1}^m D_i$ cannot determine all indices in $[k]$ (since the dimension of $\{v_i : i \in T\}$, where $v_1, \ldots, v_n$ denote the columns of G, is less than k, it cannot span all the standard basis vectors), so it doesn't determine at least s indices of the message. However, it does determine the indices $1, 2, \ldots, m$. Hence, we obtain $m \leqslant k - s$, which implies that $s \leqslant k - k/\ell + 1$. □

4 Explicit Unshardable Codes

Building on the intuition laid out in Section 1.2, the key idea in proving our main theorem (Theorem 1.3) is to find a basis change so that any given linear code with some reasonable nice properties (such as having good distance as we will see) can be turned into a code with high unshardability (over an extension field). However, we also need the basis change to be such that it keeps the code locally decodable. The following notion of a *sparse rebasing* (Definition 4.1) encapsulates these two requirements.

Definition 4.1 (Sparse rebasing)**.** *Let $q = 2^b$ for $b \in \mathbb{N}$. We say that an invertible matrix $B \in \mathbb{F}_q^{k \times k}$ is a* (ℓ_1, s)-sparse rebasing *of $A \in \mathbb{F}_2^{k \times n}$ if the following conditions hold:*

1. *(Sparsity) Each column of B can be expressed as an $\mathbb{F}_q$-linear combination of some ℓ_1 columns of A, and*
2. *(Distance for Boolean inputs) For all $\mathbf{z} \in \mathbb{F}_2^k \setminus \{\mathbf{0}\}$, we have $|\mathbf{z}^\top B| \geqslant s$.*

We highlight the fact that the above notion involves both the fields $\mathbb{F}_2$ and $\mathbb{F}_q \supseteq \mathbb{F}_2$. In particular, Item 1 (which will be used to derive local decodability) involves looking at $\mathbb{F}_q$-linear spans of the columns of A, whereas Item 2 (which will be used to derive unshardability) involves looking at $\mathbb{F}_2$-linear spans of the rows of B (as indicated by the test vectors $\mathbf{z}$ being Boolean). Indeed if the matrix B were to be Boolean, Item 2 cannot hold for invertible matrices. Going to the extension field $\mathbb{F}_q$ from $\mathbb{F}_2$ gives us the additional flexibility needed for our construction.

4.1 A Generic Transformation

We show that any code with a good distance can be transformed into one that has good unshardability and locality (see Theorem 4.4). We first show that a sparse rebasing immediately yields unshardability:

Lemma 4.2. *If an invertible matrix $B \in \mathbb{F}_q^{k\times n}$ is a (ℓ_1, s)-sparse rebasing of $A \in \mathbb{F}_2^{k\times n}$, then $G = B^{-1}A$ is s-unshardable.*

Proof. Using Proposition 3.1, it suffices to show that for all subsets $T \subseteq [n]$ such that $\mathrm{rank}(G_{\cdot,T}) < k$, there exists a vector $\mathbf{m} \in \mathbb{F}_q^k$ such that $|\mathbf{m}| \geqslant s$ and $\mathbf{m}^\top G_{\cdot,T} = \mathbf{0}^\top$. For this, we first note that

$$G_{\cdot,T} = B^{-1} \cdot A_{\cdot,T},$$

so $\mathrm{rank}_{\mathbb{F}_q}(A_{\cdot,T}) = \mathrm{rank}_{\mathbb{F}_q}(B \cdot G_{\cdot,T}) < k$ as B is invertible. Recall that A has entries from $\mathbb{F}_2$. We now argue that $\mathrm{rank}_{\mathbb{F}_2}(A_{\cdot,T}) < k$. Otherwise, there exists $T' \subseteq T$ of size k such that $\mathrm{rank}_{\mathbb{F}_2}(A_{\cdot,T'}) = k$, which means that its determinant is $\det_{\mathbb{F}_2}(A_{\cdot,T'}) \neq 0$. In particular, we obtain $\det_{\mathbb{F}_2}(A_{\cdot,T'}) = 1$ and thus, $\det_{\mathbb{F}_q}(A_{\cdot,T'}) = 1 \in \mathbb{F}_q$. Therefore, $A_{\cdot,T'}$ (and hence $A_{\cdot,T}$) is of rank equal to k, which is a contradiction. Therefore, $\mathrm{rank}_{\mathbb{F}_2}(A_{\cdot,T}) < k$. By rank-nullity, there exists $\mathbf{z} \in \mathbb{F}_2^k \setminus \{\mathbf{0}\}$ such that

$$\mathbf{z}^\top A_{\cdot,T} = \mathbf{0}^\top. \tag{7}$$

Let $\mathbf{m} := B^\top \mathbf{z} \in \mathbb{F}_q^k$. From Definition 4.1, we have

$$|\mathbf{m}| = |\mathbf{z}^\top B| \geqslant s.$$

Using (7), we also have

$$\mathbf{m}^\top G_{\cdot,T} = (\mathbf{z}^\top B) \cdot G_{\cdot,T} = \mathbf{z}^\top A_{\cdot,T} = \mathbf{0}^\top.$$

Hence, G is s-unshardable. □

We now claim that one can always find a sparse rebasing if A has good distance:

Lemma 4.3. *Let $A \in \mathbb{F}_2^k \to \mathbb{F}_2^n$ be a linear code of rate at least R and relative distance at least $\delta > 0$. Then there exists $q \leqslant 2^{O(1/(R\delta\varepsilon))}$ that is a power of 2 and an invertible matrix $B \in \mathbb{F}_q^{k\times k}$ that is a (ℓ_1, s)-sparse rebasing of A with $\ell_1 = O(1/(R\delta\varepsilon))$ and $s = (1-\varepsilon)k$.*

Moreover, B can be computed explicitly given A.

With the above two lemmas in place, we prove that any code with a good relative distance yields unshardable codes:

Theorem 4.4. *Given a linear code $A : \mathbb{F}_2^k \to \mathbb{F}_2^n$ of rate $R < 1$ and relative distance $\delta > 0$, there exists $q \leqslant 2^{O(1/(R\delta\varepsilon))}$ and an explicit linear code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(O(1/(R\delta\varepsilon)), (1-\varepsilon)k)$-unshardable.*

Proof. We take

$$G = B^{-1}A,$$

where $B \in \mathbb{F}_q^{k\times k}$ is a $(\ell_1, (1-\varepsilon)k)$-sparse rebasing of A given by Lemma 4.3, where $\ell_1 = O(1/(R\delta\varepsilon))$. In particular, the rate of G is also at least R.

Moreover, it is clear from Lemma 4.2 that G is $((1-\varepsilon)k)$-unshardable. Using the sparsity property of B, we now observe that each column of B can be spanned by some ℓ_1 columns of A. In turn, this implies that for all $i \in [k]$, the basis vector $\mathbf{1}_i$ is spanned by some ℓ_1 columns of $G = B^{-1}A$. Hence, G is ℓ_1-locally decodable (from Observation 2.2). Therefore, G is $(O(1/(R\delta\varepsilon)), (1-\varepsilon)k)$-unshardable. □

Using the above theorem, we immediately obtain our main theorem as stated in the introduction:

Theorem 1.3 (Explicit unshardable codes)**.** *For every $R < 1$ and $\varepsilon > 0$, there exists $q \leqslant 2^{O_R(1/\varepsilon)}$ and an explicit infinite family of linear codes $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ of rate at least R that are $(O_R(1/\varepsilon), (1-\varepsilon)k)$-unshardable.*

Proof. We apply Theorem 4.4 with the code $A : \mathbb{F}_2^k \to \mathbb{F}_2^n$ being an explicit linear code of rate at least R and relative distance $\delta \geqslant \Omega_\varepsilon(1)$; it is well-known that such codes can be constructed explicitly for every constant $R < 1$ (see e.g. [GRS23]). □

We now prove the key lemma asserting that a sparse rebasing always exists (Lemma 4.3).

Proof of Lemma 4.3. Let $q = 2^b$ for a sufficiently large $b = O_{R,\varepsilon}(1)$ that is to be fixed later. We first note that we can assume that A is systematic without loss of generality. Indeed, since $\mathrm{rank}_{\mathbb{F}_2}(A) = k$ we can always find an invertible matrix $E \in \mathbb{F}_2^{k\times k}$ such that $A' := EA$ is systematic and has relative distance at least δ. Now, if $B' \in \mathbb{F}_q^{k\times k}$ is a (ℓ_1, s)-sparse rebasing of A', using Definition 4.1, it follows that $B := E^{-1}B'$ is a (ℓ_1, s)-sparse rebasing of A.

We will now model the "sparsity" part of the sparse rebasing B by letting

$$B = A \cdot S, \tag{8}$$

where $S \in \mathbb{F}_q^{n\times k}$ is column-sparse: in particular, it has at most ℓ_1 non-zero elements in each column. For any such S, we observe that each column of B is a $\mathbb{F}_q$-linear combination of some ℓ_1 columns of A. Our goal now is to find such an S such that the "distance" part of the sparse rebasing (i.e., Item 2 of Definition 4.1) holds. Towards this, we show the following key claim:

Claim 4.5. *Suppose the non-zero elements in each column of S are $\mathbb{F}_2$-linearly independent and that the bipartite graph $H : [n] \times [k] \to \{0,1\}$ defined by: $H(j,i) = 1$ if $S(j,i) \neq 0$ and 0 otherwise, is a $(\geqslant \delta n, s)$-vertex expander[10]. Then $|\mathbf{z}^\top B| \geqslant s$ for all $\mathbf{z} \in \mathbb{F}_2^k \setminus \{\mathbf{0}\}$.*

Proof. Note that $\mathbf{c} := \mathbf{z}^\top A \in \mathbb{F}_2^n$ has Hamming weight at least δn since A has relative distance δ. For $i \in [k]$, we have

$$(\mathbf{z}^\top B)_i = (\mathbf{z}^\top AS)_i = (\mathbf{c} \cdot S)_i = \sum_{j=1}^{n} c_j \cdot S(j,i) = \sum_{j \in T} S(j,i),$$

where $T \subseteq [n]$ denotes the indicator subset of $\mathbf{c}$. By the $\mathbb{F}_2$-linear independence assumption on the non-zero entries of $S_{\cdot,i}$ and the definition of H, we have the implication:

$$L(i) \cap T \neq \emptyset \implies (\mathbf{z}^\top B)_i \neq 0, \tag{9}$$

where $L(i) \subseteq [n]$ denotes the set of left-neighbors of a right-vertex $i \in [k]$ in H. Therefore, $|\mathbf{z}^\top B|$ is lower bounded by the number of right-neighbors of T, which is at least s, using the vertex expansion of H. This finishes the proof of Claim 4.5. □

[10] See Lemma 2.16 for a definition and construction of such expanders.

Therefore, it suffices to find a matrix S such that the premise of Claim 4.5 holds (and also ensure B is invertible). We first take $H : [n] \times [k] \to \{0,1\}$ to be a $(\geqslant \delta n, (1-\varepsilon)k)$-vertex expander as given by Lemma 2.16. Let the largest degree of a right-vertex of H be denoted by

$$\ell_1 = O(1/(R\delta\varepsilon)).$$

Now, we construct S from H by labeling the edges of H carefully. In other words, we let $S \in \mathbb{F}_q^{n\times k}$ be such that $S(j,i) \neq 0$ iff $H(j,i) = 1$ and call $S(j,i)$ the *label* of H. Assuming $q \geqslant 2^{\ell_1+1}$ (which can be ensured by taking $b = O(1/(R\delta\varepsilon))$), let

$$\zeta_1, \ldots, \zeta_{\ell_1+1} \in \mathbb{F}_q \tag{10}$$

be $\mathbb{F}_2$-linearly independent field elements[11]. Labeling the edges incident on each right-vertex using these field elements, we obtain the desired condition that the non-zero elements in each column of S are $\mathbb{F}_2$-linearly independent. However, we also need to make sure that the labeling is such that $B = AS$ is invertible.

For this, we will use the fact that there exists a matching in H saturating its right vertices (i.e., Lemma 2.16). Suppose $M := \{(i,i) : i \in [k]\}$ is such a matching without loss of generality[12]. We first give the labeling of the edges of H except those in M. In particular, we fix the labels of the edges incident on a right vertex $i \in [k]$, except the edge $(i,i) \in M$, using pairwise distinct labels from $\{\zeta_1, \ldots, \zeta_{\ell_1-1}\}$ in an arbitrary way [13]. Now, we claim that there exists a labeling of the edges in M using only the field elements ζ_{ℓ_1} and ζ_{ℓ_1+1} such that B is invertible. Note that for any such assignment, the premise of Claim 4.5 still holds. Let $\lambda_i \in \{\zeta_{\ell_1}, \zeta_{\ell_1+1}\}$ denote the label of the edge (i,i) that we wish to find. Observe that the matrix S now is of the form

$$S = \begin{bmatrix} D \\ \mathbf{0} \end{bmatrix} + S_0,$$

where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$ and $S_0 \in \mathbb{F}_q^{n\times k}$ is some fixed matrix. Hence, by inspecting (8) and using the fact that A is systematic, we have

$$B = D + B_0,$$

for some fixed $B_0 \in \mathbb{F}_q^{k\times k}$ (indeed, $B_0 = AS_0$). The following claim now finishes the construction of B.

Claim 4.6. *For all $B_0 \in \mathbb{F}_q^{k\times k}$ and $\zeta \neq \zeta' \in \mathbb{F}_q$, there exist $\lambda_1, \ldots, \lambda_k \in \{\zeta, \zeta'\}$ such that $B = D + B_0$ is invertible, where $D = \mathrm{diag}(\lambda_1, \ldots, \lambda_k)$.*

Proof. The proof is by induction on k. The base case $k = 1$ is true since for all $a_0 \in \mathbb{F}_q$, $\det(B) = \det(\lambda_1 + a_0) = \lambda_1 + a_0 \neq 0$, for at least one value of $\lambda_1 \in \{\zeta, \zeta'\}$. Now, suppose $k > 1$ and assume the claim holds for $k-1 \times k-1$ matrices. In particular, let $\widetilde{B_0} \in \mathbb{F}_q^{k-1\times k-1}$ denote the submatrix

[11] See Definition 2.15 for the definition and finding such elements.

[12] This is indeed without loss of generality since there are polynomial time algorithms to compute maximum matchings.

[13] Recall that the right-degree is at most ℓ_1 so this is possible.

of B_0 corresponding to the first $k-1$ rows and columns. By applying induction hypothesis, fix $\lambda_1, \ldots, \lambda_{k-1} \in \{\zeta, \zeta'\}$ such that

$$\widetilde{B} := \mathrm{diag}(\lambda_1, \ldots, \lambda_{k-1}) + \widetilde{B_0}$$

is invertible. Once $\lambda_1, \ldots, \lambda_{k-1}$ are fixed, note that $\det(B)$ is a linear polynomial in the variable λ_k. Furthermore, the coefficient of λ_k is exactly $\det(\widetilde{B}) \neq 0$. Hence, we conclude that there is at most one value of $\lambda_k \in \mathbb{F}_q \supseteq \{\zeta, \zeta'\}$ such that $\det(B) = 0$. This finishes the proof of Claim 4.6. □

Now, we put everything together to complete the proof that the matrix B is the desired sparse rebasing of A. First, we note that the matrix B given by the above discussion is indeed a $(\ell_1, (1-\varepsilon)k)$-sparse rebasing of A by applying Claim 4.5. Finally, we remark that the matrix B can be explicitly computed since constructions of the matrix S (which depends on the bipartite graph H given by Lemma 2.16 and $\mathbb{F}_2$-linearly independent field elements used in (10)) and the matrix D from the proof of Claim 4.6 can all be made explicit. Thus we can construct a sparse rebasing of a given code explicitly, proving Lemma 4.3.

□

5 Local Decodability Under Errors

In this section, we present constructions of codes with linear (in message length) unshardability while being locally decodable from a constant fraction errors in sublinear amount of queries.

5.1 High-Rate Regime

We provide explicit constructions of asymptotically good LDCs with near-maximal unshardability (Theorem 1.5). In fact, we prove a more general theorem that converts any arbitrary linear LDC[14] into an unshardable one:

Theorem 5.1. *Given a (δ, ℓ_0)-locally correctable code $A : \mathbb{F}_2^k \to \mathbb{F}_2^n$ of rate $R < 1$ and relative distance at least $\delta > 0$, there exists $q \leqslant 2^{1/(R\delta\varepsilon)}$ and an explicit linear code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(\delta, \ell, (1-\varepsilon)k)$-unshardable, where $\ell = O(\ell_0 \cdot (R\delta\varepsilon)^{-2} \cdot \log(1/(R\delta\varepsilon)))$.*

We immediately obtain our main result as a corollary:

Theorem 1.5 (High-rate unshardable LDCs)**.** *For every $R < 1$ and $\varepsilon > 0$, there exist $\delta = \delta(R) > 0$, $q \leqslant 2^{O_R(1/\varepsilon)}$, and an explicit infinite family of linear codes $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ of rate at least R that are $(\delta, \ell, (1-\varepsilon)k)$-unshardable, where $\ell = O_R(\exp(\sqrt{\log n \log\log n}) \cdot \varepsilon^{-2} \cdot \log(1/\varepsilon))$.*

Proof. We instantiate Theorem 5.1 with the LCC code $A : \mathbb{F}_2^k \to \mathbb{F}_2^n$ being the LCCs of [KMRS17] given by Theorem 2.5. That is, we have that A is $(\Omega_\varepsilon(1), \ell_0, s(1-\varepsilon)k)$-unshardable for some $\ell_0 = O_R(\exp(\sqrt{\log n \log\log n}))$. Hence, we get that G is indeed $(\delta, \ell, (1-\varepsilon)k)$-unshardable for $\ell = O(\ell_0 \cdot (R\delta\varepsilon)^{-2} \cdot \log(1/(R\delta\varepsilon))) = O_R(\exp(\sqrt{\log n \log\log n}) \cdot \varepsilon^{-2} \cdot \log(1/\varepsilon))$. □

Thus, we turn to the proof of the black-box reduction (Theorem 5.1):

[14] As indicated by the theorem statement, the construction becomes slightly easier if it is further locally correctable. For a construction that only uses local decodability, see Theorem 1.7.

Proof of Theorem 5.1. The idea is to perform the generic transformation discussed in Section 4 and argue that the local correcting properties of A are sufficient to get local decodability of G (under errors). That is, we claim that

$$G = B^{-1}A$$

is the desired unshardable LDC where $B \in \mathbb{F}_q^{k\times k}$ is a $(\ell_1, (1-\varepsilon)k)$-sparse rebasing of A given by Lemma 4.3, where $b, \ell_1 = O(1/(R\delta\varepsilon))$, and $q = 2^b$. In other words, for a vector $\widetilde{\mathbf{m}} \in \mathbb{F}_q^k$, we treat $\mathbf{m} = B^\top \widetilde{\mathbf{m}}$ as the message and $\mathbf{c} = \widetilde{\mathbf{m}}^\top A$ as the corresponding codeword.

We immediately observe that G is $((1-\varepsilon)k)$-unshardable using Lemma 4.2. We now show that G is (δ, ℓ)-locally decodable. Let $\mathcal{L}'$ be the local corrector corresponding to A, i.e., we have for all $\mathbf{m}' \in \mathbb{F}_q^k, \mathbf{w}' \in \mathbb{F}_2^n$ such that $\delta(\mathbf{w}', \mathbf{m}'^\top A) \leqslant \delta$ and $i \in [n]$:

$$\Pr\left[\mathcal{L}'^{\mathbf{w}'}(i) = \mathbf{m}'^\top \mathbf{b}_i\right] \geqslant 1 - \frac{1}{10b\ell_1}, \tag{11}$$

where $\mathbf{b}_1, \ldots, \mathbf{b}_k$ denote the columns of B and $\mathbf{a}_1, \ldots, \mathbf{a}_n$ denote the columns of A. We then claim that the following (Algorithm 1, denoted $\mathcal{L}$) is a local decoder for G.

Algorithm 1: Local decoder of G

Input: Oracle access to $\mathbf{w} \in \mathbb{F}_q^n$, index $i \in [k]$

Find $T \subseteq [n]$ of size ℓ_1 and $(\lambda_j)_{j\in T}$ in $\mathbb{F}_q$ guaranteed by Definition 4.1, i.e.: $\mathbf{b}_i = \sum_{j\in T} \lambda_j \cdot \mathbf{a}_j$.

Let $W \in \mathbb{F}_2^{b\times n}$ be defined by $W_{\cdot,j} \leftarrow \rho(w_j)$ for $j \in [n]$, where $\rho : \mathbb{F}_q \to \mathbb{F}_2^b$ is a $\mathbb{F}_2$-linear bijection.

for $1 \leqslant j' \leqslant b$ **do**

  $\mathbf{w}^{(j')} \leftarrow W_{j',\cdot}$ (i.e., the j'-th row of W).

for $j \in T$ **do**

  $c'_j \leftarrow \rho^{-1}(\mathcal{L}'^{\mathbf{w}^{(1)}}(j), \ldots, \mathcal{L}'^{\mathbf{w}^{(b)}}(j))$.

return $\sum_{j\in T} \lambda_j \cdot c'_j$.

In order to show correctness of the above local decoder $\mathcal{L}$, let $\mathbf{m} = B^\top \widetilde{\mathbf{m}} \in \mathbb{F}_q^k$ and $\mathbf{c} = \mathbf{m}^\top G = \widetilde{\mathbf{m}}^\top A$ be a message-codeword pair such that $\delta(\mathbf{w}, \mathbf{c}) \leqslant \delta$. We will prove that $\mathcal{L}$ returns $m_i = \mathbf{b}_i^\top \widetilde{\mathbf{m}}$ with high probability. Let $\widetilde{M} \in \mathbb{F}_2^{b\times k}$ and $\widetilde{C} \in \mathbb{F}_2^{b\times n}$ be formed by applying $\rho(\cdot)$ to $\widetilde{\mathbf{m}}$ and $\mathbf{c}$ coordinate-wise. Since ρ is $\mathbb{F}_2$-linear, we obtain for all $1 \leqslant j' \leqslant b$:

$$\widetilde{\mathbf{m}}^{(j')} A = \mathbf{c}^{(j')},$$

where $\widetilde{\mathbf{m}}^{(j')} := \widetilde{M}_{j',\cdot}$ and $\mathbf{c}^{(j')} := \widetilde{C}_{j',\cdot}$. Moreover, since $\delta(\mathbf{w}^{(j')}, \mathbf{c}^{(j')}) \leqslant \delta(\mathbf{w}, \mathbf{c}) \leqslant \delta$ for $1 \leqslant j' \leqslant b$, using (11), we have for all $j \in T$ and $1 \leqslant j' \leqslant b$ that

$$\Pr\left[\mathcal{L}'^{\mathbf{w}^{(j')}}(j) = {c^{(j')}}_j\right] \geqslant 1 - \frac{1}{10b\ell_1}.$$

Now, by a union bound over $j \in T$ and $1 \leqslant j' \leqslant b$ and using $\mathbb{F}_2$-linearity of the map ρ, we have for all $j \in T$:

$$c'_j = \rho^{-1}({c^{(1)}}_j, \ldots, {c^{(b)}}_j) = c_j,$$

with probability at least

$$1 - |T| \cdot b \cdot \frac{1}{10b\ell_1} \geqslant \frac{3}{4}.$$

Therefore, with the same probability, the output of Algorithm 1 (i.e., $\mathcal{L}^{\mathbf{w}}(i)$) is equal to

$$\sum_{j \in T} \lambda_j \cdot c_j = \sum_{j \in T} \lambda_j \cdot \widetilde{\mathbf{m}}^\top \mathbf{a}_j = \widetilde{\mathbf{m}}^\top \mathbf{b}_i = m_i,$$

the i-th message symbol. Finally note that each $\mathcal{L}'$ call in Algorithm 1 makes $O(\ell_0 \log(b\ell_1))$ oracle queries. Thus, $\mathcal{L}$ makes $\ell = O(b\ell_0\ell_1 \log(b\ell_1)) = O(\ell_0 \cdot (R\delta\varepsilon)^{-2} \cdot \log(1/(R\delta\varepsilon)))$ oracle queries. Hence, this concludes the proof of Theorem 5.1 that G is $(\delta, \ell, (1-\varepsilon)k))$-unshardable. □

5.2 Constant-Query Regime

We also construct constant query LDCs of subexponential length with near-maximal unshardability:

Theorem 1.7 (Constant-query unshardable LDCs)**.** *For every $r \in \mathbb{N}$ and $\varepsilon > 0$, there exist $\delta = \delta(r) > 0$, $q \leqslant 2^{O(1/\varepsilon)}$, and an explicit infinite family of linear codes $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that are $(\delta, \ell, (1-\varepsilon)k)$-unshardable, where $\ell = O(2^r \cdot \varepsilon^{-2} \cdot \log(1/\varepsilon))$ and $n = \exp(\exp(\sqrt[r]{\log k \log\log^{r-1} k}))$.*

Proof. The high level strategy is similar to our proof of Theorem 1.5 where we essentially use a sparse rebasing to make a given locally correctable code unshardable. However, we note that in the above proof, the assumption of A being an LCC was too strong. Indeed, in this proof, we will use an LDC over a larger message length to effectively accomplish the needed local decoding. In particular, we use the same notion of sparse rebasing (Definition 4.1) but with an additional "intermediate code". More specifically, let $k' = 2k$ and $D : \mathbb{F}_2^k \to \mathbb{F}_2^{k'}$ be an explicit *systematic* code of relative distance $\Omega(1)$. Moreover, suppose $A : \mathbb{F}_2^{k'} \to \mathbb{F}_2^n$ is a systematic (δ, ℓ_0)-locally decodable code of [Efr09] given by Theorem 2.6. Thus we have

$$\ell_0 = O(2^r) \quad \text{and} \quad n = \exp(\exp(\sqrt[r]{\log k' \log\log^{r-1} k'})). \tag{12}$$

Let $B \in \mathbb{F}_q^{k \times k}$ be a (ℓ_1, s)-sparse rebasing of the code D given by Lemma 4.3: in particular, we have $b, \ell_1 = O(1/\varepsilon), q = 2^b$, and $s = (1-\varepsilon)k$. We then claim that

$$G = B^{-1} D A$$

is the desired code.

We start by showing that G has good unshardability. For this, we will show that B is actually a (ℓ_1, s)-sparse rebasing of DB (which is a $k \times n$ matrix over $\mathbb{F}_2$). We note that distance condition of Definition 4.1 only depends on the matrix B so it continues to hold when D is replaced with DB. It remains to show that the sparsity condition also holds: Indeed, this follows by noting that the first k' columns of DB are exactly the columns of D, since A is systematic. Hence, we immediately conclude (by applying Lemma 4.2) that G is s-unshardable.

We now show that G is (δ, ℓ)-locally decodable. Let $\mathcal{L}'$ be the local decoder corresponding to A, i.e., we have for all $\mathbf{m}' \in \mathbb{F}_q^{k'}, \mathbf{w}' \in \mathbb{F}_2^n$ such that $\delta(\mathbf{w}', \mathbf{m}'^\top A) \leqslant \delta$ and $i \in [k']$:

$$\Pr\left[\mathcal{L}'^{\mathbf{w}'}(i) = m'_i\right] \geqslant 1 - \frac{1}{10b\ell_1}. \tag{13}$$

We then claim that the following (Algorithm 2, denoted $\mathcal{L}$) is a local decoder for G. Let $\mathbf{b}_1, \ldots, \mathbf{b}_k$ denote the columns of B and $\mathbf{d}_1, \ldots, \mathbf{d}_{k'}$ denote the columns of D.

Algorithm 2: Local decoder of G

Input: Oracle access to $\mathbf{w} \in \mathbb{F}_q^n$, index $i \in [k]$

Find $T \subseteq [k']$ of size ℓ_1 and $(\lambda_j)_{j \in T}$ in $\mathbb{F}_q$ guaranteed by Definition 4.1, i.e.: $\mathbf{b}_i = \sum_{j \in T} \lambda_j \cdot \mathbf{d}_j$.

Let $W \in \mathbb{F}_2^{b \times n}$ be defined by $W_{\cdot,j} \leftarrow \rho(w_j)$ for $j \in [n]$, where $\rho : \mathbb{F}_q \to \mathbb{F}_2^b$ is a $\mathbb{F}_2$-linear bijection.

for $1 \leqslant j' \leqslant b$ **do**

  $\mathbf{w}^{(j')} \leftarrow W_{j',\cdot}$ (i.e., the j'-th row of W).

for $j \in T$ **do**

  $c'_j \leftarrow \rho^{-1}(\mathcal{L}'^{\mathbf{w}^{(1)}}(j), \ldots, \mathcal{L}'^{\mathbf{w}^{(b)}}(j))$.

return $\sum_{j \in T} \lambda_j \cdot c'_j$.

In order to show correctness of the above local decoder $\mathcal{L}$, let $\mathbf{m} = B^\top \widetilde{\mathbf{m}} \in \mathbb{F}_q^k$ and $\mathbf{c} = \mathbf{m}^\top G = \widetilde{\mathbf{m}}^\top DB$ be a message-codeword pair such that $\delta(\mathbf{w}, \mathbf{c}) \leqslant \delta$. We will prove that $\mathcal{L}$ returns $m_i = \mathbf{b}_i^\top \widetilde{\mathbf{m}}$ with high probability. Let $\widetilde{M} \in \mathbb{F}_2^{b \times k}$ and $\widetilde{C} \in \mathbb{F}_2^{b \times n}$ be formed by applying $\rho(\cdot)$ to $\widetilde{\mathbf{m}}$ and $\mathbf{c}$ coordinate-wise. Since ρ is $\mathbb{F}_2$-linear, we obtain for all $1 \leqslant j' \leqslant b$:

$$\widetilde{\mathbf{m}}^{(j')} DA = \mathbf{c}^{(j')},$$

where $\widetilde{\mathbf{m}}^{(j')} := \widetilde{M}_{j',\cdot}$ and $\mathbf{c}^{(j')} := \widetilde{C}_{j',\cdot}$. Moreover, since $\delta(\mathbf{w}^{(j')}, \mathbf{c}^{(j')}) \leqslant \delta(\mathbf{w}, \mathbf{c}) \leqslant \delta$ for $1 \leqslant j' \leqslant b$, using (13), we have for all $j \in T$ and $1 \leqslant j' \leqslant b$ that

$$\Pr\left[\mathcal{L}'^{\mathbf{w}^{(j')}}(j) = (\widetilde{\mathbf{m}}^{(j')} D)_j\right] \geqslant 1 - \frac{1}{10 b \ell_1}.$$

Now, by a union bound over $j \in T$ and $1 \leqslant j' \leqslant b$ and using $\mathbb{F}_2$-linearity of the map ρ, we have for all $j \in T$:

$$c'_j = \rho^{-1}(\widetilde{\mathbf{m}}^{(1)} \mathbf{d}_j, \ldots, \widetilde{\mathbf{m}}^{(b)} \mathbf{d}_j) = \widetilde{\mathbf{m}}^\top \mathbf{d}_j,$$

with probability at least

$$1 - |T| \cdot b \cdot \frac{1}{10 b \ell_1} \geqslant \frac{3}{4}.$$

Therefore, with the same probability, the output of $\mathcal{L}^{\mathbf{w}}(i)$ is equal to

$$\sum_{j \in T} \lambda_j \cdot c'_j = \sum_{j \in T} \lambda_j \cdot \widetilde{\mathbf{m}}^\top \mathbf{d}_j = \widetilde{\mathbf{m}}^\top \mathbf{b}_i = m_i.$$

Note that each call to $\mathcal{L}'$ in Algorithm 2 makes $O(\ell_0 \log(b\ell_1))$ oracle queries. Finally, using (12), note that $\mathcal{L}$ makes $\ell = O(b \ell_0 \ell_1 \log(b\ell_1)) = O(2^r \cdot \varepsilon^{-2} \cdot \log(1/\varepsilon))$ oracle queries and the blocklength of G is $n = \exp(\exp(\sqrt[r]{\log k \log\log^{r-1} k}))$ as desired. Hence, this concludes the proof of Theorem 1.7 that G is $(\delta, \ell, (1-\varepsilon)k))$-unshardable. □

6 Superregularity-based constructions

Note that the constructions discussed in the above sections (say Theorem 1.3) do not yield anything for small number of queries like 2, since the locality (even though constant) depends on the relative distance of the underlying $\mathbb{F}_2$ base code used. Here, we give simpler constructions of unshardable codes over fields of linear size, with locality 2 (under no errors) and unshardability $\Omega(k)$. All the constructions in this section are under the weaker notion of local decoding without errors, and are built on the superregularity connection from Proposition 3.1. In particular, they circumvent the use of extension fields and expanders as done in Section 4. In the later subsections, we present several improvements to this base construction.

6.1 $(2, \Omega(k))$-Unshardable Codes

We first show the existence of $(2, \Omega(k))$-unshardable codes and then make the construct explicit later (in Theorem 6.3).

Theorem 6.1. *Let $k \geqslant 1$, $n = 2k$ and $q > 2k4^k$ be a prime power. There exists $G \in \mathbb{F}_q^{k \times n}$ that is $(2, \lceil (k+1)/2 \rceil)$-unshardable.*

Proof. We take

$$G = \begin{bmatrix} A & A + I \end{bmatrix},$$

where $A \in \mathbb{F}_q^{k \times k}$ is a matrix that we will choose later and $I \in \mathbb{F}_q^{k \times k}$ is the identity matrix. That is, denoting the columns of A by $(\mathbf{a}_i)_{i \in [k]}$, the columns of G (denoted by $(\mathbf{g}_i)_{i \in [2k]}$) are given as:

$$\mathbf{g}_i = \begin{cases} \mathbf{a}_i \text{ for } 1 \leqslant i \leqslant k \text{ and} \\ \mathbf{a}_{i-k} + \mathbf{1}_{i-k} \text{ for } k < i \leqslant 2k. \end{cases}$$

It is clear that G is locally decodable with 2 queries by Observation 2.2, since we have

$$\mathbf{1}_i = (\mathbf{a}_i + \mathbf{1}_i) - (\mathbf{a}_i) = \mathbf{g}_{k+i} - \mathbf{g}_i,$$

for all $i \in [k]$.

We will now pick A as given by the following lemma.

Lemma 6.2. *There exists $A \in \mathbb{F}_q^{k \times k}$ such that both A and $A + I$ are superregular.*

To show that the unshardability is at least $s := \lceil (k+1)/2 \rceil$ for the matrix G given by the above A, using Proposition 3.1, it suffices to show that G is $(s-1, k-1)$-superregular. That is, for a subset of rows $R \subseteq [k]$ of size $s-1$ and a subset of columns $T \subseteq [n]$ of size $k-1$ of G, we will show that the corresponding submatrix $G_{R,T}$ is full row rank. Let

$$G_{R,T} = \begin{bmatrix} A' & B' \end{bmatrix},$$

where A' is some submatrix of A and B' is some submatrix of $A + I$. Since $|T| = k - 1$, we note that either A' or B' has at least $\lceil (k-1)/2 \rceil = s - 1$ columns. Without loss of generality[15], assume that A' has at least $s - 1$ columns. Then, by Lemma 6.2, there exists a square submatrix of A' of order $s - 1$ that is full rank. Hence, $G_{R,T}$ must also be of full rank. This finishes the proof of Theorem 6.1. □

[15] The other case is similar.

We now give a proof of Lemma 6.2.

Proof of Lemma 6.2. The proof is by a straightforward application of the probabilistic method. We note for every non-empty $R, C \subseteq [k]$ of equal size, that the determinant of the corresponding submatrix $A_{R,C}$ is a non-zero polynomial in the entries of A of degree at most k. Hence, applying the polynomial identity lemma (Fact 2.12) and taking a union bound over all choices of R and C, we have that

$$\Pr_A \left[\exists R, C \subseteq [k] : |R| = |C| \neq 0 \text{ and } \operatorname{rank}(A_{R,C}) \neq |R|\right] \leqslant 2^k \cdot 2^k \cdot \frac{k}{q}.$$

Since the distribution of $A + I$ is uniform for a uniform random choice of A, we obtain the same bound for $A + I$:

$$\Pr_A \left[\exists R, C \subseteq [k] : |R| = |C| \neq 0 \text{ and } \operatorname{rank}((A+I)_{R,C}) \neq |R|\right] \leqslant 2^k \cdot 2^k \cdot \frac{k}{q}.$$

By a union bound, we thus have, over a uniform random choice of $A \in \mathbb{F}_q^{k \times k}$, that the probability that there is some square submatrix of A or $A + I$ that is not full rank is at most

$$2 \cdot 2^k \cdot 2^k \cdot \frac{k}{q} < 1,$$

by our choice of q. Hence, a matrix $A \in \mathbb{F}_q^{k \times k}$ with the claimed property exists. □

In this section, we obtain several improvements to the above construction in various axes including rate, distance, better unshardability, and field size.

Now, we turn the construction of Theorem 6.1 into an explicit code (by losing a constant factor in the unshardability).

Theorem 6.3 (2-query unshardable codes)**.** *Let $k \geqslant 1, n = 2k$, and $q \geqslant n$ be a prime power. There exists an explicit code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(2, \lceil k/3 \rceil)$-unshardable.*

Proof. Similar to the proof of Theorem 6.1, we take

$$G = \begin{bmatrix} A & A + I \end{bmatrix},$$

where A is an explicit superregular matrix. However, unlike the proof of Theorem 6.1, we don't impose any condition for $A+I$ to be superregular. We first note that G is always 2-locally decodable regardless of the choice of A. We will show that G is $\lceil k/3 \rceil$-unshardable.

Using Proposition 3.1, it suffices to prove that $G_{R,T}$ is full rank for all subsets of rows $R \subseteq [k]$ of size $s-1$ and subsets of columns $T \subseteq [n]$ of size $k-1$, in order to conclude that G is s-unshardable, where $s := \lceil k/3 \rceil$. Let $T = T_1 \cup (k + T_2)$ where $T_1 \subseteq [k]$ and $T_2 \subseteq [k]$ correspond to the columns of T corresponding to the first and second block respectively. Observe that

$$G_{R,T} = \begin{bmatrix} A_{R,T_1} & (A+I)_{R,T_2} \end{bmatrix}.$$

Now, since I is the identity matrix, we observe that $A_{R,T_2 \setminus R}$ is a submatrix of $(A+I)_{R,T_2}$. Hence, $G_{R,T}$ contains as a submatrices A_{R,T_1} and $A_{R,T_2 \setminus R}$. We have two cases:

- **Case 1:** $|T_1| \geqslant s-1$. Since $|R| = s-1$ and A is $(s-1, k-1)$-superregular, we conclude that A_{R,T_1}, and hence $G_{R,T}$, has full row rank.
- **Case 2:** $|T_1| < s-1$. We have

$$|T_2 \setminus R| \geqslant |T_2| - |R| = k - 1 - |T_1| - |R| \geqslant k - 2s + 2 \geqslant s,$$

where the last inequality is using $s = \lceil k/3 \rceil$. Hence, in particular, $A_{R,T_2 \setminus R}$ (being a submatrix of A) is full row rank as $|R| < s$. Therefore, $G_{R,T}$ is also full row rank.

This concludes the proof of Theorem 6.3, that G is $(2, \lceil k/3 \rceil)$-unshardable. □

6.2 Optimal Locality-Unshardability Trade-Off

Note that the above subsection cannot achieve unshardability beyond $\approx k/2$ since it only makes 2 queries (see Theorem 1.2). Our constructions from Section 4 are also not optimal in terms of the locality they achieve (considering constant factors). Here, we remedy these two drawbacks by achieving an optimal trade-off between the locality and unshardability for all regimes of these two parameters (existentially).

Theorem 6.4 (Optimal locality-unshardability)**.** *Let $k \geqslant 1, \ell \geqslant 2$, $n = \ell k$ and $q \leqslant \ell^{O(k)}$ be a sufficiently large prime power. There exists a code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(\ell, (\frac{\ell-1}{\ell}) \cdot (k-1))$-unshardable.*

In particular, for every $\varepsilon > 0$, there exist $(1/\varepsilon + O(1), (1-\varepsilon)k)$-unshardable codes.

Proof. The proof is an extension of that of Theorem 6.1 where G is formed by two "blocks". Here we take G to be a matrix formed by ℓ blocks as follows:

$$G = \begin{bmatrix} A_1 & A_2 & \cdots & A_{\ell-1} & A_\ell \end{bmatrix}, \tag{14}$$

where $A_1, \ldots A_{\ell-1} \in \mathbb{F}_q^{k \times k}$ will be fixed later and $A_\ell := A_1 + \cdots + A_{\ell-1} + I$. We immediately note that G is ℓ-query locally decodable by construction (using Observation 2.2). In order to show that G has good unshardability, we need to define the notion of G being *t-wise superregular*, which means that every collection of t "blocks" of G form a superregular matrix (where a "block" refers to A_i for $i \in [\ell]$). With this definition, we have the following lemma (analogous to Lemma 6.2).

Lemma 6.5. *There exists a matrix $G \in \mathbb{F}_q^{k \times \ell k}$ of the form (14) that is $(\ell-1)$-wise superregular.*

Assuming the above lemma holds, let $R \subseteq [k]$ and $T \subseteq [n]$ be row and column indices of sizes $s-1$ and $k-1$ respectively, where $s := (\frac{\ell-1}{\ell}) \cdot (k-1)$. Let $T = T_1 \cup \cdots \cup T_\ell$, where T_i corresponds to the columns in the i-th block of G (i.e., A_i). By Proposition 3.1, it suffices to show that the submatrix $G_{R,T}$ is full rank. By the pigeon-hole principle, we observe that there exists a block $i^* \in [\ell]$ such that $|T_{i^*}| \leqslant \frac{k-1}{\ell}$. In other words, $|T \setminus T_{i^*}| \geqslant s$. Now, note that $T \setminus T_{i^*}$ only contains indices of columns of G that are not in the i^*-th block. Since G is $(\ell-1)$-wise superregular, we conclude that $G_{R,T \setminus T_{i^*}}$ is full row rank (since $|R| < s$). Hence, $G_{R,T}$ is also full row rank. □

We now prove Lemma 6.5.

Proof of Lemma 6.5. Similar to Lemma 6.2, we use the probabilistic method. We note that for $A_1, \ldots, A_{\ell-1} \in \mathbb{F}_q^{k \times k}$ uniformly and independently chosen (and $A_\ell = A_1 + \cdots + A_{\ell-1} + I$), that the matrix G is "$(\ell-1)$-wise uniform", i.e., the marginal distribution of any $\ell - 1$ blocks of G is uniform. Hence, using the polynomial identity lemma (Fact 2.12) and a union bound over suitable submatrices of G, the probability that G is *not* $(\ell-1)$-wise superregular is at most

$$\ell \cdot \binom{(\ell-1)k}{k} \cdot 2^{2k} \cdot \frac{k}{q} \leqslant \frac{\ell^{O(k)}}{q} < 1,$$

by taking q suitably large. □

6.3 Improving Rate and Distance (MDS)

We note that the 2-query constructions discussed above have a rate of at most $1/2$ (i.e., $n \geqslant 2k$). The distance of the code is unclear. In this subsection, we show how to get a rate approaching one, admittedly at the expense of a worse unshardability, and also achieve the best possible distance (i.e., MDS property). In particular, we show the following.

Proposition 6.6. *Let $c \geqslant 1$, k be a multiple of c, $n = k + k/c$, and $q \geqslant n$ be a prime power. There exists an explicit code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(2, (k-1)/(2c+2))$-unshardable.*

In particular, for every $\varepsilon > 0$, there exist explicit $(2, \Omega(\varepsilon k))$-unshardable codes of rate $1 - \Theta(\varepsilon)$.

Proof. The proof closely follows that of Theorem 6.3. Let $k = c \cdot k'$ where $k' \in \mathbb{N}$. We take

$$G = \begin{bmatrix} A & B + I \end{bmatrix},$$

where A is a $k \times k'$ matrix that we will choose later and B is a $k \times k$ matrix given by $B = \begin{bmatrix} A & A & \ldots & A \end{bmatrix}$ (repeated c many times). Note that the resulting blocklength is $n = k' + k = k + k/c$, as desired. We now take

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_c \end{bmatrix},$$

where $A_1, A_2, \ldots, A_c$ are $k' \times k'$ matrices such that A is superregular; such a matrix can be constructed explicitly using Lemma 2.14. That is, we have

$$G = \begin{bmatrix} A_1 & A_1 + I & A_1 & \ldots & A_1 \\ A_2 & A_2 & A_2 + I & \ldots & A_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_c & A_c & A_c & \ldots & A_c + I \end{bmatrix}.$$

In order to argue about the locality and unshardability of G, we divide the message indices $[k]$ into c many contiguous blocks, and similarly we divide the columns of G into $c + 1$ many contiguous

blocks (and index the blocks as $0, 1, \ldots, c$ from left to right). By inspecting the matrix G, we note that for every message index $i \in [k]$, if i is the j-th block, we have

$$\mathbf{1}_i = \mathbf{g}_{k'+i} - \mathbf{g}_j,$$

where $\mathbf{g}_1, \ldots, \mathbf{g}_n$ denote the columns of G. Hence G can be locally decoded in 2 queries. We now show that the unshardability of G is at least $(k-1)/(2c+2)$.

Let $T \subseteq [n]$ be of size $k-1$ and $s := (k-1)/(2c+2)$. By Proposition 3.1, it suffices to show that for an arbitrary subset of rows $R \subseteq [k]$ of size at most s that $G_{R,T}$ has full rank. First, note that there exists at least one block of columns of G, say $j \in \{0, 1, \ldots, c\}$ that T intersects in at least $(k-1)/(c+1)$ indices (call these indices $T' \subseteq [n]$). If $j = 0$, we observe that $G_{R,T'}$ is a submatrix of the matrix A. Thus, by our assumption on the matrix A, we have that $G_{R,T'}$ is full rank. Now, suppose $j > 0$. Since each row of $G_{R,T'}$ differs from the corresponding row of $A_{R,T'}$ on at most one index, we observe that there exists $T'' \subseteq T'$ of size at least

$$|T'| - |R| \geqslant (k-1)/(c+1) - (k-1)/(2c+2) = s,$$

such that $G_{R,T''} = A_{R,T''}$. Then, since $A_{R,T''}$ (being a submatrix of A) must have full rank, we obtain that $G_{R,T''}$ (and hence $G_{R,T}$) must be full rank. This finishes the proof that G has unshardability at least $(k-1)/(2c+2)$. □

Now, we show that there exist unshardable codes achieving rate close to 1 with good distance. In fact, we show that one can guarantee the MDS property on top of locality and unshardability[16].

Theorem 6.7 (Optimal rate-distance (MDS))**.** *Let $c \geqslant 1$, k be a multiple of c, $n = k + k/c$, and $q = 2^{O(k)}$ be a sufficiently large prime power. Then there exists an MDS code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(2, (k-1)/(2c+2))$-unshardable.*

In particular, for every $\varepsilon > 0$, there exist $(2, \Omega(\varepsilon k))$-unshardable MDS codes of rate $1 - \Theta(\varepsilon)$.

Proof. The idea is to show that the construction of Proposition 6.6 is actually an MDS code for a random choice of the matrix A, with high probability. We take

$$G = \begin{bmatrix} A_1 & A_1 + I & A_1 & \ldots & A_1 \\ A_2 & A_2 & A_2 + I & \ldots & A_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_c & A_c & A_c & \ldots & A_c + I \end{bmatrix},$$

where $k' := k/c$ and $A_1, \ldots, A_c \in \mathbb{F}_q^{k' \times k'}$ are chosen uniformly and independently. Note that G is always 2-locally decodable. By a similar argument as in Proposition 6.6, we note that G is $((k-1)/(2c+2))$-unshardable if the matrix $[A_1^\top \cdots A_c^\top]$ is superregular, which happens with probability at least

$$1 - 4^k \cdot \frac{k}{q} > \frac{1}{2}. \tag{15}$$

[16] Without the MDS requirement, one can easily get *some* distance, by following along the same lines as Theorem 6.1, except by taking $G = [B \;\; A \;\; A + I]$ instead, where $B \in \mathbb{F}_q^{k \times (n-2k)}$ and $A \in \mathbb{F}_q^{k \times k}$ are such that $[I \;\; B \;\; A]$ is an MDS code.

It remains to show that G is also an MDS code with high probability. For this, we show that each $k \times k$ minor of G is symbolically non-zero. More specifically, let $S_0, S_1, \ldots, S_c \subseteq [k']$ be arbitrary subsets whose sizes sum to k corresponding to some k columns of G. We will show that there exists a choice of vectors $(\mathbf{a}_i)_{i\in[k']}$ in $\mathbb{F}_q^{k'}$ such that

$$\{\mathbf{a}_i : i \in S_0\} \cup \{\mathbf{a}_i + \mathbf{1}_i : i \in S_1\} \cup \cdots \cup \{\mathbf{a}_i + \mathbf{1}_{(c-1)k'+i} : i \in S_c\} \tag{16}$$

are linearly independent. By performing elementary column operations $\{\mathbf{a}, \mathbf{a} + \mathbf{1}_i\} \mapsto \{\mathbf{a}, \mathbf{1}_i\}$, (16) is equivalent to the following vectors being linearly independent:

$$\{\mathbf{a}_i : i \in S_0\} \cup \bigcup_{j\in[c]} \{\mathbf{1}_{(j-1)k'+i} : i \in S_j \cap S_0\} \cup \bigcup_{j\in[c]} \{\mathbf{a}_i + \mathbf{1}_{(j-1)k'+i} : i \in S_j \setminus S_0\}. \tag{17}$$

Now, we observe that taking $\mathbf{a}_i = \mathbf{0}$ for all $i \in \bigcup_{j\in[c]} S_j \setminus S_0$ results in pairwise disjoint standard basis vectors along with $(\mathbf{a}_i)_{i\in S_0}$. Hence, there exists vectors as in (16) that are linearly independent. Applying the polynomial distance lemma (Fact 2.12) for the determinant, we thus conclude that the probability that a particular $k \times k$ submatrix of G is not full rank is at most k/q. Union bounding over all such submatrices, we obtain a total error of at most

$$\binom{n}{k} \cdot \frac{k}{q} \leqslant \frac{2^{O(k)}}{q} < \frac{1}{2}. \tag{18}$$

Using (15) and (18), this completes the proof of existence of a code G that is MDS as well as being $(2, (k-1)/(2c+2))$-unshardable. □

6.4 Small Fields

The results of Theorem 6.4 achieving optimal unshardability for a given locality parameter, only apply for fields of size depending on the message length k. In this subsection, we provide a way to get similar results over fields of size independent of n (again, non-explicitly).

Theorem 6.8 (Constant-sized fields)**.** *For every $\varepsilon > 0$, there exists a prime power $q = (1/\varepsilon)^{O(1/\varepsilon)}$ such that for all integers $k \geqslant 1$, there exists a code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(O(1/\varepsilon), (1-\varepsilon)k)$-unshardable, where $n = O(k/\varepsilon)$.*

Proof. We extend the proof of Theorem 6.4 and take G to be a matrix formed by $\ell := \lceil c/\varepsilon \rceil$ (where c is a constant that we will choose later) blocks as follows:

$$G = \begin{bmatrix} A_1 & A_2 & \cdots & A_{\ell-1} & A_\ell \end{bmatrix}, \tag{19}$$

where $A_1, \ldots A_{\ell-1} \in \mathbb{F}_q^{k\times k}$ are chosen independently and uniformly at random and $A_\ell := A_1 + \cdots + A_{\ell-1} + I$. Clearly, G has locality ℓ. We now show that it has unshardability $s := \lceil (1-\varepsilon)k \rceil$ with positive probability. We now deviate from the proof of Theorem 6.4 by a more careful bound on the error probability. Because of Proposition 3.1, recall that it suffices to show that G is $(s-1, k-1)$-superregular. However, using a pigeon-hole argument as in Theorem 6.4, it is enough to show that G is $(\ell-1)$-wise $(s-1, t)$-superregular[17] where $t := (\frac{\ell-1}{\ell}) \cdot (k-1)$. Note that the marginal distribution of G restricted to any $\ell - 1$ blocks is uniformly random. Hence, the following fact will be useful.

[17] which means that every choice of $\ell - 1$ matrices among $(A_i)_{i\in[\ell]}$ form a $k \times (\ell-1)k$ matrix that is $(s-1, t)$-superregular.

Fact 6.9 (e.g. [Bel93])**.** *For $s < t$, the probability of a uniformly random $s \times t$ matrix over $\mathbb{F}_q$ being full rank is equal to*

$$\prod_{i=0}^{s-1}\left(1-\frac{1}{q^{t-i}}\right) \geqslant 1-\frac{s}{q^{t-s+1}}.$$

Now, applying a union bound over all $s \times t$ submatrices of G entirely contained in some $\ell - 1$ blocks, we obtain that the probability that G is *not* $(\ell-1)$-wise $(s-1,t)$-superregular is at most

$$\ell \cdot \binom{k}{s} \cdot \binom{(\ell-1)k}{t} \cdot \frac{s}{q^{t-s+1}} \leqslant \frac{\ell^{O(k)}}{q^{t-s+1}} < 1,$$

using the fact that $t - s \geqslant \varepsilon k/2$ (taking $c = 2$ and assuming $k \geqslant 10\varepsilon$) and taking suitable $q = (1/\varepsilon)^{\Theta(1/\varepsilon)}$. Hence, this completes the proof of Theorem 6.8. □

We note that the above proofs (including Theorem 1.3 and Theorem 6.8) only work for sufficiently large fields and in particular, do not yield anything for the binary alphabet $\mathbb{F}_2$. We show below how to get unshardable codes over all fields (existentially).

Theorem 6.10 (All fields)**.** *For every finite field $\mathbb{F}_q$, there exist infinite values of k such that there exists $n \leqslant 2k$ and a code $G : \mathbb{F}_q^k \to \mathbb{F}_q^n$ that is $(2, \Omega(k))$-unshardable.*

Proof. We remark that the proof of Theorem 6.8 does not immediately extend to very small fields since the union bound is not strong enough. Regardless, the construction achieving rate close to 1 (i.e., Proposition 6.6) does extend as we prove below. In particular, we let $c \geqslant 1$ be a sufficiently large constant dividing k, $n := k + k/c$, and take

$$G = \begin{bmatrix} A_1 & A_1 + I & A_1 & \dots & A_1 \\ A_2 & A_2 & A_2 + I & \dots & A_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ A_c & A_c & A_c & \dots & A_c + I \end{bmatrix},$$

where $A_1, \dots, A_c$ are $k' \times k'$ matrices chosen uniformly and independently at random, where $k' := k/c$. Using Observation 2.2, G is always 2-locally decodable. We now show that G is $\Omega(k)$-unshardable with positive probability. In particular, by an identical argument as Proposition 6.6, it suffices to show that the matrix

$$A = \begin{bmatrix} A_1 \\ A_2 \\ \vdots \\ A_c \end{bmatrix}$$

is $(s-1, \lfloor (k-1)/(c+1) \rfloor)$-superregular to conclude that G is s-unshardable (where s is such that $s \leqslant \lfloor (k-1)/(2c+2) \rfloor$). Applying a union bound over all $(s-1) \times \lfloor (k-1)/(c+1) \rfloor$ submatrices of A and using Fact 6.9, we get that the probability that G is *not* s-unshardable is at most

$$\binom{k}{s} \cdot \binom{k/c}{\lfloor (k-1)/(c+1) \rfloor} \cdot \frac{s}{q^{\lfloor (k-1)/(c+1) \rfloor - s + 1}}. \tag{20}$$

We will now set $s := \lfloor (k-1)/(2c+2)^2 \rfloor$ and argue that the above quantity is less than 1. First, we have

$$\binom{k}{s} \leqslant \left(\frac{ek}{s}\right)^s \leqslant c^{O(k/c^2)}.$$

By similar calculations, the second factor is

$$\binom{k/c}{\lfloor (k-1)/(c+1) \rfloor} = \binom{k/c}{k/c - \lfloor (k-1)/(c+1) \rfloor} \leqslant c^{O(k/c^2)}.$$

Finally, we have

$$\frac{s}{q^{\lfloor (k-1)/(c+1) \rfloor - s + 1}} \leqslant \frac{k}{2^{\Omega(k/c)}}.$$

Combining the above three bounds, we get that (20) is at most

$$c^{O(k/c^2)} \cdot \frac{k}{2^{\Omega(k/c)}} \leqslant \frac{1}{2^{\Omega(k/c - \log k - k \log c/c^2)}} < 1,$$

where the last inequality follows by choosing $c = O(1)$ to be sufficiently large. Therefore, there exists a code G that is $s = \Omega(k)$-unshardable concluding Theorem 6.10. □

7 Ruling Out Local-Local Unshardability

In this section, we discuss a variant of our definition of unshardability, which we call "local-local unshardability," and show a negative result for the same.

For a map C, we say T *ℓ-locally determines* i w.r.t. C if there exists $D \subseteq T$ of size $|D| \leqslant \ell$ such that D determines i w.r.t. C. We now define a variant of unshardability where if one party fails to *locally* decode then the same will be true for many other parties.

Definition 7.1 (Local-Local Unshardability)**.** *Let $n, k \in \mathbb{N}$ and $C : \Sigma^k \to \Sigma^n$. We say C is a* (ℓ, s)-local-local unshardable code *if the following holds:*

1. *C is ℓ-locally decodable under no errors (as in Definition 2.1).*
2. *For every $T \subseteq [n]$, if there exists an index $i \in [k]$ that is not* ℓ-locally *determined by T then there are at least s indices in $[k]$ that are also not* ℓ-locally *determined by T.*

We note that it is apriori unclear how the above definition compares to our original definition of unshardability since both the premise and conclusion of Item 2 in the above definition are weaker than those of Definition 2.7. Regardless, we are able to essentially rule out the existence of interesting local-local unshardable codes.

Theorem 7.2. *Any $C : \Sigma^k \to \Sigma^n$ that is a (ℓ, s)-local-local unshardable code must have $s \leqslant \ell$.*

Proof. Fix any $C : \Sigma^k \to \Sigma^n$ that satisfies condition (1) of the above definition. Choose the minimal subset $T^* \subseteq [n]$ such that condition (1) still holds for $C|_{T^*}$. For this T^*, every $i \in [k]$ is determined by some $D_i \subseteq T^*$ of size $|D_i| \leqslant \ell$. We will find the $j^* \in [n]$ that is "touched" by the fewest D_i's, that is, j^* minimizes $\varphi(j) := |\{i : j \in D_i\}|$ over all $j \in [n]$. Using the upper bound on the size of each D_i, by an averaging argument we have $\varphi(j^*) \leqslant k\ell/|T^*| \leqslant \ell$ where the final inequality follows since $|T^*| \geqslant k$ because $C|_{T^*}$ is injective.

Now set $T' = T^* \setminus \{j^*\}$. By minimality of T^* we have that there exists $i^* \in [k]$ that is not locally determined by any $D \subseteq T'$ of size $\leqslant \ell$. On the other hand, there are at least $(k-\ell)$-many i's such that D_i avoids j^* so $D_i \subseteq T'$; these i's remain ℓ-locally determined by T'. We conclude that $s \leqslant \ell$. □

It is easy to tightly meet this lower bound with the "batching" style construction that we briefly described in Section 1. That is, simply group the input into ℓ-size blocks and apply a superregular linear map block-wise. This is unsurprising as the batching technique essentially makes local and global determination equivalent.

AI Disclosure. The main technical ideas and constructions of Section 4 and consequently Section 5 using the notion of a *sparse rebasing w.r.t. extension fields* (Definition 4.1) were found by ChatGPT Sol-5.6 prompted on an initial draft of our characterizations and constructions from Section 3 and Section 6. The authors verified, digested, and contextualized the AI proofs and take full responsibility for the results. More details can be found in Section 1.2. AI tools were used to assist with formatting and organizing the manuscript, and correcting typos. No AI tools were used in modeling or defining unshardable codes.

References

[ABNNR92] Noga Alon, Jehoshua Bruck, Joseph Naor, Moni Naor, and Ron M Roth. "Construction of asymptotically good low-rate error-correcting codes through pseudo-random graphs". In: *IEEE Transactions on information theory* 38.2 (1992), pp. 509–516 (cit. on p. 7).

[AEL95] Noga Alon, Jeff Edmonds, and Michael Luby. "Linear time erasure codes with nearly optimal recovery". In: *Proceedings of IEEE 36th Annual Foundations of Computer Science*. 1995, pp. 512–519. DOI: 10.1109/SFCS.1995.492581 (cit. on p. 7).

[Bel93] Eric David Belsley. *Rates of convergence of Markov chains related to association schemes*. Harvard University, 1993 (cit. on p. 28).

[CY24] Gil Cohen and Tal Yankovitz. "Asymptotically-Good RLCCs with $(\log n)^{2+o(1)}$ Queries". In: *39th Computational Complexity Conference (CCC 2024)*. Schloss Dagstuhl–Leibniz-Zentrum für Informatik. 2024, 8:1–8:16. DOI: 10.4230/LIPIcs.CCC.2024.8 (cit. on p. 5).

[Coh16] Michael B Cohen. "Ramanujan graphs in polynomial time". In: *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE. 2016, pp. 276–281 (cit. on p. 12).

[DL78] Richard A. DeMillo and Richard J. Lipton. "A probabilistic remark on algebraic program testing". In: *Information Processing Letters* 7.4 (1978), pp. 193–195. DOI: 10.1016/0020-0190(78)90067-4 (cit. on p. 10).

[Efr09] Klim Efremenko. "3-query locally decodable codes of subexponential length". In: *Proceedings of the forty-first annual ACM symposium on Theory of computing*. 2009, pp. 39–44 (cit. on pp. 5, 9, 20).

[Gal62] Robert Gallager. "Low-density parity-check codes". In: *IRE Transactions on information theory* 8.1 (1962), pp. 21–28 (cit. on p. 7).

[GRS23] Venkatesan Guruswami, Atri Rudra, and Madhu Sudan. *Essential Coding Theory (Book draft)*. 2023. URL: http://www.cse.buffalo.edu/atri/courses/coding-theory/book (cit. on p. 16).

[KMRS17] Swastik Kopparty, Or Meir, Noga Ron-Zewi, and Shubhangi Saraf. "High-rate locally correctable and locally testable codes with sub-polynomial query complexity". In: *Journal of the ACM (JACM)* 64.2 (2017), pp. 1–42 (cit. on pp. 5, 9, 18).

[KM24] Vinayak M Kumar and Geoffrey Mon. "Relaxed local correctability from local testing". In: *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*. 2024, pp. 1585–1593 (cit. on p. 5).

[MSS18] Adam W Marcus, Daniel A Spielman, and Nikhil Srivastava. "Interlacing families IV: Bipartite Ramanujan graphs of all sizes". In: *SIAM Journal on computing* 47.6 (2018), pp. 2488–2509 (cit. on p. 12).

[RL89] Ron M Roth and Abraham Lempel. "On MDS codes via Cauchy matrices". In: *IEEE transactions on information theory* 35.6 (1989), pp. 1314–1319 (cit. on p. 11).

[Sch80] Jacob T. Schwartz. "Fast Probabilistic Algorithms for Verification of Polynomial Identities". In: *Journal of the ACM* 27.4 (1980), pp. 701–717. DOI: 10.1145/322217.322225 (cit. on p. 10).

[SS96] Michael Sipser and Daniel A Spielman. "Expander Codes". In: *IEEE Transactions on Information Theory* 42.6 (1996), pp. 1710–1722 (cit. on p. 7).

[Tan81] R. Michael Tanner. "A Recursive Approach to Low Complexity Codes". In: *IEEE Transactions on Information Theory* 27.5 (1981), pp. 533–547 (cit. on p. 7).

[Vad12] Salil P. Vadhan. "Pseudorandomness". In: *Foundations and Trends® in Theoretical Computer Science* 7.1–3 (2012), pp. 1–336. ISSN: 1551-305X. DOI: 10.1561/0400000010. URL: http://dx.doi.org/10.1561/0400000010 (cit. on p. 12).

[Zip79] Richard Zippel. "Probabilistic algorithms for sparse polynomials". In: *Symbolic and Algebraic Computation*. Springer Berlin Heidelberg, 1979, pp. 216–226. DOI: 10.1007/3-540-09519-5_73 (cit. on p. 10).

A Unshardability of Tensor-Product Reed-Solomon Codes

In this section, we show the challenge in constructing unshardable codes using a tensor-product of Reed-Solomon codes (equivalently multivariate polynomials with an individual degree bound). While such codes can be locally decoded in $\ell = o(k)$ queries, obtaining unshardability appears to be challenging. In particular, we show that for the natural setting of treating the message symbols as low-*individual*-degree evaluations over a "grid", there exist settings where unshardability remains $s = O(\ell)$.

We start with a few definitions. For positive integers d, m, let $\mathcal{P}_d(\mathbb{F}_q^m)$ (or simply $\mathcal{P}_d$ if the rest of the parameters are known from context) denote the set of m-variate polynomials over $\mathbb{F}_q$ of individual degree at most d. We say that $I \subseteq \mathbb{F}_q^m$ is an *interpolating set* for $\mathcal{P}_d(\mathbb{F}_q^m)$ if for every assignment $f : I \to \mathbb{F}_q$, there exists a unique polynomial $P \in \mathcal{P}_d(\mathbb{F}_q^m)$ such that $f(\mathbf{a}) = P(\mathbf{a})$ for

all $\mathbf{a} \in I$. A concrete interpolating set that is the focus of this section is the "grid" $I = S^2$, with $|S| = d+1$.

We are now ready to define polynomial evaluation codes.

Definition A.1 (Polynomial evaluation codes)**.** *For an interpolating set $I \subseteq \mathcal{P}_d(\mathbb{F}_q^m)$ and an* evaluation set *$E \subseteq \mathbb{F}_q^m$, we define the evaluation code $\mathcal{C}_d(I,E) : \mathbb{F}_q^I \to \mathbb{F}_q^E$ (or just $\mathcal{C}_d$ for simplicity) as follows: For a given $f \in \mathbb{F}^I$, let $P \in \mathcal{P}_d$ be the unique polynomial such that $P(\mathbf{a}) = f(\mathbf{a})$ for all $\mathbf{a} \in I$. We then let*

$$\mathcal{C}_d(f) = (P(\mathbf{b}))_{\mathbf{b} \in E}.$$

Note that $\mathcal{C}_d(I,E)$ has message length $k = |I|$ and blocklength $n = |E|$. We mainly focus on $m = 1$ and $m = 2$. We begin by recalling our discussion from Section 1 that when $m = 1$ (i.e., for (a variant of) Reed-Solomon codes) we achieve maximal unshardability (but no locality):

Claim A.2 ($m = 1$)**.** *Let $I \subseteq \mathbb{F}_q$ be of size $d+1$ and $E \subseteq \overline{I}$ be of size at least $d+1$. Then $\mathcal{C}_d(I,E)$ is $(d+1, d+1)$-unshardable.*

We now prove that for $m = 2$, there exist choices of $I = S^2$ such that $\mathcal{C}_d$ has unshardability at most $O(\sqrt{k})$. Combined with the fact that $\mathcal{C}_d$ is $O(d) = O(\sqrt{k})$ establishes the main conclusion of this section.

Claim A.3. *Let S be a subfield of $\mathbb{F}_q$ of size $d+1$ and $q \geqslant 4d+3$. Then, for $I = S^2$ and $E \supseteq \overline{S}^2$, if $\mathcal{C}_d(I,E)$ is (ℓ, s)-unshardable, then we have $s \leqslant d+1 = \sqrt{|I|}$.*

Proof. For $\alpha \in \mathbb{F}_q$, let $L_\alpha := \{(a_1, a_2) \in \mathbb{F}_q^2 : a_1 + a_2 = \alpha\}$. Letting $T := \left(\cup_{\alpha \in S \setminus \{0\}} L_\alpha \right) \cap \overline{S}^2$, we first note that T does not determine $(0,0) \in I$ since $\prod_{\alpha \in S \setminus \{0\}} (X + Y - \alpha)$ (which is a non-zero polynomial in $\mathcal{P}_d$) vanishes on T but not at $(0,0)$. We will show that T doesn't determines at most $d+1$ points in S^2. Let $P(X,Y) \in \mathcal{P}_d$ be a polynomial of individual degree at most d such that $P(a_1, a_2) = 0$ for all $(a_1, a_2) \in T$. We will show that $P(a_1, a_2) = 0$ for all $(a_1, a_2) \in S^2$ such that $a_1 + a_2 \neq 0$. In turn, this would establish that T determines all $(a_1, a_2) \in S^2$ such that $a_1 + a_2 \neq 0$ and hence the unshardability s must be at most

$$|S^2| - |\{(a_1, a_2) \in S^2 : a_1 + a_2 \neq 0\}| = (d+1)^2 - d(d+1) = d+1.$$

Fix $\alpha \in S \setminus \{0\}$ arbitrarily. Since P vanishes on $L_\alpha \cap \overline{S}^2$, we have that $P(X, \alpha - X) = 0$ for all $X \in \overline{S}$ such that $\alpha - X \in \overline{S}$. As $|S| = d+1$, there are at most $2(d+1)$ values of $X \in \mathbb{F}_q$ such that $X \notin \overline{S}$ or $\alpha - X \notin \overline{S}$. Hence there are at least $q - 2(d+1) \geqslant 2d+1$ values of X such that $X \in \overline{S}$ and $\alpha - X \in \overline{S}$. In other words, $P(X, \alpha - X)$ has at least $2d+1$ zeroes. However, note that $P(X, \alpha - X)$ is a polynomial of degree at most $2d$ (since P has individual degree at most d). Thus, $P(X, \alpha - X) \equiv 0$ identically and hence $(X + Y - \alpha)$ divides $P(X,Y)$. Therefore, $P(a_1, a_2) = 0$ for $(a_1, a_2) \in S^2$ such that $a_1 + a_2 = \alpha$. Since this holds for all $\alpha \in S \setminus \{0\}$ and S is a subfield, we have that $P(a_1, a_2) = 0$ for all $(a_1, a_2) \in S^2$ such that $a_1 + a_2 \neq 0$. This finishes the proof. □

ECCC ISSN 1433-8092
https://eccc.weizmann.ac.il