

Error-Correcting Graph Codes

Swastik Kopparty*

Aditya Potukuchi†

Harry Sha‡

November 19, 2024

Abstract

In this paper, we construct *Error-Correcting Graph Codes*. An error-correcting graph code of distance δ is a family C of graphs on a common vertex set of size n , such that if we start with any graph in C , we would have to modify the neighborhoods of at least δn vertices in order to obtain some other graph in C . This is a natural graph generalization of the standard Hamming distance error-correcting codes for binary strings.

Yohananov and Yaakobi were the first to construct codes in this metric, constructing good codes for $\delta < 1/2$, and optimal codes for a large-alphabet analogue. We extend their work by showing

- a.) Combinatorial results determining the optimal rate vs. distance trade-off nonconstructively.
- b.) Graph code analogues of Reed-Solomon codes and code concatenation, leading to positive distance codes for all rates and positive rate codes for all distances.
- c.) Graph code analogues of dual-BCH codes, yielding large codes with distance $\delta = 1 - o(1)$.

This gives an explicit “graph code of Ramsey graphs”.

Several recent works, starting with the paper of Alon, Gujgiczer, Körner, Milojević, and Simonyi, have studied more general graph codes; where the symmetric difference between any two graphs in the code is required to have some desired property. Error-correcting graph codes are a particularly interesting instantiation of this concept.

*Department of Mathematics and Department of Computer Science, University of Toronto. Research supported by an NSERC Discovery Grant. Email: swastik.kopparty@utoronto.ca

†Department of Computer Science, York University. Supported in part by NSERC Discovery grants RGPIN-2023-05087 & DGECR-2023-00408 Email: apotu@yorku.ca

‡Department of Computer Science, University of Toronto. Email: shaharry@cs.toronto.edu

Contents

1	Introduction	3
1.1	Rate versus distance and prior work	4
1.2	Results	5
1.3	Techniques	5
1.4	Concluding thoughts and questions	6
2	Graph codes: Basics	7
3	Existence of optimal graph codes	8
4	Explicit graph codes for high distance: Concatenated Codes	10
4.1	Symmetric concatenation	12
4.2	Concatenated graph codes	13
4.3	Multiple concatenation	14
5	Explicit graph codes with very high distance: dual-BCH Codes	15
5.1	Warmup: a graph code with dimension $\Omega(\log n)$	15
5.2	Larger dimension	17
6	Appendix	23
6.1	Justensen-like code	23

1 Introduction

In this paper, we study *Error-Correcting Graph Codes*. These are large families of undirected graphs on the same vertex set such that any two graphs in the family are “far apart” in a natural graph distance. Informally, the graph distance between two graphs on the same vertex set of size n measures the minimum number of vertices one needs to delete to make the resulting graphs identical (not just isomorphic). This can also be thought of as (1) the number of vertices whose neighborhoods one has to modify to go from one graph to another, (2) the vertex cover number of the symmetric difference of the two graphs, or (3) n minus the largest independent set in the symmetric difference of the two graphs.

Definition: (Graph Distance) Given two graphs G and H on vertex set $[n]$, the graph distance $d_{\text{graph}}(G, H)$ is the size of the smallest set $S \subseteq [n]$ such that $G[\overline{S}] = H[\overline{S}]$.

This is a very natural metric and encompasses deep information about graphs. For example, note the following two simple facts: (1) the graph distance of a graph from the empty graph is n minus the independence number of the graph. (2) the graph distance of a graph from the complete graph is n minus the clique number of the graph. Thus, the answer to the question: “how far can a graph be from both the empty graph and the complete graph?” is precisely the question of finding the right bound for the diagonal Ramsey numbers; the answer is $n - O(\log n)$.

Codes in the graph-distance metric were defined and initially studied by Yohananov and Yaakobi [YY19], and Yohananov, Efron, and Yaakobi [YFY20]. Their setting allows for arbitrary symbols $\alpha \in \mathbb{F}_q$ to be written on every edge of K_n ($q = 2$ corresponds to graph codes), including self-loops. They gave a complete understanding of the combinatorial limits for large q , optimal explicit constructions in the large q setting, and asymptotically good explicit constructions for the $q = 2$ setting for distance at most $n/2$.

Error-correcting graph codes also fall into the general framework of graph codes defined by Alon, Gijgiczer, Körner, Milojević, and Simonyi [Alo+23], where for a fixed family \mathcal{F} of graphs, one seeks a large code C of graphs on the same n -vertex set such that the symmetric difference of any two graphs in C does not lie in \mathcal{F} . This class of problems was studied for a wide variety of natural \mathcal{F} in a number of recent works [Alo+23; Alo23b; Alo23a]. As discussed in [Alo23b], for a suitable choice of \mathcal{F} , graph codes become equivalent to classical Hamming distance codes.

Additionally, the notion of graph distance arises in the definition of node differential privacy (see, for example [CZ13; Hay+09; Kas+13; JSW24]). One instantiation of this setting is a graph encoding a social network where vertices correspond to people and edges correspond to social connections. The goal for node differential privacy is to design an algorithm \mathcal{A} that approximately computes certain statistics of the graph (such as counting edges, triangles, and connected components) while maintaining each individual’s privacy. Here, privacy is ensured by requiring that the output distribution on a certain graph does not change by much when any one vertex is deleted. In other words, for any graphs G, H of graph distance 1, the output distribution of \mathcal{A} on the G and H should be similar. Graph distances greater than 1 are then considered in the continual release model where the graph varies over time as studied in [JSW24].

Finally, we note that error-correcting codes are pseudorandom objects, and the connection to Ramsey graphs suggests that error-correcting graph codes might be closely related to pseudorandom graphs. Thus, the problem of studying and explicitly constructing a pseudorandom family of pseudorandom graphs is interesting in its own right.

1.1 Rate versus distance and prior work

Similar to the Hamming setting, we briefly define the dimension, rate, and distance of a graph code C on n vertices where each edge is allowed to have an arbitrary symbol $\alpha \in \mathbb{F}_q$ written on it.

- **Dimension.** $k = \log_q(|C|)$.
- **Rate.** $R = \log_q(|C|)/\binom{n}{2}$.
- **Distance.** The distance of a code is the largest d such that $d_{\text{graph}}(G, H) \geq d$ for each $G, H \in C$ such that $G \neq H$. The relative distance δ is d/n .

Unless specified otherwise, we will always be interested in the asymptotics of the above parameters as $n \rightarrow \infty$. All construction of graph codes in this paper will be linear subspaces of $\mathbb{F}_q^{\binom{n}{2}}$. A construction of graph codes is called explicit if a basis for the space can be computed in time $\text{poly}(n)$. A construction of graph codes is called strongly explicit if the (i, j) entry of the e 'th basis element can be computed in time $\text{poly}(\log(n))$ when given i, j, e as input.

Yohananov and Yaakobi [YY19] showed, by an argument similar to the Singleton bound, that for a graph code of dimension k and distance d , we have

$$k \leq \binom{n-d+1}{2}. \quad (1)$$

In terms of rate R , and relative distance δ , we have

$$R \leq (1 - \delta)^2 + o(1). \quad (2)$$

They [YY19] (Construction 3) then showed, by explicit construction, that this bound is tight for large q . The main ingredient is a way to convert Hamming distance into graph distance using the tensor product. A similar code will be an important ingredient in our constructions.

[YY19] (Construction 5) also gave explicit constructions of codes over the binary alphabet for $\delta \leq 1/2$. This leverages a connection to symmetric array codes and a construction by Schmidt [Sch15] related to the rank metric. The trade-off achieved, $R \sim 1 - 2\delta$, is close to the upper bound (2) for δ very small.

Finally, we note that [YY19] and [YFY20] also give optimal constructions in the $\delta = O(1/n)$ regime.

These constructions¹ are summarized in Table 1.

¹Note that all of the constructions in Table 1 are for undirected graphs where self-loops are allowed. Since these are linear codes, one can impose n linear constraints to get codes with no self-loops (zeros on the diagonals of the adjacency matrix). Since the block length of these codes is $\binom{n}{2}$, adding these constraints changes the rate by $o(1)$ while preserving explicitness.

Adding the 0-diagonal constraint to the symmetric array codes while preserving *strong explicitness* is more subtle. Alessandro Neri [Ner] observed that this can be done using results of Gow and Quinlan (Theorem 6 and Theorem 7 of [GQ09]), which themselves are related to results of Delsarte and Goethals [DG75].

Name	δ	Field	Tradeoff
[YY19]	≤ 1	$\mathbb{F}_q, q \geq n - 1$	$R = (1 - \delta)^2$ (optimal)
[YY19]	$\leq 1/2$	\mathbb{F}_2	$R = 1 - 2\delta$
[YY19],[YFY20]	$= O(1/n)$	\mathbb{F}_2	optimal

Table 1: Summary of constructions in [YY19], and [YFY20]

1.2 Results

Our main results are:

- 1.) We observe that there are binary graph codes achieving $R = (1 - \delta)^2 - o(1)$ for any constant $\delta \in (0, 1)$. Thus $R = (1 - \delta)^2 - o(1)$ (or $\delta = 1 - \sqrt{R} - o(1)$) is the optimal R vs δ tradeoff even for the binary alphabet ([YY19] showed this for large alphabet).
- 2.) We give constructions of graph codes that have positive constant R for all constant $\delta < 1$. We give an explicit construction with $\delta = 1 - O(R^{1/6})$, and a strongly explicit construction with $\delta = 1 - O(R^{1/8})$. We also give a quasi-polynomial time explicit construction achieving $\delta = 1 - O(R^{1/4})$. Compare this with the optimal nonconstructive tradeoff of $\delta = 1 - R^{1/2}$.

Although these codes are not optimal, they are the first binary error-correcting graph codes achieving $\delta > 1/2$ with a constant rate.

- 3.) We give (strongly) explicit constructions of graph codes with very high distance $\delta = 1 - O(n^{-\epsilon})$ and inverse polynomial rate $R = \Omega(n^{\epsilon-1/2})$ for constant $\epsilon > 0$. This gives a “graph code of Ramsey graphs,” as will be discussed later.

Independent work: Pat Devlin and Cole Franks [Dev] independently proposed the study of graph error-correcting codes under this metric, determined the optimal R vs δ tradeoff for binary codes, and gave some weaker explicit constructions of graph codes that worked for certain ranges of R and δ .

1.3 Techniques

We now discuss our techniques. We often specify graphs by their adjacency matrices, viewed as matrices with \mathbb{F}_2 entries.

Our nonconstructive existence result is a straightforward application of the probabilistic method. We consider a uniformly random \mathbb{F}_2 -linear subspace of the \mathbb{F}_2 -linear space of symmetric 0-diagonal $n \times n$ matrices (i.e., the space of all adjacency matrices of graphs); this turns out to give a good graph code with optimal R vs δ tradeoff.

To get explicit constructions of asymptotically good codes for any constant $\delta \in (0, 1)$, we use several ideas.

- a.) We start with a slight variation of [YY19] (Construction 3), which gives a way to get a good graph code from a classical Hamming-distance linear code $C \subseteq \mathbb{F}_2^n$. We first consider the tensor code $C \otimes C$, where the elements are matrices whose rows and columns are codewords of C . A-priori, C could contain matrices that are neither symmetric nor have a 0 diagonal. But interestingly, if we consider the set C^* of all matrices in $C \otimes C$ that are symmetric and have 0 diagonal, then C^* is a linear space with quite large dimension. In particular, if the

classical Hamming distance code C has positive rate, then so does the graph code C^* . We call this construction $C^* = \text{STCZD}(C)$ (Symmetric Tensor Code with Zero Diagonals).

It turns out that if C has good relative distance (in the Hamming metric), then $\text{STCZD}(C)$ has good distance in the graph metric. However, the relative graph distance of $\text{STCZD}(C)$ such a code is bounded by the relative distance of C – and since C is a binary code, this is at most $1/2$.

- b.) Now, we bring in another idea from the Hamming code world: code concatenation. Instead of constructing a graph code of symmetric zero-diagonal matrices over \mathbb{F}_2 , we instead construct a “large-alphabet graph code” of symmetric zero-diagonal matrices over \mathbb{F}_q for some large $q = 2^t$ and then try to reduce the alphabet size down to 2 by replacing the q -ary symbols with \mathbb{F}_2 -matrices with suitable properties.

Applying the analog of STCZD to a large alphabet code allows one to get large-alphabet graph codes with large δ , approaching 1 (since over large alphabets, Hamming distance codes can have length approaching 1). Using Reed-Solomon codes as these large alphabet codes also allows us to make the STCZD construction strongly explicit. Furthermore, when applied to Reed-Solomon codes, these codes have a natural direct description: these are the evaluation tables of low-degree bivariate polynomials $P(X, Y)$ on product sets $S \times S$ that are (1) symmetric (to get a symmetric matrix), and (2) multiples of $(X - Y)^2$ (to get zero diagonal).

- c.) What remains now is to develop the right kind of concatenation so that the resulting graph code has good distance. This turns out to be subtle and requires an “inner code” with a stronger “directed graph distance” property with δ nearly 1. Fortunately, this inner code we seek is very slowly growing in size, and we may find it by brute force search. This concludes our description of our explicit construction of graph codes with δ approaching 1 and positive constant R .

Finally, we discuss our constructions for very high distances, $\delta = 1 - o(1)$. In this regime, as mentioned earlier, this is related to constructions of Ramsey graphs, a difficult problem in pseudorandomness with a long history. Our constructions work up to $\delta = 1 - \Omega\left(\frac{1}{\sqrt{n}}\right)$; concretely, we get a large linear space of graphs such that all graphs in the family have no clique or independent set of size $\Omega(\sqrt{n})$. The construction is based on polynomials over finite fields of characteristic 2: When $n = 2^t$, we consider a linear space of certain low degree univariate polynomials $f(X)$ over \mathbb{F}_{2^t} and create the \mathbb{F}_2 matrix with rows and columns indexed by \mathbb{F}_{2^t} whose x, y entry is $\text{Tr}(f(x + y))$. Here Tr is the finite field trace map from \mathbb{F}_{2^t} to \mathbb{F}_2 . The use of Tr of polynomials is inspired by the construction of dual-BCH codes. We then show that any such matrix has no large clique or independent set unless $\text{Tr} \circ f$ is identically 0 or identically 1 (corresponding to the empty and complete graphs, respectively). The proof uses the Weil bounds on character sums and a Fourier analytic approach to bound the independence number for the graphs.

Our constructions are listed in Table 2.

1.4 Concluding thoughts and questions

The most interesting question in this context is obtaining explicit graph code constructions with optimal R vs δ tradeoff. While we have several constructions achieving nontrivial parameters in various regimes, it would even be interesting to get the right asymptotic behavior for the endpoints

Name	Approximate Tradeoffs	Explicit/Strongly Explicit?
Random Linear Codes (Proposition 2.4)	$R = (1 - \delta)^2 - o(1)$	No/No
Concatenated RS Tensor Codes (Code 3)	$R = (1 - \sqrt{\delta})^4 - o(1)$	Almost Yes ² /No
Double Concatenated RS Tensor Codes	$R = (1 - \delta^{1/3})^6 - o(1)$	Yes/No
Triple Concatenated RS Tensor Codes (Code 4)	$R = (1 - \delta^{1/4})^8 - o(1)$	Yes/Yes
Dual BCH Codes (Code 6)	$R = \log(n)(1 - \delta)/\sqrt{n}$	Yes/Yes

Table 2: A list of constructions of error-correcting graph codes in this paper. A particular focus is the regime where δ is very close to 1.

with δ approaching 1. The setting of large δ (including $\delta = 1 - o(1)$) seems especially challenging, given the connection with the notorious problem of constructing Ramsey graphs.

Another interesting question is to get decoding algorithms for graph codes. For a certain graph code C , if we are given a graph that is promised to be close in graph distance to some graph G in C . Then, can we efficiently find G ?

A more general context relevant to error-correcting graph codes is the error correction of strings under more general error patterns. Suppose we have a collection of subsets $S_i \subseteq [m]$ for $i \in [t]$, where $\bigcup_i S_i = [m]$. These S_i denote the corruption zones; a single “corruption” of a string $z \in \{0, 1\}^m$ entails, for some $i \in [t]$, changing $z|_{S_i}$ to something arbitrary in $\{0, 1\}^{S_i}$. We want to design a code $C \subseteq \{0, 1\}^m$ such that starting at any $x \in C$ if we do fewer than d corruptions to x , we do not end up at any $y \in C$ with $y \neq x$. When the S_i are all of size b and form a partition of $[m]$ into $t = m/b$ parts, then such a code is exactly the same as a classical Hamming distance code an alphabet of size 2^b . Error-correcting graph codes give a first step into the challenging setting where the S_i all pairwise intersect - here we have $m = \binom{n}{2}$, $t = n$, the S_i (which correspond to all edges incident on a given vertex) all have size $n - 1$, and every pair S_i and S_j intersect in exactly 1 element. It would be interesting to develop this theory - to both find the limits of what is achievable and to develop techniques for constructing codes against this error model.

Finally, there are many other themes from classical coding theory that could make sense to study in the context of graph codes and graph distance, including in the context of sublinear time algorithms. It would be interesting to explore this.

Organization of this paper: We set up basic notions in Section 2. We show the existence of optimal graph codes in Section 3. In Section 4 we construct asymptotically good codes. Finally, in Section 5, we show explicit constructions of graph codes with very high distance.

2 Graph codes: Basics

Definitions and notations. All graphs will be simple, undirected graphs on the vertex set $[n]$ unless otherwise noted. For any graph G , use A_G to denote the adjacency matrix of G , and view A_G as an element of the vector space $\mathbb{F}_2^{\binom{n}{2}}$. For two graphs G, H , let $G \Delta H$ be the symmetric difference of the two graphs, i.e. $A_{G \Delta H} = A_G - A_H$. For a subset $S \subseteq [n]$, we use $G[S]$ to denote the subgraph of G induced by the vertex set S . If A is a $n \times n$ matrix and $S, T \subseteq [n]$, let $A_{S,T}$ be the sub-matrix indexed by S on the rows and T on the columns. For $x \in [0, 1]$, we use $h_2(x) = -x \log_2 x - (1 - x) \log_2 (1 - x)$ to denote the binary entropy function.

Definition 2.1 (Graph distance and relative graph distance).

- The *graph distance* between two graphs G and H , denoted by $d_{\text{graph}}(G, H)$, is the smallest $d \in \mathbb{N}$ such that there is a set $S \subseteq [n]$, $|S| = d$, and $G[[n] \setminus S] = H[[n] \setminus S]$.
- The *relative graph distance*, or simply *relative distance*, between G and H is denoted by $\delta_{\text{graph}}(G, H)$, and is the quantity $\frac{d_{\text{graph}}(G, H)}{n}$.

In the above definition, we require that the graphs $G[\bar{S}]$ and $H[\bar{S}]$ be identical and not just isomorphic. Lemma 2.2 describes several equivalent characterizations of graph distance.

Proposition 2.2 (Alternate characterizations of d_{graph}). *Suppose G and H are graphs on the same vertex set. Then*

1. $d_{\text{graph}}(G, H)$ is the minimum vertex cover size of $G \Delta H$.
2. $d_{\text{graph}}(G, H)$ is the minimum number of vertices whose neighborhoods you need to edit to transform G into H .
3. $d_{\text{graph}}(G, H)$ is the minimum number of vertices whose neighborhoods you need to edit to transform $G \Delta H$ into the empty graph.

Note that d_{graph} is a metric (see [YY19], Lemma 5).

Definition 2.3 (Graph code). We say that a set $C \subseteq 2^{\binom{[n]}{2}}$ is a *graph code* on $[n]$ with distance d if for every pair of graph $G, H \in C$, we have that $d_{\text{graph}}(G, H) \geq d$.

- The *rate* of C , denoted by R_C , is the quantity $\frac{\log_2(|C|)}{\binom{n}{2}} \geq \frac{2 \log_2(|C|)}{n^2}$.
- The *distance* (resp. *relative distance*) of C , denoted by d_C (resp. δ_C), is the quantity $\min_{\substack{G, H \in C \\ G \neq H}} d_{\text{graph}}(G, H)$ (resp. $\min_{\substack{G, H \in C \\ G \neq H}} \delta_{\text{graph}}(G, H)$).

Upper bound. As noted in [YY19], the Singleton bound can be used to obtain an upper bound on the rate of a graph code. We include a proof here for completeness.

Proposition 2.4. *Any graph code with relative distance δ has dimension at most $\binom{n(1-\delta)+1}{2}$.*

Proof. Consider any graph code C of relative distance δ . Let $A \subseteq [n]$ be any subset of at most $\delta n - 1$ vertices. For any two distinct $G_1, G_2 \in C$, we have that the graphs induced on the vertices outside A , $G_1[[n] \setminus A]$ and $G_2[[n] \setminus A]$, are different. Indeed, since otherwise, A is a vertex cover of $G_1 \Delta G_2$, contradicting the relative distance assumption. So, we have that

$$|C| \leq 2^{\binom{n(1-\delta)+1}{2}}.$$

□

Expressed in terms of rate and distance, Proposition 2.4 implies that for constant relative distance δ , $R \leq (1 - \delta)^2 + O(1/n)$.

3 Existence of optimal graph codes

As with other objects in the theory of error-correcting codes, the first question we seek to answer relates to the optimal rate-distance tradeoff.

In contrast to the Hamming world, we find that random linear codes meet the Singleton bound in the graph distance.

Proposition 3.1. *Let $\delta \in (0, 1)$. Then, there exists a linear graph code with distance greater than δ and dimension at least*

$$\max \left\{ \binom{n(1-\delta)}{2} - h_2(\delta)n - 2, 0 \right\}.$$

Proof. We only consider the case when $\binom{n(1-\delta)}{2} - h_2(\delta)n - 2 > 0$, and prove this by a probabilistic construction. Let $\mathbf{G}_{n,1/2}$ be the Erdős-Rényi random graph distribution where the vertices are $[n]$, and each of the $\binom{n}{2}$ possible edges are selected independently with probability $1/2$. Let $k = \binom{n(1-\delta)}{2} - h_2(\delta)n - 2$, and let G_1, \dots, G_k be graphs chosen independently from $\mathbf{G}_{n,1/2}$. Consider the \mathbb{F}_2 -linear space $C = \text{span}\{A_{G_1}, \dots, A_{G_k}\}$. We wish to show that C has distance at least δn with high probability.

For $\vec{\alpha} \in \mathbb{F}_2^k$, let $H_{\vec{\alpha}}$ be the graph with the adjacency matrix $\sum_{i=1}^k \alpha_i A_{G_i}$. Recalling the definition of graph distance, we need that for any distinct $\vec{\alpha}, \vec{\beta} \in \mathbb{F}_2^k$, $H_{\vec{\alpha}} \Delta H_{\vec{\beta}}$ must have minimum vertex cover size at greater than δn . Since $H_{\vec{\alpha}} \Delta H_{\vec{\beta}} = H_{\vec{\alpha}-\vec{\beta}}$, it suffices to show that for any non-zero $\vec{\alpha} \in \mathbb{F}_2^k$, that $H_{\vec{\alpha}}$ has no vertex cover of size δn .

For any $\vec{\alpha} \in \mathbb{F}_2^k \setminus \{\vec{0}\}$, $H_{\vec{\alpha}}$ has the same law as $\mathbf{G}_{n, \frac{1}{2}}$. Let $B_{\vec{\alpha}}$ be the event that $H_{\vec{\alpha}}$ has a vertex cover of size δn . We have

$$\begin{aligned} \Pr(B_{\vec{\alpha}}) &= \Pr \left(\exists S \in \binom{[n]}{\delta n} : G[[n] \setminus S] = \text{is the empty graph} \right) \\ &\leq \binom{n}{\delta n} \cdot 2^{-\binom{n(1-\delta)}{2}} \\ &\leq 2^{-(\binom{n(1-\delta)}{2} + h_2(\delta)n)}, \end{aligned}$$

where the first inequality uses the union bound over subsets of size δn , and the fact that there are $\binom{n(1-\delta)}{2}$ edges outside of a vertex cover that all have to be unselected. Then, union bounding over the choices of $\vec{\alpha}$, we get

$$\Pr \left(\bigcup_{\alpha \in \mathbb{F}_2^k \setminus \{\vec{0}\}} B_{\vec{\alpha}} \right) \leq 2^{k - (\binom{n(1-\delta)}{2} + h_2(\delta)n)} \leq 1/4.$$

Therefore, with probability at least $3/4$, there does not exist $\alpha \in \mathbb{F}_2^k$, such that $\alpha \neq \vec{0}$, and H_{α} has a vertex cover of size δn , and hence C is a code with relative distance greater than δ .

Finally, note that if H_{α} does not have a vertex cover of size δn , then H_{α} is not the empty graph. Thus, the fact that there does not exist $\alpha \in \mathbb{F}_2^k$, $\alpha \neq \vec{0}$ such that H_{α} has a vertex cover of size δn implies that G_1, \dots, G_k are linearly independent. Hence, the dimension of C is $k = \binom{n(1-\delta)}{2} - h_2(\delta)n - 2$, as required. \square

As a result, we have the following corollaries.

Corollary 3.2. *For any constant $\delta \in (0, 1)$, there exist optimal linear graph codes with relative distance at least δ .*

Corollary 3.3. *For any constant $c > 2$, there exists a linear graph code with dimension at least $\Omega(\log^2 n)$ and relative distance at least $\delta = 1 - c \cdot \frac{\log n}{n}$.*

4 Explicit graph codes for high distance: Concatenated Codes

To get explicit codes of distance $\delta > 1/2$, we start with Construction 3 of [YY19]. The construction utilizes the tensor product code introduced by [Vol65], where elements of the code are matrices where all rows and columns are codewords over some base Hamming code. The elements of the tensor code are then the adjacency matrices of graphs. Since we consider undirected graphs and do not allow self-loops, we take the subcode of the tensor code containing only symmetric matrices with zeros on the diagonal.

Definition 4.1 (Symmetric Tensor Code with Zeros on the Diagonal). Let C be a code over \mathbb{F}_q . The symmetric tensor code with zeros on the diagonal built on C denoted $\text{STCZD}(C)$ is the set of matrices A over $\mathbb{F}_q^{n \times n}$ such that (1) A is symmetric, (2) the rows and columns of A are codewords of C , and (3) the entries on the diagonal are all 0.

Properties of elements of Tensor Product Codes that are symmetric and zero-diagonal were also previously studied in the context of constructing a gap-preserving reduction from SAT to the Minimum Distance of a Code problem by Austrin and Khot [AK14].

We will also define another notion of distance that will be useful later.

Definition 4.2 (Directed graph distance). Let A , and B be $n \times n$ matrices over some field. Define the *directed graph distance* denoted $d_{\text{directed}}(G, H)$ to be the minimum d such that there exists sets $S, T \subseteq [n]$ of size d where $(A - B)_{\bar{S}, \bar{T}} = \mathbf{0}$.

For weighted, directed graphs, G , and H , abbreviate $d_{\text{directed}}(A_G, A_H) = d_{\text{directed}}(G, H)$. To better distinguish between d_{directed} and d_{graph} , we sometimes refer to d_{graph} as the *undirected* graph distance.

When G and H are weighted directed graphs, $d_{\text{directed}}(A_G, A_H)$ can be viewed as the minimum d such that you can go from G to H by editing the incoming edges of d vertices and the outgoing edges of d vertices. The main difference between directed and undirected graph distance is that directed graph distance allows the subset of rows and the subset of columns to be edited to be different. Insisting that $S = T$ in the definition for directed graph distance recovers the undirected graph distance. From this, it easily follows that if G and H are undirected graphs, then $d_{\text{directed}}(G, H) \leq d_{\text{graph}}(G, H)$. Thus, to find codes with high graph distance, it suffices to find codes with large directed graph distance, where all the elements are adjacency matrices of undirected, unweighted graphs (i.e., 0/1 matrices that are symmetric and zero diagonal). Note that when discussing rate directed graph codes C , we are referring to the quantity $\log_q(|C|)/n^2$ instead of $\log_q(|C|)/\binom{n}{2}$.

In the next lemma, we show several properties of $\text{STCZD}(C)$. Most importantly, the Hamming distance of C translates to the directed graph distance of $\text{STCZD}(C)$.

Lemma 4.3. *Let C be a linear $[n, k, d]_q$ -code, then $\text{STCZD}(C) \subseteq \mathbb{F}_q^{n \times n}$ is linear, has dimension at least $\binom{k+1}{2} - n$, and has directed graph distance d .*

Proof. Let C be a linear $[n, k, d]_q$ -code, and let $C' = \text{STCZD}(C)$. C' is linear because C is linear, and the sum of symmetric matrices is symmetric.

WLOG, we assume that C is systematic, i.e., it has $k \times n$ generator matrix $G = [I|A]$, where I is the $k \times k$ identity and A is a $k \times (n - k)$ matrix. Then, for every $X \in \mathbb{F}_q^{k \times k}$, the following has rows and columns belonging to C

$$G^T X G = \begin{bmatrix} X & XA \\ A^T X & A^T X A \end{bmatrix}.$$

Furthermore, $G^T X G$ is symmetric and has zeros on the diagonal iff X is symmetric, X has zeros on the diagonal, and $A^T X A$ has zeros on the diagonal. This imposes $\binom{k+1}{2} + (n - k)$ linear constraints on the entries of X . Thus, the subspace of X for which $G^T X G \in C'$ has dimension at least $k^2 - \binom{k+1}{2} - (n - k) = \binom{k+1}{2} - n$.

Since C' is linear, to show the distance property, it suffices to show that $d_{\text{directed}}(A, \mathbf{0}) \geq d$ for every non-zero $A \in C'$. Let $A \in C'$ be a non-zero element of C' , we'll show that for any $S, T \subseteq [n]$, with $|S| < d$, and $|T| < d$, $A_{\overline{S}, \overline{T}} \neq \mathbf{0}$.

Since A is non-zero, there is some non-zero entry A_{ij} . Since the rows are elements of a linear code of distance d , the Hamming weight of the i th row is at least d . Since $|T| < d$, there is some $j' \notin T$ such that $A_{ij'}$ is non-zero. Then, the j' th column is also a non-zero codeword of C , so it has a Hamming weight of at least d . Since $|S| < d$, there is some $i' \notin S$ such that $A_{i'j'}$ is non-zero. Thus, $A_{\overline{S}, \overline{T}} \neq \mathbf{0}$. \square

Remark 1. A simple calculation shows that if C has constant rate, R , then $\text{STCZD}(C)$ has rate $R^2/2 - o(1)$ as a directed graph code.

Given this lemma (and using the fact that $d_{\text{graph}} \geq d_{\text{directed}}$), for any **binary** code $C \subseteq \mathbb{F}_2^n$ with rate R and relative distance δ , $\text{STCZD}(C)$ is a (undirected) graph code with rate $R^2 - o(1)$, and relative distance δ . Thus, if C has rate distance tradeoff $R = f(\delta)$, then $\text{STCZD}(C)$ has rate distance tradeoff $R = f(\delta)^2 - o(1)$. Immediately, we get that taking the STCZD of any asymptotically good binary code yields an asymptotically good graph code.

There are two problems with this construction. Firstly, these codes may not be strongly explicit. Secondly, the Plotkin bound [Pl60] implies that any binary code with distance $> 1/2$ has vanishing rate. So this falls short of our goal of obtaining strongly explicit, asymptotically good codes with $\delta > 1/2$.

We will address the first problem by showing that if the base code is a Reed Solomon code [RS60], then there is a large subcode that is strongly explicit.

Code 1 (Reed Solomon Code $\text{RS}(n, R, q)$ [RS60]). *The Reed Solomon Code with parameters n , R , and q , where $q \geq n$, is a code over \mathbb{F}_q^n with rate R and distance $1 - R$.*

Lemma 4.4. *Let $C \in \text{RS}(n, R, q)$ where $Rn = k - 1$. Then, there exists a strongly explicit subcode $S \subseteq \text{STCZD}(C)$ such that the dimension of S is at least $\binom{k-1}{2}$.*

Proof. Essentially, we will evaluate symmetric polynomials that are a multiple of $(X - Y)^2$ on a $n \times n$ grid.

Suppose $h(X, Y)$ is a symmetric polynomial of individual degree at most $k - 3$, and let M be the evaluations of $f(X, Y) = (X - Y)^2 h(X, Y)$ on a $n \times n$ grid. M is symmetric and has zeros on the diagonal. Furthermore, for a fixed value, y , $f(X, y)$ is a univariate polynomial in X of degree at most $k - 1$, and hence the column indexed by y is an element of a Reed Solomon code of dimension k , and block length n . Similarly, the rows are also elements of the same code. Thus $M \in \text{STCZD}$.

Let S be the space of bivariate symmetric polynomials of degree at most $k-3$. For $a, b \in \mathbb{N}$, define polynomials $p_{a,b}(X, Y) = X^a Y^b + X^b Y^a$. Notice that $p_{a,b}$ is symmetric, and furthermore the set

$$\{p_{a,b} : 0 \leq a < b \leq k-3\} \cup \{X^i Y^i : i \in \{0, 1, \dots, k-3\}\},$$

is linearly independent. Thus $\dim(S) = \binom{k-2}{2} + k-2 = \binom{k-1}{2}$, as desired. \square

To extend this construction to the setting of $\delta > 1/2$, we use the concatenation paradigm from standard error-correcting code theory, initially introduced by Forney [For65].

We will start with a code over a large alphabet and then concatenate with an inner code, which will be an optimal directed graph code.

Lemma 4.5. *For any $\epsilon > 0$, and sufficiently large n , for any $k < \epsilon^2 n^2 - 2n$, there exists a linear directed graph code over \mathbb{F}_2 of dimension k and distance at least $(1 - \epsilon)n$.*

The proof is standard and similar to that of Proposition 3.1. So we will omit it.

Code 2 (Optimal Directed Graph Code $\text{Opt}(\epsilon, n, k)$). *Require $k < \epsilon^2 n^2 - 2n$. Refer to a code with the properties in Lemma 4.5 as $\text{Opt}(\epsilon, n, k)$.*

4.1 Symmetric concatenation

Since our inner code is not guaranteed to be symmetric, simply replacing each field element in the outer code with its encoding might result in an asymmetric matrix. To remedy this, we transpose the encoding for entries below the diagonal. This is made formal below.

Definition 4.6 (Symmetric Concatenation). Let q, Q be prime powers, and n, N be positive integers. Let $C_{out} \subseteq \mathbb{F}_Q^{N \times N}$ and $C_{in} \subseteq \mathbb{F}_q^{n \times n}$ such that $|C_{in}| = Q$. Define $C_{in} \circ C_{out} \subseteq \mathbb{F}_q^{nN \times nN}$ to be the code obtained by taking codewords of C_{out} and replacing each symbol of the outer alphabet with by their encodings under C_{in} if they lie above or on the diagonal, and with the transpose of their encodings if they lie below the diagonal.

Figure 1 visualizes an example of symmetric concatenation. We now show that distance and dimension concatenate exactly like it does for standard error-correcting codes.

Lemma 4.7. *Suppose C_{in} and C_{out} are linear codes as in the previous definition with directed graph distance d and D , respectively. Let k be the dimension of C_{in} , and K be the dimension of C_{out} . Note $|C_{in}| = q^k = Q$. Then $C_{in} \circ C_{out}$ is linear and has distance at least dD , and dimension kK .*

Proof. Let $C = C_{in} \circ C_{out}$. First note that $C_{in} \circ C_{out}$ can be made linear by using a \mathbb{F}_q -linear map from \mathbb{F}_{q^k} to \mathbb{F}_q^k before encoding with the inner code.

Consider a non-zero outer codeword O , and let A be the codeword after concatenation. Let $S, T \subseteq [nN]$ be of size less than dD . We'll show that $A_{\overline{S}, \overline{T}} \neq \mathbf{0}$. Partition A into $N \times N$ blocks, where the (I, J) 'th block for $I, J \in [N]$, is the $n \times n$ matrix encoding the symbol at O_{IJ} . Identify the indices $[nN]$ with $[N] \times [n]$ where the tuple (I, i) corresponds to the i 'th index in the I 'th block.

For $I \in [N]$, let $S_I = \{i \in [n] : (I, i) \in S\}$ be the set rows in S in the I 'th block. Define T_J similarly. Let $S_{\geq} = \{I \in [N] : |S_I| \geq d\}$, be the set of blocks in which there are at least d elements in S , and similarly define T_{\geq} , $S_{<}$, and $T_{<}$.

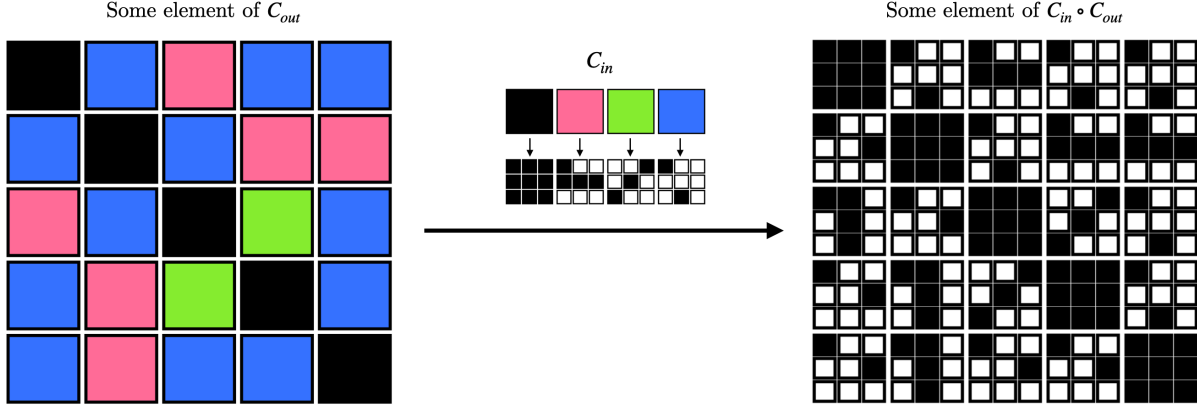


Figure 1: Example of symmetric concatenation. An outer codeword is shown on the left, with field elements represented as different colors. The concatenation with the inner code is shown to the right. Black squares represent 0, and white squares represent 1s.

Since $\sum_{I \in [N]} |S_I| < dD$, $\sum_{J \in [N]} |T_J| < dD$, we have $|S_{\geq}| < D$, and $|T_{\geq}| < D$. Since the outer code has directed distance D , $O_{S_{<}, T_{<}} \neq \mathbf{0}$, so there exists $I \in S_{<}$, and $J \in T_{<}$ such that $O_{I,J}$ is non-zero. So, the (I, J) 'th block of A is a non-zero codeword or the transpose of a non-zero codeword of C_{in} . Let us call it X , and suppose that $X \in C_{in}$.

Since $|S_I| < d$, and $|T_J| < d$, and the inner code has distance at least d , we have that $X_{\overline{S_I}, \overline{T_J}} \neq \mathbf{0}$. To finish the proof, note that $d_{\text{directed}}(X, \mathbf{0}) = d_{\text{directed}}(X^T, \mathbf{0})$ by switching the roles of S and T in the definition of directed graph distance.

For the claim about dimension, note that the number of codewords in C is the number of codewords in C_{out} , which is Q^K . The dimension of C is then $K \log_q(Q) = Kk$.

□

Additionally, it is clear from the definition of symmetric concatenation that if C_{out} is symmetric and zero-diagonal, so is $C_{in} \circ C_{out}$.

Remark 2. Lemma 4.7 also holds for the standard definition of concatenation (without transposing blocks below the diagonal). However, we will not need this fact.

4.2 Concatenated graph codes

We can instantiate the concatenated code using Reed Solomon codes.

Code 3 (Concatenated Code $C_{RS}(\epsilon, n, k, N, \rho)$). Let $Q = 2^k$ be the size of the alphabet of the outer code. Let $\epsilon, \rho \in (0, 1)$, and n, k, N , to be integers satisfying $k < \epsilon^2 n^2 - 2n$, and $N \leq Q$. Then,

$$C_{RS} = \text{Opt}(\epsilon, n, k) \circ \text{STCZD}(\text{RS}(N, \rho, Q)).$$

The following theorem follows directly from Lemmas 4.3 and 4.7. As a reminder, here we are considering the rate of the codes as a (undirected) graph code.

Theorem 4.8. *Let ϵ, n, k, N, ρ be parameters satisfying the requirements listed in Code 3, then $C_{\text{RS}}(\epsilon, n, k, N, \rho)$ is a graph code with rate $\epsilon^2 \rho^2 - o(1)$, and relative distance $(1 - \epsilon)(1 - \rho)$.*

Note that using this construction, we can get asymptotically good codes for any constant rate and distance - including for distances $> 1/2$, which was not obtained in any of the previous constructions. We get $R = (1 - \sqrt{\delta})^4 - o(1)$ by setting $\epsilon = \rho$.

One drawback of this construction is that it is not strongly explicit or even explicit. The outer code can be made strongly explicit using Lemma 4.4; however, the inner code was an optimal code, which we obtained by a randomized construction. The complexity of searching for such a code by brute force is too large. In particular, the optimal code has dimension $\epsilon^2 n^2$ and block length n^2 . Since we need the size of the code to be equal to the size of the outer alphabet, we have $N = 2^{\epsilon^2 n^2}$, so $n = \sqrt{\log(N)}/\epsilon$. Then, there are at least $2^{\epsilon^2 n^4} = 2^{\log(N)^2/\epsilon^2}$ generator matrices to search over. Thus, we cannot find such a code efficiently.

To address this, we reduce the search space by concatenating *multiple times*. The resulting code will have a slightly worse distance/rate tradeoff but will still be asymptotically good for any constant distance or rate.

We note that C_{RS} can also be made strongly explicit using a Justensen-like construction. However, although this code is again asymptotically good, it has distance bounded away from 1. We present this construction in the Appendix.

4.3 Multiple concatenation

While concatenating twice suffices to obtain an explicit code, it is not clear that the obtained code is strongly explicit. This may be addressed by concatenating three times at the cost of slightly weaker parameters. Here, we will also use the tensor product code as a building block. For any linear code $C \subseteq \mathbb{F}_q^n$, let $\text{TC}(C)$ be the tensor product code of C . As a reminder, $\text{TC}(C)$ is the code consisting of matrices $A \in \mathbb{F}_q^{n \times n}$ such that each row and each column of A are elements of C .

Remark 3. Suppose C is a linear code with distance d and rate R . Then, it follows from the proof of Lemma 4.3 that $\text{TC}(C)$ has directed graph distance at least d . It is also well known that $\text{TC}(C)$ has rate R^2 .

Below, we present the analysis for triple concatenation.

Code 4 (Triple Concatenation $C_{\text{Trip}}(\rho, N)$). *For $\rho \in (0, 1)$ and an integer N , let C be the subcode of $\text{STCZD}(\text{RS}(N, \rho, N))$ in Lemma 4.4. Then*

$$C_{\text{Trip}} = \text{Opt}(N_3, \rho) \circ \text{TC}(\text{RS}(N_2, \rho, N_2)) \circ \text{TC}(\text{RS}(N_1, \rho, N_1)) \circ C,$$

where N_1, N_2 and N_3 are picked to make the concatenation work, i.e., $|\text{Opt}(N_3, \rho)| \geq N_2$, and so on.

Notice that only the outer-most code needs to be symmetric and have zero diagonal since we use the symmetric concatenation operation (entries below the diagonal will be transposed). Thus, using the Tensor Product Code for the two middle codes (instead of STCZD) is sufficient and saves a factor of 2 (each time) on the rate.

Theorem 4.9. *Let N be a positive integer and $\rho \in (0, 1)$. Then $C_{\text{Trip}}(\rho, N)$ has distance at least $(1 - \rho)^4$, and rate ρ^8 . Furthermore, $C_{\text{Trip}}(\rho, N)$ is strongly explicit.*

Proof. The claims about rate and distance follow directly from Lemma 4.7.

We'll now show that this code is strongly explicit. The outermost code C is strongly explicit, and the two codes in the middle built on Reed Solomon codes are also strongly explicit. The idea is that the concatenation steps in the middle allow us to shrink the alphabet size from N to (less than) $\log(\log(N))$. Searching for optimal codes of this size can be done easily by brute force.

The dimension of $\text{TC}(\text{RS}(N_1, \rho, N_1))$ is $(\rho N_1)^2$, so the number of codewords is $N_1^{(\rho N_1)^2}$, and for the concatenation to work, we need this to be at least N . That is, we need $(\rho N_1)^2 \log(N_1) \geq \log(N)$, which we can get easily by setting $N_1 = O(\log(N))$. For the same reason, we can take $N_2 = O(\log \log(N))$.

This is now small enough to do a brute-force search for an optimal code. The inner-most code has dimension $\rho^2 N_3^2$, so we need $2^{\rho^2 N_3^2} = N_2$, or $N_3 = O(\sqrt{\log(N_2)})$. There are then $\rho^2 N_3^4$ possible generator matrices to search over. So the total number of codes we will need to search over is at most $2^{\rho^2 N_3^4} = 2^{O(\log \log \log(N)^2)} = 2^{o(\log \log(N))} = O(\log(N))$. \square

The tradeoff for this code is then

$$R = (1 - \delta^{1/4})^8.$$

Thus, we get *strongly explicit* asymptotically good codes for any constant distance or rate.

Concatenating twice would suffice if we just wanted explicit codes (instead of strongly explicit). In particular, the search space for the inner-most code has size

$$2^{O(\log \log(N)^2)} = 2^{o(\log(N))},$$

which is smaller than any polynomial in N but not polylogarithmic. The corresponding tradeoff for the double concatenated code is $R = (1 - \delta^{1/3})^6$.

5 Explicit graph codes with very high distance: dual-BCH Codes

In this section, we give explicit constructions of graph codes for the setting of very high distance ($\delta = 1 - o(1)$). As noted earlier, when the complete graph and the empty graph are part of the code, this is a generalization of the problem of constructing explicit Ramsey graphs (i.e., graphs with no large clique or independent set), which corresponds to graph codes of size at least 3.

Our main result here is an explicit construction of a graph code with distance $1 - \frac{n^\epsilon}{n^{1/2}}$ and dimension $\Omega(n^\epsilon \log n)$, for all $\epsilon \in [0, 1/2)$.

Theorem 5.1. *For all d , there is a strongly explicit construction of a code with dimension $\Omega(d \log n)$ and distance $n - O(d\sqrt{n})$.*

In analogy with the situation for Hamming-distance codes, these are the dual-BCH codes of the graph-distance world.

5.1 Warmup: a graph code with dimension $\Omega(\log n)$

As a warmup, we first construct code with distance $1 - \frac{1}{n^\epsilon}$ with growing dimension.

Let $n = 2^t$. Let $\text{Tr} : \mathbb{F}_{2^t} \rightarrow \mathbb{F}_2$ denote the finite field trace map. Concretely, it is given by

$$\text{Tr}(x) = x + x^2 + x^4 + \dots + x^{2^i} + \dots + x^{2^{t-1}}$$

For each $\alpha \in \mathbb{F}_{2^t}$, consider the matrix $M_\alpha \in \mathbb{F}_2^{n \times n}$, where the rows and columns of M_α are indexed by elements of \mathbb{F}_{2^t} , given by:

$$(M_\alpha)_{x,y} = \text{Tr}(\alpha \cdot (x+y)^3).$$

Note that each M_α is symmetric. Consider the code

Code 5. For n of the form 2^t , let us define the family of codes

$$C_{\text{Warmup}} = \{M_\alpha \mid \alpha \in \mathbb{F}_{2^t}\}.$$

We have that C_{Warmup} is a linear code of dimension $t = \log_2 n$.

Theorem 5.2. The distance of C_{Warmup} is at least $1 - O(n^{-1/2})$.

Proof. Fix any $\alpha \in \mathbb{F}_{2^t}$. Let $S \subseteq \mathbb{F}_{2^t}$ be an arbitrary subset of vertices. It suffices to show that if S is bigger than $\Omega(n^{1/2}) = \Omega(2^{t/2})$, then there exist some $x, y \in S$ such that

$$\text{Tr}(\alpha \cdot (x+y)^3) = 1.$$

Suppose not. Then we have:

$$\sum_{x,y \in S} (-1)^{\text{Tr}(\alpha(x+y)^3)} = |S|^2.$$

By Cauchy-Schwarz, we get:

$$\begin{aligned} |S|^4 &= \left(\sum_{x \in S} \sum_{y \in S} (-1)^{\text{Tr}(\alpha(x+y)^3)} \right)^2 \\ &\leq \left(\sum_{x \in S} \left(\sum_{y \in S} (-1)^{\text{Tr}(\alpha(x+y)^3)} \right)^2 \right) \cdot |S| \\ &\leq \left(\sum_{x \in \mathbb{F}_{2^t}} \left(\sum_{y \in S} (-1)^{\text{Tr}(\alpha(x+y)^3)} \right)^2 \right) \cdot |S| \\ &= \left(\sum_{x \in \mathbb{F}_{2^t}} \sum_{y_1, y_2 \in S} (-1)^{\text{Tr}(\alpha((x+y_1)^3 + (x+y_2)^3))} \right) \cdot |S| \\ &\leq \left(\sum_{y_1, y_2} \left| \sum_{x \in \mathbb{F}_{2^t}} (-1)^{\text{Tr}(\alpha((x+y_1)^3 + (x+y_2)^3))} \right| \right) \cdot |S|. \end{aligned}$$

For $y_1, y_2 \in \mathbb{F}_{2^t}$, let $P_{y_1, y_2}(X)$ be the polynomial

$$\begin{aligned} P_{y_1, y_2}(X) &= \alpha \cdot ((X + y_1)^3 + (X + y_2)^3) \\ &= \alpha \cdot ((y_1 + y_2)X^2 + (y_1^2 + y_2^2)X + (y_1^3 + y_2^3)). \end{aligned}$$

The key observation is that for most $(y_1, y_2) \in S^2$, the trace of the polynomial $P_{y_1, y_2}(X)$ is a nonconstant \mathbb{F}_2 -linear function, and thus the inner sum:

$$\sum_{x \in \mathbb{F}_{2^t}} (-1)^{\text{Tr}(P_{y_1, y_2}(x))}$$

equals 0.

Lemma 5.3. *If $P(X) = aX^2 + bX + c \in \mathbb{F}_{2^t}[X]$, then*

$$\text{Tr} \circ P : \mathbb{F}_{2^t} \rightarrow \mathbb{F}_2$$

is a nonconstant \mathbb{F}_2 -linear function unless $a = b^2$.

The proof is standard, and we omit it.

By the lemma, we get that there are at most $4|S|$ choices of (y_1, y_2) such that the inner sum is non-zero (namely those $(y_1, y_2) \in S^2$ for which $\alpha(y_1 + y_2) = (\alpha(y_1^2 + y_2^2))^2$, which are few in number by the Schwartz-Zippel lemma).

Thus we get:

$$|S|^4 \leq 4|S|^2 \cdot 2^t,$$

from which we get $|S| \leq O(2^{t/2})$, as desired. \square

5.2 Larger dimension

We now see how to get graph codes of distance $1 - \frac{1}{n^\epsilon}$ with $\epsilon < \frac{1}{2}$ and larger rate.

For a polynomial $f(X) \in \mathbb{F}_{2^t}[X]$, let M_f be $n \times n$ matrix with rows and columns indexed by \mathbb{F}_{2^t} for which:

$$(M_f)_{x,y} = \text{Tr}(f(x+y)).$$

Let W_d be the \mathbb{F}_{2^t} -linear space of all polynomials $f(X)$ of the form:

$$f(X) = \sum_{3 \leq 2i+1 \leq d} \alpha_i X^{2i+1},$$

where the $\alpha_i \in \mathbb{F}_{2^t}$.

Then, let us define our construction.

Code 6. *For n of the form 2^t and $d \leq n^{1/2}$, let us define the family of codes*

$$C_{\text{DualBCH}}(n, d) = \{M_f : f \in W_d\}.$$

Theorem 5.4. *We have that $C_{\text{DualBCH}}(n, d)$ is a linear graph code of distance $1 - O(dn^{-1/2})$ and dimension $\Omega(dn) = \Omega(d \log n)$.*

Proof. The proof is very similar to the proof of Theorem 5.2. Consider any $M_f \in C_{\text{DualBCH}}(n, d)$ with $f \neq 0$. It suffices to show that the independence number³ of M_f is $O(dn^{1/2})$.

Assume that $S \subseteq \mathbb{F}_{2^t}$ is an independent set in M_f . Then

$$|S|^2 = \sum_{x,y \in S} (-1)^{\text{Tr}(f(x+y))}. \quad (3)$$

³An essentially identical proof shows that the clique number also has the same bound. The only change is to replace the LHS of (3) by $-(|S|^2 - |S|)$, and this sign change does not affect anything later because we immediately apply Cauchy-Schwarz to get (4).

This justifies our referring to this code as a “code of Ramsey graphs”.

As in the proof of Theorem 5.2, by the Cauchy-Schwartz inequality and some simple manipulations, we get:

$$|S|^4 \leq \left(\sum_{y_1, y_2 \in S} \left| \sum_{x \in \mathbb{F}_{2^t}} (-1)^{\text{Tr}(P_{y_1, y_2}(x))} \right| \right) \cdot |S|, \quad (4)$$

where:

$$P_{y_1, y_2}(X) = f(X + y_1) - f(X + y_2).$$

At this point, we need an upper bound in the inner sum:

$$U_{y_1, y_2} = \left| \sum_{x \in \mathbb{F}_{2^t}} (-1)^{\text{Tr}(P_{y_1, y_2}(x))} \right|.$$

To get this, we will invoke the deep and powerful Weil bound:

Theorem 5.5 ([Sch76], Chapter II, Theorem 2E). *Suppose $P(X) \in \mathbb{F}_{2^t}(X)$ is a nonzero polynomial of odd degree with degree at most d . Then:*

$$\left| \sum_{x \in \mathbb{F}_{2^t}} (-1)^{\text{Tr}(P(x))} \right| \leq O(d2^{t/2}).$$

We will use this to show that all but a few pairs $(y_1, y_2) \in S^2$, U_{y_1, y_2} are small.

Lemma 5.6. *For all but $d|S|$ pairs $(y_1, y_2) \in S^2$,*

$$U_{y_1, y_2} \leq O(d2^{t/2}).$$

is at most $d|S|$.

Assuming this for the moment, we can proceed with Equation (4):

$$\begin{aligned} |S|^4 &\leq \left(d|S| \cdot 2^t + |S|^2 \cdot O(d \cdot 2^{t/2}) \right) \cdot |S| \\ &= d|S|^2 2^t + O(d|S|^3 2^{t/2}). \end{aligned}$$

Thus:

$$|S|^2 \leq d2^t + O(d|S|2^{t/2}),$$

which implies that $|S| \leq O(d \cdot 2^{t/2})$, as desired.

□

Proof of Lemma 5.6

Proof. Theorem 5.5 only applies to polynomials with odd degree. We first recall a standard trick involving the Tr map to deduce consequences for arbitrary degree polynomials.

Note that $\text{Tr}(a^2) = \text{Tr}(a)$ for all $a \in \mathbb{F}_{2^t}$, and that every element of \mathbb{F}_{2^t} has a square root. Thus for any positive degree monomial $M(X) = \alpha X^i$, where $i = j \cdot 2^k$ with j odd, the equality:

$$\text{Tr}(M(x)) = \text{Tr}(\widetilde{M}(x))$$

for each $x \in \mathbb{F}_{2^t}$, where $\widetilde{M}(X)$ is the odd degree monomial given by:

$$\widetilde{M}(X) = \alpha^{1/2^k} X^j.$$

Extending by linearity, this allows us to associate, to every polynomial $P(X) \in \mathbb{F}_{2^t}[X]$, a polynomial $\widetilde{P}(X)$ with

$$\text{Tr}(P(x)) = \text{Tr}(\widetilde{P}(x))$$

for all $x \in \mathbb{F}_{2^t}$, and where every monomial of $\widetilde{P}(X)$ (except possibly the constant term) has odd degree.

The key observation is that whenever $\widetilde{P}_{y_1, y_2}(X)$ is nonconstant, it has odd degree, and so we can apply the Weil bound. In this case, since:

$$\text{Tr}(P_{y_1, y_2}(x)) = \text{Tr}(\widetilde{P}_{y_1, y_2}(x))$$

for each $x \in \mathbb{F}_{2^t}$, we get:

$$U_{y_1, y_2} = \left| \sum_{x \in \mathbb{F}_{2^t}} (-1)^{\text{Tr}(\widetilde{P}_{y_1, y_2}(x))} \right| \tag{5}$$

$$\leq O(d \cdot 2^{t/2}), \tag{6}$$

where the last step follows from the Weil bound (Theorem 5.5).

Thus, we simply need to show that there are at most $d|S|$ pairs $(y_1, y_2) \in S^2$ for which $\widetilde{P}_{y_1, y_2}(X)$ is a constant.

Suppose $f(X)$ has degree exactly $2e + 1$. Let α be the coefficient of X^{2e+1} in $f(X)$.

Define $\gamma_i(Y) \in \mathbb{F}_q[Y]$ by:

$$f(X + Y) = \sum_{j=0}^{2e+1} \gamma_j(Y) X^j.$$

Note that $\deg(\gamma_i(Y)) \leq 2e + 1 - i$. It is easy to check that $\gamma_{2e+1}(Y) = \alpha$ and $\gamma_{2d}(Y) = \alpha Y$.

Then we have:

$$\begin{aligned} P_{y_1, y_2}(X) &= f(X + y_1) - f(X + y_2) \\ &= \sum_{i \leq 2e} (\gamma_i(y_1) - \gamma_i(y_2)) X^i. \end{aligned}$$

Then, by definition,

$$\tilde{P}_{y_1, y_2}(X) = \sum_{\substack{j \leq 2e-1 \\ j \text{ odd}}} \left(\sum_{\substack{k \geq 0 \\ j2^k \leq 2e}} (\gamma_{j \cdot 2^k}(y_1) - \gamma_{j \cdot 2^k}(y_2))^{\frac{1}{2^k}} \right) X^j.$$

We are trying to show that this is nonconstant for most y_1, y_2 . We will do this by identifying a monomial of positive degree, which often has a nonzero coefficient. Let $e = j_0 \cdot 2^{k_0}$ with j odd. We will focus on the coefficient of X^{k_0} . It equals:

$$(\gamma_{2e}(y_1) - \gamma_{2e}(y_2))^{1/2^{k_0+1}} + \left(\sum_{0 \leq k \leq k_0} (\gamma_{j \cdot 2^k}(y_1) - \gamma_{j \cdot 2^k}(y_2))^{\frac{1}{2^k}} \right).$$

By linearity of the map $a \mapsto a^{1/2^k}$, this can be expressed in the form $Q(y_1^{1/2^{k_0+1}}, y_2^{1/2^{k_0+1}})$, where $Q(Z_1, Z_2)$ is a bivariate polynomial of degree at most $2e$. Furthermore, using the fact that $\gamma_{2e}(Y) = \alpha Y$, the homogeneous part of $Q(Z_1, Z_2)$ of degree 1 exactly equals:

$$\alpha^{1/2^{k_0+1}}(Z_1 - Z_2),$$

which is nonzero. Thus $Q(Z_1, Z_2)$ is a nonzero polynomial.

Thus by the Schwartz-Zippel lemma, for $T = \{y^{1/2^{k_0+1}} \mid y \in S\}$, there are at most $2e|T| \leq d|S|$ values of $(z_1, z_2) \in T^2$ such that $Q(z_1, z_2) = 0$. Thus there are at most $d|S|$ values of $(y_1, y_2) \in S^2$ for which the coefficient of X^{k_0} in $\tilde{P}_{y_1, y_2}(X)$ is 0. Whenever it is nonzero, Equation (6) bounding U_{y_1, y_2} applies, and we get the desired result. □

Acknowledgements

We are grateful to Mike Saks, Shubhangi Saraf, and Pat Devlin for valuable discussions. We thank Alessandro Neri for very valuable remarks and references about rank metric codes and their connection to error-correcting graph codes. We thank Lev Yohananov and Eitan Yaakobi for bringing [YY19; YEY20] to our attention.

References

- [Alo23a] Noga Alon. “Connectivity Graph-Codes”. In: *arXiv preprint arXiv:2308.07653* (2023) (cit. on p. 3).
- [Alo23b] Noga Alon. “Graph-codes”. In: *arXiv preprint arXiv:2301.13305* (2023) (cit. on p. 3).
- [Alo+23] Noga Alon, Anna Gujgiczer, János Körner, Aleksa Milojević, and Gábor Simonyi. “Structured codes of graphs”. In: *SIAM Journal on Discrete Mathematics* 37.1 (2023), pp. 379–403 (cit. on p. 3).
- [AK14] Per Austrin and Subhash Khot. “A Simple Deterministic Reduction for the Gap Minimum Distance of Code Problem”. In: *IEEE Transactions on Information Theory* 60.10 (Oct. 2014), pp. 6636–6645. ISSN: 1557-9654. DOI: [10.1109/TIT.2014.2340869](https://doi.org/10.1109/TIT.2014.2340869). URL: <https://ieeexplore.ieee.org/document/6868217> (visited on 11/10/2023) (cit. on p. 10).

- [CZ13] Shixi Chen and Shuigeng Zhou. “Recursive Mechanism: Towards Node Differential Privacy and Unrestricted Joins”. In: *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*. SIGMOD ’13. New York, NY, USA: Association for Computing Machinery, June 22, 2013, pp. 653–664. ISBN: 978-1-4503-2037-5. DOI: [10.1145/2463676.2465304](https://doi.org/10.1145/2463676.2465304). URL: <https://doi.org/10.1145/2463676.2465304> (visited on 07/18/2024) (cit. on p. 3).
- [DG75] Philippe Delsarte and Jean-Marie Goethals. “Alternating bilinear forms over $GF(q)$ ”. In: *Journal of Combinatorial Theory, Series A* 19.1 (1975), pp. 26–50 (cit. on p. 4).
- [Dev] Pat Devlin. Personal Communication (cit. on p. 5).
- [For65] G David Forney. *Concatenated Codes*. Tech. rep. Massachusetts Institute of Technology, Research Laboratory of Electronics, 1965 (cit. on p. 12).
- [GQ09] Rod Gow and Rachel Quinlan. “Galois extensions and subspaces of alternating bilinear forms with special rank properties”. In: *Linear algebra and its applications* 430.8-9 (2009), pp. 2212–2224 (cit. on p. 4).
- [Hay+09] Michael Hay, Chao Li, Gerome Miklau, and David Jensen. “Accurate Estimation of the Degree Distribution of Private Networks”. In: *2009 Ninth IEEE International Conference on Data Mining*. 2009 Ninth IEEE International Conference on Data Mining. Dec. 2009, pp. 169–178. DOI: [10.1109/ICDM.2009.11](https://doi.org/10.1109/ICDM.2009.11). URL: <https://ieeexplore.ieee.org/document/5360242> (visited on 07/18/2024) (cit. on p. 3).
- [JSW24] Palak Jain, Adam Smith, and Connor Wagaman. *Time-Aware Projections: Truly Node-Private Graph Statistics under Continual Observation*. Mar. 7, 2024. DOI: [10.48550/arXiv.2403.04630](https://doi.org/10.48550/arXiv.2403.04630). arXiv: [2403.04630 \[cs\]](https://arxiv.org/abs/2403.04630). URL: <http://arxiv.org/abs/2403.04630> (visited on 07/18/2024). Pre-published (cit. on p. 3).
- [Jus72] J. Justesen. “Class of Constructive Asymptotically Good Algebraic Codes”. In: *IEEE Transactions on Information Theory* 18.5 (Sept. 1972), pp. 652–656. ISSN: 1557-9654. DOI: [10.1109/TIT.1972.1054893](https://doi.org/10.1109/TIT.1972.1054893). URL: <https://ieeexplore.ieee.org/document/1054893> (visited on 11/11/2023) (cit. on p. 23).
- [Kas+13] Shiva Prasad Kasiviswanathan, Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. “Analyzing Graphs with Node Differential Privacy”. In: *Proceedings of the 10th Theory of Cryptography Conference on Theory of Cryptography*. TCC’13. Berlin, Heidelberg: Springer-Verlag, Mar. 3, 2013, pp. 457–476. ISBN: 978-3-642-36593-5. DOI: [10.1007/978-3-642-36594-2_26](https://doi.org/10.1007/978-3-642-36594-2_26). URL: https://doi.org/10.1007/978-3-642-36594-2_26 (visited on 07/18/2024) (cit. on p. 3).
- [Mas63] James L Massey. *Threshold Decoding*. Tech. rep. Massachusetts Institute of Technology, Research Laboratory of Electronics, 1963 (cit. on p. 23).
- [Ner] Alessandro Neri. Personal Communication (cit. on p. 4).
- [Plo60] M. Plotkin. “Binary Codes with Specified Minimum Distance”. In: *IRE Transactions on Information Theory* 6.4 (Sept. 1960), pp. 445–450. ISSN: 2168-2712. DOI: [10.1109/TIT.1960.1057584](https://doi.org/10.1109/TIT.1960.1057584). URL: <https://ieeexplore.ieee.org/document/1057584> (visited on 11/11/2023) (cit. on p. 11).
- [RS60] I. S. Reed and G. Solomon. “Polynomial Codes Over Certain Finite Fields”. In: *Journal of the Society for Industrial and Applied Mathematics* 8.2 (June 1960), pp. 300–304. ISSN: 0368-4245, 2168-3484. DOI: [10.1137/0108018](https://doi.org/10.1137/0108018). URL: <http://epubs.siam.org/doi/10.1137/0108018> (visited on 07/04/2022) (cit. on p. 11).
- [Sch15] Kai-Uwe Schmidt. “Symmetric Bilinear Forms over Finite Fields with Applications to Coding Theory”. In: *Journal of Algebraic Combinatorics* 42.2 (Sept. 1, 2015), pp. 635–670. ISSN: 1572-9192. DOI: [10.1007/s10801-015-0595-0](https://doi.org/10.1007/s10801-015-0595-0). URL: <https://doi.org/10.1007/s10801-015-0595-0> (visited on 07/16/2024) (cit. on p. 4).

- [Sch76] Wolfgang M. Schmidt. “Equations over Finite Fields: An Elementary Approach”. In: 1976. URL: <https://api.semanticscholar.org/CorpusID:119017517> (cit. on p. 18).
- [Wol65] J. Wolf. “On Codes Derivable from the Tensor Product of Check Matrices”. In: *IEEE Transactions on Information Theory* 11.2 (Apr. 1965), pp. 281–284. ISSN: 1557-9654. DOI: [10.1109/TIT.1965.1053771](https://doi.org/10.1109/TIT.1965.1053771) (cit. on p. 10).
- [YFY20] Lev Yohananov, Yuval Efron, and Eitan Yaakobi. “Double and Triple Node-Erasure-Correcting Codes Over Complete Graphs”. In: *IEEE Transactions on Information Theory* 66.7 (July 2020), pp. 4089–4103. ISSN: 1557-9654. DOI: [10.1109/TIT.2020.2971997](https://doi.org/10.1109/TIT.2020.2971997). URL: <https://ieeexplore.ieee.org/document/8985404/?arnumber=8985404> (visited on 07/16/2024) (cit. on pp. 3–5, 20).
- [YY19] Lev Yohananov and Eitan Yaakobi. “Codes for Graph Erasures”. In: *IEEE Transactions on Information Theory* 65.9 (Sept. 2019), pp. 5433–5453. ISSN: 1557-9654. DOI: [10.1109/TIT.2019.2910040](https://doi.org/10.1109/TIT.2019.2910040). URL: <https://ieeexplore.ieee.org/document/8685206/?arnumber=8685206> (visited on 07/16/2024) (cit. on pp. 3–5, 8, 10, 20).

6 Appendix

6.1 Justensen-like code

The construction in this example is inspired by the Justensen code [Jus72], which uses an ensemble of codes for the inner code instead of a single inner code. Justensen uses an ensemble known as the Wozencraft Ensemble [Mas63] with the following properties.

Theorem 6.1 (Wozencraft Ensemble). *For every large enough k , there exists codes $C^{(1)}, C^{(2)}, \dots, C^{(2^k-1)}$ over $\mathbb{F}_2^{2^k}$ with rate $1/2$, where $1 - \epsilon$ fraction of them have distance at least $H_2^{-1}(1/2 - \epsilon)$.*

Since our goal is graph distance, we use the STCZD operation to convert the Wozencraft Ensemble from codes over strings with good Hamming distance to codes of matrices with good graph distance.

Lemma 6.2 (Wozencraft Ensemble Modification). *For any $\epsilon > 0$, and large enough k , there exists codes $D^{(1)}, D^{(2)}, \dots, D^{(2^k-1)}$ over $\mathbb{F}_2^{2^k \times 2^k}$. View these as directed graph codes. Then these codes have rate $1/8$, and at least a $1 - \epsilon$ fraction of them have distance at least $H_2^{-1}(1/2 - \epsilon)$.*

Proof. Let $C^{(1)}, C^{(2)}, \dots, C^{(N)}$ be the Wozencraft Ensemble. For each $I \in [N]$, define $D^{(I)} = \text{STCZD}(C^{(I)})$. Note that each $D^{(I)}$ is a code over $\mathbb{F}_2^{2^k \times 2^k}$. Note that by lemma Lemma 4.3, each of the codes has rate $1/8$. Since the STCZD operation translates Hamming distance to directed graph distance, we also have the same guarantee as the original Wozencraft Ensemble - at least $(1 - \epsilon)$ fraction of the codes have distance at least $H_2^{-1}(1/2 - \epsilon)$. \square

Concatenating $\text{STCZD}(RS)$ with the modified Wozencraft Ensemble in a particular arrangement yields our next construction.

Code 7 (Justensen-like $C_{\text{Justensen}}(\epsilon, k, \rho)$). *Require $\epsilon, \rho \in (0, 1)$. Let $Q = 2^k$, and $N = 2^k - 1$.*

Let $D^{(1)}, D^{(2)}, \dots, D^{2^k-1}$ be the modified Wozencraft Ensemble Lemma 6.2.

Then $C_{\text{Justensen}}$ is the code where for each element of $A \in \text{STCZD}(RS(N, \rho, Q))$, for each $I, J \in [N]$, we replace the symbol at A_{IJ} with its encoding under $D^{(\min(I, J))}$. If $J < I$, we transpose the encoding (to keep the matrix symmetric).

Figure 2 shows where each inner code is applied.

$$\begin{bmatrix} - & D^{(1)} & D^{(1)} & D^{(1)} & D^{(1)} & D^{(1)} \\ D^{(1)T} & - & D^{(2)} & D^{(2)} & D^{(2)} & D^{(2)} \\ D^{(1)T} & D^{(2)T} & - & D^{(3)} & D^{(3)} & D^{(3)} \\ D^{(1)T} & D^{(2)T} & D^{(3)T} & - & D^{(4)} & D^{(4)} \\ D^{(1)T} & D^{(2)T} & D^{(3)T} & D^{(4)T} & - & D^{(5)} \\ D^{(1)T} & D^{(2)T} & D^{(3)T} & D^{(4)T} & D^{(5)T} & - \end{bmatrix}$$

Figure 2: Inner code arrangement for $C_{\text{Justensen}}$

Theorem 6.3. *For any $\epsilon, \rho \in (0, 1)$, and k , a sufficiently large integer, $C_{\text{Justensen}}(\epsilon, \rho, k)$ is a strongly explicit linear graph code with rate $\rho^2/8 - o(1)$, and distance at least $(1 - \rho - \epsilon)H^{-1}(1/2 - \epsilon)$.*

Proof. Let $N = 2^k - 1$ and $n = 2k$ be the side lengths of the inner and outer codes, respectively. First note that $C_{\text{Justensen}}$ is a linear graph code over $\mathbb{F}_2^{nN \times nN}$, since both the inner and outer codes are linear, and we can apply a \mathbb{F}_2 linear map from $\mathbb{F}_{2^k} \rightarrow \mathbb{F}_2^k$ before encoding with the inner code.

Rate. By Lemma 4.3, the outer code, $\text{STCZD}(\text{RS}(N, \rho, Q))$, has rate $\rho^2/2 - o(1)$, and by Lemma 6.2, the inner codes have rate $1/8 - o(1)$. Thus, the rate is $\rho^2/8 - o(1)$ as an undirected graph code.

Distance. Let O be a non-zero outer codeword. For convenience, let $d = H^{-1}(1/2 - \epsilon)$, and $n = 2k$ be the side length of the inner code. We claim the distance is at least $(1 - \rho - \epsilon)d$.

Call $I \in [N]$, bad if the distance of $D^{(I)} < d$, and good otherwise. Let $B \subseteq [N]$ be the subset of bad indices. By the guarantee of the Wozencraft ensemble, we know that $|B| < \epsilon N$. Since O_{IJ} gets encoded with $\min(I, J)$, if $I, J \notin B$, then O_{IJ} is encoded with an inner code of distance at least d .

Define S_I, T_J as in the proof of Lemma 4.7. Let $S_{\geq} = \{I : |S_I| \geq d_{in} \text{ and } I \notin B\}$. Similarly, define T_{\geq} . Then $|S_{\geq}|, |T_{\geq}| < (1 - \rho - \epsilon)N$. Then $|S_{\geq} \cup B|, |T_{\geq} \cup B| < (1 - \rho)N$. Since this is less than the outer distance of the code, we have that $O_{IJ} \neq 0$ for some $I \notin S_{\geq} \cup B$, and $J \notin T_{\geq} \cup B$. In other words, $|S_I| < d$, $|T_J| < d_{in}$, and O_{IJ} is encoded with a code of directed graph distance at least d . Thus, by the inner distance, there remains a non-zero element in the (I, J) th block outside of S_I , and T_J . \square