Graph-Codes

Noga Alon *

Abstract

The symmetric difference of two graphs G_1, G_2 on the same set of vertices $[n] = \{1, 2, \ldots, n\}$ is the graph on $[n]$ whose set of edges are all edges that belong to exactly one of the two graphs G_1, G_2. Let H be a fixed graph with an even (positive) number of edges, and let $D_H(n)$ denote the maximum possible cardinality of a family of graphs on $[n]$ containing no two members whose symmetric difference is a copy of H. Is it true that $D_H(n) = o(2^{\binom{n}{2}})$ for any such H? We discuss this problem, compute the value of $D_H(n)$ up to a constant factor for stars and matchings, and discuss several variants of the problem including ones that have been considered in earlier work.

1 Introduction

1.1 The problem

The *symmetric difference* of two graph $G_1 = (V, E_1)$ and $G_2 = (V, E_2)$ on the same set of vertices V is the graph $(V, E_1 \oplus E_2)$ where $E_1 \oplus E_2$ is the symmetric difference between E_1 and E_2, that is, the set of all edges that belong to exactly one of the two graphs. Put $V = [n] = \{1, 2, \ldots, n\}$ and let $\mathcal{H}$ be a family of graphs on the set of vertices $[n]$ which is closed under isomorphism. A collection of graphs $\mathcal{F}$ on $[n]$ is called an *$\mathcal{H}$-(graph)-code* if it contains no two members whose symmetric difference is a graph in $\mathcal{H}$. For the special case that $\mathcal{H}$ contains all copies of a single graph H on $[n]$ this is called an H-code. Here we are interested in the maximum possible cardinality of such codes for various families $\mathcal{H}$. Let $D_{\mathcal{H}}(n)$ denote this maximum, and let

$$d_{\mathcal{H}}(n) = \frac{D_{\mathcal{H}}(n)}{2^{\binom{n}{2}}}$$

denote the maximum possible fraction of the total number of graphs on $[n]$ in an $\mathcal{H}$-code. If $\mathcal{H}$ consists of all graphs isomorphic to one graph H, we denote $d_{\mathcal{H}}(n)$ by $d_H(n)$. Note that if $\mathcal{H}$ consists of all graphs with less than d edges, then $D_{\mathcal{H}}(n)$ is simply the maximum

*Princeton University, Princeton, NJ, USA and Tel Aviv University, Tel Aviv, Israel. Email: nalon@math.princeton.edu. Research supported in part by NSF grant DMS-2154082 and by USA-Israel BSF grant 2018267.

possible cardinality of a binary code of length $\binom{n}{2}$ and minimum distance at least d. This motivates the terminology "graph-codes" used here.

The case $\mathcal{H} = \mathcal{K}$ where $\mathcal{K}$ is the family of all cliques is of particular interest. This case is motivated by a conjecture of Gowers raised in his blog post [10] in 2009 and is discussed briefly in the comments of that blog. If $\mathcal{H}$ consists of all graphs with independence number at most 2, then $d_{\mathcal{H}}(n) \geq 1/8$ for all $n \geq 3$, as shown by the family of all graphs on $[n]$ containing a triangle on the set of vertices $\{1, 2, 3\}$. An interesting result of Ellis, Filmus and Friedgut [7], settling a conjecture of Simonovits and Sós, asserts that this is tight for all $n \geq 3$. The corresponding result, that $d_{\mathcal{H}'}(n) = 1/2^6$ for all $n \geq 4$, where $\mathcal{H}'$ is the family of all graphs with independence number at most 3, is proved in [4]. A more systematic study of the parameters $D_{\mathcal{H}}(n)$ and $d_{\mathcal{H}}(n)$ for various families of graphs $\mathcal{H}$ appears in the recent paper [1]. The families $\mathcal{H}$ considered in this work include the family of all disconnected graphs, the family of all graphs that are not 2-connected, the family of all non-Hamiltonian graphs and the family of all graphs that contain or do not contain a spanning star. Additional families studied are all graphs that contain an induced or non-induced copy of a fixed graph T, or all graphs that do not contain such a subgraph.

In this note we focus on the case that $\mathcal{H}$ consists of a single graph H and the case that $\mathcal{H}$ is the family of all cliques, or all cliques up to a prescribed size. Note that trivially, if every member of $\mathcal{H}$ has an odd number of edges then $d_{\mathcal{H}}(n) \geq \frac{1}{2}$ as the family of all graphs on $[n]$ with an even number of edges forms an $\mathcal{H}$-code.

This suggests the following intriguing question.

Question 1.1. *Let $\mathcal{H}$ be a family of graphs closed under isomorphism. Is it true that $d_{\mathcal{H}}(n)$ tends to 0 as n tends to infinity if and only if $\mathcal{H}$ contains a graph with an even number of edges ? Equivalently: is it true that for any fixed graph H with an even number of edges, $d_H(n)$ tends to 0 as n tends to infinity ?*

We also study the linear variant of these problems, where the $\mathcal{H}$-codes considered are restricted to linear subspaces, that is, to families of graphs on $[n]$ closed under symmetric difference.

1.2 Results

Recall that $\mathcal{K}$ is the family of all cliques. Let $\mathcal{K}(r)$ denote the set of all cliques on at most r vertices. Let $K_{1,t}$ denote the star with t edges and let M_t denote the matching of t edges.

Theorem 1.2. *For every positive integer k,*

$$d_{K_{1,2k}}(n) = \Theta_k(1/n^k) \quad \textit{and} \quad d_{M_{2k}}(n) = \Theta_k(1/n^k).$$

Proposition 1.3. *For every integer $r \geq 1$,*

$$d_{\mathcal{K}(4r+3)}(n) \geq \Omega(\frac{1}{n^r}).$$

Proposition 1.4. *For the family $\mathcal{K}$ of all cliques, $d_{\mathcal{K}}(n) \geq \frac{1}{2^{[n/2]}}$.*

Proposition 1.5. *Let H be a fixed graph obtained from two copies of a graph H' by identifying the vertices of an independent set of H'. Then*

$$d_H(n) \leq \frac{|V(H)|}{n} \quad \textit{for all} \quad n \geq |V(H)|.$$

In particular, $d_H(n)$ tends to 0 as n tends to infinity.

Remark: All lower bounds are proved by constructing linear codes, that is, families of graphs closed under symmetric difference. Using a simple Ramsey-theoretic argument it is not difficult to show that for any such linear code the maximum possible cardinality is at most a fraction $O(\log\log n/\log n)$ of all graphs on n vertices whenever the defining family contains a fixed graph with an even number of edges.

Since all lower bounds are obtained by what may be called linear graph-codes one can study this separately, as done for standard error correcting codes. For the family of all cliques $\mathcal{K}$ we get here an exact result (strengthening the assertion of Proposition 1.4).

Theorem 1.6. *For any $n \geq 2$, the minimum possible co-dimension of a linear space of graphs on n vertices that contains no member of $\mathcal{K}$ is exactly $[n/2]$.*

2 Proofs

2.1 Upper bounds

For a family of graphs $\mathcal{H}$ and an integer n, the Cayley graph $C(n,\mathcal{H})$ is the graph whose vertices are all graphs on the n vertices $[n]$, where two are adjacent iff their symmetric difference is a member of $\mathcal{H}$. This is clearly a Cayley graph over the elementary abelian 2-group Z_2^N with $N = \binom{n}{2}$. The function $D_{\mathcal{H}}(n)$ is just the independence number of this graph, $d_{\mathcal{H}}(n)$ is the so called independence ratio. Since the graph $C(n,\mathcal{H})$ is vertex transitive, its independence ratio is exactly the reciprocal of its fractional chromatic number. See, for example, [11] for some background about this notion. A simple property of the fractional chromatic number of a graph is that it is at least that of any subgraph of it. This implies that in order to prove an upper bound of β for the independence ratio of the Cayley graph above it suffices to exhibit a set S of vertices that contains no independent set of size larger than $\beta|S|$. This applies also to weighted sets of vertices, but we will not use weights here.

Proof of Proposition 1.5: Let $a+b$ denote the number of vertices of H' where b is the size of its independent set so that H is obtained from two copies of H' by identifying the

vertices in this independent set. Thus the number of vertices of H is $2a+b$. Consider the following set of $m = \lfloor (n-b)/a \rfloor$ copies of H' on subsets of the vertex set $[n]$. All of them contain the same independent set on the vertices $\{n-b+1, n-b+2, \ldots, n\}$, and the additional vertices of copy number i are the vertices $\{(i-1)a+1, (i-1)a+2, \ldots, ia\}$, where $1 \leq i \leq m$. Each of these copies can be viewed as a vertex of the Cayley graph $C = C(n, \{H\})$. Since the symmetric difference of every pair of such copies forms a copy of H, this set forms a clique of size m in C, implying that $d_H(n) \leq \frac{1}{m} \leq |V(H)|/n$. $\square$

The proofs of Theorem 1.2 for stars and for matchings are very similar. We describe the proof for stars and briefly mention the modification needed for matchings. The upper bound in Theorem 1.2 for the star $K_{1,2}$ is a special case of the result above (with H' being a single edge). The upper bound for any prime k can be proved using the following result of Frankl and Wilson.

Theorem 2.1 ([9]). *Let p be a prime, and let $a_0, a_1, \ldots, a_r$ be distinct residue classes modulo p. Let $\mathcal{F}$ be a family of subsets of $[n]$ and suppose that $|F| \equiv a_0 \bmod p$ for all $F \in \mathcal{F}$ and that for every two distinct $F_1, F_2 \in \mathcal{F}$, $|F_1 \cap F_2| \equiv a_i \bmod p$ for some $1 \leq i \leq r$. Then $|\mathcal{F}| \leq \sum_{i=0}^{r} \binom{n}{i}$.*

Suppose k is a prime, $n \geq 2k$ and consider the family $\mathcal{G}$ of all stars $K_{1,2k-1}$ with center 1 and $2k-1$ leaves among the vertices $\{2, 3, \ldots, n\}$. Thus $|\mathcal{G}| = \binom{n-1}{2k-1}$. If two such stars share exactly $k-1$ common leaves then their symmetric difference is a copy of $K_{1,2k}$. A subset of $\mathcal{G}$ which is independent in the Cayley graph $C(n, K_{1,2k})$ corresponds to a collection of subsets of the set $\{2, 3, \ldots, n\}$, each of size $2k-1$, where the intersection of no two of these subsets is of cardinality $k-1$. Therefore, each of these sets is of cardinality -1 modulo k and no intersection is of cardinality -1 modulo k. By the Frankl-Wilson Theorem (Theorem 2.1) the cardinality of such a family is at most $\sum_{i=0}^{k-1} \binom{n-1}{i}$. Therefore, for every prime k,

$$d_{K_{1,2k}}(n) \leq \frac{\sum_{i=0}^{k-1} \binom{n-1}{i}}{\binom{n-1}{2k-1}} \leq O_k(\frac{1}{n^k}).$$

In order to prove the upper bound for all k we need the following result of Frankl and Füredi.

Theorem 2.2 ([8]). *For every fixed positive integers $\ell > \ell_1 + \ell_2$ there exist $n_0 = n_0(\ell)$ and $d_\ell > 0$ so that for all $n > n_0$, if $\mathcal{F}$ is a family of ℓ-subsets of $[n]$ in which the intersection of each pair of distinct members is of cardinality either at least $\ell - \ell_1$ or strictly smaller than ℓ_2, then*

$$|\mathcal{F}| \leq d_\ell \cdot n^{\max\{\ell_1, \ell_2\}}.$$

Proof of Theorem 1.2, upper bound: The proof for stars is essentially identical to the one described above for prime k, using Theorem 2.2 instead of Theorem 2.1. Let

$\mathcal{G}$ be the family of all stars $K_{1,2k-1}$ with center 1 and $2k-1$ leaves among the vertices $\{2,3,\ldots,n\}$. Thus $|\mathcal{G}| = \binom{n-1}{2k-1}$. If two such stars share exactly $k-1$ common leaves then their symmetric difference is a copy of $K_{1,2k}$. Therefore, by Theorem 2.2 above with $\ell = 2k-1, \ell_1 = \ell_2 = k-1$, the maximum cardinality of a subset of $\mathcal{G}$ which is independent in the Cayley graph $C(n, K_{1,2k})$ is at most some $c_k(n-1)^{k-1}$ for all sufficiently large n. This supplies the required upper bound

$$\frac{c_k(n-1)^k}{|\mathcal{G}|} \leq O_k(\frac{1}{n^k}),$$

for $d_{K_{1,2k}}(n)$. The proof for matchings is similar, starting with the family of all subsets of cardinality $2k-1$ of a fixed matching of cardinality $\lfloor n/2 \rfloor$. The symmetric difference of any two matchings that share exactly $k-1$ common edges is a copy of M_{2k}. Thus the proof can proceed exactly as in the case of stars. □

2.2 Lower bounds

In order to lower bound the independence number of a Cayley graph $C = C(n, \mathcal{H})$ it suffices to upper bound its chromatic number. One way to do so is to assign to each edge e of the complete graph on $[n]$ a vector $v_e \in Z_2^r$ for some r, so that for every $H \in \mathcal{H}$, $\sum_{e \in E(H)} v_e \neq 0$, where the sum is computed in Z_2^r. Given these vectors, we can assign to each graph G on $[n]$ the color $\sum_{e \in E(G)} v_e$ (computed, of course, in Z_2^r). This is clearly a proper coloring of C by at most 2^r colors. Note that the matrix whose columns are the $\binom{n}{2}$ vectors v_e is the analogue of the parity-check matrix of a linear error correcting code in the traditional theory of codes, and the color defined above is the analogue of the syndrome of a word, see, e.g., [12] for more information about these basic notions. The same book contains also a discussion of BCH codes which are used in the proofs below.

Proof of Theorem 1.2, lower bound: For stars, it suffices to show that the chromatic number of the Cayley graph $C = C(n, K_{1,2k})$ is at most $O(n^k)$. Let s be the smallest integer so that $2^s - 1 \geq n$. As shown by the columns of the parity check matrix of a BCH-code with designed distance $2k+1$ there is a collection S of $2^s - 1$ binary vectors of length $r = ks$ so that no sum of at most $2k$ of them (in Z_2^{ks}) is the zero vector. Fix a numbering of these vectors and a proper edge-coloring c of K_n by n colors. For each edge e let v_e be the vector of S with number $c(e)$. This gives the desired lower bound for stars. For matchings we use essentially the same construction, starting with a (non-proper) edge coloring of K_n by n colors in which each color class forms a star. □

Proof of Proposition 1.3, lower bound: As in the previous proof, but the initial edge-coloring now is defined by $c(ij) = i$ for all $i < j$ and the binary vectors selected are taken from the columns of the parity check matrix of a code with designed distance $2r+2$.

These are binary vectors of length $rs+1$, where $s = \lceil \log_2 n \rceil$. Let U be the set of vertices of a clique of size at least 2 and at most $4r+3$. Then U contains at least 1 and at most $2r+1$ vertices i for which there is an odd number of vertices of U with index strictly larger than i. Therefore the sum of vectors corresponding to the edges of the clique on U is equal to a sum of at most $2r+1$ column vectors of the parity check matrix, which is nonzero. □

Proof of Proposition 1.4, lower bound: This follows from the construction in the proof of Theorem 1.6 described in the next section.

3 Linear graph-codes

Proof of Theorem 1.6: The theorem is equivalent to the statement that for all $n \geq 2$ the minimum possible $r = r(n)$ so that there are graphs $G_1, \ldots, G_r$ on the vertex set $[n]$ such that every clique on a subset of cardinality at least 2 of $[n]$ contains an odd number of edges of at least one graph G_i, is $r = [n/2]$. Indeed, the code is simply the set of all graphs that have an even number of edges of each G_i. Therefore, a graph belongs to the code if and only if the characteristic vector of the set of its edges is orthogonal (over Z_2) to the characteristic vectors of the sets of edges of all the graphs G_i. These r graphs can be chosen so that the above r vectors form a basis for the orthogonal complement of the code.

In order to show that the minimum possible value of r is $\lfloor n/2 \rfloor$ it clearly suffices to prove the upper bound for odd n (that imply the result for $n-1$) and the lower bound for even n (implying the result for $n+1$). The upper bound is described in what follows. Let $n \geq 3$ be odd. Split the numbers $[n-1] = \{1, 2, \ldots, n-1\}$ into the $(n-1)/2$ blocks $B_i = \{2i-1, 2i\}$ for $1 \leq i \leq (n-1)/2$. Let G_i be the graph consisting of all edges of the $n-2i$ triangles with a common base B_i on the vertices $B_i \cup \{j\}$ for $2i < j \leq n$. Our family of graphs is the set of these $(n-1)/2$ graphs G_i. Let K be an arbitrary clique on a subset A of at least 2 vertices in $[n]$. If A contains a full block B_i for some i, then it contains exactly $2x+1$ edges of G_i, where x is the cardinality of the intersection of A with $\{2i+1, 2i+2, \ldots, n\}$. As this is odd for all $x \geq 0$ we may assume that A contains no block B_i. In this case, let j be the second largest element in A (recall that $|A| \geq 2$). Clearly $j \leq n-1$, hence it is contained in one of the blocks B_i. But in this case G_i contains exactly one edge of the clique K, completing the proof of the upper bound. Note that it is simple to give additional constructions with the same properties as any set of graphs that spans the same subspace as the graphs above will do. In particular, we can replace one of the graphs G_i by the complete graph K_n, which is the sum of all graphs G_i.

To prove the lower bound assume n is even and let $G_1, \ldots G_{n/2-1}$ be a family of $n/2-1$ graphs on $[n]$. We have to show that there is a clique on at least 2 vertices containing an

even number of edges of each G_i. We show that in fact there is such a clique on an even number of vertices. To do so we apply the classical theorem of Chevalley and Warning (cf., e.g., [2] or [15]). Recall that it asserts that any system of polynomials with n variables over a finite field in which the number of variables exceeds the sum of the degrees, which admits a solution, must admit another one (in fact, the number of solutions is divisible by the characteristics). Associate each vertex i with a variable x_i over Z_2 and consider the following homogeneous system of polynomial equations over Z_2. For each graph G_s in our family,

$$\sum_{ij \in E(G_s)} x_i x_j = 0.$$

In addition, add the linear equation $\sum_{i=1}^{n} x_i = 0$.

The sum of the degrees of the polynomials here is $2(n/2-1)+1 = n-1$, which is smaller than the number of variables. Since the system is homogeneous it admits the trivial solution $x_i = 0$ for all i. Any other solution (which exists by the Chevalley Warning Theorem) gives a clique on the set of vertices $\{i : x_i = 1\}$ which is nonempty, of even cardinality, and contains an even number of edges (possibly zero) of each G_i. This establishes the lower bound and completes the proof of Theorem 1.6. □

4 Concluding remarks and open problems

- Question 1.1, which is equivalent to the problem of deciding whether or not for any fixed nonempty graph H with an even number of edges $d_H(n)$ tends to 0 as n tends to infinity, remains wide open.

  An interesting special case is whether or not $d_{K_4}(n) = o(1)$. It is also interesting to decide whether or not $d_{K_4}(n) \geq \frac{1}{n^{o(1)}}$. It is not difficult to show that the latter can be deduced from the existence of an edge coloring of K_n by $n^{o(1)}$ colors with no copy of K_4 in which every color appears an even number of times. Indeed, such a coloring together with the columns of the parity check matrix of a BCH code with designed distance 7 supplies the lower bound above using the reasoning in the proofs of some of the results here. We have learned recently from Zach Hunter and Dhruv Mubayi that such an edge coloring is described in [6], modifying the constructions in [13], [5]. Therefore $d_{K_4}(n) \geq \frac{1}{n^{o(1)}}$. Even more recently, Bennett, Heath and Zerbib [3] found a similar coloring for K_5, implying that $d_{K_5}(n) \geq \frac{1}{n^{o(1)}}$.

- Gowers conjectured in [10] that any family of a constant fraction of all graphs on $[n]$, where n is sufficiently large, contains two graphs G_1, G_2 such that G_2 is a subgraph of G_1 and the symmetric difference of the two graphs (that is, the set of all edges of G_1 that are not in G_2) forms a clique. This is clearly stronger than the conjecture that $d_{\mathcal{K}}(n)$ tends to 0 as n tends to infinity, which is also open. As explained in

[10] the question of Gowers can be viewed as the first unknown case of a polynomial version of the density Hales-Jewett Theorem.

- As mentioned in the remark following the statement of Proposition 1.5, it is not difficult to show that for every graph H with an even number of edges the maximum possible cardinality of a *linear* family of graphs on $[n]$ in which no symmetric difference is a copy of H, is $o(2^{\binom{n}{2}})$. As the proof applies Ramsey's Theorem, it provides very weak bounds. It will be interesting to establish tighter bounds for the linear case. Theorem 1.6 provides an example of a tight result of this form.

- The problem considered above can be extended to hypergraphs. More generally, it can be extended to other versions of problems about binary codes, where the coordinates of each codeword are indexed by the elements of some combinatorial structure, and the forbidden symmetric differences correspond to a prescribed family of substructures. Here is an example of a problem of this type. What is the maximum possible cardinality of a collection of binary vectors whose coordinates are indexed by the elements of the ordered set $[n]$, where no symmetric difference of two distinct members of the collection forms an interval of length which is a cube of an integer? The corresponding Cayley graph here has 2^n vertices, and it is triangle-free by Fermat's last Theorem for cubes. Its independence number, which is the answer to the question above, is $o(2^n)$. Indeed, this follows from the Furstenberg-Sárközy Theorem and its extensions [14], by considering the maximum possible cardinality of an independent set in the induced subgraph on the set of all vertices that are characteristic vectors of an interval $[i] = \{1, \ldots, i\}$ for $0 \leq i \leq n$.

- Some of the discussion here suggests the problem of determining or estimating the smallest number of colors in an edge coloring of K_n in which every copy of a given graph H (or every copy of any member of a prescribed family $\mathcal{H}$ of graphs) intersects at least one of the color classes by an odd number of edges. This appears to be an interesting variant of classical questions in Ramsey Theory and deserves further study.

Acknowledgment I would like to thank Zach Hunter and Dhruv Mubayi for helpful comments and in particular for telling me about [6].

References

[1] N. Alon, A. Gujgiczer, J. Körner, A. Milojević and G. Simonyi, Structured codes of graphs, SIAM J. Discrete Math., to appear.

[2] Z. I. Borevich and I. R. Shafarevich, *Number Theory*, Academic Press, New York, 1966.

[3] P. Bennett, E. Heath and S. Zerbib, Edge-coloring a graph G so that every copy of a graph H has an odd color class, arXiv:2307.01314, 2023.

[4] A. Berger and Y. Zhao, K_4-intersecting families of graphs, J. Combin. Theory Ser. B 163 (2023), 112–132.

[5] D. Conlon, J. Fox, C. Lee and B. Sudakov, The Erdős-Gyárfás problem on generalized Ramsey numbers, Proc. London Math. Soc. 110 (2015), 1–18.

[6] A. Cameron and E. Heath, New upper bounds for the Erdős-Gyárfás problem on generalized Ramsey numbers, Combinatorics, Probability, and Computing (2022), 1–14.

[7] D. Ellis, Y. Filmus and E. Friedgut, Triangle-intersecting families of graphs, Journal of the European Mathematical Society 14 (2012), No. 3, 841–885.

[8] P. Frankl and Z. Füredi, Forbidding just one intersection, J. Combin. Theory Ser. A 39 (1985), no. 2, 160–176.

[9] P. Frankl and R. M. Wilson, Intersection theorems with geometric consequences, Combinatorica 1 (1981), 357–368.

[10] W. T. Gowers, https://gowers.wordpress.com/2009/11/14/the-first-unknown-case-of-polynomial-dhj/

[11] Pavol Hell and Jaroslav Nešetřil, *Graphs and homomorphisms*, Oxford lecture series in mathematics and its applications 28, Oxford University Press 2004.

[12] F. MacWilliams and N. Sloane, *The Theory of Error-Correcting Codes*, I. North-Holland Mathematical Library, Vol. 16. North-Holland Publishing Co., Amsterdam-New York-Oxford (1977).

[13] D. Mubayi, Edge-coloring cliques with three colors on all 4-cliques, Combinatorica 18 (1998), no. 2, 293–296.

[14] A. Sárközy, On difference sets of sequences of integers. III. Acta Math. Acad. Sci. Hungar. 31 (1978), no. 3-4, 355–386.

[15] W. M. Schmidt, *Equations over Finite Fields, an Elementary Approach*, Springer Verlag Lecture Notes in Math., 1976.