Independent sets in the discrete hypercube

David Galvin*

January 9, 2019

Abstract

In this expository note we describe a proof due to A. Sapozhenko that the number of independent sets in the discrete d-dimensional hypercube Q_d is asymptotically $2\sqrt{e}2^{2^{d-1}}$ as d tends to infinity.

1 Introduction

The focus of this note is the discrete hypercube Q_d. This is the graph on vertex set $\{0,1\}^d$ with two strings adjacent if they differ on exactly one coordinate. It is a d-regular bipartite graph with bipartition classes $\mathcal{E}$ and $\mathcal{O}$, where $\mathcal{E}$ is the set of vertices with an even number of 1's. Note that $|\mathcal{E}| = |\mathcal{O}| = 2^{d-1}$. (For graph theory basics, see e.g. [2], [5]).

An *independent set* in Q_d is a set of vertices no two of which are adjacent. Write $\mathcal{I}(Q_d)$ for the set of independent sets in Q_d. A trivial lower bound on $|\mathcal{I}(Q_d)|$ is $2^{2^{d-1}}$: each of the $2^{2^{d-1}}$ subsets of $\mathcal{E}$ is an independent set. A beautiful result of Korshunov and Sapozhenko [8] asserts that this trivial bound is not far off the truth.

Theorem 1.1.

$$|\mathcal{I}(Q_d)| = 2\sqrt{e}(1+o(1))2^{2^{d-1}} \quad as \quad d \to \infty.$$

The purpose of this expository note is to describe Sapozhenko's simplification [11] of the proof of Theorem 1.1. The simplified proof of Theorem 1.1 depends on a technical lemma bounding the number of subsets of $\mathcal{E}$ of a given size whose neighbourhood in $\mathcal{O}$ is of a given size. This lemma, which appears in [10], is stated in (close to) full generality in Section 3 and its proof is given in Section 5. Section 2 establishes notation and gathers together in a single place all the tools that we need. Mostly these are in the form of references, but one important tool, a familiar isoperimetric inequality in the hypercube, is proven in full, since we are not aware of an explicit presentation of it in the literature in English. The proof of Theorem 1.1 appears in Section 4.

*Department of Mathematics, University of Notre Dame, Notre Dame IN 46556; dgalvin1@nd.edu.
Research supported in part by the Simons foundation.

To improve the trivial lower bound $|\mathcal{I}(Q_d)| \geq 2^{2^{d-1}}$ to that given by Theorem 1.1, we consider not just independent sets which are confined purely to either $\mathcal{E}$ or $\mathcal{O}$. It is easy to see that there are

$$2^{d-1}2^{2^{d-1}-d} = \frac{1}{2}2^{2^{d-1}}$$

independent sets that have just one vertex from $\mathcal{O}$, and more generally approximately

$$\frac{\left(\frac{1}{2}\right)^k}{k!}2^{2^{d-1}}$$

independent sets which have exactly k non-nearby vertices from $\mathcal{O}$, for small k (by "non-nearby" it is meant that there are no common neighbours between pairs of the vertices). Indeed, there are approximately $\binom{2^{d-1}}{k}$ ways to choose the k vertices from $\mathcal{O}$; these vertices together have a neighbourhood of size kd, so there are $2^{2^{d-1}-kd}$ extensions of the k vertices to an independent set. Summing over k from 0 to any $\omega(d)$ we obtain the $\sqrt{e}$ in the lower bound.

To motivate the upper bound, consider what happens when we count independent sets that have exactly two nearby vertices from $\mathcal{O}$, (i.e., two vertices with a common neighbour). There are approximately $d^2 2^d$ choices for this pair (as opposed to approximately 2^{2d} choices for a pair of vertices without a common neighbour), since once the first vertex has been chosen the second must come from the approximately d^2 vertices at distance two from the first. There are approximately $2^{2^{d-1}-2d}$ extensions of the pair to an independent set (roughly the same as the number of extensions in the case of the pair of vertices without a common neighbour). The critical point here is that a pair of vertices from $\mathcal{O}$ has a neighbourhood size of approximately $2d$, whether or not the vertices are nearby. This is because a pair of vertices in Q_d has at most two common neighbours. Thus we get an additional contribution of approximately $2^{2^{d-1}-d}$ to the count of independent sets from those sets with two nearby vertices from $\mathcal{O}$ (negligible compared to the addition contribution of approximately $\frac{1}{4}2^{2^{d-1}}$ to the count from those sets with two non-nearby vertices from $\mathcal{O}$). The main work in the upper bound is the correct extension of this observation to the observation that the only non-negligible contribution to the count is from independent sets that on one side consist of a set of vertices with non-overlapping neighbourhoods. This in turn amounts to showing that there is a negligible contribution from those independent sets which are "2-connected" on one side (i.e., are such that between any two vertices on one side, there is a path in the cube every second vertex of which passes through the independent set.) This, finally, entails proving the technical lemma bounding the number of subsets of $\mathcal{E}$ of a given size whose neighbourhood in $\mathcal{O}$ is of a given size.

2 Notation and tools

2.1 Notation

Let Σ be a d-regular bipartite graph on vertex set V with bipartition classes X and Y. For $A \subseteq V$ we write $N(A)$ for the set of vertices outside V that are neighbours of a vertex in A, and $N(u)$ for $N(\{u\})$. We write $\rho(u, v)$ for the length of the shortest u-v path in Σ.

We define the *closure* of A to be

$$[A] = \{v \in V : N(v) \subseteq N(A)\}$$

and say that A is *closed* if $[A] = A$. We say that $A \subseteq X$ is *small* if $[A] \leq |X|/2$.

We say that A is *k-linked* if for every $u, v \in A$ there is a sequence $u = u_0, u_1, \ldots, u_l = v$ in A with $\rho(u_i, u_{i+1}) \leq k$ for $i = 0, \ldots, l-1$. Note that if A is 2-linked, then so is $[A]$. For any k, a set A can be decomposed into its maximal k-linked subsets; we refer to these as the *k-components* of A.

For $u \in V$ and $A, B \subseteq V$ we write $\nabla(A)$ for the set of edges having one end in A and $\nabla(A, B)$ for the set of edges having one end in each of A, B; $N_B(u) = N(u) \cap B$, $N_B(A) = N(A) \cap B$ and $d_B(u) = |N_B(u)|$. Set $\rho(u, A) = \min_{w \in A}\{\rho(u, w)\}$.

Given $A \subseteq X$ we always write G for $N(A)$, a for $|[A]|$, g for $|G|$ and set $t = g - a$.

Throughout we use log for the base 2 logarithm. We do not track constants in the proofs; all implied constants in O and Ω notation are independent of d. We will always assume that d is sufficiently large to support our assertions.

2.2 Tools

The following easy lemma is from [10].

Lemma 2.1. *If A is k-linked, and $T \subseteq V$ is such that $\rho(u, T) \leq l$ for each $u \in A$ and $\rho(v, A) \leq l$ for each $v \in T$, then T is $(k + 2l)$-linked.*

We need a lemma that bounds the number of k-linked subsets of V. The infinite Δ-branching rooted tree contains precisely

$$\frac{\binom{\Delta n}{n}}{(\Delta - 1)n + 1}$$

rooted subtrees with n vertices (see e.g. Exercise 11 (p. 396) of [6]). This implies that the number of n-vertex subsets of V which contain a fixed vertex and induce a connected subgraph is at most $(e\Delta)^n$. We will use the following easy corollary which follows from the fact that a k-linked subset of Σ is connected in a graph with all degrees $O(d^{k+1})$.

Lemma 2.2. *For each fixed k, the number of k-linked subsets of V of size n which contain a fixed vertex is at most $2^{O(n \log d)}$.*

The next lemma is a special case of a fundamental result due to Lovász [9] and Stein [12]. For a bipartite graph Γ with bipartition $P \cup Q$, we say that $Q' \subseteq Q$ *covers* P if each $p \in P$ has a neighbour in Q.

Lemma 2.3. *If Γ as above satisfies $|N(p)| \geq a$ for each $p \in P$ and $|N(q)| \leq b$ for each $q \in Q$, then P is covered by some $Q' \subseteq Q$ with*

$$|Q'| \leq (|Q|/a)(1 + \ln b).$$

We also use a result concerning the sums of binomial coefficients which follows from the Chernoff bounds [4] (see also [3], p.11):

$$\sum_{i=0}^{[cN]} \binom{N}{i} \leq 2^{H(c)N} \qquad \text{for } c \leq \tfrac{1}{2}, \tag{1}$$

where $H(x) = -x \log x - (1-x)\log(1-x)$ is the usual binary entropy function and $[x]$ denotes the integer part of x. Also, throughout we will use (usually without comment) a simple observation about sums of binomial coefficients: if $k = o(n)$, we have

$$\begin{aligned} \sum_{i \leq k} \binom{n}{i} &\leq (1 + O(k/n)) \binom{n}{k} \\ &\leq (1 + O(k/n))(en/k)^k \\ &\leq \exp_2 \left\{ (1 + o(1)) k \log(n/k) \right\}. \end{aligned}$$

2.3 Isoperimetry in the cube

A *Hamming ball centered at* x_0 in Q_d is any set of vertices B satisfying

$$\{u \in V(Q_d) \ : \ \rho(u, x_0) \leq k\} \subseteq B \subseteq \{u \in V(Q_d) \ : \ \rho(u, x_0) \leq k+1\}$$

for some $k < d$, where ρ is the usual graph distance (which in this case coincides with the Hamming distance on $\{0,1\}^d$). An *even* (resp. *odd*) *Hamming ball* is a set of vertices of the form $B \cap \mathcal{E}$ (resp. $B \cap \mathcal{O}$) for some Hamming ball B. We use the following result of Körner and Wei [7]. (A similar isoperimetric bound of Bezrukov [1] would also suffice).

Lemma 2.4. *For every $C \subseteq \mathcal{E}$ (resp. $\mathcal{O}$) and $D \subseteq V(Q_d)$, there exists an even (resp. odd) Hamming ball C' and a set D' such that $|C'| = |C|$, $|D'| = |D|$ and $\rho(C', D') \geq \rho(C, D)$.*

The following is a well-known isoperimetric inequality in Q_d (see, e.g. [8, Lemma 1.3]).

Claim 2.5. *For $A \subseteq \mathcal{E}$ (or $\mathcal{O}$) with $|A| \leq M/2$ we have*

$$\frac{|N(A)| - |A|}{|N(A)|} = \Omega(1/\sqrt{d}).$$

Proof: Without loss of generality, we may assume that $A \subseteq \mathcal{E}$. Let such an A be given. Applying Lemma 2.4 with $C = A$ and $D = V(Q_d) \setminus (A \cup N(A))$, we find that there exists an even Hamming ball A' with $|A'| = |A|$ and $|N(A)| \geq |N(A')|$. So we may assume that A is an even Hamming ball.

We consider only the case where A is centered at an even vertex, without loss of generality $\underline{0} := \{0, \ldots, 0\}$, the other case being similar. In this case,

$$\{v \in \mathcal{E} \ : \ \rho(v, \underline{0}) \leq k\} \subseteq A \subseteq \{v \in \mathcal{E} \ : \ \rho(v, \underline{0}) \leq k+2\}$$

for some even $k \leq d/2$ (the bound on k coming from the fact that $|A| \leq M/2$). For each $0 \leq i \leq (k+2)/2$, set $B_i = A \cap \{v : \rho(v, \underline{0}) = 2i\}$, and $N^+(B_i) = N(B_i) \cap \{u : \rho(u, \underline{0}) = 2i+1\}$. It is clear that $N(A) = \cup_{0 \leq i \leq (k+2)/2} N^+(B_i)$ and that

$$\cup_{0 \leq i \leq k/2} N^+(B_i) = \{v \in \mathcal{O} \ : \ \rho(v, \underline{0}) \leq k+1\}.$$

Also, observe that for all i

$$\frac{|B_i|}{|N^+(B_i)|} \leq \frac{2i+1}{d-2i}, \tag{2}$$

from which it follows that

$$|N^+(B_{(k+2)/2})| - |B_{(k+2)/2}| \geq \frac{-10}{d-4} |N^+(B_{(k+2)/2})| \geq \frac{-20}{d} \binom{d}{k+3}.$$

Indeed, (2) is an equality except when $i = (k+2)/2$, in which case it follows from the fact that each vertex in $B_{(k+2)/2}$ has exactly $d - (k+2)$ neighbours in $N^+(B_{(k+2)/2})$, and each vertex in $N^+(B_{(k+2)/2})$ has at most $(k+2)+1$ neighbours in $B_{(k+2)/2}$.

We deal first with the case $k \leq d/4$. In this case, (2) gives

$$\frac{|B_i|}{|N^+(B_i)|} \leq \frac{2}{3}$$

and so $(|N(A)| - |A|)/|N(A)| \geq 1/3$.

For $d/4 < k \leq d/2$ and c any constant, we claim that

$$\sum_{i=0}^{k+c} \binom{d}{i} = O\left(\sqrt{d} \binom{d}{k+c}\right), \tag{3}$$

where the constant in the O depends on c. For $k + c \leq d/2$, this follows from (1):

$$\sum_{i=0}^{k+c} \binom{d}{i} \leq 2^{H(\frac{k+c}{d})d} = O\left(\sqrt{d} \binom{d}{k+c}\right),$$

the equality being an easy consequence of Stirling's approximation, while for $k + c \geq d/2$ we have

$$\sum_{i=0}^{k+c} \binom{d}{i} \leq 2^d = O\left(\sqrt{d} \binom{d}{k+c}\right),$$

once again from Stirling's approximation. For $d/4 < k \leq d/2$, we therefore have

$$\begin{aligned}
|N(A)| - |A| &= |N^+(B_{(k+2)/2})| - |B_{(k+2)/2}| + \sum_{i=0}^{k/2} \binom{d}{2i+1} - \binom{d}{2i} \\
&\geq \frac{-20}{d} \binom{d}{k+3} + \sum_{i=0}^{k/2} \binom{d-1}{2i} \frac{d(d-1)}{(2i+1)(d-2i)} \\
&\geq \frac{8(d-1)}{3(d+2)} \binom{d-1}{k} - \frac{20}{d} \binom{d}{k+3} \\
&= \Omega\left(\binom{d}{k+3}\right), \qquad (4)
\end{aligned}$$

and

$$\begin{aligned} |N(A)| &= |N^+(B_{(k+2)/2})| + \sum_{i=0}^{k/2} \binom{d}{2i+1} \\ &\leq \sum_{i=0}^{k+3} \binom{d}{i} = O\left(\sqrt{d}\binom{d}{k+3}\right), \end{aligned} \tag{5}$$

the last equality following from (3). Combining (4) and (5), we may now conclude that

$$\frac{|N(A)| - |A|}{|N(A)|} = \Omega\left(\frac{1}{\sqrt{d}}\right).$$

□

Considerably stronger results can be obtained if we impose more conditions on $|A|$. In that direction we need only the following simple lemma.

Lemma 2.6. *For $A \subseteq \mathcal{E}$ (or $\mathcal{O}$),*

$$\text{if } |A| < d^{O(1)}, \text{ then } |A| \leq O(1/d)|N(A)|.$$

Proof: If $|A| < d^{O(1)}$, then we have $k = O(1)$ in the notation of Claim 2.5, and repeating the argument of that lemma we get $|A| \leq O(1/d)|N(A)|$. □

3 Sapozhenko's graph lemma

Let Σ be a d-regular, bipartite graph with bipartition classes X and Y. The *co-degree* of Σ is the maximum of $|N(x) \cap N(y)|$ over all pairs $(x, y) \in Y \times Y$. For each a and g and $v \in \mathcal{O}$, set

$$\mathcal{G}(a, g, v) = \{A \subseteq X \text{ 2-linked} : |[A]| = a,\ |G| = g \text{ and } v \in G\}.$$

Lemma 3.1. *For each pair of constants $c, \Delta_2 > 0$ there is a constant $c' = c'(c, \Delta_2) > 0$ such that the following holds. If Σ is a d-regular bipartite graph with partition classes X and Y and with co-degree Δ_2, then for $g \geq d^4$ and*

$$t > \frac{cg \log^3 d}{d^2} \tag{6}$$

we have

$$|\mathcal{G}(a, g, v)| \leq 2^{g - \frac{c't}{\log d}}.$$

4 Proof of Theorem 1.1

The key observation is the following consequence of Lemma 3.1.

Corollary 4.1.

$$\sum_{A \subseteq \mathcal{E},\ A \text{ small}} 2^{-|N(A)|} = (1+o(1))\sqrt{e} \quad \text{as } d \to \infty.$$

Proof: The main point is the easy observation that if A has 2-components $A_1, \ldots, A_k$ then $|N(A)| = |N(A_1)| + \ldots + |N(A_k)|$. Armed with this, we have $\sum_{A \subseteq \mathcal{E},\ A \text{ small}} 2^{-|N(A)|}$

$$\begin{aligned}
&= \sum_k \sum_{A \subseteq \mathcal{E} \text{ small with 2-components } A_1, \ldots, A_k} 2^{-|N(A_1)|-\ldots-|N(A_k)|} \\
&\leq \sum_k \frac{1}{k!} \left(\sum_{A \subseteq \mathcal{E} \text{ small, 2-linked, } |A| \geq 1} 2^{-|N(A)|} \right)^k \\
&\leq \exp\left\{ \sum_{A \subseteq \mathcal{E} \text{ small, 2-linked, } |A| \geq 1} 2^{-|N(A)|} \right\} \\
&= \exp\left\{ \sum_{A \subseteq \mathcal{E},\ |A|=1} 2^{-d} + \sum_{A \subseteq \mathcal{E} \text{ small, 2-linked, } |A| \geq 2} 2^{-|N(A)|} \right\} \\
&= \exp\left\{ \frac{1}{2} + \sum_{A \subseteq \mathcal{E} \text{ small, 2-linked, } |A| \geq 2} 2^{-|N(A)|} \right\}.
\end{aligned}$$

The result will follow if we show that

$$\sum_{A \subseteq \mathcal{E} \text{ small, 2-linked, } |A| \geq 2} 2^{-|N(A)|} = o(1). \tag{7}$$

Say that an $A \subseteq \mathcal{E}$ with A small and 2-linked is *of type I* if $|N(A)| \leq d^4$ and of *of type II* otherwise. We consider the portions of the sum in (7) corresponding to types I and II A's separately. In each case we partition the set of A's according to the sizes of $[A]$ and $N(A)$ and the first vertex $v \in N(A)$ (in some fixed ordering of $\mathcal{O}$).

If A is of type I then by Lemma 2.6, $|[A]| \leq O(|N(A)|/d)$ and so we may bound

$$|\mathcal{G}(a,g,v)| \leq d 2^{O(a \log d)} 2^a = 2^{O(a \log d)} = 2^{O(g \log d/d)}. \tag{8}$$

The factor of d comes from choosing a neighbour of v to be a fixed vertex in $[A]$, the factor of $2^{(a \log d)}$ is from Lemma 2.2 and corresponds to a choice of $[A]$ (noting that $[A]$ is 2-linked) and the factor of 2^a comes from choosing A as a subset of $[A]$. We therefore have

$$\begin{aligned}
\sum_{A \text{ of type I}} 2^{-|N(A)|} &\leq \sum_{2 \leq a < g,\ g \leq d^4,\ v} |\mathcal{G}(a,g,v)| 2^{-g} \\
&\leq 2^d d^4 \sum_{g \geq d} 2^{O(g \log d/d) - g} \qquad (9) \\
&= e^{-\Omega(d)}. \qquad (10)
\end{aligned}$$

The factor of 2^d in (9) comes from choosing v while the factor of d^4 is for the choice of a.

For A of type II (summing only over those values of a, g and v with $\mathcal{G}(a,g,v) \neq \emptyset$ and with the inequalities justified below)

$$\begin{aligned}
\sum_{A \text{ of type II}} 2^{-|N(A)|} &= \sum_{a \le g,\ g \ge d^4,\ v} |\mathcal{G}(a,g,v)| 2^{-g} \\
&\le \sum_{a \le g,\ g \ge d^4,\ v} 2^{-\Omega\left(\frac{t}{\log d}\right)} \qquad (11) \\
&\le 2^{3d} 2^{-\Omega\left(\frac{d^{7/2}}{\log d}\right)} \qquad (12) \\
&= e^{-\Omega(d)}. \qquad (13)
\end{aligned}$$

Here (11) is from Lemma 3.1. To see that the application is valid, note that (6) is given by Claim 2.5, and that $\Delta_2(Q_d) = 2$. In (12) we use the fact that there are fewer than 2^d choices for each of a, g and v and we bound

$$t = g\left(\frac{g-a}{g}\right) \ge \Omega\left(\frac{g}{\sqrt{d}}\right) \ge \Omega(d^{7/2}),$$

the first inequality following from Claim 2.5 and the second from the fact that $g \ge d^4$.

Combining (10) and (13), we have (7). □

We may now swiftly put an upper bound on $|\mathcal{I}(Q_d)|$. Indeed, it is easy to check that for any $I \in \mathcal{I}(Q_d)$ we have that $[I \cap \mathcal{E}]$ and $[I \cap \mathcal{O}]$ have no edges between them, and so at least one of $I \cap \mathcal{E}$, $I \cap \mathcal{E}$ is small. Noting that once we have chosen a small $A \subseteq \mathcal{E}$ there are $2^{2^{d-1}-|N(A)|}$ ways to complete this choice to an independent set by selecting an arbitrary subset of the non-neighbours of A on $\mathcal{O}$, by symmetry it follows that

$$\begin{aligned}
|\mathcal{I}(Q_d)| &\le 2\sum_{A \subseteq \mathcal{E}} 2^{2^{d-1}-|N(A)|} \\
&= 2(1+o(1))\sqrt{e}\,2^{2^{d-1}},
\end{aligned}$$

the second inequality coming from Corollary 4.1.

To obtain the matching lower bound and complete the proof of Theorem 1.1, let $f(k)$ denote the number of subsets S of $\mathcal{E}$ of size k which satisfy the condition that $|N(S)| = kd$ (the maximum possible; achievable only if the elements of S have pairwise disjoint neighbourhoods). Noting that for each $v \in \mathcal{E}$ there are $\binom{d}{2}$ vertices at distance 2 from v (and exactly one at distance 0), for $k \le d$ (say) we have

$$\begin{aligned}
f(k) &\ge \frac{1}{k!}\prod_{j=0}^{k-1}\left(2^{d-1} - (j-1)\left(\binom{d}{2}+1\right)\right) \\
&\ge (1-o(d^{-1}))\frac{\left(2^{d-1}\right)^k}{k!} \qquad \text{as } d \to \infty.
\end{aligned}$$

We will get our lower bound by considering those independent sets which have k vertices with non-overlapping neighbourhoods on one side, and are arbitrary on the other side, for

$k \le d$. In the first inequality in this count, the final term of 2^{2d^2} upper bounds the overcount; the contribution from those sets which have at most d vertices from $\mathcal{E}$ and at most d from $\mathcal{O}$.

$$\begin{aligned}
|\mathcal{I}(Q_d)| &\geq 2\left(\sum_{k=0}^{d} f(k)2^{2^{d-1}-kd}\right) - 2^{2d^2} \\
&\geq 2^{2^{d-1}+1}\left(\sum_{k=0}^{d}(1-o(d^{-1}))\frac{\left(2^{d-1}\right)^k}{k!}(1-o(d^{-1}))2^{-kd}\right) - 2^{2d^2} \\
&\geq 2(1-o(1))2^{2^{d-1}}\sum_{k=0}^{d}\frac{\left(\frac{1}{2}\right)^k}{k!} \\
&\geq 2(1-o(1))\sqrt{e}2^{2^{d-1}}.
\end{aligned}$$

We have shown, as intended, that

$$|\mathcal{I}(Q_d)| \sim 2\sqrt{e}2^{2^{d-1}} \qquad \text{as } d \to \infty.$$

5 Proof of Lemma 3.1

We bound $|\mathcal{G}(a,g,v)|$ using two notions of "approximation". These are introduced in Section 5.1, and in this section we also state the three "approximation" lemmas that we will use to obtain the results discussed above. Section 5.2 gives the proof of Lemma 3.1, modulo the approximation lemmata, while Sections 5.3, 5.4 and 5.5 are then devoted to the proofs of the approximation lemmata.

5.1 Approximation

The first notion of approximation depends on a parameter φ, $1 \le \varphi \le d-1$. Set

$$G^{\varphi} = \{y \in G : d_{[A]}(y) > \varphi\}.$$

Definition 5.1. A *φ-approximation* for $A \subseteq X$ is an $F' \subseteq Y$ satisfying

$$G^{\varphi} \subseteq F' \subseteq G \tag{14}$$

and

$$N(F') \supseteq [A] \tag{15}$$

The second depends on a parameter ψ, $1 \le \psi \le d-1$.

Definition 5.2. A *ψ-approximation* for $A \subseteq X$ is a pair $(F,S) \in 2^Y \times 2^X$ satisfying

$$F \subseteq G, \quad S \supseteq [A], \tag{16}$$

$$d_F(u) \geq d - \psi \quad \forall u \in S \tag{17}$$

and

$$d_{X\setminus S}(v) \geq d - \psi \quad \forall v \in Y \setminus F. \tag{18}$$

Before continuing, we note a property of ψ-approximations that will be of use later.

Lemma 5.3. *If (F, S) is a ψ-approximation for $A \in \mathcal{G}$ then*

$$|S| \leq |F| + 2t\psi/(d - \psi). \tag{19}$$

Proof: Observe that $|\nabla(S, G)|$ is bounded above by $d|F| + \psi|G \setminus F|$ and below by $d|[A]| + (d - \psi)|S \setminus [A]| = d|S| - \psi|S \setminus [A]|$, giving

$$|S| \leq |F| + \psi|(G \setminus F) \cup (S \setminus [A])|/d,$$

and that each $u \in (G \setminus F) \cup (S \setminus [A])$ contributes at least $d - \psi$ edges to $\nabla(G, X \setminus [A])$, a set of size td, giving

$$|(G \setminus F) \cup (S \setminus A)| \leq 2td/(d - \psi).$$

These two observations together give (19). □

In what follows we write $\mathcal{G}$ for $\mathcal{G}(a, g, v)$. We will bound $|\mathcal{G}|$ by combining the following three lemmata.

Lemma 5.4. *For $g > d^4$ there is a family $\mathcal{V} = \mathcal{V}(\varphi) \subseteq 2^Y$ with*

$$|\mathcal{V}| \leq \begin{cases} 2^{O(g \log^2 d/(\varphi d)) + O(t \log^2 d/\varphi)} & \text{if } t < O(g(d - \varphi)/(\varphi d)) \text{ and} \\ 2^{O(t \log^2 d/(d - \varphi)) + O(t \log^2 d/\varphi)} & \text{if } t > \Omega(g(d - \varphi)/(\varphi d)) \end{cases} \tag{20}$$

such that each $A \in \mathcal{G}$ has a φ-approximation in $\mathcal{V}$.

Lemma 5.5. *For any $F' \in \mathcal{V}(\varphi)$ and $1 \leq \psi \leq d - 1$ there is a family $\mathcal{W} = \mathcal{W}(F', \varphi, \psi) \subseteq 2^Y \times 2^X$ with*

$$|\mathcal{W}| \leq 2^{O(td \log d/((d-\varphi)\psi)) + O(td \log d/((d-\psi)\psi))} \tag{21}$$

such that any $A \in \mathcal{G}$ for which F' is a φ-approximation has a ψ-approximation in $\mathcal{W}$.

Lemma 5.6. *Given $1 \leq \psi \leq d - 1$ and $\gamma > 0$, for each $(F, S) \in 2^Y \times 2^X$ that satisfies (19) there are at most*

$$\max\left\{2^{g-\gamma t}, 2^{g-t+O((t\psi/(d-\psi)+\gamma t)\log d)}\right\} \tag{22}$$

A's in $\mathcal{G}$ satisfying $F \subseteq G$ and $S \supseteq [A]$.

5.2 Derivation of Lemma 3.1

For t satisfying (6), we obtain

$$|\mathcal{G}| < 2^{g - \Omega(t/\log d)}$$

by taking $\gamma = c/\log d$, $\psi = c'd/\log d$ (for suitably chosen constants c, c') and, for example, $\varphi = d/2$.

5.3 Proof of the φ-approximation lemma

Our φ-approximation $F' = F'(A)$ for a particular $A \in \mathcal{G}$ will consist of three pieces. The first of these is $N(N_{[A]}(T_0))$ where T_0 is a small subset of G for which $N(N_{[A]}(T_0))$ contains most of G^φ and for which $\Omega := \nabla(T_0, X \setminus [A])$ is also small. (A suitably chosen random T_0 does both of these.) The second piece is $T_0' := G^\varphi \setminus N(N_{[A]}(T_0))$. Setting $L = N(N_{[A]}(T_0)) \cup T_0'$, we have $L \supseteq G^\varphi$. The final piece, T_1, is a small subset of $G \setminus L$ whose neigbourhood includes $[A] \setminus N(L)$ (we use Lemma 2.3 to bound $|T_1|$). Clearly $F' = N(N_{[A]}(T_0)) \cup T_0' \cup T_1$ is a φ-approximation for A. We then take $\mathcal{U}$ to be the collection of F''s that are produced in this way as we run over all possible $A \in \mathcal{G}$.

To control $|\mathcal{U}|$ we observe that in this procedure each F' is given by the quadruple (T_0, T_0', T_1, Ω), where $T := T_0 \cup T_0' \cup T_1$ is a small 8-linked subset of G and Ω is a small subset $\nabla(T_0)$. Lemma 2.2 bounds the number of possible T's, and direct calculations limit the number of choices for T_0, T_1 and Ω given T_0.

Fix $A \in \mathcal{G}$. Set $p = 20\Delta_2 \log d/(\varphi d)$.

Claim 5.7. *There is a $T_0 \subseteq G$ such that $v \in T_0$ and*

$$|T_0| \leq 4gp \tag{23}$$

$$|\nabla(T_0, X \setminus [A])| \leq 4tdp \tag{24}$$

and

$$|G^\varphi \setminus N(N_{[A]}(T_0))| \leq 3g/d^{10}. \tag{25}$$

Proof: Construct a random subset S of G by putting each $y \in G$ in S with probability p, these choices made independently. Clearly

$$\mathbf{E}(|S|) = gp \tag{26}$$

and since $|\nabla(G, X \setminus [A])| = td$,

$$\mathbf{E}(|\nabla(S, X \setminus [A])|) = tdp. \tag{27}$$

By the co-degree condition, for $y \in G^\varphi$ we have

$$|N(N_{[A]}(\{y\}))| \geq \frac{\varphi d}{2\Delta_2}$$

and so

$$\begin{aligned}
\mathbf{E}(|G^\varphi \setminus N(N_{[A]}(S))|) &= \sum_{y \in G^\varphi} \mathbf{P}(y \notin N(N_{[A]}(S))) \\
&= \sum_{y \in G^\varphi} \mathbf{P}(N(N_{[A]}(\{y\})) \cap T_0 = \emptyset) \\
&\leq g(1-p)^{\frac{\varphi d}{2\Delta_2}} \\
&\leq g/d^{10}
\end{aligned} \tag{28}$$

Combining (26), (27) and (28) and using Markov's inequality we find that there is at least one $T_0^{initial} \subseteq G$ satisfying

$$|T_0| \leq 3gp, \quad |\nabla(T_0, X \setminus [A])| \leq 3tdp$$

and (25). Now note that $p > \Omega(\log d/d^2)$, so for $g > d^4$ we have (as usual, for sufficiently large d)

$$gp \geq 1 \quad \text{and} \quad tdp \geq d. \tag{29}$$

Set $T_0 = T_0^{initial} \cup \{v\}$. By (29) T_0 satisfies (23) and (24), and it inherits (25) from $T_0^{initial}$. □

Set $T_0' = G^\varphi \setminus N(N_{[A]}(T_0))$, $\Omega = \nabla(T_0, X \setminus [A])$ and $L = N(N_{[A]}(T_0)) \cup T_0'$. Let $T_1 \subseteq G \setminus L$ be a cover of minimum size of $[A] \setminus N(L)$ in the graph induced by $(G \setminus L) \cup ([A] \setminus N(L))$. Then $F' = L \cup T_1$ is a φ-approximation for A.

Before estimating how many sets F' might be produced in this way as we run over all $A \in \mathcal{G}$, we make some observations about the sets described above.

First, note that by Lemma 2.1 F' is 4-linked (A is 2-linked, every $x \in A$ is at distance 1 from F' and every $y \in F'$ is at distance 1 from A) and so, again by Lemma 2.1, $T = T_0 \cup T_0' \cup T_1$ is 8-linked (every $y \in T$ is at distance 2 from something in F' and every $y \in F'$ is at distance 2 from something in T).

By (23) we have $|T_0| \leq O(g \log d/(\varphi d))$, by (25) $|T_0'| \leq O(g/d^{10})$, and by (24) $|\Omega| \leq O(t \log d/\varphi)$.

To bound $|T_1|$, note that $|G \setminus L| \leq td/(d-\varphi)$ (each vertex in $G \setminus L$ is in $G \setminus G^\varphi$ and so contributes at least $(d-\varphi)$ edges to $\nabla(G, X \setminus [A])$, a set of size td), $d_{[A] \setminus N(L)}(u) \leq d$ for each $u \in G \setminus L$, and $d_{G \setminus L}(v) = d$ for each $v \in [A] \setminus N(L)$. So by Lemma 2.3, $|T_1| \leq (t/(d-\varphi))(1 + \ln d) = O(t \log d/(d-\varphi))$.

Combining these observations, we get that T is an 8-linked subset of Y with $|T| = O(g \log d/(\varphi d))$ (if $t < O(g(d-\varphi)/(\varphi d))$) and $|T| = O(t \log d/(d-\varphi))$ (if $t > \Omega(g(d-\varphi)/(\varphi d))$). We deal with these two cases separately.

If $t < O(g(d-\varphi)/(\varphi d))$ we apply Lemma 2.2 to find that there are $2^{O(g \log^2 d/(\varphi d))}$ possible choices for T (note that $v_0 \in T$). Once T has been chosen, there are a further $2^{O(g \log^2 d/(\varphi d))}$ choices for $T_0 \subseteq T$, the same number of choices for $T_1 \subseteq T$ and $\sum_{i \leq O(t \log d/\varphi)} \binom{|\nabla(T_0)|}{i} = 2^{O(t \log^2 d/\varphi)}$ choices for Ω. So the total number of choices for the quadruple (T_0, T_0', T_1, Ω) is

$$2^{O(g \log^2 d/(\varphi d)) + O(t \log^2 d/\varphi)}.$$

If $t > \Omega(g(d-\varphi)/(\varphi d))$ we apply Lemma 2.2 to find that there are $2^{O(t \log^2 d/(d-\varphi))}$ possible choices for T. Once T has been chosen, there are a further $2^{O(t \log^2 d/(d-\varphi))}$ choices for $T_0 \subseteq T$, the same number of choices for $T_1 \subseteq T$ and (as before) $2^{O(t \log^2 d/\varphi)}$ choices for Ω. So the total number of choices for the quadruple (T_0, T_0', T_1, Ω) is

$$2^{O(t \log^2 d/(d-\varphi)) + O(t \log^2 d/\varphi)}.$$

Once T_0, T_1 and Ω have been chosen, F' is completely determined, and (20) follows.

5.4 Proof of the ψ-approximation lemma

Fix a linear ordering $\ll$ of V. Given $A \in \mathcal{G}$ for which F' is a φ-approximation, we produce a ψ-approximation (F, S) for A via the following algorithm.

Step 1: If $\{u \in [A] : d_{G \setminus F'}(u) > \psi\} \neq \emptyset$, pick the smallest (with respect to $\ll$) u in this set and update F' by $F' \longleftarrow F' \cup N(u)$. Repeat this until $\{u \in [A] : d_{G \setminus F'}(u) > \psi\} = \emptyset$. Then set $F'' = F'$ and $S'' = \{u \in X : d_{F''}(u) \geq d - \psi\}$ and go to Step 2.

Step 2: If $\{w \in Y \setminus G : d_{S''}(w) > \psi\} \neq \emptyset$, pick the smallest (with respect to $\ll$) w in this set and update S'' by $S'' \longleftarrow S'' \setminus N(w)$. Repeat this until $\{w \in Y \setminus G : d_{S''}(w) > \psi\} = \emptyset$. Then set $S = S''$ and $F = F'' \cup \{w \in Y : d_S(w) > \psi\}$ and stop.

Claim 5.8. *The output of this algorithm is a ψ-approximation for A.*

Proof: To see that $F \subseteq G$ and $S \supseteq [A]$, first observe that $F'' \subseteq G$ (an immediate consequence of $F' \subseteq G$ and the procedure in Step 1) and that $S'' \supseteq [A]$ (or Step 1 would not have terminated). We then have $S \supseteq [A]$ since Step 2 deletes from S'' only neighbours of $Y \setminus G$, and $F \subseteq G$ since the vertices added to F'' at the end of Step 2 are all in G (or Step 2 would not have terminated). This gives (16).

To verify (17) and (18), note that $d_{F''}(u) \geq d - \psi \ \forall u \in S''$ by definition, $S \subseteq S''$, and $F \supseteq F''$, so that $d_F(u) \geq d - \psi \ \forall u \in S$, and if $w \in Y \setminus F$ then $d_S(w) \leq \psi$ (by Step 2), so that $d_{X \setminus S}(w) \geq d - \psi \ \forall w \in Y \setminus F$. □

Claim 5.9. *The algorithm described above has at most*

$$2^{O(td \log d/((d-\varphi)\psi)) + O(td \log d/((d-\psi)\psi))}$$

outputs as the input runs over those $A \in \mathcal{G}$ for which F' is a φ-approximation.

Taking $\mathcal{W}$ to be the set of all possible outputs of the algorithm, the lemma follows.

Proof of Claim 5.9: The output of the algorithm is determined by the set of u's whose neighbourhoods are added to F' in Step 1, and the set of w's whose neighbourhoods are removed from S'' in Step 2.

Initially, $|G \setminus F'| \leq td/(d - \varphi)$ (each vertex in $G \setminus F'$ is in $G \setminus G^{\varphi}$ and so contributes at least $(d - \varphi)$ edges to $\nabla(G, X \setminus [A])$, a set of size td). Each iteration in Step 1 removes at least ψ vertices from $G \setminus F'$ and so there can be at most $td/((d - \varphi)\psi)$ iterations. The u's in Step 1 are all drawn from $[A]$ and hence $N(F')$, a set of size at most dg. So the total number of outputs for Step 1 is at most

$$\sum_{i \leq td/((d-\varphi)\psi)} \binom{dg}{i} = 2^{O(td \log d/((d-\varphi)\psi))}.$$

We perform a similar analysis on Step 2. Each $u \in S'' \setminus [A]$ contributes more than $d - \psi$ edges to $\nabla(G, X \setminus [A])$, so initially $|S'' \setminus [A]| \leq td/(d - \psi)$. Each w used in Step 2 reduces this by at least ψ, so there are at most $td/((d - \psi)\psi)$ iterations. Each w is drawn from $N(S'')$, a set which is contained in the fourth neighbourhood of F' and so has size at most $d^4 g$. So the total number of outputs for Step 2 is

$$2^{O(td \log d/((d-\psi)\psi))}.$$

The claim follows. □

5.5 Proof of the reconstruction lemma

Say that S is *small* if $|S| < g - \gamma t$ and *large* otherwise. We can obtain all $A \in \mathcal{G}$ for which $F \subseteq G$ and $S \supseteq [A]$ as follows.

If S is small, we specify of A by picking a subset of S. If S is large, we first specify of G. Note that by (19) and the definition of large we have $|G \setminus F| < 2t\psi/(d-\psi) + \gamma t$ and that $G \setminus F \subseteq N(S) \setminus F$, so we specify G by picking a subset of $N(S) \setminus F$ of size at most $2t\psi/(d-\psi) + \gamma t$ (this is our choice of $G \setminus F$). Then, noting that $[A]$ is determined by G, we specify of A by picking a subset of $[A]$.

This procedure produces all possible A's (and lots more besides). We now bound the total number of outputs.

If S is small then the total number of possibilities for A is at most

$$2^{g-\gamma t}. \tag{30}$$

We have

$$|N(S) \setminus F| \leq d|S| \leq 3d^2 g$$

so that if S is large, the total number of possibilities for $|G \setminus F|$ is at most

$$\sum_{i<2t\psi/(d-\psi)+\gamma t} \binom{3d^2 g}{i} \leq 2^{O((t\psi/(d-\psi)+\gamma t)\log d)}.$$

and so the total number of possibilities for A is at most

$$2^{g-t+O((t\psi/(d-\psi)+\gamma t)\log d)}. \tag{31}$$

The lemma follows from (30) and (31).

References Cited

[1] S. Bezrukov, On minimization of the surrounding of subsets in Hamming space, in *Kombinatorno-Algebraic Methods in Applied Mathematics*, Gorky University Press, 1985. (Russian)

[2] B. Bollobás, *Modern Graph Theory*, Springer, New York, 1998.

[3] B. Bollobás, *Random Graphs,* Cambridge University Press, Cambridge, 2001.

[4] H. Chernoff, A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations, *Ann. Math. Statistics* **23** (1952), 493–507.

[5] R. Diestel, *Graph Theory*, Springer, New York, 1997.

[6] D. Knuth, *The Art of Computer Programming* (Vol. I), Addison Wesley (1969).

[7] J. Körner and V. Wei, Odd and even Hamming spheres also have minimum boundary, *Discrete Math.* **51** (1984), 147–165.

[8] A. D. Korshunov and A. A. Sapozhenko, The number of binary codes with distance 2, *Problemy Kibernet.* **40** (1983), 111–130. (Russian)

[9] L. Lovász, On the ratio of optimal integral and fractional covers, *Discrete Math.* **13** (1975), 383–390.

[10] A. A. Sapozhenko, On the number of connected subsets with given cardinality of the boundary in bipartite graphs, *Metody Diskret. Analiz.* **45** (1987), 42–70. (Russian)

[11] A. A. Sapozhenko, The number of antichains in ranked partially ordered sets, *Diskret. Mat.* **1** (1989), 74–93. (Russian; translation in *Discrete Math. Appl.* **1** (1991), 35–58)

[12] S. K. Stein, Two combinatorial covering theorems, *J. Combin. Th. Ser. A* **16** (1974), 391–397.