Combinatorics, Probability and Computing (2017) 26, 248–266. © Cambridge University Press 2016
doi:10.1017/S0963548316000304

On the Chromatic Number of Random Cayley Graphs

BEN GREEN[†]

Mathematical Institute, Andrew Wiles Building, Radcliffe Observatory Quarter, Woodstock Road,
Oxford OX2 6GG, UK
(e-mail: ben.green@maths.ox.ac.uk)

Received 8 August 2013; revised 22 July 2016; first published online 9 September 2016

To Béla Bollobás on his 70th birthday

Let G be an abelian group of cardinality n, where $\mathrm{hcf}(n,6)=1$, and let A be a random subset of G. Form a graph Γ_A on vertex set G by joining x to y if and only if $x+y\in A$. Then, with high probability as $n\to\infty$, the chromatic number $\chi(\Gamma_A)$ is at most $(1+o(1))\frac{n}{2\log_2 n}$. This is asymptotically sharp when $G=\mathbb{Z}/n\mathbb{Z}$, n prime.

2010 *Mathematics subject classification*: Primary 05C15
Secondary 11P70

1. Introduction

A celebrated result of Bollobás [5] asserts that a random graph from the $G(n,1/2)$ model has chromatic number $(1+o(1))\frac{n}{2\log_2 n}$ w.h.p.[1] Our aim in this note is to prove that $(1+o(1))\frac{n}{2\log_2 n}$ is an upper bound for the chromatic number of random Cayley sum graphs on G,[2] where G is an abelian group of order n where $\mathrm{hcf}(n,6)=1$. If $A\subset G$ is a set then we define its *Cayley sum graph* Γ_A to be the graph on vertex set G in which i is joined to j if and only if $i+j\in A$.

Theorem 1.1. *Let G be an abelian group of order n, and suppose that $A\subset G$ is selected uniformly at random from all subsets of G. Then, w.h.p. over n with* $\mathrm{hcf}(n,6)=1$, *the chromatic number $\chi(\Gamma_A)$ of Γ_A is at most $(1+o(1))\frac{n}{2\log_2 n}$.*

[†] The author is supported by ERC Starting Grant 274938 (Approximate Algebraic Structure) as well as by an Investigator Award from the Simons Foundation.

[1] w.h.p. = with high probability. Throughout the paper, this will mean that the claimed statement holds with probability tending to 1 as $n\to\infty$, sometimes with an additional constraint on n which will be explicitly noted.

[2] It is more usual in the literature to take A to be symmetric and to join i to j if and only if $i-j\in A$. This gives the true Cayley graph as opposed to the Cayley sum graph. There is little genuine difference between the two models, and the Cayley sum graphs are slightly simpler to handle notationally, which is why we have preferred them.

What is meant by this is as follows: for every $\varepsilon > 0$, the probability that $\chi(\Gamma_A) \leqslant (1+\varepsilon)\frac{n}{2\log_2 n}$ tends to 1 as $n \to \infty$ through values of n coprime to 6.

The condition that $\mathrm{hcf}(n,6) = 1$ could probably be removed, but to do so would involve a number of non-trivial modifications to certain parts of the argument. This is particularly so with regard to the definition of *dissociativity*, which assumes some importance later on. Since the case of greatest interest is probably $G = \mathbb{Z}/n\mathbb{Z}$, n prime, we have chosen not to do this additional work here. In this case, it follows from work of Morris and the author [8] that we have a corresponding lower bound $\chi(\Gamma_A) \geqslant (1-o(1))\frac{n}{2\log_2 n}$ w.h.p. To see this, note that we have $\chi(\Gamma)\omega(\Gamma^c) \geqslant n$ for any graph Γ on n vertices, where Γ^c denotes the complement of Γ and ω is the clique number. Indeed, if Γ can be k-coloured then there is a set of vertices of size at least n/k, all of which get the same colour and which must therefore be independent. An independent set of vertices is the same thing as a clique in Γ^c. The stated lower bound on $\chi(\Gamma_A)$ is a consequence of this observation and the main result of [8], which implies that $\omega(\Gamma_A^c) = \omega(\Gamma_{A^c}) \leqslant (2+o(1))\log_2 n$ w.h.p.

A k-colouring of Γ_A corresponds precisely to a partition $G = X_1 \cup \cdots \cup X_k$ with the property that $X_i \hat{+} X_i \subset A^c$ for all i, where $X \hat{+} X$ denotes the restricted sumset $\{x + x' : x, x' \in X, x \neq x'\}$. Let us record the arithmetic formulation of the upper bound in Theorem 1.1 as a separate proposition.

Proposition 1.2. *Suppose that $A \subset G$ is a random set and let $r = (1+\varepsilon)\frac{n}{2\log_2 n}$. Then w.h.p. over n with* $\mathrm{hcf}(n,6) = 1$ *there is a partition $G = X_1 \cup \cdots \cup X_r$ such that $X_i \hat{+} X_i \subset A$ for all i.*

Note that we wrote A instead of A^c, since if A is a random set then so is its complement.

The reader should be aware that many of the key ideas in the proof of this proposition, and hence of Theorem 1.1, have exact parallels in Bollobás's paper [5], though we will not always draw attention to these explicitly. However a number of quite non-trivial technical obstacles must be overcome in this arithmetic setting, and herein lies all of the novelty of the present work.

Previous results. A comprehensive resource for questions concerning the clique number of random Cayley graphs is Alon's paper [2]. Alon considers different groups (not necessarily abelian) and random sets A of different sizes. He notes in [2, Theorem 2.1(i)] that as a consequence of an earlier result of his, joint with Krivelevich and Sudakov [3], we have $\chi(\Gamma_A) \ll \frac{n}{\log n}$ almost surely if $A \subset G$ is a random set of size $n/2$. (Here, and henceforth in the paper, $X \ll Y$ means that $X \leqslant CY$ for some absolute constant C.) The same method gives a similar bound when $A \subset G$ is selected uniformly from all subsets of G.

This whole argument, which is phrased in terms of graph eigenvalues, translates rather succinctly to the arithmetic setting considered here if one uses a little Fourier analysis (which amounts to essentially the same thing). We give this argument in Appendix B, obtaining the bound $\chi(\Gamma_A) \leqslant (2+o(1))\frac{n}{\log_2 n}$ for a random set A by these methods. Note that this is four times the bound of Theorem 1.1.

Let us note, however, that in the argument of [3] the set A is only required to be *pseudorandom* in the sense of having no large Fourier coefficients (the bound being weaker the larger these coefficients are). In particular it would apply when $A = Q$ is the set of quadratic residues modulo n when n is prime. In this case Γ_Q is called the *Paley sum graph*, and it follows from [3] that

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

$\chi(\Gamma_Q) \leqslant (2+o(1))\log_2 n$, a result that appears to be the best known for this problem. By contrast, our result gives nothing new about this specific graph.

Another observation of Alon is [2, Proposition 4.5], which notes a consequence of an observation of mine from [7]: if $G = \mathbb{F}_2^m$ and $n = 2^m$ then w.h.p. (as n ranges over powers of 2) $\chi(\Gamma_A) \ll \frac{n}{\log n \log\log n}$. Thus the chromatic number of a random Cayley graph can depend on the underlying group.

2. Preliminary manœuvres

We begin with the observation that Proposition 1.2 is implied by the following result. Here, and for the rest of the paper, we write $E[X] := X \hat{+} X$ (the letter E is supposed to denote 'edge').

Proposition 2.1. *Suppose that $A \subset G$ is a random set. Then w.h.p. there is a partition $G = X_1 \cup \cdots \cup X_r \cup X_*$ such that $|X_i| = (2+o(1))\log_2 n$ uniformly in i, $E[X_i] \subset A$, and $|X_*| = o\big(\frac{n}{\log n}\big)$.*

Proposition 1.2 follows upon splitting the exceptional set X_* into singletons. Proposition 2.1 follows in turn by a repeated application of the next statement.

Proposition 2.2. *Suppose that $A \subset G$ is a random set. Then w.h.p. every set $S \subset G$ with $|S| = \frac{n}{\log^2 n}$ contains a set X with $|X| = (2+o(1))\log_2 n$ and $E[X] \subset A$.*

Remark. The role of quantity $\frac{n}{\log^2 n}$ here is simply to be something concrete that is $o\big(\frac{n}{\log n}\big)$.

Perhaps the most obvious approach to proving Proposition 2.2 would be to obtain a strong upper bound on the probability that there does *not* exist such a set X for a *fixed* S, and then use the union bound over all $S \subset G$ of size $\frac{n}{\log^2 n}$. Unfortunately, this approach is too crude, since there are various difficulties in obtaining a good upper bound for an arbitrary set S. We must instead pass to a thinner class of sets S' satisfying some useful technical properties. The following definition and lemma make this possible.

Definition 1. We say that a set $S' \subset G$ is *useful* if it enjoys the following properties.

(i) (*Cardinality*)

$$2^{-12}\frac{n}{\log^{20} n} < |S'| \leqslant \frac{n}{\log^{20} n}.$$

(ii) (*Good clique size*) There is some integer $k = k_{S'}$, $k = (2+o(1))\log_2 n$, such that

$$\frac{n}{2\log^8 n} < \binom{|S'|}{k} 2^{-\binom{k}{2}} \leqslant \frac{n}{\log^8 n}.$$

(iii) (*Lack of structure*) For at least 90% of all sets $X \subset S'$ of size k, the following is true: the number of quadruples $(x_1, x_2, s_1, s_2) \in X \times X \times S' \times S'$ with $x_1 \neq x_2$, $s_1 \neq s_2$ and $x_1 + x_2 = s_1 + s_2$ is at most $\frac{1}{\log^{15} n}|S'|$.

Lemma 2.3. *Suppose that $S \subset G$ has size $\frac{n}{\log^2 n}$. Then there is a set $S' \subset S$ that is useful.*

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Before proceeding to the proof of this lemma, we offer some explanatory remarks concerning parts (ii) and (iii) of Definition 1. Concerning (ii), let us assume (a true statement, in fact, as we shall see below in Lemma 4.2) that 'most' sets $X \subset S'$ of size k have $|E[X]| = \binom{k}{2}$. Then $\binom{|S'|}{k}2^{-\binom{k}{2}}$ is roughly the expected number of $X \subset S'$ with $|X| = k$ and $E[X] \subset A$. At a later point in the argument it will be important to have an integer value of k for which this number is of controlled size, slightly less than n (in fact $\asymp \frac{n}{\log^8 n}$,[3] though there is some flexibility in this choice). Unfortunately, we have

$$\frac{\binom{|S'|}{k}2^{-\binom{k}{2}}}{\binom{|S'|}{k+1}2^{-\binom{k+1}{2}}} = \frac{(k+1)2^k}{|S'|-k}.$$

If $|S'|$ is of order $n^{1+o(1)}$ and $k = (2+o(1))\log_2 n$, this will be roughly $n^{1+o(1)}$. Hence the values of $\binom{|S'|}{k}2^{-\binom{k}{2}}$ are rather widely spaced as k varies, and unless $|S'|$ is chosen quite judiciously there will be no value of k with the property we require. Item (ii) of Definition 1 is devoted to making just such a judicious choice.

Concerning part (iii) of Definition 1, let us note that a trivial upper bound for the number of quadruples (x_1, x_2, s_1, s_2) with $x_1 \neq x_2$, $s_1 \neq s_2$ and $x_1 + x_2 = s_1 + s_2$ is $O(|S'|\log^2 n)$, since any choice of x_1, x_2, s_1 uniquely determines s_2. However, this trivial bound is not necessarily a very sharp one, since we will not usually have $x_1 + x_2 - s_1 \in S'$ unless S' has some particular additive structure. Item (iii) is asserting, in a certain technical sense, that we can assume S' does not have too much structure. We will only make use of (iii) at one later point in the argument, namely in the proof of Lemma 5.1, but it will be quite crucial there.

Remark. The use of additively unstructured sets in contexts like this goes back at least as far as [1].

In view of Lemma 2.3, to prove Proposition 2.2 it is enough to establish the following.

Proposition 2.4. *Suppose that $A \subset G$ is a random set. Then w.h.p. every useful set $S' \subset G$ contains a set X with $|X| = k_{S'}$ and $E[X] \subset A$.*

We will prove this by taking a union bound over all useful sets S'. Since any useful set has cardinality at most $\frac{n}{\log^{20} n}$ (by item (i) of Definition 1) and $\binom{n}{n/\log^{20} n} = e^{o(n/\log^{19} n)}$, it suffices to prove the following.

Proposition 2.5. *Let n be sufficiently large, and let $S \subset G$ be useful in the sense of Definition 1. Let $k_S = (2+o(1))\log_2 n$ be as in part (ii) of Definition 1. Suppose that $A \subset G$ is a random set. Then, with probability at least $1 - e^{-n/\log^{19} n}$, there is a set $X \subset S$ with $|X| = k_S$ and $E[X] \subset A$.*

The proof of this bound, which by the reductions just given implies our main theorem, will occupy our attention for most of the rest of the paper. To conclude this section, we prove Lemma 2.3.

[3] $X \asymp Y$ means that X and Y have the same order of magnitude in the sense that $c_1 Y \leqslant X \leqslant c_2 Y$ for absolute constants $c_1, c_2 > 0$.

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Proof of Lemma 2.3. The strategy will be to choose $S' \subset S$ to be a random subset of size $\asymp n/\log^{20} n$. The lack of structure condition of part (iii) of Definition 1 will almost surely be satisfied, but part (ii) will not. However, we can pass to a further subset, also of size $\asymp n/\log^{20} n$, which does enjoy this property. Doing so does not do substantial damage to part (iii).

We turn to the details. Let $\varepsilon = \frac{1}{2\log^{18} n}$ and select a set $T \subset S$ at random by picking each element of S independently at random with probability ε. By standard tail estimates such as [4, Theorem A.1.4], we have

$$\mathbb{P}\left(||T| - \varepsilon|S|| \geqslant \frac{1}{2}\varepsilon|S|\right) < 2e^{-C\varepsilon^2|S|} < \frac{1}{2}. \tag{2.1}$$

By an *additive quadruple* in S we mean a quadruple (s_1, s_2, s_3, s_4) with $s_1 \neq s_2$, $s_3 \neq s_4$ and $s_1 + s_2 = s_3 + s_4$. The number of such additive quadruples is certainly less than $|S|^3$, since s_1, s_2, s_3 determine s_4. Divide the additive quadruples into two classes: the *non-degenerate* ones in which s_1, s_2, s_3, s_4 are all distinct, and the *degenerate* ones in which either $s_1 = s_3$ or $s_1 = s_4$. There are at most $2|S|^2$ degenerate quadruples. A degenerate quadruple lies in T with probability ε^2, whereas a non-degenerate one lies in T with probability ε^4. Therefore the expected number of additive quadruples in T is less than $\varepsilon^4|S|^3 + 2\varepsilon^2|S|^2$, which is less than $2\varepsilon^4|S|^3$ if N is large. By Markov's inequality, the number of additive quadruples in T is less than $4\varepsilon^4|S|^3$ with probability at least $1/2$. Noting that the number of additive quadruples is equal to

$$\sum_{x,x' \in T : x \neq x'} r_T(x + x'),$$

where $r_T(\xi)$ denotes the number of ways of writing ξ as the sum of two *distinct* elements x, x' of T, we see from this and (2.1) that there is some T for which

$$\frac{1}{2}\varepsilon|S| \leqslant |T| \leqslant \frac{3}{2}\varepsilon|S| \tag{2.2}$$

and

$$\sum_{x,x' \in T, x \neq x'} r_T(x + x') \leqslant 4\varepsilon^4|S|^3. \tag{2.3}$$

These properties will help us satisfy parts (i) and (iii) of Definition 1. Let us leave them aside for now and concentrate on part (ii). For this we require the following lemma.

Lemma 2.6. *Let M be a sufficiently large integer. Suppose that D is a real number satisfying $1 \leqslant D \leqslant M^2$. Then there is some integer $M' = M'(D)$, $2^{-10}M \leqslant M' \leqslant M$, and an integer $k = (2 + o(1))\log_2 M$ such that $D \leqslant \binom{M}{k}2^{-\binom{k}{2}} \leqslant 2D$.*

Proof. In the proof of this lemma we write $F(k, M) := \binom{M}{k}2^{-\binom{k}{2}}$. First of all note that if $M > M_0(\varepsilon)$ is sufficiently large then

$$F(\lceil(2 - \varepsilon)\log_2 M\rceil, M) > M^{10} \tag{2.4}$$

and

$$F(\lfloor(2 + \varepsilon)\log_2 M\rfloor, M) < M^{-10} = o(1). \tag{2.5}$$

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

We leave the straightforward confirmation of these facts to the reader. Thus if there is a value of k such that $D \leqslant F(k,M) \leqslant 2D$ with D in the stated range, then it automatically satisfies $k = (2+o(1))\log_2 M$.

Note also that if $\log_2 M \leqslant k \leqslant 3\log_2 M$ then

$$1 \leqslant \frac{F(k,M)}{F(k+1,M)} = \frac{k+1}{M-k}2^k \leqslant M^3.$$

Thus, in view of (2.4) and (2.5), there is certainly some k in this range such that

$$M^2 \leqslant F(k,M) \leqslant M^5.$$

Now we fix this k and start decreasing M. Note that

$$\frac{F(k,M)}{F(k,M-1)} = \frac{M}{M-k} < 2,$$

and so by decreasing M one by one we do hit some M' for which $D \leqslant F(k,M') \leqslant 2D$. We must give a lower bound for M'. To do this, note that by Lemma C.1 we have

$$\frac{F(k,2t)}{F(k,t)} \geqslant 2^k > M^{1/2}$$

for any $t \geqslant k$, and therefore

$$\frac{F(k,2^{10}t)}{F(k,t)} > M^5.$$

It follows that $M' \geqslant 2^{-10}M$, concluding the proof of Lemma 2.6. □

Let us return now to the proof of Lemma 2.3. Recall that we had isolated a subset $T \subset S$ satisfying (2.2) and (2.3), where $\varepsilon := \frac{1}{2\log^{18} n}$. By Lemma 2.6 applied with $M = |T|$ and $D := \frac{n}{2\log^8 n}$, we can pass to a subset $S' \subset T$ which satisfies parts (i) and (ii) of Definition 1 for some $k = (2+o(1))\log_2 n$ satisfying $D \leqslant \binom{M}{k}2^{-\binom{k}{2}} \leqslant 2D$. We claim that S' also satisfies property (iii) of that lemma.

To this end, note that (2.2) and (2.3) together with the bounds for $|S'|$ imply that

$$\sum_{x,x' \in S', x \neq x'} r_{S'}(x+x') \leqslant \sum_{x,x' \in T, x \neq x'} r_T(x+x') \leqslant C\varepsilon|S'|^3 \ll \frac{1}{\log^{18} n}|S'|^3. \tag{2.6}$$

Let $X \subset S'$ be a random subset of S' of size k. Then, by (2.6),

$$\begin{aligned}\mathbb{E}_X \sum_{x,x' \in X, x \neq x'} r_{S'}(x+x') &= \sum_{x,x' \in S', x \neq x'} \mathbb{P}_X(x,x' \in X) r_{S'}(x+x') \\ &= \left(\frac{k}{|S'|}\right)^2 \sum_{x,x' \in S', x \neq x'} r_{S'}(x+x') \\ &\ll \left(\frac{k(k-1)}{|S'|(|S'|-1)}\right) \cdot \frac{1}{\log^{18} n}|S'|^3 \ll \frac{1}{\log^{16} n}|S'|.\end{aligned}$$

By Markov's inequality, we have

$$\sum_{x,x' \in X, x \neq x'} r_{S'}(x+x') \ll \frac{1}{\log^{16} n}|S'|$$

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

for at least 90% of all $X \subset S'$ of size k, which implies item (iii) of Lemma 2.3 for large n. □

3. The exposure martingale

We now explain the main outline of the proof of Proposition 2.5. First, let us recall the statement (the reader may care to recall the definition of S being *useful*, which is given in Definition 1, but the specifics of that definition are not important in this section).

Proposition 2.5. Let n be sufficiently large, and let $S \subset G$ be useful. Let $k_S = (2+o(1))\log_2 n$ be as in part (ii) of Definition 1. Suppose that $A \subset G$ is a random set. Then, with probability at least $1 - e^{-n/\log^{19} n}$, there is a set $X \subset S$ with $|X| = k_S$ and $E[X] \subset A$.

In what follows, we write $k = k_S$ for short.

For the rest of this section our notation will be as in this proposition. Let Ω be the probability space consisting of all subsets of G, each occurring with equal probability 2^{-n}. Thus A is drawn at random from Ω. Let the random variable $f : \Omega \to \mathbb{N}$ be defined as follows: $f(A)$ is the maximum value of r for which there exist sets $X_i \subset S$, $i = 1,\dots,r$, with $|X_i| = k$ and such that the $E[X_i]$ are disjoint subsets of A. The task of establishing Proposition 2.5 is equivalent to showing that

$$\mathbb{P}(f(A) = 0) \leqslant e^{-n/\log^{19} n} \tag{3.1}$$

if n is sufficiently large.

Let $g_1,\dots,g_n$ be some arbitrary enumeration of the elements of G. Let $\mathcal{F}_j$ be the sub-σ-algebra of 2^{Ω} generated by sets of the form

$$\{A \in \Omega : 1_A(g_1) = \varepsilon_1, \dots 1_A(g_j) = \varepsilon_j\} \quad \text{for } (\varepsilon_1,\dots,\varepsilon_j) \in \{0,1\}^j,$$

and consider the random variables $Z_j := \mathbb{E}(f(A)|\mathcal{F}_j)$. We have the nesting

$$\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}_n,$$

where $\mathcal{F} \subset \mathcal{F}'$ means that $\mathcal{F}'$ is a refinement of $\mathcal{F}$. The sequence Z_j is a Doob martingale and we have $Z_0 = \mathbb{E}f(A)$ and $Z_n = f(A)$. Furthermore, flipping the value of $1_A(g_j)$ cannot change $f(A)$ by more than 1, since if $g_j \in A$ then removing g_j cannot destroy the containment $E[X_i] \subset A$ for more than one value of i, on account of the sets $E[X_i]$ being disjoint. It follows that the martingale $(Z_j)_{j=0}^n$ enjoys the Lipschitz property $|Z_{j-1} - Z_j| \leqslant 1$ for $j = 1,2,\dots,n$, and therefore we may apply Azuma's inequality [4, Corollary 7.2.2] to conclude that

$$\mathbb{P}(|f(A) - \mathbb{E}f(A)| \geqslant t) \leqslant 2e^{-t^2/2n}.$$

In order to establish (3.1), and hence Proposition 2.5 and our main theorem, it is enough to prove that

$$\mathbb{E}f(A) \geqslant n/\log^9 n \tag{3.2}$$

provided that n is sufficiently large.

4. In search of edge-disjoint cliques

In the next two sections we always assume that $n = |G|$ is sufficiently large and that $\mathrm{hcf}(n,6) = 1$.

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Our remaining task is to prove (3.2). Let us recall the setup: we have a fixed useful set $S \subset G$, satisfying parts (i), (ii) and (iii) of Definition 1, and we wish to give a lower bound for the expectation of $f(A)$, where A ranges uniformly over all subsets of G. Here, $f(A)$ is the maximum value of r for which there exist sets $X_i \subset S$, $i = 1, \dots, r$, with $|X_i| = k$ and such that the $E[X_i]$ are disjoint subsets of A. Recall that $k = k_S = (2 + o(1)) \log_2 n$ is such that item (ii) of Lemma 2.3 holds.

Define $\tilde{f}(A)$ to be the number of sets $X \subset S$ such that the following hold.

(i) $|X| = k$.
(ii) $E[X] \subset A$.
(iii) If $Y \subset S$ is any other set, distinct from X, with $|Y| = k$ and $E[Y] \subset A$ then $E[X] \cap E[Y] = \emptyset$.

We clearly have $f(A) \geqslant \tilde{f}(A)$, and so it suffices to obtain a lower bound for $\mathbb{E}\tilde{f}(A)$. In fact, we shall consider the following technical variant of $\tilde{f}(A)$: define $\tilde{\tilde{f}}(A)$ to be the number of sets $X \subset S$ such that the following hold.

(i) $|X| = k$.
(ii) $E[X] \subset A$.
(iii) (*Lack of structure with respect to S*) The number of quadruples $(x_1, x_2, s_1, s_2) \in X \times X \times S \times S$ with $x_1 \neq x_2$, $s_1 \neq s_2$ and $x_1 + x_2 = s_1 + s_2$ is at most $\frac{1}{\log^{15} n}|S|$.
(iv) (*Dissociativity*) If $x_1, \dots, x_4, x_1', \dots, x_4' \in X$ and $x_1 + x_2 + x_3 + x_4 = x_1' + x_2' + x_3' + x_4'$, then x_1, x_2, x_3, x_4 are a permutation of x_1', x_2', x_3', x_4'.
(v) If $Y \subset S$ is any other dissociated set with $|Y| = k$ and $E[Y] \subset A$, then $E[X] \cap E[Y] = \emptyset$.

Note that $f(A) \geqslant \tilde{\tilde{f}}(A)$, so it suffices to get a lower bound for $\mathbb{E}\tilde{\tilde{f}}(A)$. Specifically, to conclude (3.2) and hence our main theorem, we need only prove the following.

Proposition 4.1. *Let $S \subset G$ be a fixed useful set. Let $A \subset G$ be chosen at random, and let $\tilde{\tilde{f}}(A)$ be as above. Then $\mathbb{E}\tilde{\tilde{f}}(A) \geqslant n/\log^9 n$.*

To prove this proposition we will find lower bounds for:

(1) the number of dissociated sets $X \subset S$ with $|X| = k$ and having lack of structure with respect to S;
(2) if X is such a set, the probability that $E[X] \subset A$;
(3) the conditional probability that there is another dissociated set $Y \subset S$ with $|Y| = k$, $E[Y] \subset A$ and $E[X] \cap E[Y] \neq \emptyset$.

Point (2) is actually rather easy: if X is dissociated then all sums $x + x'$ with $x \neq x'$ are distinct apart from the trivial equalities $x + x' = x' + x$, and so $|E[X]| = \binom{k}{2}$. Therefore if A is a random set then

$$\mathbb{P}(E[X] \subset A) = 2^{-\binom{k}{2}}. \tag{4.1}$$

We turn now to point (1), which is also quite straightforward, the point being that a random subset consisting of k elements of S is almost certain to be dissociated and to have lack of structure with respect to S. We formulate this in a lemma.

Lemma 4.2. *Let $S \subset G$ be useful and let $k = k_S$ be as above. Then at least 80% of all sets $X \subset S$ with $|X| = k$ satisfy both lack of structure with respect to S and dissociativity.*

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Proof. It is enough to show that at least 90% of sets $X \subset S$ with $|X| = k$ are dissociated, since it is part of the definition of S being useful (Definition 1 part (iii)) that at least 90% of such sets satisfy lack of structure with respect to S. To prove that this is so, we select the elements of X one at a time, without replacement and with an order, and ask what might happen to prevent X being dissociated. If we have selected j elements $\{x_1, \dots, x_j\}$ then the next element x must not give rise to a non-trivial solution to such equations as $x_{i_1} + x_{i_2} + x_{i_3} + x_{i_4} = x_{i_5} + x_{i_6} + x_{i_7} + x$. The number of such equations is no more than $2^8 j^7$, and each forbids a unique value of x. Another example of such an equation is $x_{i_1} + x + x + x = x_{i_2} + x_{i_3} + x_{i_4}$, which forbids a unique value of x because, since $\mathrm{hcf}(n, 6) = 1$, G has no 3-torsion.[4] Therefore, if $j < k$ then the number of choices for x_{j+1} is at least $|S| - 2^8 j^7 \geqslant |S| - 2^{15} \log^7 n$. Since $|S| \asymp \frac{n}{\log^{20} n}$, we very comfortably have

$$(|S| - 2^{15} \log^7 n)^k \geqslant \frac{9}{10}|S|^k \geqslant \frac{9}{10} k! \binom{|S|}{k},$$

and this implies the claimed result upon dividing through by $k!$ to take account of the fact that we counted ordered k-tuples rather than sets. □

We turn now to point (3). We shall show in the next section that this conditional probability is at most $1/2$. This finishes the task of proving Proposition 4.1 and hence the main theorem, because by combining these estimates for (1), (2) and (3) we find that the expected number of X satisfying (i)–(v) in the definition of $\tilde{\tilde{f}}(A)$ above is at least

$$0.8 \binom{|S|}{k} \times 2^{-\binom{k}{2}} \times \frac{1}{2} \asymp \frac{n}{\log^8 n},$$

this last estimate following from the ‘good clique size’ property, part (ii) of Definition 1.

5. Intersecting arithmetic cliques

At the end of the last section we had reduced the proof of our main theorem to the following fact, which gives the required bound for the conditional probability (3).

Lemma 5.1. *Suppose that $S \subset G$ is useful and that $X \subset S$ has $|X| = k$, is dissociated, and has lack of structure with respect to S. Then, conditioned upon the event that $E[X] \subset A$, the probability that there is some other dissociated set $Y \subset S$ with $|Y| = k$, $E[Y] \subset A$ and $E[X] \cap E[Y] \neq \emptyset$ is at most $1/2$.*

We will divide into a number of cases according to the value of $\ell := |E[X] \cap E[Y]|$. The case $\ell = 1$ is somewhat special, so we handle it separately. This is the only point in the argument at which the lack of structure condition (Definition 1 part (iii)) is required.

Lemma 5.2. *Suppose that $S \subset G$ is useful and that $X \subset S$ has $|X| = k$, is dissociated, and has lack of structure with respect to S. Then, conditioned upon the event that $E[X] \subset A$, the probability*

[4] If G has 2-, 3- or 4-torsion then more delicate arguments, including a change in the definition of dissociativity, are required.

that there is some other dissociated set $Y \subset S$ *with* $|Y| = k$, $E[Y] \subset A$ *and* $|E[X] \cap E[Y]| = 1$ *is at most* $1/10$.

Proof. If $|E[X] \cap E[Y]| = 1$ then there is a pair $(x_1, x_2) \in X \times X$, $x_1 \neq x_2$, such that $x_1 + x_2 \in E[Y]$. Thus Y contains two distinct elements $s_1, s_2 \in S$ with $s_1 + s_2 = x_1 + x_2$. The total number of choices for this pair of elements (across all choices of x_1, x_2) is bounded by $\frac{1}{\log^{15} n}|S|$, by the assumption that X has lack of structure with respect to S. The number of choices for the remaining elements of Y is at most $\binom{|S|}{k-2}$, and hence there are at most

$$\frac{1}{\log^{15} n}|S| \cdot \binom{|S|}{k-2}$$

choices for Y in total. For each such choice, the probability that $E[Y] \subset A$ given that $E[X] \subset A$ is precisely $2^{1-\binom{k}{2}}$, since Y is dissociated and hence $|E[Y]| = \binom{k}{2}$.

Therefore the probability we seek to bound is at most

$$\begin{aligned} &\frac{2}{\log^{15} n}|S| \cdot \binom{|S|}{k-2} \cdot 2^{-\binom{k}{2}} \\ &\qquad = \frac{2|S|k(k-1)}{(|S|-k+1)(|S|-k+2)\log^{15} n} \cdot \binom{|S|}{k} 2^{-\binom{k}{2}}. \end{aligned}$$

Recalling that $|S| \asymp \frac{n}{\log^{20} n}$, $k \asymp \log n$ and that

$$\binom{|S|}{k} 2^{-\binom{k}{2}} \asymp \frac{n}{\log^8 n}$$

(the 'good clique size' property, Definition 1 part (ii)), we see that this probability is $\ll \frac{1}{\log n} = o(1)$, as required. □

For the remainder of this section, then, we assume that $\ell = |E[X] \cap E[Y]| \geqslant 2$. To study this situation we introduce some further notation.

Apart from ℓ, another key quantity will be d, defined as follows. Fix, once and for all, an arbitrary total ordering $\prec$ on G. Then any set $Y \subset S$ with $|Y| = k$ may be totally ordered as $y_1 \prec \cdots \prec y_k$. Define a graph Γ_Y on vertex set $[k]$ by joining i and j by an edge if and only if $y_i + y_j \in E[X]$. The number of edges in Γ_Y is precisely ℓ; write d for the number of connected components (including isolated vertices) in Γ_Y.

There is an important relation between ℓ and d.

Lemma 5.3. *We have* $\ell \leqslant \frac{1}{2}(k-d+1)(k-d)$, *and* $\binom{k}{2} - \ell \geqslant (d-1)\big(k - \frac{1}{2}d\big)$.

Proof. The second statement is obviously equivalent to the first, but we have stated it separately for convenience. Suppose that the components of Γ have sizes (number of vertices) $x_1, \dots, x_d$, thus $x_1 + \cdots + x_d = k$ and $x_1, \dots, x_d \geqslant 1$. Then the number of edges in Γ is at most $\sum_i \binom{x_i}{2}$. By convexity this is largest when $x_1 = k-d+1$ and $x_2 = \cdots = x_d = 1$, in which case the number of edges is $\frac{1}{2}(k-d+1)(k-d)$. □

An important notion will be that of the *skeleton* $\mathrm{sk}(\Gamma_Y)$ of Γ_Y. Given a graph Γ on vertex set $[k]$, its skeleton $\mathrm{sk}(\Gamma) \subset \Gamma$ is a forest (union of trees) whose connected components are precisely

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

the connected components of Γ. There are, in general, many choices of a skeleton $\mathrm{sk}(\Gamma)$ for each Γ, but we make an arbitrary one.

Turning back to our main task, we now dispense with the case in which ℓ is not extremely close to $\binom{k}{2}$.

Lemma 5.4. *Suppose that $S \subset G$ is useful and that $X \subset S$ is a dissociated set with $|X| = k$. Then, conditioned upon the event that $E[X] \subset A$, the probability that there is some other dissociated set $Y \subset S$ with $|Y| = k$, $E[Y] \subset A$ and $2 \leqslant |E[X] \cap E[Y]| \leqslant \binom{k}{2} - k^{4/3}$ is at most* $1/10$.

Proof. For each ℓ we enumerate the number of sets Y for which $|E[X] \cap E[Y]| = \ell$ according to their skeleton $\sigma := \mathrm{sk}(\Gamma_Y)$. Since this skeleton has at most $\min(\ell, k)$ edges, a very crude bound for the number of choices for σ is

$$\sum_{i=0}^{\min(\ell,k)} \binom{k}{2}^i < k^{2\min(\ell,k)}.$$

If σ has d connected components, then an upper bound for the number of choices for Y is at most $\binom{|S|}{d} k^{2\min(\ell,k)}$. To see this, first select vertices $i_1, \dots, i_d \in [k]$, one in each of the d connected components, and assign the values of y_{i_j} arbitrarily subject to the order relation that $y_i \prec y_{i'}$ if $i < i'$. These elements must be elements of S, so there are $\binom{|S|}{d}$ choices. Some further vertices will be joined to an i_j by an edge of σ. Suppose, for example, that i_* is joined to i_1 by such an edge. Then $y_{i_*} + y_{i_1} \in E[X]$, and so there are at most $\binom{k}{2} < k^2$ choices for y_{i_*}. Similarly, some further vertices will be joined to i_* by an edge of σ, and so on. By repeating this process we will eventually assign all of the values $y_1, \dots, y_k$, and the number of choices that has been made is at most the number of edges of σ, which is at most $\min(\ell, k)$.

Putting these observations together, the number of choices for Y, for a fixed given ℓ, is at most $\binom{|S|}{d} k^{4\min(k,\ell)}$. For each such choice, the probability that $E[Y] \subset A$, given that $E[X] \subset A$, is $2^{-\binom{k}{2}+\ell}$. Summing over the choices of Y for which $|E[X] \cap E[Y]| = \ell$, we get a total contribution of at most

$$\binom{|S|}{d} k^{4\min(k,\ell)} \cdot 2^{-\binom{k}{2}+\ell}. \tag{5.1}$$

We handle this differently according to the size of ℓ.

Suppose first that $2 \leqslant \ell \leqslant k$. Then by Lemma C.3 we may bound (5.1) above by

$$\binom{|S|}{k} 2^{-\binom{k}{2}} \cdot 2^{\ell} k^{4\ell} \left(\frac{k}{|S| - k} \right)^{k-d}.$$

The fact that S is useful tells us that

$$\binom{|S|}{k} 2^{-\binom{k}{2}} \asymp \frac{n}{\log^8 n} < n.$$

Furthermore, the first relation in Lemma 5.3 implies that $k - d \geqslant \sqrt{\ell}$. Since $|S| \asymp n^{1+o(1)}$ and $k < 3\log_2 n$, these facts together allow us to bound (5.1) above by

$$\frac{(6\log_2 n)^{4\ell} n}{n^{(1-o(1))\sqrt{\ell}}},$$

which is $\ll n^{-c}$ for the stated range of ℓ (and in fact for ℓ up to about $\frac{k^2}{\log^{O(1)} k}$).

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

If $k < \ell \leqslant \binom{k}{2} - k^{4/3}$ then we replace $\min(k,\ell)$ by k, so the quantity (5.1) that we wish to bound is

$$\binom{|S|}{d} k^{4k} 2^{-\binom{k}{2}+\ell}. \tag{5.2}$$

Since S is useful, we have $\binom{|S|}{k} = 2^{\binom{k}{2}} n^{1+o(1)}$, and so from Lemma C.2 and the fact that $k \sim 2\log_2 N$, we have

$$\binom{|S|}{d} \leqslant k^d \binom{|S|}{k}^{d/k} \leqslant 2^{\frac{1}{2}d(k-1)} k^d n^{O(d/k)} \leqslant 2^{\frac{1}{2}d(k-1)} (2k)^{O(d)}.$$

Since $d \leqslant k$, an upper bound for (5.2) is therefore

$$2^{\frac{1}{2}dk+\ell-\binom{k}{2}+O(k\log k)}. \tag{5.3}$$

Since $\ell - \binom{k}{2} \leqslant -k^{4/3}$, we get an upper bound of $2^{-\frac{1}{2}k^{4/3}}$ if $d < k^{1/10}$, and this is certainly acceptable when summed over all ℓ. If $d > k^{1/10}$, we instead apply the second bound $\binom{k}{2} - \ell \geqslant (d-1)\big(k - \frac{1}{2}d\big)$ from Lemma 5.3. This implies that (5.3) is bounded above by

$$2^{-\frac{1}{2}d(k-d)+O(k\log k)}. \tag{5.4}$$

As above (a consequence of the first bound in Lemma 5.3) we have $k - d \geqslant \sqrt{\ell} > \sqrt{k} > k^{1/10}$, and so $k^{1/10} \leqslant d \leqslant k - k^{1/10}$. But in this range we have $d(k-d) \gg k^{11/10}$. Thus (5.4) is bounded by $2^{-ck^{11/10}}$, which is again acceptable when summed over ℓ. This concludes the proof. □

Finally, we need to consider the possibility that $\binom{k}{2} - k^{4/3} < \ell \leqslant \binom{k}{2}$.

Lemma 5.5. *Suppose that $S \subset G$ is useful and that $X \subset S$ is a dissociated set with $|X| = k$. Then, conditioned upon the event that $E[X] \subset A$, the probability that there is some other dissociated set $Y \subset S$ with $|Y| = k$, $E[Y] \subset A$ and*

$$\binom{k}{2} - k^{4/3} < |E[X] \cap E[Y]| \leqslant \binom{k}{2}$$

is at most $1/10$.

In the regime covered here, the argument used in proving Lemma 5.4 breaks down. The key new observation here is that if $\ell = |E[X] \cap E[Y]|$ is nearly $\binom{k}{2}$, then in fact X and Y have substantial overlap as well, at least if X and Y are sufficiently dissociated. To prepare the ground for proving this, we first establish a couple of lemmas.

The first of these, which is really the key, has a tedious but basically straightforward proof which we outsource to Appendix A.

Lemma 5.6 (K_5 lemma). *Suppose that $X, Z \subset G$ are dissociated, that $|Z| \geqslant 5$ and that $E[Z] \subset E[X]$. Then $Z \subset X$.*

The next lemma is a graph-theoretic fact of a fairly standard type.

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Lemma 5.7. *Suppose that Γ is a graph on k vertices with ℓ edges. Then all but at most*

$$5\sqrt{\binom{k}{2}-\ell}$$

vertices of Γ lie in a subgraph of Γ isomorphic to the complete graph K_5.

Proof. The result is trivial if $\ell = \binom{k}{2}$, so suppose $\ell \leqslant \binom{k}{2} - 1$. Let $V \subset [k]$ be the set of vertices not contained in any copy of K_5 in Γ. If $|V| \leqslant 5$ then the result is immediate. Otherwise, certainly $\Gamma|_V$ does not contain any copy of K_5 and hence, by Turán's theorem, this graph has at most $\frac{3}{8}|V|^2$ edges. Hence $\binom{k}{2} - \ell$, the number of edges in the complement of Γ, is at least

$$\binom{|V|}{2} - \frac{3}{8}|V|^2 \geqslant \frac{1}{25}|V|^2.$$

The result follows. □

Corollary 5.8. *Suppose that X, Y are dissociated sets with $|X| = k$ and $|E[X] \cap E[Y]| = \ell$. Then*

$$|X \cap Y| \geqslant k - 5\sqrt{\binom{k}{2} - \ell}.$$

Corollary 5.9. *Suppose that X, Y are distinct dissociated sets of size k and that k is large. Then*

$$|E[X] \cap E[Y]| \leqslant \binom{k}{2} - \frac{1}{2}k.$$

Proof. Suppose the result is false. Then by Corollary 5.8 we have

$$|X \cap Y| \geqslant k - \frac{5}{\sqrt{2}}\sqrt{k} > k - 4\sqrt{k}.$$

Since X and Y are distinct and have the same size k, there is some y that lies in Y but not in X. By Lemma 5.6, y is not joined in Γ_Y to more than 3 of the vertices corresponding to $X \cap Y$ by an edge. Indeed, if it was joined to 4 such vertices y_1, y_2, y_3, y_4, then we could apply Lemma 5.6 with $Z := \{y_1, y_2, y_3, y_4, y\}$, concluding that $Z \subset X$ and in particular that $y \in X$, contrary to assumption. It follows that the degree of y in Γ_Y is no more than $3 + 4\sqrt{k}$. Thus the complement of Γ_Y contains at least $k - 1 - (3 + 4\sqrt{k}) > \frac{1}{2}k$ edges (if n, and hence k, is large enough). □

Proof of Lemma 5.5. In the statement of Lemma 5.5 we assumed that

$$\binom{k}{2} - k^{4/3} < \ell \leqslant \binom{k}{2},$$

but Corollary 5.9 in fact allows us to assume the stronger upper bound

$$\ell \leqslant \binom{k}{2} - \frac{1}{2}k.$$

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Recall that the graph Γ_Y is defined as follows. It is a graph on vertex set $[k]$, with i joined to j by an edge if and only if $y_i + y_j \in E[X]$. Recall also that d is the number of connected components of Γ_Y.

We claim that the number of choices of Y is at most $\binom{|S|}{d-1}2^{o(k)}$. To see this, we consider a variant of the skeleton of Γ_Y, which we will again call σ. By Corollary 5.8, the graph Γ_Y has one extremely large component containing a clique Ω of size at least $k - 5k^{2/3}$; every element of Y assigned to a vertex in Ω is an element of X. Take σ to be any collection of $\leqslant 5k^{2/3}$ edges such that the edges of Ω and σ span all the connected components of Γ_Y. Such a collection σ may be found using a greedy algorithm. The number of choices for σ is clearly at most

$$\binom{k}{2}^{5k^{2/3}} = 2^{o(k)},$$

and the number of choices for Ω is also $2^{o(k)}$.

Given Ω and σ, we must assign the set Y. All the vertices in Ω must be elements of X, so the number of choices for these vertices is at most the number of subsets of X of size at least $k - 5k^{2/3}$, which is again $2^{o(k)}$. In each of the other $d-1$ connected components, select one vertex. The values of these vertices must be elements of S, and they must obey the order relation that $y_i \prec y_j$ if $i < j$, so this gives at most $\binom{|S|}{d-1}$ choices.

The remaining unassigned vertices are connected to vertices already assigned (that is, to vertices in Ω or to the $d-1$ special vertices) by paths in σ. Each time we take an edge of σ from an assigned vertex i to a currently unassigned one j, the fact that $y_i + y_j \in E[X]$ gives us at most $\binom{k}{2}$ choices for y_j. Therefore the number of possible assignments of the remaining vertices, of which there are $k - |\Omega| - (d-1) \leqslant 5k^{2/3}$, is at most

$$\binom{k}{2}^{5k^{2/3}} = 2^{o(k)},$$

thereby concluding the proof of the claim.

Now from Lemma 5.3 we have

$$(d-1)\left(k - \frac{1}{2}d\right) \leqslant \binom{k}{2} - \ell < k^{4/3}; \tag{5.5}$$

since $d \leqslant k$, this immediately leads to the stronger bound

$$d \ll k^{1/3}. \tag{5.6}$$

Equations (5.5) and (5.6) together imply that

$$\frac{1}{2}(d-1)k \leqslant \left(\frac{1}{2} + o(1)\right)\left(\binom{k}{2} - \ell\right). \tag{5.7}$$

Since S is useful, we have

$$\binom{|S|}{k} 2^{-\binom{k}{2}} \leqslant n,$$

and so by Lemma C.2 we have

$$\binom{|S|}{d-1} \leqslant k^{d-1}\binom{|S|}{k}^{(d-1)/k} \leqslant 2^{\frac{1}{2}(d-1)k}k^{d-1}n^{(d-1)/k}$$
$$\leqslant 2^{\frac{1}{2}(d-1)k}k^{O(d)} \leqslant 2^{(1/2+o(1))(\binom{k}{2}-\ell)}k^d,$$

by (5.7). By (5.6), this is

$$2^{(1/2+o(1))(\binom{k}{2}-\ell)+o(k)}.$$

It follows from this and the earlier claim about the number of choices of Y that the probability that there is some $Y \subset S$, distinct from X, such that $|E[X] \cap E[Y]| = \ell$ is bounded by

$$2^{-(1/2-o(1))(\binom{k}{2}-\ell)+o(k)}.$$

When summed over the range

$$\binom{k}{2} - k^{4/3} < \ell \leqslant \binom{k}{2} - \frac{1}{2}k,$$

this is $o(1)$, as required. □

6. Further questions

For a wide selection of further questions we refer the reader to [2] or [6]. Here are two further questions.

Question 1. What is an asymptotic for $\chi(\Gamma_A)$, almost surely, when $A \subset \mathbb{F}_2^m$ is selected at random?

Question 2. If G is an abelian group of size n and if $A \subset G$ is selected at random, is

$$\chi(\Gamma_A)\omega(\Gamma_A) = (1+o(1))n$$

almost surely?

Appendix A: Proof of the K_5 lemma

Let us begin by recalling the statement of the K_5 lemma, Lemma 5.6.

Lemma 5.6. Suppose that $X, Z \subset G$ are dissociated, that $|Z| \geqslant 5$ and that $E[Z] \subset E[X]$. Then $Z \subset X$.

Proof. The proof of this is somewhat tedious, though straightforward. We begin by looking at sets Z of size 4 for which $E[Z] \subset E[X]$. We claim that such sets are of two types: *type I* in which $Z \subset X$, and *type II* in which $z_i = g - x_i$ for $i = 1,2,3,4$, where $2g = x_1 + x_2 + x_3 + x_4$ and the x_i are all elements of X.

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

Let us prove this claim. Suppose that $Z = \{z_1, z_2, z_3, z_4\}$ and that $z_i + z_j = x_{ij} + x'_{ij}$. We have the relations

$$x_{12} + x'_{12} + x_{34} + x'_{34} = x_{13} + x'_{13} + x_{24} + x'_{24} = x_{14} + x'_{14} + x_{23} + x'_{23}. \tag{6.1}$$

By dissociativity, $x_{12} \in \{x_{14}, x'_{14}, x_{23}, x'_{23}\}$, and there is no loss of generality in assuming that $x_{12} = x_{14}$. Thus

$$x'_{12} + x_{34} + x'_{34} = x'_{14} + x_{23} + x'_{23}.$$

By dissociativity, $x'_{12} \in \{x'_{14}, x_{23}, x'_{23}\}$. Now $z_2 \neq z_4$ and so $x'_{14} - x'_{12} = z_4 - z_2 \neq 0$. Hence we may assume without loss of generality that $x'_{12} = x'_{23}$. Dissociativity and (6.1) also imply that $x_{12} \in \{x_{13}, x'_{13}, x_{24}, x'_{24}\}$. There are two essentially different cases: *case 1* in which $x_{12} = x_{13}$, and *case 2* in which $x_{12} = x_{24}$.

Suppose we are in *case 1*. Then (6.1) implies that $x'_{12} + x_{34} + x'_{34} = x'_{13} + x_{24} + x'_{24}$, and so by dissociativity $x'_{12} \in \{x'_{13}, x_{24}, x'_{24}\}$. Since $z_2 \neq z_3$, we cannot have $x'_{12} = x'_{13}$. Without loss of generality, then, $x'_{12} = x'_{24}$. Referring back to (6.1), we see that $x_{34} + x'_{34} = x_{24} + x'_{13} = x_{23} + x'_{14}$. By dissociativity we have $x'_{13} \in \{x_{23}, x'_{14}\}$. However, if $x'_{13} = x'_{14}$ then we would have $z_3 = z_4$, a contradiction, and therefore $x'_{13} = x_{23}$ and $x'_{14} = x_{24}$. Writing $x_1 = x_{12}$, $x_2 = x'_{12}$, $x_3 = x_{23}$, $x_4 = x_{24}$, the above relations imply that $x_i + x_j = z_i + z_j$ for all distinct $i, j \in \{1,2,3,4\}$. Writing $w_i := x_i - z_i$, it follows that $w_i + w_j = 0$ for distinct i, j. This immediately implies that all of the w_i are zero. This is the *type I* situation.

Suppose now that we are in *case 2*, that is to say, $x_{12} = x_{24}$. Now (6.1) implies that $x_{23} + x'_{14} = x_{34} + x'_{34}$. By dissociativity we have $\{x_{23}, x'_{14}\} = \{x_{34}, x'_{34}\}$, and there is no loss of generality in assuming that $x_{23} = x_{34}$ and $x'_{14} = x'_{34}$. We also have, from (6.1), $x'_{24} + x_{13} + x'_{13} = x'_{12} + x_{23} + x'_{14}$ and so, by dissociativity, $x'_{24} \in \{x'_{12}, x_{23}, x'_{14}\}$. Since the z_i are all distinct we cannot have either $x'_{24} = x'_{12}$ or $x'_{24} = x_{23}$, so we must have $x'_{24} = x'_{14}$. Writing $x_1 = x_{23}$, $x_2 = x'_{14}$, $x_3 = x_{12}$ and $x_4 = x'_{12}$, we have $z_i + z_j + x_i + x_j = s$ for all distinct $i, j \in \{1,2,3,4\}$, where $s = x_1 + x_2 + x_3 + x_4$. Writing $w_i := x_i + z_i$, this implies that $w_i + w_j = s$ for all $i \neq j$. This easily implies that all of the w_i are equal to some g and that $2g = s$. This is the *type II* situation. This completes the proof of the claim.

Suppose now that $|Z| \geqslant 5$ and that $E[Z] \subset E[X]$. Every set $Z' \subset Z$ with $|Z'| = 4$ is of type I or II. We claim that if there is *any* set of type I then in fact the whole of Z is a subset of X, thereby concluding the proof. Suppose that Z' is of type I, thus $Z' = \{x_1, x_2, x_3, x_4\}$ with the x_i being elements of X. Suppose that $z \in Z \setminus Z'$. Then we have $z + x_i = x'_i + x''_i$ for $i = 1,2,3,4$ and some x'_i, x''_i. In particular $x_1 + x'_2 + x''_2 = x_2 + x'_1 + x''_1$, and so by dissociativity we have, without loss of generality, $x_1 = x'_1$. It then follows immediately that $z = x''_1$, thereby establishing the claim.

It remains to examine the possibility that every subset $Z' \subset Z$ with $|Z'| = 4$ is of type II. We claim this case cannot occur. We may suppose that Z' consists of elements $g - x_i$, $i = 1,2,3,4$, where $2g = x_1 + x_2 + x_3 + x_4$. Let $Z'' \subset Z$ be a different subset of size 4, intersecting Z' in $\{g - x_1, g - x_2, g - x_3\}$. Since this set is also of type II, we may label it so that it consists of elements $g' - x'_i$ with $2g' = x'_1 + x'_2 + x'_3 + x'_4$, where $g - x_i = g' - x'_i$ for $i = 1,2,3$. This last relation certainly implies that $x_1 + x'_2 = x'_1 + x_2$ and hence, by dissociativity, that $x_1 \in \{x'_1, x_2\}$. Since the elements of Z' are distinct, we cannot have $x_1 = x_2$. Therefore $x_1 = x'_1$, from which it follows that $g = g'$. It then follows that $x_2 = x'_2$ and $x_3 = x'_3$. Finally, since $x'_1 + x'_2 + x'_3 + x'_4 = 2g' = 2g = x_1 + x_2 + x_3 + x_4$, it follows that $x_4 = x_4'$. But then $g - x_4 = g' - x_4'$ lies in both Z' and Z'', contrary to assumption. This contradiction establishes the claim. □

Appendix B: On a result of Alon, Krivelevich and Sudakov

In this appendix we give a short proof that $\chi(\Gamma_A) \leqslant (2+o(1))\frac{n}{\log_2 n}$ almost surely if A is a random subset of an abelian group G, $|G| = n$. The argument is basically that of Alon, Krivelevich and Sudakov, but because we are dealing with Cayley sum graphs rather than arbitrary regular graphs, we can use a concise Fourier argument instead of an eigenvalue argument.

We assume some familiarity with the notation of the discrete Fourier transform as discussed in [9, Chapter 4], for example. Here we will be writing

$$\hat{f}(\gamma) := \mathbb{E}_{x\in G} f(x)\overline{\gamma(x)}$$

for $\gamma \in G^*$.

Proposition B.1. *Suppose that* $\sup_{\gamma\neq 1} |\hat{1}_A(\gamma)| \leqslant n^{-\eta}$. *Then*

$$\chi(\Gamma_A) \ll \left(\frac{1}{\eta} - o(1)\right)\frac{n}{\log_2 n}.$$

Proof. First of all note that

$$\sum_{x,x'\in S} 1_A(x+x') = n^2 \sum_{\gamma} \hat{1}_S(\gamma)^2 \hat{1}_A(\overline{\gamma}) = \frac{1}{2}|S|^2 + n^2 \sum_{\gamma\neq 1} \hat{1}_S(\gamma)^2 \hat{1}_A(\overline{\gamma}).$$

If $S \subset G$ has $|S| > n^{1-\eta}\log n$ (say) then the error term here can be efficiently bounded by Parseval's identity and the triangle inequality:

$$n^2 \left|\sum_{\gamma\neq 1} \hat{1}_S(\gamma)^2 \hat{1}_A(\overline{\gamma})\right| \leqslant n^{2-\eta} \sum_{\gamma} |\hat{1}_S(\gamma)|^2 = n^{2-\eta}\frac{|S|}{n} = o(|S|^2).$$

Thus

$$\sum_{x,x'\in S} 1_A(x+x') = \left(\frac{1}{2} + o(1)\right)|S|^2,$$

and in particular there is some $x \in S$ such that $x + x' \notin A$ for all x' in some subset $S' \subset S$ of size at least $(1/2 - o(1))|S|$.

By repeated application of this, it follows that any set $S \subset G$ of size at least $\frac{n}{\log^2 n}$ has a subset X of size at least $(\eta - o(1))\log_2 n$ with the property that $X \hat{+} X$ is disjoint from A.

From this the result follows straightforwardly by iteration, as in Section 2. □

If $A \subset G$ is a random set then almost surely we have

$$\sup_{\gamma\neq 1} |\hat{1}_A(\gamma)| \ll n^{-1/2+o(1)}.$$

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

This follows from a standard application of Bernstein's large deviation bound for each individual $\gamma \in G^*$, followed by a union bound over all $\gamma \neq 1$; see for example [9, Lemma 4.16]. Combining this with Proposition B.1 tells us that indeed $\chi(\Gamma_A) \ll (2+o(1))\frac{n}{\log_2 n}$ almost surely.

Appendix C: Some bounds on binomial coefficients

In this appendix we collect some bounds on binomial coefficients. These are of standard type, and we have often given crude bounds sufficient for our purposes rather than the strongest possible estimates.

Lemma C.1. *Let $n \geqslant k \geqslant 1$ be integers. Then*

$$\binom{2n}{k} \geqslant 2^k \binom{n}{k}.$$

Proof. We have

$$\frac{\binom{2n}{k}}{\binom{n}{k}} = \frac{2n}{n} \cdot \frac{2n-1}{n-1} \cdots \frac{2n-k+1}{n-k+1} \geqslant 2^k,$$

as required. □

Lemma C.2. *Let $n \geqslant k \geqslant d \geqslant 1$ be integers. Then*

$$\binom{n}{d} \leqslant k^d \binom{n}{k}^{d/k}.$$

Proof. First note that $n^k \leqslant k^k \binom{n}{k}$. Indeed,

$$\binom{n}{k} = \frac{n}{k} \cdot \frac{n-1}{k-1} \cdots \frac{n-k+1}{1} \geqslant \left(\frac{n}{k}\right)^k.$$

Therefore we have

$$\binom{n}{d} \leqslant n^d = (n^k)^{d/k} \leqslant \left(k^k \binom{n}{k}\right)^{d/k},$$

which is the stated bound. □

Lemma C.3. *Let $n \geqslant k \geqslant d \geqslant 1$ be integers. Then*

$$\binom{n}{d} \leqslant \left(\frac{k}{n-k}\right)^{k-d} \binom{n}{k}.$$

Proof. We have

$$\binom{n}{d} = \binom{n}{k} \cdot \frac{d+1}{n-d} \cdot \frac{d+2}{n-d-1} \cdots \frac{k}{n-k+1} \leqslant \left(\frac{k}{n-k}\right)^{k-d} \binom{n}{k}.$$ □

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press

References

[1] Agarwal, P. K., Alon, N., Aronov, B. and Suri, S. (1994) Can visibility graphs be represented compactly? In *ACM Symposium on Computational Geometry: San Diego, CA, 1993*. *Discrete Comput. Geom.* **12** 347–365.

[2] Alon, N. (2013) The chromatic number of random Cayley graphs. *European J. Combin.* **34** 1232–1243.

[3] Alon, N., Krivelevich, M. and Sudakov, B. (1999) List coloring of random and pseudo-random graphs. *Combinatorica* **19** 453–472.

[4] Alon, N. and Spencer, J. H. (2000) *The Probabilistic Method*, second edition, Wiley Interscience.

[5] Bollobás, B. (1988) The chromatic number of random graphs. *Combinatorica* **8** 49–55.

[6] Christophides, D. Random Cayley graphs. In *Midsummer Combinatorial Workshop 2011*, to appear. `http://www.christofides.org/Papers/mcw11.pdf`

[7] Green, B. J. (2005) Counting sets with small sumset, and the clique number of random Cayley graphs. *Combinatorica* **25** 307–326.

[8] Green, B. J. and Morris, R. (2016) Counting sets with small sumset and applications. *Combinatorica* **36** 129–159.

[9] Tao, T. C. and Vu, V. H. (2006) *Additive Combinatorics*, Vol. 105 of *Cambridge Studies in Advanced Mathematics*, Cambridge University Press.

https://doi.org/10.1017/S0963548316000304 Published online by Cambridge University Press