Instituto Nacional de Matemática Pura e Aplicada

Sparse random subsets of combinatorial objects

Maurício Collares Neto

Doctoral Thesis

Advisor: Robert Morris

November 2014

AGRADECIMENTOS

Primeiramente, agradeço à minha família. Em especial, gostaria de agradecer muito à minha mãe, Creusa, minhas irmãs, Marília e Sofia, e minha avó Lila por me apoiarem constantemente e por incentivarem meus estudos. Foi muito difícil sacrificar tempo com vocês, e agradeço por entenderem.

A meu orientador, Rob Morris, que, além de ser criativo e inteligente, é amigável, prestativo, paciente e empolgado para dividir tal criatividade e inteligência com seus alunos e colaboradores. Foi uma honra poder trabalhar contigo durante meu doutorado.

À Alice, cujo amor e amizade tornaram o fim do doutorado mais fácil e fizeram de 2014 um ano muito especial.

Ao grupo de combinatória do IMPA: Ao professor Roberto Imbuzeiro e ao Neal, Paul e Simon pelos muitos momentos de aprendizado. Ao professor Gugu, sempre muito prestativo, pelas muitas conversas e dúvidas tiradas. Graças ao esforço dos professores Gugu e Roberto Imbuzeiro, foi possível a realização de um exame de qualificação em combinatória, fato pelo qual sou muito grato.

Ao Bruno, amigo que esteve presente em muitos dos momentos, bons e ruins, da minha jornada de pós-graduando, desde os tempos de Análise na Reta. Ao Ricardo e Susana, pela alegria e amizade que tornaram minha vida no Rio muito mais agradável. Às demais pessoas com quem morei no Rio, em especial a Alan Reyes, Liev, Giulia, Luiz Paulo e Mateus. Ao Lucas, Karina, Sergio e Javier pela ajuda salvadora na época do exame.

À várias pessoas de Sergipe: Ao professor Valdenberg por um trabalho admirável de ensino e divulgação da matemática moderna. Aos colegas de Projeto, em particular a Reinan e Ana Maria. Ao professor Fábio, que me incentivou no ensino médio e me apresentou à matemática. Aos professores Lino e Dora, por acreditarem em mim desde bem cedo.

Ao IMPA, pelo excelente ambiente de trabalho. Minha reação inicial, ao chegar aqui pela primeira vez, foi sair correndo feliz pela biblioteca; apesar de não ter mais essa reação específica, este local, sua diversidade cultural e matemática continuam sendo fonte de inspiração e empolgação. Aos funcionários do IMPA, pela competência, dedicação e simpatia, mesmo depois de eu perder o prazo de inscrição pela n-ésima vez.

À CAPES e ao IMPA, por proporcionarem a oportunidade de concentrar-me nos estudos sem preocupações financeiras.

ABSTRACT

This thesis is concerned with the extremal properties and typical structure of sparse random combinatorial objects.

The first chapter, which is joint work with Morris, deals with a sparse random variant of a generalisation of Sperner's theorem. Denoting by $\mathcal{P}(n,p)$ the p-random subset of the power set of $\{1,\ldots,n\}$, we show that, if $pn \to \infty$, the largest subset of $\mathcal{P}(n,p)$ containing no k-chain has size $(k-1+o(1))p\binom{n}{n/2}$ with high probability. The case $k=2$ confirms a conjecture of Osthus.

The second chapter, which is joint work with Bushaw, Morris and Smith, focuses on a probabilistic result in additive combinatorics. We determine, for any even-order abelian group G, a sharp threshold for the following property: Each maximum-size sum-free subset of a p-random subset of G is contained in a maximum-size sum-free subset of the whole of G. This strengthens a result of Balogh, Morris and Samotij.

The third chapter, which is joint work with Balogh, Bushaw, Liu, Morris and Sharifzadeh, contains a result on the typical structure of graphs in a certain family. We prove that, for $r \leqslant (\log n)^{1/4}$, almost every K_{r+1}-free graph on n vertices is r-partite. This generalizes a result of Kolaitis, Prömel and Rothschild, who obtained the same result for fixed r, and strengthens a result of Mousset, Nenadov and Steger, who computed the number of K_{r+1}-free graphs for the same range of r.

Keywords: sparse random problems, probabilistic combinatorics, hypergraph container method, thresholds.

RESUMO

Esta tese lida com propriedades extremais e com a estrutura típica de objetos combinatoriais esparsos aleatórios.

O primeiro capítulo, relativo a trabalho conjunto com Morris, trata de uma versão esparsa aleatória de uma generalização do teorema de Sperner. Denotando por $\mathcal{P}(n, p)$ o conjunto p-aleatório da família de todos os subconjuntos de $\{1, \dots, n\}$, mostramos que, se $pn \to \infty$, o maior subconjunto de $\mathcal{P}(n, p)$ sem k-cadeias tem tamanho $(k - 1 + o(1))p\binom{n}{n/2}$ com alta probabilidade. O caso $k = 2$ confirma uma conjectura de Osthus.

O segundo capítulo, relativo a trabalho conjunto com Bushaw, Morris e Smith, foca num resultado probabilístico em combinatória aditiva. Determinamos, para qualquer grupo abeliano G de ordem par, um limiar *sharp* para a seguinte propriedade: Todo subconjunto sem somas de tamanho máximo de um subconjunto p-aleatório de G está contido num subconjunto sem somas de tamanho máximo relativo a todo o G. Tal teorema melhora um resultado de Balogh, Morris e Samotij.

O terceiro capítulo, relativo a trabalho conjunto com Balogh, Bushaw, Liu, Morris e Sharifzadeh, contém um resultado sobre a estrutura típica dos grafos de uma certa família. Provamos que, para $r \leqslant (\log n)^{1/4}$, quase todo grafo (com n vértices) sem K_{r+1} é r-partido. Isso generaliza um resultado de Kolaitis, Prömel e Rothschild, que mostraram o mesmo resultado no caso em que r é fixo, e melhora um resultado de Mousset, Nenadov e Steger, que computaram o número de grafos sem K_{r+1} com as mesmas restrições sobre r que usamos no nosso teorema.

Palavras-chave: problemas esparsos aleatórios, combinatória probabilística, método dos containers em hipergrafos, limiares.

INTRODUCTION

A vibrant area of research in combinatorics, especially in the last 20 years, concerns the formulation and proof of probabilistic and counting analogues of classical extremal results such as Turán's theorem and Szemerédi's theorem. More concretely, we are interested in problems such as "what is the typical structure of a subset of $[n]$ containing no solution to a given equation?" and "what is the largest H-free subgraph of $G_{n,p}$?".

A representative example of a "typical structure" result in the context of graphs is a theorem of Erdős, Kleitman and Rothschild, which states that almost all triangle-free graphs are bipartite. Usually, a first step towards determining the precise structure of objects avoiding some forbidden structure is to prove a counting result, i.e. to try to determine asymptotic bounds on the total number of such objects with and without the relevant structure. The famous conjecture of Cameron and Erdős [21], which states that there aren't many more subsets of $[n]$ with no solution to the equation $x + y = z$ than the obvious ones, provides an example of this.

The study of such questions gave rise to the so-called "sparse random problems", which deal with proving extremal and Ramsey-type results on a sparse random ground set. For example, suppose a particular property holds for the largest H-free subgraph of K_n. For what values of p can we prove that it also holds for the largest H-free subgraph of $G_{n,p}$?

We will now describe some relevant problems in more detail. In order to do so, we must go back and begin at the beginning.

1.1 EXTREMAL PROBLEMS

The first example of an extremal result in combinatorics is provided by a century-old result known as Mantel's theorem [67].

Theorem 1.1.1 (Mantel, 1907)**.** *Any n-vertex graph G with more than $\lfloor n/2 \rfloor \lceil n/2 \rceil$ edges contains a triangle.*

Although the above theorem is elementary, it is a starting point for several generalisations of great importance in the field. For example, the following famous result of Turán [88] replaces "triangle" in the above theorem by larger complete graphs. It is considered by many (e.g. [16]) to be the founding theorem of extremal graph theory.[1]

[1]Erdős almost initiated this area five years earlier when he, in the course of proving a number-

Theorem 1.1.2 (Turán, 1941). *A K_{r+1}-free graph G on n vertices has at most $e(T_{n,r})$ vertices, where $T_{n,r}$ is the r-partite complete graph with parts of size as equal as possible.*

Turán's theorem is the first step towards asymptotically determining the extremal numbers $\mathrm{ex}(n, H)$, which are defined to be the maximum number of edges in a n-vertex graph containing no copy of H. In 1946, Erdős and Stone made major progress in this task by computing the extremal number of any graph in terms of its chromatic number.

Theorem 1.1.3 (Erdős-Stone, 1946). *Let H be a graph with $\chi(H) \geqslant 2$. Then*

$$\mathrm{ex}(n, H) = \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right)\binom{n}{2}$$

This completely determines the asymptotic order of growth of $\mathrm{ex}(n, H)$ for every graph with $\chi(H) \geqslant 3$. Determining the asymptotic order of growth of $\mathrm{ex}(n, H)$ in the bipartite case is a major open problem, currently solved only in a few particular cases.

Between the formulation of Mantel's theorem and the founding of extremal graph theory, another important area of extremal combinatorics was born through the proof, by Sperner, of a Mantel-like result. A k-*chain* is simply a k-tuple of nested sets $A_1 \subsetneq \cdots \subsetneq A_k$, and we say a family $\mathcal{A}$ of sets is an *antichain* if it contains no 2-chain.

Theorem 1.1.4 (Sperner, 1928). *Let $\mathcal{A} \subset \mathcal{P}([n])$ be an antichain. Then $|\mathcal{A}| \leqslant \binom{n}{\lfloor n/2 \rfloor}$.*

Much in the same way Mantel's theorem extends to larger cliques, this admits a generalisation to larger chains, as was shown by Erdős [29]. In order to state it, note that $\mathcal{P}([n])$ can be decomposed into $n+1$ subfamilies of equally-sized subsets, and call the largest k such subfamilies[2] the k *middle layers of* $\mathcal{P}([n])$.

Theorem 1.1.5 (Erdős, 1945). *Any family of sets $\mathcal{A} \subset \mathcal{P}([n])$ containing more elements than the $k-1$ middle layers of $\mathcal{P}([n])$ contains a k-chain. In particular,*

$$|\mathcal{A}| \leqslant (k-1)\binom{n}{n/2}.$$

Although it is slightly out of the scope of this thesis to discuss many of the deterministic results from the beautiful area of Ramsey theory, we must mention the following cornerstone of combinatorics. This result, along with further developments by Erdős

theoretical result, showed that a n-vertex C_4-free graph can have at most $O\left(n^{3/2}\right)$ edges. Here's how he tells the story in [32]: "Being struck by a curious blindness and lack of imagination, I did not at the time extend the problem from C_4 to other graphs and thus missed founding an interesting and fruitful new branch of graph theory".

[2]Some subfamilies will have equal cardinality. An ordering of those can be chosen arbitrarily.

and Székeres [39], is responsible for kickstarting some of the most interesting branches of modern combinatorics.

Theorem 1.1.6 (Ramsey, 1930). *For any $s, t \in \mathbb{N}$, there exists an R such that any red-blue colouring of the edges of the complete graph K_R contains either a red K_s or a blue K_t.*

Ramsey-type theorems (which guarantee the existence of a monochromatic structure no matter the way an object is coloured) are often studied in parallel to density-type theorems (which guarantee the existence of a structure in any sufficiently dense object), and as such they will feature prominently below, when we describe the origins of sparse random problems.

Extremal and Ramsey-type problems in arithmetic combinatorics

Combinatorial problems are of great importance in number theory, and this was true long before the term "arithmetic combinatorics" was coined. One of the first problems in the area was studied by Schur [83], who proved the following theorem and used it as a tool to show that, for every $n \in \mathbb{N}$, the equation $a^n + b^n \equiv c^n \pmod{q}$ has a non-trivial solution for infinitely many primes q.

Theorem 1.1.7 (Schur, 1917). *Any finite colouring of $\mathbb{N}$ contains a monochromatic solution to the equation $x + y = z$.*

The above result does not follow from a density-type result, in the sense that it is not enough to look at the largest colour class to find a monochromatic solution to the equation. Indeed, the odds have density $1/2$ and are *sum-free*, that is, admit no solution to the equation $x + y = z$.

A related Ramsey-type result was conjectured by Baudet and proved by van der Waerden [90] with the help of Artin and Schreier[3]. Unlike Schur's theorem, this result admits a non-trivial density version, as is now well-known (see Theorem 1.1.9 below).

Theorem 1.1.8 (van der Waerden, 1927). *Any finite colouring of $\mathbb{N}$ contains a k-term monochromatic arithmetic progression for any $k \in \mathbb{N}$.*

A system of linear equations $Ax = 0$ is *partition-regular* if every finite colouring of $\mathbb{N}$ contains a monochromatic solution of $Ax = 0$. In 1933, Rado [72] generalised both theorems by providing a linear-algebraic condition for a system of linear equations to be partition-regular.

[3] See [91], reprinted in [85], for a fascinating account of the discovery of this proof.

Arguably the most famous problem in the area is the density version of van der Waerden's result, conjectured by Erdős and Turán [40] in 1936 and now known as Szemerédi's theorem. Roth [77] first proved it for $k = 3$ using Fourier analysis, and Szemerédi [87] gave a fully combinatorial proof of the general case using his Regularity Lemma.

Theorem 1.1.9 (Szemerédi, 1975)**.** *Any set* $A \in \mathbb{N}$ *satisfying*

$$\limsup_{n\to\infty} \frac{|A \cap [n]|}{n} > 0$$

contains a k*-term arithmetic progression for any* $k \in \mathbb{N}$*.*

Frankl, Graham and Rödl [42] noticed that Szemerédi's theorem implies that an irreducible partition-regular system admits a monochromatic solution in sets of arbitrarily small density if and only if the columns of A sum up to the zero vector.

It is difficult to pass up the opportunity to mention that Erdős [31] conjectured the following generalisation of Szemerédi's theorem, as a possible way of proving that the primes contain arbitrarily long arithmetic progressions. The general conjecture, stated below, is one of the main open problems in combinatorics.

Conjecture 1.1.10 (Erdős)**.** *Any set* $A \subset \mathbb{N}$ *satisfying*

$$\sum_{a\in A} \frac{1}{a} = \infty$$

contains a k*-term arithmetic progression for any* $k \in \mathbb{N}$*.*

A major result of Green and Tao [50] showed that this conjecture is true if A is the set of prime numbers (the original motivation for the conjecture). They did so by showing so-called "transference principles", inspiring a major development in the study of sparse random problems we will discuss in Section 1.5.

1.2 STABILITY RESULTS AND TYPICAL-STRUCTURE PROBLEMS

A different direction in which to take extremal results such as that of Turán is to prove a *stability theorem*, that is, to show that an object almost as big as the extremal example must possess some additional structure. For example, Erdős and Simonovits (see [84]) showed the example of Turán has the stability property: Any n-vertex K_{r+1}-free graph whose density is sufficiently close to maximum looks like $T_{n,r}$.

Theorem 1.2.1 (Erdős–Simonovits, 1968)**.** *For every $r \in \mathbb{N}$ and $\varepsilon > 0$, there exists a $\delta > 0$ such that every K_{r+1}-free n-vertex graph G with*

$$e(G) \geqslant \left(1 - \frac{1}{r} - \delta\right)\binom{n}{2}$$

can be turned into $T_{n,r}$ by adding and removing at most εn^2 edges.

The above theorem of Erdős and Simonovits just provides a characterisation of all *close-to-extremal* K_{r+1}-free graphs, mainly because there is very little we can say structure-wise for *all* K_{r+1}-free graphs, especially ones with few edges. Surprisingly, however, Erdős, Kleitman and Rothschild [36] showed that it is possible to provide a finer characterisation if we discard an asymptotically negligible proportion of such graphs.

Theorem 1.2.2 (Erdős–Kleitman–Rothschild, 1976)**.** *Almost all n-vertex triangle-free graphs are bipartite.*

As we remarked earlier, the first step towards showing that a result holds for almost all graphs is to prove a *counting result*. For example, Erdős, Kleitman and Rothschild showed, in the same paper, the following result for $r \geqslant 2$.

Theorem 1.2.3 (Erdős–Kleitman–Rothschild, 1976)**.** *For $r \geqslant 2$, let $F_{r+1}(n)$ be the number of n-vertex graphs containing no K_{r+1}. Then*

$$\log_2 F_{r+1}(n) = \left(1 - \frac{1}{r}\right)\binom{n}{2} + o(n^2). \tag{1.1}$$

In particular, if $M_r(n)$ denotes the number of n-vertex r-partite graphs, then

$$\lim_{n\to\infty} \frac{\log_2 F_{r+1}(n)}{\log_2 M_r(n)} = 1.$$

This result is, as expected, weaker than an "almost all"-type result. Kolaitis, Prömel and Rothschild [60] announced in 1985 the stronger typical-structure result (see [61] for their proof).

Theorem 1.2.4 (Kolaitis–Prömel–Rothschild, 1987)**.** *Let $r \geqslant 3$ be fixed. Then almost all K_{r+1}-free graphs on n vertices are r-partite. That is, $F_{r+1}(n) = (1 + o(1)) \cdot M_r(n)$.*

The history of this problem goes on, and it again exemplifies the interplay between stability, counting and typical-structure problems. In 1986, Erdős, Frankl, and Rödl [33] asked if forbidding a graph H with $\chi(H) = r$ has the same effect as forbidding a copy of K_r. They first showed a stability theorem under such conditions (cf. Theorem 1.2.1).

Theorem 1.2.5 (Erdős–Frankl–Rödl, 1986)**.** *Let* H *be a graph with* $\chi(H) = r+1 \geqslant 3$*. For every* $\varepsilon > 0$*, there exists a* $\delta > 0$ *such that every* H*-free* n*-vertex graph* G *with*

$$e(G) \geqslant \left(1 - \frac{1}{r} - \delta\right)\binom{n}{2}$$

can be made r*-partite by the deletion of at most* εn^2 *edges.*

Using this stability result, they showed the following counting result, which generalises Theorem 1.2.3.

Theorem 1.2.6 (Erdős–Frankl–Rödl, 1986)**.** *Let* H *be a fixed graph with* $\chi(H) \geqslant 3$*. The number of* H*-free* n*-vertex graphs is*

$$2^{(1+o(1))\,\mathrm{ex}(n,H)} = 2^{\left(1-\frac{1}{\chi(H)-1}\right)\binom{n}{2}+o(n^2)}.$$

This result was improved by Balogh, Bollobás and Simonovits [10], who showed that, for any graph H with $\chi(H) \geqslant 3$, there exists $\alpha > 0$ such that one may replace the $o(n^2)$ term in the exponent of the above theorem by an $O(n^{2-\alpha}\log n)$ term. They also conjectured that this result is sharp, which they later proved in [11] by establishing a strong Erdős–Simonovits-type result.

In a 2013 preprint, Mousset, Nenadov and Steger [70] generalised the counting result of Erdős, Kleitman and Rothschild, Theorem 1.2.3, by allowing the clique size to grow. Note that, in order to get a non-trivial statement in the case where the clique size goes to infinity, the error term must not overshadow the $\frac{1}{r}\binom{n}{2}$ term in equation (1.1). Mousset, Nenadov and Steger proved the result below, which has this property.

Theorem 1.2.7 (Mousset–Nenadov–Steger, 2013+)**.** *Let* $r = r(n) \leqslant (\log n)^{1/4}$*. Then, denoting by* $F_{r+1}(n)$ *the number of* K_{r+1}*-free graphs on* n *vertices, we have*

$$\log_2 F_{r+1}(n) = \left(1 - \frac{1}{r}\right)\binom{n}{2} + o\left(\frac{n^2}{r}\right).$$

In Chapter 4, which is joint work with József Balogh, Neal Bushaw, Hong Liu, Robert Morris and Maryam Sharifzadeh, we strengthen this result by extending the typical-structure result of Kolaitis, Prömel and Rothschild, Theorem 1.2.4, to non-constant values of r in a similar manner. This is the content of Theorem 4.1.1, which we reproduce below.

Theorem 1.2.8 (Balogh–Bushaw–CN–Liu–Morris–Sharifzadeh, 2014+)**.** *Let* $r = r(n) \in \mathbb{N}$ *be a function satisfying* $r \leqslant (\log n)^{1/4}$ *for every* $n \in \mathbb{N}$*. Then almost all* K_{r+1}*-free graphs on* n *vertices are* r*-partite. Formally,*

$$\lim_{n\to\infty} \frac{F_{r+1}(n)}{M_r(n)} = 1,$$

where $F_{r+1}(n)$ *denotes the number of* K_{r+1}*-free graphs and* $M_r(n)$ *denotes the number of* r*-partite graphs on* n *vertices.*

Our result uses the method of Balogh, Bollobás and Simonovits together with hypergraph container methods. We also prove and use a new supersaturation result, Theorem 4.1.2 (reproduced below), which is a generalisation of Theorem 1.2.1. It allows us to find many copies of K_{r+1} in any graph G which cannot be made r-partite by the deletion of few edges. It is optimal up to a factor of e^r.

Theorem 1.2.9 (Balogh–Bushaw–CN–Liu–Morris–Sharifzadeh, 2014+)**.** *For every* $n, r, t \in \mathbb{N}$*, the following holds. Every graph* G *on* n *vertices which cannot be made* r*-partite by the deletion of* t *edges contains at least*

$$\frac{n^{r-1}}{e^{2r} \cdot r!} \left(e(G) + t - \left(1 - \frac{1}{r}\right) \frac{n^2}{2} \right)$$

copies of K_{r+1}*.*

Typical-structure problems in arithmetic combinatorics

Problems about typical structure also play a central role in many other areas. For example, one of the most famous problems in arithmetic combinatorics is the Cameron–Erdős conjecture [21], which states that the number of sum-free subsets of $[n]$ is $O(2^{n/2})$. The conjecture is optimal up to a constant factor, as all subsets of $\{1, 3, \ldots\} \cap [n]$ and of $\{\lfloor n/2 \rfloor + 1, \ldots, n\}$ are sum-free. It was proven by Sapozhenko [79] (see also [80]) and independently by Green [48]. More precisely, they proved the following typical-structure result[4].

Theorem 1.2.10 (Sapozhenko, 2003 and Green, 2004)**.** *Almost all sum-free subsets of* $[n]$ *either consist entirely of odd numbers or are contained in* $\{\lceil (n+1)/3 \rceil, \ldots, n\}$*.*

[4] We note that Alon, Balogh, Morris and Samotij [4] recently proved a stronger result, characterising the typical structure of a sum-free m-subset of $[n]$.

Since Cameron and Erdős had previously shown that the number of sum-free elements of $\{\lceil (n+1)/3 \rceil, \ldots, n\}$ is exactly $c_n 2^{n/2}$, where $(c_{2n})_{n=1}^{\infty}$ and $(c_{2n-1})_{n=1}^{\infty}$ are convergent sequences, the typical-structure result of Sapozhenko and Green implies the Cameron–Erdős conjecture.

A related counting result for finite groups was first shown by Alon [1]. He proved that the size of $SF(G)$, the family of all sum-free sets of a n-element abelian group G, is at most $2^{(1/2+o(1))n}$. A sharper bound was proven in the abelian case by Lev, Łuczak and Schoen [63].

Theorem 1.2.11 (Lev–Łuczak–Schoen, 2001)**.** *There exists $\delta > 0$ with the following property. Let G be an n-element abelian group with canonical decomposition*

$$G \cong \mathbb{Z}_{2^{a_1}} \oplus \cdots \oplus \mathbb{Z}_{2^{a_k}} \oplus J,$$

where $1 \leqslant a_1 \leqslant \cdots \leqslant a_k$ and J is an odd-order group. Then

$$|SF(G)| = \left(2^k - 1\right) 2^{n/2} + O\left(2^{(1/2-\delta)n}\right).$$

While proving the above result, they showed the following stability theorem for sum-free subsets (see also [49]). We say a triple (x, y, z) is a *Schur triple* if $x + y = z$.

Theorem 1.2.12 (Lev–Łuczak–Schoen, 2001)**.** *Let G be an n-element abelian group. Any set $A \subset G$ with size $|A| \geqslant (1/3 + \varepsilon)n$ and at most $\varepsilon^3 n^2/27$ Schur triples contains a sum-free subset S with $|A \setminus S| \leqslant \varepsilon n$.*

It turns out that sum-free subsets in a particular type of group satisfy an even stronger stability property, shown by Green and Ruzsa [49]: Unlike in Theorem 1.2.1, we don't need to delete any elements to turn close-to-extremal sets into extremal ones. We say a group G is of *type* I if $|G|$ is divisible by a prime $q \equiv 2 \pmod 3$, and we say G is of *type* $I(q)$ if q is the smallest such prime.

Theorem 1.2.13 (Green–Ruzsa, 2005)**.** *Let G be an n-element abelian group of type $I(q)$. If $A \subset G$ is a sum-free set satisfying*

$$|A| \geqslant \left(\frac{1}{3} + \frac{1}{3(q+1)}\right) n,$$

then A is contained in a maximum-size sum-free subset of G.

In order to clarify the size condition on the above theorem, we note that an older result of Diananda and Yap [28] shows that maximum sum-free subsets of an abelian group G of type $I(q)$ have size $\left(\frac{1}{3} + \frac{1}{3q}\right)|G|$.

1.3 SPARSE RANDOM PROBLEMS

In a nutshell, sparse random problems consist of generalising classic extremal and Ramsey-type results to a probabilistic set-up. The story of such problems, however, begins with an entirely deterministic folklore fact from graph theory.

Fact 1.3.1. *Any graph* G *contains a bipartite subgraph* H *with* $e(H) \geqslant e(G)/2$.

In 1983, Erdős and Nešetřil (see [30]) asked whether this result could be strengthened if we forbid copies of K_4.

Question 1.3.2 (Erdős–Nešetřil, 1983)**.** *Does every* K_4*-free graph* G *contain a bipartite subgraph* H *with* $e(G) \geqslant C \cdot e(H)$*, for some constant* $C > 1/2$*?*

Frankl and Rödl [43] answered this *in the negative* by taking $G_{n,p(n)}$ for $p(n) = n^{-1/2+\varepsilon}$ and, in their words, showing that "random graphs behave as complete graphs, i.e. they are like sparse complete graphs". More technically, they proved the following.

Theorem 1.3.3 (Frankl–Rödl, 1986)**.** *For any small* $\varepsilon > 0$*, let* $p(n) = n^{-1/2+\varepsilon}$*. Then the largest triangle-free subgraph of* $G_{n,p(n)}$ *has density* $\frac{1}{2} + o(1)$ *with high probability.*

Since replacing $G_{n,p(n)}$ in the above by K_n leads to an asymptotic version of Mantel's theorem, we will call this a *sparse random analogue* of Mantel's result. In fact, for this value of $p(n)$, the largest K_4-free subgraph of $G_{n,p(n)}$ satisfies the same property, which implies that Question 1.3.2 has a negative answer.

Two key aspects of the above theorem warrant more investigation. We will take a more careful look at them in the next two subsections.

Thresholds

A property of graphs is merely a subfamily of the family of all graphs. A function $f(n)$ is a threshold function for a property $\mathcal{A}$ if

$$\lim_{n\to\infty} \mathbb{P}(G_{n,p(n)} \in \mathcal{A}) = \begin{cases} 0 & \text{if} \quad p(n)/f(n) \to 0 \\ 1 & \text{if} \quad p(n)/f(n) \to \infty. \end{cases}$$

Moreover, we say a property $\mathcal{A}$ is non-trivial if some but not all graphs satisfy it.

The word *sparse* in the context of sparse random problems refers to the fact that, in this area, most interesting thresholds are functions $p(n)$ that go to zero as n goes to infinity. In fact, the case where p is constant is so uncommon that the parameter n is often omitted, and so $p(n)$ is usually just denoted by p. We will follow this practice from now on.

In 1987, Bollobás and Thomason [15] proved a general and important result about thresholds. They showed that *every* non-trivial monotone increasing[5] property has a threshold. In fact, they showed that $f(n) = \sup\{p : \mathbb{P}(G_{n,p} \in \mathcal{A}) \leqslant 1/2\}$ is always a threshold function.

Many of the global problems in combinatorics, such as the connectivity problem studied by Erdős and Rényi [37] in their seminal paper on random graphs, admit a stronger notion of threshold: a so-called *sharp threshold*. Formally, a threshold $f(n)$ is sharp if, for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(G_{n,p} \in \mathcal{A}) = \begin{cases} 0 & \text{if} \quad p \leqslant (1-\varepsilon)f(n) \\ 1 & \text{if} \quad p \geqslant (1+\varepsilon)f(n). \end{cases}$$

However, other properties, such as that of containing a fixed subgraph H, only admit *coarse* (i.e., not sharp) thresholds, as was first shown by Bollobás [17]. In a spectacular breakthrough, Friedgut [44] characterised the graph properties (families) that do not admit sharp thresholds. In his words, his theorem essentially means that "a family with a coarse threshold can be approximated by a family whose minimal graphs are all small". Thus, all properties with coarse thresholds are essentially local ones.

Obviously, properties also make sense in contexts other than graph theory. It is trivial to generalise all of the above definitions to those contexts.

Asymptotic and precise problems

The theorem proved by Frankl and Rödl is a sparse random version of Mantel's theorem, but only asymptotically. Indeed, the latter says the largest triangle-free subgraph and the largest bipartite subgraph of K_n have *precisely* the same size, with no error term.

Babai, Simonovits and Spencer [6] were the first to show a precise version of Theorem 1.3.3. We state a slightly weaker version of their result for simplicity.

Theorem 1.3.4 (Babai–Simonovits–Spencer, 1990)**.** *For* $p \geqslant 1/2$, *any maximum-size triangle-free subgraph of* $G_{n,p}$ *is bipartite with high probability.*

[5] We say a property $\mathcal{A}$ is monotone increasing if $G \in \mathcal{A}$ and $G \subset H$ imply $H \in \mathcal{A}$.

Precise versions of sparse random results generally use the asymptotic version as a starting point, and they are usually much harder to prove, involving technical estimates of an ad-hoc nature and using only standard probabilistic tools (such as Janson's inequality). In fact, the threshold for the above theorem was only determined (up to a constant factor) in 2014 by DeMarco and Kahn [27].

Ramsey problems

In the field of Ramsey theory, sparse random problems also play a significant role. In stating several results, we will use the following standard notation.

Definition 1.3.5. Let F and G be two graphs and $r \in \mathbb{N}$. The statement "any r-colouring of the edges of F contains a monochromatic copy of G" will be denoted by $F \to (G)_r$. The same statement for vertex colourings will be denoted by $F \xrightarrow{v} (G)_r$.

In 1967, Erdős and Hajnal [34] (see also [35]) proposed the following question.

Question 1.3.6 (Erdős–Hajnal, 1967). *For every $k \geqslant 3$ and $r \geqslant 2$, does there exist a K_{k+1}-free graph G satisfying $G \to (K_k)_r$?*

Folkman [41] proved the vertex-colouring version of the above result, and used this to provide an affirmative answer to the question of Erdős and Hajnal for the case $r = 2$. The given proof was a complicated inductive construction. This, in the words of Rödl and Ruciński, "made everyone believe that such graphs are very rare".

Although the method of proof of Theorem 1.3.3 also showed that $G(n, p) \to (K_3)_2$ holds with high probability for $p = n^{-1/2+\varepsilon}$, the first paper focusing solely on sparse random Ramsey properties was published by Łuczak, Ruciński and Voigt [66].

In it, they showed that, for the property $G_{n,p} \xrightarrow{v} (G)_r$, the change of behaviour happens when a typical vertex of $G_{n,p}$ is contained in a constant number of copies of G. To be more precise, define

$$m^*(G) = \max_{\substack{H \subset G \\ |H|>1}} \frac{e(H)}{v(H) - 1}$$

so that, for any given subgraph $H \subset G$, the expected number of copies of H containing a vertex v is $\Omega(p^{e(H)} n^{v(H)-1}) = \Omega\left(p^{m^*(G)} n\right)^{v(H)-1}$. They showed the following.

Theorem 1.3.7 (Łuczak–Ruciński–Voigt, 1992). *For any non-trivial graph* G *and* $r \geqslant 2$, *there exist positive constants* c *and* C *such that*

$$\lim_{n\to\infty} \mathbb{P}(G_{n,p} \xrightarrow{v} (G)_r) = \begin{cases} 0 & \textit{if} \quad p \leqslant cn^{-1/m^*(G)} \\ 1 & \textit{if} \quad p \geqslant Cn^{-1/m^*(G)}. \end{cases}$$

The above is one of the first results that fully represents the spirit of sparse random questions. They also showed that the result of Frankl and Rödl, Theorem 1.3.3, is essentially sharp by computing the actual threshold for the property $G(n,p) \to (K_3)_2$. This was then extended to an arbitrary number of colours by Rödl and Ruciński [74] in 1994, and to arbitrary graphs (in place of K_3) by the same authors [75] in 1995. This last general version is as follows, where, analogously to the vertex case,

$$m_2(G) = \max_{\substack{H\subset G\\ |H|>2}} \frac{e(H)-1}{v(H)-2}$$

is such that, for any subgraph $H \subset G$, a typical edge of $G_{n,p}$ is contained in at least $\Omega(p^{e(H)-1}n^{v(H)-2}) = \Omega\left(p^{m_2(G)}n\right)^{v(H)-2}$ copies of H.

Theorem 1.3.8 (Rödl–Ruciński, 1995). *For any graph* G *which is not a star forest*[6] *and any* $r \geqslant 2$, *there exist positive constants* c *and* C *such that*

$$\lim_{n\to\infty} \mathbb{P}(G_{n,p} \to (G)_r) = \begin{cases} 0 & \textit{if} \quad p \leqslant cn^{-1/m_2(G)} \\ 1 & \textit{if} \quad p \geqslant Cn^{-1/m_2(G)}. \end{cases}$$

In the same paper, they also showed a sparse random version of van der Waerden's theorem. Generalising the arrow notation, the statement "any r-colouring of F contains a monochromatic k-term arithmetic progression" by $F \to (AP_k)_r$.

Theorem 1.3.9 (Rödl–Ruciński, 1995). *For any* $k \geqslant 3$ *and* $r \geqslant 2$, *there exist positive constants* c *and* C *such that*

$$\lim_{n\to\infty} \mathbb{P}([n]_p \to (AP_k)_r) = \begin{cases} 0 & \textit{if} \quad p \leqslant cn^{-1/(k-1)} \\ 1 & \textit{if} \quad p \geqslant Cn^{-1/(k-1)}. \end{cases}$$

The sparse Turán problem

In 1995, Haxell, Kohayawaka and Łuczak [52] conjectured a sparse random version of the Erdős–Stone theorem, which they then proved for all cycles in two papers [52, 53].

[6]That is, a disjoint union of graphs of the form $K_{1,t}$ for some $t \in \mathbb{N}$.

The statement of the conjecture, now known as the sparse Turán problem, is as follows[7]. We denote by $\mathrm{ex}(G_{n,p}, H)$ the number of edges of the largest H-free subgraph of $G_{n,p}$.

Conjecture 1.3.10 (Haxell–Kohayakawa–Łuczak, 1995)**.** *Let H be a non-trivial graph, and let $0 < p \leqslant 1$ be such that $pn^{1/m_2(H)} \to \infty$. Then, with high probability,*

$$\mathrm{ex}(G_{n,p}, H) = \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right) e(G_{n,p}).$$

In 1997, Kohayakawa, Łuczak and Rödl [58] proved Conjecture 1.3.10 for $H = K_4$ (the $H = K_3$ case was essentially shown by Frankl and Rödl). In the same paper, they also conjectured the following sparse random version of Erdős–Simonovits stability theorem, and proved it for $H = K_3$.

Conjecture 1.3.11 (Kohayakawa–Łuczak–Rödl, 1997)**.** *Let H be a non-trivial graph, and let $0 < p \leqslant 1$ be such that $pn^{1/m_2(H)} \to \infty$. For every $\varepsilon > 0$ there exists a $\delta > 0$ such that, with high probability, any H-free subgraph $J \subset G_{n,p}$ with*

$$e(J) \geqslant \left(1 - \frac{1}{\chi(H) - 1} - \delta\right) e(G_{n,p})$$

can be made $(\chi(H) - 1)$-partite by the deletion of at most $\varepsilon e(G_{n,p})$ edges.

A common thread among all of the mentioned partial solutions for the above conjectures is a sparse version of the Szemerédi Regularity Lemma, independently discovered by Kohayakawa and Rödl [55]. With this technique, Kohayakawa, Łuczak and Rödl [59] were able to prove another major result: a sparse random version of Roth's theorem (cf. Theorem 1.3.9).

Theorem 1.3.12 (Kohayakawa–Łuczak–Rödl, 1996)**.** *For any $\alpha > 0$, there exists $C > 0$ such that, for $p \geqslant Cn^{-1/2}$, any subset $A \subset [n]_p$ of size $|A| \geqslant \alpha \cdot |[n]_p|$ contains a 3-term arithmetic progression with high probability.*

The KŁR results brought renewed interest to the area, effectively starting the systematic study of sparse random problems. The result is a vibrant and fruitful research area, which is going strong to this day. For readers interested in knowing more about sparse random problems in the context of graph theory, we recommend the excellent survey of Rödl and Schacht [76].

[7]Below the conjectured threshold, the 0-statements for these results are easy: As explained before, there are fewer copies of H than edges in $G_{n,p}$ when $pn^{1/m_2(H)} \to 0$, meaning that all copies of H can be deleted without affecting the asymptotic number of edges. Therefore the result, if true, would be best possible.

A sparse random Sperner theorem

Recall that Sperner's theorem, Theorem 1.1.4, says that no antichain of $\mathcal{P}([n])$ is larger than its middle layer. In 2000, Osthus [71] proved the following sparse analogue of Sperner's result. We denote by $\mathcal{P}(n, p)$ the p-random subset of $\mathcal{P}([n])$.

Theorem 1.3.13 (Osthus, 2000). *Let $0 \leqslant p = p(n) \leqslant 1$ be such that $pn/\log n \to \infty$. The largest subset $\mathcal{A} \subset \mathcal{P}(n, p)$ containing no 2-chains has size*

$$|\mathcal{A}| = (1 + o(1))\, p \binom{n}{n/2}.$$

with high probability.

He also observed that, for $pn \to C$, a second moment calculation tells us that $\mathcal{P}(n, p)$ contains an antichain of size

$$\left(1 + e^{-C/2} + o(1)\right) p \binom{n}{n/2}$$

with high probability. The gap between the two restrictions on the probability function led Osthus to conjecture that the conclusion of Theorem 1.3.13 is also true whenever $pn \to \infty$. By his observation, the conjecture is equivalent to saying that $f(n) = 1/n$ is a (coarse) threshold function for the sparse random version of Sperner's theorem.

In Chapter 2, which is joint work with Robert Morris, we generalise Osthus' conjecture to the k-chains case and prove the $1/n$ threshold for every $k \in \mathbb{N}$, effectively showing a sparse random analogue of Theorem 1.1.5. Our result, Theorem 2.1.1, is reproduced below.

Theorem 1.3.14 (CN–Morris, 2014+). *Let $2 \leqslant k \in \mathbb{N}$, and let $p = p(n)$ be such that $pn \to \infty$. Then the largest subset $\mathcal{A} \subset \mathcal{P}(n, p)$ containing no k-chain has size*

$$|\mathcal{A}| = (k - 1 + o(1))\, p \binom{n}{n/2}$$

with high probability.

In the following sections, we will look at fundamental advances in the area of sparse random problems in the last five years.

1.4 INDEPENDENT SETS IN HYPERGRAPHS

Recall that an *independent set* for a hypergraph $\mathcal{H}$ is a set $A \subset V(\mathcal{H})$ for which $e(\mathcal{H}[A]) = 0$, that is, a set A which contains no edge $e \in E(\mathcal{H})$.

Many of the combinatorial problems we saw can be restated as problems about independent sets in hypergraph families. The advantage of doing so might not be obvious at first, but we will study general theorems in the next few sections which allow us to deduce sparse random versions of problems about independent sets in hypergraphs. In fact, since all of the discussed combinatorial results are asymptotic and deal with increasingly large ground sets, the natural object to study for this purpose is a hypergraph family $(\mathcal{H}_n)_{n=1}^{\infty}$ instead of a single hypergraph.

In order to show that such restatements are possible, we will focus on two representative examples, the sparse Szemerédi and sparse Turán problems.

- The case of k-term arithmetic progressions (k-APs) is the simplest to model as an independent set problem. We first construct a family of k-uniform hypergraphs as follows:

  $$\begin{aligned} V(\mathcal{H}_n) &:= [n] \\ E(\mathcal{H}_n) &:= \{\{a_1, a_2, \ldots, a_k\} : (a_1, \ldots, a_k) \text{ is a } k\text{-AP}\} \end{aligned}$$

  Notice that Szemerédi's theorem (Theorem 1.1.9) says that for every $\delta > 0$ there exists $n_0 \in \mathbb{N}$ such that, for $n \geqslant n_0$, any set $X \subset V(\mathcal{H}_n)$ with $|X| \geqslant \delta v(\mathcal{H}_n)$ is not independent. In other words, any sequence of independent sets $I_n \subset V(\mathcal{H}_n)$ satisfies $|I_n| = o(n)$.

- The case of H-free graphs just has slightly more complicated terminology. Remember that we choose *edges* at random in the $G_{n,p}$ model, and for this reason, the vertices of our constructed hypergraph will be pairs of vertices representing possible edges of $G_{n,p}$.

  Also, since we want to talk about independent sets, the edges of our hypergraph must encode the restrictions of our problem, which are forbidden copies of H. Thus, we will let $m = e(H)$ and construct a m-uniform hypergraph in which the edges of $\mathcal{H}$ represent copies of H in K_n.

  $$\begin{aligned} V(\mathcal{H}_n) &= E(K_n) = \binom{[n]}{2} \\ E(\mathcal{H}_n) &= \{\{e_1, \ldots, e_m\} \text{ forming a copy of } H\} \end{aligned}$$

Notice that the theorem of Erdős and Stone (Theorem 1.1.3) states that for every $\delta > 0$ there exists $n_0 \in \mathbb{N}$ such that, for $n \geqslant n_0$, any set $X \subset V(\mathcal{H}_n)$ with $|X| \geqslant \left(1 - \frac{1}{\chi(H)-1} + \delta\right) v(\mathcal{H}_n)$ is not independent.

The sparse random version of both problems can be restated as follows: For appropriate p and large enough n, the corresponding statement about independent sets still holds with high probability if we replace $\mathcal{H}_n$ by the hypergraph induced by a p-random subset of $V(\mathcal{H}_n)$.

1.5 TRANSFERENCE THEOREMS

In 2009 and 2010, Schacht [82] and Conlon and Gowers [24] independently obtained powerful theorems which allow robust (in a sense we will clarify later) extremal results to be transferred to the sparse setting, therefore resolving a large number of open questions in the field. For example, they proved[8] Conjecture 1.3.10.

Theorem 1.5.1 (Schacht, 2009+, Conlon–Gowers, 2010+)**.** *Let* H *be a non-empty graph and* $\varepsilon > 0$*. Then there exist positive constants* c *and* C *such that*

$$\lim_{n\to\infty} \mathbb{P}\left(\operatorname{ex}(G_{n,p}, H) \leqslant \left(1 - \frac{1}{\chi(H)-1} + \varepsilon\right) e(G_{n,p})\right) = \begin{cases} 0 & \textit{if} \quad p \leqslant cn^{1/m_2(H)} \\ 1 & \textit{if} \quad p \geqslant Cn^{1/m_2(H)}. \end{cases}$$

Both methods require robust versions of the dense combinatorial result (also called “supersaturation results”), that is, a version that guarantees the existence of not just one but *many* copies of the desired structure whenever the ground set is a bit bigger than the extremal example. Such theorems usually follow by simple averaging arguments. For example, a robust version of Szemerédi’s theorem may be obtained by applying the averaging argument of Varnavides [89]. It looks like this.

Theorem 1.5.2 (Szemerédi, 1974)**.** *For every* $\varepsilon > 0$ *and* $k \in \mathbb{N}$*, there exists* $\delta > 0$ *such that, for large* n*, any set* $A \subset [n]$ *with* $|A| \geqslant \varepsilon n$ *has* δn^2 k*-term arithmetic progressions.*

Despite this commonality, the methods of Conlon–Gowers and Schacht use very different techniques, each one having its own strengths and weaknesses. We will briefly discuss the two methods below.

[8] To be precise, Conlon and Gowers [24] proved their density and stability results for strictly balanced graphs, those graphs G for which $m_2(G) > m_2(H)$ for any $H \subset G$. According to their paper, it is possible to adapt their idea to obtain the general result when $p \geqslant n^{-1/m_2(H)} \cdot (\log n)^c$, for some $c > 0$. The method of Schacht fully proves the stated results.

The work of Conlon and Gowers [24] is based on a functional transference principle[9]. This requires restating the relevant combinatorial theorem in terms of $[0,1]$-valued functions (instead of $\{0,1\}$-valued functions, i.e. sets) in discrete spaces. As an example, here is the functional version of Szemerédi's theorem. For convenience, we state the result in Z_n instead of in $[n]$.

Theorem 1.5.3 (Szemerédi, 1974). *For every $\varepsilon > 0$ and $k \in \mathbb{N}$, there exists $\delta > 0$ such that any function $g\colon \mathbb{Z}_n \to [0,1]$ with $\frac{1}{n}\sum_{x=0}^{n-1} g(x) \geqslant \varepsilon$ satisfies the inequality*

$$\frac{1}{n^2}\sum_{x=0}^{n-1}\sum_{d=0}^{n-1} g(x)g(x+d)\cdots g(x+(k-1)d) \geqslant \delta$$

Given a function supported on a sparse random set, their principle proceeds by associating to it a function corresponding to a dense object. The dense function has global *and local* properties which are close to the sparse one, meaning that dense results lead to proofs for the sparse case under some technical probability conditions. For example, the above functional restatement gives rise to the following theorem.

Theorem 1.5.4 (Schacht, 2009+, Conlon–Gowers, 2010+). *For any $\varepsilon > 0$ and $k \geqslant 2$, there exists $C > 0$ such that, for $p \geqslant Cn^{-1/(k-1)}$, any subset $A \subset [n]_p$ of size $|A| \geqslant \varepsilon \cdot |[n]_p|$ contains a k-term arithmetic progression with high probability.*

In the language of hypergraph families (see Section 1.4), Schacht quantified the concept of robustness by saying a property is robust if, for some α, any vertex set of density larger than α contains a positive proportion of the edges of $\mathcal{H}_n$. Formally, this is his definition.

Definition 1.5.5 (Schacht, 2009+). For $\alpha \geqslant 0$, a k-uniform hypergraph family $(\mathcal{H}_n)_{n=1}^{\infty}$ is α-dense if for every $\varepsilon > 0$ there exists $\delta > 0$ with the following property: Any set $U \subset V(\mathcal{H}_n)$ with $|U| \geqslant (\alpha+\varepsilon)v(\mathcal{H}_n)$ satisfies $e(\mathcal{H}_n[U]) \geqslant \delta e(\mathcal{H}_n)$ for large enough n.

In the case of arithmetic progressions, for example, Szemerédi's theorem says precisely that the corresponding hypergraph family is 0-dense.

In order to relate overlaps among edges of the hypergraph family to probabilistic restrictions for the 1-statement of a sparse random analogue, Schacht introduced the concept of (K,p)-boundedness. We note that, as usual, the probabilities p and q are functions of n (see Section 1.3).

[9] Transference principles have their roots in the previously mentioned work of Green and Tao [50]. A good introduction to the subject is the paper of Gowers [47].

Definition 1.5.6 (Schacht, 2009+)**.** For $0 \leqslant p \leqslant 1$ and $K \geqslant 1$, a k-uniform hypergraph family is (K, p)-bounded if for every $1 \leqslant i < k$ and every $q \geqslant p$,

$$\mathbb{E}\left(\sum_{v \in V(\mathcal{H}_n)} \deg_i(v, V(\mathcal{H}_n)_q)^2\right) \leqslant Kq^{2i} \frac{e(\mathcal{H}_n))^2}{v(\mathcal{H}_n)}$$

holds for sufficiently large n, where $V(\mathcal{H}_n)_q$ is a q-random subset of $V(\mathcal{H}_n)$ and

$$\deg_i(v, U) = |\{e \in E(\mathcal{H}_n) : |e \cap (U \setminus \{v\})| \geqslant i \text{ and } v \in e\}| .$$

Having those definitions, we can summarize Schacht's theorem in the following manner: The 1-statement of a sparse random analogue of a combinatorial result is true for $p \geqslant C(K, \alpha) \cdot p'$ whenever the corresponding hypergraph family is α-dense and (K, p')-bounded for some K and α.

For many problems, the definition of (K, p)-boundedness leads to thresholds which are optimal up to a constant. For example, a standard calculation shows that the hypergraph family for the sparse Szemerédi problem is $(K, n^{-1/(k-1)})$-bounded for large enough K. Also, the family for the sparse Turán problem is $(K, n^{-1/m_2(H)})$-bounded for large enough K.

Schacht's proof gives exponential bounds on the probability of failure, whereas the functional approach of Conlon and Gowers provides only polynomial bounds. On the other hand, the functional transference theorem of Conlon and Gowers also reflects local properties, and is thus strong enough to prove asymptotic counting results. In fact, some two-sided counting results can only hold with polynomially-decaying failure probabilities: Theorem 1.6 (ii) of [25], proved via the method of Conlon and Gowers, is one such result.

Having mentioned this, we take an opportunistic pause to marvel at one of their stability results, which we stated before as Conjecture 1.3.11. Since Samotij [78] refined the method of Schacht to allow for obtaining stability results, we attribute the result to both approaches.

Theorem 1.5.7 (Schacht, 2009+, Conlon–Gowers, 2010+ and Samotij, 2014)**.** *Let H be a graph with maximum degree at least two, and let $\varepsilon > 0$. Then there exist positive constants C and δ such that, for $p \geqslant Cn^{1/m_2(H)}$, every H-free subgraph of $G_{n,p}$ with more than*

$$\left(1 - \frac{1}{\chi(H) - 1} - \delta\right) e(G_{n,p})$$

edges can be made $(\chi(H) - 1)$-partite by the removal of at most $\varepsilon e(G)$ edges.

A sparse stability theorem for sum-free subsets

With the above techniques, the stability theorem for sum-free subsets in abelian groups, Theorem 1.2.13, can also be naturally transferred to a sparse random context. This was first shown by Balogh, Morris and Samotij in [13] (see also [78]).

Theorem 1.5.8 (Balogh–Morris–Samotij, 2014)**.** *For any small* $\delta > 0$ *and prime number* $q \equiv 2 \pmod 3$, *there exist constants* $C > 0$ *and* $\varepsilon > 0$ *with the following property. Let* G *be any* n*-element abelian group of type* $I(q)$*. If*

$$p \geqslant Cn^{-1/2}$$

then, with high probability, for every sum-free subset $B \subset G_p$ *with*

$$|B| \geqslant \left(\frac{1}{3} + \frac{1}{3q} - \varepsilon\right) p|G|$$

there exists a maximum-size sum-free subset $\mathcal{O}$ *of* G *with* $|B \setminus \mathcal{O}| \leqslant \delta pn$.

Motivated by the above result, we say a subset A of a group G is *sum-free good* if some maximum-size sum-free subset of A is contained in a maximum-size sum-free subset of G. Roughly speaking, this precise property (in the sense of Section 1.3) means that A “inherits” its sum-free subset structure from G.

In Chapter 3, which is joint work with Neal Bushaw, Robert Morris and Paul Smith, we deal with even-order abelian groups. For such a group G, we extend the above by calculating the sharp threshold for the property of a p-random subset of G being sum-free good with high probability. The location of the sharp threshold depends on the number of order 2 elements of G (see Fact 3.2.2).

It is arguably easier to introduce the results of Chapter 3 by talking about a sequence of groups $\mathbf{G} = (G_n)_{n=1}^{\infty}$, and we will do so in this Introduction, although our result is slightly more general. We start with this definition, which provides the constant used in our sharp threshold. For an abelian group G', let $r(G')$ be the number of elements x of G' satisfying $x = -x$.

Definition 1.5.9 (Bushaw–CN–Morris–Smith, 2013+)**.** A group sequence $\mathbf{G} = (G_n)_{n=1}^{\infty}$ is *well-behaved* if the two limits below exist.

$$\alpha(\mathbf{G}) := \lim_{n\to\infty} \frac{\log r(G_n)}{\log(|G_n|/2)} \qquad \text{and} \qquad \beta(\mathbf{G}) := \lim_{n\to\infty} \frac{r(G_n)}{|G_n|/2}$$

For a well-behaved sequence $\mathbf{G}$, define moreover

$$\lambda(\mathbf{G}) := \begin{cases} 1/3 & \text{if} \quad \alpha(\mathbf{G}) \leqslant 5/6 \\ \alpha(\mathbf{G}) - 1/2 & \text{if} \quad \alpha(\mathbf{G}) > 5/6 \text{ and } \beta(\mathbf{G}) = 0 \\ 2/\big(4 - \beta(\mathbf{G})\big) & \text{if} \quad \beta(\mathbf{G}) > 0. \end{cases}$$

In this context, our result is the following.

Theorem 1.5.10 (Bushaw–CN–Morris–Smith, 2013+). *Let $\boldsymbol{G} = (G_n)_{n=1}^{\infty}$ be a well-behaved sequence of even-order abelian groups, let $p \in (0,1)$ with $p \geqslant (\log n)^2/n$ and let A_n be a p-random subset of G_n. Then, for every $\varepsilon > 0$,*

$$\mathbb{P}(A_n \textit{ is sum-free good}) \to \begin{cases} 0 & \text{if} \quad p \leqslant (\lambda(\boldsymbol{G}) - \varepsilon)\sqrt{\frac{\log n}{n}} \\ 1 & \text{if} \quad p \geqslant (\lambda(\boldsymbol{G}) + \varepsilon)\sqrt{\frac{\log n}{n}}. \end{cases}$$

This answers a natural question arising from the previous work of Balogh, Morris and Samotij [13], who established the existence of a threshold for any sequence of even-order abelian groups[10] but showed a sharp threshold only for the particular case $G_n = \mathbb{Z}_{2n}$.

1.6 HYPERGRAPH CONTAINERS

In 2012, a different approach to the above problems was found by Balogh, Morris and Samotij [8] and independently by Saxton and Thomason [81], who provided a simple but powerful characterisation of the independent sets in a hypergraph. They showed that, if the edges of a hypergraph $\mathcal{H}$ are "well-distributed" (in a sense made precise below), then the independent sets of $\mathcal{H}$ are "clustered", in the sense that there exists a small family $\mathcal{C} \subset \mathcal{P}(V(\mathcal{H}))$ of *containers*, each significantly smaller than $V(\mathcal{H})$, such that every independent set of $\mathcal{H}$ is contained in some $C \in \mathcal{C}$.

We will state the main lemma in the notation of Balogh, Morris and Samotij [8, Proposition 3.1]. The version of Saxton and Thomason [81, Theorem 2.5] replaces the maximum degree conditions by a more elaborate version of uniformity expressed in terms of co-degree functions[11].

[10] The result they proved was slightly more general and worked for all sequences of groups of type $I(q)$, for any fixed prime $q \equiv 2 \pmod 3$. We focused on the even-order case because we thought it contained most of the essential difficulties of the problem.

[11] So far, there are no known applications for which only one of the versions work.

The Hypergraph Container Lemma. *For every* $k \in \mathbb{N}$ *and* $c > 0$*, there exists a* $\delta > 0$ *such that the following holds. Let* $\tau \in (0,1)$ *and suppose that* $\mathcal{H}$ *is a nonempty* k*-uniform hypergraph on* N *vertices such that*

$$\Delta_\ell(\mathcal{H}) \leqslant c \cdot \tau^{\ell-1} \frac{e(\mathcal{H})}{N} \tag{1.2}$$

for every $1 \leqslant \ell \leqslant k$*, where*

$$\Delta_\ell(\mathcal{H}) := \max_{|T|=\ell} |\{e \in E(\mathcal{H}) : T \subset e\}| \,.$$

Then there exist a family $\mathcal{C}$ *of subsets of* $V(\mathcal{H})$*, and a function* $f \colon \mathcal{P}(V(\mathcal{H})) \to \mathcal{C}$ *such that:*

(a) *For every independent set* I *there exists* $T \subset I$ *with* $|T| \leqslant k \cdot \tau N$ *and* $I \subset f(T)$*,*

(b) $|C| \leqslant (1-\delta)N$ *for every* $C \in \mathcal{C}$*.*

The power of this method comes from the fact that the Hypergraph Container Lemma can be iterated. That is, if we have any supersaturation result and we let $\mathcal{F}$ denote the family of hypergraphs for which this result holds, we can repeatedly apply the lemma to ensure none of the containers are in $\mathcal{F}$. Once this happens, the containers will typically either be small enough to easily count, or have a very special structure which we can exploit.

There are several differences between this result and the transference theorems discussed in the previous section. Although containers are frequently used to prove asymptotic results, its non-asymptotic nature provides a great deal of flexibility. For example, all of the constants in the theorem can be explicitly estimated, which allows for results in which the size of the hypergraph is not fixed. All of these advantages will be used to our favor in Chapter 4.

Moreover, this statement provides a meaningful deterministic counting assertion and not just a probabilistic one. Thus, it should come as no surprise that it is also significantly easier to obtain counting results for several of the problems we discussed. In fact, once the counting is done, a sparse random result often follows from a simple application of a concentration inequality such as Chernoff's (or even Markov's).

1.7 ORGANISATION OF THIS THESIS

The rest of this thesis is organised as follows. A simple description of each of the main results proved can be found in the sections above, near the corresponding theorem statement. A more detailed introduction to problem can be found in the corresponding chapter.

In Chapter 2, which is joint work with Robert Morris, we prove Theorem 1.3.14, which determines the threshold for the sparse random Sperner theorem, verifying a conjecture of Osthus, and generalises it to k-chains.

In Chapter 3, which is joint work with Neal Bushaw, Robert Morris and Paul Smith, we prove a slightly more general version of Theorem 1.5.10, which determines the sharp threshold for a problem related to sum-free sets of abelian groups.

In Chapter 4, which is joint work with József Balogh, Neal Bushaw, Hong Liu, Robert Morris and Maryam Sharifzadeh, we prove Theorem 1.2.8, which shows that asymptotically almost all K_{r+1}-free graphs are r-partite even in the case of unbounded $r = r(n)$. Along with this, we prove a new supersaturation result, Theorem 1.2.9, which guarantees the existence of many copies of K_r in graphs which cannot be made r-partite by the deletion of few edges.

MAXIMUM-SIZE ANTICHAINS IN RANDOM SETS

The work in this chapter is joint with Robert Morris. It is adapted from a preprint [23].

2.1 INTRODUCTION

One of the cornerstones of extremal set theory is the famous theorem of Sperner [86], who proved in 1928 that the largest antichain in $\mathcal{P}(n)$, the family of all subsets of $\{1, \dots, n\}$, has size $\binom{n}{n/2}$. In 1945, Erdős [29] generalised this result by showing that any family of sets larger than the $k-1$ middle layers of $\mathcal{P}(n)$ contains a k-chain.

The study of the random set-system $\mathcal{P}(n, p)$ was initiated in 1961 by Rényi [73], who determined the threshold for the event that $\mathcal{P}(n, p)$ is an antichain. More recently, Kreuter [62] and Kohayakawa, Kreuter and Osthus [57] studied the length of the longest chain in $\mathcal{P}(n, p)$, and Kohayakawa and Kreuter [56] and Osthus [71] studied the size of the largest antichain. In particular, Osthus [71] proved that (2.1) holds in the case $k = 2$ if $pn \gg \log n$, and conjectured that $pn \gg 1$ is sufficient. We note that this conjecture has also been proved independently by Balogh, Mycroft and Treglown [9], who moreover obtained a corresponding result for sparser random set systems, though again only in the case $k = 2$.

In this chapter we will prove a sparse random analogue of Erdős' theorem. More precisely, for every function $p \gg 1/n$ we will determine, with high probability, the (asymptotic) size of the largest subset of $\mathcal{P}(n, p)$, the p-random subset[1] of $\mathcal{P}(n)$, containing no k-chain. In the case $k = 2$, this confirms a conjecture of Osthus [71].

Theorem 2.1.1. *Let $2 \leqslant k \in \mathbb{N}$, let $p = p(n)$ be such that $pn \to \infty$. Then the largest subset $\mathcal{A} \subset \mathcal{P}(n, p)$ containing no k-chain has size*

$$|\mathcal{A}| = \big(k - 1 + o(1)\big)p\binom{n}{n/2} \tag{2.1}$$

with high probability as $n \to \infty$.

We remark that the bound on p is best possible, since the result fails to hold whenever $pn \to C$. Indeed, in this case Osthus [71] showed that, with high probability, the two middle layers of $\mathcal{P}(n, p)$ contain an antichain $\mathcal{A}$ of size $\big(1 + e^{-C/2} + o(1)\big)p\binom{n}{n/2}$; adding $k-2$ further layers to $\mathcal{A}$ gives a set of size $\big(k-1+e^{-C/2}+o(1)\big)p\binom{n}{n/2}$ containing no k-chains.

[1] That is, $\mathcal{P}(n, p)$ is a random variable such that $\mathbb{P}(A \in \mathcal{P}(n, p)) = p$ for each $A \in \mathcal{P}(n)$, and such events are independent for different values of A.

In order to effectively apply the hypergraph container method (see Section 2.2), one requires a so-called 'balanced supersaturation theorem', and the proof of such a result (see Theorem 2.1.2, below) is the main innovation of this work. An 'unbalanced' supersaturation theorem (giving a lower bound on the number of k-chains, but not controlling the distribution of these chains) was proved by Kleitman [54] in the case $k = 2$, and by Das, Gan and Sudakov [26] in general. More precisely, the authors of [26] used the permutation method pioneered by Katona and LYMB[2] in order to show that a family with t more elements than the extremal example above contains $\Omega\big(tn^{k-1}\big)$ k-chains. One of the key ideas from [26] will also play an important role in our proof, see Lemma 2.3.4 below.

In order to state our balanced supersaturation theorem, we will need a couple of simple definitions. For each $k \geqslant 2$ and $n \in \mathbb{N}$, let $\mathcal{G}_k = \mathcal{G}_k(n)$ denote the k-uniform hypergraph on vertex set $\mathcal{P}(n)$ whose edges encode k-chains, i.e., $\{F_1, \ldots, F_k\} \in E(\mathcal{G}_k)$ if and only if $F_1 \supsetneq \cdots \supsetneq F_k$ for some ordering of the elements. Given $\mathcal{F} \subset \mathcal{P}(n)$, we write $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ to denote that $\mathcal{H}$ is a k-uniform hypergraph with vertex set $\mathcal{F}$ whose edges are all members of $E(\mathcal{G}_k)$. For each $\ell \in [k]$, we write $\Delta_\ell(\mathcal{H})$ for the maximum degree of an ℓ-set in $\mathcal{H}$, that is

$$\Delta_\ell(\mathcal{H}) = \max\big\{d_{\mathcal{H}}(L) \,:\, L \subset V(\mathcal{H}),\, |L| = \ell\big\},$$

where $d_{\mathcal{H}}(L) = \big|\big\{\mathcal{A} \in E(\mathcal{H}) \,:\, L \subset \mathcal{A}\big\}\big|$. We also write $\mathcal{I}(\mathcal{H})$ for the collection of independent sets of $\mathcal{H}$, and $\alpha(\mathcal{H})$ for the size of the largest member of $\mathcal{I}(\mathcal{H})$.

We can now state the key new tool that we will use to prove Theorem 2.1.1. It says that a family with slightly more than $\alpha(\mathcal{G}_k) = \big(k-1+o(1)\big)\binom{n}{n/2}$ elements not only contains many k-chains, but that these chains can be chosen to be fairly 'evenly distributed' over $\mathcal{P}(n)$.

Theorem 2.1.2. *For every $k \geqslant 2$ and $\alpha > 0$, there exists $\delta = \delta(\alpha, k) > 0$ such that the following holds. Let $n \in \mathbb{N}$ and $\mathcal{F} \subset \mathcal{P}(n)$ satisfy $|\mathcal{F}| \geqslant (k - 1 + \alpha)\binom{n}{n/2}$, and suppose that $\delta^{-1} \leqslant m \leqslant \binom{|F|}{|G|}$ for every $F, G \in \mathcal{F}$ with $F \supsetneq G$. Then there exists $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ satisfying*

(a) $e(\mathcal{H}) \geqslant \delta^k m^{k-1}\binom{n}{n/2}$,

(b) $\Delta_\ell(\mathcal{H}) \leqslant (\delta m)^{k-\ell}$ *for every* $1 \leqslant \ell \leqslant k$.

We remark that the bounds in Theorem 2.1.2 are all close to best possible. To see this set $m = n/3$ and consider the $k - 1$ middle layers of the hypercube, together

[2] The acronym LYMB refers to Lubell [65], Yamamoto [93], Mešalkin [68] and Bollobás [14]. It often causes spelling confusion due to the silent B.

with $\alpha\binom{n}{n/2}$ elements from the next layer up. Then $\mathcal{G}_k[\mathcal{F}]$ has $O\big(n^{k-1}\binom{n}{n/2}\big)$ edges and $\Delta_\ell\big(\mathcal{G}_k[\mathcal{F}]\big) = \Omega(n^{k-\ell})$ for every $1 \leqslant \ell \leqslant k$. The technical assumption $m \leqslant \binom{|F|}{|G|}$ for every $F, G \in \mathcal{F}$ with $F \supsetneq G$ will be useful because it will allow us to deduce sufficiently strong bounds both when $|\mathcal{F}|$ is close to $\alpha(\mathcal{G}_k)$, and when it is much larger, see Section 2.2.

The rest of this chapter is organised as follows. In Section 2.2 we apply the hypergraph container method together with Theorem 2.1.2 to obtain a collection of containers tuned to our needs (see Corollary 2.2.3). In Section 2.3 we prove Theorem 2.1.2, and in Section 2.4 we perform the necessary technical computations in order to deduce Theorem 2.1.1.

2.2 HYPERGRAPH CONTAINERS

In this section, we will apply the powerful method of hypergraph containers described in Section 1.6. We will use the Hypergraph Container Lemma together with Theorem 2.1.2 to deduce that there exists a relatively small family of "containers", each not too large, which cover the family $\mathcal{I}(\mathcal{H})$ of independent sets of a k-uniform hypergraph $\mathcal{H} \subset \mathcal{G}_k$.

In more detail, we first apply the Hypergraph Container Lemma to the hypergraph $\mathcal{G}_k$, to obtain a large family $\mathcal{C}_1$ of containers, each of size at most $(1-\delta)2^n$. We then apply the lemma again, for each $\mathcal{F} \in \mathcal{C}_1$ with $|\mathcal{F}| \geqslant (k-1+\alpha)\binom{n}{n/2}$ (for some small $\alpha > 0$), to the hypergraph $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ given by Theorem 2.1.2. We repeat this process until all containers have size at most $(k-1+\alpha)\binom{n}{n/2}$. The conditions (a) and (b) in Theorem 2.1.2 allow us to check that (1.2) holds for a suitable value of τ, and hence to count the containers in our final collection. See [69] for a similar application of the container lemma in the context of C_{2k}-free graphs.

In order to further motivate the statement of Theorem 2.1.2 (and the technical condition $m \leqslant \binom{|F|}{|G|}$ for every $F, G \in \mathcal{F}$ with $F \supsetneq G$), we will next deduce from it the following two lemmas, which we will use to check the condition (1.2) from the Hypergraph Container Lemma. The first shows that we can take $\tau = 1/n$ when $\mathcal{F}$ is slightly larger than $\alpha(\mathcal{G}_k)$.

Lemma 2.2.1. *For every $k \geqslant 2$ and $\alpha > 0$, there exists $c = c(\alpha, k) > 0$ such that the following holds. Let $n \in \mathbb{N}$ be sufficiently large and $\mathcal{F} \subset \mathcal{P}(n)$ satisfy $(k-1+\alpha)\binom{n}{n/2} \leqslant |\mathcal{F}| \leqslant 3k\binom{n}{n/2}$. Then there exists a nonempty $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ satisfying*

$$\Delta_\ell(\mathcal{H}) \leqslant \frac{c}{n^{\ell-1}} \cdot \frac{e(\mathcal{H})}{|\mathcal{F}|}$$

for every $1 \leqslant \ell \leqslant k$.

Proof. First, observe that (by adjusting α slightly) we may assume that $|F| \geqslant n/3$ for every $F \in \mathcal{F}$, since the number of sets smaller than this is much smaller than $\binom{n}{n/2}$. Thus, applying Theorem 2.1.2 with $m = n/3$, it follows that there exists a hypergraph $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ and a constant $\delta = \delta(\alpha, k) > 0$ with $e(\mathcal{H}) \geqslant \delta^k n^{k-1}\binom{n}{n/2}$ and $\Delta_\ell(\mathcal{H}) \leqslant (3\delta n)^{k-\ell}$ for every $1 \leqslant \ell \leqslant k$. It follows that

$$\Delta_\ell(\mathcal{H}) \leqslant (3\delta n)^{k-\ell} = \frac{3^{k-\ell+1}k}{\delta^\ell n^{\ell-1}} \cdot \frac{\delta^k n^{k-1}\binom{n}{n/2}}{3k\binom{n}{n/2}} \leqslant \frac{c}{n^{\ell-1}} \cdot \frac{e(\mathcal{H})}{|\mathcal{F}|},$$

where $c = 3^k k \cdot (1/\delta)^k$, as required. □

The next lemma shows that if $|\mathcal{F}|$ is larger, then we can in fact take τ much smaller.

Lemma 2.2.2. *For every* $k \geqslant 2$*, there exists* $c = c(k) > 0$ *such that the following holds. Let* $n \in \mathbb{N}$ *be sufficiently large and* $\mathcal{F} \subset \mathcal{P}(n)$ *satisfy* $|\mathcal{F}| \geqslant 3k\binom{n}{n/2}$*. Then there exists a nonempty* $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ *satisfying*

$$\Delta_\ell(\mathcal{H}) \leqslant \frac{c}{n^{3\ell-3}} \cdot \frac{e(\mathcal{H})}{|\mathcal{F}|}$$

for every $1 \leqslant \ell \leqslant k$.

Proof. First, choose an arbitrary partition $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1 \cup \cdots \cup \mathcal{F}_t$ such that $|\mathcal{F}_i| = 3k\binom{n}{n/2}$ for every $i \in [t]$ and $|\mathcal{F}_0| < 3k\binom{n}{n/2}$. Fix $i \in [t]$, and observe that, by the pigeonhole principle, there are at least $k\binom{n}{n/2}$ elements of $\mathcal{F}_i$ whose sizes have the same remainder modulo 3. Let $\mathcal{F}_i'$ be a collection of $\big(k - o(1)\big)\binom{n}{n/2}$ such elements, all of size at least $n/3$, and note that $\binom{|F|}{|G|} \geqslant \binom{n/3}{3}$ for every $F, G \in \mathcal{F}_i'$ with $F \supsetneq G$. Thus, applying Theorem 2.1.2 with $m = \binom{n/3}{3}$, it follows that there exists a hypergraph $\mathcal{H}_i \subset \mathcal{G}_k[\mathcal{F}_i']$ and a constant $\delta = \delta(k) > 0$ such that $e(\mathcal{H}_i) = \delta^{k+1} n^{3k-3}\binom{n}{n/2}$ and

$$\Delta_\ell(\mathcal{H}_i) \leqslant (\delta n^3)^{k-\ell} = \frac{k}{\delta^\ell n^{3\ell-3}} \cdot \frac{\delta^k n^{3k-3}\binom{n}{n/2}}{k\binom{n}{n/2}} \leqslant \frac{c'}{n^{3\ell-3}} \cdot \frac{e(\mathcal{H}_i)}{|\mathcal{F}_i'|}$$

for some $c' = c'(k)$ and every $1 \leqslant \ell \leqslant k$. Let $\mathcal{H} = \mathcal{H}_1 \cup \cdots \cup \mathcal{H}_t$, and observe that

$$\Delta_\ell(\mathcal{H}) \leqslant \max_{1 \leqslant i \leqslant t} \big\{\Delta_\ell(\mathcal{H}_i)\big\} \leqslant \frac{c}{n^{3\ell-3}} \cdot \frac{e(\mathcal{H})}{|\mathcal{F}|}$$

as claimed, since $e(\mathcal{H}) = t \cdot e(\mathcal{H}_i)$ and $|\mathcal{F}| = O(t \cdot |\mathcal{F}_i'|)$ for every $i \in [t]$. □

Motivated by the above bounds, fix $\tau\colon \mathcal{P}(n) \to \mathbb{R}$ to be the function defined by

$$\tau(A) := \begin{cases} n^{-1} & \text{if } |A| \leqslant 3k\binom{n}{n/2} \\ n^{-3} & \text{otherwise.} \end{cases} \tag{2.2}$$

We can now specialise the Hypergraph Container Lemma to our application by combining it with Lemma 2.2.1 and Lemma 2.2.2. The following corollary will be used in Section 2.4 to count the containers of a given size produced by repeated applications of the Hypergraph Container Lemma, see Theorem 2.4.2.

Corollary 2.2.3. *For every* $2 \leqslant k \in \mathbb{N}$ *and* $\alpha > 0$, *there exists* $\delta = \delta(\alpha, k) > 0$ *such that the following holds. Let* $n \in \mathbb{N}$ *be sufficiently large and* $C \subset \mathcal{P}(n)$ *with* $|C| \geqslant (k - 1 + \alpha)\binom{n}{n/2}$. *Then there exists a collection* $\mathcal{C} \subset \mathcal{P}(C)$ *and a function* $f\colon \mathcal{P}(C) \to \mathcal{C}$ *such that*

(a) *For every* $I \in \mathcal{I}(\mathcal{G}_k[C])$, *there exists* T *with* $|T| \leqslant k \cdot \tau(C)|C|$ *and* $T \subset I \subset f(T)$.

(b) $|C'| \leqslant (1 - \delta)|C|$ *for every* $C' \in \mathcal{C}$.

Proof. Apply the Hypergraph Container Lemma to the hypergraph $\mathcal{H} \subset \mathcal{G}_k[C]$ given by Lemma 2.2.1 (if $|C| \leqslant 3k\binom{n}{n/2}$), or by Lemma 2.2.2 (otherwise), and observe that (for a suitable choice of the constant c) the inequality (1.2) holds with $\tau = \tau(C)$ for every $1 \leqslant \ell \leqslant k$. It follows immediately that there exist a family $\mathcal{C}$ of subsets of C, and a function $f\colon \mathcal{P}(C) \to \mathcal{C}$ such that (a) and (b) hold, as required. $\square$

2.3 BALANCED SUPERSATURATION

In this section, we will prove Theorem 2.1.2 by constructing $\mathcal{H}$ one edge at a time. More precisely, starting with $\mathcal{H} = \emptyset$, we will repeatedly apply the following lemma, adding new edges to $\mathcal{H}$ until the conditions of Theorem 2.1.2 are satisfied.

Lemma 2.3.1. *For every* $k \geqslant 2$ *and* $\alpha > 0$, *there exists* $\delta = \delta(\alpha, k) > 0$ *such that the following holds. Let* $n \in \mathbb{N}$ *and* $\mathcal{F} \subset \mathcal{P}(n)$ *satisfy* $|\mathcal{F}| \geqslant (k - 1 + \alpha)\binom{n}{n/2}$, *and suppose that* $\delta^{-1} \leqslant m \leqslant \binom{|F|}{|G|}$ *for every* $F, G \in \mathcal{F}$ *with* $F \supsetneq G$. *If* $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ *is a hypergraph satisfying*

(a) $e(\mathcal{H}) \leqslant \delta^k m^{k-1}\binom{n}{n/2}$,

(b) $\Delta_\ell(\mathcal{H}) \leqslant (\delta m)^{k-\ell}$ *for every* $\ell \in [k]$.

then there exists an edge $f \in \mathcal{G}_k[\mathcal{F}] \setminus \mathcal{H}$ *for which* $\Delta_\ell(\{f\} \cup \mathcal{H}) \leqslant (\delta m)^{k-\ell}$ *for every* $\ell \in [k]$.

The rest of this section will be dedicated to proving the above lemma, so from now on let us fix $\alpha > 0$ and $k \geqslant 2$, and choose $\delta > 0$ sufficiently small and $m \geqslant \delta^{-1}$. Moreover, let us fix $n \in \mathbb{N}$, a set $\mathcal{F} \subset \mathcal{P}(n)$ and a hypergraph $\mathcal{H} \subset \mathcal{G}_k[\mathcal{F}]$ satisfying the conditions of the lemma. The degree function of $\mathcal{H}$ will simply be denoted by d, for simplicity.

We say that a non-empty set $\mathcal{A} \subset \mathcal{P}(n)$ is *saturated* if $d(\mathcal{A}) = \lfloor (\delta m)^{k-|\mathcal{A}|} \rfloor$, that is, if no edge of $\mathcal{F}$ containing this set can be added to the hypergraph $\mathcal{H}$ without violating condition (b). A set $\mathcal{A} \subset \mathcal{P}(n)$ is *bad* if it contains a saturated set, and it is *good* otherwise. With this terminology, the conclusion of Lemma 2.3.1 is that $\mathcal{G}_k[\mathcal{F}]$ contains a good edge. Indeed, since a good edge $f \in \mathcal{G}_k[\mathcal{F}]$ is not saturated, then $d(f) < 1$, and so $f \notin \mathcal{H}$.

The following easy lemma will be a crucial tool in the proof of Lemma 2.3.1. It says that there are not too many ways to turn a good family bad.

Lemma 2.3.2. *For any good $\mathcal{A} \subset \mathcal{P}(n)$, there are at most $2^{|\mathcal{A}|} \cdot 2\delta k m$ sets $F \in \mathcal{P}(n)$ for which $\{F\} \cup \mathcal{A}$ is bad and $\{F\}$ is not saturated.*

Proof. The result follows from a simple double-counting argument, which we spell out below. Since $\mathcal{A}$ is good, any saturated subset of $\{F\} \cup \mathcal{A}$ must contain F. In other words, any F such that $\{F\} \cup \mathcal{A}$ is bad belongs to

$$S(\mathcal{B}) = \{F \in \mathcal{P}(n) : d(\{F\} \cup \mathcal{B}) = \lfloor (\delta m)^{k-|\mathcal{B}|-1} \rfloor\}$$

for some $\mathcal{B} \subset \mathcal{A}$. Moreover, if $\{F\}$ is not saturated, then $\mathcal{B}$ cannot be empty. Therefore, it is enough to bound the size of $S(\mathcal{B})$ when $\mathcal{B}$ is non-empty. We do so by noting that

$$|S(\mathcal{B})| \lfloor (\delta m)^{k-|\mathcal{B}|-1} \rfloor = \sum_{F \in S(\mathcal{B})} d(\{F\} \cup \mathcal{B}) \leqslant k d(\mathcal{B}) \leqslant k (\delta m)^{k-|\mathcal{B}|},$$

where the first inequality is true because each edge of $\mathcal{H}$ containing $\mathcal{B}$ contributes at most k to the sum. Since $m \geqslant \delta^{-1}$, we obtain $|S(\mathcal{B})| \leqslant 2\delta k m$. The claimed bound now follows by summing over all choices of $\mathcal{B}$. $\square$

Similarly, writing $S = S(\emptyset)$ for the family of saturated sets, we have $|S| \lfloor (\delta m)^{k-1} \rfloor \leqslant k \cdot e(\mathcal{H})$. By condition (a) and the bound $m \geqslant \delta^{-1}$, it follows that $|S| \leqslant 2\delta k \binom{n}{n/2}$. Thus, by adjusting α slightly if necessary, we can remove the elements of S from $\mathcal{F}$. Therefore, from now on we will assume that $\mathcal{F}$ contains no saturated vertices.

We will next sketch the proof of Lemma 2.3.1. The key idea is that if we choose F_1 to be of *minimal cardinality* such that the "density" of k-chains below F_1 (see Definition 2.3.3) is bigger than α/k (see Lemma 2.3.6), then only few of those k-chains will be bad, and hence at least one of them will be good. In order to bound the density of bad k-chains below F_1, let us define a chain $F_1 \supsetneq \cdots \supsetneq F_\ell$ to be *critical* if $\{F_1, \dots, F_{\ell-1}\}$ is good but $\{F_1, \dots, F_\ell\}$ is not. We will use Lemma 2.3.2 to show that the density of critical ℓ-chains is small (see Lemma 2.3.7). We will then use the minimality of F_1 to deduce that the operation of extending critical ℓ-chains to bad k-chains only increases the density by a bounded factor.

In order to make the above sketch more precise, let us next formalise the notion of density that we will use. This definition is inspired by the work of Das, Gan and Sudakov [26], see Lemma 2.3.4 below. We remark that, despite its name, the ℓ-chain density of a set is not bounded above by 1, and in fact can be as large as $\Omega(n^{\ell-1})$.

Definition 2.3.3. The ℓ-chain density of a set $F_1 \in \mathcal{F}$, denoted by $c_\ell(F_1)$, is given by

$$c_\ell(F_1) := \sum_{\substack{F_2,\dots,F_\ell \in \mathcal{F} \\ F_1 \supsetneq F_2 \supsetneq \cdots \supsetneq F_\ell}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_{\ell-1}|}{|F_\ell|}^{-1}$$

In particular, $c_1(F) = 1$ for all $F \in \mathcal{F}$.

The following lemma is essentially due to Das, Gan and Sudakov [26]. Since it was not explicitly stated in their paper, we will give the proof for completeness.

Lemma 2.3.4 (Das, Gan and Sudakov). *For any fixed* $1 \leqslant i < j \leqslant k$, *we have*

$$\sum_{F \in \mathcal{F}} \frac{1}{\binom{n}{|F|}} \big(c_i(F) - c_j(F) \big) \leqslant \max_{s \in \mathbb{N}} \binom{s}{i} - \binom{s}{j}.$$

Proof. Following the permutation method, say a permutation π of $[n]$ *contains* a set F if $F = \{\pi(1), \dots, \pi(|F|)\}$. Moreover, say it contains a chain if it contains all sets of the chain. Note that the number of permutations containing a given chain $F_1 \supsetneq \cdots \supsetneq F_\ell$ is

$$(n - |F_1|)! \cdot |F_1 \setminus F_2|! \cdots |F_{\ell-1} \setminus F_\ell|! \cdot |F_\ell|! = n! \cdot \binom{n}{|F_1|}^{-1} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_{\ell-1}|}{|F_\ell|}^{-1},$$

and so, denoting by $X_\ell(\pi)$ the number of ℓ-chains contained in π, the expected value of X_ℓ with respect to the uniform probability measure on the set of permutations is

$$\mathbb{E}(X_\ell) = \sum_{\substack{F_1,\dots,F_\ell \in \mathcal{F} \\ F_1 \supsetneq \dots \supsetneq F_\ell}} \binom{n}{|F_1|}^{-1} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_{\ell-1}|}{|F_\ell|}^{-1} = \sum_{F_1 \in \mathcal{F}} c_\ell(F_1) \Big/ \binom{n}{|F_1|}.$$

On the other hand, since sets contained in a single permutation always form a chain, $X_\ell(\pi)$ equals $\binom{s}{\ell}$, where s is the number of elements of $\mathcal{F}$ contained in π. We deduce that

$$X_i(\pi) - X_j(\pi) \leqslant \max_{s\in\mathbb{N}} \binom{s}{i} - \binom{s}{j},$$

and the conclusion follows by taking the expected value of both sides. □

A very useful feature of Lemma 2.3.4 is that the upper bound it provides does not depend on n. We will next use this to show that ℓ-chain densities cannot decrease too quickly as a function of ℓ, and hence that it is enough to upper bound the k-chain density of a set whenever we want an upper bound for all of its lower densities.

Lemma 2.3.5. *For every* $F \in \mathcal{F}$ *and* $1 \leqslant \ell < k$, *we have* $c_\ell(F) \leqslant c_k(F) + 4^k$.

Proof. The result is trivial for $\ell = 1$, as $c_1(F) = 1$. For $\ell \geqslant 2$, we can use the identity

$$c_\ell(F) = \sum_{\substack{F_2\in\mathcal{F}\\ F\supsetneq F_2}} \binom{|F|}{|F_2|}^{-1} \sum_{\substack{F_3,\ldots,F_k\in\mathcal{F}\\ F_2\supsetneq\cdots\supsetneq F_\ell}} \binom{|F_2|}{|F_3|}^{-1} \cdots \binom{|F_{\ell-1}|}{|F_\ell|}^{-1} = \sum_{F\supsetneq F_2\in\mathcal{F}} c_{\ell-1}(F_2) \Big/ \binom{|F|}{|F_2|}$$

together with Lemma 2.3.4 (applied to the hypercube of subsets of F) to obtain

$$c_\ell(F) - c_k(F) = \sum_{F\supsetneq F_2\in\mathcal{F}} \frac{1}{\binom{|F|}{|F_2|}} (c_{\ell-1}(F_2) - c_{k-1}(F_2)) \leqslant \max_{s\in\mathbb{N}} \binom{s}{\ell-1} - \binom{s}{k-1}.$$

Since the function being maximised is negative for all $s \geqslant 2k-1$, the right side is at most $\binom{2k-1}{\ell-1} \leqslant 4^k$, which proves the result. □

Lemma 2.3.4 also allows us to deduce that at least one element of our family has large k-chain density, as we show in the following pigeonhole-like observation.

Lemma 2.3.6. *If* $0 \leqslant \alpha \leqslant 1$ *and* $|\mathcal{F}| \geqslant (k-1+\alpha)\binom{n}{n/2}$, *then* $\max_F c_k(F) \geqslant \alpha/k$.

Proof. By Lemma 2.3.4 with $i = 1$ and $j = k$, and since $c_1(F) = 1$, we have

$$\sum_{F\in\mathcal{F}} \frac{1}{\binom{n}{|F|}}(1 - c_k(F)) \leqslant \max_{s\in\mathbb{N}} \binom{s}{1} - \binom{s}{k} = k-1.$$

However, if the desired conclusion were not true, we would have

$$\sum_{F\in\mathcal{F}} \frac{1}{\binom{n}{|F|}} (1 - c_k(F)) > \sum_{F\in\mathcal{F}} \frac{1}{\binom{n}{n/2}} \left(1 - \frac{\alpha}{k}\right) \geqslant (k-1+\alpha) \cdot \frac{k-\alpha}{k} \geqslant k-1,$$

where, for the last step, note that equality holds when $\alpha \in \{0, 1\}$. □

Finally, we will need the following lemma, which bounds the density of critical ℓ-chains. It is a simple consequence of Lemma 2.3.2 and our assumption that $m \leqslant \binom{|F|}{|G|}$ for every $F, G \in \mathcal{F}$ with $F \supsetneq G$.

Lemma 2.3.7. *For every* $F_1 \in \mathcal{F}$ *and* $1 \leqslant \ell < k$,

$$\sum_{\substack{F_2,\ldots,F_{\ell+1}\in\mathcal{F} \\ F_1\supsetneq\cdots\supsetneq F_{\ell+1} \text{ critical}}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_\ell|}{|F_{\ell+1}|}^{-1} \leqslant 2^\ell \cdot 2\delta k \cdot c_\ell(F_1) \tag{2.3}$$

Proof. Recall that if $F_1 \supsetneq \cdots \supsetneq F_{\ell+1}$ is critical, then $\{F_1, \ldots, F_\ell\}$ is good but $\{F_1, \ldots, F_{\ell+1}\}$ is not. By Lemma 2.3.2, it follows that the left-hand side of (2.3) is at most

$$\sum_{\substack{F_2,\ldots,F_\ell\in\mathcal{F} \\ F_1\supsetneq\cdots\supsetneq F_\ell}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_{\ell-1}|}{|F_\ell|}^{-1} \cdot 2^\ell \cdot 2\delta k m \cdot \max_{F_\ell\supsetneq F_{\ell+1}\in\mathcal{F}} \binom{|F_\ell|}{|F_{\ell+1}|}^{-1}.$$

The result then follows from our upper bound on m and the definition of $c_\ell(F_1)$. $\square$

We are now ready to carry out the plan outlined above, and prove Lemma 2.3.1.

Proof of Lemma 2.3.1. We may assume, without loss of generality, that $0 < \alpha < 1$. Let F_1 be of minimal cardinality such that $c_k(F_1) \geqslant \alpha/k$ (note that at least one such F_1 exists, by Lemma 2.3.6). We claim that

$$\sum_{\substack{F_2,\ldots,F_k\in\mathcal{F} \\ F_1\supsetneq\cdots\supsetneq F_k \text{ bad}}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_{k-1}|}{|F_k|}^{-1} \leqslant \frac{c_k(F_1)}{2}, \tag{2.4}$$

which immediately implies that the total k-chain density of *good* chains is positive, and therefore that at least one good chain exists. In order to prove (2.4), notice that every bad k-chain $F_1 \supsetneq \cdots \supsetneq F_k$ is associated with a unique $1 \leqslant \ell < k$ such that $F_1 \supsetneq \cdots \supsetneq F_{\ell+1}$ is critical. As such, we can write the left side of (2.4) as

$$\sum_{\ell=1}^{k-1} \left(\sum_{\substack{F_2,\ldots,F_{\ell+1}\in\mathcal{F} \\ F_1\supsetneq\cdots\supsetneq F_{\ell+1} \text{ critical}}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_\ell|}{|F_{\ell+1}|}^{-1} \cdot c_{k-\ell}(F_{\ell+1}) \right).$$

We will proceed by bounding each term of the outer sum separately, so fix $1 \leqslant \ell < k$. By Lemma 2.3.5 and the minimality of F_1, we have $c_{k-\ell}(F_{\ell+1}) \leqslant c_k(F_{\ell+1}) + 4^k < \alpha/k + 4^k < 5^k$. Using this bound and Lemma 2.3.7, we obtain

$$\sum_{\substack{F_2,\ldots,F_{\ell+1}\in\mathcal{F} \\ F_1\supsetneq\cdots\supsetneq F_{\ell+1} \text{ critical}}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_\ell|}{|F_{\ell+1}|}^{-1} \cdot c_{k-\ell}(F_{\ell+1}) \leqslant 2^\ell \cdot 2\delta k \cdot c_\ell(F_1) \cdot 5^k. \tag{2.5}$$

Using Lemma 2.3.5 once again for the bound $c_\ell(F_1) \leqslant c_k(F_1) + 4^k$ and summing (2.5) over $1 \leqslant \ell < k$, we conclude that

$$\sum_{\substack{F_2,\dots,F_k \in \mathcal{F} \\ F_1 \supsetneq \cdots \supsetneq F_k \text{ bad}}} \binom{|F_1|}{|F_2|}^{-1} \cdots \binom{|F_{k-1}|}{|F_k|}^{-1} = \delta \cdot 2^{O(k)} \cdot (c_k(F_1) + 4^k) = \frac{\delta \cdot 2^{O(k)}}{\alpha} \cdot c_k(F_1),$$

since $c_k(F_1) \geqslant \alpha/k$. The right side can be made less than $c_k(F_1)/2$ by choosing δ to be small (only as a function of α and k), and so the proof is complete. □

2.4 PROOF OF THEOREM 2.1.1

In this section we will deduce Theorem 2.1.1 from the results of the previous two sections. More precisely, we will use Corollary 2.2.3 to prove a 'fingerprint theorem' (Theorem 2.4.2, below), which easily implies Theorem 2.1.1. A coloured vertex set is simply a set $\mathcal{A} \subset \mathcal{P}(n)$ together with a function $c \colon \mathcal{A} \to \mathbb{N}$. We will need the following definition.

Definition 2.4.1. A fingerprint of $\mathcal{G}_k$ is a family $\mathcal{S}$ of coloured vertex sets, together with:

(a) A fingerprint function $T \colon \mathcal{I}(\mathcal{G}_k) \to \mathcal{S}$ with $T(I) \subset I$ for every $I \in \mathcal{I}(\mathcal{G}_k)$.

(b) A container function $C \colon \mathcal{S} \to \mathcal{P}(V(\mathcal{G}_k))$ such that $I \subset C(T(I))$ for every $I \in \mathcal{I}(\mathcal{G}_k)$.

Each $S \in \mathcal{S}$ should be thought of as a sequence of sets given by repeated application of the Hypergraph Container Lemma. The container function is obtained by applying the sequence of functions f given by these repeated applications. We will prove the following theorem.

Theorem 2.4.2. *For every $k \geqslant 2$ and $\varepsilon > 0$, there exist a constant $K = K(\varepsilon, k) > 0$ and a fingerprint $(\mathcal{S}, T, C)$ of $\mathcal{G}_k$ such that the following hold:*

(a) *Every $S \in \mathcal{S}$ satisfies $|S| \leqslant \frac{K}{n}\binom{n}{n/2}$;*

(b) *The number of sets of size s in $\mathcal{S}$ is at most*

$$\left(\frac{K\binom{n}{n/2}}{s}\right)^s \exp\left(\frac{K}{n}\binom{n}{n/2}\right);$$

(c) $|C(T(I))| \leqslant (k - 1 + \varepsilon)\binom{n}{n/2}$ *for every $I \in \mathcal{I}(\mathcal{H})$.*

Before proving Theorem 2.4.2, let us see how it implies Theorem 2.1.1.

Proof of Theorem 2.1.1. Let $k \geqslant 2$ and $\varepsilon > 0$ be arbitrary, and let $K = K(\varepsilon, k) > 0$ and $(\mathcal{S}, T, C)$ be the constant and fingerprint given by Theorem 2.4.2. Let $n \in \mathbb{N}$ be sufficiently large, and note that $pn \geqslant K\varepsilon^{-1}$, since $pn \to \infty$. If $I \subset \mathcal{P}(n,p)$ is an independent set of $\mathcal{G}_k$ of size at least $(k-1+3\varepsilon)p\binom{n}{n/2}$, then it follows that $T(I) \subset \mathcal{P}(n,p)$ and

$$\big|C(T(I)) \cap \mathcal{P}(n,p)\big| \geqslant (k-1+3\varepsilon)p\binom{n}{n/2}.$$

Let X be the number of elements of $\mathcal{S}$ for which these two properties hold. Then

$$\mathbb{E}(X) \leqslant \sum_{A \in \mathcal{S}} \mathbb{P}\big(A \subset \mathcal{P}(n,p)\big) \cdot \mathbb{P}\left(\big|(C(A) \setminus A) \cap \mathcal{P}(n,p)\big| \geqslant (k-1+2\varepsilon)p\binom{n}{n/2}\right),$$

where we used that $|A| \leqslant \varepsilon p\binom{n}{n/2}$ by the lower bound on pn and Theorem 2.4.2 (a). Hence, by the properties of $(\mathcal{S}, T, C)$ guaranteed by Theorem 2.4.2, and Chernoff's inequality,

$$\begin{aligned}\mathbb{E}(X) &\leqslant \sum_{s=1}^{\frac{K}{n}\binom{n}{n/2}} \left(\frac{K\binom{n}{n/2}}{s}\right)^s \exp\left(\frac{K}{n}\binom{n}{n/2}\right) \cdot p^s \cdot \exp\left(-\varepsilon^2 p\binom{n}{n/2}\right) \\ &\leqslant \frac{K}{n}\binom{n}{n/2} \exp\left(\frac{K\log(pn)}{n}\binom{n}{n/2} + \frac{K}{n}\binom{n}{n/2} - \varepsilon^2 p\binom{n}{n/2}\right),\end{aligned}$$

since the summand is increasing in s on the interval $\big(0, (Kp/e)\binom{n}{n/2}\big)$, and $K/n \ll Kp/e$. Therefore, by Markov's inequality, and since $pn \gg \log(pn) \gg 1$, we have

$$\mathbb{P}\left(\alpha\big(\mathcal{P}(n,p)\big) \geqslant (k-1+3\varepsilon)p\binom{n}{n/2}\right) \leqslant \exp\left(-\frac{\varepsilon^2 p}{2}\binom{n}{n/2}\right) \to 0$$

as $n \to \infty$, as required. □

It only remains to prove Theorem 2.4.2. We will use a straightforward but technical lemma.

Lemma 2.4.3. *Let* $M > 0$*,* $s > 0$ *and* $0 < \delta < 1$*. For any finite sequence* $(a_1, \ldots, a_m)$ *of real numbers summing to* s *such that* $1 \leqslant a_j \leqslant (1-\delta)^j M$ *for each* $j \in [m]$*, we have*

$$s \log s \leqslant \sum_{j=1}^{m} a_j \log a_j + O(M).$$

Proof. Fix $m \in \mathbb{N}$ and note that, by compactness, we can assume that the sequence $(a_1, \ldots, a_m)$ achieves the minimum of $\sum_{j=1}^m x_j \log x_j$ subject to the given conditions. Let

$$J_1 = \{j \in [m] : a_j < (1-\delta)^j M\}$$

and $J_2 = [m] \setminus J_1$; define also $s_i = \sum_{j \in J_i} a_j$ for $i \in \{1,2\}$. The convexity of $x \log x$ implies that all of the elements of the subsequence $(a_j)_{j \in J_1}$ are equal and that $J_1 = [t]$ for some $t \in \{0, \dots, m\}$, so that $s_1 \leqslant t(1-\delta)^t M$. Note that $s = \sum_j a_j = O(M)$ and

$$s_2 \log M - \sum_{j \in J_2} a_j \log a_j = \sum_{j \in J_2} a_j \log \frac{M}{a_j} \leqslant \sum_{j=1}^{\infty} (1-\delta)^j M \log \frac{1}{(1-\delta)^j} = O(M).$$

We are done if $t = 0$, so assume $t \geqslant 1$. By convexity, $s \log s \leqslant s_1 \log s_1 + s_2 \log s_2 + s \log 2$. Hence, recalling that $a_1 = \ldots = a_t = s_1/t$, we have

$$\begin{aligned} s \log s &\leqslant s_1 \log \frac{s_1}{t} + s_1 \log t + s_2 \log s_2 + O(M) \\ &\leqslant \sum_{j \in J_1} a_j \log a_j + t(1-\delta)^t M \log t + \sum_{j \in J_2} a_j \log a_j + O(M) \\ &= \sum_{j=1}^{m} a_j \log a_j + O(M), \end{aligned}$$

as claimed. $\square$

We are now ready to prove the 'fingerprint theorem', and thus complete the proof of Theorem 2.1.1.

Proof of Theorem 2.4.2. Let $k \geqslant 2$ and $\varepsilon > 0$ be arbitrary, let $\delta = \delta(\varepsilon, k) > 0$ be given by Corollary 2.2.3, choose a large constant $K = K(\varepsilon, k, \delta)$, and let $n \in \mathbb{N}$ be sufficiently large. For a given $I \in \mathcal{I}(\mathcal{G}_k)$, we will apply Corollary 2.2.3 a certain number of times, which we will denote by $m = m(I)$, to construct two sequences of sets $C_1, \dots, C_{m+1}$ and $T_1, \dots, T_m$. The construction will inductively maintain the following properties:

1. $I \subset C_{i+1} \cup T_1 \cup \cdots \cup T_i$,
2. The sets $C_{i+1}, T_1, \dots, T_i$ are pairwise disjoint,
3. C_{i+1} only depends on C_i and T_i,
4. $|C_{i+1}| \leqslant (1-\delta)|C_i|$.

To do this, first set $C_1 := \mathcal{P}(n)$. As long as $|C_i| \geqslant (k-1+\varepsilon)\binom{n}{n/2}$, let $T_i \subset I \cap C_i$ and f_i be given by Corollary 2.2.3, and set $C_{i+1} := f_i(T_i) \setminus T_i \subset C_i \setminus T_i$. We stop when we can no longer apply Corollary 2.2.3, that is, when $|C_{m+1}| < (k-1+\varepsilon)\binom{n}{n/2}$.

We define our fingerprint $(\mathcal{S}, T, C)$ of $\mathcal{G}_k$ by setting

$$T(I) := (T_1, \dots, T_m) \qquad \text{and} \qquad C(T(I)) := C_{m+1} \cup T_1 \cup \cdots \cup T_m,$$

and letting $\mathcal{S} := \{T(I) : I \in \mathcal{I}(\mathcal{G}_k)\}$. Note that Property 3 implies C is well-defined, while Property 1 guarantees that it is a container function.

In order to check that the constructed fingerprint satisfies the conditions of the theorem, we first bound the sizes of the fingerprints and the number of iterations of the above procedure. To do so, let $2 \leqslant m_0 \leqslant m$ be minimal such that $|C_{m_0}| \leqslant 3k\binom{n}{n/2}$, and observe that, by Property 4 and the definition (2.2) of $\tau(A)$,

$$\tau(C_i)|C_i| \leqslant \begin{cases} n^{-3} \cdot 2^n & \text{if } i < m_0, \\ n^{-1} \cdot (1-\delta)^{i-m_0} \cdot 3k\binom{n}{n/2} & \text{otherwise.} \end{cases} \tag{2.6}$$

The geometric decay of $|C_i|$ moreover immediately implies that $m = O(\log n)$. We thus obtain

$$\sum_{i=1}^{m_0-1} \tau(C_i)|C_i| \leqslant \frac{m \cdot 2^n}{n^3} \ll \frac{1}{n^2}\binom{n}{n/2} \qquad \text{and} \qquad \sum_{i=m_0}^{m} \tau(C_i)|C_i| = \frac{O(1)}{n}\binom{n}{n/2}. \tag{2.7}$$

Since $|T(I)| = \sum_{i=1}^{m} |T_i| \leqslant \sum_{i=1}^{m} k\tau(C_i)|C_i|$, adding the two bounds immediately proves (a). Also, since n is sufficiently large,

$$|C(T(I))| = |C_{m+1}| + |T_1 \cup \cdots \cup T_m| \leqslant (k - 1 + 2\varepsilon)\binom{n}{n/2},$$

which proves (c), since $\varepsilon > 0$ was arbitrary.

It only remains to prove (b), which follows using Lemma 2.4.3. The first step is to partition the collection of s-sets in $\mathcal{S}$ into subfamilies $\mathcal{S}(\hat{m}_0, \mathbf{t})$, where for given $\hat{m}_0 \in \mathbb{N}$ and $\mathbf{t} = (t_1, \dots, t_{\hat{m}}) \in \mathbb{N}^{\hat{m}}$, we define $\mathcal{S}(\hat{m}_0, \mathbf{t})$ to be set of all $(T_1, \dots, T_{\hat{m}}) \in \mathcal{S}$ such that $\hat{m}_0$ is the smallest integer for which $|C_{\hat{m}_0}| \leqslant 3k\binom{n}{n/2}$ and moreover $|T_i| = t_i$ for each $i \in [\hat{m}]$.

In order to bound the number of elements of $\mathcal{S}(\hat{m}_0, \mathbf{t})$ of size s, set $s_1 = \sum_{i=\hat{m}_0}^{\hat{m}} t_i$, and observe that

$$\sum_{i=\hat{m}_0}^{\hat{m}} t_i \log \frac{1}{t_i} \leqslant s_1 \log \frac{1}{s_1} + \frac{O(1)}{n}\binom{n}{n/2}, \tag{2.8}$$

by Lemma 2.4.3 and the second bound in (2.6). Since each T_i is a subset of the corresponding C_i, we can use the trivial bound $|C_i| \leqslant 2^n$ and the definition of $\hat{m}_0$ to write

$$\begin{aligned}
\left|\mathcal{S}(\hat{m}_0, \mathbf{t})\right| &\leqslant \prod_{i=1}^{\hat{m}_0 - 1} \binom{2^n}{t_i} \prod_{i=\hat{m}_0}^{\hat{m}} \binom{3k\binom{n}{n/2}}{t_i} \\
&\leqslant \left(\prod_{i=1}^{\hat{m}_0-1} 2^{t_i n}\right) \left(\left[3ek \cdot \binom{n}{n/2}\right]^{s_1} \prod_{i=\hat{m}_0}^{\hat{m}} \left(\frac{1}{t_i}\right)^{t_i}\right) \\
&\leqslant \left(\frac{K\binom{n}{n/2}}{s_1}\right)^{s_1} \exp\left(\frac{K}{n}\binom{n}{n/2}\right)
\end{aligned}$$

where the final step follows from the first sum in (2.7) and from applying the exponential function to (2.8). Finally, note that the right-hand side is monotone in s_1 on the interval $\left(0, K\binom{n}{n/2}/e\right)$, and we can therefore replace s_1 by s. Summing over the (at most $n^{O(n)}$) choices of $\mathbf{t}$, $\hat{m}_0$ and $\hat{m}$, the claimed bound follows. □

THE SHARP THRESHOLD FOR MAXIMUM-SIZE SUM-FREE SUBSETS IN EVEN-ORDER ABELIAN GROUPS

The work in this chapter is joint with Neal Bushaw, Robert Morris and Paul Smith. It is adapted from an article [20] which will appear in *Combinatorics, Probability & Computing*.

3.1 INTRODUCTION

In this chapter we will determine the sharp threshold for the maximum sum-free subset problem in an arbitrary even-order abelian group. Our main theorem improves some recent results of Balogh, Morris and Samotij [13], who resolved the case $G = \mathbb{Z}_{2n}$, and obtained weaker bounds in the general setting.

We consider the following question: How large is a maximum-size sum-free set in a p-random subset of an abelian group? For the group $\mathbb{Z}_{2n}$, this problem was resolved (asymptotically) by Conlon and Gowers [24] and Schacht [82], who determined the following threshold:

$$\max\big\{|B| \,:\, B \subset A = (\mathbb{Z}_{2n})_p \text{ is sum-free}\big\} \;=\; \begin{cases} \big(1+o(1)\big)\cdot 2pn & \text{if} \quad p \ll 1/\sqrt{n} \\ \big(1/2+o(1)\big)\cdot 2pn & \text{if} \quad p \gg 1/\sqrt{n} \end{cases} \tag{3.1}$$

with high probability as $n \to \infty$. More precisely, one can show using the methods of [24, 82] (see [13, 78]), and also using those of [8, 81], that (with high probability) the maximum-size sum-free subsets of A contain only $o(pn)$ even numbers. Moreover, a corresponding result holds for any even-order abelian group. This fact will be a key tool in the proof below.

We will be interested in the following more precise question, which was first studied by Balogh, Morris and Samotij [13]. Given an even-order abelian group G, note that the maximum-size sum-free subsets of G are exactly the odd cosets of subgroups of index 2, and that a p-random subset $A \subset G$ has a sum-free subset of (expected) size

$$\max\big\{|A \cap \mathcal{O}| \,:\, \mathcal{O} \text{ is the odd coset of a subgroup of index } 2\big\} \;\geqslant\; \left(\frac{1}{2}+o(1)\right) p|G|. \tag{3.2}$$

For which functions $p = p(n)$ is it true that, with high probability, the size of the largest sum-free subset of A is equal to the left-hand side of (3.2)? In other words, for which densities does the *exact* extremal result in G transfer to the sparse random setting? It was shown in [13] that the threshold for this property is $\left(\frac{\log n}{n}\right)^{1/2}$ for every even-order[1] abelian group, and moreover that there is a sharp threshold at $\left(\frac{\log n}{3n}\right)^{1/2}$ in the group $\mathbb{Z}_{2n}$. In other words, writing $\mathrm{SF}(A)$ for the collection of maximum-size sum-free subsets of A, and $\mathcal{O}_{2n}$ for the set of odd numbers in $\mathbb{Z}_{2n}$, they proved that for every $\varepsilon > 0$,

$$\mathbb{P}\Big(\mathrm{SF}\left((\mathbb{Z}_{2n})_p\right) = \left\{(\mathbb{Z}_{2n})_p \cap \mathcal{O}_{2n}\right\}\Big) \to \begin{cases} 0 & \text{if} \quad p \leqslant (1-\varepsilon)\sqrt{\frac{\log n}{3n}} \\ 1 & \text{if} \quad p \geqslant (1+\varepsilon)\sqrt{\frac{\log n}{3n}} \end{cases}$$

as $n \to \infty$. For more on the general theory of the existence of (sharp) thresholds, we refer the reader to [15, 44, 51], and to [45] for an example involving monochromatic triangles.

Since Balogh, Morris and Samotij [13] were able to prove such a sharp threshold for the group $\mathbb{Z}_{2n}$, but only a weaker threshold result for other even-order abelian groups, it is natural to ask whether one can also obtain a more precise result in the general setting. In this chapter we answer this question in the affirmative, by determining the sharp threshold for every even-order abelian group. In order to state our main theorem, we shall need the following function, which determines the location of the sharp threshold.

Definition 3.1.1. Given an abelian group G with $|G| = 2n$, let $r(G)$ denote the number of elements $x \in G$ such that $x = -x$, and set

$$\alpha(G) := \frac{\log r(G)}{\log n} \qquad \text{and} \qquad \beta(G) := \frac{r(G)}{n}.$$

Now, given $\delta > 0$, define $\lambda^{(\delta)}(G)$ as follows:

$$\lambda^{(\delta)}(G) := \begin{cases} 1/3 & \text{if} \quad \alpha(G) \leqslant 5/6 \\ \alpha(G) - 1/2 & \text{if} \quad \alpha(G) > 5/6 \text{ and } \beta(G) < \delta \\ 2/\big(4 - \beta(G)\big) & \text{if} \quad \beta(G) \geqslant \delta. \end{cases}$$

We encourage the reader to think of δ as a function going to zero slowly, and n as a function going to infinity much faster. The following theorem is our main result.

Theorem 3.1.2. *For every $\varepsilon > 0$, and every sufficiently small $0 < \delta < \delta_0(\varepsilon)$, there exists $n_0(\varepsilon, \delta) \in \mathbb{N}$ such that the following holds for every $n \geqslant n_0(\varepsilon, \delta)$. Let G be an abelian group of*

[1]In fact Theorem 1.1 of [13] is more general: it determines the threshold for any abelian group whose order has a (fixed) prime factor q with $q \equiv 2 \pmod 3$. Here, as before, we set $|G| = qn$.

order $2n$*, and let* $p \in (0,1)$ *with* $p \geqslant (\log n)^2/n$*. If* A *is a* p*-random subset of* G*, then*

$$\mathbb{P}\Big(A \cap \mathcal{O} \in \mathrm{SF}(A) \text{ for some } \mathcal{O} \in \mathrm{SF}(G)\Big) = \begin{cases} o(1) & \text{if} \quad p \leqslant \big(1-\varepsilon\big)\sqrt{\lambda^{(\delta)}(G)\frac{\log n}{n}} \\ 1+o(1) & \text{if} \quad p \geqslant \big(1+\varepsilon\big)\sqrt{\lambda^{(\delta)}(G)\frac{\log n}{n}}. \end{cases}$$

Here, as usual, $o(1)$ denotes a function that tends to zero as $n \to \infty$. We shall refer to the two bounds as the 0- and 1-statements respectively.

The proof of Theorem 3.1.2 uses the method of [13], but we will require several substantial new ideas in order to overcome various obstacles which do not occur in the case $G = \mathbb{Z}_{2n}$. Many of these arise from the fact that $\mathrm{SF}(G)$ can be quite large (as big as $|G|$ in the case of the hypercube), which means that we must obtain much stronger bounds than in [13] if we wish to apply the union bound. For the 0-statement we shall do this using a recent concentration inequality of Warnke [92], which allows us to deduce for almost all $\mathcal{O} \in \mathrm{SF}(G)$ that, with very high probability, the set $A \cap \mathcal{O}$ is not a maximal sum-free set. For the 1-statement, however, such a straightforward strategy is not feasible, since the threshold for the event that $A \cap \mathcal{O}$ is maximal for *every* odd coset $\mathcal{O} \in \mathrm{SF}(G)$ is not given by $\lambda^{(\delta)}(G)$.

In order to avoid this problem, we need to show that $A \cap \mathcal{O}$ is a maximal sum-free set for each $\mathcal{O} \in \mathrm{SF}(G)$ such that $|A \cap \mathcal{O}|$ is maximal. Unfortunately, conditioning on the size of $A \cap \mathcal{O}$ introduces significant dependence between odd cosets, and our first attempts to prove the 1-statement failed as a consequence. We resolve this issue by fixing the *number* of elements of A (i.e., coupling with the hypergeometric distribution), which essentially eliminates the positive correlation between the quantities $|A \cap \mathcal{O}|$ for different cosets.

A third issue involves the analysis of the Cayley graphs $\mathcal{G}_S$ for each $S \subset \mathcal{E}$, where $\mathcal{E}$ is a subgroup of index 2, $V(\mathcal{G}_S) = \mathcal{O}$ (the corresponding odd coset) and $xy \in E(\mathcal{G}_S)$ if either $x + y \in S$ or $x - y \in S$. Although counting the edges in these graphs precisely is not entirely trivial, we are fortunate that we can absorb most of the resulting mess into an error term. However, we still need to do some rather careful (and delicate) counting of the number of sets S that contain a given number of edges of $\mathcal{H}_W$, the Cayley graph of the set $W = \{a + a : a \in \mathcal{O}\}$, since this controls the size of $e(\mathcal{G}_S)$, see Section 3.3.

The remainder of the chapter is organised as follows. In Section 3.2, we collect some probabilistic tools and simple group-theoretic facts that will be needed later. In Section 3.3 we analyse the Cayley graph $\mathcal{G}_S$ for each set $S \subset \mathcal{E}$, where $\mathcal{E}$ is a subgroup of index 2, and count the number of such sets S whose Cayley graph has fewer edges than expected. In Section 3.4 we deduce the 0-statement from Warnke's concentration inequality (see Section 3.2), together with some of the more straightforward bounds from Section 3.3. Finally, in Section 3.5 we prove the 1-statement of Theorem 3.1.2 using the method of [13], combined with the coupling argument and careful counting described above. We end the chapter with a short Appendix, which contains a somewhat technical calculation involving the hypergeometric distribution.

We will also recall the FKG inequality and the concentration inequalities of Warnke and Janson, and state some simple facts about abelian groups that will be useful in the proof.

3.2 PRELIMINARIES

In this section, we will recall the FKG inequality and the concentration inequalities of Warnke and Janson, and state some simple facts about abelian groups that will be useful later on.

Probabilistic tools

Recently, Warnke [92] showed a powerful concentration inequality which improves martingale concentration methods. The main advantage of his method is that it relaxes the Lipschitz condition by allowing us to specify an event Γ for which we know the Lipschitz constant is smaller than the worst-case bound. In many combinatorial applications (see the article of Warnke [92] for examples), this improvement is substantial.

Here, we state a simpler version of this inequality which will be our main tool for the 0-statement in Section 3.4.

Warnke's inequality. *Given* $N \in \mathbb{N}$*, let* $\Gamma \subset \{0,1\}^N$ *be an event and* $f\colon \{0,1\}^N \to \mathbb{R}$ *be a function. Let* $p > 0$ *and* $X = (X_1, \ldots, X_N)$*, where* $X_k \in \{0,1\}$ *and* $\mathbb{P}(X_k = 1) = p$ *for each* $k \in [N]$*, all independently, and set* $\mu = \mathbb{E}\left[f(X)\right]$*. Suppose that, for some* $c, d > 0$*,*

$$|f(x) - f(y)| \leqslant \begin{cases} c & \text{if } x \in \Gamma, \\ d & \text{otherwise} \end{cases}$$

whenever $x, y \in \{0,1\}^N$ *with* $|x - y| = 1$*, and let* $\gamma \in (0,1)$*.*

There exists an event $\mathcal{B} = \mathcal{B}(\Gamma, \gamma) \subset \{0,1\}^N$*, with* $\neg\mathcal{B} \subset \Gamma$*, such that*

$$\mathbb{P}\left(X \in \mathcal{B}\right) \leqslant \frac{N}{\gamma} \cdot \mathbb{P}\left(X \notin \Gamma\right),$$

and moreover, setting $C = c + \gamma\big(d - c\big)$*, we have*

$$\mathbb{P}\left(f(X) \leqslant \mu - t \textit{ and } \neg\mathcal{B}\right) \;\leqslant\; \exp\left(-\frac{t^2}{2C^2pN + Ct}\right)$$

for any $t \geqslant 0$*.*

We also recall two well-known probabilistic inequalities: Janson's inequality and the FKG inequality. We refer the reader to [5] for various more general statements and their proofs.

Janson's inequality. *Suppose that* $\{B_i\}_{i \in I}$ *is a family of subsets of a finite set* X *and let* $p \in [0,1]$*. Let*

$$\mu = \sum_{i \in I} p^{|B_i|}, \quad \textit{and} \quad \Delta = \sum_{i \sim j} p^{|B_i \cup B_j|},$$

where $i \sim j$ *denotes the fact that* $i \neq j$ *and* $B_i \cap B_j \neq \emptyset$*. Then,*

$$\mathbb{P}\left(B_i \not\subset X_p \textit{ for all } i \in I\right) \leqslant e^{-\mu + \Delta}.$$

Furthermore, if $2c\mu \leqslant \Delta$ *with* $c \leqslant 1/4$*, then*

$$\mathbb{P}\left(B_i \not\subset X_p \textit{ for all } i \in I\right) \leqslant e^{-c\mu^2/\Delta}.$$

The FKG inequality. *Suppose that* $\{B_i\}_{i \in I}$ *is a family of subsets of a finite set* X *and let* $p \in [0,1]$*. Then*

$$\mathbb{P}\left(B_i \not\subset X_p \textit{ for all } i \in I\right) \;\geqslant\; \prod_{i \in I} \mathbb{P}\left(B_i \not\subset X_p\right).$$

Another key probabilistic component, which will be of great importance in the proof of the 1-statement, is the asymptotic stability theorem for even-order groups proved by Balogh, Morris and Samotij, which already appeared as Theorem 1.5.8 in the introduction.

Group-theoretic facts

In order to avoid repetition, we shall assume throughout the chapter that G is a finite abelian group of order $2n$. Given a subset $X \subset G$, we write

- $R(X)$ for the collection of elements $x \in X$ for which $x = -x$, and $r(X) = |R(X)|$.
- $m(X)$ for number of two-element subsets of X that are of the form $\{x, -x\}$.

We will need a few basic facts about finite abelian groups. The first one is well-known.

Fact 3.2.1. *There exist integers* $1 \leqslant a_1 \leqslant \ldots \leqslant a_k$ *and an odd-order group* J *such that*

$$G \cong \mathbb{Z}_{2^{a_1}} \oplus \cdots \oplus \mathbb{Z}_{2^{a_k}} \oplus J.$$

The second fact we need is a characterisation of the index 2 subgroups of G.

Fact 3.2.2. *Let* $I \subset \{1, \ldots, k\}$. *Writing* $x \in G$ *as* $(x_1, \ldots, x_k, y)$ *via the isomorphism of Fact 3.2.1, the subgroup* $H_I = \left\{x \in G : \sum_{i \in I} x_i \equiv 0 \pmod 2\right\}$ *is isomorphic to*

$$\mathbb{Z}_{2^{a_1}} \oplus \cdots \oplus \mathbb{Z}_{2^{a_i - 1}} \oplus \cdots \oplus \mathbb{Z}_{2^{a_k}} \oplus J,$$

where $i = \min I$. *Moreover, every subgroup of* G *of index* 2 *is equal to* H_I *for some* $I \neq \emptyset$.

Proof. Without loss of generality, assume that $J = \{0\}$ (and thus omit the last coordinate of elements of G) and $I = \{1, \ldots, k\}$. Then the image of the (injective) homomorphism

$$\begin{aligned} f \colon H_I &\to \mathbb{Z}_{2^{a_1}} \oplus \cdots \oplus \mathbb{Z}_{2^{a_k}} \\ (x_1, \ldots, x_k) &\mapsto (x_1 + \cdots + x_k, x_2, \ldots, x_k) \end{aligned}$$

consists of the elements of G whose first coordinate is even. Observe that the addition above is well-defined because there is a natural projection from $\mathbb{Z}_{2^{a_i}}$ to $\mathbb{Z}_{2^{a_1}}$ for any $1 \leqslant i \leqslant k$.

Conversely, given a subgroup H of index 2, observe that $\mathbb{1}_{H^c}$ is a homomorphism onto $\mathbb{Z}_2$, which implies that $\mathbb{1}_{H^c}(x_1, \ldots, x_k) \equiv \sum_{i=1}^{k} x_i \mathbb{1}_{H^c}(e_i) \equiv \sum_{i : e_i \notin H} x_i \pmod 2$, and thus $H = \left\{x \in G : \sum_{i : e_i \notin H} x_i \equiv 0 \pmod 2\right\}$. $\square$

Note that Fact 3.2.2 implies that G has exactly $r(G) - 1$ index 2 subgroups. Finally, we make a simple but useful observation.

Fact 3.2.3. *For any subgroup* H *of* G *of index* 2, *either* $r(H) = r(G)$ *or* $r(H) = r(G \setminus H)$.

Proof. For any $x \in R(G \setminus H)$, $y \mapsto y + x$ is a bijection between $R(H)$ and $R(G \setminus H)$. $\square$

3.3 EDGE COUNTS IN CAYLEY GRAPHS

In order to bound the probability of the event "$A \cap \mathcal{O} \in \mathrm{SF}(A)$" for some fixed maximum-size sum-free set $\mathcal{O} \in \mathrm{SF}(G)$ and its corresponding set of evens $\mathcal{E} = G \setminus \mathcal{O}$, we will need to consider events of the form

$$\text{“}\big((A \cap \mathcal{O}) \cup S\big) \setminus T \text{ is sum-free"}$$

where $S \subset A \cap \mathcal{E}$, $T \subset A \cap \mathcal{O}$ and $|S| \geqslant |T|$. This event is contained in the event that $(A \cap \mathcal{O}) \setminus T$ is an independent set in the Cayley graph $\mathcal{G}_S$, defined below, and to bound its probability we will need to analyse carefully the number of edges in this Cayley graph for each such set S of evens. In particular, there may be an exceptional collection of sets S with too few edges for our purposes (that is, for our application of the union bound over all sets S), and we will need to bound the size of this collection.

Let us begin by stating precisely the main results we will prove in this section. We fix throughout an arbitrary $\varepsilon > 0$, a sufficiently small $\delta > 0$ and a sufficiently large $n \in \mathbb{N}$.[2] We also fix an abelian group G of order $2n$, an odd coset $\mathcal{O} \in \mathrm{SF}(G)$, and its corresponding set of evens $\mathcal{E} = G \setminus \mathcal{O}$, which is a subgroup of G of index 2. For each set $S \subset \mathcal{E}$, we define the Cayley graph $\mathcal{G}_S$ of S to have vertex set $\mathcal{O}$ and edge set

$$E(\mathcal{G}_S) = \left\{ \{y, z\} \in \binom{\mathcal{O}}{2} \, : \, y + z \in S \text{ or } y - z \in S \right\},$$

where (for simplicity) we do not permit $\mathcal{G}_S$ to have loops. Recall that we write $r(X)$ for the number of order 2 elements in $X \subset G$, and $m(X)$ for the number of pairs $\{x, -x\} \subset X$.

We will prove the following propositions.

Proposition 3.3.1. *Let* $k \in \mathbb{N}$*. For every* $0 \notin S \subset \mathcal{E}$ *with* $|S| = k$ *and* $m(S) = 0$*, we have*

$$\left(\frac{3k - r(S)}{2} \right) n - O\big(r(G) \cdot k^2\big) \leqslant e(\mathcal{G}_S) \leqslant \left(\frac{3k - r(S)}{2} \right) n.$$

Moreover, if $r(G) \leqslant \delta n$ *and* $4\delta \leqslant \alpha \leqslant 1$*, then there are at most* $\big(6/\delta^2\big)^k \big(n/k\big)^{k - (\alpha/2 - \delta)k}$ *sets* $0 \notin S \subset \mathcal{E}$ *with*

$$e(\mathcal{G}_S) \leqslant \left(\frac{3k - r(S)}{2} - \alpha k \right) n$$

such that $|S| = k$ *and* $m(S) = 0$.

When $r(G) \geqslant \delta n$ the edge counts are slightly different.

Proposition 3.3.2. *If* $r(G) \geqslant \delta n$*, then, for every* $k \in \mathbb{N}$ *and* $0 \leqslant s \leqslant k$*, there are at most* $\big(12/\delta\big)^k \big(n/k\big)^s$ *sets* $0 \notin S \subset \mathcal{E}$ *with*

$$e(\mathcal{G}_S) < \big(s + 1\big) \left(n - \frac{r(\mathcal{O})}{2} \right) \tag{3.3}$$

such that $|S| = k$ *and* $m(S) = 0$.

[2]We think of δ as a function of n which tends to zero sufficiently slowly as $n \to \infty$.

In order to prove Propositions 3.3.1 and 3.3.2, we will first count edges in $\mathcal{G}_x = \mathcal{G}_{\{x\}}$ for each $x \in \mathcal{E}$, and then study the intersections between these graphs. These will depend on the parameter $r(S)$, as the reader can see from the statement. However, they will also depend on the intersection of S with the set

$$W = \{a + a : a \in \mathcal{O}\},$$

and with its Cayley graph. We will use several times the fact that $|W| = n/r(\mathcal{E})$.

Edge counts in $\mathcal{G}_x$

We begin with the relatively simple task of counting the edges in the Cayley graph of a single vertex x. To be precise, we will prove the following lemma.

Lemma 3.3.3. *For every* $0 \neq x \in \mathcal{E}$,

$$e(\mathcal{G}_x) = n - \frac{r(\mathcal{O})}{2} - \frac{r(\mathcal{E})}{2}\mathbb{1}\big[x \in W\big] + \left(\frac{n - r(\mathcal{O})}{2}\right)\mathbb{1}\big[x \notin R(G)\big],$$

and $\Delta(\mathcal{G}_x) \leqslant 3$.

Proof. Let us denote by $\mathcal{G}_x^+$ the edges of the form $x = y + z$, and by $\mathcal{G}_x^-$ the edges of the form $x = y - z$, so $\mathcal{G}_x = \mathcal{G}_x^+ \cup \mathcal{G}_x^-$. Note first that the graph $\mathcal{G}_x^-$ has a very simple structure, since every vertex has degree either one or two. More precisely, if $x \notin R(G)$ then it is a union of cycles, and so $e(\mathcal{G}_x^-) = n$; if $x \in R(G)$ then it is a matching, and so $e(\mathcal{G}_x^-) = n/2$.

In order to count the edges of $\mathcal{G}_x^+ \setminus \mathcal{G}_x^-$, let us partition the vertex set $\mathcal{O}$ into (up to) four parts, as follows:

(a) Set $O_1 = \{a \in \mathcal{O} : a + a = x\}$. If $|O_1| \neq 0$, then $x \in W$, and moreover $|O_1| = r(\mathcal{E})$, since the property $a \in O_1$ is invariant under the addition of an order 2 element. Moreover O_1 contains no edges of $\mathcal{G}_x^+$, and $O_1 \cap R(\mathcal{O}) = \emptyset$, since $x \neq 0$.

(b) Set $O_2 = R(\mathcal{O})$, the collection of order 2 elements in $\mathcal{O}$. If $x \in R(G)$ then O_2 induces a matching in $\mathcal{G}_x^+$, since $a \in R(\mathcal{O})$ if and only if $b = x - a \in R(\mathcal{O})$.

(c) Set $O_3 = \{b \in \mathcal{O} \setminus O_2 : x - b \in R(\mathcal{O})\}$, and observe that if $x \in R(G)$ then $|O_3| = 0$ (as above), whereas if $x \notin R(G)$ then $|O_3| = |O_2|$, since if $a \in R(\mathcal{O})$ then $b = x - a \notin R(\mathcal{O})$. Moreover $\mathcal{G}_x^+$ contains one edge for each element of O_3.

(d) Set $O_4 = \mathcal{O} \setminus \big(O_1 \cup O_2 \cup O_3\big)$, and note that $\mathcal{G}_x^+$ induces a perfect matching on O_4.

Now, observe that an edge of $\mathcal{G}_x^+$ is also contained in $\mathcal{G}_x^-$ if and only if it has an endpoint in $R(G)$, since if $a + b = x$ then $b \in R(G)$ if and only if $a - b = x$. Therefore

$$e(\mathcal{G}_x) \; = \; \big(1 + \mathbb{1}\big[x \notin R(G)\big]\big)\frac{n}{2} + \frac{|O_4|}{2}$$

and

$$|O_4| \; = \; n - \mathbb{1}\big[x \in W\big]r(\mathcal{E}) - \big(1 + \mathbb{1}\big[x \notin R(G)\big]\big)r(\mathcal{O}),$$

and so the lemma follows. □

Lemma 3.3.3 has the following simple consequence, which we shall use several times.

Observation 3.3.4. *For every* $0 \neq x \in \mathcal{E}$*, we have* $e(\mathcal{G}_x) \geqslant \max\{n - r(G), n/2\}$*. Moreover, if* $0 \notin S \subset \mathcal{E}$ *satisfies* $m(S) = 0$*, then* $e(\mathcal{G}_S) \geqslant \sum_{x \in S} e(\mathcal{G}_x)/2$.

Proof. If $x \neq 0$, Lemma 3.3.3 implies that

$$e(\mathcal{G}_x) \geqslant n - \frac{r(\mathcal{O})}{2} - \frac{r(\mathcal{E})}{2}\mathbb{1}[x \in W]$$

and, in particular, $e(\mathcal{G}_x) \geqslant n - r(G)$. In addition, either $r(\mathcal{O}) \leqslant r(\mathcal{E}) \leqslant n/2$ or $|W| = n/r(\mathcal{E}) = 1$, and so $e(\mathcal{G}_x) \geqslant n/2$. Further, when $m(S) = 0$, the set $\{x \in S : \{a, b\} \in E(G_x)\}$ contains at most two elements for any edge $\{a, b\}$. □

Before continuing to the proof of Proposition 3.3.1, let us note how to obtain (heuristically) the function $\lambda^{(\delta)}(G)$ from Lemma 3.3.3. We call an element $0 \neq x \in \mathcal{E}$ *safe* if $(A \cap \mathcal{O}) \cup \{x\}$ is sum-free, and let $S^{\mathcal{E}}(A)$ denote the collection of safe elements in $\mathcal{E}$. Note that an element $x \in \mathcal{E}$ is safe if[3] and only if $A \cap \mathcal{O}$ is an independent set in $\mathcal{G}_x$.

We need one more definition, whose slightly odd appearance will be motivated by the lemmas below.

Definition 3.3.5. A subgroup $\mathcal{E} \subset G$ is *nice* if either $r(G) \leqslant \delta n$ or $r(\mathcal{O}) = r(\mathcal{E})$.

The next lemma says that almost all index 2 subgroups are nice.

Lemma 3.3.6. G *has at most* $2/\delta$ *index* 2 *subgroups that are not nice.*

[3]This is only true if we ignore sums of the form $x = y + y$. However, such sums will never play a significant role in any of the calculations below.

Proof. Clearly if $r(G) \leqslant \delta n$ then all subgroups are nice, so let us assume $r(G) \geqslant \delta n$. By Fact 3.2.1, we can write $G \cong \mathbb{Z}_2^k \oplus H$, where $H = \mathbb{Z}_{2^{a_1}} \oplus \cdots \oplus \mathbb{Z}_{2^{a_\ell}} \oplus J$ with $2 \leqslant a_1 \leqslant \cdots \leqslant a_\ell$ and $|J|$ odd. Since $r(G) = 2^{k+\ell}$ and $|G| \geqslant 2^{k+2\ell}$, Fact 3.2.2 implies that there are at most $2^\ell \leqslant 2/\delta$ subgroups $\mathcal{E} \subset G$ of index 2 that are not isomorphic to $\mathbb{Z}_2^{k-1} \oplus H$. But if $\mathcal{E} \cong \mathbb{Z}_2^{k-1} \oplus H$, then $r(\mathcal{O}) = r(\mathcal{E})$, as required. □

We now prove the following bound on the expected number of safe elements, which we will use in the proof of the 0-statement of Theorem 3.1.2.

Lemma 3.3.7. *If* $\frac{\log n}{n} \ll p \leqslant (1-\varepsilon)\sqrt{\lambda^{(\delta)}(G)\frac{\log n}{n}}$ *and* $\mathcal{E}$ *is nice, then*

$$\mathbb{E}\left[|S^{\mathcal{E}}(A)|\right] \gg \frac{\log n}{p}.$$

Proof. Suppose first that $r(G) \leqslant \delta n$, and to simplify the notation let us write $\delta = o(1)$ (as noted above, we may assume that this holds as $n \to \infty$), and thus $r(G) = o(n)$. It follows from Lemma 3.3.3 that

$$e(\mathcal{G}_x) = \begin{cases} n + o(n) & \text{if} \quad x \in R(G) \\ 3n/2 + o(n) & \text{if} \quad x \notin R(G). \end{cases} \tag{3.4}$$

Now, by the FKG inequality, the expected number of safe elements $x \in \mathcal{E}$ is at least

$$\mathbb{E}[|S^{\mathcal{E}}(A)|] \geqslant \sum_{x \in \mathcal{E}} \left(1-p^2\right)^{e(\mathcal{G}_x)} \geqslant r(\mathcal{E})e^{-p^2(n+o(n))} + \big(n - r(\mathcal{E})\big)e^{-p^2(3n/2+o(n))} \gg \frac{\log n}{p}.$$

To see the final step, it suffices to check that the claimed inequality holds at the end-points of the claimed range of p, since xe^{-cx^2} is unimodal. At the lower end this is immediate; at the upper end, note that $e^{-p^2 n} \geqslant n^{(1-\varepsilon)^2\lambda^{(\delta)}(G)}$ and $r(\mathcal{E}) = n^{\alpha(G)+o(1)}$, and that

$$\max\left\{\alpha(G) - \lambda^{(\delta)}(G),\, 1 - \frac{3\lambda^{(\delta)}(G)}{2}\right\} = \frac{1}{2},$$

since $\lambda^{(\delta)}(G) = \max\left\{1/3, \alpha(G) - 1/2\right\}$.

When $r(G) \geqslant \delta n$, the (asymptotic) number of edges of $\mathcal{G}_x$ depends on both whether $x \in R(G)$ and whether $x \in W$. Indeed, the following table summarises the content of Lemma 3.3.3.

	$x \in R(G)$	$x \notin R(G)$
$x \in W$	$n - \frac{r(\mathcal{O})}{2} - \frac{r(\mathcal{E})}{2}$	$\frac{3n}{2} - r(\mathcal{O}) - \frac{r(\mathcal{E})}{2}$
$x \notin W$	$n - \frac{r(\mathcal{O})}{2}$	$\frac{3n}{2} - r(\mathcal{O})$

Table 3.1: Summary of Lemma 3.3.3

Fortunately, however, $|W| = n/r(\mathcal{E}) = O(1/\delta)$. We can therefore easily deduce a lower bound on $\mathbb{E}[|S^{\mathcal{E}}(A)|]$ for nice subgroups. Indeed, since $r(\mathcal{O}) = r(\mathcal{E}) = \beta(G)n/2$, and again using the unimodality of xe^{-cx^2}, it follows from Table 3.1 above that

$$\mathbb{E}\left[|S^{\mathcal{E}}(A)|\right] \geqslant \sum_{x \in R(\mathcal{E})} \left(1-p^2\right)^{e(\mathcal{G}_x)} = \Omega\Big(r(\mathcal{E})e^{-p^2(n-r(\mathcal{O})/2)}\Big) \gg \frac{\log n}{p}, \tag{3.5}$$

as required, where the last step follows since $1 - \big(1 - \beta(G)/4\big)\lambda^{(\delta)}(G) = 1/2$. □

Intersections between the graphs $\mathcal{G}_x$ and edge counts in $\mathcal{G}_S$

We now return to the proof of Proposition 3.3.1. In order to deduce the claimed bounds on $e(\mathcal{G}_S)$, we will need to control the size of the intersections between different graphs $\mathcal{G}_x$. Recall that we have fixed an odd coset $\mathcal{O} \in \mathrm{SF}(G)$, and that $W = \{a + a : a \in \mathcal{O}\}$. The following observation is key.

Observation 3.3.8. *Let $x, y \in \mathcal{E}$ with $x \notin \{y, -y\}$. If $E(\mathcal{G}_x) \cap E(\mathcal{G}_y) \neq \emptyset$, then $x + y \in W$.*

Proof. Suppose the edge $\{a, b\}$ lies in both $\mathcal{G}_x$ and $\mathcal{G}_y$. Then, without loss of generality, we have $a + b = x$ and $a - b = y$, and so $x + y = a + a$, as claimed. □

Moreover, we can bound the size of each intersection.

Observation 3.3.9. $\big|E(\mathcal{G}_x) \cap E(\mathcal{G}_y)\big| \leqslant 2 \cdot r(\mathcal{E})$ *for every $x, y \in \mathcal{E}$ with $x \notin \{y, -y\}$.*

Proof. Consider $\{a, b\}, \{c, d\} \in E(\mathcal{G}_x) \cap E(\mathcal{G}_y)$. Since $x \notin \{y, -y\}$, we may assume that $\{a + b, a - b\} = \{x, y\} = \{c + d, c - d\}$. It follows that $a + a = x + y = c + c$, and thus $c - a \in R(\mathcal{E})$. Moreover $d \in \{x - c, y - c\}$, and therefore, given $\{a, b\}$, there are at most $2 \cdot r(\mathcal{E})$ choices for $\{c, d\}$, as claimed. □

Let us denote by $\mathcal{H}_W$ the graph on vertex set $\mathcal{E}$ with edge set $\{xy : x + y \in W\}$, and note that we have $\Delta\big(\mathcal{H}_W\big) \leqslant d$, where $d := |W| = n/r(\mathcal{E})$. By Observations 3.3.8 and 3.3.9, we have

$$\sum_{x,y \in S,\, x \neq y} \big|E(\mathcal{G}_x) \cap E(\mathcal{G}_y)\big| \leqslant 2 \cdot r(\mathcal{E}) \cdot e(\mathcal{H}_W[S]) \tag{3.6}$$

for every $S \subset \mathcal{E}$ with $m(S) = 0$. Since, by Lemma 3.3.3, we have good bounds on the sum of $e(\mathcal{G}_x)$ over $x \in S$, the following lemma is all we need to complete the proof of Proposition 3.3.1.

Lemma 3.3.10. *For every $\delta \leqslant a \leqslant 1/2$, there are at most $\big(6/\delta^2\big)^k \big(n/k\big)^{k-(1-\delta)ak}$ sets $S \subset \mathcal{E}$ with $|S| = k$ and*

$$e\big(\mathcal{H}_W[S]\big) \geqslant \frac{akn}{r(\mathcal{E})}. \tag{3.7}$$

Proof. We shall first bound the number of sequences $(v_1, \dots, v_k) \in \mathcal{E}^k$ such that the set $S = \{v_1, \dots, v_k\}$ satisfies $|S| = k$ and (3.7). Given such a sequence, let us say (for each $j \in [k]$) that the vertex v_j is of 'low degree' if it is connected (by edges of $\mathcal{H}_W$) to fewer than $\delta a d = \delta a n/r(\mathcal{E})$ vertices of the set $\{v_1, \dots, v_{j-1}\}$, and say it is of high degree otherwise.

Since $\Delta(\mathcal{H}_W) \leqslant d$, it follows from (3.7) that in each such sequence there must be at least $(1-\delta)ak$ high-degree vertices, since the low-degree vertices contribute fewer than δakd edges. Moreover, since there are at most $(j-1)d < kd$ edges of $\mathcal{H}_W$ leaving the set $\{v_1, \dots, v_{j-1}\}$, there are at most $k/\delta a$ choices for a high-degree vertex, given the collection of vertices which have already been chosen.

Now, given a set $J \subset [k]$ of size at least $(1-\delta)ak$, corresponding to the positions of vertices which are required to have high degree, there are at most

$$\left(\frac{k}{\delta a}\right)^{|J|} n^{k-|J|}$$

possible sequences, and this value is maximised when $|J|$ is minimised. Therefore, considering all possible choices for J, it follows that there are at most

$$2^k \left(\frac{k}{\delta a}\right)^{(1-\delta)ak} n^{k-(1-\delta)ak}$$

sequences with the desired properties.

Finally, note that each set appears exactly $k!$ times as a sequence, and therefore the number of sets $S \subset \mathcal{E}$ with $|S| = k$ satisfying (3.7) is at most

$$\left(\frac{2e}{k}\right)^{k}\left(\frac{k}{\delta a}\right)^{(1-\delta)ak} n^{k-(1-\delta)ak} \leqslant \left(\frac{2e}{\delta^2}\right)^{k}\left(\frac{n}{k}\right)^{k-(1-\delta)ak},$$

since $a \geqslant \delta$, as required. □

We are now ready to prove the two propositions.

Proof of Proposition 3.3.1. Let $0 \notin S \subset \mathcal{E}$ with $|S| = k$ and $m(S) = 0$. By Lemma 3.3.3 and (3.6), and noting that $|W| = n/r(\mathcal{E})$, we have

$$\begin{aligned} e(\mathcal{G}_S) &\geqslant \sum_{x \in S}\left(n - \frac{r(\mathcal{O})}{2} - \frac{r(\mathcal{E})}{2}\mathbb{1}\big[x \in W\big] + \left(\frac{n - r(\mathcal{O})}{2}\right)\mathbb{1}\big[x \notin R(G)\big]\right) - 2 \cdot r(\mathcal{E})e(\mathcal{H}_W[S]) \\ &\geqslant k\big(n - r(G)\big) + \left(\frac{n - r(\mathcal{O})}{2}\right)\big(k - r(S)\big) - 2 \cdot r(\mathcal{E})e(\mathcal{H}_W[S]) \\ &\geqslant \left(\frac{3k - r(S)}{2}\right)n - O\big(r(G) \cdot k^2\big), \end{aligned}$$

as required, and the upper bound follows similarly. Moreover, the same calculation implies that if $e(\mathcal{G}_S) \leqslant \big(\frac{3k - r(S)}{2} - ak\big)n$ and $r(G) \leqslant \delta n$, then

$$e\big(\mathcal{H}_W[S]\big) \geqslant \frac{(a - 3\delta/4)kn}{2 \cdot r(\mathcal{E})},$$

and by Lemma 3.3.10 there are at most $\big(6/\delta^2\big)\big(n/k\big)^{k-(a/2-\delta)k}$ such sets $S \subset \mathcal{E}$ with $|S| = k$. □

Proof of Proposition 3.3.2. The proof is similar to that of Lemma 3.3.10, but for completeness we give the details. We will count sequences $(v_1, \dots, v_k) \in \mathcal{E}^k$ such that the set $S = \{v_1, \dots, v_k\}$ satisfies $|S| = k$ and (3.3). Let $S_j = \{v_1, \dots, v_j\}$, and observe that, since $m(S) = 0$, each $0 \neq x \notin W$ that sends no edges of $\mathcal{H}_W$ into S_j adds at least $n - r(\mathcal{O})/2$ edges to $\mathcal{G}_S$, by Lemma 3.3.3 (see Table 3.1) and Observation 3.3.8. There are therefore at most s such 'bad' vertices, since $e(\mathcal{G}_s) < (s+1)(n - r(\mathcal{O})/2)$.

Now, since $\Delta(\mathcal{H}_W) \leqslant |W| = n/r(\mathcal{E}) \leqslant 2/\delta$ and $|S_j| = j < k$, it follows that there are at most $2k/\delta$ vertices in $W \cup N_{\mathcal{H}_W}(S_j)$, and hence at most this many choices for each 'good' vertex. Note that there are at most 2^k choices for the indices j such that v_j is bad, and each set S is counted $k!$ times as a sequence. Thus, the number of sets $0 \notin S \subset \mathcal{E}$ with $|S| = k$ satisfying (3.3) is at most

$$\frac{2^k}{k!} \cdot \left(\frac{2k}{\delta}\right)^{k-s} n^s \leqslant \left(\frac{4e}{\delta}\right)^{k}\left(\frac{n}{k}\right)^{s},$$

as claimed. □

3.4 Proof of the 0-statement

In this section we will prove that if $A \subset G$ is a p-random set and

$$\frac{\log n}{n} \ll p \leqslant (1-\varepsilon)\sqrt{\lambda^{(\delta)}(G)\frac{\log n}{n}}, \tag{3.8}$$

then $A \cap \mathcal{O} \notin \mathrm{SF}(A)$ for every $\mathcal{O} \in \mathrm{SF}(G)$ with high probability as $n \to \infty$. The main step will be proving the following proposition.[4]

Proposition 3.4.1. *For every $\varepsilon > 0$, the following holds for every sufficiently large $n \in \mathbb{N}$. Let G be an abelian group of order $2n$, let $\mathcal{O} \in \mathrm{SF}(G)$ and suppose that $\mathcal{E} = G \setminus \mathcal{O}$ is nice and that $p \in (0,1)$ satisfies (3.8). If A is a p-random subset of G, then*

$$\mathbb{P}\left(A \cap \mathcal{O} \in \mathrm{SF}(A)\right) \leqslant \frac{1}{n^2}.$$

Recall also that at most $O(1/\delta)$ of the index 2 subgroups of G are not nice. We will use the following simple-sounding lemma to deal with these subgroups.

Lemma 3.4.2. *Let $\mathcal{M}$ denote the collection of odd cosets $\mathcal{O} \in \mathrm{SF}(G)$ such that $|A\cap\mathcal{O}|$ is maximal. Then with high probability there is an $\mathcal{O} \in \mathcal{M}$ such that $\mathcal{E} = G \setminus \mathcal{O}$ is nice.*

The proof of Lemma 3.4.2, although not difficult, is surprisingly technical, and so we shall postpone it to the appendix. Note that the 0-statement in Theorem 3.1.2 follows from Proposition 3.4.1 and Lemma 3.4.2 by taking a union bound over nice subgroups.

Recall that an element $x \in \mathcal{E}$ is called *safe* if $(A \cap \mathcal{O}) \cup \{x\}$ is sum-free, and that $S^{\mathcal{E}}(A)$ denotes the collection of safe elements in $\mathcal{E}$. We will bound the probability of the event $A \cap \mathcal{O} \in \mathrm{SF}(A)$ by the probability that there exists no safe element $x \in A \cap \mathcal{E}$. Since the random variable $S^{\mathcal{E}}(A)$ is independent of the set $A \cap \mathcal{E}$, it follows that

$$\mathbb{P}\left(\Big(A \cap \mathcal{O} \in \mathrm{SF}(A)\Big) \cap \left(|S^{\mathcal{E}}(A)| \geqslant \frac{3\log n}{p}\right)\right) \leqslant \left(1-p\right)^{(3\log n)/p} \leqslant \frac{1}{n^3}, \tag{3.9}$$

and so it is enough to consider the event that $|S^{\mathcal{E}}(A)| \leqslant (3\log n)/p$.

We will bound the probability of this event using Warnke's concentration inequality, which was stated in Section 3.2. The first step – showing that $|S^{\mathcal{E}}(A)|$ has large expected value – was already carried out in the previous section. Indeed, we have

$$\mathbb{E}\left[|S^{\mathcal{E}}(A)|\right] \gg \frac{\log n}{p} \tag{3.10}$$

[4] We remark that the bound $1/n^2$ could easily be replaced by $1/n^C$ for any $C > 0$.

whenever $p \in (0,1)$ satisfies (3.8), by Lemma 3.3.7. Our main task will be to prove the following lemma, which shows that $|S^{\mathcal{E}}(A)|$ is concentrated around its expected value.

Lemma 3.4.3. *If* $p \in (0,1)$ *satisfies* (3.8), *then*

$$\mathbb{P}\left(|S^{\mathcal{E}}(A)| \leqslant \frac{\mathbb{E}\left[|S^{\mathcal{E}}(A)|\right]}{2}\right) \leqslant \frac{1}{n^3}.$$

We will prove Lemma 3.4.3 by applying Warnke's inequality to the function $A \mapsto |S^{\mathcal{E}}(A)|$. In order to do so, we need to define an event $\Gamma \subset \mathcal{P}(\mathcal{O})$, and prove the 'typical Lipschitz condition'

$$\big||S^{\mathcal{E}}(A)| - |S^{\mathcal{E}}(B)|\big| \leqslant \begin{cases} c(\mathcal{E},p) := n^{-(1/4+\delta)} \cdot \mathbb{E}\left[S^{\mathcal{E}}(A)\right] & \text{if } A \in \Gamma, \\ n & \text{otherwise} \end{cases} \tag{3.11}$$

for every $A, B \subset \mathcal{O}$ with $|A \triangle B| = 1$ (note that $c(\mathcal{E},p) \gg 1$, by (3.10)). We define the event Γ so that (3.11) holds by definition:

$$\Gamma := \Big\{A \subset \mathcal{O} \,:\, \max\left\{\big||S^{\mathcal{E}}(A)| - |S^{\mathcal{E}}(B)|\big| : |A \triangle B| = 1\right\} \leqslant c(\mathcal{E},p)\Big\}. \tag{3.12}$$

We would like to show that $\mathbb{P}\left(A \notin \Gamma\right) \leqslant n^{-5}$, since this will imply the desired upper bound on the probability of the event $\mathcal{B}$ given by Warnke's inequality.

The main technical step in the proof of Lemma 3.4.3 is proving such a bound on the probability that $A \notin \Gamma$. To do so, note first that if $A \notin \Gamma$ then there exists $u \in \mathcal{O}$ such that $\big||S^{\mathcal{E}}(A)| - |S^{\mathcal{E}}(A \Delta \{u\})|\big| > c(\mathcal{E},p)$. Let $\Gamma^c(u)$ be the set of choices of A for which this property holds, so that $\Gamma^c = \bigcup_{u \in \mathcal{O}} \Gamma^c(u)$, and note that, by symmetry,[5]

$$\mathbb{P}\left(A \in \Gamma^c(u) \,\middle|\, u \in A\right) = \mathbb{P}\left(A \in \Gamma^c(u) \,\middle|\, u \notin A\right). \tag{3.13}$$

We will bound $\mathbb{P}\left(A \in \Gamma^c(u)\right)$ for each fixed $u \in \mathcal{O}$, and then sum over u.

Motivated by (3.13), let us fix $u \in \mathcal{O}$, assume that $u \notin A$, and write

$$Y_u^{\mathcal{E}}(A) = S^{\mathcal{E}}(A) \setminus S^{\mathcal{E}}(A \cup \{u\}).$$

Observe that $A \in \Gamma^c(u)$ if and only if $|Y_u^{\mathcal{E}}(A)| > c(\mathcal{E},p)$. We will prove the following lemma.

Lemma 3.4.4. *For every* k *satisfying* $25 < k \leqslant \sqrt{1/\delta}$,

$$\mathbb{P}\left(A \notin \Gamma\right) \leqslant c(\mathcal{E},p)^{-k} \sum_{u \in \mathcal{O}} \mathbb{E}\left[\left|Y_u^{\mathcal{E}}(A)\right|^k\right] \ll \frac{1}{n^5}$$

as $n \to \infty$.

[5]Indeed, if $B = A\Delta\{u\}$ then $A \in \Gamma^c(u) \Leftrightarrow B \in \Gamma^c(u) \Leftrightarrow \big||S^{\mathcal{E}}(A)| - |S^{\mathcal{E}}(B)|\big| > c(\mathcal{E},p)$.

Note that the first inequality follows from the comments above and Markov's inequality. The intuition behind the second inequality is based on our expectation that $|Y_u^{\mathcal{E}}(A)| = \Theta(p|S^{\mathcal{E}}(A)|)$, and that the events $\{z \in Y_u^{\mathcal{E}}(A) : z \in \mathcal{E}\}$ are more or less independent of one another. We expect $|Y_u^{\mathcal{E}}(A)|$ to take roughly this value since $Y_u^{\mathcal{E}}(A) \subset S^{\mathcal{E}}(A)$, and moreover for each $z \in Y_u^{\mathcal{E}}(A)$ there is a $v \in \mathcal{O}$ with $uv \in E(\mathcal{G}_z)$ such that $v \in A$.

In order to make this argument precise, the following notion will be crucial. Fix $u \in \mathcal{O}$, and say that a set $0 \neq Z \subset \mathcal{E}$ is *covered* by $Y \subset \mathcal{O}$ if for each $z \in Z$ there is a $y \in Y$ such that $uy \in E(\mathcal{G}_z)$. Say that Z is *cover-maximal* if $|Y| \geqslant |Z|$ for every set Y that covers Z, and for each $Z \subset \mathcal{E}$ choose a maximum-size cover-maximal subset $g(Z) \subset Z$. Note that since any singleton in Z is cover-maximal, $g(Z)$ is non-empty. The following lemma is key.

Lemma 3.4.5. *For each $Z \subset \mathcal{E}$, there are at most $12^{|Z|}$ sets $Z' \subset \mathcal{E}$ such that $g(Z') = Z$.*

Proof. Consider a set $Z' \subset \mathcal{E}$ such that $g(Z') = Z$. Then for any $z \in Z' \setminus Z$, there must exist some set $Y \subset \mathcal{O}$ of size $|Z|$ that covers $Z \cup \{z\}$ (and hence also covers Z), otherwise the set $Z \cup \{z\}$ contradicts the maximality in the definition of $g(Z')$.

We claim that there are at most $3^{|Z|}$ sets $Y \subset \mathcal{O}$ of size $|Z|$ covering Z. Indeed, since Z is cover-maximal, Y must contain exactly one element of $N_{\mathcal{G}_z}(u)$ for each $z \in Z$, and these neighbourhoods must be disjoint. Since $\Delta(\mathcal{G}_z) \leqslant 3$, it follows that we have at most $3^{|Y|} = 3^{|Z|}$ choices for Y. But each such set Y covers at most $3|Z|$ elements (since each is in $(Y \pm u) \cup (u - Y)$), and each $z \in Z' \setminus Z$ must be covered by some such Y, by the comments above. We therefore have at most $3^{|Z|} \cdot 2^{2|Z|} = 12^{|Z|}$ possible pre-images of Z, as claimed. □

We also need the following simple observation, which follows easily from the definition.

Observation 3.4.6. *If Z is cover-maximal and $\{a, -a\} \subset Z$, then $a = -a$.*

Proof. The element $u + a \in \mathcal{O}$ covers both a and $-a$, and so if $\{a, -a\} \subset Z$ and $a \neq -a$ then there exists a set Y with $|Y| \leqslant |Z| - 1$ which covers Z. □

We are ready to prove Lemma 3.4.4.

Proof of Lemma 3.4.4. Consider the family $\mathcal{M}_k$ of non-empty cover-maximal sets $Z \subset \mathcal{E}$ with $|Z| = k$, and note that if $Z' \subset Z$, then trivially

$$\mathbb{P}\left(Z' \subset Y_u^{\mathcal{E}}(A)\right) \geqslant \mathbb{P}\left(Z \subset Y_u^{\mathcal{E}}(A)\right).$$

Thus, by Lemma 3.4.5, we have

$$\sum_{Z \leqslant |k|} \mathbb{P}\left(Z \subset Y_u^{\mathcal{E}}(A)\right) \leqslant 12^k \sum_{\ell=1}^{k} \sum_{Z \in \mathcal{M}_\ell} \mathbb{P}\Big(\big(|A \cap N_{\mathcal{G}_Z}(u)| \geqslant |Z|\big) \cap \big(Z \subset S^{\mathcal{E}}(A)\big)\Big),$$

since each set Z contains a non-empty cover-maximal set $g(Z)$, and each such set is counted at most 12^k times. Now, since $|N_{\mathcal{G}_Z}(u)| \leqslant 3|Z|$, the right-hand side is at most

$$12^k \sum_{\ell=1}^{k} \sum_{Z \in \mathcal{M}_\ell} 2^{3\ell} p^\ell \cdot \mathbb{P}\left(Z \subset S^{\mathcal{E}}(A)\right), \tag{3.14}$$

by the FKG inequality, since $\left\{Z \subset S^{\mathcal{E}}(A)\right\}$ is decreasing in A, whereas $\big\{|A \cap N_{\mathcal{G}_Z}(u)| \geqslant |Z|\big\}$ is clearly increasing.

We will apply Janson's inequality to bound $\mathbb{P}\left(Z \subset S^{\mathcal{E}}(A)\right)$ for each $Z \in \mathcal{M}_\ell$. Note that $m(Z) = 0$, by Observation 3.4.6, and that $Z \subset S^{\mathcal{E}}(A)$ implies that $A \cap \mathcal{O}$ is an independent set in $\mathcal{G}_Z$, and suppose first that $r(G) \leqslant \delta n$. Then,

$$\mu := p^2 e(\mathcal{G}_Z) \geqslant \bigg(p^2 \sum_{z \in Z} e(\mathcal{G}_z)\bigg) - O\big(\delta \ell^2 p^2 n\big) \quad \text{and} \quad \Delta := p^3 \sum_{v \in \mathcal{O}} \binom{d_{\mathcal{G}_Z}(v)}{2} = O\big(\ell^2 p^3 n\big),$$

since $\big|E(\mathcal{G}_y) \cap E(\mathcal{G}_z)\big| \leqslant 2 \cdot r(\mathcal{E}) = O(\delta n)$ for every $y, z \in Z$ by Observation 3.3.9. Therefore, since $e(\mathcal{G}_z) \geqslant n/2$ for every $0 \neq z \in \mathcal{E}$ by Observation 3.3.4, and $\ell \leqslant k \leqslant 1/\sqrt{\delta}$, it follows by Janson's inequality and (3.8) that

$$\mathbb{P}\left(Z \subset S^{\mathcal{E}}(A)\right) \leqslant n^{O(\delta \ell^2)} \exp\bigg(-p^2 \sum_{z \in Z} e(\mathcal{G}_z)\bigg) = n^{O(\delta \ell^2)} \prod_{z \in Z} \left(1 - p^2\right)^{e(\mathcal{G}_z)},$$

since $1 - p^2 \geqslant e^{-p^2 - p^4}$ when p is sufficiently small, and $p^4 e(\mathcal{G}_z) = o(1)$. Thus

$$\begin{aligned}\sum_{Z \in \mathcal{M}_\ell} \mathbb{P}\left(Z \subset S^{\mathcal{E}}(A)\right) &\leqslant n^{O(\delta \ell^2)} \sum_{Z \in \mathcal{M}_\ell} \prod_{z \in Z} \left(1 - p^2\right)^{e(\mathcal{G}_z)} \\ &\leqslant n^{O(\delta \ell^2)} \bigg(\sum_{z \in \mathcal{E}} \left(1 - p^2\right)^{e(\mathcal{G}_z)}\bigg)^{\ell} \leqslant n^{O(\delta \ell^2)} \cdot \mathbb{E}\left[|S^{\mathcal{E}}(A)|\right]^{\ell}, \end{aligned} \tag{3.15}$$

where the final inequality follows by the FKG inequality.

On the other hand, if $r(G) \geqslant \delta n$ then, by Proposition 3.3.2, there are at most $n^{s+o(1)}$ sets $Z \subset \mathcal{E}$ with $|Z| = \ell$, $m(Z) = 0$ and

$$s\left(n - \frac{r(\mathcal{O})}{2}\right) \leqslant e(\mathcal{G}_Z) < (s+1)\left(n - \frac{r(\mathcal{O})}{2}\right).$$

Thus, applying Janson's inequality as before, we obtain[6]

$$\sum_{Z \in \mathcal{M}_\ell} \mathbb{P}\left(Z \subset S^{\mathcal{E}}(A)\right) \leqslant n^{o(1)} \sum_{s=1}^{\ell} \left(n\left(1-p^2\right)^{n-r(\mathcal{O})/2}\right)^s$$

$$= n^{o(1)}\left(1 + n\left(1-p^2\right)^{n-r(\mathcal{O})/2}\right)^\ell \leqslant n^{o(1)} \cdot \mathbb{E}\left[|S^{\mathcal{E}}(A)|\right]^\ell, \quad (3.16)$$

by (3.5). Combining (3.14), (3.15) and (3.16), it follows that

$$\mathbb{E}\left[\left|Y_u^{\mathcal{E}}(A)\right|^k\right] \leqslant n^{o(1)} \cdot \sum_{\ell=1}^{k} \sum_{Z \in \mathcal{M}_\ell} p^\ell \cdot \mathbb{P}\left(Z \subset S^{\mathcal{E}}(A)\right) \leqslant n^{O(\delta k^2)}\left(p \cdot \mathbb{E}\left[|S^{\mathcal{E}}(A)|\right]\right)^k,$$

and the lemma follows, since $c(\mathcal{E},p)^{-1} \cdot p \cdot \mathbb{E}\left[|S^{\mathcal{E}}(A)|\right] \ll n^{-1/5-\varepsilon}$. □

It is now straightforward to deduce Lemma 3.4.3, and hence Proposition 3.4.1.

Proof of Lemma 3.4.3. We apply Warnke's inequality to the function $A \mapsto |S^{\mathcal{E}}(A)|$ and the event Γ defined in (3.12), with

$$c = c(\mathcal{E},p) \gg 1, \qquad d = n, \qquad \gamma = \frac{c(\mathcal{E},p)}{n} \qquad \text{and} \qquad t = \frac{\mathbb{E}\left[|S^{\mathcal{E}}(A)|\right]}{2}.$$

We obtain an event $\mathcal{B}$ such that

$$\mathbb{P}\left(A \in \mathcal{B}\right) \leqslant \frac{n^2}{c(\mathcal{E},p)} \cdot \mathbb{P}\left(A \notin \Gamma\right) \ll \frac{1}{n^3},$$

where the last inequality follows by Lemma 3.4.4, such that

$$\mathbb{P}\left(|S^{\mathcal{E}}(A)| \leqslant \frac{\mathbb{E}\left[|S^{\mathcal{E}}(A)|\right]}{2}\right) \leqslant \mathbb{P}\left(A \in \mathcal{B}\right) + \exp\left(-\frac{t^2}{4c(\mathcal{E},p)^2pn + 2c(\mathcal{E},p)t}\right)$$

$$\leqslant \frac{o(1)}{n^3} + \exp\left(-n^\delta\right) \leqslant \frac{1}{n^3},$$

as required. □

Proof of Proposition 3.4.1. We split the event $A \cap \mathcal{O} \in \mathrm{SF}(A)$ into two parts, depending on whether or not $|S^{\mathcal{E}}(A)| \leqslant (3\log n)/p$. By Lemmas 3.3.7 and 3.4.3, the probability that $|S^{\mathcal{E}}(A)| \leqslant (3\log n)/p$ is at most $1/n^3$. On the other hand, by (3.9), the probability that $A \cap \mathcal{O} \in \mathrm{SF}(A)$ and $|S^{\mathcal{E}}(A)| \geqslant (3\log n)/p$ is at most $1/n^3$. Therefore

$$\mathbb{P}\left(A \cap \mathcal{O} \in \mathrm{SF}(A)\right) \leqslant \mathbb{P}\left(|S^{\mathcal{E}}(A)| \leqslant \frac{3\log n}{p}\right) + \frac{1}{n^3} \leqslant \frac{1}{n^2},$$

as required. □

[6]When $s = \ell$, we trivially bound the number of sets Z such that $e(\mathcal{G}_Z) \geqslant \ell\left(n - \frac{r(\mathcal{O})}{2}\right)$ by n^ℓ.

The 0-statement now follows immediately.

Proof of the 0*-statement in Theorem 3.1.2.* Recall that an abelian group G has at most $|G|$ index 2 subgroups. Thus, by Proposition 3.4.1 and the union bound, it follows that with high probability $A \cap \mathcal{O} \notin SF(A)$ whenever $\mathcal{E} = G \setminus \mathcal{O}$ is nice. However, by Lemma 3.4.2, with high probability there is an odd coset $\mathcal{O} \in SF(G)$ such that $|A \cap \mathcal{O}|$ is maximal and $\mathcal{E} = G \setminus \mathcal{O}$ is nice. Hence with high probability $A \cap \mathcal{O} \notin SF(A)$ for every $\mathcal{O} \in SF(G)$, as required. □

3.5 Proof of the 1-statement

In this section we will prove that if $A \subset G$ is a p-random set and

$$p \geqslant (1+\varepsilon)\sqrt{\lambda^{(\delta)}(G)\frac{\log n}{n}},$$

then every $B \in SF(A)$ is equal to $A \cap \mathcal{O}$ for some $\mathcal{O} \in SF(G)$, with high probability as $n \to \infty$. The proof has three steps: an application of Theorem 1.5.8 to obtain an asymptotic version, an argument for a given odd coset $\mathcal{O} \in SF(G)$, using the method of [13] (see Lemma 3.5.1), and a comparison with the hypergeometric distribution, which allows us to a partition the odd cosets depending on the size of $A \cap \mathcal{O}$ (see Lemma 3.5.2). Recall throughout that we have already fixed an arbitrary $\varepsilon > 0$, a sufficiently small $\delta > 0$ and a sufficiently large $n \in \mathbb{N}$.

We begin by proving the statement we will require for a given odd coset $\mathcal{O} \in SF(G)$. For each $k \in \mathbb{N}$, let $\mathcal{B}_k^{\mathcal{O}}(A)$ denote the event that there exist sets $S \subset A \cap \mathcal{E}$ and $T \subset A \cap \mathcal{O}$, with $|S| = k \geqslant |T|$, such that $\big((A \cap \mathcal{O}) \cup S\big) \setminus T$ is sum-free.

Lemma 3.5.1. *Let* G *be an abelian group of order* $2n$*, and let* $\mathcal{O} \in SF(G)$*. Suppose that*

$$p \geqslant (1+\varepsilon)\sqrt{\lambda^{(\delta)}(G)\frac{\log n}{n}},$$

and let $p_1 = (1-\delta)p$ *and* $p_2 = (1+\delta)p$*. Set* $A = A_1 \cup A_2$*, where* A_1 *is a* p_1*-random subset of* $\mathcal{O}$ *and* A_2 *is a* p_2*-random subset of* $\mathcal{E} = G \setminus \mathcal{O}$*. Then*

$$\mathbb{P}\left(\mathcal{B}_k^{\mathcal{O}}(A)\right) \leqslant \max\left\{n^{-\delta k}, e^{-\sqrt{n}}\right\}$$

for every $1 \leqslant k \leqslant \delta p n$*.*

Let us denote by $\mathbb{P}_{p^\pm} = \mathbb{P}_{p^\pm}^{\mathcal{O}}$ the probability distribution in Lemma 3.5.1, in which each element of $\mathcal{O}$ is chosen (independently) with probability $(1-\delta)p$ and each element of $\mathcal{E}$ is chosen with probability $(1+\delta)p$. Note that the event $\mathcal{B}_k^{\mathcal{O}}(A)$ is increasing in $A \cap \mathcal{E}$ and decreasing in $A \cap \mathcal{O}$, so $\mathbb{P}_p\left(\mathcal{B}_k^{\mathcal{O}}(A)\right) \leqslant \mathbb{P}_{p^\pm}\left(\mathcal{B}_k^{\mathcal{O}}(A)\right)$ for every $\delta \geqslant 0$.

Proof of Lemma 3.5.1. The proof of the lemma follows closely the method of Balogh, Morris and Samotij [13, Section 5], and so we shall skip some of the details. We will bound the expected number of *good triples* (S, T, U) with the following properties:

(i) $S \subset A \cap \mathcal{E}$ with $|S| = k$,

(ii) $T, U \subset A \cap \mathcal{O}$ are disjoint sets with $|U| \leqslant |T| \leqslant k$,

(iii) $(A \cap \mathcal{O}) \setminus T$ is an independent set in $\mathcal{G}_S$,

(iv) $T \subset N_{\mathcal{G}_S}(U)$.

It was shown in [13, Claim 2] that if $\mathcal{B}_k^{\mathcal{O}}(A)$ holds, then there exists such a triple. Indeed, this follows by first taking T minimal, and then taking a maximal matching M from T to $A \setminus T$ in $\mathcal{G}_S$. We set U equal to the set of vertices in $A \setminus T$ that are incident to M.

Let $Z(k, \ell, j, m, r)$ denote the number of such triples (S, T, U) with $|S| = k$, $|T| = \ell$, $|U| = j$, $m(S) = m$ and $r(S) = r$. We note that by definition $2m + r \leqslant k$, and define

$$Z_k := \sum_{\ell=0}^{k} \sum_{j=0}^{\ell} \sum_{m=0}^{k/2} \sum_{r=0}^{k-2m} Z(k, \ell, j, m, r).$$

By the discussion above,

$$\mathbb{P}\left(\mathcal{B}_k^{\mathcal{O}}(A)\right) \leqslant \mathbb{E}\left[Z_k\right] = \sum_{\ell=0}^{k} \sum_{j=0}^{\ell} \sum_{m=0}^{k/2} \sum_{r=0}^{k-2m} \mathbb{E}\left[Z(k, \ell, j, m, r)\right], \tag{3.17}$$

and therefore it will suffice to bound $\mathbb{E}[Z(k, \ell, j, m, r)]$ for each k, ℓ, j, m and r. Let $p^2 n = C \log n$, where $C \geqslant (1 + \varepsilon)\lambda^{(\delta)}(G)$. We will prove that

$$\mathbb{E}[Z(k, \ell, j, m, r)] \leqslant \begin{cases} n^{-\delta k} & \text{if} \quad k \leqslant \delta/p \\ e^{-\sqrt{n}} & \text{otherwise.} \end{cases} \tag{3.18}$$

Let us fix k, ℓ, j, m and r, and count the triples (S, T, U) that contribute to $Z(k, \ell, j, m, r)$. First, for each $S \subset \mathcal{E}$ and $\ell, j \in \mathbb{N}$, let $W(S, \ell, j)$ denote the number of disjoint pairs (T, U) such that $T, U \subset A \cap \mathcal{O}$ and $T \subset N_{\mathcal{G}_S}(U)$, with $|T| = \ell$ and $|U| = j$. It was proved in [13] that if $|S| = k$ and $0 \leqslant j \leqslant \ell \leqslant k \leqslant \delta p n$, then

$$\mathbb{E}\left[W(S, \ell, j)\right] \leqslant (3e^2 p^2 n)^k \ll \left(C \log n\right)^{2k} = n^{o(k)}$$

assuming that $C = n^{o(1)}$, as we may since the case $C \gg 1$ was already dealt with in [13].[7]

Let $\S(k, m, r)$ denote the collection of sets $S \subset \mathcal{E}$ with $|S| = k$, $m(S) = m$ and $r(S) = r$. If (S, T, U) is good, then no edge of the graph

$$\mathcal{G}_{S,T,U} := \mathcal{G}_S\big[\mathcal{O} \setminus (T \cup U)\big]$$

has both its endpoints in A. Since the vertex set of $\mathcal{G}_{S,T,U}$ is disjoint from $S \cup T \cup U$, it follows that the events $e(\mathcal{G}_{S,T,U}[A]) = 0$ and $S \cup T \cup U \subset A$ are independent. Therefore,

$$\begin{aligned}\mathbb{E}\big[Z(k,\ell,j,m,r)\big] &\leqslant \sum_{S \in \S(k,m,r)} \mathbb{P}(S \subset A) \cdot \mathbb{E}\big[W(S,\ell,j)\big] \cdot \max_{T,U}\Big\{\mathbb{P}\Big(e\big(\mathcal{G}_{S,T,U}[A]\big) = 0\Big)\Big\} \\ &\leqslant p^k \cdot n^{o(k)} \sum_{S \in \S(k,m,r)} \max_{T,U}\Big\{\mathbb{P}\Big(e\big(\mathcal{G}_{S,T,U}[A]\big) = 0\Big)\Big\}, \qquad (3.19)\end{aligned}$$

where the maximum is taken over all pairs (T, U) as in the definition of $W(S, \ell, j)$. We will bound the probability that A is an independent set in $\mathcal{G}_{S,T,U}$ using Janson's inequality. Indeed, let

$$\mu := p^2 e\big(\mathcal{G}_{S,T,U}\big) \qquad \text{and} \qquad \Delta := \sum_{v \in \mathcal{O}\setminus(T\cup U)} p^3 \binom{d(v)}{2},$$

where $d(v)$ denotes the degree of v in $\mathcal{G}_{S,T,U}$.

We break into two cases, depending on the number of elements of order 2 in G.

Case 1: $r(G) \leqslant \delta n$.

For each $S \in \S(k, m, r)$ let us choose a subset $\hat{S} \subset S$ with $|\hat{S}| = k - m$, $r(\hat{S}) = r$ and $m(\hat{S}) = 0$. Applying Proposition 3.3.1 to $\hat{S}$, it follows that

$$e(\mathcal{G}_{S,T,U}) \geqslant e(\mathcal{G}_{\hat{S}}) - O(k^2) \geqslant \left(\frac{3(k-m)-r}{2}\right) n - O\big(r(G)\cdot k^2\big), \qquad (3.20)$$

and that, for every $4\delta \leqslant \alpha \leqslant 1$, the number of sets $\hat{S} \in \S(k - m, 0, r)$ with

$$e(\mathcal{G}_{\hat{S}}) \leqslant \left(\frac{3(k-m)-r}{2} - \alpha k\right) n \qquad (3.21)$$

is at most $\big(6/\delta^2\big)^k \big(n/k\big)^{k-(\alpha/2-\delta)k}$. Moreover, for each such set $\hat{S}$ there are at most 2^k corresponding sets $S \in \S(k, m, r)$. There are three sub-cases to consider:

[7] Alternatively, we may simply carry this factor of C^{2k} through the proof, and perform an easy but tedious calculation later on.

(a) Suppose first that $k \leqslant \min\{\sqrt{\delta}/p,\ \delta n/r(G)\}$. Then, by (3.20),

$$\mu \geqslant \left(\frac{3(k-m)-r}{2} - O(\delta k)\right) p^2 n \qquad \text{and} \qquad \Delta = O(k^2p^3n) = O(\sqrt{\delta}kp^2n),$$

since $d(v) \leqslant 3k$ for every $v \in V(\mathcal{G}_{S,T,U})$. Thus, by Janson's inequality, it follows that

$$\mathbb{P}\left(e\big(\mathcal{G}_{S,T,U}[A]\big) = 0\right) \leqslant \exp\left(-\left(\frac{3(k-m)-r}{2} - O(\sqrt{\delta}k)\right)p^2n\right),$$

and hence, by (3.19),

$$\mathbb{E}\left[Z(k,\ell,j,m,r)\right] \leqslant p^k \cdot r(\mathcal{E})^r \cdot n^{k-m-r+o(k)} \cdot \exp\left(-\left(\frac{3(k-m)-r}{2} - O(\sqrt{\delta}k)\right)p^2n\right).$$

Since $p = n^{-1/2+o(1)}$, $r(G) = n^{\alpha(G)+o(1)}$ and $p^2n = C\log n$, it follows that

$$\begin{aligned}\frac{\log \mathbb{E}\left[Z(k,\ell,j,m,r)\right]}{\log n} &\leqslant \frac{k}{2} - m - \big(1-\alpha(G)\big)r - C\left(\frac{3(k-m)-r}{2} - O(\sqrt{\delta}k)\right) + o(k)\\ &\leqslant \left(\frac{1-3C}{2}\right)k - \left(\frac{2-3C}{2}\right)m + \left(\alpha(G) - \frac{2-C}{2}\right)r + O(C\sqrt{\delta}k) \leqslant -\frac{\varepsilon k}{4}.\end{aligned}$$

Indeed, the second term is decreasing in m for all $C \leqslant 2/3$,[8] and we have (considering the cases $r = 0$ and $r = k$ separately) $\frac{1-3C}{2} \leqslant -\varepsilon/2$ and $\frac{1-3C}{2} + \alpha(G) - \frac{2-C}{2} \leqslant -\varepsilon/3$, since (by assumption) we have $C \geqslant (1+\varepsilon)\max\{1/3, \alpha(G)-1/2\}$.

(b) Next, suppose that $k \geqslant \delta n/r(G)$ but $k \leqslant \sqrt{\delta}/p$. We partition the space according to the size of $e(\mathcal{G}_{\hat{S}})$: to be precise, we define $i = i(\hat{S})$ by the inequalities

$$e(\mathcal{G}_{\hat{S}}) \in \left(\frac{3(k-m)-r}{2} - \delta(2i \pm 1)(k-m)\right)n.$$

Since $(1-\delta)(k-m)n/2 \leqslant e(\mathcal{G}_{\hat{S}}) \leqslant (3(k-m)n - r)/2$ by Observation 3.3.4 and Proposition 3.3.1, we have $0 \leqslant 2\delta i(k-m) \leqslant (1+\delta)(k-m) - r/2$ for every set $\hat{S}$. Summing over i[9], applying Janson's inequality as in case (a), and using (3.21), we obtain

$$\mathbb{E}\left[Z(k,\ell,j,m,r)\right] \leqslant n^{O(\sqrt{\delta}k)} \sum_{i \geqslant 3} p^k \left(\frac{n}{k}\right)^{k-m-a_i/2} \exp\left(-\left(\frac{3(k-m)-r}{2} - a_i\right)p^2n\right),$$

where $a_i = 2\delta i(k-m)$. Substituting $p = n^{-1/2+o(1)}$ and $p^2n = C\log n$, and using the bound $k \geqslant n^{1-\alpha(G)+o(1)}$, it follows that

$$\frac{\log \mathbb{E}\left[Z(k,\ell,j,m,r)\right]}{\log n} \leqslant \max_a \left\{-\frac{k}{2} + \alpha(G)\left(k-m-\frac{a}{2}\right) - C\left(\frac{3(k-m)-r}{2} - a\right)\right\} + O(\sqrt{\delta}k).$$

[8] If $C \geqslant 2/3$ then simply note that the previous line is decreasing in C, since $3(k-m)-r \geqslant 2k-m \geqslant k$.
[9] The case $i = O(1)$ was already covered by the proof in part (a).

To bound the right-hand side, it suffices to check the extremal points. When $a = 0$, we note that $r \leqslant k - m$ and $\alpha(G) - C \leqslant 1/2 - \varepsilon/3$ to obtain a bound of

$$-k + 2(\alpha(G) - C)(k - m) + O\big(\sqrt{\delta}k\big) \leqslant -\frac{\varepsilon k}{4}.$$

At the other extreme, when $a = (1+\delta)(k-m) - r/2$, we obtain analogously that

$$-k + (\alpha(G) - C)(k - m) + \frac{\alpha(G)k}{2} + O\big(\sqrt{\delta}k\big) \leqslant -\frac{\varepsilon k}{4}.$$

(c) Finally, suppose that $k \geqslant \sqrt{\delta}/p$. Note first that $e(\mathcal{G}_{S,T,U}) \geqslant e(\mathcal{G}_{\hat{S}}) - O(k^2) = \Omega(kn)$. The inequality here is as in (3.20), whereas the equality is by Observation 3.3.4. We thus have

$$\frac{\mu}{\Delta} = O\left(\frac{n}{p \cdot e(\mathcal{G}_{S,T,U})}\right) = O\left(\frac{1}{\sqrt{\delta}}\right) \qquad \text{and} \qquad \frac{\mu^2}{\Delta} = \Omega\left(p \cdot \frac{e(\mathcal{G}_{S,T,U})^2}{k^2 n}\right) = \Omega(pn).$$

This follows because $\Delta = O(k^2p^3n)$, since $d(v) \leqslant 3k$ for every $v \in V(\mathcal{G}_{S,T,U})$, and $\Delta = \Omega\big(p^3 e(\mathcal{G}_{S,T,U})^2/n\big)$, by convexity. Janson's inequality then implies that

$$\mathbb{P}\Big(e\big(\mathcal{G}_{S,T,U}[A]\big) = 0\Big) \;=\; e^{-\Omega(pn\sqrt{\delta})},$$

from which it follows immediately that

$$\begin{aligned}\mathbb{E}\big[Z(k,\ell,j,m,r)\big] \;&\leqslant\; p^{k+\ell+j}\binom{n}{k}\binom{n}{\ell}\binom{n}{j}e^{-\Omega(pn\sqrt{\delta})}\\ &\leqslant\; p^{3k}\binom{n}{k}^3 e^{-\Omega(pn\sqrt{\delta})} \;\leqslant\; e^{-\Omega(pn\sqrt{\delta})} \;\leqslant\; e^{-2\sqrt{n}},\end{aligned}$$

since $k \leqslant \delta pn$. This completes the proof of (3.18) in the case $r(G) \leqslant \delta n$.

Case 2: $r(G) \geqslant \delta n$.

We now repeat the calculation above, replacing the bounds of Proposition 3.3.1 with those of Proposition 3.3.2. Suppose first that $k \leqslant \sqrt{\delta}/p$, and partition the space according to the maximum $s \in \{0, \dots, k\}$ such that

$$e(\mathcal{G}_{\hat{S}}) \;\geqslant\; s\left(n - \frac{r(\mathcal{O})}{2}\right).$$

By Proposition 3.3.2, there are at most $(12/\delta)^k(n/k)^s \;=\; O\big(n^{s+\sqrt{\delta}k}\big)$ such sets S with $|S| = k$. Applying Janson's inequality, we obtain[10]

$$\mathbb{E}\big[Z(k,\ell,j,m,r)\big] \;\leqslant\; n^{O(\sqrt{\delta}k)}\sum_{s=0}^{k} p^k \cdot n^s \cdot \exp\left(-p^2 s\left(n - \frac{r(\mathcal{O})}{2}\right)\right),$$

[10]When $s = k$, we trivially bound the number of sets Z such that $e(\mathcal{G}_{\hat{S}}) \geqslant k\left(n - \frac{r(\mathcal{O})}{2}\right)$ by n^k.

and hence

$$\frac{\log \mathbb{E}\left[Z(k,\ell,j,m,r)\right]}{\log n} \leqslant \max_s \left\{ s - Cs\left(\frac{4-\beta(G)}{4}\right)\right\} - \frac{k}{2} + O\big(\sqrt{\delta}k\big) \leqslant -\frac{\varepsilon k}{4}$$

since $C \geqslant (1+\varepsilon)\cdot 2/\big(4-\beta(G)\big)$. The case $k \geqslant \sqrt{\delta}/p$ is exactly the same as case (c), above.

Having bounded $\mathbb{E}\left[Z(k,\ell,j,m,r)\right]$ in all cases, the result now follows easily by summing over ℓ, j, m and r. Indeed, by (3.17), we have

$$\mathbb{P}\left(\mathcal{B}_k^{\mathcal{O}}(A)\right) \leqslant \sum_{\ell=0}^{k}\sum_{j=0}^{\ell}\sum_{m=0}^{k/2}\sum_{r=0}^{k-2m} \mathbb{E}\left[Z(k,\ell,j,m,r)\right] \leqslant \max\left\{n^{-\delta k}, e^{-\sqrt{n}}\right\}$$

as claimed. This completes the proof of the lemma. □

In order to deduce the 1-statement in Theorem 3.1.2 from Lemma 3.5.1, we cannot simply apply the union bound over odd cosets $\mathcal{O} \in \mathrm{SF}(G)$, since an even-order abelian group G can have as many as $|G|$ distinct maximum-size sum-free subsets. On the other hand, Lemma 3.5.1 (together with Theorem 1.5.8) does imply that the maximum-size sum-free subset of A contains (with high probability) only $O(1)$ even elements, and moreover that any *given* collection of $n^{o(1)}$ odd cosets are all likely to be 'locally' maximal.

Motivated by these observations, it is natural to attempt to partition the odd cosets into two classes, depending on whether or not $|A \cap \mathcal{O}|$ is within $O(1)$ of $\max_{\mathcal{O}'} |A \cap \mathcal{O}'|$. However, the random variables $\{|A \cap \mathcal{O}'| : \mathcal{O}' \in \mathrm{SF}(G)\}$ are highly correlated with one another, due to the large (size $n/2$) overlap between different odd cosets, and for this reason the maximum is not easy to control.[11]

We resolve this problem by coupling with the hypergeometric distribution, for which the positive correlation between the variables $|A \cap \mathcal{O}|$ is greatly diminished. (In fact, these variables are roughly pairwise independent of one another.) For each $0 \leqslant m \leqslant 2n$, let $\mathbb{P}_m$ denote the probability measure on subsets of G obtained by choosing each subset of size m with equal probability. Note that, since any pair of distinct subgroups $\mathcal{E}, \mathcal{E}' \subset G$ of index 2 intersect in a subgroup of index 4, the information that $|A \cap \mathcal{O}| \geqslant a$ (and therefore $|A \cap \mathcal{E}| \leqslant m - a$) has very little influence on the probability that $|A \cap \mathcal{O}'| \geqslant a$.

[11]The behaviour of the random variable $\max_{\mathcal{O}'} |A \cap \mathcal{O}'|$ is in fact somewhat mysterious, and we believe that it merits further investigation.

This crucial property of the hypergeometric distribution is captured by the following lemma. Given $k \in \mathbb{N}$ and an odd coset $\mathcal{O} \in \mathrm{SF}(G)$, define $M_k^{\mathcal{O}}(A)$ to be the event that $|A \cap \mathcal{O}| \geqslant k$, and let

$$X_k(A) := \sum_{\mathcal{O} \in \mathrm{SF}(G)} \mathbb{1}\big[M_k^{\mathcal{O}}(A)\big]$$

denote the number of odd cosets $\mathcal{O} \in \mathrm{SF}(G)$ for which $|A \cap \mathcal{O}| \geqslant k$.

Lemma 3.5.2. *Fix $\gamma > 0$ and $h \in \mathbb{N}$, and let $1 \ll m \leqslant 2n$. There exists $b = b(G, m) \in [m]$ such that the following holds. If A is chosen according to $\mathbb{P}_m$, then*

(a) $\mathbb{E}\big[X_b(A)\big] \leqslant n^{\gamma}$ *and*

(b) $X_{b+h}(A) \geqslant 1$ *with high probability.*

The proof of Lemma 3.5.2 involves some straightforward but technical approximations of binomial coefficients, and so we defer it to an Appendix.

Let us denote by $\mathcal{C}_k^{\mathcal{O}}(A)$ the event that $|A \cap \mathcal{O}'| < |A \cap \mathcal{O}| + k$ for every $\mathcal{O}' \in \mathrm{SF}(G)$. We are now ready to complete the proof of our main theorem.

Proof of the 1-statement in Theorem 3.1.2. Let $\varepsilon > 0$ be arbitrary, and let $0 < \delta < \delta_0(\varepsilon)$ be sufficiently small and $n \geqslant n_0(\varepsilon, \delta)$ be sufficiently large. Let G be an abelian group with $2n$ elements, let $C \geqslant (1+\varepsilon)\lambda^{(\delta)}(G)$, set

$$p = \sqrt{\frac{C \log n}{n}},$$

and let A be a p-random subset of G. We shall prove that, with high probability as $n \to \infty$, we have $A \cap \mathcal{O} \in \mathrm{SF}(A)$ for some $\mathcal{O} \in \mathrm{SF}(G)$.

Indeed, let $B \in \mathrm{SF}(A)$ be a maximum-size sum-free subset of A, and note that, by Chernoff's inequality, and since $A \cap \mathcal{O}$ is sum-free for every $\mathcal{O} \in \mathrm{SF}(G)$, we have

$$|B| \geqslant \left(\frac{1}{2} - \delta\right) p|G| \tag{3.22}$$

with high probability as $n \to \infty$. Therefore, applying Theorem 1.5.8, we deduce[12] that,

[12]Note that $p \geqslant C/\sqrt{n}$ since $n \geqslant n_0(\varepsilon, \delta)$ is sufficiently large.

with high probability, we have $|B \setminus \mathcal{O}| \leqslant \delta pn$ for some $\mathcal{O} \in \mathrm{SF}(G)$. Therefore,

$$\mathbb{P}_p\bigg(\bigcap_{\mathcal{O}\in\mathrm{SF}(G)}\{A\cap\mathcal{O}\notin\mathrm{SF}(A)\}\bigg) \leqslant \mathbb{P}_p\bigg(\bigcup_{\mathcal{O}\in\mathrm{SF}(G)}\bigcup_{k=1}^{\delta pn}\Big(\mathcal{B}_k^{\mathcal{O}}(A)\cap\mathcal{C}_k^{\mathcal{O}}(A)\Big)\bigg) + o(1)$$
$$\leqslant \sum_{m=(1-\delta^2)2pn}^{(1+\delta^2)2pn}\mathbb{P}_m\bigg(\bigcup_{\mathcal{O}\in\mathrm{SF}(G)}\bigcup_{k=1}^{\delta pn}\Big(\mathcal{B}_k^{\mathcal{O}}(A)\cap\mathcal{C}_k^{\mathcal{O}}(A)\Big)\bigg)\cdot\mathbb{P}_p\big(|A|=m\big) + o(1), \tag{3.23}$$

where we again used Chernoff's inequality. Let $b = b(G,m) \in [m]$ be given by Lemma 3.5.2 (with $h = 1/\delta^2$) so, with high probability, we have $|A \cap \mathcal{O}'| \geqslant b + 1/\delta^2$ for some $\mathcal{O}' \in \mathrm{SF}(G)$. Note that if such an $\mathcal{O}'$ exists, then $\mathcal{C}_k^{\mathcal{O}}(A)$ implies that either $|A\cap\mathcal{O}| \geqslant b$ or $k \geqslant 1/\delta^2$.

Let us first bound the probability when $k \geqslant 1/\delta^2$. Indeed, by Hoeffding's inequality (see, e.g., [22]), we have

$$\mathbb{P}_m\big(\mathcal{B}_k^{\mathcal{O}}(A)\big) = \sum_{i=m/2-\delta^2 m}^{m/2+\delta^2 m}\mathbb{P}_m\Big(\mathcal{B}_k^{\mathcal{O}}(A)\,\Big|\,|A\cap\mathcal{E}|=i\Big)\,\mathbb{P}_m\big(|A\cap\mathcal{E}|=i\big) + o\bigg(\frac{1}{n^3}\bigg). \tag{3.24}$$

Moreover the event $\mathcal{B}_k^{\mathcal{O}}(A)$ is increasing in $A\cap\mathcal{E}$ and decreasing in $A\cap\mathcal{O}$, and therefore (recalling from Lemma 3.5.1 the definition of $\mathbb{P}_{p^{\pm}}$), we have

$$\mathbb{P}_m\Big(\mathcal{B}_k^{\mathcal{O}}(A)\,\Big|\,|A\cap\mathcal{E}|=i\Big) \leqslant \mathbb{P}_{p^{\pm}}\Big(\mathcal{B}_k^{\mathcal{O}}(A)\,\Big|\,\big(|A\cap\mathcal{E}|\geqslant i\big)\cap\big(|A\cap\mathcal{O}|\leqslant m-i\big)\Big)$$
$$\leqslant 2\cdot\mathbb{P}_{p^{\pm}}\big(\mathcal{B}_k^{\mathcal{O}}(A)\big) \leqslant 2\cdot n^{-1/\delta} \ll \frac{1}{n^3} \tag{3.25}$$

for every $k \geqslant 1/\delta^2$, by Lemma 3.5.1. Indeed, the first inequality follows since $p^{\pm}$ chooses sets A uniformly given $|A\cap\mathcal{E}|$ and $|A\cap\mathcal{O}|$. To see the second inequality, simply note that $\mathbb{P}_{p^{\pm}}\big((|A\cap\mathcal{E}|\geqslant i)\cap(|A\cap\mathcal{O}|\leqslant m-i)\big) \geqslant 1/2$ for every $i \leqslant m/2+\delta^2 m \leqslant pn + 3\delta^2 pn$.

Next, let us bound the probability when $|A\cap\mathcal{O}| \geqslant b$. Similarly to above, we have

$$\mathbb{P}_m\Big(\mathcal{B}_k^{\mathcal{O}}(A)\cap\big(|A\cap\mathcal{O}|\geqslant b\big)\Big) = \sum_{i=0}^{m-b}\mathbb{P}_m\Big(\mathcal{B}_k^{\mathcal{O}}(A)\,\Big|\,|A\cap\mathcal{E}|=i\Big)\cdot\mathbb{P}_m\big(|A\cap\mathcal{E}|=i\big),$$

and moreover

$$\mathbb{P}_m\Big(\mathcal{B}_k^{\mathcal{O}}(A)\,\Big|\,|A\cap\mathcal{E}|=i\Big) \leqslant 2\cdot\mathbb{P}_{p^{\pm}}\big(\mathcal{B}_k^{\mathcal{O}}(A)\big) \leqslant 2\cdot n^{-\delta},$$

for every $k \geqslant 1$, by (3.25) and Lemma 3.5.1, and

$$\mathbb{E}_m\big[X_b(A)\big] = \sum_{\mathcal{O}\in\mathrm{SF}(G)}\sum_{i=0}^{m-b}\mathbb{P}_m\big(|A\cap\mathcal{E}|=i\big) \leqslant n^{\delta/2},$$

by Lemma 3.5.2(a). Therefore

$$\sum_{\mathcal{O}\in \mathrm{SF}(G)} \mathbb{P}_m\left(\mathcal{B}_k^{\mathcal{O}}(A) \cap \big(|A\cap\mathcal{O}| \geqslant b\big)\right) \leqslant 2\cdot n^{-\delta/2} \tag{3.26}$$

for every $k \geqslant 1$. Combining (3.24), (3.25) and (3.26), it follows that

$$\mathbb{P}_m\left(\bigcup_{\mathcal{O}\in \mathrm{SF}(G)} \bigcup_{k=1}^{\delta p n}\left(\mathcal{B}_k^{\mathcal{O}}(A)\cap \mathcal{C}_k^{\mathcal{O}}(A)\right)\right) \leqslant 2\cdot \sum_{k=1}^{1/\delta^2} n^{-\delta/2} + \sum_{k=1/\delta^2}^{\delta p n} \sum_{\mathcal{O}\in \mathrm{SF}(G)} \frac{1}{n^3} \leqslant n^{-\delta/3}$$

for every $m \in (1\pm\delta^2)2pn$, and every sufficiently large n. Hence, by (3.23), we have

$$\mathbb{P}_p\left(\bigcap_{\mathcal{O}\in \mathrm{SF}(G)} \big\{A\cap\mathcal{O} \notin \mathrm{SF}(A)\big\}\right) = o(1),$$

as required. □

3.A Appendix: Lemmas on the hypergeometric distribution

In this Appendix we will prove Lemmas 3.4.2 and 3.5.2. We begin with the latter.

Proof of Lemma 3.5.2

We are required to prove that there exists $b = b(G,m) \in [m]$ with the following properties: at most $n^{o(1)}$ odd cosets are expected to contain at least b elements of A, but with high probability some odd coset contains at least $b+\omega$ elements of A, where $\omega \to \infty$ as $n \to \infty$. For the proof, it will be convenient to shift the notation by $m/2$ as follows: For each $k \in \mathbb{N}$ and each $\mathcal{O} \in \mathrm{SF}(G)$, let us denote by $M_k^{\mathcal{O}}(A)$ the event that $|A\cap\mathcal{O}| \geqslant m/2+k$, and by

$$X_k(A) = \sum_{\mathcal{O}\in \mathrm{SF}(G)} \mathbb{1}\big[M_k^{\mathcal{O}}(A)\big]$$

the number of odd cosets $\mathcal{O} \in \mathrm{SF}(G)$ for which $|A\cap\mathcal{O}| \geqslant m/2+k$.

The main step in the proof of Lemma 3.5.2 is the following bound on the correlation between the events $M_k^{\mathcal{O}}(A)$. Here, and throughout this Appendix, we write $x \sim y$ to mean that $x/y \to 1$ under the given asymptotics.

Lemma 3.A.1. *Let $\mathcal{O},\mathcal{O}' \in \mathrm{SF}(G)$ be distinct odd cosets, and let $k,m \in \mathbb{N}$ be such that $1 \ll k \ll m \ll k^2$. Then*

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\cap M_k^{\mathcal{O}'}(A)\right) \sim \mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right)^2$$

as $n \to \infty$.

We begin by calculating $\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right)$ asymptotically, using the following simple bounds.

Lemma 3.A.2. *Let* $a, b, N \in \mathbb{N}$ *with* $b^{3/2} \ll a \ll N$. *Then*

$$\frac{\binom{N}{a+b}\binom{N}{a-b}}{\binom{2N}{2a}} \sim \frac{1}{\sqrt{\pi a}} \exp\left(-\frac{b^2}{a}\right)$$

as $a, N \to \infty$.

Proof. This is nothing more than an application of Stirling's formula

$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n,$$

and the partial Taylor series

$$\left|\log(1+x) - x + \frac{x^2}{2}\right| \leqslant O\big(|x|^3\big),$$

which is valid for all sufficiently small $|x|$. □

Let us denote by $\hat{M}_x^{\mathcal{O}}(A)$ the event that $|A \cap \mathcal{O}| = m/2 + x$, so $M_k^{\mathcal{O}}(A) = \bigcup_{x \geqslant k} \hat{M}_x^{\mathcal{O}}(A)$.

Lemma 3.A.3. *For every* $\mathcal{O} \in \mathrm{SF}(G)$,

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right) \sim \sqrt{\frac{2}{\pi m}} \sum_{x \geqslant k} \exp\left(-\frac{2x^2}{m}\right).$$

Proof. Observe that

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right) = \sum_{x \geqslant k} \mathbb{P}_m\left(\hat{M}_x^{\mathcal{O}}(A)\right) = \sum_{x \geqslant k} \frac{\binom{n}{m/2+x}\binom{n}{m/2-x}}{\binom{2n}{m}}.$$

The result now follows by applying Lemma 3.A.2 with $N = n$, $a = m/2$ and $b = x$. □

The following bounds now follow easily.

Lemma 3.A.4. *For every* $\mathcal{O} \in \mathrm{SF}(G)$,

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right) = \Theta\left(\frac{\sqrt{m}}{k} \exp\left(-\frac{2k^2}{m}\right)\right).$$

Proof. By Lemma 3.A.3, we have

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right) = \Theta\left(\frac{1}{\sqrt{m}}\exp\left(-\frac{2k^2}{m}\right)\sum_{x\geqslant 0}\exp\left(-\frac{4kx}{m}-\frac{2x^2}{m}\right)\right).$$

Now, the asymptotics $k \ll m \ll k^2$ imply that

$$\sum_{x\geqslant 0}\exp\left(-\frac{4kx}{m}-\frac{2x^2}{m}\right) = \Theta\left(\frac{m}{k}\right),$$

and the lemma follows immediately. □

When bounding the probability of $M_k^{\mathcal{O}}(A)\cap M_k^{\mathcal{O}'}(A)$, the following notation will be useful. Set

$$\Lambda := \left\{(x,y,z)\in\mathbb{Z}^3 \,:\, x+y\geqslant k,\ x+z\geqslant k\right\},$$

and given $\mathcal{O},\mathcal{O}'\in \mathrm{SF}(G)$ and $x,y,z\in\mathbb{Z}$, denote by $\hat{M}_{x,y,z}^{\mathcal{O},\mathcal{O}'}(A)$ the event that

$$|A\cap\mathcal{O}\cap\mathcal{O}'| = \frac{m}{4}+x,\quad |A\cap\mathcal{O}\cap\mathcal{E}'| = \frac{m}{4}+y,\quad\text{and}\quad |A\cap\mathcal{O}'\cap\mathcal{E}| = \frac{m}{4}+z,$$

where as usual $\mathcal{E} = G\setminus\mathcal{O}$ and $\mathcal{E}' = G\setminus\mathcal{O}'$.

Lemma 3.A.5. *Let $\mathcal{O},\mathcal{O}'\in \mathrm{SF}(G)$ be distinct odd cosets. Then*

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\cap M_k^{\mathcal{O}'}(A)\right) \sim \frac{4\sqrt{2}}{(\pi m)^{3/2}}\sum_{(x,y,z)\in\Lambda}\exp\left(-\frac{2}{m}\left((x+y)^2+(x+z)^2+(y+z)^2\right)\right).$$

Proof. Note first that

$$\mathbb{P}_m\left(\hat{M}_{x,y,z}^{\mathcal{O},\mathcal{O}'}(A)\right) = \frac{\binom{n}{m/2+x+y}\binom{n}{m/2-x-y}}{\binom{2n}{m}}\frac{\binom{n/2}{m/4+x}\binom{n/2}{m/4+y}}{\binom{n}{m/2+x+y}}\frac{\binom{n/2}{m/4+z}\binom{n/2}{m/4-x-y-z}}{\binom{n}{m/2-x-y}}.$$

By Lemma 3.A.2, this is asymptotically equal to

$$\sqrt{\frac{2}{\pi m}}\exp\left(-\frac{2(x+y)^2}{m}\right)\sqrt{\frac{4}{\pi m}}\exp\left(-\frac{(x-y)^2}{m}\right)\sqrt{\frac{4}{\pi m}}\exp\left(-\frac{(x+y+2z)^2}{m}\right),$$

and this expression is equal to

$$\frac{4\sqrt{2}}{(\pi m)^{3/2}}\exp\left(-\frac{2}{m}\left((x+y)^2+(x+z)^2+(y+z)^2\right)\right).$$

Thus,

$$\begin{aligned}\mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\cap M_k^{\mathcal{O}'}(A)\right) &= \sum_{(x,y,z)\in\Lambda}\mathbb{P}_m\left(\hat{M}_{x,y,z}^{\mathcal{O},\mathcal{O}'}(A)\right)\\ &\sim \frac{4\sqrt{2}}{(\pi m)^{3/2}}\sum_{(x,y,z)\in\Lambda}\exp\left(-\frac{2}{m}\left((x+y)^2+(x+z)^2+(y+z)^2\right)\right),\end{aligned}$$

as claimed. □

We are almost ready to prove Lemma 3.A.1; we need one more well-known fact.

Fact 3.A.6.
$$\sum_{x\in\mathbb{Z}} \exp\left(-\frac{2x^2}{m}\right) \sim \sqrt{\frac{\pi m}{2}}$$
as $m \to \infty$.

Proof of Lemma 3.A.1. Observe that (a, b, c) is equal to $(x+y, x+z, y+z)$ for some triple (x, y, z) if and only if $a + b + c$ is even and
$$(x, y, z) = \left(\frac{a+b-c}{2}, \frac{c+a-b}{2}, \frac{b+c-a}{2}\right).$$
Letting
$$\Lambda' := \left\{(a, b, c) \in \mathbb{Z}^3 \, : \, a \geqslant k, \; b \geqslant k, \; a+b+c \text{ even}\right\},$$
it follows that
$$\sum_{(x,y,z)\in\Lambda} \exp\left(-\frac{2}{m}\left((x+y)^2 + (x+z)^2 + (y+z)^2\right)\right) = \sum_{(a,b,c)\in\Lambda'} \exp\left(-\frac{2}{m}(a^2+b^2+c^2)\right).$$
We may split up the right-hand side into separate sums according to the parity of $a+b$, and hence of c. Doing this, we may rewrite the sum as
$$\sum_{\substack{a\geqslant k,\, b\geqslant k,\\ a+b \text{ even}}} \exp\left(-\frac{2(a^2+b^2)}{m}\right) \sum_{c \text{ even}} \exp\left(-\frac{2c^2}{m}\right) + \sum_{\substack{a\geqslant k,\, b\geqslant k,\\ a+b \text{ odd}}} \exp\left(-\frac{2(a^2+b^2)}{m}\right) \sum_{c \text{ odd}} \exp\left(-\frac{2c^2}{m}\right).$$
Since m is large, we have
$$\sum_{c \text{ odd}} \exp\left(-\frac{2c^2}{m}\right) \sim \sum_{c \text{ even}} \exp\left(-\frac{2c^2}{m}\right) \sim \frac{1}{2}\sum_{c} \exp\left(-\frac{2c^2}{m}\right) \sim \frac{1}{2}\sqrt{\frac{\pi m}{2}},$$
where we have used Fact 3.A.6 for the final estimate. We also have
$$\sum_{a\geqslant k,\, b\geqslant k} \exp\left(-\frac{2(a^2+b^2)}{m}\right) = \left(\sum_{a\geqslant k} \exp\left(-\frac{2a^2}{m}\right)\right)^2 \sim \frac{\pi m}{2} \cdot \mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right)^2$$
for an arbitrary odd coset $\mathcal{O} \in \mathrm{SF}(G)$, by Lemma 3.A.3. Putting all this together, we conclude that
$$\sum_{(x,y,z)\in\Lambda} \exp\left(-\frac{2}{m}\left((x+y)^2 + (x+z)^2 + (y+z)^2\right)\right) \sim \frac{(\pi m)^{3/2}}{4\sqrt{2}} \cdot \mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right)^2. \quad (3.27)$$

We may now use our estimate for $\mathbb{P}_m\left(M_k^{\mathcal{O}}(A) \cap M_k^{\mathcal{O}'}(A)\right)$ from Lemma 3.A.5. Together with (3.27), this implies that

$$\mathbb{P}_m\left(M_k^{\mathcal{O}}(A) \cap M_k^{\mathcal{O}'}(A)\right) \sim \mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right)^2,$$

as required. $\square$

Lemma 3.5.2 now follows by a straightforward application of the second moment method. For completeness we give the details.

Lemma 3.A.7. *If* $\mathbb{E}\left[X_k\right] \gg 1$, *then* $X_k \geqslant 1$ *with high probability.*

Proof. We have

$$\begin{aligned} \operatorname{Var}\left(X_k\right) &= \mathbb{E}\left[X_k^2\right] - \mathbb{E}\left[X_k\right]^2 = \sum_{\mathcal{O},\mathcal{O}' \in SF(G)} \mathbb{P}_m\left(M_k^{\mathcal{O}}(A) \cap M_k^{\mathcal{O}'}(A)\right) - \mathbb{E}\left[X_k\right]^2 \\ &= \mathbb{E}\left[X_k\right] + \sum_{\mathcal{O} \neq \mathcal{O}'} \mathbb{P}_m\left(M_k^{\mathcal{O}}(A) \cap M_k^{\mathcal{O}'}(A)\right) - \mathbb{E}\left[X_k\right]^2 \\ &= \mathbb{E}\left[X_k\right] + \big(1 + o(1)\big) \sum_{\mathcal{O} \neq \mathcal{O}'} \mathbb{P}_m\left(M_k^{\mathcal{O}}(A)\right)^2 - \mathbb{E}\left[X_k\right]^2, \end{aligned}$$

by Lemma 3.A.1. Therefore,

$$\operatorname{Var}\left(X_k\right) \leqslant \mathbb{E}\left[X_k\right] + \big(1 + o(1)\big)\,\mathbb{E}\left[X_k\right]^2 - \mathbb{E}\left[X_k\right]^2 = o\big(\mathbb{E}\left[X_k\right]^2\big).$$

Hence, by Chebyshev's inequality, we have $X_k \geqslant 1$ with high probability as $n \to \infty$. $\square$

It only remains to show that $\mathbb{E}\left[X_k\right]$ does not decay too quickly.

Lemma 3.A.8. *For every constant* $h > 0$, *we have*

$$\big|\,\mathbb{E}\left[X_k\right] - \mathbb{E}\left[X_{k+h}\right]\big| = o\big(\mathbb{E}\left[X_k\right]\big).$$

Proof. By Lemma 3.A.4, we have

$$\mathbb{E}\,X_k = \Omega\left(\frac{r(G)\sqrt{m}}{k} \exp\left(-\frac{2k^2}{m}\right)\right).$$

whereas, by Lemma 3.A.3, we have

$$\mathbb{E}\left[X_k\right] - \mathbb{E}\left[X_{k+h}\right] = O\left(\frac{r(G)}{\sqrt{m}} \sum_{x=k}^{k+h} \exp\left(-\frac{2x^2}{m}\right)\right) = O\left(\frac{r(G)}{\sqrt{m}} \exp\left(-\frac{2k^2}{m}\right)\right).$$

Since we assumed that $k \ll m$, the lemma follows. $\square$

Proof of Lemma 3.5.2. If $r(G) \leqslant n^\gamma$ then the lemma is trivial (set $b = 0$), so assume that $r(G) > n^\gamma$ and let $b = b(G, m)$ be minimal such that $\mathbb{E}\left[X_b(A)\right] \leqslant n^\gamma$. It follows that $\mathbb{E}\left[X_{b+h}(A)\right] \gg 1$, by Lemma 3.A.8, and hence that $X_{b+h}(A) \geqslant 1$ with high probability, by Lemma 3.A.7, as required. $\square$

Proof of Lemma 3.4.2

Let G be an even-order abelian group, and note that the lemma is trivial if $r(G) \leqslant \delta n$. Recall that $\mathcal{M}$ denotes the collection of odd cosets $\mathcal{O} \in SF(G)$ such that $|A \cap \mathcal{O}|$ is maximal. We are required to prove that with high probability there is an $\mathcal{O} \in \mathcal{M}$ such that $\mathcal{E} = G \setminus \mathcal{O}$ is nice. This is an immediate consequence of the following lemma. Recall that $\omega = \omega(n)$ is a function such that $\omega \to \infty$ slowly as $n \to \infty$.

Lemma 3.A.9. *With high probability, the following hold:*

(a) $|A \cap \mathcal{O}| \leqslant pn + \omega\sqrt{pn}$ *for every subgroup* $\mathcal{E} = G \setminus \mathcal{O}$ *which is not nice.*

(b) *There exists a nice subgroup* $\mathcal{E} = G \setminus \mathcal{O}$ *such that* $|A \cap \mathcal{O}| \geqslant pn + \omega\sqrt{pn}$.

Proof. Part (a) follows from Chernoff's inequality and the union bound, since there are at most $O(1/\delta)$ subgroups that are not nice. To prove part (b), we again couple with the hypergeometric distribution, and apply Lemma 3.5.2. Indeed, we have $|A| \geqslant 2pn - \omega\sqrt{pn}$ with high probability, and for each $m \geqslant 2pn - \omega\sqrt{pn}$ there exists a $b = b(G, m)$ such that $\mathbb{E}\left[X_b(A)\right] \leqslant \sqrt{n}$ and $X_b(A) \geqslant 1$ with high probability in $\mathbb{P}_m$. But, by Lemma 3.A.4, we have $\mathbb{E}\left[X_b(A)\right] = n^{1+o(1)}$ for $b = pn + \omega\sqrt{pn}$, and so we are done. $\square$

THE TYPICAL STRUCTURE OF GRAPHS WITH NO LARGE CLIQUES

The work in this chapter is joint with József Balogh, Neal Bushaw, Hong Liu, Robert Morris and Maryam Sharifzadeh. It is adapted from a preprint version [7].

4.1 INTRODUCTION

In this chapter we extend the result of Kolaitis, Prömel and Rothschild to K_{r+1}-free graphs, where $r = r(n)$ is a function which is allowed to grow with n. More precisely, we prove the following theorem.

Theorem 4.1.1. *Let $r = r(n) \in \mathbb{N}_0$ be a function satisfying $r \leqslant (\log n)^{1/4}$ for every $n \in \mathbb{N}$. Then almost all K_{r+1}-free graphs on n vertices are r-partite.*

Note that if $r \geqslant 2\log_2 n$ then almost all graphs are K_{r+1}-free (and almost none are r-partite if $r \ll n/\log n$), so the bound on r in Theorem 4.1.1 is not far from being best possible. It would be extremely interesting (and likely very difficult) to determine the largest $\alpha \in [1/4, 1]$ such that the theorem holds for some function $r = (\log n)^{\alpha+o(1)}$. It may well be the case that this supremum is equal to 1, though we are not prepared to state this as a conjecture.

Theorem 4.1.1 improves a recent result of Mousset, Nenadov and Steger [70], who showed that, for the same[1] family of functions $r = r(n)$, the number of n-vertex K_{r+1}-free graphs is

$$2^{t_r(n)+o(n^2/r)}, \tag{4.1}$$

where $t_r(n) = \mathrm{ex}(n, K_{r+1})$ denotes the number of edges of the Turán graph, the r-partite graph on n vertices with the maximum possible number of edges. The problem for H-free graphs with $v(H) \to \infty$ as $n \to \infty$ was first studied by Bollobás and Nikiforov [18], who proved bounds corresponding to (4.1) whenever $v(H) = o(\log n)$ and $\chi(H_n) = r+1$ is fixed. For more precise bounds for a fixed forbidden graph H, see [10], and for similar bounds in the hereditary (i.e., induced-H-free) setting, see [2, 12, 19] and the references therein.

The proof of Theorem 4.1.1 has three main ingredients. The first is the so-called 'hypergraph container method', which was recently developed by Balogh, Morris and Samotij [8], and independently by Saxton and Thomason [81]. This method was used by

[1] In fact, a very slightly weaker theorem was stated in [70], but a little additional case analysis easily gives the result for all $r \leqslant (\log n)^{1/4}$.

Mousset, Nenadov and Steger to prove Theorem 4.3.2, below, from which they deduced the bound (4.1) using a supersaturation theorem of Lovász and Simonovits [64].

In order to obtain the much more precise result stated in Theorem 4.1.1, we will use the method of Balogh, Bollobás and Simonovits [10, 11], who determined the structure of almost all H-free graphs for every fixed graph H. This powerful technique (see Sections 4.4 and 4.5) allows one to compare the number of K_{r+1}-free graphs that are 'close' to being r-partite, with the total number of K_{r+1}-free graphs.

The missing ingredient is the main new contribution of this work. In order to deduce from Theorem 4.3.2 a bound on the number of K_{r+1}-free graphs that are 'far' from being r-partite, we will need an analogue of the Lovász–Simonovits supersaturation result, mentioned above, for the well-known stability theorem of Erdős and Simonovits [38]. Although a weak such analogue can easily be obtained via the regularity lemma, this gives bounds which are far from sufficient for our purposes. Instead we will adapt a recent argument due to Füredi [46] in order to prove the following close-to-best-possible such result. We say that a graph G is t*-far from being* r*-partite*[2] if $\chi(G') > r$ for every subgraph $G' \subset G$ with $e(G') > e(G) - t$.

Theorem 4.1.2. *For every* $n, r, t \in \mathbb{N}$*, the following holds. Every graph* G *on* n *vertices which is* t*-far from being* r*-partite contains at least*

$$\frac{n^{r-1}}{e^{2r} \cdot r!} \left(e(G) + t - \left(1 - \frac{1}{r}\right) \frac{n^2}{2} \right)$$

copies of K_{r+1}.

Note that the graph obtained by adding t edges to the Turán graph $T_r(n)$ is t-far from being r-partite and has roughly $t \cdot (n/r)^{r-1}$ copies of K_{r+1}, so Theorem 4.1.2 is sharp to within a factor of roughly e^r. We prove this supersaturated stability theorem in Section 4.2, and use it in Section 4.3 to count the K_{r+1}-free graphs that are n^{2-1/r^2}-far from being r-partite. We prove various simple properties of almost all K_{r+1}-free graphs in Section 4.4, and finally, in Section 4.5, we use the Balogh–Bollobás–Simonovits method to deduce Theorem 4.1.1.

4.2 A SUPERSATURATED ERDŐS-SIMONOVITS STABILITY THEOREM

In this section, we prove our 'supersaturated stability theorem' for K_{r+1}-free graphs. As noted in the Introduction, we do so by adapting a proof of Füredi [46].

[2] Similarly, we say that G is t-close to being r-partite if it is not t-far from being r-partite.

Given a graph G, a vertex $v \in V(G)$ and an integer $m \in \mathbb{N}$, let us write $K_m(G)$ for the number of m-cliques in G, and $K_m(v)$ for the number of such m-cliques containing v.

Proof of Theorem 4.1.2. We will prove by induction on r that

$$K_{r+1}(G) \geqslant \frac{n^{r-1}}{c(r)} \left(e(G) + t - \left(1 - \frac{1}{r}\right) \frac{n^2}{2} \right), \tag{4.2}$$

where $c(r) := 2(r+1)^{r-1} r^{r-1}/r!$, for every graph G on n vertices which is t-far from being r-partite. Since $c(r) \leqslant e^{2r} r!$, the theorem follows from (4.2).

Note first that the theorem holds in the case $r = 1$, since a graph is t-far from being 1-partite if and only if $e(G) \geqslant t$, and hence G has more than $\frac{e(G)+t}{2}$ copies of K_2, as required. So let $r \geqslant 2$ and assume that the result holds for $r-1$. Let $n, t \in \mathbb{N}$, and let G be a graph that is t-far from being r-partite.

First, for each $v \in V(G)$, set $B_v = N(v)$ (the set of neighbours of v in G) and $A_v = V(G) \setminus B_v$, and observe that

$$\sum_{u \in A_v} d(u) = e(G) + e(A_v) - e(B_v), \tag{4.3}$$

where $e(X)$ denotes the number of edges in the graph $G[X]$. Now, the graph $G[B_v]$ is $\big(t - e(A_v)\big)$-far from being $(r-1)$-partite, and so, by the induction hypothesis,

$$K_{r+1}(v) \geqslant \frac{|B_v|^{r-2}}{c(r-1)} \left(e(B_v) + t - e(A_v) - \left(1 - \frac{1}{r-1}\right) \frac{|B_v|^2}{2} \right), \tag{4.4}$$

since each copy of K_r in $G[B_v]$ corresponds to a copy of K_{r+1} in G that contains v.

Combining (4.3) and (4.4), noting that $|B_v| = d(v)$, and summing over v, it follows that

$$(r+1) \cdot K_{r+1}(G) \geqslant \sum_{v \in V(G)} \frac{d(v)^{r-2}}{c(r-1)} \left(e(G) + t - \sum_{u \in A_v} d(u) - \left(1 - \frac{1}{r-1}\right) \frac{d(v)^2}{2} \right).$$

We claim that

$$\sum_{v \in V(G)} \sum_{u \in A_v} d(u) d(v)^{r-2} \leqslant \sum_{v \in V(G)} \sum_{u \in A_v} d(v)^{r-1} = \sum_{v \in V(G)} d(v)^{r-1} \big(n - d(v)\big). \tag{4.5}$$

Indeed, let $X = \big\{(v,u) : v \in V(G),\ u \in A_v\big\}$ denote the set of ordered pairs in the sum above, and note that $(v,u) \in X$ if and only if $uv \notin E(G)$. Since X is symmetric, the inequality in (4.5) now follows immediately for $r = 2$, and by the Cauchy-Schwarz inequality

$$\sum_{(v,u) \in X} d(u) d(v) \leqslant \left(\sum_{(v,u) \in X} d(u)^2 \right)^{1/2} \left(\sum_{(v,u) \in X} d(v)^2 \right)^{1/2}$$

for $r = 3$. For $r \geqslant 4$, applying Hölder's inequality with $p = r-2$ and $q = (r-2)/(r-3)$ gives

$$\sum_{(v,u)\in X} d(u)d(v)^{r-2} \leqslant \left(\sum_{(v,u)\in X} d(u)^{r-2}d(v) \right)^{1/p} \left(\sum_{(v,u)\in X} d(v)^{r-1} \right)^{1/q},$$

since $\left(r - 2 - \frac{1}{r-2}\right)\frac{r-2}{r-3} = \frac{r^2-4r+3}{r-3} = r - 1$. Once again using the symmetry of X, and noting that $1 - 1/p = 1/q$, the claimed inequality (4.5) follows.

Combining the inequalities above, we obtain

$$(r+1) \cdot K_{r+1}(G) \geqslant \sum_{v\in V(G)} \frac{d(v)^{r-2}}{c(r-1)} \left(e(G) + t - d(v)n + \left(1 + \frac{1}{r-1}\right) \frac{d(v)^2}{2} \right).$$

Since the factor in parentheses is minimised when $d(v) = \frac{r-1}{r} \cdot n$, it follows that

$$(r+1) \cdot K_{r+1}(G) \geqslant \sum_{v\in V(G)} \frac{d(v)^{r-2}}{c(r-1)} \left(e(G) + t - \left(1 - \frac{1}{r}\right) \frac{n^2}{2} \right).$$

Finally, note that every graph G is $\big(e(G)/r\big)$-close to being r-partite (take a random partition), and hence we may assume that $\big(1+\frac{1}{r}\big)e(G) \geqslant \big(1-\frac{1}{r}\big)\frac{n^2}{2}$, since otherwise the theorem is trivial. Thus, by the convexity of x^{r-2},

$$\sum_{v\in V(G)} d(v)^{r-2} \geqslant n \cdot \left(\frac{2e(G)}{n}\right)^{r-2} \geqslant \left(\frac{r-1}{r+1}\right)^{r-2} n^{r-1},$$

and so, since $c(r-1) \cdot (r+1)^{r-1} = c(r) \cdot (r-1)^{r-2}$, it follows that

$$K_{r+1}(G) \geqslant \frac{n^{r-1}}{c(r)} \left(e(G) + t - \left(1 - \frac{1}{r}\right) \frac{n^2}{2} \right),$$

as claimed. □

4.3 AN APPROXIMATE STRUCTURAL RESULT

In this section we will prove the following approximate version of Theorem 4.1.1.

Theorem 4.3.1. *Let $r = r(n) \in \mathbb{N}$ be a function satisfying $r \leqslant (\log n)^{1/4}$ for each $n \in \mathbb{N}$. Then almost all K_{r+1}-free graphs on n vertices are n^{2-1/r^2}-close to being r-partite.*

Theorem 4.3.1 is a straightforward consequence of Theorem 4.1.2 and the following theorem, which was proved by Mousset, Nenadov and Steger [70] using the hypergraph container method of Balogh, Morris and Samotij [8] and Saxton and Thomason [81]. The following theorem is slightly stronger than the result stated in [70], but follows easily from essentially the same proof.

Theorem 4.3.2. *Let $r = r(n) \in \mathbb{N}$ be a function satisfying $r \leqslant (\log n)^{1/4}$ for each sufficiently large $n \in \mathbb{N}$. There exists a collection $\mathcal{C}$ of graphs such that the following hold:*

(a) *every K_{r+1}-free graph on n vertices is a subgraph of some $G \in \mathcal{C}_n$,*

(b) *$K_{r+1}(G) \leqslant n^{r+1-2/r^2}$ for every $G \in \mathcal{C}_n$, and*

(c) *$|\mathcal{C}_n| \leqslant \exp\left(n^{2-2/r^2}\right)$,*

where $\mathcal{C}_n = \left\{G \in \mathcal{C} : v(G) = n\right\}$.

Deducing Theorem 4.3.1 from Theorems 4.1.2 and 4.3.2 is straightforward.

Proof of Theorem 4.3.1. For each $t \in \mathbb{N}$, set

$$\mathcal{F}_t = \left\{G : e(G) \geqslant \left(1 - \frac{1}{r}\right)\frac{|G|^2}{2} - \frac{t}{2} \text{ and } G \text{ is } t\text{-far from being } r\text{-partite}\right\},$$

and observe that if $G \in \mathcal{F}_t$, then

$$K_{r+1}(G) \geqslant \frac{|G|^{r-1} \cdot t}{e^{2r+1} \cdot r!},$$

by Theorem 4.1.2. Therefore, letting $\mathcal{C}$ be the collection of graphs given by Theorem 4.3.2, and setting $t = n^{2-1/r^2}$, it follows from property (b) and the bound $r \leqslant (\log n)^{1/4}$ that $\mathcal{C}_n \cap \mathcal{F}_t = \emptyset$.

Now, for each K_{r+1}-free graph G on n vertices that is n^{2-1/r^2}-far from being r-partite, we have $G \in C$ for some $C \in \mathcal{C}_n$, and by the observations above and the definition of $\mathcal{F}_t$, it follows that

$$e(C) \leqslant \left(1 - \frac{1}{r}\right)\frac{n^2}{2} - \frac{t}{2}.$$

Therefore, summing over all such containers, the number of such graphs is at most

$$\exp\left(n^{2-2/r^2}\right) \cdot 2^{t_r(n) - t/2} \ll 2^{t_r(n) - t/4},$$

which is clearly smaller than the number of K_{r+1}-free graphs on n vertices, as required. □

4.4 SOME PROPERTIES OF A TYPICAL K_{r+1}-FREE GRAPH

In this section we will prove some useful structural properties of almost all K_{r+1}-free graphs. These structural properties will allow us (in Section 4.5) to count the K_{r+1}-free graphs that are close to being r-partite, and hence to complete the proof of Theorem 4.1.1. We emphasise that the lemmas in this section were all proved for fixed $r \in \mathbb{N}$

in [10], and no extra ideas are required in order to extend their proofs to our more general setting.

Let us fix throughout this section a function $2 \leqslant r = r(n) \leqslant (\log n)^{1/4}$, and let us denote by $\mathcal{G}$ the collection of K_{r+1}-free graphs on n vertices that are n^{2-1/r^2}-close to being r-partite. We begin with two simple definitions.

Definition 4.4.1 (Optimal partitions)**.** An r-partition $(U_1, \ldots, U_r)$ of the vertex set of a graph G is called *optimal* if the number of interior edges, $\sum_{i=1}^{r} e(U_i)$, is minimised.

Definition 4.4.2 (Uniformly dense graphs)**.** We say that a graph G is *uniformly dense* if for every optimal r-partition $(U_1, \ldots, U_r)$ and every pair $\{i, j\} \subset [r]$, we have

$$e(A, B) > \frac{|A||B|}{32} \tag{4.6}$$

for every $A \subset U_i$ and $B \subset U_j$ with $|A| = |B| \geqslant 2^{-8r} n$.

Lemma 4.4.3. *The number of graphs in $\mathcal{G}$ that are not uniformly dense is at most*

$$2^{t_r(n) - 2^{-17r} n^2},$$

and therefore almost all K_{r+1}-free graphs are uniformly dense.

Proof. In order to count such graphs, we first choose the optimal partition $\mathcal{U} = (U_1, \ldots, U_r)$, the pair $\{i, j\} \subset [r]$, and the sets $A \subset U_i$ and $B \subset U_j$ for which (4.6) fails. We then choose the edges between A and B, and finally the remaining edges. Note first that we have at most r^n choices for $\mathcal{U}$, at most r^2 choices for $\{i, j\}$, and at most 2^{2n} choices for the pair (A, B).

Now, the number of choices for the edges between A and B is at most

$$\sum_{k=0}^{|A||B|/32} \binom{|A||B|}{k} \leqslant n^2 (32e)^{|A||B|/32} \leqslant 2^{|A||B|/4},$$

and the number of choices for the remaining edges is at most

$$2^{t_r(n) - |A||B|} \binom{n^2}{n^{2-1/r^2}} \leqslant 2^{t_r(n) - |A||B|} \exp\left(n^{2-1/r^2} \log n\right) \leqslant 2^{t_r(n) - |A||B|/2},$$

since $\mathcal{U}$ is optimal, $|A||B| \geqslant 2^{-16r} n^2$, and each $G \in \mathcal{G}$ is n^{2-1/r^2}-close to being r-partite.

It follows that the number of graphs in $\mathcal{G}$ that are not uniformly dense is at most

$$r^{n+2} \cdot 2^{2n} \cdot 2^{t_r(n) - |A||B|/4} \leqslant 2^{t_r(n) - 2^{-17r} n^2},$$

as claimed. □

Our next definition controls the maximum degree inside the parts of an optimal partition.

Definition 4.4.4 (Internally sparse graphs)**.** A graph G is said to be *internally sparse* if, for every optimal partition $\mathcal{U} = (U_1, \ldots, U_r)$ of G, we have

$$\Delta\big(G[U_i]\big) \leqslant 2^{-3r}n. \tag{4.7}$$

for every $1 \leqslant i \leqslant r$. Otherwise we say that G is *internally dense*.

Lemma 4.4.5. *If $G \in \mathcal{G}$ is internally dense then it is not uniformly dense.*

We will prove Lemma 4.4.5 using the following embedding lemma[3] from [3].

Lemma 4.4.6. *Let $0 < \alpha < 1$, G be a graph, and $W_1, \ldots, W_r \subset V(G)$ be disjoint sets of vertices. Suppose that for every pair $\{i, j\} \subset [r]$ and every pair of sets $A \subset W_i$ and $B \subset W_j$ with $|A| \geqslant \alpha^r |W_i|$ and $|B| \geqslant \alpha^r |W_j|$, we have $e(A, B) > \alpha |A||B|$.*

Then G contains a copy of K_r with one vertex in each set W_j.

Proof of Lemma 4.4.5. Suppose for a contradiction that $G \in \mathcal{G}$ is both internally dense and uniformly dense. Let $\mathcal{U} = (U_1, \ldots, U_r)$ be the optimal partition given by Definition 4.4.4, and suppose that $v \in U_1$ has degree at least $2^{-3r}n$ in $G[U_1]$. For each $i \in [r]$, let $W_i = N(v) \cap U_i$, and observe that $|W_i| \geqslant 2^{-3r}n$, since $\mathcal{U}$ is optimal.

Observe that $W_1, \ldots, W_r$ satisfy the conditions of Lemma 4.4.6 with $\alpha = 1/32$, since G is uniformly dense, so $e(A, B) > |A||B|/32$ for every pair $\{i, j\} \subset [r]$, and every $A \subset U_i$ and $B \subset U_j$ with $|A| = |B| \geqslant 2^{-8r}n$. Thus, by Lemma 4.4.6, there exists a copy of K_r in the neighbourhood of v, which (including v) gives a copy of K_{r+1} in G. But this is a contradiction, since our graph is K_{r+1}-free, and so every internally dense graph $G \in \mathcal{G}$ is not uniformly dense, as claimed. $\square$

Our final definition controls the sizes of the parts in an optimal partition.

Definition 4.4.7 (Balanced graphs)**.** A graph G is said to be *balanced* if, for every optimal partition $\mathcal{U} = (U_1, \ldots, U_r)$ of G, we have

$$\frac{n}{r} - 2^{-3r}n \leqslant |U_i| \leqslant \frac{n}{r} + 2^{-3r}n \tag{4.8}$$

for every $1 \leqslant i \leqslant r$. Otherwise we say that G is *unbalanced*.

[3] In fact, the version stated here is slightly more general than [3, Lemma 3.1], but follows from exactly the same proof.

Lemma 4.4.8. *The number of unbalanced graphs in $\mathcal{G}$ is at most*

$$2^{t_r(n)-2^{-8r}n^2},$$

and therefore almost all K_{r+1}-free graphs are balanced.

Proof. Let $G \in \mathcal{G}$ be an unbalanced graph, and let $\mathcal{U} = (U_1, \ldots, U_r)$ be an optimal partition of G for which (4.8) fails. Note that

$$\sum_{i=1}^{r-1} \sum_{j=i+1}^{r} |U_i||U_j| \leqslant t_r(n) - 2^{-7r}n^2,$$

since moving a vertex from a set of size at least $n/r + a$ to one of size $n/r - b$ creates at least $a + b$ new potential cross edges. The number of such graphs $G \in \mathcal{G}$ is therefore at most

$$r^n \cdot 2^{t_r(n)-2^{-7r}n^2} \cdot \binom{n^2}{n^{2-1/r^2}} \leqslant 2^{t_r(n)-2^{-8r}n^2},$$

as claimed. □

4.5 THE PROOF OF THEOREM 4.1.1

In this section we will deduce Theorem 4.1.1 from Theorem 4.3.1, using the method of Balogh, Bollobás and Simonovits [10, 11]. Recall from the previous section that almost all K_{r+1}-free graphs are uniformly dense, internally sparse and balanced.

Let us fix throughout this section a function $2 \leqslant r = r(n) \leqslant (\log n)^{1/4}$.

Definition 4.5.1. Let $\mathcal{Q}(n, r)$ denote the collection of K_{r+1}-free graphs on n vertices that are not r-partite, but are n^{2-1/r^2}-close to being r-partite, and are moreover uniformly dense, internally sparse and balanced.

Let $\mathcal{K}(n, r)$ denote the collection of K_{r+1}-free graphs on n vertices. We will prove the following proposition, which completes the proof of Theorem 4.1.1.

Proposition 4.5.2. *For every sufficiently large $n \in \mathbb{N}$,*

$$|\mathcal{Q}(n, r)| \leqslant 2^{-2^{-10r}n} \cdot |\mathcal{K}(n, r)|.$$

The idea of the proof is as follows. We will define a collection of bipartite graphs F_m (see Definition 4.5.8) with parts $\mathcal{Q}(n, r, m)$ and $\mathcal{K}(n, r)$, where the sets $\mathcal{Q}(n, r, m)$ form a partition of $\mathcal{Q}(n, r)$ (see Definitions 4.5.4 and 4.5.5). These bipartite graphs will have the following property: the degree in F_m of each $G \in \mathcal{Q}(n,r,m)$ will be significantly larger than the degree of each $G \in \mathcal{K}(n,r)$ (see Lemmas 4.5.10 and 4.5.12). The result will then follow by double counting the edges of each F_m and summing over m.

In order to define $\mathcal{Q}(n,r,m)$ and F_m, we will need the following simple concept.

Definition 4.5.3 (Bad sets)**.** Let G be a graph and let $U \subset V(G)$. A set of r vertices $R \subset V(G) \setminus U$ is said to be *bad* towards U if it has no common neighbour in U.

In the following definition we may choose the partition $\mathcal{U}$ and the sets $X^{(1)}, \dots, X^{(r)}$ arbitrarily, subject to the given conditions.

Definition 4.5.4. For each $G \in \mathcal{Q}(n,r)$, fix an optimal partition $\mathcal{U} = (U_1, \dots, U_r)$ of $V(G)$, and for each $j \in [r]$ choose a maximal collection of vertex-disjoint sets $X^{(j)} = \big\{R_1^{(j)}, \dots, R_{\ell(j)}^{(j)}\big\}$ such that $R_i^{(j)}$ is bad towards U_j for each $i \in [\ell(j)]$. We define

$$m(G) \,:=\, \max\big\{\ell(j) \,:\, j \in [r]\big\},$$

let $j(G)$ denote the smallest j for which this maximum is attained, and set

$$X(G) \,:=\, R_1^{(j(G))} \cup \dots \cup R_{\ell(j(G))}^{(j(G))}.$$

With this definition in place, it is natural to partition $\mathcal{Q}(n,r)$ by the size of $m(G)$.

Definition 4.5.5. For each $m \in \mathbb{N}$, we define

$$\mathcal{Q}(n,r,m) \,=\, \big\{G \in \mathcal{Q}(n,r) \,:\, m(G) = m\big\}.$$

Before continuing, let us note a simple but key fact.

Lemma 4.5.6. $m(G) \geqslant 1$ *for every* $G \in \mathcal{Q}(n,r)$.

Proof. This follows from the fact that G is not r-partite. Indeed, suppose that $m(G) = 0$ and let $x_0x_1 \in E(G[U_1])$ be an 'interior' edge of G with respect to $\mathcal{U}$. Since there are no bad r-sets towards U_j for any $j \in [r]$, we can recursively choose vertices $x_j \in U_j$ such that $\{x_0, \dots, x_j\}$ forms a clique. But this is a contradiction, since G is K_{r+1}-free. □

In order to establish an upper bound on those m which we need to consider, we count those graphs in $\mathcal{Q}(n,r)$ for which $m(G)$ is large.

Lemma 4.5.7. *If* $m \geqslant 2^{-6r}n$, *then*

$$|\mathcal{Q}(n,r,m)| \,\leqslant\, 2^{t_r(n) - mn/2^{3r}}.$$

Proof. Let $m \geqslant 2^{-6r}n$, and consider the number of ways of constructing a graph $G \in \mathcal{Q}(n,r,m)$. We have at most r^n choices for the partition $\mathcal{U}$, at most $\binom{n}{r}^m$ choices for the set $X(G)$, and r choices for $j = j(G)$. Moreover, we have at most

$$2^{t_r(n)-|U_j||X(G)|}\big(2^r-1\big)^{|U_j||X(G)|/r} \leqslant 2^{t_r(n)-mn/2^{2r}}$$

choices for the edges between different parts of $\mathcal{U}$, since $X(G)$ is composed of r-sets that are bad towards U_j, and G is balanced. Finally, we have at most $n^{O(n^{2-1/r^2})}$ choices for the edges inside parts of $\mathcal{U}$, since G is n^{2-1/r^2}-close to being r-partite.

It follows that

$$|\mathcal{Q}(n,r,m)| \leqslant r^n \cdot \binom{n}{r}^m \cdot r \cdot n^{O(n^{2-1/r^2})} \cdot 2^{t_r(n)-mn/2^{2r}} \leqslant 2^{t_r(n)-mn/2^{3r}}$$

as required, since $m \geqslant 2^{-6r}n$, so $n^{2-1/r^2}\log n \ll 2^{-3r}mn$. □

From now on, let us fix a function $1 \leqslant m = m(n) \leqslant 2^{-6r}n$. We are ready to define the bipartite graph F_m.

Definition 4.5.8. Define a map $\Phi_m \colon \mathcal{Q}(n,r,m) \to 2^{\mathcal{K}(n,r)}$ by placing $H \in \Phi_m(G)$ if and only if H can be constructed from G by first removing all edges of G that are incident to $X(G)$, and then adding an arbitrary subset of the edges between $X(G)$ and $V(G) \setminus \big(X(G) \cup U_{j(G)}\big)$.

Let F_m be the bipartite graph with edge set $\{(G,H) : H \in \Phi_m(G)\}$.

We first observe that the map Φ_m is well-defined.

Lemma 4.5.9. *If* $G \in \mathcal{Q}(n,r,m)$ *and* $H \in \Phi_m(G)$, *then* H *is* K_{r+1}*-free.*

Proof. This follows easily from the fact that G is K_{r+1}-free, and the maximality of $X(G)$. Indeed, if there exists a copy of K_{r+1} in H, then it must contain a vertex of $X(G)$, and therefore it must contain no other vertices of $X(G) \cup U_{j(G)}$. Hence it contains exactly r vertices of $V(G) \setminus \big(X(G) \cup U_{j(G)}\big)$, and by the maximality of $X(G)$ these have a common neighbour in $U_{j(G)}$. But this contradicts our assumption that G is K_{r+1}-free, as required. □

We are now ready to prove our first bound on the degrees in F_m.

Lemma 4.5.10. *For every* $G \in \mathcal{Q}(n,r,m)$,

$$\log_2 |\Phi_m(G)| \geqslant \left(1 - \frac{1}{r} - \frac{1}{2^{3r}} - \frac{mr}{n}\right) mnr.$$

Proof. This follows immediately from the fact that G is balanced. Indeed, we have two choices for each of the

$$|X(G)| \cdot \big|V(G) \setminus \big(X(G) \cup U_{j(G)}\big)\big| \geqslant mr \cdot \left(1 - \frac{1}{r} - \frac{1}{2^{3r}} - \frac{mr}{n}\right) n \tag{4.9}$$

potential edges between $X(G)$ and $V(G) \setminus \big(X(G) \cup U_{j(G)}\big)$. $\square$

In order to bound the degrees in F_m of vertices in $\mathcal{K}(n,r)$, we will need the following lemma, which counts the optimal partitions in the neighbourhood of such a vertex. We note that here, the upper bound on m from Lemma 4.5.7 is crucial.

Lemma 4.5.11. *For each $H \in \mathcal{K}(n,r)$, there are at most $2^{n/2^{4r}}$ distinct partitions $\mathcal{U}$ of $V(H)$ such that $\mathcal{U}$ is an optimal partition of some graph $G \in \Phi_m^{-1}(H)$.*

Proof. We will use the fact that each $G \in \Phi_m^{-1}(H)$ is uniformly dense and n^{2-1/r^2}-close to being r-partite to show that the optimal partitions in question must be 'close' to one another.

To be precise, let $G_1, G_2 \in \Phi_m^{-1}(H)$, and let $\mathcal{U} = (U_1, \ldots, U_r)$ be an optimal partition of G_1 and $\mathcal{V} = (V_1, \ldots, V_r)$ be an optimal partition of G_2. We claim that

$$\big|\big\{ j \in [r] \,:\, |U_i \cap V_j| > 2^{-6r} n + 2mr \big\}\big| \leqslant 1$$

for every $i \in [r]$. Indeed, suppose that

$$\big|U_i \cap V_j\big| > 2^{-6r} n + 2mr \qquad \text{and} \qquad \big|U_i \cap V_{j'}\big| > 2^{-6r} n + 2mr,$$

set $A = \big(U_i \cap V_j\big) \setminus \big(X(G_1) \cup X(G_2)\big)$ and $B = (U_i \cap V_{j'}) \setminus \big(X(G_1) \cup X(G_2)\big)$, and note that, since G_2 is uniformly dense, we have $e_{G_2}(A,B) > |A||B|/32 > 2^{-12r-5} n^2$. But these edges are all contained in U_i, so this contradicts the fact that G_1 is n^{2-1/r^2}-close to being r-partite, as required.

It follows that (by renumbering the parts if necessary) we have

$$\big|U_i \setminus V_i\big| \leqslant r \cdot \big(2^{-6r} n + 2mr\big) \leqslant 2^{-5r} n$$

for every $i \in [r]$, where second inequality follows since $m \leqslant 2^{-6r} n$. Set $D_i = U_i \setminus V_i$, and observe that the partition $\mathcal{U}$ and the collection $(D_1, \ldots, D_r)$ together determine $\mathcal{V}$. It follows that the number of optimal partitions is at most

$$\left(\sum_{k=0}^{2^{-5r} n} \binom{n}{k} \right)^r \leqslant n^r \cdot \binom{n}{2^{-5r} n}^r \leqslant 2^{r \log n} \cdot \big(e 2^{5r}\big)^{r 2^{-5r} n} \leqslant 2^{n/2^{4r}},$$

as required. $\square$

We can now bound the degrees on the right. Recall than in Definition 4.5.4 we chose a 'canonical' optimal partition for each graph $G \in \mathcal{Q}(n, r)$.

Lemma 4.5.12. *We have*

$$\log_2 \left|\Phi_m^{-1}(H)\right| \leqslant \left(1 - \frac{1}{r} - \frac{1}{2^{2r}}\right) mnr$$

for every $H \in \mathcal{K}(n, r)$.

Proof. First let us fix a partition $\mathcal{U} = (U_1, \ldots, U_r)$, and count the number of graphs $G \in \mathcal{Q}(n, r, m)$ with $\Phi_m(G) = H$ whose optimal partition is $\mathcal{U}$. To do so, first note that we have $\binom{n}{r}^m$ choices for $X(G)$, and at most r choices for $j = j(G)$. Now, since G is internally sparse and balanced, each vertex $v \in X(G)$ has at most $2^{-3r}n$ neighbours in its own part of $\mathcal{U}$, and $\big||U_i| - n/r\big| \leqslant n/2^{3r}$ for each $i \in [r]$. Thus we have at most

$$\binom{n}{2^{-3r}n} \cdot 2^{(1-2/r+1/2^{3r})n} \leqslant 2^{(1-2/r+1/2^{2r})n}$$

choices for the edges between each vertex $v \in X(G)$ and $V(G) \setminus U_j$. Finally, by the definition of bad sets, and since G is balanced, we have at most

$$(2^r - 1)^{(1/r+1/2^{3r})mn} \leqslant 2^{(1/r-3/2^{2r})mnr}$$

choices for the edges between $X(G)$ and U_j.

Since, by Lemma 4.5.11, we have at most $2^{n/2^{4r}}$ choices for the partition $\mathcal{U}$, it follows that

$$\begin{aligned}\log_2 \left|\Phi_m^{-1}(H)\right| &\leqslant mr \log n + \log r + \left(1 - \frac{2}{r} + \frac{1}{2^{2r}} + \frac{1}{r} - \frac{3}{2^{2r}} + \frac{1}{2^{4r}}\right) mnr \\ &\leqslant \left(1 - \frac{1}{r} - \frac{1}{2^{2r}}\right) mnr,\end{aligned}$$

as claimed. □

Finally we put the pieces together and prove Proposition 4.5.2.

Proof of Proposition 4.5.2. We claim first that

$$|\mathcal{Q}(n, r, m)| \leqslant 2^{-2^{-9r}mnr} \cdot |\mathcal{K}(n, r)| \tag{4.10}$$

for every $m \leqslant 2^{-6r}n$. To prove this, we simply double count the edges of F_m, using Lemmas 4.5.10 and 4.5.12. Indeed, we have

$$\log_2 \left(\frac{|\mathcal{Q}(n, r, m)|}{|\mathcal{K}(n, r)|}\right) \leqslant \left(1 - \frac{1}{r} - \frac{1}{2^{2r}}\right) mnr - \left(1 - \frac{1}{r} - \frac{1}{2^{3r}} - \frac{mr}{n}\right) mnr,$$

which implies (4.10) since $m \leqslant 2^{-6r}n$.

Summing (4.10) over m, and recalling that G is n^{2-1/r^2}-close to being r-partite, we obtain

$$|\mathcal{Q}(n,r)| \leqslant \sum_{m=1}^{2^{-6r}n} 2^{-2^{-9r}mnr} \cdot |\mathcal{K}(n,r)| + \sum_{m=2^{-6r}n}^{n} 2^{t_r(n)-mn/2^{3r}} \leqslant 2^{-2^{-10r}n} \cdot |\mathcal{K}(n,r)|,$$

by Lemmas 4.5.6 and 4.5.7, as required. □

Finally, let us deduce Theorem 4.1.1.

Proof of Theorem 4.1.1. By Theorem 4.3.1, almost all K_{r+1}-free graphs on n vertices are n^{2-1/r^2}-close to r-partite. We further showed in Lemmas 4.4.3, 4.4.5, and 4.4.8 that almost all of these graphs are either r-partite, or in $\mathcal{Q}(n,r)$. Since by Proposition 4.5.2, for sufficiently large n, the size of $\mathcal{Q}(n,r)$ is (almost) exponentially small compared to $\mathcal{K}(n,r)$, it follows that almost all K_{r+1}-free graphs are r-partite, as required. □

[1] N. Alon. "Independent sets in regular graphs and sum-free subsets of finite groups". *Israel J. Math.* 73.2 (1991), pp. 247–256.

[2] N. Alon, J. Balogh, B. Bollobás, and R. Morris. "The structure of almost all graphs in a hereditary property". *J. Combin. Theory Ser. B* 101.2 (2011), pp. 85–110.

[3] N. Alon, J. Balogh, P. Keevash, and B. Sudakov. "The number of edge colorings with no monochromatic cliques". *J. London Math. Soc. (2)* 70.2 (2004), pp. 273–288.

[4] N. Alon, J. Balogh, R. Morris, and W. Samotij. "A refinement of the Cameron-Erdős conjecture". *Proc. Lond. Math. Soc. (3)* 108.1 (2014), pp. 44–72.

[5] N. Alon and J. H. Spencer. *The probabilistic method*. 3rd ed. Wiley-Interscience Series in Discrete Mathematics and Optimization. John Wiley & Sons, New Jersey, 2008, pp. xviii+352.

[6] L. Babai, M. Simonovits, and J. Spencer. "Extremal subgraphs of random graphs". *J. Graph Theory* 14.5 (1990), pp. 599–622.

[7] J. Balogh, N. Bushaw, M. Collares Neto, H. Liu, R. Morris, and M. Sharifzadeh. "The typical structure of graphs with no large cliques". Preprint (2014). arXiv: `1406.6961 [math.CO]`.

[8] J. Balogh, R. Morris, and W. Samotij. "Independent sets in hypergraphs". Preprint (2012). arXiv: `1204.6530 [math.CO]`.

[9] J. Balogh, R. Mycroft, and A. Treglown. "A random version of Sperner's theorem". Preprint (2014). arXiv: `1404.5079 [math.CO]`.

[10] J. Balogh, B. Bollobás, and M. Simonovits. "The number of graphs without forbidden subgraphs". *J. Combin. Theory Ser. B* 91.1 (2004), pp. 1–24.

[11] J. Balogh, B. Bollobás, and M. Simonovits. "The typical structure of graphs without given excluded subgraphs". *Random Structures Algorithms* 34.3 (2009), pp. 305–318.

[12] J. Balogh and J. Butterfield. "Excluding induced subgraphs: critical graphs". *Random Structures Algorithms* 38.1-2 (2011), pp. 100–120.

[13] J. Balogh, R. Morris, and W. Samotij. "Random sum-free subsets of abelian groups". *Israel J. Math.* 199.2 (2014), pp. 651–685.

[14] B. Bollobás. "On generalized graphs". *Acta Math. Acad. Sci. Hungar* 16 (1965), pp. 447–452.

[15] B. Bollobás and A. Thomason. "Threshold functions". *Combinatorica* 7.1 (1987), pp. 35–38.

[16] B. Bollobás. *Extremal graph theory*. Reprint of the 1978 original. Dover Publications, Inc., Mineola, 2004, pp. xx+488.

[17] B. Bollobás. "Threshold functions for small subgraphs". *Math. Proc. Cambridge Philos. Soc.* 90.2 (1981), pp. 197–206.

[18] B. Bollobás and V. Nikiforov. "The number of graphs with large forbidden subgraphs". *European J. Combin.* 31.8 (2010), pp. 1964–1968.

[19] B. Bollobás and A. Thomason. "Projections of bodies and hereditary properties of hypergraphs". *Bull. London Math. Soc.* 27.5 (1995), pp. 417–424.

[20] N. Bushaw, M. Collares Neto, R. Morris, and P. Smith. "The sharp threshold for maximum-size sum-free subsets in even-order abelian groups". Preprint (2013). arXiv: `1310.3236 [math.CO]`.

[21] P. J. Cameron and P. Erdős. "On the number of sets of integers with various properties". *Number theory (Banff, AB, 1988)*. de Gruyter, Berlin, 1990, pp. 61–79.

[22] V. Chvátal. "The tail of the hypergeometric distribution". *Discrete Math.* 25.3 (1979), pp. 285–287.

[23] M. Collares Neto and R. Morris. "Maximum-size antichains in random sets". Preprint (2014). arXiv: `1404.5258 [math.CO]`.

[24] D. Conlon and W. T. Gowers. "Combinatorial theorems in sparse random sets". Preprint (2010). arXiv: `1011.4310 [math.CO]`.

[25] D. Conlon, W. T. Gowers, W. Samotij, and M. Schacht. "On the KŁR conjecture in random graphs". *Israel J. Math.* 203.1 (2014), pp. 535–580.

[26] S. Das, W. Gan, and B. Sudakov. "Sperner's Theorem and a Problem of Erdős-Katona-Kleitman". Preprint (2013). arXiv: `1302.5210 [math.CO]`.

[27] B. DeMarco and J. Kahn. "Mantel's theorem for random graphs". *Random Structures Algorithms* (2014). To appear.

[28] P. H. Diananda and H. P. Yap. "Maximal sum-free sets of elements of finite groups". *Proc. Japan Acad.* 45 (1969), pp. 1–5.

[29] P. Erdős. “On a lemma of Littlewood and Offord”. *Bull. Amer. Math. Soc.* 51 (1945), pp. 898–902.

[30] P. Erdős. “On some of my conjectures in number theory and combinatorics”. *Proceedings of the fourteenth Southeastern conference on combinatorics, graph theory and computing (Boca Raton, Fla., 1983)*. Vol. 39. 1983, pp. 3–19.

[31] P. Erdős. “Problems and results on combinatorial number theory. II”. *J. Indian Math. Soc. (N.S.)* 40.1–4 (1976), 285–298 (1977).

[32] P. Erdős. “Some recent progress on extremal problems in graph theory”. *Proceedings of the Sixth Southeastern Conference on Combinatorics, Graph Theory and Computing (Florida Atlantic Univ., Boca Raton, Fla., 1975)*. Utilitas Math., Winnipeg, Man., 1975, 3–14. Congressus Numerantium, No. XIV.

[33] P. Erdős, P. Frankl, and V. Rödl. “The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent”. *Graphs Combin.* 2.2 (1986), pp. 113–121.

[34] P. Erdős and A. Hajnal. “On decomposition of graphs”. *Acta Math. Acad. Sci. Hungar.* 18 (1967), pp. 359–377.

[35] P. Erdős and A. Hajnal. “Problems and results in finite and infinite combinatorial analysis.” *Ann. New York Acad. Sci.* 175 (1970), pp. 115–124.

[36] P. Erdős, D. J. Kleitman, and B. L. Rothschild. “Asymptotic enumeration of K_n-free graphs”. *Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973), Tomo II*. Accad. Naz. Lincei, Rome, 1976, 19–27. Atti dei Convegni Lincei, No. 17.

[37] P. Erdős and A. Rényi. “On random graphs. I”. *Publ. Math. Debrecen* 6 (1959), pp. 290–297.

[38] P. Erdős and M. Simonovits. “A limit theorem in graph theory”. *Studia Sci. Math. Hungar* 1 (1966), pp. 51–57.

[39] P. Erdős and G. Szekeres. “A combinatorial problem in geometry”. *Compositio Math.* 2 (1935), pp. 463–470.

[40] P. Erdös and P. Turán. “On Some Sequences of Integers”. *J. London Math. Soc.* S1-11.4 (1936), p. 261.

[41] J. Folkman. “Graphs with monochromatic complete subgraphs in every edge coloring.” *SIAM J. Appl. Math.* 18 (1970), pp. 19–24.

[42] P. Frankl, R. L. Graham, and V. Rödl. "Quantitative theorems for regular systems of equations". *J. Combin. Theory Ser. A* 47.2 (1988), pp. 246–261.

[43] P. Frankl and V. Rödl. "Large triangle-free subgraphs in graphs without K_4". *Graphs Combin.* 2.2 (1986), pp. 135–144.

[44] E. Friedgut. "Sharp thresholds of graph properties, and the k-sat problem". *J. Amer. Math. Soc.* 12.4 (1999). With an appendix by Jean Bourgain, pp. 1017–1054.

[45] E. Friedgut, V. Rödl, A. Ruciński, and P. Tetali. "A sharp threshold for random graphs with a monochromatic triangle in every edge coloring". *Mem. Amer. Math. Soc.* 179.845 (2006), pp. vi+66.

[46] Z. Füredi. *A proof of the stability of extremal graphs: Simonovits' stability from Szemerédi's regularity*. 2010. URL: `http://www.renyi.hu/conferences/sze70/Talks/furedi.pdf`.

[47] W. T. Gowers. "Decompositions, approximate structure, transference, and the Hahn-Banach theorem". *Bull. Lond. Math. Soc.* 42.4 (2010), pp. 573–606.

[48] B. Green. "The Cameron-Erdős conjecture". *Bull. London Math. Soc.* 36.6 (2004), pp. 769–778.

[49] B. Green and I. Z. Ruzsa. "Sum-free sets in abelian groups". *Israel J. Math.* 147 (2005), pp. 157–188.

[50] B. Green and T. Tao. "The primes contain arbitrarily long arithmetic progressions". *Ann. of Math. (2)* 167.2 (2008), pp. 481–547.

[51] H. Hatami. "A structure theorem for Boolean functions with small total influences". *Ann. of Math. (2)* 176.1 (2012), pp. 509–533.

[52] P. E. Haxell, Y. Kohayakawa, and T. Łuczak. "Turán's extremal problem in random graphs: forbidding even cycles". *J. Combin. Theory Ser. B* 64.2 (1995), pp. 273–287.

[53] P. E. Haxell, Y. Kohayakawa, and T. Łuczak. "Turán's extremal problem in random graphs: forbidding odd cycles". *Combinatorica* 16.1 (1996), pp. 107–122.

[54] D. Kleitman. "A conjecture of Erdős-Katona on commensurable pairs among subsets of an n-set". *Theory of Graphs (Proc. Colloq., Tihany, 1966)*. Academic Press, New York, 1968, pp. 215–218.

[55] Y. Kohayakawa. "Szemerédi's regularity lemma for sparse graphs". *Foundations of computational mathematics (Rio de Janeiro, 1997)*. Springer, Berlin, 1997, pp. 216–230.

[56] Y. Kohayakawa and B. Kreuter. "The width of random subsets of Boolean lattices". *J. Combin. Theory Ser. A* 100.2 (2002), pp. 376–386.

[57] Y. Kohayakawa, B. Kreuter, and D. Osthus. "The length of random subsets of Boolean lattices". *Random Structures Algorithms* 16.2 (2000), pp. 177–194.

[58] Y. Kohayakawa, T. Łuczak, and V. Rödl. "On K^4-free subgraphs of random graphs". *Combinatorica* 17.2 (1997), pp. 173–213.

[59] Y. Kohayakawa, T. Łuczak, and V. Rödl. "Arithmetic progressions of length three in subsets of a random set". *Acta Arith.* 75.2 (1996), pp. 133–163.

[60] P. G. Kolaitis, H. J. Prömel, and B. L. Rothschild. "Asymptotic enumeration and a 0-1 law for m-clique free graphs". *Bull. Amer. Math. Soc. (N.S.)* 13.2 (1985), pp. 160–162.

[61] P. G. Kolaitis, H. J. Prömel, and B. L. Rothschild. "K_{l+1}-free graphs: asymptotic structure and a 0-1 law". *Trans. Amer. Math. Soc.* 303.2 (1987), pp. 637–671.

[62] B. Kreuter. "Small sublattices in random subsets of Boolean lattices". *Proceedings of the Eighth International Conference "Random Structures and Algorithms" (Poznan, 1997)*. Vol. 13. 3-4. 1998, pp. 383–407.

[63] V. F. Lev, T. Łuczak, and T. Schoen. "Sum-free sets in abelian groups". *Israel J. Math.* 125 (2001), pp. 347–367.

[64] L. Lovász and M. Simonovits. "On the number of complete subgraphs of a graph. II". *Studies in pure mathematics*. Birkhäuser, Basel, 1983, pp. 459–495.

[65] D. Lubell. "A short proof of Sperner's lemma". *J. Combinatorial Theory* 1 (1966), p. 299.

[66] T. Łuczak, A. Ruciński, and B. Voigt. "Ramsey properties of random graphs". *J. Combin. Theory Ser. B* 56.1 (1992), pp. 55–68.

[67] W. Mantel. "Problem 28". *Wiskundige Opgaven* 10 (1907), pp. 60–61.

[68] L. D. Mešalkin. "A generalization of Sperner's theorem on the number of subsets of a finite set". *Teor. Verojatnost. i Primenen* 8 (1963), pp. 219–220.

[69] R. Morris and D. Saxton. "The number of $C_{2\ell}$-free graphs". Preprint (2013). arXiv: `1309.2927 [math.CO]`.

[70] F. Mousset, R. Nenadov, and A. Steger. "On the number of graphs without large cliques". Preprint (2013). arXiv: `1312.1143 [math.CO]`.

[71] D. Osthus. "Maximum antichains in random subsets of a finite set". *J. Combin. Theory Ser. A* 90.2 (2000), pp. 336–346.

[72] R. Rado. "Studien zur Kombinatorik". *Math. Z.* 36.1 (1933), pp. 424–470.

[73] A. Rényi. "On random subsets of a finite set". *Mathematica (Cluj)* 3 (1961), pp. 355–362.

[74] V. Rödl and A. Ruciński. "Random graphs with monochromatic triangles in every edge coloring". *Random Structures Algorithms* 5.2 (1994), pp. 253–270.

[75] V. Rödl and A. Ruciński. "Threshold functions for Ramsey properties". *J. Amer. Math. Soc.* 8.4 (1995), pp. 917–942.

[76] V. Rödl and M. Schacht. "Extremal results in random graphs". *Erdös centennial*. Vol. 25. Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest, 2013, pp. 535–583.

[77] K. F. Roth. "On certain sets of integers". *J. London Math. Soc.* 28 (1953), pp. 104–109.

[78] W. Samotij. "Stability results for random discrete structures". *Random Structures Algorithms* 44.3 (2014), pp. 269–289.

[79] A. A. Sapozhenko. "The Cameron-Erdős conjecture". *Dokl. Akad. Nauk* 393.6 (2003), pp. 749–752.

[80] A. A. Sapozhenko. "The Cameron-Erdős conjecture". *Discrete Math.* 308.19 (2008), pp. 4361–4369.

[81] D. Saxton and A. Thomason. "Hypergraph containers". Preprint (2012). arXiv: `1204.6595 [math.CO]`.

[82] M. Schacht. "Extremal results for random discrete structures" (2009). Submitted.

[83] I. Schur. "Über die Kongruenz $x^m + y^m \equiv z^m \pmod{p}$". *Jahresber. Dtsch. Math.-Ver.* 25 (1917), pp. 114–117.

[84] M. Simonovits. "A method for solving extremal problems in graph theory, stability problems". *Theory of Graphs (Proc. Colloq., Tihany, 1966)*. Academic Press, New York, 1968, pp. 279–319.

[85] A. Soifer. *The mathematical coloring book*. Mathematics of coloring and the colorful life of its creators, With forewords by Branko Grünbaum, Peter D. Johnson, Jr. and Cecil Rousseau. Springer, New York, 2009, pp. xxx+607.

[86] E. Sperner. "Ein Satz über Untermengen einer endlichen Menge". *Math. Z.* 27.1 (1928), pp. 544–548.

[87] E. Szemerédi. "On sets of integers containing no k elements in arithmetic progression". *Acta Arith.* 27 (1975). Collection of articles in memory of Juriĭ Vladimirovič Linnik, pp. 199–245.

[88] P. Turán. "Egy gráfelméleti szélsőérték feladatról". *Mat. Fiz. Lapok* 48 (1941), pp. 436–452.

[89] P. Varnavides. "On certain sets of positive density". *J. London Math. Soc.* 34 (1959), pp. 358–360.

[90] B. L. van der Waerden. "Beweis einer Baudetschen Vermutung". *Nieuw Arch. Wisk.* 15 (1927), pp. 212–216.

[91] B. L. van der Waerden. "How the proof of Baudet's conjecture was found". *Studies in Pure Mathematics (Presented to Richard Rado)*. Academic Press, London, 1971, pp. 251–260.

[92] L. Warnke. "On the method of typical bounded differences". Preprint (2012). arXiv: `1212.5796 [math.CO]`.

[93] K. Yamamoto. "Logarithmic order of free distributive lattice". *J. Math. Soc. Japan* 6 (1954), pp. 343–353.