ON SOME PROBLEMS IN EXTREMAL, PROBABILISTIC AND ENUMERATIVE COMBINATORICS

BY

ZSOLT ADAM WAGNER

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Mathematics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2018

Urbana, Illinois

Doctoral Committee:

Professor Alexandr V. Kostochka, Chair
Professor József Balogh, Director of Research
Assistant Professor Anush Tserunyan
Dr. Mikhail Lavrov

Abstract

This is a study of a small selection of problems from various areas of Combinatorics and Graph Theory, a fast developing field that provides a diverse spectrum of powerful tools with numerous applications to computer science, optimization theory and economics. In this thesis, we focus on extremal, probabilistic and enumerative problems in this field.

A central theorem in combinatorics is Sperner's Theorem, which determines the maximum size of a family $\mathcal{F} \subseteq \mathcal{P}(n)$ that does not contain a 2-chain $F_1 \subsetneq F_2$. Erdős later extended this result and determined the largest family not containing a k-chain $F_1 \subsetneq \ldots \subsetneq F_k$. Erdős and Katona and later Kleitman asked how many such chains must appear in families whose size is larger than the corresponding extremal result. In Chapter 2 we answer their question for all families of size at most $(1-\varepsilon)2^n$, provided n is sufficiently larger compared to k and ε.

The result of Chapter 2 is an example of a *supersaturation*, or *Erdős–Rademacher* type result, which seeks to answer how many forbidden objects must appear in a set whose size is larger than the corresponding result. These supersaturation results are a key ingredient to a very recently discovered proof method, called the Container method. Chapters 3 and 4 show various examples of this method in action. In Chapter 3 we, among others, give tight bounds on the logarithm of the number of t-error correcting codes and illustrate how the Container method can be used to prove random analoges of classical extremal results. In Chapter 4 we solve a conjecture of Burosch–Demetrovics–Katona–Kleitman–Sapozhenko about estimating the number of families in $\{0,1\}^n$ which do not contain two sets and their union.

In Chapter 5 we improve an old result of Erdős and Spencer. Folkman's theorem asserts that for each $k \in \mathbb{N}$, there exists a natural number $n = F(k)$ such that whenever the elements of $[n]$ are two-colored, there exists a set $A \subset [n]$ of size k with the property that all the sums of the form $\sum_{x \in B} x$, where B is a nonempty subset of A, are contained in $[n]$ and have the same color. In 1989, Erdős and Spencer showed that $F(k) \geq 2^{ck^2/\log k}$, where $c > 0$ is an absolute constant; here, we improve this bound significantly by showing that $F(k) \geq 2^{2^{k-1}/k}$ for all $k \in \mathbb{N}$.

Fox–Grinshpun–Pach showed that every 3-coloring of the complete graph on n vertices without a rainbow triangle contains a clique of size $\Omega\left(n^{1/3}\log^2 n\right)$ which uses at most two colors, and this bound is tight up to the constant factor. We show that if instead of looking for large cliques one only tries to find subgraphs of large chromatic number, one can do much better. In Chapter 6 we show, amongst others, that every such coloring contains a 2-colored subgraph with chromatic number at least $n^{2/3}$, and this is best possible. As a direct corollary of our result we obtain a generalisation of the celebrated theorem of Erdős-Szekeres, which states that any sequence of n numbers contains a monotone subsequence of length at least $\sqrt{n}$.

To my beloved parents,

Ilona Kocsi and Hartmut Wagner

Acknowledgments

This project would not have been possible without the support of many people. Many thanks to my advisor, József Balogh, who read my numerous revisions and helped make some sense of the confusion. Also thanks to my committee members, Alexandr V. Kostochka, Anush Tserunyan, and Mikhail Lavrov, who offered guidance and support. And finally, thanks to my parents and numerous friends who endured this long process with me, always offering support and love.

Table of Contents

List of Symbols

$\mathbb{N}$ .... the set of natural numbers starting with 1

$[n]$ .... $\{1, 2, \ldots, n\}$

$\mathcal{A} - \mathcal{B}$ .... set difference

Graphs

$V(G)$ .... the set of vertices of G

$E(G)$ .... the set of edges of G

$v(G)$ .... the number of vertices in G

$e(G)$ .... the number of edges in G

$E(X, Y)$ .... the set of edges between X and Y

$e(X, Y)$ .... the number of edges between X and Y

$N_G(v)$, $N(v)$ .... the set of neighbors of v in G

$d_G(v)$, $d(v)$ .... the degree of v in G

$\Delta(G)$ .... maximum degree of G

$K_{r,s}$ .... the complete bipartite graph with parts of size r and s

H-free .... not containing a (not nec. induced) copy of H

$G[X]$ .... graph induced by the vertex set X

$\mathrm{ex}(n, H)$ .... the maximum number of edges in an H-free n-vertex graph

$\mathrm{mis}(G)$ .... the number of maximal independent sets in G

$\chi(G)$ .... the chromatic number of G

$\chi'(G)$ .... the edge-chromatic number of G

Chapter 1

Introduction

Given an object H and a finite set $\mathcal{S}$, it is natural to ask the following questions:

1. How large is the largest subset of $\mathcal{S}$ that does not contain (an isomorphic copy of) H?

2. If the size of a subset is slightly larger than the above extremal threshold, at least how many copies of H must it contain? Given an integer M, amongst all subsets of $\mathcal{S}$ of size M, which subsets contain the fewest copies of H?

3. How many subsets of $\mathcal{S}$ do not contain a copy of H?

Answering Question 1 is the main goal of extremal combinatorics and extremal graph theory. Chapters 5 and 6 deal with two problems of this kind. Question 2 is often referred to as a *supersaturation* or *Erdős–Rademacher* type problem. It is often the case that as soon as the size of the set ever so slightly exceeds the extremal threshold for containing H, it suddenly must contain *many* copies of the forbidden object H. We address examples of this phenomenon in Chapters 2, 3, 4. Question 3 is the main question in enumerative combinatorics. The main tool we will explore in this thesis is the recently discovered powerful Container method, see Chapters 3 and 4.

1.1 Extremal problems in graph theory and combinatorics

1.1.1 Families without long chains

One of the cornerstone results of extremal combinatorics is Sperner's theorem [101]. The Boolean lattice is the poset $\mathcal{P}(n) = 2^{[n]}$ with the relation $A \leq B$ if $A \subseteq B$. Throughout the thesis we will often identify $2^{[n]}$ with $\{0,1\}^n$ by mapping every set to its characteristic vector. In a poset P we say that two elements $a, b \in P$ are *comparable* if either $a \leq b$ or $b \leq a$. An *antichain* in a poset P is a subset $C \subset P$ so that no two elements of C are comparable.

Theorem 1.1.1 (Sperner)**.** *The largest antichain in $\mathcal{P}(n)$ has size $\binom{n}{\lfloor n/2 \rfloor}$.*

Given a poset P, a set $C \subset P$ is a *chain* if any two of its elements are comparable. We call C a *k-chain* if $|C| = k$. So an antichain is a set that does not contain a 2-chain. Erdős showed [34] that Sperner's theorem can be generalized to bound the size of families without a k-chain.

Theorem 1.1.2 (Erdős)**.** *Let $k, n \in \mathbb{N}^+$. The size of the largest k-chain free family in $\mathcal{P}(n)$ is equal to the sum of the $k-1$ largest binomical coefficients of the form $\binom{n}{i}$.*

1.1.2 Ramsey theory

Ramsey theory is a branch of extremal combinatorics, that studies the conditions under which order must appear in disorder. Problems in Ramsey theory typically ask a question of the form: how large an ordered structure can we find in a large chaotic object? It begins with the innocuous looking observation that no matter how the edges of the complete graph on six vertices, denoted by K_6, are colored red and blue with each edge receiving one color, the resulting coloring must contain a monochromatic triangle. Ramsey's theorem [92], for two colors, states that instead of triangles we can hope to find arbitrarily large monochromatic cliques, provided the complete graph whose edges we color is large enough.

Theorem 1.1.3 (Ramsey)**.** *For every $n \in \mathbb{N}$ there exists an $N \in \mathbb{N}$ such that whenever the edges of K_N are two-colored, there exists a monochromatic K_n.*

We denote by $R(n)$ the smallest N for which Theorem 1.1.3 holds. The upper bound

$$R(n) \leq \binom{2n-2}{n-1} \leq 4^n$$

follows from a short pigeonhole argument due to Erdős and Szekeres [42]. The lower bound

$$\sqrt{2}^n \leq R(n)$$

is due to Erdős [35] and was instrumental in his introduction of the probabilistic method. Despite a considerable effort in the past seven decades, the two constants $\sqrt{2}$ and 4 in the lower and upper bounds were not improved. The current best bounds are due to Spencer [100] and Conlon [28].

Another classical theorem in the area is Van der Waerden's theorem [107], which states that whenever $[N]$ is r-colored, there is a monochromatic arithmetic progression of length k, provided N is sufficiently large compared to k and r.

Theorem 1.1.4 (Van der Waerden). *For every $r, k \in \mathbb{N}$ there exists an $N \in \mathbb{N}$ such that the following holds. Whenever $[N]$ is r-colored, there is a monochromatic arithmetic progression of length k.*

We will talk more about the different structures one can find in colorings of graphs and integers, and the bounds on the corresponding numbers, in Chapter 5.

1.2 The Container Method

The Container Method is a recently-developed powerful technique for bounding the number (and controlling the typical structure) of finite objects with forbidden substructures. This technique exploits a subtle clustering phenomenon exhibited by the independent sets of uniform hypergraphs, provided their edges are sufficiently evenly distributed. It provides a relatively small family of 'containers' for the independent sets, each of which contains few edges and hence are themselves close to being an independent set. This method was developed by Balogh–Morris–Samotij [8] and by Saxton–Thomason [98]. We defer the discussion of the graph container algorithm and the statement of (a version of) the hypergraph container lemma for 3-uniform hypergraphs to Chapters 3 and 4. We will not attempt to cover the extensive literature on this topic, instead we will cherry-pick some easy applications to illustrate the importance of the method and refer the reader to the excellent recent survey [9] for more details.

Dedekind suggested [31] the following problem:

Question 1.2.1 (Dedekind). *Let $\mathcal{S} = \{\mathcal{F} \subset \mathcal{P}(n) : \mathcal{F}$ is an antichain$\}$. How big is $|\mathcal{F}|$?*

The largest antichain in $\mathcal{P}(n)$ has size $\binom{n}{\lfloor n/2 \rfloor}$ and every subset of an antichain is itself an antichain. It follows that the number of antichains is at least $2^{\binom{n}{\lfloor n/2 \rfloor}}$. Kleitman [70] showed that this is not too far from the truth, by proving that the number of antichains is $2^{(1+o(1))\binom{n}{n/2}}$. The precise asymptotics were later found by Korshunov [75].

Obtaining Kleitman's result using the method of containers is an easy exercise. We first create a graph G on vertex set $\mathcal{P}(n)$ by connecting two sets by an edge if they form a comparable pair. Then Dedekind's problem is equivalent to counting the number of independent sets in G. Sperner's theorem implies that the largest independent set has size $\binom{n}{\lfloor n/2 \rfloor}$, hence the total number of independent sets is at least $2^{\binom{n}{\lfloor n/2 \rfloor}}$.

The next key ingredient is a supersaturation-, or Erdős–Rademacher type statement showing that any subset of $V(G)$ of size at least $(1+\epsilon)\binom{n}{\lfloor n/2 \rfloor}$ contains many more than $\varepsilon\binom{n}{\lfloor n/2 \rfloor}$ comparable pairs (which is the trivial lower bound). Fortunately we can apply another result of Kleitman [70] that essentially states that every such set contains at least n times more comparable pairs than the trivial bound.

The final step is to apply the Container Lemma to obtain a collection of *containers* $C_1, \dots, C_m \subset V(G)$ such that each independent set is a subset of some container, the containers have size at most $(1+o(1))\binom{n}{\lfloor n/2 \rfloor}$ and the number of containers is at most $2^{o\left(\binom{n}{\lfloor n/2 \rfloor}\right)}$. Putting everything together, the number of antichains in $\mathcal{P}(n)$ is at most the number of containers times the number of subsets of each container, which gives us the upper bound of $2^{(1+o(1))\binom{n}{\lfloor n/2 \rfloor}}$. For more details, we refer the reader to Chapter 3.

In order to give further motivation for the method, we present one more easy application. Let $\mathcal{H}$ be the hypergraph that encodes triangles in K_n. That is, the vertex set of $\mathcal{H}$ is $\binom{[n]}{2}$, i.e. the edge set of K_n. The edge set of $\mathcal{H}$ corresponds to triangles in K_n. So $\mathcal{H}$ is a 3-uniform hypergraph with $\binom{n}{3}$ edges and $\binom{n}{2}$ vertices. Directly applying the Hypergraph Container Lemma to $\mathcal{H}$ produces the following theorem, see e.g. [9].

Theorem 1.2.2. *For each $\varepsilon > 0$ there exists $C > 0$ such that the following holds. For each $n \in \mathbb{N}$ there exists a collection $\mathcal{G}$ of graphs on n vertices, with*

$$|\mathcal{G}| \leq n^{Cn^{3/2}},$$

such that

(a) each $G \in \mathcal{G}$ contains fewer than εn^3 triangles, and

(b) each triangle-free graph on n vertices is contained in some $G \in \mathcal{G}$.

In order to translate this container result to an enumerative result, we again need a supersaturation statement. An averaging argument shows that a graph with fewer than εn^3 triangles cannot have more than $(1+\varepsilon')n^2/4$ edges, for some ε' depending on ε. Hence the number N of triangle-free graphs on n vertices is at most

$$N \leq n^{O\left(n^{3/2}\right)} \cdot 2^{(1+o(1))n^2/4} = 2^{(1+o(1))n^2/4}.$$

This result is sharp up to the $o(1)$ factor, as any subset of a complete balanced bipartite graph is triangle-free. While this result is certainly not new (see [37]), it illustrates the versatility of this method. The survey [9] contains many more striking applications, in this introduction, and even in the later chapters of this thesis, we will only scratch the very surface of these results.

1.3 Supersaturation

A closely related problem to the ones studied is determining the minimum number of copies of a particular substructure in a combinatorial object of given size. These so-called *supersaturation* problems can be traced back to a result of Rademacher (see [36]) who showed that if a graph has more than $n^2/4$ edges (the maximum possible number of edges a triangle-free graph on n vertices can have) then it contains at least $\lfloor n/2 \rfloor$ triangles. For many more examples of supersaturations results and their applications in combinatorics we refer the reader to [91] and the references therein.

Many extremal results can be turned into counting results using the method of containers, we explore several examples of this in Chapters 3 and 4. The key intermediate step in all these proofs is a supersaturation result, which makes the study of these problems interesting in their own right. In Chapter 2 we investigate one particularly challenging supersaturation problem. It was Erdős and Katona, and later Kleitman, who asked at least how many k-chains must appear in a subset of $\mathcal{P}(n)$ of prescriped size. This problem was resolved by Kleitman in 1966 (see [70]), but in general the problem remained open until very recently Samotij provided [96] a full solution to the problem. We describe the history and our contribution to Kleitman's conjecture, which states that the optimal families are those which are obtained by taking sets as close as possible to the middle layer, in Chapter 2.

In Chapter 3 we explore supersaturation problems related to error-correcting codes, analoges of Kleitman's theorem, Katona's t-intersection theorem and related problems. We explore how to combine these supersaturation results with the container method to obtain enumerative and sparse random analoges of classical extremal results.

1.4 Organization of this thesis

Chapter 2 focusses on a supersaturation problem of Erdős, Katona and Kleitman. Chapter 3 focusses on some applications of the graph container method, where the difficulty in the proofs lies in the supersaturation lemma, and the application of the container method is fairly straightforward. Chapter 4 discusses a result where the main hurdle was formulating a new version of the container lemma. Chapters 5 and 6 discuss two results in Ramsey theory.

Chapter 2

Kleitman's conjecture and k-chains in the Boolean lattice

The results of this chapter are joint work with József Balogh [16].

A central theorem in combinatorics is Sperner's Theorem, which determines the maximum size of a family $\mathcal{F} \subseteq \mathcal{P}(n)$ that does not contain a 2-chain $F_1 \subsetneq F_2$. Erdős later extended this result and determined the largest family not containing a k-chain $F_1 \subsetneq \ldots \subsetneq F_k$. Erdős and Katona and later Kleitman asked how many such chains must appear in families whose size is larger than the corresponding extremal result.

This question was resolved for 2-chains by Kleitman in 1966, who showed that amongst families of size M in $\mathcal{P}(n)$, the number of 2-chains is minimized by a family whose sets are taken as close to the middle layer as possible. He also conjectured that the same conclusion should hold for all k, not just 2. The best result on this question is due to Das, Gan and Sudakov who showed that Kleitman's conjecture holds for families whose size is at most the size of the $k+1$ middle layers of $\mathcal{P}(n)$, provided $k \leq n-6$. Our main result is that for every fixed k and $\varepsilon > 0$, if n is sufficiently large then Kleitman's conjecture holds for families of size at most $(1-\varepsilon)2^n$, thereby establishing Kleitman's conjecture asymptotically. Our proof is based on ideas of Kleitman and Das, Gan and Sudakov. Several open problems are also given.

2.1 Introduction

Denote by $\Sigma(n,r)$ the size of the r largest layers in $\mathcal{P}(n)$, that is, $\Sigma(n,r) = \sum_{i=\lceil \frac{n-r+1}{2} \rceil}^{\lceil \frac{n+r-1}{2} \rceil} \binom{n}{i}$. Sperner's Theorem [101], a cornerstone result in extremal combinatorics from 1928, states that the size of the largest family $\mathcal{F} \subseteq \mathcal{P}(n)$ that does not contain two sets $F_1, F_2 \in \mathcal{F}$ with $F_1 \subsetneq F_2$ is $\binom{n}{\lfloor n/2 \rfloor}$. This result was extended by Erdős [34], who showed that the size of the largest family without a k-chain, that is, k sets $F_1 \subsetneq \ldots \subsetneq F_k$, is the sum of the $k-1$ largest binomial coefficients, $\Sigma(n, k-1)$.

The following natural question was first posed by Erdős and Katona and then extended by Kleitman some fifty years ago. Given a family $\mathcal{F}$ of s subsets of $[n]$, how many k-chains must $\mathcal{F}$ contain? We denote this minimum by $c_k(n,s)$, and determine it for a wide range of values of k and s. For $k=2$ this question was completely resolved by Kleitman [70]. We say that a family $\mathcal{F} \subseteq \mathcal{P}(n)$ is *centered* if for any two sets $A, B \subseteq [n]$ with $A \in \mathcal{F}$ and $B \notin \mathcal{F}$ we have that $|n/2 - |A|| \leq |n/2 - |B||$, and if $|n/2 - |A|| = |n/2 - |B||$ then we have $|A| \geq |B|$. That is, if $\mathcal{F}$ is constructed by "taking sets whose size is as close to $n/2$ as possible" (and if two layers have the same size we fill up the top one first). Equipped with this definition, Kleitman's theorem is as follows.

Theorem 2.1.1 (Kleitman [70])**.** *Let $n, M > 0$ be integers. Amongst families $\mathcal{F} \subseteq \mathcal{P}(n)$ of size M, the number of 2-chains in $\mathcal{F}$ is minimized by a centered family.*

Note that Theorem 2.1.1 does not claim that centered families are the *only* families achieving the minimum, which is not quite true (but close!). The families achieving minimum in Theorem 2.1.1 have been completely characterized by Das–Gan–Sudakov [30].

In the present chapter we are interested in what happens for $k > 2$. Kleitman conjectured that the conclusion of Theorem 2.1.1 should hold for $k > 2$ as well, that is, the number of k-chains in $\mathcal{F}$ is minimized if $\mathcal{F}$ is obtained by taking sets whose size is as close to $n/2$ as possible.

Conjecture 2.1.2 (Kleitman, [39, 70])**.** *Let $n, M > 0$ and $k \geq 2$ be integers. Amongst families $\mathcal{F} \subseteq \mathcal{P}(n)$ of size M, the number of k-chains in $\mathcal{F}$ is minimized by a centered family.*

Dove–Griggs–Kang–Sereni [32] and independently Das–Gan–Sudakov [30] proved that Kleitman's conjecture is true for families whose size is at most the size of the k middle layers. For a family $\mathcal{F} \subseteq \mathcal{P}(n)$, write $c_k(\mathcal{F})$ for the number of k-chains contained in $\mathcal{F}$.

Theorem 2.1.3 (Das–Gan–Sudakov [30], Dove–Griggs–Kang–Sereni [32])**.** *Let $k, M, n > 0$ with $M \leq \Sigma(n, k)$. Amongst families $\mathcal{F} \subseteq \mathcal{P}(n)$ of size M, the function $c_k(\mathcal{F})$ is minimized by centered families.*

The first set of authors obtained stability versions of the above theorem as well. The best result on Kleitman's conjecture prior to our result was due to Das–Gan-Sudakov [30], who showed that Kleitman's conjecture holds for family sizes at most the middle $k+1$ layers, provided $k \leq n - 6$.

Theorem 2.1.4 (Das–Gan–Sudakov)**.** *Let $n \geq 15$, $M \leq \Sigma(n, k+1)$ and $k \leq n - 6$. Amongst families $\mathcal{F} \subseteq \mathcal{P}(n)$ of size M, the function $c_k(\mathcal{F})$ is minimized by centered families.*

Once again they actually obtained slightly stronger results, providing stability results for families for which $c_k(\mathcal{F})$ is close to the minimum. For more on the history and motivation of this problem we refer the reader to the very well-written introduction of [30].

Our main result can be viewed as an asymptotic solution to Kleitman's conjecture. We show that Kleitman's conjecture is true for all $M \leq (1-\varepsilon)2^n$, provided n is sufficiently large compared to ε and k.

Theorem 2.1.5. *For every k and $\varepsilon > 0$ there exists an $n_0 = n_0(k,\varepsilon)$ such that if $n \geq n_0$ and $M \leq (1-\varepsilon)2^n$ then amongst families $\mathcal{F} \subset \mathcal{P}(n)$ of size M, the function $c_k(\mathcal{F})$ is minimized by centered families.*

Our proof consists of two main parts. First we show that amongst families that are contained in the middle roughly $\sqrt{n \log n}$ layers, centered families are the best (i.e. they have the smallest $c_k(\mathcal{F})$). This part is based on the symmetric chain decomposition construction of de Bruijn–Tengbergen–Kruyswijk [21] and ideas of Kleitman [69] and contains most of the new ideas of this chapter. The second part of the proof is then showing that an optimal family cannot contain sets that are too small or too large. Our method of proving this is mostly based on ideas of Das–Gan–Sudakov [30]. Throughout this chapter we make no effort to optimize the value of $n_0(k,\varepsilon)$. For the corresponding maximization question, i.e. determining the maximum possible number of comparable pairs amongst families of size M in $\mathcal{P}(n)$ we refer the reader to [3].

We note that very recently Samotij [96] used different methods to prove Conjecture 2.1.2 in its full generality:

Theorem 2.1.6 (Samotij [96])**.** *Kleitman's conjecture is true for all n, M, k.*

2.2 Set-up

Our main goal in this chapter is to prove Theorem 2.1.5. Hence throughout the chapter we consider k and $\varepsilon > 0$ to be fixed. We set n_0 to be sufficiently large so that all following inequalities hold and want to show that for any $n \geq n_0$ the conclusion of Theorem 2.1.5 holds. For that we fix an arbitrary $M \leq (1-\varepsilon)2^n$. By Theorem 2.1.4 we know that the conclusion of Theorem 2.1.5 holds if $M \leq \Sigma(n, k+1)$ hence we will always assume

$$M > \Sigma(n, k+1), \tag{2.2.1}$$

recalling that $\Sigma(n, s)$ is the total size of the s biggest layers in $\mathcal{P}(n)$. Let r be defined as the unique integer such that

$$\Sigma(n, r-1) < M \leq \Sigma(n, r).$$

That is, we wish to show that one of the optimal families will fully contain the $r-1$ middle layers, and some elements from a neighboring layer. We observe that since for n large enough

$$\Sigma(n, \sqrt{n}\log^{1/20} n) > \left(1 - \frac{\varepsilon}{2}\right) 2^n,$$

we have

$$r \le \sqrt{n}\log^{1/10} n. \tag{2.2.2}$$

(We note that the conditions ensure that in fact $r \le C_\varepsilon \sqrt{n}$. We found it preferable to introduce the polylogarithmic term to avoid introducing constants depending on ε.) Moreover we will assume that

$$k \le \log^{1/100} n. \tag{2.2.3}$$

Throughout the chapter most propositions will aim to show that given certain conditions, centered families minimize the number of k-chains. Note that every centered family $\mathcal{F} \subset \mathcal{P}(n)$ of size M contains the same number of k-chains. It will be convenient for us to pick for each positive integer Q one specific centered family of size Q, that we will call $\mathcal{G}_Q$, and show that $\mathcal{G}_M$ minimizes the number of k-chains. Note that if $\mathcal{F}$ is centered then there exists at most one $j \in [n]$ such that $\emptyset \neq \mathcal{F} \cap \binom{[n]}{j} \neq \binom{[n]}{j}$, and we call this j the *partial layer* of $\mathcal{F}$ if it exists. Moreover if $Q > 0$ is a fixed integer then every centered family $\mathcal{F}$ of size Q in $\mathcal{P}(n)$ has the same partial layer j and the same intersection sizes with all layers. Given Q, n the only free choice one has when specifying a centered family of size Q in $\mathcal{P}(n)$ is what to do on the partial layer. A natural choice for $\mathcal{G}_Q$ is to choose an initial segment of the partial layer according to some total order on the elements of $\mathcal{P}(n)$. What ordering we pick makes absolutely no difference in the proof - but we believe it could be helpful for the reader to pick a specific total order. The *lexicographic order* $<_{lex}$ on $\mathcal{P}(n)$ is defined as follows. If $|A| < |B|$ then $A <_{lex} B$. Otherwise if $|A| = |B|$ then if the smallest element of $A \Delta B$ is in A then $A <_{lex} B$, otherwise $B \le_{lex} A$. For any positive integer Q let $\mathcal{G}_Q$ be the centered family of size Q in $\mathcal{P}(n)$ whose intersection with its partial layer j is an initial segment of the lexicographic ordering of $\binom{[n]}{j}$.

We will need to deal with families which are contained in a subset of $\mathcal{P}(n)$, for these it will be useful to extend the above definitions in a natural way. Given a family $\mathcal{P}' \subseteq \mathcal{P}(n)$, say that a family $\mathcal{F} \subseteq \mathcal{P}'$ *is centered in* $\mathcal{P}'$ if for any two sets $A, B \in \mathcal{P}'$ with $A \in \mathcal{F}$ and $B \notin \mathcal{F}$ we have that $|n/2 - |A|| \le |n/2 - |B||$, and if $|n/2 - |A|| = |n/2 - |B||$ then we have $|A| \ge |B|$. That is, if $\mathcal{F}$ is constructed by "taking sets whose size is as close to $n/2$ as possible in $\mathcal{P}'$" (and if two layers have the same size we fill up the top one first). A different way to say the same definition is as follows: a family $\mathcal{F} \subseteq \mathcal{P}'$ is centered in $\mathcal{P}'$ if it is an intersection of a centered family (in $\mathcal{P}(n)$) with $\mathcal{P}'$.

For a positive integer Q define $\mathcal{G}_{\mathcal{P}',Q}$ to be the family of size Q which is centered in $\mathcal{P}'$ and whose intersection with its partial layer is an initial segment of the restriction of $<_{lex}$ to $\mathcal{P}'$. So the family $\mathcal{G}_Q$ defined above equals $\mathcal{G}_{\mathcal{P}(n),Q}$.

A family $\mathcal{A} = \{A_1, \dots, A_\ell\} \subset \mathcal{P}(n)$ is a *chain* if $A_1 \subsetneq \dots \subsetneq A_\ell$. We say $\mathcal{A}$ is a chain with *step sizes* $a_1, \dots, a_{\ell-1}$ if $|A_{i+1} \setminus A_i| = a_i$ for all $i \in [\ell - 1]$. For a family $\mathcal{F} \subseteq \mathcal{P}(n)$ and integers $a_1, \dots, a_{k-1} \geq 1$, define

$$\Phi^*(\mathcal{F}, a_1, \dots, a_{k-1}) := \{(A_1, \dots, A_k) \in \mathcal{F}^k : A_1 \subsetneq \dots \subsetneq A_k, \text{ and } |A_{i+1} \setminus A_i| = a_i \text{ for all } i \in [k-1]\},$$

the set of k-chains with precisely these step sizes in $\mathcal{F}$. Given a k-chain $\mathcal{A} = \{A_1, \dots, A_k\}$ with $A_1 \subsetneq \dots \subsetneq A_k$, define

$$d(\mathcal{A}) := \max\{||A_k| - n/2|, ||A_1| - n/2|\}.$$

For every fixed $\mathbf{a} = (a_1, \dots, a_{k-1})$ we fix a total order $<^*_{\mathcal{F},\mathbf{a}}$ on $\Phi^*(\mathcal{F}, \mathbf{a})$ that satisfies the following property:

For every positive integer Q the family $\Phi^*(\mathcal{G}_{\mathcal{F},Q}, \mathbf{a})$ is an initial segment of the order $<^*_{\mathcal{F},\mathbf{a}}$.

Note that such an ordering $<^*_{\mathcal{F},\mathbf{a}}$ exists because $\mathcal{G}_{\mathcal{F},Q} \subsetneq \mathcal{G}_{\mathcal{F},Q+1}$ for all $0 \leq Q \leq |\mathcal{F}| - 1$.

Notation. Wherever possible we use standard notation and for the variable names we aim to follow the notation of [30]. There are two notational oddities that we feel we should mention. Firstly, for chains we use cursive capital letters, e.g. $\mathcal{A}, \mathcal{B}$, etc. - however, later in the chapter we will deal with hypergraphs on vertex set $\mathcal{P}(n)$ with edges corresponding to some chains, whence we will refer to the edges as e, f, etc. Several times we will, without mentioning this explicitly, make use of the natural correspondence between such edges and chains and hence occasionally label chains as e, f, etc. wherever this does not create confusion. Secondly, since we often consider the step sizes $a_1, \dots, a_{k-1}$ of a chain, for sake of brevity and cleanliness we will sometimes abbreviate this list as $\mathbf{a}$, with the understanding that $\mathbf{a} = (a_1, \dots, a_{k-1})$. We will always assume without mentioning it explicitly that the variable $\mathbf{a}$ refers to a list $(a_1, \dots, a_{k-1})$ of positive integers corresponding to the step sizes of some chain. Moreover, whenever the variable $\mathbf{a}$ and the list $a_1, \dots, a_{k-1}$ or $\{a_i\}_{i=1}^{k-1}$ are used in the same context they will refer to the same thing.

2.3 Families close to being centered

Set

$$d = \lfloor 10k\sqrt{n \log n} \rfloor$$

and write $\mathcal{P}_{n,d}$ for the union of the d middle layers in $\mathcal{P}(n)$, that is, for the family of sets $A \in \mathcal{P}(n)$ with $\lceil \frac{n-d+1}{2} \rceil \le |A| \le \lceil \frac{n+d-1}{2} \rceil$. Recall that we fixed an M at the very beginning of Section 2.2, which denotes the size of the families we will ultimately be interested in. Our goal in this section is to show the following proposition:

Proposition 2.3.1. *Amongst all families $\mathcal{F} \subseteq \mathcal{P}_{n,d}$ of size M, the number of k-chains in $\mathcal{F}$ is minimized when $\mathcal{F} = \mathcal{G}_M$.*

Once again we do not claim that $\mathcal{G}_M$ is the only family minimizing the number of k-chains. Once we have shown this proposition the only remaining step is to show that an optimal family cannot contain sets that are very far from the middle layer. This will be done later, in Section 2.4.

The proof of Proposition 2.3.1 uses the standard technique of compressions. Given a suboptimal family we show that we can apply some operations to it to make it better (in a sense defined later). One of the main ideas of the proof is that instead of moving the sets in the family (as in standard compression techniques), we view the family as a collection of chains and apply compression to the chains instead of directly to the family. One interesting aspect of this compression is that if we apply it to a family $\mathcal{F}$ we get an object that does not usually correspond to a family $\mathcal{F}' \subset \mathcal{P}(n)$ - instead the object we obtain after compressing a family will be a subgraph (equipped with a measure) of a weighted hypergraph, whose edges correspond to chains in $\mathcal{P}(n)$. In this hypergraph the induced subhypergraphs correspond to our usual families, but in order to make our compression methods work we have to leave the world of standard families and enter the realm of these more general objects (which we will refer to as *measured subhypergraphs*). Hence in order to prove Proposition 3.1 we will in fact show that amongst all such objects that have the same 'size' as our family $\mathcal{F}$, the ones corresponding to centered families cannot be improved by compressions and then deduce Proposition 2.3.1 from this.

2.3.1 Definitions

We say $\mathcal{A} = \{A_1, \ldots, A_\ell\}$ is a chain with *step sizes* $a_1, \ldots, a_{\ell-1}$ if $|A_{i+1} \setminus A_i| = a_i$ for every $i \in [\ell - 1]$. It has step sizes *at least* $a_1, \ldots, a_{\ell-1}$ if $|A_{i+1} \setminus A_i| \ge a_i$ for all $i \in [\ell - 1]$. The *height* of the chain $\mathcal{A}$ is defined as $h(\mathcal{A}) := |A_\ell \setminus A_1|$. It is called a *downward* chain if $||A_\ell| - n/2| \ge ||A_1| - n/2|$, otherwise we call it an *upward* chain. We call $\mathcal{A}$ a *skipless chain* if it is a chain and $|A_{i+1} \setminus A_i| = 1$ for all $i \in [\ell - 1]$. Moreover $\mathcal{A}$ is a *symmetric chain* if it is a skipless chain and $n/2 - |A_1| = |A_\ell| - n/2$. That is, a symmetric chain starts at some level s, ends at level $n - s$ and contains precisely one set from each level in between. A *symmetric chain decomposition* (SCD in short) of $\mathcal{P}(n)$ is a partition of $\mathcal{P}(n)$ into disjoint symmetric chains, whose union is the entire $\mathcal{P}(n)$. It is not a priori obvious that an SCD of $\mathcal{P}(n)$ should exist for all n - this was showed by de Bruijn–Tengbergen–Kruyswijk [21]. Note that as every symmetric chain intersects the middle layer in precisely one element, every SCD of $\mathcal{P}(n)$ consists of precisely $N := \binom{n}{\lfloor n/2 \rfloor}$ chains.

Let $\mathcal{A} \subset \mathcal{P}(n)$ be a chain and $\mathcal{X} = \{X_1, \ldots, X_N\}$ be an SCD of $\mathcal{P}(n)$. We say $\mathcal{X}$ *contains* $\mathcal{A}$ if there exists an $i \in [N]$ such that every set in $\mathcal{A}$ is contained in the chain X_i. Note that whenever $\pi \in S_n$ is a permutation of $[n]$ then π induces a permutation on the subsets of $[n]$ and if $\mathcal{X}$ is an SCD then so is $\pi(\mathcal{X})$. For a chain $\mathcal{A} \subset \mathcal{P}(n)$ we define its *weight* $\omega(\mathcal{A})$ to be the probability that $\mathcal{A}$ is contained in $\pi(\mathcal{X})$ where π is chosen uniformly at random from S_n, the set of permutations of $[n]$. This weight is independent of the choice of the SCD $\mathcal{X}$ and it can be calculated easily, as shown by the following proposition.

Proposition 2.3.2. *Let ℓ be an arbitrary positive integer and let $\mathcal{A} = \{A_1, \ldots, A_\ell\} \subset \mathcal{P}(n)$ be a chain with $A_1 \subsetneq \ldots \subsetneq A_\ell$. If $\mathcal{A}$ is a downward chain then*

$$\omega(\mathcal{A}) = \prod_{i=1}^{\ell-1} \binom{|A_{i+1}|}{|A_i|}^{-1} = \frac{|A_1|!}{|A_\ell|!} \prod_{i=1}^{\ell-1} |A_{i+1} \setminus A_i|!.$$

If $\mathcal{A}$ is an upward chain then

$$\omega(\mathcal{A}) = \prod_{i=1}^{\ell-1} \binom{n - |A_i|}{n - |A_{i+1}|}^{-1} = \frac{(n - |A_\ell|)!}{(n - |A_1|)!} \prod_{i=1}^{\ell-1} |A_{i+1} \setminus A_i|!.$$

We note that this weight function could have also been defined as follows. The weight of a chain $\mathcal{A} \subset \mathcal{P}(n)$ is the probability that it is contained in the SCD $\mathcal{X}$, where $\mathcal{X}$ has been chosen uniformly at random from the set of *all* SCDs of $\mathcal{P}(n)$ (rather than the set of permutations of one fixed $\mathcal{X}$). Both definitions work equally well for the purposes of this chapter, we chose to use the definition from the previous paragraph for it is easier to see why the explicit formula in Proposition 2.3.2 holds.

2.3.2 Properties of the weight function

There are two reasons for why we chose this probability for the weight $\omega(\mathcal{A})$ of a set. The first one is that it will imply that, under suitable conditions, if $\mathcal{A}, \mathcal{B}$ are two chains with $h(\mathcal{A}) < h(\mathcal{B})$ then we will have

$\omega(\mathcal{A}) \gg \omega(\mathcal{B})$. The second reason is that it will allow us to formulate a natural and best possible weighted supersaturation statement, essentially showing that centered families minimize the total weight of k-chains that they contain. The hard part will be to show that this implies that they also minimize the *number* of k-chains.

We start by proving the formulae claimed in the previous subsection.

Proof of Proposition 2.3.2. Let $\mathcal{A} = \{A_1, \ldots, A_\ell\} \subset \mathcal{P}(n)$ be a downward chain with $A_1 \subsetneq \ldots \subsetneq A_\ell$, the proof of the upward case is identical. Let $\mathcal{X}$ be a fixed SCD of $\mathcal{P}(n)$, let $\pi \in S_n$ be a permutation chosen uniformly at random from S_n and let X be the chain in $\pi(\mathcal{X})$ that contains A_ℓ. Since X is a symmetric chain and $\mathcal{A}$ is downward we have that for each $i \in [\ell]$, the chain X contains precisely one element of size $|A_i|$ (and possibly some others). Let E_i be the event that $A_i \in X$. Then

$$\begin{aligned}
P(E_1 \cap \ldots \cap E_\ell) &= P(E_\ell)P(E_1 \cap \ldots \cap E_{\ell-1}|E_\ell) \\
&= P(E_\ell)P(E_{\ell-1}|E_\ell)P(E_1 \cap \ldots \cap E_{\ell-2}|E_{\ell-1} \cap E_\ell) \\
&= P(E_\ell)P(E_{\ell-1}|E_\ell)P(E_1 \cap \ldots \cap E_{\ell-2}|E_{\ell-1}) \\
&= \ldots \\
&= P(E_\ell) \cdot P(E_{\ell-1}|E_\ell) \cdot P(E_{\ell-2}|E_{\ell-1}) \cdot \ldots \cdot P(E_1|E_2) \\
&= 1 \cdot \binom{|A_\ell|}{|A_{\ell-1}|}^{-1} \cdot \binom{|A_{\ell-1}|}{|A_{\ell-2}|}^{-1} \cdot \ldots \cdot \binom{|A_2|}{|A_1|}^{-1}.
\end{aligned}$$

$\square$

Note that if $\mathcal{A}, \mathcal{B}$ are two downward ℓ-chains with $|A_\ell| = |B_\ell|$ and they have the same step sizes (but possibly in a different order) then they have the same weight. Let us continue with the next claimed property of the weight function. Let $\mathcal{A}$ and $\mathcal{B}$ be two ℓ-chains with $h(\mathcal{A}) < h(\mathcal{B})$, recalling the definition that if $\mathcal{A}$ is an ℓ-chain then $h(\mathcal{A}) = |A_\ell \setminus A_1|$. Note that it is not always the case that $\omega(\mathcal{A}) > \omega(\mathcal{B})$ - for instance the chain $\{\emptyset, [n]\}$ has maximal weight ($= 1$) and maximal height. But if we avoid wandering too far off from the middle layer then our claim will hold. We emphasize that all the following propositions and lemmata are only valid if n is sufficiently large, as explained in Section 2.2.

Proposition 2.3.3. *Let $a_1, \ldots, a_{k-1}$ and $b_1, \ldots, b_{k-1}$ be positive integers such that $a_i \leq b_i$ for all $i \in [k]$ and strict inequality holds for at least one i. Suppose that $\sum_i a_i \leq \sqrt{n} \log^{2/5} n$ and let $\mathcal{A}, \mathcal{B}$ be k-chains in $\mathcal{P}_{n,d}$ with step sizes $\{a_i\}_{i=1}^{k-1}$ and $\{b_i\}_{i=1}^{k-1}$ respectively. Then $\omega(\mathcal{A}) \geq \omega(\mathcal{B}) n^{(h(\mathcal{B})-h(\mathcal{A}))/3}$.*

Proof. Without loss of generality we may assume that both chains $\mathcal{A}, \mathcal{B}$ are downward, the proof is similar if one (or both) of them is upward. Then by Proposition 2.3.2 we have

$$\omega(\mathcal{A}) = \frac{|A_1|! \prod_i a_i!}{|A_k|!}, \qquad \omega(\mathcal{B}) = \frac{|B_1|! \prod_i b_i!}{|B_k|!}.$$

Then we get (using the falling factorial notation $s_{(t)} = s(s-1) \cdot \ldots \cdot (s-t+1)$)

$$\begin{aligned}
\frac{\omega(\mathcal{B})}{\omega(\mathcal{A})} &= \frac{\prod_i b_i!}{\prod_j a_j!} \frac{|A_k|_{(h(\mathcal{A}))}}{|B_k|_{(h(\mathcal{B}))}} \leq d^{h(\mathcal{B})-h(\mathcal{A})} \left(\frac{|A_k|}{\frac{n}{2} - d} \right)^{h(\mathcal{A})} \left(\frac{1}{n/3} \right)^{h(\mathcal{B})-h(\mathcal{A})} \\
&\leq n^{-0.49(h(\mathcal{B})-h(\mathcal{A}))} \left(1 + \frac{30k\sqrt{n \log n}}{n/2} \right)^{\sqrt{n} \log^{2/5} n} \\
&\leq n^{-0.49(h(\mathcal{B})-h(\mathcal{A}))} e^{60 \log^{0.91} n} \leq n^{-(h(\mathcal{B})-h(\mathcal{A}))/3},
\end{aligned}$$

where in the first line we used that $b_i \leq d$ for all i and that $|B_1| \geq \frac{n}{2} - d \geq n/3$, in the second line we used that $|A_k| \leq \frac{n}{2} + d$, (2.2.3) and that $d = \lfloor 10k\sqrt{n \log n} \rfloor$, and in the last line we used (2.2.3). □

We further show that if two chains have the same step sizes then their weight decreases with their distance from the middle layer. Given a k-chain $\mathcal{A} = \{A_1, \ldots, A_k\}$ with $A_1 \subsetneq \ldots \subsetneq A_k$, recall the definition

$$d(\mathcal{A}) := \max\{||A_k| - n/2|, ||A_1| - n/2|\}.$$

Lemma 2.3.4. *Given positive integers $a_1, \ldots, a_{k-1}$, let $\mathcal{A}, \mathcal{B}$ be two k-chains in $\mathcal{P}(n)$ with step sizes $a_1, \ldots, a_{k-1}$, satisfying $d(\mathcal{A}) > d(\mathcal{B})$. Then $\omega(\mathcal{A}) < \omega(\mathcal{B})$ and in fact $\omega(\mathcal{B})/\omega(\mathcal{A}) \geq 1 + h(\mathcal{A})/n$.*

Proof. We assume that both chains are downward, the other cases are handled similarly. The weight of a chain $\mathcal{A}$ is given by

$$\omega(\mathcal{A}) = \frac{\prod a_i!}{|A_k|_{(h(\mathcal{A}))}}$$

which, if the a_i's and hence $h(\mathcal{A})$ are fixed, is a decreasing function of $|A_k|$. The ratio $\omega(\mathcal{B})/\omega(\mathcal{A})$ is bounded below by

$$\frac{\omega(\mathcal{B})}{\omega(\mathcal{A})} \geq \left(\frac{|A_k|}{|B_k|} \right)^{h(\mathcal{A})} \geq \left(\frac{|B_k| + 1}{|B_k|} \right)^{h(\mathcal{A})} \geq 1 + \frac{h(\mathcal{A})}{n}.$$

□

Lemma 2.3.5. *Let $a_1, \ldots, a_{k-1}$ and $b_1, \ldots, b_{k-1}$ be positive integers such that $a_i = b_i$ for all but one $i \in [k]$, and if j is the index where the two sequences differ then $b_j = a_j + 1$. Suppose that $\mathcal{A}, \mathcal{B}$ are k-chains in $\mathcal{P}_{n,d}$*

with step sizes $\{a_i\}_{i=1}^{k-1}$ *and* $\{b_i\}_{i=1}^{k-1}$ *respectively such that* $d(\mathcal{A}) \le \lceil (h(\mathcal{A})+1)/2 \rceil$*. Then* $\omega(\mathcal{A}) \ge \omega(\mathcal{B}) n^{1/3}$*.*

Proof. Without loss of generality we may assume that both chains $\mathcal{A}, \mathcal{B}$ are downward, the proof is similar if one (or both) of them is upward. Note that the condition $d(\mathcal{A}) \le \lceil (h(\mathcal{A})+1)/2 \rceil$ implies that $d(\mathcal{B}) \ge d(\mathcal{A})$ and so $|B_k| \ge |A_k|$. Then by Proposition 2.3.2 we have

$$\omega(\mathcal{A}) = \frac{|A_1|! \prod_i a_i!}{|A_k|!}, \qquad \omega(\mathcal{B}) = \frac{|B_1|! b_j \prod_i a_i!}{|B_k|!}.$$

Then, using that $b_j \le d = \lfloor 10k\sqrt{n \log n} \rfloor$, we have

$$\frac{\omega(\mathcal{B})}{\omega(\mathcal{A})} = \frac{|A_k|_{(h(\mathcal{A}))} b_j}{|B_k|_{(h(\mathcal{A})+1)}} \le \frac{b_j}{|B_k| - h(\mathcal{A})} \left(\frac{|A_k|}{|B_k|} \right)^{h(\mathcal{A})} \le \frac{20k\sqrt{n \log n}}{n/4} \cdot 1 \le n^{-1/3}.$$

□

Finally we prove in this subsection a weighted supersaturation result for families whose size exceeds $\Sigma(n, k-1)$. For a family $\mathcal{F} \subseteq \mathcal{P}(n)$ and integers $a_1, \dots, a_{k-1} \ge 1$, define

$$\Phi(\mathcal{F}, a_1, \dots, a_{k-1}) := \{(A_1, \dots, A_k) \in \mathcal{F}^k : A_1 \subsetneq \dots \subsetneq A_k, \text{ and } |A_{i+1} \setminus A_i| \ge a_i \text{ for all } i \in [k-1]\}.$$

Now let

$$\mathcal{W}_{a_1,\dots,a_{k-1}}(\mathcal{F}) := \sum_{(A_1,\dots,A_k) \in \Phi(\mathcal{F}, a_1, \dots, a_{k-1})} \omega(A_1, \dots, A_k).$$

Using these definitions we can state the promised supersaturation lemma.

Lemma 2.3.6. *Let* $Q, a_1, \dots, a_{k-1}$ *be positive integers and let* $\mathcal{F} \subset \mathcal{P}(n)$ *be a family of size* Q*. Then*

$$\mathcal{W}_{a_1,\dots,a_{k-1}}(\mathcal{F}) \ge \mathcal{W}_{a_1,\dots,a_{k-1}}(\mathcal{G}_Q).$$

Proof. Let $\mathcal{X}_1, \mathcal{X}_2$ be two arbitrary SCDs with chains $A_1, \dots, A_N$ and $B_1, \dots, B_N$ respectively, where $N = \binom{n}{\lfloor n/2 \rfloor}$, and consider the two multisets of integers $\mathcal{A} = \{|A_i \cap \mathcal{G}_Q| : i \in [N]\}$ and $\mathcal{B} = \{|B_i \cap \mathcal{G}_Q| : i \in [N]\}$. Then the two multisets are the same. Indeed, let A be an element of G_Q of maximum size and B of minimum size and assume that $|A| - n/2 > n/2 - |B|$ (other cases are similar). Then the number $|A| - |B| + 1$ appears $|G_Q \cap \binom{[n]}{|A|}|$ times in $\mathcal{A}$, the number $|A| - |B|$ appears $\binom{n}{|A|-1} - |G_Q \cap \binom{[n]}{|A|}|$ times, etc.

Let $f(p)$ be the least possible number of k-chains with step sizes at least $a_1, \dots, a_{k-1}$ contained in a chain of length p. Then $f(p)$ is exactly equal to the number of k-chains with step sizes at least $a_1, \dots, a_{k-1}$

contained in a *skipless* chain of length p. Note that then $f(p+1)-f(p)$ counts the number of k-chains with step sizes at least $a_1,\dots,a_{k-1}$ contained in a skipless chain of length $p+1$ that contain the bottom element of the skipless chain. Hence $f(p+1)-f(p) \geq f(p)-f(p-1)$ for all p and thus $f(p)$ is a convex function of p. Hence every SCD contains at least as many k-chains with step sizes at least $a_1,\dots,a_{k-1}$ from $\mathcal{F}$ as it does from $\mathcal{G}_Q$ where the intersection sizes with the chains are distributed as evenly as possible amongst the $\binom{n}{n/2}$ chains of the SCD, and all intersections are skipless chains. Here "as evenly as possible" means that there exists an integer ℓ such that if A_i is a chain in the SCD with $|A_i| < \ell$ then $A_i \subset \mathcal{G}_Q$ and if $|A_i| \geq \ell$ then $|A_i \cap \mathcal{G}_Q| \in \{\ell-1, \ell\}$.

Take a random SCD $\mathcal{X}$ and count the number of k-chains with step sizes at least $a_1,\dots,a_{k-1}$ in $\mathcal{F}$ that are contained in $\mathcal{X}$, call this number $x(\mathcal{F},\mathcal{X})$ and similarly define $x(\mathcal{G}_Q,\mathcal{X})$. Then by the above argument we had for every $\mathcal{X}$ that $x(\mathcal{F},\mathcal{X}) \geq x(\mathcal{G}_Q,\mathcal{X})$. Every k-chain is contained in $\mathcal{X}$ with probability equal to its weight. Taking expectations we have

$$\mathcal{W}_{a_1,\dots,a_{k-1}}(\mathcal{F}) = \mathbf{E}(x(\mathcal{F},\mathcal{X})) \geq \mathbf{E}(x(\mathcal{G}_Q,\mathcal{X})) = \mathcal{W}_{a_1,\dots,a_{k-1}}(\mathcal{G}_Q).$$

$\square$

To conclude this subsection we briefly indicate how Lemma 2.3.6 implies for example a special case of Theorem 2.1.1, stating that if a family $\mathcal{F}$ has $\binom{n}{\lfloor n/2\rfloor} + x \leq \Sigma(n,2)$ elements then it contains at least $x\lfloor 1+n/2\rfloor$ comparable pairs. Indeed if we set $Q = \binom{n}{\lfloor n/2\rfloor} + x$ then Lemma 2.3.6 states that $\mathcal{W}_1(\mathcal{F}) \geq x$. But every comparable pair except for the pair $\{\emptyset,[n]\}$ has weight at most $\lfloor 1+n/2\rfloor^{-1}$. Moreover the only comparable pairs of such maximum weight are the ones centered on the two middle layers, hence it is best to take such pairs greedily (i.e. take those pairs first which have the largest weight). Hence the result follows if we can show that e.g. an optimal family cannot contain the empty set.

The above paragraph illustrates some of the main ideas of the proof of the main result. We start with a collection of inequalities given to us by Lemma 2.3.6. We will claim that satisfying these inequalities greedily is the best one can do, assuming the optimal family cannot contain any small sets. The last step is then to show that this is indeed the case, i.e. if a family contains very small sets then it is bound to contain many more k-chains than $\mathcal{G}_M$.

2.3.3 Solving Kleitman's conjecture in $\mathcal{P}_{n,d}$

We are now ready to prove Proposition 2.3.1, in fact we will prove something more. We define the weighted hypergraph $\mathcal{H} = \mathcal{H}_{n,d,k}$ to be the k-uniform hypergraph on vertex set $V(\mathcal{H}) = \mathcal{P}_{n,d}$, edges corresponding to k-chains in $\mathcal{P}_{n,d}$ and the weight of an edge is given by the weight of the k-chain. A function $f : E(\mathcal{H}) \to [0,1]$ is called a *measured subhypergraph* of $\mathcal{H}$ and for an edge e we call $f(e)$ the measure of e.

Note that every family $\mathcal{F}$ corresponds to a measured subhypergraph $f_{\mathcal{F}}$ given by $f(e) = 1$ if the k-chain e is contained in $\mathcal{F}$, and $f(e) = 0$ otherwise. That is, $f_{\mathcal{F}}$ is the *characteristic function* corresponding to the k-chains in the family $\mathcal{F}$. We say that a measured subhypergraph f is *Q-good* if it satisfies the conclusion of Lemma 2.3.6, that is, if for all positive integers $a_1, \ldots, a_{k-1}$ we have

$$\sum_{(A_1,\ldots,A_k)\in\Phi(\mathcal{P}_{n,d},a_1,\ldots,a_{k-1})} \omega(A_1,\ldots,A_k) f(A_1,\ldots,A_k) \geq \mathcal{W}_{a_1,\ldots,a_{k-1}}(\mathcal{G}_Q).$$

Note that by Lemma 2.3.6 if $\mathcal{F}$ is a family of size at least M in $\mathcal{P}_{n,d}$ then the corresponding characteristic function $f_{\mathcal{F}}$ is M-good. The *size* of a measured subhypergraph f is defined as

$$|f| = \sum_{e\in E(\mathcal{H})} f(e).$$

Recall the definition of $\Phi^*(\mathcal{F}, \mathbf{a})$ and $<^*_{\mathcal{F},\mathbf{a}}$ from Section 2.2. For a family $\mathcal{F} \subseteq \mathcal{P}_{n,d}$, a measured subhypergraph f and a vector of positive integers $\mathbf{a} = (a_1, \ldots, a_{k-1})$ denote by $f_{\mathcal{F},\mathbf{a}}$ the restriction of f to the subhypergraph of $\mathcal{H}_{n,d,k}$ whose edges are the elements of $\Phi^*(\mathcal{F}, \mathbf{a})$. We say that f is $(\mathcal{F}, \mathbf{a})$-*compressed* if there is a chain $\mathcal{A} \in \Phi^*(\mathcal{F}, \mathbf{a})$ such that if $\mathcal{B} <^*_{\mathcal{F},\mathbf{a}} \mathcal{A}$ then $f(\mathcal{B}) = 1$ and if $\mathcal{A} <^*_{\mathcal{F},\mathbf{a}} \mathcal{B}$ then $f(\mathcal{B}) = 0$. Similarly define for a family $\mathcal{F} \subseteq \mathcal{P}_{n,d}$, a measured subhypergraph f and vector $\mathbf{a} = (a_1, \ldots, a_{k-1})$ the $(\mathcal{F}, \mathbf{a})$-compression of f, which is also a measured subhypergraph, denoted by $c[f, \mathcal{F}, \mathbf{a}]$, as follows.

- If $e \notin \Phi^*(\mathcal{F}, \mathbf{a})$ then $c[f, \mathcal{F}, \mathbf{a}](e) = f(e)$.
- $c[f, \mathcal{F}, \mathbf{a}]$ is $(\mathcal{F}, \mathbf{a})$-compressed.
- $|c[f, \mathcal{F}, \mathbf{a}]_{\mathcal{F},\mathbf{a}}| = |f_{\mathcal{F},\mathbf{a}}|$.

Observe that we always have $|f| = |c[f, \mathcal{F}, \mathbf{a}]|$, i.e. compression does not change the size of f. We say f is *completely compressed* if f is $(\mathcal{P}_{n,d}, \mathbf{a})$-compressed for every vector of positive integers $\mathbf{a} = (a_1, \ldots, a_{k-1})$.

Example. Let $n = 10$ and $k = 2$. Define the families $\mathcal{F}_1 = \binom{[n]}{4} \cup \binom{[n]}{6}$, $\mathcal{F}_2 = \mathcal{F}_1 \cup \binom{[n]}{7}$ and $\mathcal{F}_3 = \mathcal{F}_2 \cup \binom{[n]}{8}$. Let $\mathbf{a} = (2)$, i.e. we consider comparable pairs with set difference 2. Let $f_{\mathcal{F}_1}, f_{\mathcal{F}_2}, f_{\mathcal{F}_3}$ be the corresponding characteristic functions. Then $f_{\mathcal{F}_1}$ is $(\mathcal{P}(n), \mathbf{a})$-compressed, as the only comparable pairs with set difference 2 in $\mathcal{F}_1$ are those pairs closest possible to the middle layer, hence of largest weight. Moreover since all comparable pairs in $\mathcal{F}_1$ have set difference 2, we conclude that $f_{\mathcal{F}_1}$ is completely compressed. Since in $\mathcal{F}_2$ there are no new comparable pairs of set difference exactly 2, $f_{\mathcal{F}_2}$ is also $(\mathcal{P}(n), \mathbf{a})$-compressed. For $\mathbf{b} = (1)$ however, $f_{\mathcal{F}_2}$ is not $(\mathcal{P}(n), \mathbf{b})$-compressed, as $f_{\mathcal{F}_2}(\{123456, 1234567\}) = 1$ but e.g. $f_{\mathcal{F}_2}(\{1234, 12345\}) = 0$. Similarly $f_{\mathcal{F}_3}$ is not $(\mathcal{P}(n), \mathbf{a})$-compressed as $f_{\mathcal{F}_3}(\{123456, 12345678\}) = 1$ but $f_{\mathcal{F}_3}(12345, 1234567) = 0$. Note also that for every Q we have that the function $f_{\mathcal{G}_Q}$ corresponding to the centered family $\mathcal{G}_Q$ is completely compressed.

In the next proposition we will make use of the rearrangement inequality (see e.g. [62], Section 10.2, Theorem 368).

Proposition 2.3.7 (Rearrangement inequality)**.** *Given numbers $0 \leq x_1 \leq \ldots \leq x_m$ and $0 \leq y_1 \leq \ldots \leq y_m$ and a permutation $\pi \in S_m$ we have that*

$$\sum_i x_i y_i \geq \sum_i x_i y_{\pi(i)}.$$

Proposition 2.3.8. *Let $Q > 0$ be an integer, $\mathbf{a} = (a_1, \ldots, a_{k-1})$ be a vector of positive integers, f a Q-good measured subhypergraph and $\mathcal{F} \subset \mathcal{P}_{n,d}$. Then $c[f, \mathcal{F}, \mathbf{a}]$ is Q-good.*

Proof. We only need to prove that if we denote $\mathcal{X} := \Phi^*(\mathcal{F}, \mathbf{a})$ then

$$\sum_{e \in \mathcal{X}} c[f, \mathcal{F}, \mathbf{a}](e) \cdot \omega(e) \geq \sum_{e \in \mathcal{X}} f(e)\omega(e).$$

This follows from a simple property of the ordering $<^*_{\mathcal{F},\mathbf{a}}$: note that by the definition of $<^*_{\mathcal{F},\mathbf{a}}$ and Lemma 2.3.4 we have that if $\mathcal{A}, \mathcal{B}$ are two chains in $\mathcal{X}$ with $d(\mathcal{A}) > d(\mathcal{B})$ then $\mathcal{B} <^*_{\mathcal{F},\mathbf{a}} \mathcal{A}$. Hence by Lemma 2.3.4 $c[f, \mathcal{F}, \mathbf{a}]_{\mathcal{F},\mathbf{a}}$ greedily assigns measure 1 to the edges in $\mathcal{X}$ of largest weight until it has allocated a total measure equal to $|f_{\mathcal{F},\mathbf{a}}|$. Since the summation goes over $\mathcal{X}$ both functions in the above inequality can be replaced by their restrictions to $\mathcal{X}$ and then the claim follows from the rearrangement inequality (Proposition 2.3.7). □

Instead of proving Proposition 2.3.1 directly we will show the following stronger statement. As is often the case, the stronger statement will be easier and more natural to prove.

Proposition 2.3.9. *Amongst all M-good measured subhypergraphs, $f_{\mathcal{G}_M}$ has the smallest size.*

Proof. The collection of M-good measured subhypergraphs forms a closed subset of the compact set $[0,1]^{E(\mathcal{H})}$, so the restriction of $|\cdot|$ to this subset attains its minimum. Hence it suffices to show that for any M-good f

we have either $|f| = |f_{\mathcal{G}_M}|$ or we can find an M-good f' with $|f'| < |f|$. Recall that $\Phi^*(\mathcal{P}_{n,d}, a_1, \dots, a_{k-1})$ is defined to be the collection of all k-chains with step sizes precisely $a_1, \dots, a_{k-1}$ contained in $\mathcal{P}_{n,d}$. By Proposition 2.3.8 it suffices to consider those measured subhypergraphs which are completely compressed.

Let g be an M-good measured subhypergraph. For a list of positive integers $\mathbf{a} = (a_1, \dots, a_{k-1})$ write $g_{\mathbf{a}}$ for the restriction of g to the set $\Phi^*(\mathcal{P}_{n,d}, \mathbf{a})$, and similarly let $f_{\mathbf{a}}$ be the restriction of $f_{\mathcal{G}_M}$ to the same set. Let p be the smallest positive integer for which there exist positive integers $a_1, \dots, a_{k-1}$ with $\sum_i a_i = p$ such that $|g_{\mathbf{a}}| > |f_{\mathbf{a}}|$. We split into two cases according to whether such a p exists or not.

Case 1: If such a p exists then pick $a_1, \dots, a_{k-1}$ with $\sum_i a_i = p$ and $|g_{\mathbf{a}}| > |f_{\mathbf{a}}|$. Note that both $g_{\mathbf{a}}$ and $f_{\mathbf{a}}$ are $(\mathcal{P}_{n,d}, \mathbf{a})$-compressed: $g_{\mathbf{a}}$ is because as said before, by Proposition 2.3.8 it suffices to consider completely compressed measured subhypergraphs, and $f_{\mathbf{a}}$ is because of how we defined $<^*_{\mathcal{P}_{n,d},\mathbf{a}}$. Note that this implies that $g_{\mathbf{a}}(e) \geq f_{\mathbf{a}}(e)$ for all $e \in E(\mathcal{H})$, and there exists at least one $e^* \in E(\mathcal{H})$ such that $g_{\mathbf{a}}(e^*) > f_{\mathbf{a}}(e^*)$. Let $\varepsilon' := g(e^*) - f(e^*)$.

Define the following collection of $(k-1)$-sequences obtained from $a_1, \dots, a_{k-1}$ by decreasing one of the a_i's by one, assuming $a_i \neq 1$:

$$\mathcal{A}_{\mathbf{a}} := \{(a_1, a_2, \dots, a_{i-1}, a_i - 1, a_{i+1}, \dots, a_{k-1}) : i \in [k-1], a_i \geq 2\}.$$

Observe that by the choice of p, for every $\mathbf{b} = (b_1, \dots, b_{k-1}) \in \mathcal{A}_{\mathbf{a}}$ and for every $e \in E(\mathcal{H})$ that corresponds to a k-chain with step sizes exactly $b_1, \dots, b_{k-1}$ we have $g(e) \leq f_{\mathcal{G}_M}(e)$. This is because $|g_{\mathbf{b}}| \leq |f_{\mathbf{b}}|$ and both $g_{\mathbf{b}}$ and $f_{\mathbf{b}}$ are $(\mathcal{P}_{n,d}, \mathbf{a})$-compressed. Now for every $\mathbf{b} \in \mathcal{A}_{\mathbf{a}}$ pick an $e_{\mathbf{b}} \in E(\mathcal{H})$ of largest possible weight that corresponds to a k-chain with step sizes exactly $b_1, \dots, b_{k-1}$ and $g(e_{\mathbf{b}}) = 0$ and denote the collection of these at most $k-1$ edges by F. Choosing such edges is possible since $\mathcal{G}_M$ is contained in $\mathcal{P}_{n,r}$ and $r \ll d$.

Define a measured subhypergraph g' as follows.

$$g'(e) = \begin{cases} \frac{\varepsilon'}{2k} & : e \in F, \\ g(e) - \varepsilon' & : e = e^*, \\ g(e) & \text{otherwise.} \end{cases}$$

Observe that

$$|g'| = |g| - \varepsilon' + |F|\frac{\varepsilon'}{2k} \leq |g| - \varepsilon'/2 < |g|,$$

hence (recalling the first paragraph of this proof) it suffices to show that g' is M-good. Pick any positive integers $b_1, \ldots, b_{k-1}$, and we will show that

$$\sum_{(A_1,\ldots,A_k)\in\Phi(\mathcal{P}_{n,d},b_1,\ldots,b_{k-1})} \omega(A_1,\ldots,A_k)g'(A_1,\ldots,A_k) \geq \mathcal{W}_{b_1,\ldots,b_{k-1}}(\mathcal{G}_M). \tag{2.3.1}$$

If for some i we have $b_i > a_i$ then the changes we have made to g did not affect this inequality, and since g was M-good, (2.3.1) still holds for g'. If $b_i = a_i$ for all $i \in [k-1]$ then (2.3.1) holds by choice of ε'. Now suppose that there exists some $j \in [k-1]$ such that $b_j \leq a_j - 1$. Let $e_j \in F$ be the edge defined for the sequence $(a_1, \ldots, a_{j-1}, a_j - 1, a_{j+1}, \ldots, a_{k-1})$ above. If $h(e^*) \leq \sqrt{n}\log^{1/5} n$ then by Proposition 2.3.3 we have $\omega(e^*)n^{1/3} \leq \omega(e_j)$. If $h(e^*) \geq \sqrt{n}\log^{1/5} n$ then by (2.2.2) we may take the e_j's to have as small $d(e_j)$ as possible (and hence maximising their weight by Lemma 2.3.4) since none of the chains of height at least $\sqrt{n}\log^{1/5} n$ are present in $\mathcal{G}_M$, and by Lemma 2.3.5 we also have $\omega(e^*)n^{1/3} \leq \omega(e_j)$. So

$$\sum_{(A_1,\ldots,A_k)\in\Phi(\mathcal{P}_{n,d},b_1,\ldots,b_{k-1})} \omega(A_1,\ldots,A_k)\left(g'(A_1,\ldots,A_k) - g(A_1,\ldots,A_k)\right) \geq \frac{\varepsilon'}{2k}\omega(e_j) - \varepsilon'\omega(e^*) > 0.$$

Since g was M-good we conclude that g' also satisfies (2.3.1) and so is M-good. This completes the proof of the first case.

Case 2: For the second case we suppose such a p does not exist, i.e. for every list of positive integers $\mathbf{a} = (a_1, \ldots, a_{k-1})$ we have $|g_{\mathbf{a}}| \leq |f_{\mathbf{a}}|$. We claim that then $|g_{\mathbf{a}}| = |f_{\mathbf{a}}|$ for all sequences $a_1, \ldots, a_{k-1}$ and this will finish the proof as then $|g| = |f_{\mathcal{G}_M}|$. Suppose this is not true and let q be the largest positive integer such that there exists a list of integers $\mathbf{a} = (a_1, \ldots, a_{k-1})$ with $\sum_i a_i = q$ and $|g_{\mathbf{a}}| < |f_{\mathbf{a}}|$. Pick such an $\mathbf{a}$. Note that by the choice of q and since g is completely compressed we have that if $\mathbf{b} = (b_1, \ldots, b_{k-1})$ is a list such that $a_i \leq b_i$ for all $i \in [k-1]$ and e is any edge then $g_{\mathbf{b}}(e) = f_{\mathbf{b}}(e)$. Moreover since $|g_{\mathbf{a}}| < |f_{\mathbf{a}}|$ there exists an edge $e^* \in \Phi^*(\mathcal{P}_{n,d}, \mathbf{a})$ such that $g_{\mathbf{a}}(e^*) < f_{\mathbf{a}}(e^*)$. We have

$$\begin{aligned}
\sum_{\mathcal{A}\in\Phi(\mathcal{P}_{n,d},\mathbf{a})} \omega(\mathcal{A})g(\mathcal{A}) &= \sum_{\mathcal{A}\in\Phi^*(\mathcal{P}_{n,d},\mathbf{a})} \omega(\mathcal{A})g(\mathcal{A}) + \sum_{\mathbf{b}:\mathbf{b}>\mathbf{a}} \sum_{\mathcal{A}\in\Phi^*(\mathcal{P}_{n,d},\mathbf{b})} \omega(\mathcal{A})g(\mathcal{A}) \\
&= \sum_{\mathcal{A}\in\Phi^*(\mathcal{P}_{n,d},\mathbf{a})} \omega(\mathcal{A})g(\mathcal{A}) + \sum_{\mathbf{b}:\mathbf{b}>\mathbf{a}} \sum_{\mathcal{A}\in\Phi^*(\mathcal{P}_{n,d},\mathbf{b})} \omega(\mathcal{A})f_{\mathcal{G}_M}(\mathcal{A}) \\
&< \sum_{\mathcal{A}\in\Phi^*(\mathcal{P}_{n,d},\mathbf{a})} \omega(\mathcal{A})f_{\mathcal{G}_M}(\mathcal{A}) + \sum_{\mathbf{b}:\mathbf{b}>\mathbf{a}} \sum_{\mathcal{A}\in\Phi^*(\mathcal{P}_{n,d},\mathbf{b})} \omega(\mathcal{A})f_{\mathcal{G}_M}(\mathcal{A}) \\
&= \sum_{\mathcal{A}\in\Phi(\mathcal{P}_{n,d},\mathbf{a})} \omega(\mathcal{A})f_{\mathcal{G}_M}(\mathcal{A}) = \sum_{\mathcal{A}\in\Phi(\mathcal{G}_M,\mathbf{a})} \omega(\mathcal{A}) = \mathcal{W}_{\mathbf{a}}(\mathcal{G}_M).
\end{aligned}$$

Hence by Lemma 2.3.6 the measured subhypergraph g is not M-good, contradicting our assumptions. This completes the proof of Proposition 2.3.9. $\square$

Proof of Proposition 2.3.1. Let $\mathcal{F} \subseteq \mathcal{P}_{n,d}$ be a family of size M. Then $f_{\mathcal{F}}$ is M-good by Lemma 2.3.6, hence by Proposition 2.3.9 we have $|f_{\mathcal{G}_M}| \leq |f_{\mathcal{F}}|$, implying by definition that $\mathcal{F}$ contains at least as many k-chains as $\mathcal{G}_M$. $\square$

2.3.4 Non-centered families in $\mathcal{P}_{n,d}$

In the previous subsections we have shown that amongst families contained in $\mathcal{P}_{n,d}$, centered families are the best (i.e. given the size they minimize the number of k-chains). In the next section our goal will be to show that an optimal family cannot contain sets from outside of $\mathcal{P}_{n,d}$. For that we will make use of a proposition stating that if a family of size M is contained in $\mathcal{P}_{n,d}$, but misses some number of elements from the middle layers (and hence it is not centered) then this family contains significantly more k-chains than $\mathcal{G}_M$. This technique was used by Das–Gan–Sudakov [30] to prove Theorem 2.1.4.

Let $\mathcal{C} \subset \mathcal{P}_{n,r-1}$ be a family of size at most $\binom{n}{\lfloor (n+r)/2 \rfloor}$. Write $\mathcal{P}' := \mathcal{P}_{n,d} \setminus \mathcal{C}$ and say that a measured subhypergraph f is contained in $\mathcal{P}'$ if it assigns zero to every k-chain that intersects $\mathcal{C}$. Define the measured hypergraph $\hat{f}_{\mathcal{C},M}$, contained in $\mathcal{P}'$, as follows.

- $\sum_{e \in \Phi^*(\mathcal{P}',\mathbf{a})} \hat{f}_{\mathcal{C},M}(e)\omega(e) = \sum_{e \in \Phi^*(\mathcal{P}_{n,d},\mathbf{a})} f_{\mathcal{G}_M}(e)\omega(e)$ for all $\mathbf{a}$, and
- $\hat{f}_{\mathcal{C},M}$ is $(\mathcal{P}',\mathbf{a})$-compressed for all $\mathbf{a}$.

That is, $\hat{f}_{\mathcal{C},M}$ is obtained by greedily taking edges of largest possible weights, avoiding $\mathcal{C}$, to satisfy the definition of being M-good. Note that the first equality in the above definition of $\hat{f}_{\mathcal{C},M}$ can be satisfied because $r \ll d$, and that $\hat{f}_{\mathcal{C},M}$ is M-good by definition.

Proposition 2.3.10. *Let $0 \leq t \leq \binom{n}{\lfloor (n+r)/2 \rfloor}$ and let $\mathcal{C}$ be a family of size t contained in $\mathcal{P}_{n,d}$. If g is an M-good measured subhypergraph contained in $\mathcal{P}' = \mathcal{P}_{n,d} \setminus \mathcal{C}$ then $|g| \geq |\hat{f}_{\mathcal{C},M}|$.*

Proof. The proof of this proposition will be essentially the same as the proof of Proposition 2.3.9, therefore we only give a sketch. By Proposition 2.3.8 we may assume that g is $(\mathcal{P}',\mathbf{a})$-compressed for every list $\mathbf{a} = (a_1, \ldots, a_{k-1})$. For ease of notation, write $f := \hat{f}_{\mathcal{C},M}$ and as before, for a list of positive integers $\mathbf{a} = (a_1, \ldots, a_{k-1})$ write $g_{\mathbf{a}}$ for the restriction of g to the set $\Phi^*(\mathcal{P}',\mathbf{a})$, and similarly let $f_{\mathbf{a}}$ be the restriction of f to the same set. Let p be the smallest positive integer for which there exist positive integers $a_1, \ldots, a_{k-1}$ with $\sum_i a_i = p$ such that $|g_{\mathbf{a}}| > |f_{\mathbf{a}}|$. We split into two cases according to whether such a p exists or not. If such a p exists then we can find an M-good measured subhypergraph g' contained in $\mathcal{P}'$ with $|g'| < |g|$

the same way as we did in the proof of Proposition 2.3.9. If such a p does not exist then we may choose the largest positive integer q such that there exists a list of integers $\mathbf{a} = (a_1, \ldots, a_{k-1})$ with $\sum_i a_i = q$ and $|g_{\mathbf{a}}| < |f_{\mathbf{a}}|$. The existence of such q would show that g is not M-good and also result in a contradiction in the same fashion as in Proposition 2.3.9, hence we conclude that $|g_{\mathbf{a}}| = |f_{\mathbf{a}}|$ for all $\mathbf{a}$ and hence $|g| = |f|$. $\square$

2.4 Excluding very small and very large sets

In this section we show that an optimal family cannot contain sets from $\mathcal{P}(n) \setminus \mathcal{P}_{n,d}$. The main ideas in this section are similar to ideas in the work of Das–Gan–Sudakov [30]. For any j let $\mathcal{H}_{j,\ell}$ be the ℓ-uniform hypergraph with vertex set $V(\mathcal{H}_{j,\ell}) = \mathcal{P}_{n,j}$, and edges corresponding to ℓ-chains. Denote $\Delta_{j,\ell}$ the maximum degree of $\mathcal{H}_{j,\ell}$.

We continue our train of thought from the previous section with the following proposition:

Proposition 2.4.1. *Let $0 \leq t \leq \binom{n}{\lceil \frac{n+d-1}{2} \rceil}$ and let $\mathcal{C}$ be a family of t elements contained in $\mathcal{P}_{n,r-2}$. Let $s = \sum_{v \in \mathcal{C}} d(v, \mathcal{H}_{r-2,k})$ be the sum of the degrees of vertices in $\mathcal{C}$ in $\mathcal{H}_{r-2,k}$. If $\mathcal{F} \subset \mathcal{P}_{n,d} \setminus \mathcal{C}$ is a family of size M then $c_k(\mathcal{F}) \geq c_k(\mathcal{G}_M) + \frac{s}{kn}$.*

Proof. By Proposition 2.3.10 we have that $c_k(\mathcal{F}) \geq |\hat{f}_{\mathcal{C},M}|$. Since $c_k(\mathcal{G}_M) = |f_{\mathcal{G}_M}|$ it suffices to show that $|\hat{f}_{\mathcal{C},M}| - |f_{\mathcal{G}_M}| \geq \frac{s}{kn}$. Let E be the collection of k-chains contained in $\mathcal{P}_{n,r-2}$ that intersect $\mathcal{C}$. Note that every element $e \in E$ is present in $\mathcal{G}_M$ but missing from $\mathcal{F}$, and in fact we have $\hat{f}_{\mathcal{C},M}(e) = 0$ and $f_{\mathcal{G}_M}(e) = 1$. The idea of the proof is that since $\mathcal{P}_{n,r-1} \subseteq \mathcal{G}_M$, every $e \in E$ had to be replaced by edges of strictly smaller weight in $\hat{f}_{\mathcal{C},M}$. By Lemma 2.3.4 we will then have that

$$|\hat{f}_{\mathcal{C},M}| \geq |f_{\mathcal{G}_M}| + \frac{1}{n} \cdot |E|. \tag{2.4.1}$$

Since $|E| \geq s/k$ this will give the required result. To see why (2.4.1) holds, for a list $\mathbf{a} = (a_1, \ldots, a_{k-1})$ let $\hat{f}_{\mathbf{a}}$ be the restriction of $\hat{f}_{\mathcal{C},M}$ to the set $\Phi^*(\mathcal{P}_{n,d} \setminus \mathcal{C}, \mathbf{a})$, let $f_{\mathbf{a}}$ be the restriction of $f_{\mathcal{G}_M}$ to the set $\Phi^*(\mathcal{P}_{n,d}, \mathbf{a})$ and denote by $E_{\mathbf{a}}$ the set $E \cap \Phi^*(\mathcal{P}_{n,r-2}, \mathbf{a})$. Using the first point of the definition of $\hat{f}_{\mathcal{C},M}$ together with Lemma 2.3.4 gives

$$|\hat{f}_{\mathbf{a}}| \geq |f_{\mathbf{a}}| + \frac{1}{n}|E_{\mathbf{a}}|.$$

Summing up over all $\mathbf{a}$ gives (2.4.1). $\square$

Let A be a set in $\mathcal{P}_{n,j}$ for some $j \geq k$, and let v be the vertex corresponding to A in $\mathcal{H}_{j,k}$. We wish to estimate the degree $d(v, \mathcal{H}_{j,k})$ of v in $\mathcal{H}_{j,k}$. Denote the smallest and largest elements' sizes of $\mathcal{P}_{n,j}$ by p_-

and p_+, thats is, $p_- = \lceil \frac{n-j+1}{2} \rceil$ and $p_+ = \lceil \frac{n+j-1}{2} \rceil$. For $q \in [k]$ let

$$S_q = \{\mathbf{a} = (a_1, \ldots, a_{k-1}) : \; a_1 + \ldots + a_{q-1} \le |A| - p_-, \; a_q + \ldots + a_{k-1} \le p_+ - |A|\}.$$

Then

$$d(v, \mathcal{H}_{j,k}) = \sum_{q=1}^{k} \sum_{\mathbf{a} \in S_q} \frac{|A|_{(a_1+\ldots+a_{q-1})}(n-|A|)_{(a_q+\ldots+a_{k-1})}}{\prod a_i!}.$$

The largest term in the second sum occurs when the numerator has j terms and the denominator is as small as possible, i.e. when $\mathbf{a}$ is such that all $a_i \in \{\lfloor (j-1)/(k-1) \rfloor, \lceil (j-1)/(k-1) \rceil\}$ and $\sum a_i = j-1$. Let $\mathbf{a}^* = (a_1^*, \ldots, a_{k-1}^*)$ be such an $\mathbf{a}$. Since $|S_q| \le n^{k-1}$ we get

$$d(v, \mathcal{H}_{j,k}) \le k n^{k-1} \frac{|A|_{(|A|-p_-)}(n-|A|)_{(p_+-|A|)}}{\prod a_i^*!}.$$

This implies that

$$\frac{\lceil \frac{n+j-1}{2} \rceil_{(j-1)}}{\prod a_i^*!} \le \Delta_{j,k} \le n^k \frac{\lceil \frac{n+j-1}{2} \rceil_{(j-1)}}{\prod a_i^*!},$$

where the lower bound comes from simply counting the number of chains with step sizes precisely $\mathbf{a}^*$ containing a fixed set of size p_+. Suppose A is such that there exists an $\mathbf{a}$ and a $q \in [k-1]$ such that $a_1 + \ldots + a_{q-1} = |A| - p_-$ and $\sum a_i = j-1$ and all $a_i \in \{\lfloor (j-1)/(k-1) \rfloor, \lceil (j-1)/(k-1) \rceil\}$. Then for $j \le r$ we get for the corresponding v that

$$d(v, \mathcal{H}_{j,k}) \ge \frac{(n/2)_{(n/2-p_-)}(n/2)_{(p_+-n/2)}}{\prod a_i^*!} \ge \Delta_{j,k} n^{-k} \left(\frac{p_-}{p_+}\right)^{r/2} \ge \Delta_{j,k} n^{-k} \left(1 - \frac{r}{n/3}\right)^{r} \ge \Delta_{j,k} n^{-k-1}. \tag{2.4.2}$$

We now show that a small change in j does not change the degrees by much. Let $\mathbf{a}^{**}$ be such that all $a_i^{**} \in \{\lfloor j/(k-1) \rfloor, \lceil j/(k-1) \rceil\}$ and $\sum a_i^{**} = j$. Then

$$\Delta_{j,k} \ge \frac{\lceil \frac{n+j-1}{2} \rceil_{(j-1)}}{\prod a_i^*!} \ge \frac{\lceil \frac{n+j+1}{2} \rceil_{(j)}}{\prod a_i^{**}!} n^{-1} \ge \Delta_{j+1,k} n^{-k-1}. \tag{2.4.3}$$

Equipped with these bounds we are now ready to tackle the main result of this section.

Proposition 2.4.2. *If $\mathcal{F} \subset \mathcal{P}(n)$ is a family of size M with $\mathcal{F} \setminus \mathcal{P}_{n,d} \neq \emptyset$ then $c_k(\mathcal{F}) > c_k(\mathcal{G}_M)$.*

Proof. Let $M' := |\mathcal{F} \cap \mathcal{P}_{n,d}|$ and define r' such that $\Sigma(n, r'-1) < M' \le \Sigma(n, r')$. Set

$$b_+ = \left\lceil \frac{n+r'-1}{2} \right\rceil, \qquad b_- = \left\lceil \frac{n-r'+1}{2} \right\rceil, \qquad c_+ = \left\lceil \frac{n+d+1}{2} \right\rceil, \qquad c_- = \left\lceil \frac{n-d-1}{2} \right\rceil.$$

Note that $\mathcal{P}_{n,r'} = \{A \in \mathcal{P}(n) : b_- \le |A| \le b_+\}$ and $\mathcal{P}(n) \setminus \mathcal{P}(n,d) = \{A \in \mathcal{P}(n) : |A| \le c_- \text{ or } |A| \ge c_+\}$. As

$$\binom{n}{\le c_-} + \binom{n}{\ge c_+} \ll \binom{n}{b_+},$$

we have $r' \in \{r-1, r\}$. Recall that by (2.2.1) we have $r \ge k+2$ and so $r' \ge k+1$. We will assume throughout the first half of the proof that $r' \ge k+2$. The proof for the case $r' = k+1$ is very similar (in fact easier), but needs to be handled separately - we will do so later.

Let $\mathcal{S}$ be the family of those sets $A \in \mathcal{P}_{n,r'-2}$ for which there exists an $\mathbf{a} = (a_1, \dots, a_{k-1})$ satisfying $\sum a_i = r' - 3$ with $a_i \ge 1$ for all i, and there exists a $q \in [k]$ with $b_- + a_1 + \ldots + a_{q-1} = |A|$ and moreover $a_i \in \{\lfloor (r'-3)/(k-1) \rfloor, \lceil (r'-3)/(k-1) \rceil\}$. Note that $\mathcal{S}$ consists of at least k complete layers in $\mathcal{P}_{n,r'-2}$ (corresponding to splitting up the distance between b_- and b_+ into $k-1$ roughly equal pieces). Observe that we used the fact that $r' \ge k+2$ here.

Let $\mathcal{A} := \mathcal{P}_{n,r'-2} \setminus \mathcal{F}$. For $j \in I = \{0, \dots, c_-\} \cup \{c_+ \dots, n\}$ let $R_j = \mathcal{F} \cap \binom{[n]}{j}$ and let h_j denote the number of k-chains in $\mathcal{F}$ which contain an element of R_j and $k-1$ elements from $\mathcal{P}_{n,r'-2}$. Hence we have by Proposition 2.4.1 that

$$c_k(\mathcal{F}) \ge c_k(\mathcal{G}_{M'}) + \frac{\sum_{v \in \mathcal{A}} d(v, \mathcal{H}_{r'-2,k})}{kn} + \sum_{j \in I} h_j.$$

Note that $c_k(\mathcal{G}_M) - c_k(\mathcal{G}_{M'}) \le (M - M')\, \Delta_{r,k}$, so it suffices to show that

$$\frac{\sum_{v \in \mathcal{A}} d(v, \mathcal{H}_{r'-2,k})}{kn} + \sum_{j \in I} h_j > (M - M')\Delta_{r,k} = \sum_{j \in I} |R_j| \Delta_{r,k}. \tag{2.4.4}$$

W.l.o.g. we assume that $\sum_{j \le c_-} |R_j| \ge \sum_{j \ge c_+} |R_j|$, the proof otherwise is identical. From now on we always assume $j \in [c_-]$, the extra factor of 2 will be dominated by larger terms in our inequalities. Define β by

$$\beta \binom{n}{\le c_-} = \sum_{j \le c_-} |R_j|.$$

Now we split into two cases. For the first case assume that $|\mathcal{S} \setminus \mathcal{F}| \ge \beta \binom{n}{b_- + 1} / n^5$. Then by (2.4.2) and (2.4.3) we get

$$\frac{\sum_{v \in \mathcal{A}} d(v, \mathcal{H}_{r'-2,k})}{kn} \ge \beta \binom{n}{b_- + 1} \Delta_{r'-2,k} n^{-k-10} \ge \beta \binom{n}{b_- + 1} \Delta_{r,k} n^{-10k}.$$

Now note that
$$\binom{n}{\le c_-} \le n^{-10k^2}\binom{n}{b_-+1},$$
and hence (2.4.4) holds in this case.

Henceforth we assume $|\mathcal{S}\setminus\mathcal{F}| \le \beta\binom{n}{b_-+1}/n^5$. Let $\mathcal{T}$ be the family of those sets in $\binom{[n]}{b_-+1}$ which are not contained in any $(k-1)$-chains in $\mathcal{F}\cap\mathcal{P}_{n,r'-2}$. In other words, if $A\in\mathcal{T}$ then every $(k-1)$-chain in $\mathcal{S}$ containing A intersects $\mathcal{S}\setminus\mathcal{F}$. Recall that $\mathcal{S}$ contains at least k complete layers and let $\mathcal{S}'$ denote the bottom $k-1$ layers from $\mathcal{S}$, so that $\mathcal{S}'$ contains all sets of sizes $b_-+1 = s_1 < s_2 < \ldots < s_{k-1} \le b_+ - 1$. For all $i\in[k-1]$, write $\mathcal{Q}_i := (\mathcal{S}'\setminus\mathcal{F})\cap\binom{[n]}{s_i}$. Let $\mathcal{T}_1 := \mathcal{T}\setminus\mathcal{Q}_1$ and for $i\in[k-1]\setminus\{1\}$ define $\mathcal{T}_i := \partial(\mathcal{T}_{i-1},s_i)\setminus\mathcal{Q}_i$, where $\partial(\mathcal{T}_{i-1},s_i)$ denotes the family of sets $A\in\binom{[n]}{s_i}$ for which there exists a set $B\in\mathcal{T}_{i-1}$ such that $B\subset A$ (i.e. the upper shadow of $\mathcal{T}_{i-1}$ on level s_i). Since every $(k-1)$-chain in $\mathcal{S}'$ that intersects $\mathcal{T}$ has to intersect $\mathcal{S}'\setminus\mathcal{F}$, we conclude that $\mathcal{T}_{k-1}=\emptyset$. For all $i\in[k-1]$ define $q_i := |\mathcal{Q}_i|\binom{n}{s_i}^{-1}$ and similarly $t_i := |\mathcal{T}_i|\binom{n}{s_i}^{-1}$. By the normalized matching property[1] of the Boolean lattice we have the following inequalities:

- $t_i \le q_{i+1} + t_{i+1}$ for all $i\in[k-3]$, and
- $t_{k-2} \le q_{k-1}$.

By summing up all these inequalities we conclude that $t_1+q_1 \le q_1+q_2+\ldots+q_{k-1}$, which since $s_{k-1}\le b_+-1$ implies that $|\mathcal{T}| \le 3\left(|\mathcal{Q}_1|+|\mathcal{Q}_2|+\ldots+|\mathcal{Q}_{k-1}|\right) = 3|\mathcal{S}'\setminus\mathcal{F}| \le 3|\mathcal{S}\setminus\mathcal{F}| \le \beta\binom{n}{b_-+1}/n^4$.

Using the definition of $\mathcal{T}$ we now have the bound
$$h_j \ge |R_j|\binom{n-j}{b_-+1-j} - |\mathcal{T}|\binom{b_-+1}{j}. \tag{2.4.5}$$

For $j\in[c_-]$ define β_j by $|R_j| = \beta_j\binom{n}{j}$. Using $\binom{n}{b_-+1}\binom{b_-+1}{j} = \binom{n}{j}\binom{n-j}{b_-+1-j}$ and that $h_j\ge 0$ we get
$$\begin{aligned} h_j &\ge \max\left\{0, \beta_j\binom{n}{j}\binom{n-j}{b_-+1-j} - \frac{\beta}{n^4}\binom{n}{j}\binom{n-j}{b_-+1-j}\right\} \\ &\ge \max\left\{0, \beta_j\binom{n}{j}\binom{n-c_-}{b_-+1-c_-} - \frac{\beta}{n^4}\binom{n}{j}\binom{n-c_-}{b_-+1-c_-}\right\}. \end{aligned}$$

Since $\sum_{j\le c_-}\beta_j\binom{n}{j} = \beta\binom{n}{\le c_-}$ we have
$$\sum_{j\le c_-} h_j \ge \binom{n-c_-}{b_-+1-c_-}\left(\sum_{j\le c_-}\binom{n}{j}\beta_j - \sum_{j\le c_-}\binom{n}{j}\frac{\beta}{n^4}\right) \ge \frac{1}{4}\binom{n-c_-}{b_-+1-c_-}\beta\binom{n}{\le c_-}.$$

[1]In our context this means that (for all p), whenever $\mathcal{C}\subseteq\binom{[n]}{p}$ and $\mathcal{C}^*$ is the subset of $\binom{[n]}{p+1}$ consisting of the elements of $\binom{[n]}{p+1}$ covering elements of $\mathcal{C}$, it holds that $|\mathcal{C}|\binom{n}{p}^{-1} \le |\mathcal{C}^*|\binom{n}{p+1}^{-1}$.

To complete the proof it only remains to show that $\binom{n-c_-}{b_-+1-c_-} \gg \Delta_{r,k}$, as then (2.4.4) holds. Note that $\Delta_{r,k} \leq n^{k+r}$ - indeed, there are at most n^k ways to choose the sizes of the k sets in a k-chain, and there are at most n^r distinct r-chains through a fixed set in $\mathcal{P}_{n,r}$. Moreover we have

$$\binom{n-c_-}{b_-+1-c_-} \geq \binom{n/2}{4k\sqrt{n\log n}} \geq n^{k\sqrt{n\log n}} \geq n^{2(k+r)},$$

and the proof of the case $r' \geq k+2$ is complete.

All that is missing now is the case $r' = k+1$ – fortunately when $r' = k+1$ we can directly apply the results of Das–Gan–Sudakov [30]. Recall that $M' = |\mathcal{F}'|$.

Theorem 2.4.3 (Corollary of Theorem 4.2 of [30])**.** *Let $\mathcal{F}' \subset \mathcal{P}_{n,d}$ be a family of size $\Sigma(n,k) \leq |\mathcal{F}'| \leq \Sigma(n,k+1)$ with at least t sets missing from the middle $k-1$ levels. Then*

$$c_k(\mathcal{F}') \geq c_k(\mathcal{G}_{M'}) + \frac{t}{n}\Delta_{k+1,k}.$$

Now we continue with the proof of Proposition 2.4.2 for the case $r' = k+1$. We follow the notation of the first half of the proof of Proposition 2.4.2. W.l.o.g. assume that $\sum_{j\leq c_-}|R_j| \leq \sum_{j\geq c_+}|R_j|$, the proof otherwise is identical to the proof of the case $r' \geq k+2$. As before, define h_j to be the number of k-chains in $\mathcal{F}$ which contain an element of R_j and $k-1$ elements from $\mathcal{P}_{n,k-1}$. Setting $\mathcal{F}' := \mathcal{F} \cap \mathcal{P}_{n,d}$ and $t := |\mathcal{P}_{n,k-1} \setminus \mathcal{F}'|$ and applying Theorem 2.4.3 we get that

$$c_k(\mathcal{F}) \geq c_k(\mathcal{G}_{M'}) + \frac{t}{n}\Delta_{k+1,k} + \sum_{j\geq c_+} h_j,$$

and hence as before it suffices to show that

$$\frac{t}{n}\Delta_{k+1,k} + \sum_{j\geq c_+} h_j \geq (M - M')\,\Delta_{k+1,k}.$$

If $t \geq n\,(M-M')$ then this inequality holds as each h_j is non-negative, hence we may assume $t \leq n\,(M-M') \leq 2n\sum_{j\geq c_+}|R_j|$. In this case we will in fact show that

$$\sum_{j\geq c_+} h_j \geq (M - M')\,\Delta_{k+1,k}.$$

Following the notation of [30], let $a = \lceil \frac{n+k}{2} \rceil$ so that the $k-1$ middle levels are those sets of sizes between

$a-k+1$ and $a-1$. As before in (2.4.5), for $j \geq c_+$ we have the lower bound

$$h_j \geq \max\left\{|R_j|\binom{j}{a-1} - t\binom{n-a+1}{j-a+1}, 0\right\}\binom{a-1}{k-2}(k-2)!.$$

Now observe that for $j \geq c_+$ we have

$$\binom{j}{a-1}\binom{n-a+1}{j-a+1}^{-1} = \frac{j!(n-j)!}{(a-1)!(n-(a-1))!} \geq \left(1+\frac{10k\sqrt{n\log n}}{n}\right)^{4k\sqrt{n\log n}} \geq n^{20k^2}.$$

Hence it suffices to show

$$\sum_{j\geq c_+} \max\left\{|R_j| - \frac{\sum_{i\geq c_+}|R_i|}{n^{19k^2}}, 0\right\}\binom{j}{a-1}\binom{a-1}{k-2}(k-2)! \geq 2\Delta_{k+1,k}\sum_{j\geq c_+}|R_j|.$$

Now since

$$\Delta_{k+1,k} = \left(\binom{a}{k-1} + \binom{a}{k}\binom{k}{2}\right)(k-1)! \leq n^5\binom{a-1}{k-2}(k-2)!,$$

and since for every $j \geq c_+$ we have $\binom{j}{a-1} \geq n^{10}$, it is enough to show

$$n^4\sum_{j\geq c_+}\left(|R_j| - \frac{\sum_{i\geq c_+}|R_i|}{n^{19k^2}}\right) \geq \sum_{j\geq c_+}|R_j|.$$

However the left hand side is at least

$$n^4\sum_{j\geq c_+}\left(|R_j| - \frac{\sum_{i\geq c_+}|R_i|}{n^{19k^2}}\right) \geq n^4\sum_{j\geq c_+}|R_j| - \frac{n^5}{n^{19k^2}}\sum_{j\geq c_+}|R_j| \gg \sum_{j\geq c_+}|R_j|,$$

and the proof is complete. □

2.5 Proof of Theorem 2.1.5

Let $\mathcal{F}$ be a family of size M. If $\mathcal{F} \not\subset \mathcal{P}_{n,d}$ then by Proposition 2.4.2 we have $c_k(\mathcal{F}) > c_k(\mathcal{G}_M)$. If on the other hand $\mathcal{F} \subseteq \mathcal{P}_{n,d}$ then by Proposition 2.3.1 we have $c_k(\mathcal{F}) \geq c_k(\mathcal{G}_M)$. Hence $\mathcal{G}_M$ minimizes the number of contained k-chains amongst families of size M in $\mathcal{P}(n)$, and the proof is complete.

2.6 Open problems

One direction that might be of interest is to extend the question of minimizing the number of k-chains to other posets, hence generalizing Kleitman's question. Instead of considering families in $\mathcal{P}(n) = \{0,1\}^n$ one could ask the same questions for $[m]^2$ or even $[m]^d$. A k-chain in $[m]^d$ is a set of k distinct points satisfying $\mathbf{a}_1 \leq \ldots \leq \mathbf{a}_k$ (where $\mathbf{b} \leq \mathbf{c}$ means $b_i \leq c_i$ for all $i \in [d]$). Solving the following problem in full generality seems hopeless, but partial results for larger m would be of much interest. Is a similar phenomenon as in Kleitman's conjecture likely to hold for these posets as well? For related results, see [11, 91].

Problem 2.6.1. *Given d, m, M, k, which sets $\mathcal{F} \subseteq [m]^d$ of size $|\mathcal{F}| = M$ minimize the number of k-chains?*

Consider the following definition of an *m-centered* set: a set $\mathcal{F} \subseteq [m]^d$ is m-centered if for all $\mathbf{a}, \mathbf{b} \in [m]^d$ with $\mathbf{a} \in \mathcal{F}$ and $\mathbf{b} \notin \mathcal{F}$ we have that

$$\left|\sum_{i=1}^{d} a_i - \frac{dm}{2}\right| \leq \left|\sum_{i=1}^{d} b_i - \frac{dm}{2}\right|,$$

and in case of equality we have $\sum a_i \geq \sum b_i$. Note that taking $m = 2$ we get our usual definition of centered families. Once again we do not make the (false) claim that m-centered sets are the *only* ones minimizing the number of k-chains. One might be tempted to conjecture the following: given m there exists a number $d_0(m)$ such that if $d \geq d_0(m)$ then the answer to Problem 2.6.1 is given by m-centered sets. Note that if we do not assume d to be large enough then this natural conjecture might fail. One small counterexample is given by the case $m = 16, d = 2, k = 2$ where the family $\mathcal{F} := \{(a_1, a_2) \in [16]^2 : |a_1 + a_2 - 16| \leq 5\}$ can be improved by letting $\mathcal{F}' := \mathcal{F} \setminus \{(5,6)\} \cup \{(10,0)\}$. As it turns out this conjecture is false for large values of d as well, see [11].

Instead of the poset $\{0,1\}^n$ we can consider the poset $[0,1]^n$. Given a subset $\mathcal{F} \subseteq [0,1]^n$ let $\mathcal{C}(\mathcal{F}, k)$ be the collection of k-chains in $\mathcal{F}$ (where a k-chain, as before, is a set of k points satisfying $\mathbf{a}_1 \leq \ldots \leq \mathbf{a}_k$). Then $\mathcal{C}(\mathcal{F}, k)$ can be regarded as a subset of $\left([0,1]^n\right)^k$. This leads to the following natural question. By the *measure* of a set $A \subset \mathbf{R}^n$ we always refer to the Lebesgue measure (or n-volume) of A and denote it by $\lambda(A)$.

Problem 2.6.2. *Given n, M, k, which measurable $A \subseteq [0,1]^n$ of measure M minimizes the volume of k-chains, i.e. $\lambda\left(\mathcal{C}(\mathcal{F}, k)\right)$?*

Consider the first non-trivial case, i.e. $n = k = 2$. For $\mathbf{x} \in [0,1]^2$ define $M(\mathbf{x}) := \{\mathbf{y} \in A : \mathbf{x} \leq \mathbf{y}\}$. Let $S(A) := \{\mathbf{x} \in A : \nexists \mathbf{y} \in A : \mathbf{y} \leq \mathbf{x}\}$. Then it seems that in one of the optimal sets A the function $f(\mathbf{x}) := \lambda(M(\mathbf{x}))$ should be constant on S. Giving a nice description of the optimal set A in Problem 2.6.2

may well turn out to be difficult. It may be possible to determine the limiting structure of the solution as n, M remain fixed and k grows to infinity. Alternatively, estimates on the minimal volume of k-chains might be of interest and easier to obtain. Let $f(n, M, k) := \inf\{\lambda(\mathcal{C}(A, k)) : A \subseteq [0,1]^n,\ \lambda(A) = M\}$, where the infimum is taken over all measurable subsets A.

Problem 2.6.3. *Determine the value of* $f\left(2, \frac{1}{2}, 2\right)$.

Chapter 3

Applications of graph containers in the Boolean lattice

The results in this chapter are joint work with József Balogh and Andrew Treglown [13].

We apply the graph container method to prove a number of counting results for the Boolean lattice $\mathcal{P}(n)$. In particular, we:

(i) Give a partial answer to a question of Sapozhenko estimating the number of t error correcting codes in $\mathcal{P}(n)$, and we also give an upper bound on the number of transportation codes;

(ii) Provide an alternative proof of Kleitman's theorem on the number of antichains in $\mathcal{P}(n)$ and give a two-coloured analogue;

(iii) Give an asymptotic formula for the number of (p, q)-tilted Sperner families in $\mathcal{P}(n)$;

(iv) Prove a random version of Katona's t-intersection theorem.

In each case, to apply the container method, we first prove corresponding supersaturation results. We also give a construction which disproves two conjectures of Ilinca and Kahn on maximal independent sets and antichains in the Boolean lattice. A number of open questions are also given.

3.1 Introduction

Many problems in combinatorics and other areas can be rephrased into questions about independent sets in (hyper)graphs. For example, Sperner's theorem [101] states that the largest antichain in the power set of $[n]$, $\mathcal{P}(n)$ has size $\binom{n}{\lfloor n/2 \rfloor}$. ($\mathcal{P}(n)$ is also refered to as the *Boolean lattice*.) Let G be the graph with vertex set $\mathcal{P}(n)$ and where A and B are adjacent if $A \subset B$ or $B \subset A$. Then equivalently, Sperner's theorem states that the largest independent set in G has size $\binom{n}{\lfloor n/2 \rfloor}$.

So-called *container results* have emerged as powerful tools for attacking problems which reduce to counting independent sets in (hyper)graphs. Roughly speaking, container results typically state that the independent sets of a given (hyper)graph H lie only in a 'small' number of subsets of the vertex set of H (referred to as *containers*), where each of these containers is an 'almost independent set'. The *graph* container method dates back to work of Kleitman and Winston [73, 74] from more than 30 years ago. Indeed, they constructed a relatively simple algorithm that can be used to produce graph container results. This algorithm will be the starting point for proving the container results of this chapter; we give a more detailed overview of the method in Section 3.3. An excellent recent survey of Samotij [95] gives several applications of this method to a range of problems in combinatorics and number theory.

The container method has also been recently generalised to hypergraphs of higher uniformity. Perhaps the first applications of the hypergraph container method appeared in [12]. Balogh, Morris and Samotij [8] and independently Saxton and Thomason [98] developed general container theorems for hypergraphs whose edge distribution satisfies certain boundedness conditions. These results have been used to tackle a range of important problems including questions arising in combinatorial number theory, Ramsey theory, positional games, list colourings of graphs and H-free graphs.

In this chapter we provide several new short applications of the graph container method to counting problems in the Boolean lattice. In Section 3.4 we asymptotically determine the number of (p, q)-tilted Sperner families in $\mathcal{P}(n)$. In Section 3.5 we give an upper bound on the number of t error correcting codes, thereby giving a partial answer to a question of Sapozhenko [97], and an upper bound on the number of so-called 2-(n, k, d)-codes. Katona's intersection theorem [66] determines the largest t-intersecting family in $\mathcal{P}(n)$. In Section 3.6 we prove a random analogue of this result. We also prove counting versions of generalisations of Sperner's theorem: we give an alternative proof of a famous result of Kleitman [71] that gives an asymptotic formula for the number of antichains in $\mathcal{P}(n)$ (see Section 3.7.1). We then prove a two-coloured generalisation of this result in Section 3.7.2. Finally, in Section 3.8 we give a construction which disproves two conjectures of Ilinca and Kahn [64] on maximal independent sets and antichains in the Boolean lattice.

Section 3.3 describes the general algorithm used for producing our graph container results. After this, each of the sections are self-contained and so can be read separately. However, there are two important themes which run throughout the chapter and which we are keen to publicise. Firstly, for the proof of each of our container theorems, the key step is to apply various *supersaturation* results. Roughly speaking, such results state that if a vertex set S in some auxiliary graph G is significantly bigger than the size of the largest independent set, then $G[S]$ contains many edges. Secondly, in some cases we need to apply a *multi-stage* version of the Kleitman–Winston algorithm (and apply more than one supersaturation result). We explain this in more detail in Section 3.3.

3.2 Notation and preliminaries

For a given $n \in \mathbb{N}$, write $[n] := \{1, \dots, n\}$. Denote S_n the set of all permutations of $[n]$. Given a set X we write $\mathcal{P}(X)$ for the set of all subsets of X. Given $k \in \mathbb{N}$, we write $\binom{X}{\leq k}$ to denote the set of all subsets of X of size at most k and define $\binom{X}{k}$ and $\binom{X}{\geq k}$ analogously. Given $n \in \mathbb{N}$, we write, for example, $\binom{n}{\geq k} := \binom{n}{k} + \binom{n}{k+1} + \cdots + \binom{n}{n}$. We say two sets A, B are *comparable* if $A \subset B$ or $B \subset A$.

Given a graph G we write $N_G(x)$ for the neighbourhood of a vertex $x \in V(G)$ and set $\deg_G(x) := |N_G(x)|$. We write $\Delta(G)$ for the *maximum degree of* G.

Throughout the chapter we omit floors and ceilings where the argument is unaffected. We write $0 < \alpha \ll \beta \ll \gamma$ to mean that we can choose the constants α, β, γ from right to left. More precisely, there are increasing functions f and g such that, given γ, whenever we choose $\beta \leq f(\gamma)$ and $\alpha \leq g(\beta)$, all calculations needed in our proof are valid. Hierarchies of other lengths are defined in the obvious way.

The following well known bounds for binomial coefficients will be useful later on.

Fact 3.2.1.
$$\binom{n}{n/2} \sim \sqrt{\frac{2}{\pi n}} 2^n.$$

Fact 3.2.2. *If* $k = (n + c\sqrt{n})/2$ *where* $c = o(n^{1/6})$ *then*

$$\binom{n}{k} \sim \binom{n}{n/2} e^{-(c^2/2)}.$$

Fact 3.2.3. *For any* $n, k \in \mathbb{N}$,
$$\binom{n}{k} \leq \left(\frac{e \cdot n}{k}\right)^k.$$

3.3 The graph container algorithm

For each of our problems, we will prove and then apply a *container* result. We will first introduce some auxiliary graph G. For example, to prove Kleitman's theorem on antichains in the Boolean lattice, we will define G to have vertex set $\mathcal{P}(n)$ where distinct A and B are adjacent if they are comparable. Most of our container results then take the following general structure: Let $I_{\max}$ denote the size of the largest independent set in G. Then there is a collection $\mathcal{F}$ of subsets of $V(G)$ such that:

(i) $|\mathcal{F}| = 2^{o(|I_{\max}|)}$;

(ii) Every independent set I in G lies in some $F \in \mathcal{F}$;

(iii) $|F| \leq (1+o(1))|I_{\max}|$ for each $F \in \mathcal{F}$.

We refer to the elements of $\mathcal{F}$ as *containers*. In some cases, when we only have an upper bound D on $|I_{\max}|$, we in fact have D instead of $|I_{\max}|$ in (i) and (iii). Typically the container result will then immediately imply our desired counting theorem. For example, in the case of Kleitman's theorem, since independent sets in G correspond to antichains in $\mathcal{P}(n)$, we have that $|I_{\max}| = \binom{n}{\lfloor n/2 \rfloor}$. Thus, (i)–(iii) imply that there are $2^{(1+o(1))\binom{n}{\lfloor n/2 \rfloor}}$ antichains in $\mathcal{P}(n)$, as desired.

To prove each of our container results we will apply the following algorithm of Kleitman and Winston [73, 74].

The graph container algorithm. Let $V := V(G)$, $n := |V|$ and fix an arbitrary total order $v_1, \dots, v_n$ of V and some $\Delta > 0$. Let I be an independent set in G. Set $G_0 := G$ and $S := \emptyset$. In Step i of the algorithm we do the following:

(a) Let u be the vertex of maximum degree in G_{i-1} (ties are broken here by our fixed total ordering);

(b) If $u \notin I$ then define $G_i := G_{i-1} \setminus \{u\}$ and move to Step $i+1$;

(c) If $u \in I$ and $\deg_{G_{i-1}}(u) \geq \Delta$ then add u to S; define $G_i := G_{i-1} \setminus (\{u\} \cup N_G(u))$ and move to Step $i+1$;

(d) If $u \in I$ and $\deg_{G_{i-1}}(u) < \Delta$ then define $f(S) := V(G_i)$ and terminate.

Note that $I \subseteq S \cup f(S)$. We sometimes refer to Δ as the *parameter* of the algorithm.

The algorithm produces a function $f : \binom{V}{\leq |V|/\Delta} \to \mathcal{P}(V)$. Indeed, the algorithm ensures that $|S| \leq |V|/\Delta$ and that f is well-defined.

Let $\mathcal{F}$ denote the collection of sets $S \cup f(S)$ for each $S \in \binom{V}{\leq |V|/\Delta}$. By construction (ii) is satisfied. There are $\binom{V}{\leq |V|/\Delta}$ containers in $\mathcal{F}$. Thus, if one chooses Δ sufficiently large we can ensure that (i) is satisfied. At the end of the algorithm, G_i has maximum degree less than Δ, so is 'sparse'. In a standard application of the algorithm, we then apply a supersaturation result to ensure that (iii) holds: roughly speaking, since G_i is sparse it cannot be too much bigger than the largest independent set in G. Hence, G_i and so $S \cup f(S)$ is not too big.

In some cases though, the value of Δ required to ensure that (i) holds is not small enough to immediately ensure (iii) also holds. That is, $\Delta(G_i) \leq \Delta$ may not imply that (iii) holds. In this case we have to analyse the algorithm more carefully. Roughly speaking, the idea is to first apply the algorithm with some relatively large parameter Δ'. This will ensure (i) holds and by applying a supersaturation result the graph G_i is not too big (though perhaps much bigger than $(1+o(1))|I_{\max}|$). We then continue the algorithm with a new, much smaller parameter Δ to ensure at the end of this process G_i is much sparser and so (via another supersaturation result) (iii) is satisfied. We will use this multi-stage approach in Section 3.5. This idea was first used only very recently in [10] to prove a random analogue of Sperner's theorem. We remark that when applying this approach in Section 3.5, we will not explicitly state it in this way (we only state the parameter Δ explicitly and then split the analysis of the algorithm in two), but the method described above is (implicitly) precisely what is happening.

3.4 Tilted Sperner families

Let $\mathcal{P}(n)$ denote the power set of $[n]$, ordered by inclusion. A subset $\mathcal{A} \subseteq \mathcal{P}(n)$ is an *antichain* if for any $A, B \in \mathcal{A}$ with $A \subseteq B$ we have $A = B$. So $\binom{[n]}{k}$ is an antichain for any $0 \leq k \leq n$. A celebrated theorem of Sperner [101] states that in fact no antichain in $\mathcal{P}(n)$ has size larger than $\binom{n}{\lfloor n/2 \rfloor}$.

Given $A, B \subseteq [n]$ the *subcube of* $\mathcal{P}(n)$ *spanned by* A *and* B consists of all subsets of $A \cup B$ that contain $A \cap B$. Kalai (see [80]) observed that $\mathcal{A}$ is an antichain precisely if it does not contain A and B such that, in the subcube of $\mathcal{P}(n)$ spanned by A and B, A is the top point and B is the bottom point. He asked what happens if one 'tilts' this condition. That is, for some $p, q \in \mathbb{N}$ we forbid A to be $p/(p+q)$ of the way up this subcube and B to be $q/(p+q)$ of the way up this subcube. More precisely, we say that $\mathcal{A} \subseteq \mathcal{P}(n)$ is a (p,q)-*tilted Sperner family* if $\mathcal{A}$ does not contain distinct A, B such that $q|A \setminus B| = p|B \setminus A|$. So the case when $p \neq 0$, $q = 0$ corresponds to antichains.

Let $p, q \in \mathbb{N}$ be coprime with $p < q$. Leader and Long [80] proved that the largest (p,q)-tilted Sperner family in $\mathcal{P}(n)$ has size $(q - p + o(1))\binom{n}{n/2}$, where the lower bound is obtained by considering the union of the $q - p$ middle layers of the Boolean lattice (see [80] for an explanation of this).

In 1897, Dedekind [31] raised the question of how many antichains there are in $\mathcal{P}(n)$. This was famously resolved asymptotically by Kleitman [71] who proved that there are in fact $2^{(1+o(1))\binom{n}{n/2}}$ antichains. In this section we prove an analogue of this result for (p,q)-tilted Sperner families.

Theorem 3.4.1. *Let $p, q \in \mathbb{N}$ be coprime with $p < q$. Then there are*

$$2^{(q-p+o(1))\binom{n}{n/2}}$$

(p,q)-tilted Sperner families in $\mathcal{P}(n)$.

To prove Theorem 3.4.1 we will apply the following *supersaturation* version of the Leader–Long theorem [80]. The proof applies the same averaging argument strategy used in [80].

Lemma 3.4.2. *Let $p, q \in \mathbb{N}$ be coprime with $p < q$. Given any $\varepsilon > 0$, there exist $\delta > 0$ and $n_0 \in \mathbb{N}$ such that the following holds. Suppose that $n \geq n_0$ and $\mathcal{A} \subseteq \mathcal{P}(n)$ such that $|\mathcal{A}| \geq (q-p+\varepsilon)\binom{n}{n/2}$. Then there are at least $\delta\binom{n}{n/2}n^{p+q}$ pairs $A, B \in \mathcal{A}$ such that $q|A \setminus B| = p|B \setminus A|$.*

We remark that the conclusion of Lemma 3.4.2 is actually somewhat stronger than what is needed in the application to the proof of Theorem 3.4.1. Indeed, for our application instead of $\delta\binom{n}{n/2}n^{p+q}$ such pairs, having only $\delta n 2^n$ would be sufficient.

Proof. Given $\varepsilon > 0$, define $\delta > 0$ and $C, n_0 \in \mathbb{N}$ such that

$$0 < 1/n_0 < \delta \ll 1/C \ll \varepsilon, 1/p, 1/q.$$

Let $n \geq n_0$ and $\mathcal{A} \subseteq \mathcal{P}(n)$ such that $|\mathcal{A}| \geq (q-p+\varepsilon)\binom{n}{n/2}$.

Let $\mathcal{A}_i$ denote the set of $A \in \mathcal{A}$ with $|A| = i$. Since $1/n_0 \ll 1/C \ll \varepsilon$,

$$\sum_{i \geq n/2 + C\sqrt{n}} \binom{n}{i} + \sum_{i \leq n/2 - C\sqrt{n}} \binom{n}{i} \leq \frac{\varepsilon}{2}\binom{n}{n/2}.$$

Thus, we may assume that $|\mathcal{A}| \geq (q-p+\varepsilon/2)\binom{n}{n/2}$ and every $A \in \mathcal{A}$ satisfies $n/2 - C\sqrt{n} \leq |A| \leq n/2 + C\sqrt{n}$.

For simplicity we may assume that $n = (p+q)m$ for some $m \in \mathbb{N}$ (the other cases follow identically). Clearly there exists $k \in [0, q-p-1]$ such that

$$\sum_{i \equiv k \mod (q-p)} |\mathcal{A}_i| \geq \left(1 + \frac{\varepsilon}{2(q-p)}\right)\binom{n}{n/2}. \tag{3.4.1}$$

Define $k' \in [0, q-p-1]$ so that $k' \equiv k - pm \mod (q-p)$.

Pick a random ordering of $[n]$ that we denote by $(a_1, \dots, a_{qm}, b_1, \dots, b_{pm})$ (this can be viewed as a permutation of $[n]$). Given this ordering, define $C_i := \{a_j : j \in [qi + k']\} \cup \{b_{j'} : j' \in [pi+1, pm]\}$ and set $\mathcal{C} := \{C_i : i \in [0, m-1]\}$. Notice that for every $i < j$ we have $|C_i| = pm + k' + (q-p)i \equiv k \mod (q-p)$ and $p|C_j \setminus C_i| = q|C_i \setminus C_j|$. Further, for each $i \equiv k \mod (q-p)$ where $pm + k' \leq n/2 - C\sqrt{n} \leq i \leq n/2 + C\sqrt{n} \leq qm + k'$, there is precisely one set $C_{i'}$ in $\mathcal{C}$ of size i.

Consider the random variable $X := |\mathcal{A} \cap \mathcal{C}|$. Let $i \equiv k \mod (q-p)$ where $n/2 - C\sqrt{n} \leq i \leq n/2 + C\sqrt{n}$, and set i' so that $i = pm + k' + (q-p)i'$. Note that each set $B \in \binom{[n]}{i}$ is equally likely to be $C_{i'}$, therefore

$\mathbb{P}[B \in \mathcal{C}] = \frac{1}{\binom{n}{i}}$. So

$$\mathbb{E}X = \sum_{i \equiv k \mod (q-p)} \frac{|\mathcal{A}_i|}{\binom{n}{i}} \geq \sum_{i \equiv k \mod (q-p)} \frac{|\mathcal{A}_i|}{\binom{n}{n/2}} \overset{(3.4.1)}{\geq} 1 + \frac{\varepsilon}{2(q-p)}. \tag{3.4.2}$$

Consider any permutation $\pi \in S_n$. Write π as $(a'_1, \ldots, a'_{qm}, b'_1, \ldots, b'_{pm})$. Define $C_{\pi,i} := \{a'_j : j \in [qi + k']\} \cup \{b'_{j'} : j' \in [pi+1, pm]\}$ and set $\mathcal{C}_\pi := \{C_{\pi,i} : i \in [0, m-1]\}$. So the set $\mathcal{C}$ is simply $\mathcal{C}_\pi$ for a randomly selected permutation π. Set $\alpha(\pi) := |\mathcal{A} \cap \mathcal{C}_\pi|$. Thus,

$$\mathbb{E}X = \frac{1}{n!} \sum_{\pi \in S_n} \alpha(\pi).$$

Together with (3.4.2) this implies that

$$\sum_{\pi \in S_n} \binom{\alpha(\pi)}{2} \geq \sum_{\pi \in S_n} (\alpha(\pi) - 1) \geq \frac{\varepsilon n!}{2(q-p)}. \tag{3.4.3}$$

We say a pair $A, B \in \mathcal{A}$ is *good* if there is some permutation π such that $A, B \in \mathcal{C}_\pi$. That is, $A = C_{\pi,i}$ and $B = C_{\pi,j}$ for some i, j and $\pi \in S_n$. In this case, since $A, B \in \mathcal{A}$, we have $n/2 - C\sqrt{n} \leq |C_{\pi,i}|, |C_{\pi,j}| \leq n/2 + C\sqrt{n}$. Further, by definition of $\mathcal{C}_\pi$:

(i) $|A \cap B| \geq n/2 - (2p+1)C\sqrt{n}$;

(ii) $q|A \setminus B| = p|B \setminus A|$ or $p|A \setminus B| = q|B \setminus A|$.

(By relabeling A, B we may assume that $q|A \setminus B| = p|B \setminus A|$.) Moreover, if A, B is good, the definition of the $\mathcal{C}_\pi$ implies that there are precisely

$$|A \cap B|!|A \setminus B|!|B \setminus A|!|\overline{A \cup B}|! \tag{3.4.4}$$

permutations π such that $A, B \in \mathcal{C}_\pi$. Additionally, the following conditions hold:

- $|A \cap B|, |\overline{A \cup B}| \leq n/2 + C\sqrt{n}$;
- $p \leq |A \setminus B| \overset{(i)}{\leq} (2p+2)C\sqrt{n}$;
- $q \leq |B \setminus A| \overset{(i)}{\leq} (2p+2)C\sqrt{n}$.

Under these constraints, an upper bound on (3.4.4) is

$$p!q!(n/2 + C\sqrt{n})!(n/2 - C\sqrt{n} - p - q)!.$$

Together with (3.4.3) this implies that there are at least

$$\frac{\varepsilon n!}{2(q-p)} \times \frac{1}{p!q!(n/2 + C\sqrt{n})!(n/2 - C\sqrt{n} - p - q)!} \geq \frac{\varepsilon n!}{2(q-p)} \times \frac{\delta^{1/2} n^{p+q}}{(n/2)!(n/2!)} \geq \delta \binom{n}{n/2} n^{p+q}$$

good pairs $A, B \in \mathcal{A}$. (In the last inequality we apply Fact 3.2.2.) Since each such pair satisfies (ii), this completes the proof. ☐

Lemma 3.4.2 can now be applied to prove the following *container* lemma which immediately implies Theorem 3.4.1.

Lemma 3.4.3. *Let $p, q \in \mathbb{N}$ be coprime with $p < q$. There is a collection $\mathcal{F} \subseteq \mathcal{P}(n)$ with the following properties:*

(i) $|\mathcal{F}| = 2^{o(1)\binom{n}{n/2}}$*;*

(ii) If $\mathcal{A} \subseteq \mathcal{P}(n)$ is a (p, q)-tilted Sperner family, then $\mathcal{A}$ is contained in some member of $\mathcal{F}$;

(iii) $|F| \leq (q - p + o(1))\binom{n}{n/2}$ *for every $F \in \mathcal{F}$.*

Proof. Let $\varepsilon > 0$ and let δ, n_0 be as in Lemma 3.4.2. Let $n \geq n_0$. Define G to be the graph with vertex set $\mathcal{P}(n)$ in which distinct sets A and B are adjacent if and only if $p|A \setminus B| = q|B \setminus A|$ or $q|A \setminus B| = p|B \setminus A|$. Thus a (p, q)-tilted Sperner family in $\mathcal{P}(n)$ is precisely an independent set in G.

Claim 3.4.4. *There exists a function $f : \binom{V(G)}{\leq 2^n/\delta n} \to \binom{V(G)}{\leq (q-p+\varepsilon)\binom{n}{n/2}}$ such that, for any independent set I in G, there is a subset $S \subseteq I$ where $S \in \binom{V(G)}{\leq 2^n/\delta n}$ and $I \subseteq S \cup f(S)$.*

To prove the claim, fix an arbitrary total order $v_1, \ldots, v_{2^n}$ on the vertices of $V(G)$. Given any independent set I in G, define $G_0 := G$, and take S to be initially empty. We add vertices to S through the following iterative process: At Step i, let u be the maximum degree vertex of G_{i-1} (with ties broken by our fixed total order). If $u \notin I$ then define $G_i := G_{i-1} \setminus \{u\}$, and proceed to Step $i + 1$. Alternatively, if $u \in I$ and $\deg_{G_{i-1}}(u) \geq \delta n$ then add u to S, define $G_i := G_{i-1} \setminus (\{u\} \cup N_G(u))$, and proceed to Step $i + 1$. Finally, if $u \in I$ and $\deg_{G_{i-1}}(u) < \delta n$, then set $f(S) := V(G_i)$ and terminate.

Observe that for any independent set I in G the process defined ensures that $S \subseteq I$ where $|S| \le 2^n/\delta n$ and $I \subseteq S \cup f(S)$. Further, at the end of the process we know that $\Delta(G_i) < \delta n$ and so $e(G_i) < \delta n 2^n < \delta \binom{n}{n/2} n^{p+q}$. Hence, Lemma 3.4.2 implies that $|f(S)| = |V(G_i)| \le (q - p + \varepsilon)\binom{n}{n/2}$.

To complete the claim we must show that f is well-defined. That is, we must check that if the process described above yields the same set S when applied to independent sets I and I', then it should also yield the same set $f(S)$. However, this is a consequence of the fact that we always chose u to be the vertex of I of maximum degree in G_{i-1}. Thus, the claim is proven.

Define $\mathcal{F}$ to be the collection of all the sets $S \cup f(S)$ for every $S \in \binom{V(G)}{\le 2^n/\delta n}$. Then (i) and (ii) hold and $|F| \le (q - p + \varepsilon)\binom{n}{n/2} + 2^n/\delta n \le (q - p + 2\varepsilon)\binom{n}{n/2}$ for every $F \in \mathcal{F}$, as desired.

□

3.5 The number of t error correcting codes and 2-(n, k, d)-codes

3.5.1 Counting t error correcting codes: Sapozhenko's question

The *Hamming distance* $d(A, B)$ between two sets $A, B \subseteq [n]$ is defined as

$$d(A, B) := |A \backslash B| + |B \backslash A|.$$

In this section we will view a subset A of $[n]$ as a string of length n over the alphabet $\{0, 1\}$ where the ith entry of the string is 1 precisely when $i \in A$. In this setting, the Hamming distance between A and B can be rewritten as

$$d(A, B) = |\{1 \le i \le n : A_i \neq B_i\}|,$$

where for example, A_i denotes the ith term in the string A. The *Hamming ball* of radius r around a set $A \subseteq [n]$ is defined as

$$B(A, r) := \{X \subseteq [n] : d(A, X) \le r\}.$$

Given a family $\mathcal{C} \subset \mathcal{P}(n)$, we say $\mathcal{C}$ is a *distance d code* if the Hamming distance between any two distinct members of $\mathcal{C}$ is at least d. Moreover $\mathcal{C}$ is said to be a *t error correcting code* if there exists a *decoding function* $\mathrm{Dec} : \{0, 1\}^n \to \mathcal{C}$ such that for every $X \in \{0, 1\}^n$ and $A \in \mathcal{C}$ with $d(X, A) \le t$ we have $\mathrm{Dec}(X) = A$. Recall that $\mathcal{C}$ is t error correcting if and only if for every pair $A, B \in \mathcal{C}$ we have $d(A, B) \ge 2t + 1$. So t error correcting codes are precisely distance $2t + 1$ codes, i.e. codes where the Hamming balls $\{B(A, t) : A \in \mathcal{C}\}$

are disjoint. Since most communication channels are subject to channel noise, which can cause errors in the transmission of messages, the additional redundancy given by error correcting codes plays a crucial role in ensuring that the receiver can recover the original message. Given the widespread usage of such codes in digital communications, natural question to ask is *how many* t error correcting codes there are in total, i.e. estimate the size of

$$|\{\mathcal{C} \subseteq \mathcal{P}(n) : \mathcal{C} \text{ is } t \text{ error correcting}\}|.$$

This problem was first raised by Sapozhenko [97]. We wish to bound the number of t error correcting codes of length n and alphabet $\{0,1\}$. An upper bound for the size of such a code is given by the Hamming bound, which gives an important limitation on the efficiency of error correcting codes. Let $V(n,t)$ be the volume of a Hamming ball of radius t in $[n]$, so $V(n,t) = \sum_{k=0}^{t} \binom{n}{k}$. Then the Hamming bound states that if $\mathcal{C}$ is t error correcting, since the Hamming balls of radius t centered at the members of $\mathcal{C}$ have to be disjoint, we have

$$|\mathcal{C}| \leq \frac{2^n}{V(n,t)}.$$

If $\mathcal{C}$ attains equality in the Hamming bound, we say $\mathcal{C}$ is a *perfect code*. Perfect codes are precisely those for which the Hamming balls centered at the codewords fill up the entire space $\{0,1\}^n$ without overlap. The trivial perfect codes are codes consisting of a single codeword (when $t = n$), or the whole of $\{0,1\}^n$ (when $t = 0$), and *repetition codes* where the same substring is repeated an odd number of times. The non-trivial perfect codes over prime-power alphabets must have the same parameters as the so-called *Hamming codes* or the *Golay codes* (see [104]).

Let $H(n,t) := \frac{2^n}{V(n,t)}$. Since every subset of a t error correcting code is also t error correcting, if $\mathcal{C}$ is t error correcting then the number of t error correcting codes is at least $2^{|\mathcal{C}|}$. In particular, if the parameters n,t are such that a perfect code exists, then the number of t error correcting codes is at least $2^{H(n,t)}$. Our first goal is to prove a corresponding upper bound:

Theorem 3.5.1. *Let $t = t(n) \ll \sqrt[3]{\frac{n}{\log^2 n}}$. Then the number of t error correcting codes is at most $2^{H(n,t)(1+o(1))}$.*

The range of t given in Theorem 3.5.1 is probably not optimal - indeed our guess is that the conclusion of Theorem 3.5.1 should hold whenever $t \ll \frac{n}{\log n}$. However, a heuristic argument suggests that if $t \gg \frac{n}{\log n}$ the conclusion of Theorem 3.5.1 may fail. Indeed, suppose one could partition $\{0,1\}^n$ into disjoint copies of balls of radius $t+1$, obtaining roughly $\frac{t}{n}H(n,t)$ balls. From each ball we can pick one element, that is either the centre of the ball or an element at distance one from the centre, giving $n+1$ choices for each ball.

Every family we obtain like this is a t error correcting code, and we have roughly $n^{\frac{t}{n}H(n,t)} \gg 2^{H(n,t)}$ such families.

Our overarching proof strategy is similar to the one used in the previous section. However, now we will employ a two phase strategy to construct our containers and as such we require two different supersaturation results. The first states that if $|\mathcal{C}|$ is slightly bigger than $H(n,t)$ then it contains many *bad pairs*, i.e. pairs at distance less than $2t+1$. Let $W(t,d)$ be the size of the intersection of two Hamming balls of radius t in n, the centers being distance d apart. So $W(t,1) \geq W(t,d)$ for all $d \geq 2$, and $W(t,1) = 2V(n-1,t-1)$. The key observation is that the volume of the intersection of two balls is significantly smaller than the volume of a single ball.

Lemma 3.5.2. *Let $\mathcal{C} \subset \mathcal{P}(n)$. If $|\mathcal{C}| \geq H(n,t) + x$ then there are at least $x\frac{n}{2t}$ pairs $A, B \in \mathcal{C}$ that have Hamming distance at most $2t$.*

Proof. For $X \in \{0,1\}^n$, let $K_X := \{A \in \mathcal{C} : d(A,X) \leq t\}$. For $k \in \mathbb{N}$ set $S_k := \{X \in \{0,1\}^n : |K_X| = k\}$. Then $\sum_k k|S_k| = |\mathcal{C}|V(n,t) \geq 2^n + xV(n,t)$. So the number of pairs in $\mathcal{C}$ of distance at most $2t$ is at least

$$\frac{1}{W(t,1)}\sum_k |S_k|\binom{k}{2} \geq \frac{1}{W(t,1)}\sum_k |S_k|(k-1) \geq x \cdot \frac{V(n,t)}{W(t,1)} = x \cdot \frac{V(n,t)}{2V(n-1,t-1)} \geq x\frac{n}{2t}.$$

$\square$

Our next supersaturation lemma considers sets of size at least $2H(n,t)$. Consider the graph G with $V(G) = \mathcal{P}(n)$, where two distinct vertices A, B are connected by an edge of colour $d(A,B)$ if they form a bad pair, i.e. their Hamming distance is at most $2t$. Define

$$\alpha := \frac{n}{10tH(n,t)}.$$

Lemma 3.5.3. *Let $\mathcal{C} \subset \mathcal{P}(n)$. If $|\mathcal{C}| \geq 2H(n,t)$, then there is an $A \in \mathcal{C}$ such that its degree in $G[\mathcal{C}]$ is at least $\alpha|\mathcal{C}|$.*

Proof. Let E_i denote the number of pairs of vertices connected by an edge of colour i in $G[\mathcal{C}]$ for all $i = 1, \ldots, 2t$, and let $E := \sum_i E_i$. Define K_X as in the proof of Lemma 3.5.2. Note that

$$\sum_{k=1}^{2t} W(t,k)E_k = \sum_{X \in \{0,1\}^n} \binom{|K_X|}{2} \tag{3.5.1}$$

since both terms count the number of pairs $(X,(A,B))$ where $X \in \{0,1\}^n$, $A, B \in \mathcal{C}$ and $d(X,A)$, $d(X,B) \leq t$. The average value of K_X over all $X \in \{0,1\}^n$ is $|\mathcal{C}|V(n,t)/2^n$. Thus,

$$\sum_{X \in \{0,1\}^n} \binom{|K_X|}{2} \geq 2^n \binom{|\mathcal{C}|V(n,t)/2^n}{2}. \tag{3.5.2}$$

Combining (3.5.1) and (3.5.2), since $|\mathcal{C}|V(n,t)/2^n \geq 2$, we have that

$$\sum_{k=1}^{2t} W(t,k)E_k \geq \frac{|\mathcal{C}|^2 V(n,t)^2}{10 \cdot 2^n}.$$

As $W(t,k) \leq W(t,1) = 2V(n-1,t-1) \leq \frac{2t}{n}V(n,t)$, we have that

$$E \geq \frac{|\mathcal{C}|^2 V(n,t)n}{20t2^n}$$

and the result follows. $\square$

Given these two supersaturation results, we are now ready to prove the following container lemma which immediately implies Theorem 3.5.1.

Lemma 3.5.4. *Let $t = t(n) \ll \sqrt[3]{\frac{n}{\log^2 n}}$. There is a collection $\mathcal{F} \subseteq \mathcal{P}(n)$ with the following properties:*

(i) $|\mathcal{F}| = 2^{o(H(n,t))}$;

(ii) If $\mathcal{C} \subseteq \mathcal{P}(n)$ is a t error correcting code, then $\mathcal{C}$ is contained in some member of $\mathcal{F}$;

(iii) $|F| \leq (1+o(1))H(n,t)$ *for every* $F \in \mathcal{F}$.

Proof. Let $0 < \varepsilon < 1$ and let n be sufficiently large. Let G be the graph with vertex set $\mathcal{P}(n)$ in which distinct sets A and B are adjacent if and only if their Hamming distance is at most $2t$. Thus a t error correcting code in $\mathcal{P}(n)$ is precisely an independent set in G.

Claim 3.5.5. *There exists a function $f : \binom{V(G)}{\leq \varepsilon \frac{H(n,t)}{t \log n}} \to \binom{V(G)}{\leq (1+\varepsilon)H(n,t)}$ such that, for any independent set I in G, there is a subset $S \subseteq I$ where $S \in \binom{V(G)}{\leq \varepsilon \frac{H(n,t)}{t \log n}}$ and $I \subseteq S \cup f(S)$.*

To prove the claim, fix an arbitrary total order $v_1, \ldots, v_{2^n}$ on the vertices of $V(G)$. Given any independent set I in G, define $G_0 := G$, and take S to be initially empty. We add vertices to S through the following iterative process: At Step i, let u be the maximum degree vertex of G_{i-1} (with ties broken by our fixed total order). If $u \notin I$ then define $G_i := G_{i-1} \setminus \{u\}$, and proceed to Step $i+1$. Alternatively, if $u \in I$ and

$\deg_{G_{i-1}}(u) \geq \varepsilon n/4t$ then add u to S, define $G_i := G_{i-1} \setminus (\{u\} \cup N_G(u))$, and proceed to Step $i+1$. Finally, if $u \in I$ and $\deg_{G_{i-1}}(u) < \varepsilon n/4t$, then set $f(S) := V(G_i)$ and terminate.

Observe that for any independent set I in G the process defined ensures that $S \subseteq I$ and $I \subseteq S \cup f(S)$. Further, at the end of the process we know that $\Delta(G_i) < \varepsilon n/4t$ and so $e(G_i) < |V(G_i)|\varepsilon n/4t$. Hence, Lemma 3.5.2 implies that $|f(S)| = |V(G_i)| \leq (1+\varepsilon)H(n,t)$. Moreover it is easy to see that f is well-defined.

To complete the proof of the claim, it remains to prove that $|S| \leq \varepsilon H(n,t)/(t \log n)$. We will distinguish two stages in the above algorithm, according to the size of $V(G_i)$. Let S_1 denote the set of vertices $u \in S$ that were added to S in some Step i of the algorithm where $|V(G_{i-1})| \geq 2H(n,t)$. Set $S_2 := S \setminus S_1$. So there is some k such that, up to and including Step k, every vertex added to S lies in S_1, and every vertex added to S after Step k lies in S_2.

By Lemma 3.5.3, for every $i \leq k$, at Step i we remove at least an α proportion of the vertices from G_{i-1} to obtain G_i. Thus, $|S_1| = k$ and $(1-\alpha)^k 2^n \leq 2H(n,t)$. Note that $\alpha \to 0$ as $n \to \infty$, so as n is sufficiently large we have that $\alpha \leq 10 \log(1/(1-\alpha))$. Therefore,

$$|S_1| \leq \frac{\log\left(\frac{2^n}{2H(n,t)}\right)}{\log\left(\frac{1}{1-\alpha}\right)} \leq 10\frac{\log V(n,t)}{\alpha} \leq 5000\frac{tH(n,t)}{n}t\log(n/t) \leq \frac{\varepsilon}{2}\frac{H(n,t)}{t\log n}.$$

Note that in the last inequality we use that $t \ll \sqrt[3]{\frac{n}{\log^2 n}}$.

After Step k we remove at least $\varepsilon n/4t$ vertices at each step, so we have

$$|S_2| \leq \frac{8tH(n,t)}{\varepsilon n} \leq \frac{\varepsilon}{2}\frac{H(n,t)}{t\log n}.$$

Hence,

$$|S| = |S_1| + |S_2| \leq \varepsilon\frac{H(n,t)}{t\log n},$$

as required. This finishes the proof of the claim.

Define $\mathcal{F}$ to be the collection of all the sets $S \cup f(S)$ for every $S \in \binom{V(G)}{\leq \varepsilon \frac{H(n,t)}{t\log n}}$. Then (ii) clearly holds. Further,

$$|\mathcal{F}| \leq \binom{2^n}{\leq \varepsilon\frac{H(n,t)}{t\log n}} \leq 2^{2\varepsilon\frac{H(n,t)}{t\log n}\log(tV(n,t)\log(n)/\varepsilon)} \leq 2^{2\varepsilon\frac{H(n,t)}{t\log n}(2t\log n + \log t + \log\log n + \log\frac{1}{\varepsilon})} \leq 2^{5\varepsilon H(n,t)}$$

and $|F| \leq (1+2\varepsilon)H(n,t)$ for all $F \in \mathcal{F}$. Since $0 < \varepsilon < 1$ was arbitrary, this proves the lemma.

□

3.5.2 Counting 2-(n,k,d)-codes

In this subsection, all pairs of sets considered are *unordered*. Let us now turn our attention to the space $\mathcal{Y}$ of pairs of disjoint k-subsets of $[n]$ (for some fixed $0 < k \leq n/2$). Given two pairs (A_1, A_2), $(B_1, B_2) \in \mathcal{Y}$, the *transportation distance*, or *Enomoto–Katona distance* is defined by

$$d((A_1, A_2), (B_1, B_2)) := \min\{|A_1 \setminus B_1| + |A_2 \setminus B_2|, |A_1 \setminus B_2| + |A_2 \setminus B_1|\}.$$

For convenience, throughout this subsection we will write *distance* when we mean transportation distance. The notion of transportation distance has been widely studied (also in a more general setting for metric spaces). See for example [108] and the introduction of [19] for background on the topic.

We say that a collection $\mathcal{C} \subseteq \mathcal{Y}$ is a 2-(n,k,d)*-code* if the distance between any two elements of $\mathcal{C}$ is at least d. Write $C(n,k,d)$ for the maximum size of a 2-(n,k,d)-code. Brightwell and Katona [20] proved that

$$C(n,k,d) \leq \frac{1}{2}\frac{n(n-1)\cdot \ldots \cdot (n-2k+d)}{\left(k(k-1)\cdot \ldots \cdot \lceil \frac{d+1}{2} \rceil\right)\left(k(k-1)\cdot \ldots \cdot \lfloor \frac{d+1}{2} \rfloor\right)} =: H(n,k,d). \tag{3.5.3}$$

Recently the value of $C(n,k,d)$ has been determined for many values of (n,k,d) (see [19, 24]). As an example, we have equality or are 'close' to equality in (3.5.3) when $k \geq 2$, $d = 2k-1$ and for certain (congruency) classes of n (see [24]). Further, the bound in (3.5.3) is asymptotically sharp for fixed k, d and $n \to \infty$ (see [19]). Our goal in this subsection is to prove the following upper bound on the number of 2-(n,k,d)-codes.

Theorem 3.5.6. *Suppose that $k = k(n) \leq n/2$ and $t = t(n) \ll \sqrt[3]{\frac{k}{\log^2 n}}$ then the number of* 2-$(n,k,2t+1)$-*codes is at most* $2^{H(n,k,2t+1)(1+o(1))}$.

Similarly to Theorem 3.5.1, we believe that the correct range of t in Theorem 3.5.6 should be $t \ll \frac{k}{\log n}$.

Given a pair $(A,B) \in \mathcal{Y}$, let $P((A,B),u)$ denote the family of pairs (U,V) where $|U| = |V| = u$ and $U \subseteq A, V \subseteq B$ or vice versa. Then $|P((A,B),u)| = \binom{k}{u}^2$. Let $\mathcal{Z}(u)$ be the space of pairs of disjoint sets of size u in $[n]$. So $|\mathcal{Z}(u)| = \frac{1}{2}\binom{n}{u}\binom{n-u}{u}$ and note that $\mathcal{Y} = \mathcal{Z}(k)$. We will refer to $P((A,B),k-t)$ as the *ball of radius t around* (A,B). In particular, for any $(A_1,B_1),(A_2,B_2)$ in $\mathcal{Y}$, if $P((A_1,B_1),k-t)$ and $P((A_2,B_2),k-t)$ intersect, then $d((A_1,A_2),(B_1,B_2)) \leq 2t$.

The proof strategy for Theorem 3.5.6 is extremely close to that of Theorem 3.5.1. The following supersaturation lemma is an analogue of Lemma 3.5.2.

Lemma 3.5.7. *Let $\mathcal{C} \subseteq \mathcal{Y}$. If $|\mathcal{C}| \geq H(n,k,2t+1)+x$ then there are at least xk/t pairs $(A_1,B_1),(A_2,B_2) \in \mathcal{C}$ at distance at most $2t$.*

Proof. Note that

$$\sum_{(A,B)\in\mathcal{C}} |P((A,B),k-t)| \geq \binom{k}{k-t}^2 (H(n,k,2t+1)+x) = \frac{1}{2}\binom{n}{k-t}\binom{n-k+t}{k-t} + x\binom{k}{k-t}^2$$
$$= |\mathcal{Z}(k-t)| + x\binom{k}{k-t}^2. \tag{3.5.4}$$

Let $W(k-t,d)$ denote the largest possible intersection of two balls $P((A_1,B_1),k-t)$ and $P((A_2,B_2),k-t)$, amongst all $(A_1,B_1),(A_2,B_2) \in \mathcal{Y}$ with $d((A_1,B_1),(A_2,B_2)) = d$. This is maximised when $A_1 = A_2$ and $|B_1 \cap B_2| = k-1$, so $W(k-t,d) \leq W(k-t,1)$ for all $d \geq 2$. Now $W(k-t,1) = \binom{k}{k-t}\binom{k-1}{k-t} = \binom{k}{k-t}^2\binom{k-1}{t-1}/\binom{k}{t}$. Combining this with (3.5.4) we see that there are at least

$$x\binom{k}{t}/\binom{k-1}{t-1} = xk/t$$

pairs $(A_1,B_1),(A_2,B_2) \in \mathcal{C}$ such that $P((A_1,B_1),k-t)$ and $P((A_2,B_2),k-t)$ intersect. Note that each such pair $(A_1,B_1),(A_2,B_2) \in \mathcal{C}$ have distance at most $2t$, as desired. $\square$

Consider the graph G with $V(G) = \mathcal{Y}$, two vertices $(A_1,B_1),(A_2,B_2)$ being connected by an edge of colour $d((A_1,B_1),(A_2,B_2))$ if they form a bad pair, i.e. their transportation distance is at most $2t$. Define the constant α by

$$\alpha := \frac{k}{10tH(n,k,2t+1)}.$$

Lemma 3.5.8. *Let $\mathcal{C} \subset \mathcal{Y}$. If $|\mathcal{C}| \geq 2H(n,k,2t+1)$, then there is a vertex $(A_1,B_1) \in \mathcal{C}$ such that its degree in $G[\mathcal{C}]$ is at least $\alpha|\mathcal{C}|$.*

Proof. One can prove the lemma by arguing in a similar way to the proof of Lemma 3.5.3. Now though given $X \in \mathcal{Z}(k-t)$ we take $K_X := \{(A,B) \in \mathcal{C} : X \in P((A,B),k-t)\}$ and $S_i := \{X \in \mathcal{Z}(k-t) : |K_X| = i\}$. By arguing as in Lemma 3.5.3 and using that $|\mathcal{Z}(k-t)| = H(n,k,2t+1)\binom{k}{k-t}^2$ we have that the number E of edges in $G[\mathcal{C}]$ satisfies

$$E \geq \frac{|\mathcal{Z}(k-t)|}{W(k-t,1)}\binom{|\mathcal{C}|/H(n,k,2t+1)}{2} \geq |\mathcal{C}|^2\frac{k}{20tH(n,k,2t+1)}$$

and the result follows. (In the last inequality we use that $W(k-t,1) = \binom{k}{k-t}^2\binom{k-1}{t-1}/\binom{k}{t}$.) $\square$

The following container lemma immediately implies Theorem 3.5.6; its proof follows the same approach used in the proof of Lemma 3.5.4.

Lemma 3.5.9. *Let $k = k(n) \leq n/2$ and $t = t(n) \ll \sqrt[3]{\frac{k}{\log^2 n}}$. There is a collection $\mathcal{F}$ of subsets of $\mathcal{Y}$ with the following properties:*

(i) $|\mathcal{F}| = 2^{o(H(n,k,2t+1))}$;

(ii) If $\mathcal{C} \subseteq \mathcal{Y}$ is a 2-$(n,k,2t+1)$-code, then $\mathcal{C}$ is contained in some member of $\mathcal{F}$;

(iii) $|F| \leq (1+o(1))H(n,k,2t+1)$ *for every* $F \in \mathcal{F}$.

Proof. Let $0 < \varepsilon < 1$ and let n be sufficiently large. Let G be the graph defined before Lemma 3.5.8.

Claim 3.5.10. *There exists a function $f : \binom{V(G)}{\leq \frac{\varepsilon H(n,k,2t+1)}{t \log n}} \to \binom{V(G)}{\leq (1+\varepsilon)H(n,k,2t+1)}$ such that, for any independent set I in G, there is a subset $S \subseteq I$ where $S \in \binom{V(G)}{\leq \frac{\varepsilon H(n,k,2t+1)}{t \log n}}$ and $I \subseteq S \cup f(S)$.*

To prove the claim we argue as in Claim 3.5.5 except that we now apply the graph container algorithm with parameter $\varepsilon k/2t$ instead of $\varepsilon n/4t$. That is, at Step i if $u \notin I$ then set $G_i = G_{i-1} \setminus \{u\}$; if $u \in I$ and $\deg_{G_{i-1}}(u) \geq \varepsilon k/2t$ we add u to S, define $G_i := G_{i-1} \setminus (\{u\} \cup N_G(u))$; if $u \in I$ and $\deg_{G_{i-1}}(u) < \varepsilon k/2t$, set $f(S) := V(G_i)$ and terminate.

As before we have that for any independent set I in G the process defined ensures that $S \subseteq I$ and $I \subseteq S \cup f(S)$. Further, at the end of the process we know that $\Delta(G_i) < \varepsilon k/2t$ and so $e(G_i) < |V(G_i)|\varepsilon k/2t$. Hence, Lemma 3.5.7 implies that $|f(S)| = |V(G_i)| \leq (1+\varepsilon)H(n,k,2t+1)$. Moreover f is well-defined.

To complete the proof of the claim, it remains to prove that $|S| \leq \varepsilon H(n,k,2t+1)/(t \log n)$. As in Claim 3.5.5 we distinguish two stages in the above algorithm, according to the size of $V(G_i)$. Let S_1 denote the set of vertices $u \in S$ that were added to S in some Step i of the algorithm where $|V(G_{i-1})| \geq 2H(n,k,2t+1)$. Set $S_2 := S \setminus S_1$. So there is some k such that, up to and including Step k, every vertex added to S lies in S_1, and every vertex added to S after Step k lies in S_2.

By Lemma 3.5.8, for every $i \leq k$, at Step i we remove at least an α proportion of the vertices from G_{i-1} to obtain G_i. Thus, $|S_1| = k$ and $(1-\alpha)^k |\mathcal{Y}| \leq 2H(n,k,2t+1)$. Note that $\alpha \to 0$ as $n \to \infty$, so as n is sufficiently large we have that $\alpha \leq 10 \log(1/(1-\alpha))$. Therefore,

$$\begin{aligned}|S_1| &\leq \frac{\log\left(\frac{|\mathcal{Y}|}{2H(n,k,2t+1)}\right)}{\log\left(\frac{1}{1-\alpha}\right)} \leq 10\frac{\log\left(\frac{(n-2k+2t)\ldots(n-2k+1)}{2(t!)^2}\right)}{\alpha} \leq 5000\frac{tH(n,k,2t+1)\log\binom{n}{t}}{k} \\ &\leq 10000\frac{t^2 \log n}{k} H(n,k,2t+1) < \frac{\varepsilon}{2}\frac{H(n,k,2t+1)}{t \log n}.\end{aligned}$$

In the first inequality we used that $|\mathcal{Y}| = \binom{n}{k}\binom{n-k}{k}/2$ and in the last inequality we use that $t \ll \sqrt[3]{\frac{k}{\log^2 n}}$.

After Step k we remove at least $\varepsilon k/2t$ vertices at each step, so we have

$$|S_2| \leq \frac{2H(n,k,2t+1)}{\frac{\varepsilon k}{2t}} = \frac{4t}{\varepsilon k} H(n,k,2t+1) \leq \frac{\varepsilon}{2} \frac{H(n,k,2t+1)}{t \log n}.$$

Hence,

$$|S| = |S_1| + |S_2| \leq \varepsilon \frac{H(n,k,2t+1)}{t \log n},$$

as required. This finishes the proof of the claim.

Define $\mathcal{F}$ to be the collection of all the sets $S \cup f(S)$ for every $S \in \binom{V(G)}{\leq \varepsilon \frac{H(n,k,2t+1)}{t \log n}}$. Then (ii) clearly holds. Further,

$$|\mathcal{F}| \leq \binom{\frac{1}{2}\binom{n}{k}\binom{n-k}{k}}{\leq \varepsilon \frac{H(n,k,2t+1)}{t \log n}} \leq 2^{2\varepsilon \frac{H(n,k,2t+1)}{t \log n} \log\left(\binom{n}{t}^2 t \log(n)/\varepsilon\right)} \leq 2^{10\varepsilon \frac{H(n,k,2t+1)}{t \log n}(t \log n + \log t + \log\log n + \log \frac{1}{\varepsilon})}$$
$$\leq 2^{20\varepsilon H(n,k,2t+1)},$$

and $|F| \leq (1+2\varepsilon)H(n,k,2t+1)$ for all $F \in \mathcal{F}$. Since $0 < \varepsilon < 1$ was arbitrary, this proves the lemma. $\square$

3.6 A random version of Katona's intersection theorem

A family $\mathcal{A} \subseteq \mathcal{P}(n)$ is *t-intersecting* if $|A \cap B| \geq t$ for all $A, B \in \mathcal{A}$. In the case when $t = 1$ we simply say that $\mathcal{A}$ is *intersecting*. Two of the most fundamental results in extremal set theory concern t-intersecting sets. The cornerstone theorem of Erdős–Ko–Rado states that for every k, t there exists an $n_0 = n_0(k,t)$ such that if $n \geq n_0$ then the largest t-intersecting k-uniform family is the trivial family, i.e., there is a t-element set which is contained in each of the sets. The other fundamental theorem is Katona's intersection theorem [66], which determines the size $K(n,t)$ of the largest t-intersecting (not necessarily uniform) family in $\mathcal{P}(n)$: it states that

$$K(n,t) = \begin{cases} \binom{n}{\geq (n+t)/2} & \text{if } 2|(n+t); \\ 2\binom{n-1}{\geq (n+t-1)/2} & \text{otherwise.} \end{cases}$$

In the case when $n+t$ is even, $\binom{[n]}{\geq (n+t)/2}$ is a t-intersecting set of size $K(n,t)$. When $n+t$ is odd, $\binom{[n]}{\geq (n+t+1)/2} \cup \binom{[n-1]}{(n+t-1)/2}$ is a t-intersecting set of size $K(n,t)$. Notice that if $t = o(\sqrt{n})$ then $K(n,t) \sim 2^{n-1}$.

Beginning with the work of Balogh, Bohman and Mubayi [5], the problem of developing a 'random' version of the Erdős–Ko–Rado theorem has received significant attention (see [5, 6, 51, 60, 61]). In this section, we raise the analogous question for Katona's intersection theorem. More precisely, let $\mathcal{P}(n,p)$ be the set obtained from $\mathcal{P}(n)$ by selecting elements randomly with probability p and independently of all other choices.

Question 3.6.1. *Suppose that $n \in \mathbb{N}$, $t = t(n) \in \mathbb{N}$ and write $K := K(n,t)$. For which values of p do we have that, with high probability, the largest t-intersecting family in $\mathcal{P}(n,p)$ has size $(1+o(1))pK$?*

The model $\mathcal{P}(n,p)$ was first investigated by Rényi [93] who determined the probability threshold for the property that $\mathcal{P}(n,p)$ is not itself an antichain, thereby answering a question of Erdős. More recently, a random version of Sperner's theorem for $\mathcal{P}(n,p)$ was obtained independently by Balogh, Mycroft and Treglown [10] and by Collares Neto and Morris [27].

In this section we give a precise answer to Question 3.6.1 in the case when $t = o(\sqrt{n})$. For intersecting families (i.e. for the $t = 1$ case), this question has also been resolved independently by Mubayi and Wang [89]. Clearly the conclusion of Question 3.6.1 is not satisfied if $p < C/2^n$ for any constant $C > 0$. The next result implies that the conclusion of Question 3.6.1 is not satisfied for $p = 2^{-\Omega(\sqrt{n}\log n)}$ and $t = o(\sqrt{n})$.

Theorem 3.6.2. *Let $p = 2^{-\Omega(\sqrt{n}\log n)}$ where $p \geq \omega(n)/2^n$ for some function $\omega(n) \to \infty$ as $n \to \infty$, and let $t = o(\sqrt{n})$. Then there exists a constant $\varepsilon > 0$ such that, with high probability, the largest t-intersecting family in $\mathcal{P}(n,p)$ has size at least $(\frac{1}{2}+\varepsilon)2^n p$.*

Proof. The choice of p and t implies that there exists a constant $a > 0$ such that $p < 2^{-a\sqrt{n}\log n}$ and $t < \frac{a}{100}\sqrt{n}$ for n sufficiently large. Define ε so that $0 < \varepsilon \ll a$.

Let $\mathcal{A}$ denote the set of elements A of $\mathcal{P}(n)$ that satisfy $n/2 - a\sqrt{n}/2 \leq |A| \leq n/2 - a\sqrt{n}/4$ and $|A \cap [n/2]| \geq n/4 + t/2$. The latter condition implies that $\mathcal{A}$ is a t-intersecting family.

Claim 3.6.3. $|\mathcal{A}| \geq 4\varepsilon 2^n$.

The claim holds since

$$\begin{aligned}
|\mathcal{A}| &= \sum_{s=a\sqrt{n}/4}^{a\sqrt{n}/2} \sum_{k=0}^{n/4-t/2-s} \binom{n/2}{n/4+t/2+k}\binom{n/2}{n/4-t/2-s-k} \\
&\geq \frac{a\sqrt{n}}{4} \sum_{k=0}^{a\sqrt{n}} \binom{n/2}{n/4+t/2+k}\binom{n/2}{n/4-t/2-a\sqrt{n}/2-k} \\
&\geq \frac{a\sqrt{n}}{4} \cdot a\sqrt{n}\binom{n/2}{n/4+2a\sqrt{n}}\binom{n/2}{n/4-2a\sqrt{n}} \geq 4\varepsilon 2^n,
\end{aligned}$$

where the last inequality follows by applying Facts 3.2.1 and 3.2.2 and since $\varepsilon \ll a$.

Write $\mathcal{A}_{\text{ex}} := \binom{[n]}{\geq n/2+t/2}$. So $\mathcal{A}_{\text{ex}}$ is a t-intersecting set. Since $t = o(\sqrt{n})$ note that $|\mathcal{A}_{\text{ex}}| \geq (1/2-\varepsilon/2)2^n$. As $p \geq \omega(n)/2^n$, by the Chernoff bound for the binomial distribution, we have that, with high probability, $\mathcal{P}(n,p)$ contains at least $(1/2-\varepsilon)p2^n$ elements from $\mathcal{A}_{\text{ex}}$. Denote this set by $\mathcal{A}_{\text{ex},p}$. We will show that, with high probability, we can add a significant number of elements from $\mathcal{A}$ to $\mathcal{A}_{\text{ex},p}$ to obtain a t-intersecting set in $\mathcal{P}(n,p)$ of size at least $(\frac{1}{2}+\varepsilon)2^n p$.

Consider any $A \in \mathcal{A}$. The number of elements $B \in \binom{[n]}{\geq n/2}$ with $|A \cap B| < t$ is

$$\begin{aligned}\sum_{k=0}^{t-1}\binom{|A|}{k}\binom{n-|A|}{\geq n/2-k} \leq \binom{|A|}{t-1}\sum_{k=0}^{t-1}\binom{n-|A|}{\geq n/2-k} &\leq \binom{n/2-a\sqrt{n}/4}{t-1}\sum_{k=0}^{t-1}\binom{n/2+a\sqrt{n}/2}{\geq n/2-k}\\ &= \binom{n/2-a\sqrt{n}/4}{t-1}\sum_{k=0}^{t-1}\binom{n/2+a\sqrt{n}/2}{\leq a\sqrt{n}/2+k}\\ &\leq 2\binom{n/2-a\sqrt{n}/4}{t-1}\binom{n/2+a\sqrt{n}/2}{a\sqrt{n}/2+t}.\end{aligned}$$

Further,

$$2\binom{n/2-a\sqrt{n}/4}{t-1}\binom{n/2+a\sqrt{n}/2}{a\sqrt{n}/2+t} \leq n^t\left(\frac{3\sqrt{n}}{a}\right)^{0.55a\sqrt{n}} \leq 2^{0.6a\sqrt{n}\log n},$$

where in the first inequality we use that $t < a\sqrt{n}/100$ and apply Fact 3.2.3.

Let $\mathcal{A}_p$ denote the set of elements $A \in \mathcal{A}$ that lie in $\mathcal{P}(n,p)$ and where $\mathcal{A}_{\text{ex},p} \cup \{A\}$ is a t-intersecting set. Thus, the probability that $A \in \mathcal{A}$ lies in $\mathcal{A}_p$ is at least

$$p(1-p)^{2^{0.6a\sqrt{n}\log n}}.$$

By X denote the size of the family $\mathcal{A}_p$. By Claim 3.6.3,

$$\mathbb{E}(X) \geq 4\varepsilon 2^n p(1-p)^{2^{0.6a\sqrt{n}\log n}} \geq 4\varepsilon p 2^n(1-p2^{0.6a\sqrt{n}\log n}) \geq 4\varepsilon p2^n(1-2^{-0.4a\sqrt{n}\log n}) \geq 3\varepsilon p 2^n,$$

where the last inequality follows since n is sufficiently large.

Write $\mathcal{A} = \{A_1, \dots, A_m\}$ and $X = \sum_{i=1}^m X_i$ where $X_i = 1$ if $A_i \in \mathcal{A}_p$ and $X_i = 0$ otherwise. Note that the random variables X_i, X_j are not independent if and only if there is some $B \in \mathcal{A}_{\text{ex}}$ such that $|B \cap A_i|, |B \cap A_j| < t$. In this case, $|B| \geq n/2$ and so $|A_i \cup A_j| \leq n/2 + 2t \leq n/2 + a\sqrt{n}/50$. Further, $|A_i|, |A_j| \geq n/2 - a\sqrt{n}/2$ and thus $|A_i \setminus A_j|, |A_j \setminus A_i| \leq a\sqrt{n}$. So given a fixed i, the number of X_js that are not independent with X_i is at most

$$\binom{n}{\leq a\sqrt{n}}\binom{|A_i|}{\leq a\sqrt{n}} \leq 2\binom{n}{a\sqrt{n}}^2 \leq 2\left(\frac{e\sqrt{n}}{a}\right)^{2a\sqrt{n}} < 2^{10a\sqrt{n}\log n}.$$

Write $i \sim j$ to mean that X_i and X_j are not independent. By abusing notation let us also write A_i to denote the event that $A_i \in \mathcal{A}_p$. Consider

$$\Delta := \sum_{i \sim j} \mathbb{P}(A_i \cap A_j).$$

For $A_i, A_j \in \mathcal{A}_p$ we require that $A_i, A_j \in \mathcal{P}(n,p)$ and so $\mathbb{P}(A_i \cap A_j) \leq p^2$. Therefore,

$$\Delta \leq \sum_{i=1}^{m} 2^{10a\sqrt{n}\log n} p^2 \leq 2^n 2^{10a\sqrt{n}\log n} p^2.$$

In particular, $\Delta = o(\mathbb{E}(X)^2)$. Thus, by applying Corollary 4.3.4 from [4] (Chebyshev's inequality) we have that, with high probability, $X \geq 2\varepsilon p 2^n$.

Note that $\mathcal{A}_p \cup \mathcal{A}_{\mathrm{ex},p}$ is a t-intersecting set in $\mathcal{P}(n,p)$ and, with high probability, it has size at least $(1/2+\varepsilon)p2^n$, as required. □

By arguing precisely as in the proof of Theorem 3.6.2 we in fact obtain the following result for $t = O(\sqrt{n})$.

Theorem 3.6.4. *Given any constant $C > 0$, there is a constant $\varepsilon > 0$ such that the following holds. Let $p < 2^{-100C\sqrt{n}\log n}$ where $p \geq \omega(n)/2^n$ for some function $\omega(n) \to \infty$ as $n \to \infty$, and let $t \leq C\sqrt{n}$. Write $K := K(n,t)$. Then there exists a constant $\varepsilon > 0$ such that, with high probability, the largest t-intersecting family in $\mathcal{P}(n,p)$ has size at least $(1+\varepsilon)pK$.*

The following result together with Theorem 3.6.2 resolves Question 3.6.1 for $t = o(\sqrt{n})$.

Theorem 3.6.5. *If $p = 2^{-o(\sqrt{n}\log n)}$ and $t = o(\sqrt{n})$ then with high probability the largest t-intersecting family in $\mathcal{P}(n,p)$ has size $(\frac{1}{2}+o(1))2^n p$.*

Proof. Note that for this range of t, with high probability, the size of the largest t-intersecting family in $\mathcal{P}(n,p)$ is at least $(\frac{1}{2}+o(1))2^n p$. Hence to prove the theorem, it suffices to show that the largest *intersecting* family in $\mathcal{P}(n,p)$ has size at most $(\frac{1}{2}+o(1))2^n p$. That is, it suffices to prove the upper bound in the theorem for $t=1$, since for any $t \geq 2$ any t-intersecting family is also 1-intersecting (i.e. intersecting).

Fix any $\delta > 0$ and define $0 < \varepsilon \ll \gamma \ll \delta$. We will show that with high probability the largest intersecting family in $\mathcal{P}(n,p)$ has size at most $(\frac{1}{2}+\delta)2^n p$.

The first step in the proof is to create a collection of *containers* that house all intersecting families. Define the graph G on vertex set $\mathcal{P}(n)$ where distinct A, B are adjacent in G precisely if $A \cap B = \emptyset$. In order to bound the size of the containers, we require a supersaturation result.

Claim 3.6.6. *If $\mathcal{F} \subseteq \mathcal{P}(n)$ where $|\mathcal{F}| \geq \sum_{k=\frac{n}{2}-C\sqrt{n}}^{n} \binom{n}{k}$ for some constant $C > 0$, then $e(G[\mathcal{F}]) \geq 2^{n+\frac{C}{20}\sqrt{n}\log n}$ and so $\Delta(G[\mathcal{F}]) \geq 2^{\frac{C}{20}\sqrt{n}\log n}$.*

A result of Frankl [44] and Ahlswede [1] implies that, given $|\mathcal{F}|$, the number of edges in $G[\mathcal{F}]$ is minimised if $\mathcal{F}$ consists of the top layers of $\mathcal{P}(n)$, and possibly one partial layer. That is, there are no $A, B \subseteq [n]$ with $|A| < |B|$ and $A \in \mathcal{F}$ but $B \notin \mathcal{F}$.

Hence we may assume that $\mathcal{F}$ consists of the top $\frac{n}{2} + C\sqrt{n} + 1$ layers of $\mathcal{P}(n)$. We will estimate the degrees of vertices in the lowest $C\sqrt{n}/2$ layers of $\mathcal{F}$. The total number of vertices in these layers is at least $\delta_1 2^n$, where $\delta_1 > 0$ is a constant dependent only on C. The degree of each vertex v in these layers is bounded below by

$$\deg_{G[\mathcal{F}]}(v) \geq \binom{n/2 + (C/2)\sqrt{n}}{n/2 - (C/2)\sqrt{n}} \geq 2^{(C\sqrt{n}\log n)/10}.$$

Thus, the number of edges in $G[\mathcal{F}]$ is at least $\delta_1 2^n 2^{(C\sqrt{n}\log n)/10}/2 \geq 2^{n+\frac{C}{20}\sqrt{n}\log n}$, thereby proving the claim.

By applying the graph container algorithm to G with parameter $2^{\varepsilon\sqrt{n}\log n}$, Claim 3.6.6 implies that there is a function $f : \binom{V(G)}{\leq 2^{n-\varepsilon\sqrt{n}\log n}} \to \binom{V(G)}{\leq (1/2+\gamma)2^n}$ such that, for any independent set I in G, there is a subset $S \subseteq I$ where $S \in \binom{V(G)}{\leq 2^{n-\varepsilon\sqrt{n}\log n}}$ and $I \subseteq S \cup f(S)$. Note here we used that

$$\sum_{k=\frac{n}{2}-20\epsilon\sqrt{n}}^{n} \binom{n}{k} \leq \left(\frac{1}{2} + \gamma\right) 2^n.$$

Let $\mathcal{F}$ be the collection of all sets $S \cup f(S)$ for all $S \in \binom{V(G)}{\leq 2^{n-\varepsilon\sqrt{n}\log n}}$. So $|F| \leq (1/2 + 2\gamma)2^n$ for every $F \in \mathcal{F}$. Further, Fact 3.2.3 implies that

$$\begin{aligned} \log|\mathcal{F}| &\leq \log\left(\sum_{a \leq 2^{n-\epsilon\sqrt{n}\log n}} \binom{2^n}{a}\right) \leq \log\left(2\left(\frac{e2^n}{2^{n-\varepsilon\sqrt{n}\log n}}\right)^{2^{n-\varepsilon\sqrt{n}\log n}}\right) \\ &\leq n2^{n-\varepsilon\sqrt{n}\log n} \leq 2^{n-\frac{\varepsilon}{2}\sqrt{n}\log n}. \end{aligned} \tag{3.6.1}$$

Given any $F \in \mathcal{F}$, by the Chernoff bound for the binomial distribution we have that

$$\mathbb{P}\big(|F \cap \mathcal{P}(n,p)| \geq (1/2 + 4\gamma)2^n p\big) \leq 2e^{-\gamma^2 2^n p/2}. \tag{3.6.2}$$

Thus, (3.6.1), (3.6.2) and the choice of p imply that with high probability $|F \cap \mathcal{P}(n,p)| \leq (1/2+\delta)2^n p$ for all $F \in \mathcal{F}$. Since every intersecting family in $\mathcal{P}(n)$ lies in some $F \in \mathcal{F}$, the theorem now follows. $\square$

3.7 Sperner's theorem revisited

3.7.1 Counting antichains in $\mathcal{P}(n)$

Sperner's theorem [101] states that the largest antichain in $\mathcal{P}(n)$ has size $\binom{n}{\lfloor n/2 \rfloor}$. It was Dedekind [31] in 1897 who first attempted to find the total number $A(n)$ of distinct antichains in $\mathcal{P}(n)$.

Since every subset of an antichain is an antichain itself, it follows that $2^{\binom{n}{\lfloor n/2 \rfloor}} \leq A(n)$. The following result of Kleitman determines $A(n)$ up to an error term in the exponent.

Theorem 3.7.1 (Kleitman [71])**.** *The number of antichains in $\mathcal{P}(n)$ is $2^{\binom{n}{\lfloor n/2 \rfloor}(1+o(1))}$.*

For further details on the history of this, and similar questions, we refer the reader to the brilliant survey by Saks [94]. Our first goal in this section is to give an alternative proof of Theorem 3.7.1 using the container method. We will apply the following supersaturation result of Kleitman [70].

Theorem 3.7.2 (Kleitman [70])**.** *Let $\mathcal{A} \subseteq \mathcal{P}(n)$ with $|\mathcal{A}| \geq \binom{n}{\lfloor n/2 \rfloor} + x$. Then $\mathcal{A}$ contains at least $(\lfloor n/2 \rfloor + 1)\, x$ pairs A, B with $A \subset B$.*

For $x \leq \binom{n}{\lfloor n/2 \rfloor + 1}$, Theorem 3.7.2 is easily seen to be optimal, by taking a full middle layer and any x sets on the layer above. Our proof of this supersaturation theorem will make use of the existence of a *symmetric chain decomposition* (or SCD) of $\mathcal{P}(n)$, given first by de Bruijn, Tengbergen and Kruyswijk [21]. An SCD $\mathcal{X}$ is a partition of $\mathcal{P}(n)$ into *symmetric chains*, i.e. chains that for some $k \leq n/2$ consist of precisely one set of each size i between k and $n-k$. The proof we give is very similar to the proof of a more general result from [32].

Proof of Theorem 3.7.2. Without loss of generality we may assume that $\emptyset, [n] \notin \mathcal{A}$. Given any SCD $\mathcal{Z}$ we say that $\mathcal{Z}$ *contains* a bad pair A, B if $A, B \in \mathcal{A}$ and there exists a chain $X \in \mathcal{Z}$ such that $A, B \in X$. Note that $\mathcal{Z}$ is a partition of $\mathcal{P}(n)$ into $\binom{n}{\lfloor n/2 \rfloor}$ chains, hence by the pigeonhole principle $\mathcal{Z}$ contains at least x bad pairs.

Fix some SCD $\mathcal{X}$. Each permutation $\pi \in S_n$ induces a permutation on the subsets of $[n]$ and hence on collections of subsets of $[n]$. In particular, $\pi(\mathcal{X})$ is a SCD. We will pick a random permutation $\pi \in S_n$ and estimate the number of bad pairs contained in $\pi(\mathcal{X})$.

Let $\mathcal{P}$ denote the set of ordered pairs $A, B \in \mathcal{A}$ where $A \subset B$. Consider any $(A, B) \in \mathcal{P}$. If $|B| \geq \lfloor n/2 \rfloor + 1$ define $\delta_A(B) := \{S \subset [n] : S \subset B, |S| = |A|\}$. Otherwise define $\delta_A(B) := \{S \subset [n] : A \subset S, |S| = |B|\}$. Since $A, B \notin \{\emptyset, [n]\}$, in both cases we have $|\delta_A(B)| \geq \binom{\lfloor n/2 \rfloor + 1}{\lfloor n/2 \rfloor}$.

If $|B| \geq \lfloor n/2 \rfloor + 1$, the probability that there is a chain $X \in \pi(\mathcal{X})$ with $S, B \in X$ is the same for all $S \in \delta_A(B)$. So the probability that there is a chain $X \in \pi(\mathcal{X})$ with $A, B \in X$ is at most $\frac{1}{\lfloor n/2 \rfloor + 1}$. Similarly if $|B| \leq \lfloor n/2 \rfloor$, the probability that there is a chain $X \in \pi(\mathcal{X})$ with $A, B \in X$ is at most $\frac{1}{\lfloor n/2 \rfloor + 1}$. Thus, the expected number of bad pairs in $\pi(\mathcal{X})$ is at most $|\mathcal{P}|/(\lfloor n/2 \rfloor + 1)$. On the other hand, as $\pi(\mathcal{X})$ is a SCD there are at least x bad pairs in $\pi(\mathcal{X})$. Hence, $|\mathcal{P}| \geq (\lfloor n/2 \rfloor + 1)\, x$, as desired. $\square$

Now Theorem 3.7.1 follows from an easy application of the container method.

Proof of Theorem 3.7.1. Let $\varepsilon > 0$ and let G be the graph with vertex set $\mathcal{P}(n)$ where A and B are adjacent precisely if $A \subset B$ or $B \subset A$. By applying the graph container algorithm to G with parameter $\varepsilon n/10$, Theorem 3.7.2 implies that we obtain a function $f : \binom{V(G)}{\leq 10 \cdot 2^n/\varepsilon n} \to \binom{V(G)}{\leq (1+\varepsilon)\binom{n}{\lfloor n/2 \rfloor}}$ such that, for any independent set I in G, there is a subset $S \subseteq I$ where $S \in \binom{V(G)}{\leq 10 \cdot 2^n/\varepsilon n}$ and $I \subseteq S \cup f(S)$.

Let $\mathcal{F}$ be the collection of all sets $S \cup f(S)$ for all $S \in \binom{V(G)}{\leq 10 \cdot 2^n/\varepsilon n}$. Then

$$|\mathcal{F}| \leq \binom{2^n}{\leq 10 \frac{2^n}{\epsilon n}} \leq 2^{20 \frac{2^n}{\epsilon n} \log n} = 2^{o\left(\binom{n}{\lfloor n/2 \rfloor}\right)}.$$

Further, every antichain is an independent set in G and therefore lies in some element of $\mathcal{F}$ and $|F| \leq (1 + 2\varepsilon)\binom{n}{\lfloor n/2 \rfloor}$ for every $F \in \mathcal{F}$. The existence of $\mathcal{F}$ immediately proves the theorem. $\square$

Let $\mathcal{F} \subseteq \mathcal{P}(n)$, and for $i \in [n]$, let B_i denote the number of comparable pairs $A, B \in \mathcal{F}$ with $|B \backslash A| = i$ and let $B_{\geq i} := B_i + B_{i+1} + \ldots + B_n$. Then by arguing as in the proof of Theorem 3.7.2 we get the following innocent-looking proposition, that eventually led us to the proof of the main results of Chapter 2.

Proposition 3.7.3. *Let $n, N, x \in \mathbb{N}$. Suppose $\mathcal{F} \subseteq \mathcal{P}(n)$ where $|\mathcal{F}| = \binom{n}{n/2} + x$ and for all $A \in \mathcal{F}$ we have $N \leq |A| \leq n - N$. Then*

$$\frac{B_{\geq N}}{\binom{\lfloor n/2 \rfloor + \lceil N/2 \rceil}{N}} + \sum_{k=1}^{N-1} \frac{B_k}{\binom{\lfloor n/2 \rfloor + \lceil k/2 \rceil}{k}} \geq x.$$

3.7.2 A two-coloured generalisation of Theorem 3.7.1

Now we turn our attention to a two-coloured generalisation of Sperner's theorem, which was discovered independently by Katona [68] and Kleitman [67]. Given a (two-)colouring of $[n]$, we say that a pair of sets $A, B \in \mathcal{P}(n)$ is *comparable with monochromatic difference* if $A \subset B$, and the difference $B \backslash A$ is monochromatic.

Theorem 3.7.4 (Katona [68], Kleitman [67])**.** *Let $\mathcal{A} \subseteq \mathcal{P}(n)$, and let $R \cup W$ be a partition of $[n]$ (i.e. a two-colouring of $[n]$ using colours Red and White). If $\mathcal{A}$ does not contain a pair of sets which are comparable with monochromatic difference then $|\mathcal{A}| \leq \binom{n}{\lfloor n/2 \rfloor}$.*

Note that setting $R = \emptyset$ in Theorem 3.7.4 gives the classical Sperner theorem. We will consider the following question: given the partition $R \cup W$, how many families are there without two comparable sets whose difference is monochromatic? Alternatively, how many families are there for which there exists a partition $R \cup W$ such that there are no comparable pairs with monochromatic difference? (The answers to these two questions are at most a factor of 2^n apart.) To answer this question using the container method, first we need a supersaturation result.

Lemma 3.7.5. *Let $\varepsilon > 0$ and n be sufficiently large. Given a partition $R \cup W = [n]$ and a family $\mathcal{F} \subset \mathcal{P}(n)$ of size $|\mathcal{F}| \geq (1+\varepsilon)\binom{n}{n/2}$, there are at least $\varepsilon \binom{n}{n/2} \frac{n^{3/4}}{4}$ comparable pairs $A, B \in \mathcal{F}$ with monochromatic difference.*

We note that the factor $n^{3/4}/4$ is far from optimal and indeed the proof below can easily be strengthened to some function of the form $n^{1-o(1)}$ instead of the $n^{3/4}$. The number of such pairs is probably at least $\varepsilon\binom{n}{n/2}\left(\lfloor \frac{n}{2} \rfloor + 1\right)$ (if $n > 5$). We could not prove this, but fortunately this weaker result suffices to prove the counting theorem later (in fact one could replace the $n^{3/4}$ by anything bigger than $c\sqrt{n}\log n$ and the calculations would still go through, with a worse $o(1)$ error term in the final result).

Proof of Lemma 3.7.5. Suppose first that $|R| \leq n^{3/4}$. For each $S \subseteq R$, let $\mathcal{F}_S := \{A \subseteq W : A \cup S \in \mathcal{F}\}$. Given any pair of sets A, B with $A \subset B$ and $A, B \in \mathcal{F}_S$ we can find a comparable pair $A \cup S, B \cup S$ in $\mathcal{F}$ with monochromatic difference. Thus, Theorem 3.7.2 implies that the number of comparable pairs in $\mathcal{F}$ with monochromatic difference is at least

$$\sum_{S \subseteq R} \left(|\mathcal{F}_S| - \binom{n-|R|}{(n-|R|)/2} \right) \left(\frac{n-|R|}{2} + 1 \right) \geq \left((1+\varepsilon)\binom{n}{n/2} - 2^{|R|}\binom{n-|R|}{(n-|R|)/2} \right) \frac{n}{3}$$
$$\geq \left((1+\varepsilon)\binom{n}{n/2} - \left(1+\frac{\varepsilon}{2}\right)\binom{n}{n/2} \right) \frac{n}{3} = \varepsilon \binom{n}{n/2} \frac{n}{6},$$

as required. Note that in the penultimate inequality we used Fact 3.2.1 and that $|R| = o(n)$.

We may therefore assume that $|R| \geq n^{3/4}$ and $|W| \geq n^{3/4}$. Remove all monochromatic elements of $\mathcal{F}$, and all those elements of $\mathcal{F}$ that contain the entire set R or the entire set W as subsets. The number of such sets is at most $\frac{\varepsilon}{2}\binom{n}{n/2}$. Following the original proof of Theorem 3.7.4, let $\mathcal{B}_W$ and $\mathcal{B}_R$ be SCDs of $\mathcal{P}(W)$ and $\mathcal{P}(R)$. Let the group $G = S_{|W|} \times S_{|R|}$ act on $R \cup W$ in the natural way, permuting the elements in the two sets. From now on for simplicity we shall refer to comparable pairs in $\mathcal{F}$ with monochromatic difference simply as *bad pairs*. We say that the pair of SCDs $\mathcal{B}_R$, $\mathcal{B}_W$ *contains* the bad pair (A, B) if there exist chains $X \in \mathcal{B}_R$ and $Y \in \mathcal{B}_W$ such that Y contains $(A \cap W, B \cap W)$ and X contains $(A \cap R, B \cap R)$.

Let $x := \frac{\varepsilon}{2}\binom{n}{n/2}$. We first show that every pair of SCDs contains at least x bad pairs. This follows instantly from the original proof of Theorem 3.7.4: suppose on the contrary, we could find a pair of SCDs $\mathcal{B}_R$ and $\mathcal{B}_W$ and a family $\mathcal{A} \subset \mathcal{F}$ of size $|\mathcal{A}| = \binom{n}{n/2} + 1$ such that the pair $\mathcal{B}_R, \mathcal{B}_W$ does not contain any bad pairs from $\mathcal{A}$. If a pair of chains $(X, Y) \in \mathcal{B}_R \times \mathcal{B}_W$ does not contain any bad pairs, then the number of sets A such that Y contains $A \cap W$ and X contains $A \cap R$ is at most $\min\{|X|, |Y|\}$. So if $X_1 \subset \ldots \subset X_t$ is a chain in $\mathcal{B}_R$ and $Y_1 \subset \ldots \subset Y_s$ is a chain in $\mathcal{B}_W$ then $\mathcal{A}$ contains at most $\min\{s, t\}$ sets of the form $X_i \cup Y_j$, which is also the number of sets of this form having size exactly $\lfloor n/2 \rfloor$. Hence $\sum_{X \in \mathcal{B}_R, Y \in \mathcal{B}_W} \min\{|X|, |Y|\} = \binom{n}{n/2}$ because both sides count the number of subsets of $[n]$ of size $\lfloor n/2 \rfloor$. Thus for every subfamily $\mathcal{A} \subseteq \mathcal{F}$ with $|\mathcal{A}| = \binom{n}{n/2} + 1$ there exists a pair of chains $X \in \mathcal{B}_R$, $Y \in \mathcal{B}_W$ containing a bad pair from $\mathcal{A}$, and the claim follows.

Choose a random element $\pi \in G$. We claim the probability that $\pi(\mathcal{B}_R, \mathcal{B}_W)$ contains a given bad pair is at most $2/n^{3/4}$. To see this, let (A, B) be a bad pair. Without loss of generality we may assume $A \subset B$ and $B \backslash A \subseteq R$. This implies that $B \cap W = A \cap W$, and since $A \neq B$ we have $A \cap R \neq B \cap R$. The probability that (A, B) is contained in the pair $\pi(\mathcal{B}_R, \mathcal{B}_W)$ of SCDs is equal to the probability that $(A \cap R, B \cap R)$ is contained in $\pi_R(\mathcal{B}_R)$ (where π_R denotes the restriction of π to the set R). We removed the monochromatic elements of $\mathcal{F}$ and those that contain R, hence $A \cap R, B \cap R \notin \{\emptyset, R\}$. Hence, defining $\delta_A(B)$ and applying the shadow argument as in the proof of Theorem 3.7.2, we get that the probability that $\pi(\mathcal{B}_R, \mathcal{B}_W)$ contains (A, B) is at most $\max\{2/|R|, 2/|W|\} \leq 2/n^{3/4}$, as claimed. Putting the last two paragraphs together, we obtain that there are at least $\varepsilon\binom{n}{n/2}\frac{n^{3/4}}{4}$ bad pairs, as required. $\square$

Now a simple application of the container method, exactly as in the proof of Theorem 3.7.1, yields a counting version of Theorem 3.7.4.

Theorem 3.7.6. *The number of families $\mathcal{F}$ for which there exists a colouring $R \cup W = [n]$ such that there is no comparable pair $A, B \in \mathcal{F}$ with monochromatic difference is $2^{\binom{n}{n/2}(1+o(1))}$.*

Proof. Fix a colouring R, W with $R \cup W = [n]$, and define a graph G on vertex set $\mathcal{P}(n)$ where two comparable sets are adjacent if their difference is monochromatic. Hence families without comparable pairs with monochromatic difference correspond to independent sets in G. Arguing precisely as in the proof of Theorem 3.7.1, we find that the number of independent sets, and hence the number of families without comparable sets with monochromatic difference, is $2^{\binom{n}{n/2}(1+o(1))}$. There are 2^n possible colourings to start with, hence the number of families for which there exists a colouring avoiding comparable sets with monochromatic difference is $2^n \cdot 2^{\binom{n}{n/2}(1+o(1))} = 2^{\binom{n}{n/2}(1+o(1))}$ as required. □

3.8 Counting maximal independent sets and antichains in the Boolean lattice

Most of this chapter dealt with finding $\alpha(n)$, the number of families in $\mathcal{P}(n)$ satisfying some given property. We applied different variations of the container method to obtain asymptotics for $\log \alpha(n)$. The reader might be curious whether it is possible to obtain precise asymptotics for $\alpha(n)$ using these (or different) methods. In general though this seems to be a much more difficult task. For example, in the problem we consider below, it is difficult to even make a firm guess on the asymptotics.

For a graph G, we say an independent set I of G is *maximal* if for any $v \in V(G) \setminus I$, we have that $I \cup \{v\}$ is not independent. Let $\text{mis}(G)$ denote the number of maximal independent sets in G. Most of the problems discussed below have their origins in [33]. Let $\mathcal{B}_{n,k}$ be the graph on vertex set $\binom{[n]}{k} \cup \binom{[n]}{k+1}$, and edges given by inclusion. Ilinca and Kahn [64] proved that $\log_2 \text{mis}(\mathcal{B}_{n,k}) = (1+o(1))\binom{n-1}{k}$. They also made the following sharp conjecture.

Conjecture 3.8.1 (Ilinca and Kahn [64])**.**

$$mis(\mathcal{B}_{n,k}) = (1 + o(1))n2^{\binom{n-1}{k}},$$

where the $o(1)$ term goes to 0 as $n \to \infty$.

The natural lower bound in Conjecture 3.8.1 follows by defining for each $i \in [n]$ an induced matching M_i in $\mathcal{B}_{n,k}$ of size $\binom{n-1}{k}$ where the edges of the matching are of the form $(B, B \cup i)$ for $B \in \binom{[n]\setminus\{i\}}{k}$. Each of the $2^{|M_i|}$ sets containing precisely one vertex from each edge in M_i extends to a maximal independent set, and each extension is different. This produces $2^{\binom{n-1}{k}}$ distinct maximal independent sets. By considering each M_i for $i \in [n]$ we obtain a list of $n2^{\binom{n-1}{k}}$ maximal independent sets, containing not too many repetitions.

It turns out however, that this construction can be tweaked to obtain significantly more maximal independent sets.

Proposition 3.8.2. *If* $|k - n/2| \leq \sqrt{n}$ *then*

$$mis(\mathcal{B}_{n,k}) = \Omega\left(n^{3/2} 2^{\binom{n-1}{k}}\right).$$

Proof strategy: We say a triple $\mathcal{T} = (B, r, s)$ is *good* if $r, s \in [n]$, $B \in \binom{[n]}{k}$, $1, r \notin B$, $r \neq 1$ and $s \in B$. For each good triple $\mathcal{T}$ we will construct a collection $f(\mathcal{T})$ of independent sets in $\mathcal{B}_{n,k}$ with

$$|f(\mathcal{T})| = 2^{\binom{n-1}{k}-n+1}.$$

For each good triple $\mathcal{T}$ we will extend all elements of $f(\mathcal{T})$ to a maximal independent set in an arbitrary way. We will show that every maximal independent set in $\mathcal{B}_{n,k}$ is obtained at most twice in this way. The number of good triples is $\Omega\left(\binom{n-1}{k}n^2\right)$, hence a simple calculation will complete the proof.

Proof. For a set C with $i \notin C$ and $j \in C$, let $C^i := C \cup \{i\}$ and $C_j := C \setminus \{j\}$. We define, for example, $C_j^{i,k}$ analogously (assuming $i, k \notin C$ and $j \in C$). Let M be the induced matching in $\mathcal{B}_{n,k}$ of size $\binom{n-1}{k-1}$ given by $M := \{(C, C^1) : 1 \notin C,\ |C| = k\}$.

Given a good triple $\mathcal{T} = (B, r, s)$, let $U(\mathcal{T}) := \{C : |C| = k+1, B_s^1 \subset C\}$ and $D(\mathcal{T}) := \{C : |C| = k, C \subset B^r\}$. Let $e := (B, B^1)$ and $f := (B_s^r, B_s^{1,r})$; notice these are two edges of M. Note that $|U(\mathcal{T})| = n - k$ and $|D(\mathcal{T})| = k + 1$. Every vertex of $D(\mathcal{T}) \cup U(\mathcal{T})$ is incident to precisely one edge of M. Moreover the only two edges that are simultaneously incident to one vertex in $D(\mathcal{T})$ and one vertex in $U(\mathcal{T})$ are e and f. The collection of independent sets $f(\mathcal{T})$ is defined as follows.

- Let B^r and B_s^1 be elements of the independent set.
- If an edge of M is not incident to any vertex in $D(\mathcal{T}) \cup U(\mathcal{T})$ then put exactly one endpoint of that edge into the independent set.
- If an edge of M is incident to precisely one vertex of $D(\mathcal{T}) \cup U(\mathcal{T})$, choose the other vertex of this edge.

Note that this gives us $\binom{n-1}{k-1} - (n-k) - k + 1$ free choices, hence $|f(\mathcal{T})| = 2^{\binom{n-1}{k}-n+1}$ as claimed. Every such set constructed is independent. Indeed, this follows since M is an induced matching, there is no edge between B^r and B^1_s, and there is no edge between B^r or B^1_s and another vertex in the set. Let $\mathcal{F}$ be the union of the $f(\mathcal{T})$s over all good triples.

Every such constructed independent set contains precisely one vertex from each edge in M except for precisely two edges (e, f from above). These two edges lie in a unique 6-cycle in $\mathcal{B}_{n,k}$ (e, f together with B^1_s and B^r), hence given any $I \in \mathcal{F}$, there are precisely two good triples giving rise to this I. Specifically, if $I \in \mathcal{F}$ arises from a good triple $\mathcal{T} = (B, r, s)$, the only other good triple that 'produces' I is $\mathcal{T}' = (B^r_s, s, r)$.

Moreover, if $I, I' \in \mathcal{F}$ where $I \neq I'$ then I and I' lie in different maximal independent sets in $\mathcal{B}_{n,k}$. Hence a maximal independent set of $\mathcal{B}_{n,k}$ is counted twice by $|\mathcal{F}|$ if it intersects M in $|M| - 2$ vertices, and not counted otherwise.

The number of good triples is

$$\binom{n-1}{k} k(n-k-1) = \Omega\left(\frac{2^n}{\sqrt{n}} n^2\right) = \Omega\left(n^{3/2} 2^n\right),$$

and the result follows.

□

In the above argument we started with a good triple and modified the independent sets from the Ilinca–Kahn construction along a 6-cycle determined by the triple. But we can get a better lower bound by starting out with a collection S of $N > 1$ good triples, as long as the sets in the triples are sufficiently far apart (Hamming distance at least 20, say) so that the modifications do not interfere with each other. Each maximal independent set is then counted at most 2^N times, and as long as N is not too large the number of choices for the N triples is at least

$$\left(\frac{n}{10}\right)^{2N} \binom{\binom{n-1}{k-1}}{N} \geq \left(\frac{n}{10}\right)^{2N} \left(\frac{2^n}{C_1 N \sqrt{n}}\right)^N \geq 2^{nN} \left(n^{3/2} N^{-1} C_2\right)^N,$$

for some absolute constants $C_1, C_2 > 0$. Each good triple decreases the number of free choices on edges of M by at most n, hence costing us a factor of 2^n. So by setting $N = n^{3/2} C_2 / 2$ we conclude the following result, which is an exponential improvement over Proposition 3.8.2:

Proposition 3.8.3. *There exists an absolute constant $C > 0$ such that whenever k, n are such that $|k - n/2| \leq \sqrt{n}$ then*

$$mis(\mathcal{B}_{n,k}) \geq 2^{\binom{n-1}{k-1} + Cn^{3/2}}.$$

□

We do not have any reason to believe that Proposition 3.8.3 gives the correct order of magnitude of $\mathrm{mis}(\mathcal{B}_{n,k})$; It would be extremely interesting to determine this. However, we suspect this may be very difficult.

Finally we note that the above result also disproves another conjecture from [64]. Write $\mathrm{ma}(\mathcal{P}(n))$ for the number of maximal antichains in $\mathcal{P}(n)$. Ilinca and Kahn [64] conjectured that $\mathrm{ma}(\mathcal{P}(n)) = \Theta\left(n2^{\binom{n-1}{\lfloor n/2 \rfloor}}\right)$. However, since $\mathrm{ma}(\mathcal{P}(n)) \geq \mathrm{mis}(\mathcal{B}_{n,k})$ for all k, Proposition 3.8.3 disproves this conjecture.

3.9 The number of metric spaces

The results of this section are based on [14]. The following result is an application of hypergraph containers, but the main difficulty of the proof is the set-up and the formulation of the right tsupersaturation statement. Hence this section can be regarded as a warm-up for the next chapter, where we delve deeper into the topic of hypergraph containers.

Our goal is to estimate the number of metric spaces on n points, where the distance between any two points lies in $\{1, \dots, r\}$ for some $r = r(n)$. This problem was considered first by Kozma–Meyerovitch–Peled–Samotij [78], who, using the regularity lemma gave an asymptotic bound on the number of such metric spaces for a fixed constant r. Recently, Mubayi–Terry [88] provided a characterisation of the typical structure of such metric spaces for a fixed constant r, while $n \to \infty$. We will be more interested in what happens if r is allowed to grow as a function of n. Our main result is the following:

Theorem 3.9.1. *Fix an arbitrary small constant $\epsilon > 0$. If*

$$r = O\left(\frac{n^{1/3}}{\log^{\frac{4}{3}+\epsilon} n}\right),$$

then the number $|\mathcal{M}_n^r|$of such metric spaces is

$$|\mathcal{M}_n^r| = \left\lceil \frac{r+1}{2} \right\rceil^{\binom{n}{2}+o(n^2)}.$$

Kozma–Meyerovitch–Peled–Samotij [78] pointed out that the discrete and the continuous problems are related. They considered the same question in the continuous case with distances in $[0, 1]$. Their entropy based approach yields Theorem 3.9.1 for $r < n^{1/8}$. The same set of authors have recently announced [78] an almost optimal estimate, showing that the number of such metric spaces satisfies

$$|\mathcal{M}_n^r| \le \left(\left(\frac{1}{2} + \frac{2}{r} + \frac{C}{\sqrt{n}}\right) r\right)^{\binom{n}{2}}.$$

At the end of this section we show how our results translate to the continuous setting.

For a positive integer r define $m(r) = \lceil \frac{r+1}{2} \rceil$. We will use an easy corollary of Mubayi–Terry ([88] Lemma 4.9):

Lemma 3.9.2. *Let $A, B, C \subset [r]$, all non-empty. Suppose the triple $\{A, B, C\}$ does not contain a non-metric triangle – that is, every triple $\{a, b, c : a \in A, b \in B, c \in C\}$ satisfies the triangle-inequality. Then if r is even we have $|A| + |B| + |C| \le 3m(r)$, and if r is odd we have $|A| + |B| + |C| \le 3m(r) + 1$.*

Let $\mathcal{H}$ be the 3-uniform hypergraph with vertex set r rows, one for each color, and $\binom{n}{2}$ columns, one for each edge of K_n. A vertex (i, f) of $\mathcal{H}$ corresponds to the event that the graph edge f has color i. Three vertices of $\mathcal{H}$ form a hyperedge when the graph edge coordinates of the vertices form a triangle in K_n while the 'colors' do not satisfy the triangle inequality. With other words, the hyperedges correspond to non-metric triangles, and independent sets having exactly one vertex from each column correspond to points of the metric polytope. Our plan is to prove a supersaturation statement, but first we need two lemmas.

The first lemma we use is due to Füredi [47]. For a graph G, write G^2 for the "proper square" of G, i.e., where xy is an edge if and only if there is a z such that xz and zy are edges in G. Write $e(G)$ for the number of edges in G.

Lemma 3.9.3. *For any graph G with n vertices, we have*

$$e(G^2) \ge e(G) - \lfloor n/2 \rfloor.$$

The second lemma bounds the size of the largest independent set in $\mathcal{H}$.

Lemma 3.9.4. *Let $S \subset V(\mathcal{H})$ have no empty columns and contain no edges in $\mathcal{H}$. Then if r is even we have $|S| \le m(r)\binom{n}{2}$, and if r is odd we have $|S| \le m(r)\binom{n}{2} + rn$.*

Proof. The even case follows directly from Lemma 3.9.2, and we note that this bound is tight – let S contain the interval $[r/2, r]$ from each column. The bound in the odd case is slightly more difficult, and we make no effort to establish a tight bound, which should probably be $|S| \le m(r)\binom{n}{2} + n/2$.

Let r be odd, and let A, B, C be three columns that form a triangle of S. Note that if for some $k \geq 1$ we have $|A| \geq m(r) + k$ and $|B| \geq m(r) + k$ then $|C| \leq m(r) - 2k + 1 \leq m(r) - k$ by Lemma 3.9.2. Write B_k for the set of columns in S of order at least $m(r) + k$ and write S_k for the set of columns in S of order at most $m(r) - k$. Let G_k be the graph on $[n]$ with edges B_k. Then by Lemma 3.9.3 we get

$$|S_k| \geq e(G_k^2) \geq e(G_k) - \lfloor n/2 \rfloor \geq |B_k| - n.$$

Hence

$$|S| - m(r)\binom{n}{2} = \sum_{k=1}^{r} k(|B_k| - |B_{k+1}|) - \sum_{k=1}^{r} k(|S_k| - |S_{k+1}|) = \sum_{k=1}^{r}(|B_k| - |S_k|) \leq nr,$$

and the result follows. ☐

Now we are ready to prove a supersaturation-like result.

Lemma 3.9.5. *Let $\epsilon > 0$ and let $S \subset V(\mathcal{H})$ with no empty columns.*

1. *If r is even and $|S| \geq (1+\epsilon)\binom{n}{2}m(r)$, then S contains at least $\frac{\epsilon}{10}\binom{n}{3}$ hyperedges.*
2. *If r is odd, $n > n_\epsilon$ sufficiently large and $|S| \geq (1+\epsilon)\binom{n}{2}m(r)$, then S contains at least $\frac{\epsilon^4}{40000}\binom{n}{3}$ hyperedges.*

Proof. Suppose first that r is even. Then there are at least $\frac{\epsilon}{10}\binom{n}{3}$ triangles in G such that the corresponding columns contain at least $(1+\epsilon/10)3m(r)$ vertices from S. Indeed, if this was not the case, then

$$\frac{\frac{\epsilon}{10}\binom{n}{3}3r + \binom{n}{3}3m(r)(1+\epsilon/10)}{n-2} = \binom{n}{2}\left(m(r) + \frac{\epsilon(r + m(r))}{10}\right) > |S|,$$

which is a contradiction. Hence part 1 of the lemma follows from Lemma 3.9.2.

Now suppose r is odd. Given $T \subset [n]$, write $f_S(T)$ for the set of vertices of S contained in the $\binom{|T|}{2}$ columns of $\mathcal{H}$ corresponding to the edges spanned by T. Set $n_0 = 20/\epsilon$, so that by Lemma 3.9.4, whenever $T \subset [n]$ with $|T| = n_0$ and $|f_S(T)| \geq m(r)\binom{n_0}{2}(1+\frac{\epsilon}{3})$ then $\mathcal{H}[f_S(T)]$ contains a hyperedge.

First, we claim that there are at least $\frac{\epsilon}{4}\binom{n}{n_0}$ choices of $T \subset [n]$ with $|T| = n_0$ and $|f_S(T)| \geq m(r)\binom{n_0}{2}(1+\frac{\epsilon}{3})$.

Indeed, if this was not the case, then we would have

$$\begin{aligned} |S| &\leq \frac{\frac{\epsilon}{4}\binom{n}{n_0}\binom{n_0}{2}r + \binom{n}{n_0}m(r)\binom{n_0}{2}(1+\frac{\epsilon}{3})}{\binom{n-2}{n_0-2}} = \binom{n}{2}\left(m(r) + \frac{\epsilon r}{4} + \frac{\epsilon m(r)}{3}\right) \\ &< \binom{n}{2}m(r)(1+\epsilon), \end{aligned} \tag{3.9.1}$$

which is not possible. So the number of hyperedges contained in S is at least

$$e(\mathcal{H}[S]) \geq \frac{\epsilon}{4}\binom{n}{n_0}/\binom{n-3}{n_0-3} \geq \frac{\epsilon}{4}\frac{\epsilon^3}{20^3}\binom{n}{3},$$

and the result follows. □

Write $\bar{d}$ for the average degree of $\mathcal{H}$, and for $j \in [3]$ define the j-th maximum co-degree

$$\Delta_j = \max\{|\{e \in E(\mathcal{H}) : \sigma \subset e\}| : \sigma \subset V(\mathcal{H}) \text{ and } |\sigma| = j\}.$$

Below, we will make use of a version of the container theorem of [8, 98], the way it was formulated by Mousset–Nenadov–Steger [87].

Theorem 3.9.6. *There exists a positive integer c such that the following holds for every positive integer N. Let $\mathcal{H}$ be a 3-uniform hypergraph of order N. Let $0 \leq p \leq 1/(3^6c)$ and $0 < \alpha < 1$ be such that $\Delta(\mathcal{H}, p) \leq \alpha/(27c)$, where*

$$\Delta(\mathcal{H}, p) = \frac{4\Delta_2}{\bar{d}p} + \frac{2\Delta_3}{\bar{d}p^2}.$$

Then there exists a collection of containers $\mathcal{C} \subset \mathcal{P}(V(\mathcal{H}))$ such that

(i) every independent set in $\mathcal{H}$ is contained in some $C \in \mathcal{C}$,

(ii) for all $C \in \mathcal{C}$ we have $e(\mathcal{H}[C]) \leq \alpha e(\mathcal{H})$, and

(iii) the number of containers satisfies

$$\log|\mathcal{C}| \leq 3^9 c(1 + \log(1/\alpha))Np\log(1/p).$$

Proof of Theorem 3.9.1. Let $\mathcal{H}$ be the hypergraph defined earlier, i.e., the 3-uniform hypergraph with vertex set formed by pairs of the r colors and the $\binom{n}{2}$ edges of K_n, with 3-edges corresponding to non-metric triangles. Let $\epsilon, \delta > 0$ be arbitrarily small constants and set $p = \frac{1}{r\log^{2+\delta} n}$ and $\alpha = \frac{10^{10}c\log^{4+2\delta} n}{n}$. In $\mathcal{H}$ we have $\Delta_1 \leq nr^2,\ \Delta_2 \leq r,\ \Delta_3 = 1,\ \bar{d} \geq r^2n/64$ and

$$\Delta(\mathcal{H}, p) \leq 4\left(\frac{64r^2\log^{2+\delta} n}{r^2n} + \frac{64r^2\log^{4+2\delta} n}{2r^2n}\right) \leq \frac{\alpha}{27c}.$$

Then Theorem 3.9.6 provides containers with

$$e(\mathcal{H}[C]) \leq \alpha e(\mathcal{H}) \leq 10^4 cr^3n^2\log^{4+2\delta} n,$$

and the number of containers is

$$\log|\mathcal{C}| \leq \frac{c3^{10}rn^2 \cdot \log n \cdot \log r \cdot \log\log n}{r\log^{2+\delta} n} = o(n^2).$$

Now assume

$$r = o\left(\frac{n^{1/3}}{\log^{(4+2\delta)/3} n}\right).$$

Then the maximum number of edges in a container is $o(n^3)$, hence by Lemma 3.9.5, and the fact that a useful container does not have an empty column, we have for n large enough,

$$|V(C)| < (1+\epsilon)m(r)\binom{n}{2}.$$

Hence, the number of colourings in a container is at most $(1+\epsilon)^{\binom{n}{2}}m(r)^{\binom{n}{2}} = m(r)^{\binom{n}{2}+o(n^2)}$. The logarithm of the number of containers is $o(n^2)$. The total number of good colourings is at most the number of containers times the maximum number of colourings in a container. Hence the total number of good colourings is

$$m(r)^{\binom{n}{2}+o(n^2)},$$

as required. □

Now we turn our attention to the continuous setting. The set-up in [78] is as follows. Given a metric space with n points and all distances being in $[0,1]$, we regard the set of distances as a vector in $[0,1]^{\binom{n}{2}}$. We will call the union of all such n points in $[0,1]^{\binom{n}{2}}$ for all finite metric spaces the *metric polytope* M_n. More precisely, the metric polytope M_n is the convex polytope in $\mathbb{R}^{\binom{n}{2}}$ defined by the inequalities $0 < d_{ij} \leq 1$ and $d_{ij} \leq d_{ik} + d_{jk}$.

Note that if $a+b \geq c$ then

$$\lceil a \rceil + \lceil b \rceil \geq \lceil c \rceil. \tag{3.9.2}$$

Theorem 3.9.7. *Fix $\delta > 0$ constant. Then for $n > n_\delta$ sufficiently large, we have*

$$(vol(M_n))^{1/\binom{n}{2}} \leq \frac{1}{2} + \frac{1}{n^{\frac{1}{6}-\delta}}.$$

Proof. First consider the discrete setting, colouring with r colours, where r is the even integer closest to $n^{\frac{1}{6}-\frac{\delta}{2}}$. W.l.o.g. $\delta < 1/4$ and set

$$1/p = n^{\frac{1}{3}-\frac{\delta}{4}}, \qquad \alpha = 300cn^{\delta-\frac{2}{3}},$$

where c is the constant from Theorem 3.9.6. Then

$$\Delta(\mathcal{H}, p) < 300\left(\frac{1}{rnp} + \frac{1}{p^2r^2n}\right) \leq 300\left(\frac{n^{1/3-\delta/4}}{n^{7/6-\delta/2}} + \frac{n^{2/3-\delta/2}}{n^{4/3-\delta}}\right) \leq \alpha,$$

and we get containers with

$$e(\mathcal{H}[C]) \leq \alpha n^3 r^3 \leq n^{3-1/6-\delta/4},$$

where the number of containers satisfies

$$\log |\mathcal{C}| \leq n^2 rp \log^3 n \leq n^{2-1/6-\delta/5}.$$

Hence, by Lemma 3.9.5, the number of vertices in a container is at most

$$|V(C)| \leq (1 + n^{-1/6-\delta/6})\binom{n}{2} m(r).$$

This implies that the number of colourings contained in a container is at most

$$\mathrm{col}(C) \leq \left(\frac{V(C)}{\binom{n}{2}}\right)^{\binom{n}{2}} \leq \left((1+n^{-1/6-\delta/6})m(r)\right)^{\binom{n}{2}} \leq m(r)^{\binom{n}{2}} e^{n^{2-1/6-\delta/7}}.$$

That is, the total number X of colourings is at most

$$X \leq m(r)^{\binom{n}{2}} e^{n^{2-1/6-\delta/7}} e^{n^{2-1/6-\delta/5}} \leq m(r)^{\binom{n}{2}} e^{n^{2-1/6-\delta/8}}.$$

Now consider colourings in the continuous setting. Cut up each edge of the cube into r pieces. We get by

(3.9.2) that

$$\begin{aligned}(\mathrm{vol}(M_n))^{1/\binom{n}{2}} &\leq \left(\frac{(m(r)+1)^{\binom{n}{2}} e^{n^{2-1/6-\delta/8}}}{r^{\binom{n}{2}}}\right)^{1/\binom{n}{2}} \\ &\leq \left(2^{-\binom{n}{2}}\left(1+\frac{4}{r}\right)^{\binom{n}{2}}\right)^{1/\binom{n}{2}}\left(1+\frac{1}{n^{1/6+\delta/9}}\right) \leq \frac{1}{2}+\frac{1}{n^{\frac{1}{6}-\delta}},\end{aligned}$$

as required. $\square$

Chapter 4

The number of union-free families

The results in this chapter are joint work with József Balogh [15].

A family of sets is union-free if there are no three distinct sets in the family such that the union of two of the sets is equal to the third set. Kleitman proved that every union-free family has size at most $(1+o(1))\binom{n}{n/2}$. Later, Burosch–Demetrovics–Katona–Kleitman–Sapozhenko asked for the number $\alpha(n)$ of such families, and they proved that $2^{\binom{n}{n/2}} \leq \alpha(n) \leq 2^{2\sqrt{2}\binom{n}{n/2}(1+o(1))}$. They conjectured that the constant $2\sqrt{2}$ can be removed in the exponent of the right hand side. We prove their conjecture by formulating a new container-type theorem for rooted hypergraphs.

4.1 Introduction

Let $\mathcal{P}(n)$ denote the family consisting of all subsets of $[n]$. Given a family $\mathcal{F} \subseteq \mathcal{P}(n)$, we say that $\mathcal{F}$ is *union-free* if there are no three distinct sets $A, B, C \in \mathcal{F}$ such that $A \cup B = C$. Kleitman was a young professor at Brandeis in the early 1960s when he stumbled across a book of open mathematical problems by Ulam [105]. The problem of determining the largest union-free family, originally raised by Erdős, appeared in this book. Erdős conjectured that no union-free family could have size larger than $O\left(\binom{n}{\lfloor n/2 \rfloor}\right)$. Kleitman [72] proved this conjecture by establishing an upper bound of $2\sqrt{2}\binom{n}{\lfloor n/2 \rfloor}$, which he later improved to $(1+o(1))\binom{n}{\lfloor n/2 \rfloor}$.

Theorem 4.1.1 (Kleitman)**.** *If $\mathcal{F} \subseteq \mathcal{P}(n)$ is a union-free family, then $|\mathcal{F}| \leq (1+o(1))\binom{n}{\lfloor n/2 \rfloor}$.*

Later, Burosch–Demetrovics–Katona–Kleitman–Sapozhenko raised [22] the problem of enumerating all union-free families in $\mathcal{P}(n)$. Let $\alpha(n) = |\{\mathcal{F} \subseteq \mathcal{P}(n) : \mathcal{F} \text{ is union-free}\}|$. Since the collection of all $\lfloor n/2 \rfloor$-sets gives rise to a union-free family, and every subfamily of a union-free family is also union-free, we have $\alpha(n) \geq 2^{\binom{n}{\lfloor n/2 \rfloor}}$. They proved the following upper bound on $\alpha(n)$:

Theorem 4.1.2 (Burosch–Demetrovics–Katona–Kleitman–Sapozhenko)**.** *The function $\alpha(n)$ satisfies*

$$2^{\binom{n}{\lfloor n/2 \rfloor}} \leq \alpha(n) \leq 2^{2\sqrt{2}\binom{n}{\lfloor n/2 \rfloor}(1+o(1))}.$$

They conjectured that the constant $2\sqrt{2}$ in the exponent of the right hand side can be removed. The main result of this chapter is that their conjecture was correct:

Theorem 4.1.3. *The function $\alpha(n)$ satisfies*

$$\alpha(n) = 2^{\binom{n}{\lfloor n/2 \rfloor}(1+o(1))}.$$

Our main tool in proving Theorem 4.1.3 is the Hypergraph Container Method, pioneered by Balogh–Morris–Samotij [8] and independently by Saxton–Thomason [98]. We create a 3-uniform hypergraph $\mathcal{H}$ with vertex set $\mathcal{P}(n)$, and sets A, B, C forming an edge if $A \cup B = C$. Now every union-free family corresponds to an independent set in $\mathcal{H}$. Hence to prove Theorem 4.1.3 we will use the Container Method to bound the number of independent sets in this hypergraph $\mathcal{H}$. The idea behind the method is that there exists a small family of vertex sets, called *containers*, which consists of sets spanning only few hyperedges, and each independent set is contained in one of them.

The main difficulty in this problem compared to the main results in [8, 98] is that here $\mathcal{H}$ does not satisfy any of the necessary co-degree conditions – it has large subgraphs where the co-degrees are comparable to the total number of edges – hence straightforward applications of the available container theorems are doomed to fail. To get around this difficulty we need a new version of the container theorem, that works well for *rooted* hypergraphs. For this theorem to be applicable one needs to prove a nonstandard version of a supersaturation theorem. The proof of the supersaturation theorem makes use of the Expander Mixing Lemma of Alon–Chung [2]. Note that in general for the container method to work one needs some type of supersaturation, which means that if vertex set U is a somewhat larger than the independence number of the hypergraph, then U contains many hyperedges. Here we need a little bit more, we need some even distribution of these hyperedges, a similar obstacle (which was handled differently) showed up in [86].

This chapter is organised as follows. In Section 4.2 we prove a new version of the Container Theorem for 3-uniform hypergraphs. In Section 4.3 we prove a supersaturated version of Theorem 4.1.1 and in Section 4.4 we combine it with our container theorem to prove Theorem 4.1.3.

4.2 Constructing containers in rooted hypergraphs

Definition 4.2.1. *A 3-uniform hypergraph $\mathcal{H}$ is* ***rooted*** *if there exists a function $f : E(\mathcal{H}) \to V(\mathcal{H})$ such that*

- *for every edge $e \in E(\mathcal{H})$ we have $f(e) \in e$, and*

- *for any two vertices u, v there is at most one edge $e \in E(\mathcal{H})$ with $u, v \in e$ and $f(e) \notin \{u, v\}$.*

If $\mathcal{H}$ is a rooted hypergraph and f is specified then we call f a ***rooting function for*** *$\mathcal{H}$ and for every edge e we call $f(e)$ the* ***head*** *of e. The* ***head-degree*** *of a vertex v is $\mathrm{hd(v)} = |\{\mathrm{e} \in \mathrm{E}(\mathcal{H}) : \mathrm{f(e)} = \mathrm{v}\}|$. The* ***head link-graph*** *of a vertex v is the graph $\mathrm{HL_v}(\mathcal{H})$ with vertex set $V(\mathcal{H})$ and edge set $\{\{u_1u_2\} : \{vu_1u_2\} \in E(\mathcal{H}), f(\{vu_1u_2\}) = v\}$.*

Definition 4.2.2. *Given a 3-uniform hypergraph $\mathcal{H}$, a subset of its vertex set $\mathcal{X} \subseteq V(\mathcal{H})$ and two positive numbers s, t, we say that a vertex $v \in \mathcal{X}$ is $(\mathcal{X}, s, t)$-**eligible** if there is a subgraph G_v of its head link graph $\mathrm{HL_v}(\mathcal{H} \cap \mathcal{X})$ with $\Delta(G_v) \leq s$ and $e(G_v) \geq t$. We say that $\mathcal{X}$ is an (s, t)-**core** if it does not contain an $(\mathcal{X}, s, t)$-eligible vertex.*

*Given $\varepsilon > 0$ and $N > 0$ we say that the hypergraph $\mathcal{H}$ is (ε, N, s, t)-**nice** if for every $\mathcal{X} \subseteq V(\mathcal{H})$ with $|\mathcal{X}| \geq (1+\varepsilon)N$, the set $\mathcal{X}$ contains an $(\mathcal{X}, s, t)$-eligible vertex.*

Some explanation might come in handy. Throughout this chapter we will mostly work with the hypergraph $\mathcal{H}$ which has vertex set $\mathcal{P}(n)$ and edge set $\{(A, B, C) : A \cup B = C\}$. We prove a container theorem for general hypergraphs, but it does no harm for the reader to think of this $\mathcal{H}$ throughout the proof. This hypergraph is rooted, since whenever $A \cup B = C$ we can let $f(A, B, C) = C$. The crucial observation is that given A, C there may be many choices for B such that $A \cup B = C$ holds (so the codegrees of the hypergraph can be very large), but given A, B there is only one C such that $A \cup B = C$ (hence in this direction, all codegrees are one). All our approaches using existing container lemmas broke down because $\mathcal{H}$ has such large codegrees - but breaking the symmetry and proving an oriented- (or rooted) version of the container lemma fixes the problem.

There is another difficulty that arises when one tries to prove a rooted container lemma - during the proof it is much harder to keep control over the degrees of the link graphs, when we reduce from 3-uniform to 2-uniform. To overcome this difficulty we need a stronger, balanced supersaturation result. A simple supersaturation result (that is not good enough for us) states that if a family has size slightly larger than the largest independent set, then it contains many edges, and hence it contains a vertex of large degree within the family. In the present chapter we will show that such a family contains a vertex that not only has large degree, but in fact one can find a dense subgraph of its link graph that is nicely distributed. Using the terms defined above, our stronger supersaturation result will show that $\mathcal{H}$ is *nice* (with some parameters), i.e. that if a family is slightly larger than the largest independent set then it is not a *core*, so it contains an *eligible* vertex (again, parameters specified later). For more details, we direct the reader to Section 4.3.

The main goal of this section is to prove the following Container Theorem. Let $H : [0,1] \to \mathbb{R}$ be the binary entropy function defined as

$$H(p) = -p\log p - (1-p)\log(1-p).$$

Theorem 4.2.3. *[Container theorem for rooted 3-uniform hypergraphs] Let $\varepsilon, s, t, N, M > 0$ be parameters satisfying*

$$\varepsilon \le 1/10, \qquad 8s \le \varepsilon t, \qquad \frac{1}{\varepsilon^2} \le s \qquad \text{and} \qquad M \ge (1+100\varepsilon)N.$$

Let $\mathcal{H}$ be a 3-uniform rooted M-vertex hypergraph $\mathcal{H}$ such that there exists a rooting function f for $\mathcal{H}$ so that $\mathcal{H}$ is (ε, N, s, t)-nice. Then there exists a family $\mathcal{C} \subseteq \mathcal{P}(V(\mathcal{H}))$ satisfying the following:

1. *For every independent set $I \subseteq V(\mathcal{H})$, there exists a $C_I \in \mathcal{C}$ such that $I \subseteq C_I$.*
2. $\log_2 |\mathcal{C}| \le \frac{2M}{\varepsilon}(H(2s/t) + H(1/4\varepsilon s))$.
3. *Every $C \in \mathcal{C}$ satisfies $|C| \le (1+100\varepsilon)N$.* □

Remark. 1. If $\mathcal{H}$ is (ε, N, s, t)-nice where $s < (8t/\varepsilon)^{1/2}$, then $\mathcal{H}$ is $(\varepsilon, N, (8t/\varepsilon)^{1/2}, t)$-nice as well, and the theorem yields a smaller family $\mathcal{C}$ if s is replaced by $(8t/\varepsilon)^{1/2}$.

2. Even though we will not need it in the present chapter, we remark that condition (2) can be replaced by the following, stronger condition, typical for container type lemmas. Define p to be the least integer such that $(1-\varepsilon/2)^p M \le N$.

2.' For every independent set $I \subseteq V(\mathcal{H})$, there exist sets $Q_I = (Q_1, Q_2, \ldots, Q_p)$ and $R_I = (R_1, R_2, \ldots, R_p)$ with $Q_i, R_i \subseteq I$ for all i, and sizes $|Q_i| \le 2sM(1-\varepsilon/2)^i/t$ and $|R_i| \le M(1-\varepsilon/2)^i/4\varepsilon s$. Moreover, the container C_I depends only on the pair (Q_I, R_I). (Note: these sets are usually referred to as *fingerprints* (or *certificates*, *tokens*) in the literature).

The following algorithm, which is useful when the codegrees can be high but the hypergraph is rooted, will be used to produce the containers:

Container algorithm

Input: Parameters $\varepsilon, s, t, N, \tau, z > 0$ satisfying

$$\tau \geq \frac{2s}{t}, \qquad \varepsilon \leq \frac{1}{10}, \qquad 4\varepsilon s \geq z, \qquad \text{and} \quad \tau + \frac{1}{z} \leq \frac{\varepsilon}{2},$$

a 3-uniform rooted hypergraph $\mathcal{H}$ on at least $(1+100\varepsilon)N$ vertices, a rooting function f such that $\mathcal{H}$ is (ε, N, s, t)-nice, and an independent set $I \subseteq V(\mathcal{H})$.

Output: A set C such that $I \subseteq C \subseteq V(\mathcal{H})$, and two sets $T, T' \subseteq I$.

Phase I:

1. Fix an arbitrary ordering of the vertices, and another arbitrary ordering of all graphs on vertex set $V(\mathcal{H})$. These will be used to break ties.
2. Set $A = V(\mathcal{H})$ (the set of *available* vertices), $T = \emptyset$ be the first *fingerprint* (or *token*) of I, and let L be the empty graph on vertex set $V(\mathcal{H})$ (the *link graph* we build). The set C will only depend on T and T'.
3. If $|A| \leq (1-\varepsilon)|V(\mathcal{H})|$ then set $C := A \cup T$ and STOP. Otherwise, set $S := \{u \in A : |N_L(u) \cap A| \geq s\}$.
4. If $A \setminus S$ is an (s,t)-core then go to Phase II.
5. Let $v \in V(\mathcal{H}[A \setminus S])$ be the largest degree vertex among $(A \setminus S, s, t)$-eligible vertices (break ties according to the ordering fixed in Step 1). If $v \notin I$ then replace A by $A \setminus \{v\}$, replace L by $L \setminus \{v\}$ and return to Step 3.
6. We have $v \in I$ and v is $(A \setminus S, s, t)$-eligible. Let G_v be a subgraph of its head link graph $\mathrm{HL}_v(\mathcal{H}) \cap A \setminus S$ with $\Delta(G_v) \leq s$ and $e(G_v) \geq t$ (break ties according to the ordering fixed in Step 1).
   (a) Set $T := T \cup \{v\}$.
   (b) Set $A := A \setminus \{v\}$.
   (c) Let $L := (L \cup G_v) \setminus \{v\}$. Note that since the input graph $\mathcal{H}$ was rooted, every pair of vertices forms an edge in the head link graph of at most one other vertex, hence L is always a simple graph.
   (d) Return to Step 3.

Phase II:

1. Initiate $T' = \emptyset$, the second *fingerprint*.

2. Let v be the largest degree vertex in L (break ties according to the ordering fixed in Step I.1). If $d_L(v) < z$ then set $C := A \cup T \cup T'$ and STOP.

3. If $v \notin I$ then replace A by $A \setminus \{v\}$ and replace L by $L \setminus \{v\}$ and go to Step 2.

4. We have $v \in I$ and $d_L(v) \geq z$. Set $T' := T' \cup \{v\}$, replace A by $A \setminus (N_L(v) \cup \{v\})$ and replace L by $L \setminus (N_L(v) \cup \{v\})$. Go to Step 2.

End of algorithm.

Observation 4.2.4. *The containers only depend on the fingerprints T, T'.*

Proof. Person A runs the algorithm with input I and gets output C, T, T'. He then tells Person B the values of T, T' and all other input parameters (including the orderings specified in Step 1), but not I. We claim that B can find the value of C. Indeed, all he has to do is to try to follow the algorithm exactly as A did. In Phase I, the only critical points are in Steps 5 and 6 where B seems to need knowledge of I to make the same decisions as A did. But actually all B needs to know is whether the vertex v is in I or not. But for this v we know that $v \in T$ iff $v \in I$, hence B can run Phase I the same way as A did (and hence at every point in Phase I B will know the values of L, S, A, etc.).

The same argument applies to Phase II. The largest degree vertex v that we consider in the algorithm is in I precisely if it gets put into the fingerprint T', hence B can recover C. $\square$

Observation 4.2.5. *At every point in the algorithm, $\Delta(L) \leq 2s$.*

Proof. Every time we change L, we add to it a graph of maximum degree at most s. But as soon as some vertex gets degree at least s we put it in S and do not touch it until its degree goes below s again. Hence in L, the maximum possible degree is at most $s + s = 2s$. $\square$

Observation 4.2.6. *[Small fingerprints.] After the algorithm stops we have $|T| \leq \tau|V(\mathcal{H})|$ and $|T'| \leq |V(\mathcal{H})|/z$.*

Proof. Suppose we have $|T| > \tau|V(\mathcal{H})|$ at some point in the algorithm. Stop the algorithm when this happens (in step $6(d)$) and count the edges in L. Every time we increase T we add at least t edges to L. The only times we delete edges from L is when we remove some vertices from $A \setminus S$. But these vertices had by definition at most s neighbours in L. Hence in total we remove at most $|V(\mathcal{H})|s$ edges from L, and so $e(L) >$

$\tau|V(\mathcal{H})|t - |V(\mathcal{H})|s$. By Observation 4.2.5, we have $\Delta(L) \leq 2s$, so $s|V(L)| \geq e(L) > \tau|V(\mathcal{H})|t - |V(\mathcal{H})|s$ whence it follows that $2s/t > \tau$ (since $L \subseteq \mathcal{H}$). As in the input we took $\tau \geq 2s/t$ we conclude that $|T| < \tau|V(\mathcal{H})|$ at every point in the algorithm.

If the algorithm stopped in Phase I then $T' = \emptyset$ and the claim follows. Otherwise, every time we put a vertex into T' we removed at least z vertices from A. Hence $|T'| \leq |V(\mathcal{H})|/z$ and the claim follows. □

Observation 4.2.7. *After the algorithm stops we have $|A| \leq (1-\varepsilon)|V(\mathcal{H})|$.*

Proof. If the algorithm terminated in Phase I then the claim follows. Now assume the algorithm entered Phase II. Denote A_1, S, L the sets A, S and graph L right before entering Phase II of the algorithm, and let A_2 be the set A at the end of the algorithm (noting that $A_2 \subseteq A_1$). Since we did not terminate after Phase I, we know that $|A_1| \geq (1-\varepsilon)|V(\mathcal{H})|$. It was specified in the input that $|V(\mathcal{H})| \geq (1+100\varepsilon)N$ and $\varepsilon \leq 1/10$, hence we conclude that

$$|A_1| \geq (1 + 89\varepsilon)N. \tag{4.2.1}$$

Since $\mathcal{H}$ is (ε, N, s, t)-nice, we have

$$|A_1 \setminus S| \leq (1+\varepsilon)N, \tag{4.2.2}$$

otherwise $A_1 \setminus S$ would not be an (s,t)-core. Since $\varepsilon \leq 1/10$ we also have

$$\frac{|S|}{|A_1|} \overset{(4.2.2)}{\geq} 1 - \frac{(1+\varepsilon)N}{|A_1|} \overset{(4.2.1)}{\geq} 1 - \frac{1+\varepsilon}{1+89\varepsilon} > 8\varepsilon. \tag{4.2.3}$$

Since $|A_1| \leq |V(\mathcal{H})|$, if it is the case that $|A_2| \leq (1-\varepsilon)|A_1|$ then we are done. Hence in what follows, we assume for contradiction that $|A_2| > (1-\varepsilon)|A_1|$, implying

$$|A_1 \setminus A_2| < \varepsilon|A_1|. \tag{4.2.4}$$

As in $L[A_2]$ every vertex has degree at most z, we get

$$e(L[A_2]) \leq \frac{|A_2|z}{2} \leq \frac{|A_1|z}{2}. \tag{4.2.5}$$

Recall that in L, every vertex has degree at most $2s$. Hence, counting those edges in L which have at least one endpoint in $A_1 \setminus A_2$ we get

$$e(L[A_1]) - e(L[A_2]) \leq |A_1 \setminus A_2| 2s \overset{(4.2.4)}{<} \varepsilon|A_1|2s. \tag{4.2.6}$$

Note also that in L, every vertex in S has degree at least s. As $e(L) = e(L[A_1])$, we have

$$e(L) \geq \frac{|S|s}{2}. \tag{4.2.7}$$

Putting the relations (4.2.5), (4.2.6) and (4.2.7) together we get

$$\frac{|S|s}{2} \leq e(L) = e(L(A_1)) \leq e(L(A_2)) + \varepsilon|A_1|2s < \frac{|A_1|z}{2} + \varepsilon|A_1|2s \overset{(4.2.3)}{\leq} \frac{|A_1|z}{2} + \frac{|S|s}{4}. \tag{4.2.8}$$

Hence

$$8\varepsilon \overset{(4.2.3)}{<} \frac{|S|}{|A_1|} \overset{(4.2.8)}{<} \frac{2z}{s},$$

which contradicts the restriction $4\varepsilon s \geq z$ on the input parameters. This completes the proof. $\square$

Observation 4.2.8. *[Small containers.] After the algorithm stops we have* $|C| \leq \left(1 - \frac{\varepsilon}{2}\right)|V(\mathcal{H})|$.

Proof. If the algorithm stopped after Phase I then $|C| \leq |A| + |T|$, and if the algorithm stopped after Phase II we get $|C| \leq |A| + |T| + |T'|$. In both cases, we have $|C| \leq (1-\varepsilon)|V(\mathcal{H})| + \tau|V(\mathcal{H})| + |V(\mathcal{H})|/z$ by Observations 4.2.6 and 4.2.7. Since $\tau + 1/z \leq \varepsilon/2$ we have $|C| \leq \left(1 - \frac{\varepsilon}{2}\right)|V(\mathcal{H})|$ as required. $\square$

Putting all these observations together, we get the following container lemma. We use the notation $\binom{M}{\leq m} = \sum_{i=0}^{m}\binom{M}{i}$.

Lemma 4.2.9. *[Container lemma for rooted 3-uniform hypergraphs] Let $\varepsilon, s, t, N, M > 0$ be parameters satisfying*

$$\varepsilon \leq 1/10, \qquad 8s \leq \varepsilon t, \qquad \frac{1}{\varepsilon^2} \leq s \qquad \text{and} \quad M \geq (1+100\varepsilon)N.$$

Let $\mathcal{H}$ be a 3-uniform rooted M-vertex hypergraph $\mathcal{H}$ such that there exists a rooting function f for $\mathcal{H}$ so that $\mathcal{H}$ is (ε, N, s, t)-nice. Then there exists a family $\mathcal{C} \subseteq \mathcal{P}(V(\mathcal{H}))$ satisfying the following:

1. *For every independent set $I \subseteq V(\mathcal{H})$, there exists a $C \in \mathcal{C}$ such that $I \subseteq C$.*
2. $|\mathcal{C}| \leq \binom{M}{\leq 2sM/t}\binom{M}{\leq M/4\varepsilon s}$.
3. *Every $C \in \mathcal{C}$ satisfies $|C| \leq (1-\varepsilon/2)n$.*

Proof. Setting $\tau = 2s/t$ and $z = 4\varepsilon s$, the claim follows from Observations 4.2.4 - 4.2.8. $\square$

We will obtain our main container theorem by iterating Lemma 4.2.9. Recall the following standard bound on the sum of binomial coefficients that for $\zeta \le 1/2$ and M

$$\binom{M}{\le \zeta M} \le 2^{H(\zeta)M}. \tag{4.2.9}$$

Proof of Theorem 4.2.3. The key observation that makes this proof work is that if $\mathcal{H}$ is (ε, N, s, t)-nice then for any $S \subseteq V(\mathcal{H})$ we have that $\mathcal{H}[S]$ is (ε, N, s, t)-nice. Hence we can iterate Lemma 4.2.9 to obtain a family $\mathcal{C}$, all of whose members have sizes less than $(1+100\varepsilon)N$.

Set

$$\tau = 2s/t, \qquad \beta = 1/(4\varepsilon s) \qquad \text{and} \quad \gamma = 1 - \varepsilon/2.$$

The size of $\mathcal{C}$ satisfies, for some N' with $N < N' < N(1+100\varepsilon)$

$$\begin{aligned}|\mathcal{C}| &\le \binom{M}{\le \tau M}\binom{M}{\le \beta M} \cdot \binom{M\gamma}{\le \tau M\gamma}\binom{M\gamma}{\le \beta M\gamma} \cdot \binom{M\gamma^2}{\le \tau M\gamma^2}\binom{M\gamma^2}{\le \beta M\gamma^2} \cdot \ldots \cdot \binom{N'}{\le \tau N'}\binom{N'}{\le \beta N'} \\ &\overset{(4.2.9)}{\le} 2^{(H(\tau)+H(\beta))\left(M+M\gamma+M\gamma^2+\ldots+N'\right)} \le 2^{2M(H(\tau)+H(\beta))/\varepsilon},\end{aligned}$$

and the result follows. $\square$

4.3 A supersaturated version of Theorem 4.1.1

In this section we will consider a family $\mathcal{F}$ of size slightly larger than the maximum size of a union-free family, say $|\mathcal{F}| = (1+\varepsilon)\binom{n}{\lfloor n/2 \rfloor}$. Then by Theorem 4.1.1 we know that $\mathcal{F}$ contains a triple A, B, C with $A \cup B = C$. With more work one can prove that $\mathcal{F}$ contains at least $\varepsilon' n^2 \binom{n}{n/2}$ such triples, where ε' is a constant depending on ε. Note also that the factor n^2 cannot be improved to $n^{2+\alpha}$ for some constant $\alpha > 0$, as if $\mathcal{F}$ is contained in the middle two layers of $\mathcal{P}(n)$ then every element in $\mathcal{F}$ has at most n subsets in $\mathcal{F}$, and hence for fixed C the equation $A \cup B = C$ has at most n^2 solutions.

Unfortunately this supersaturation is not quite strong enough for us - we not only want to find many triples in $\mathcal{F}$, but we want to find a large subset of such triples that is nicely distributed. Let $\mathcal{H}$ be the 3-uniform hypergraph on vertex set $\mathcal{P}(n)$, three sets A, B, C forming an edge with head C if $C = A \cup B$. We want to prove that $\mathcal{H}[\mathcal{F}]$ contains at least one $(\mathcal{F}, n, \varepsilon' n^2)$-eligible vertex. (Note that it is then an immediate corollary that $\mathcal{H}[\mathcal{F}]$ contains at least $\varepsilon'' n^2 \binom{n}{n/2}$ edges for some $\varepsilon'' > 0$.)

Theorem 4.3.1. *Let $0 < \varepsilon < 1/200$ be a small constant and n sufficiently large. If $|\mathcal{F}| \ge \binom{n}{\lfloor n/2 \rfloor}(1+\varepsilon)$*

then $\mathcal{H}[\mathcal{F}]$ *contains at least one* $(\mathcal{F}, n, \frac{\varepsilon^2}{10^{40}}n^2)$*-eligible vertex.*

The first ingredient in the proof is the Expander Mixing Lemma, due to Alon and Chung [2]:

Theorem 4.3.2 (Expander Mixing Lemma)**.** *Let* G *be a* D*-regular graph on* N *vertices, and let* λ *be its minimum eigenvalue. Then for all* $S \subseteq V(G)$,

$$e(G[S]) \geq \frac{D}{2N}|S|^2 + \frac{\lambda}{2N}|S|(N - |S|).$$

Denote KG(m, k) the Kneser graph with vertex set $\binom{[m]}{k}$, two k-sets being connected by an edge if the sets are disjoint. Then KG(m, k) is D-regular with $D = \binom{m-k}{k}$ and its minimum eigenvalue $\lambda = -\frac{k}{m-k}D$ (see [84]). Let $N = \binom{m}{k} = |V(KG(m,k))|$. The following is a corollary of the Expander Mixing Lemma.

Lemma 4.3.3. *Given* $\beta > 0$*, any set* S *of at least* $(1+\beta)\binom{m-1}{k-1}$ *vertices in* KG(m, k) *induces at least* $\left(1 - \frac{1}{1+\beta}\right)\frac{Dm}{N(m-k)}\binom{|S|}{2}$ *edges.* □

We will need the following easy lemma to take care of families which are densely packed on the middle layers of $\mathcal{P}(n)$. Families which are more spread out will be much harder to handle. Recall that $\mathcal{H}$ is the 3-uniform hypergraph with vertex set $\mathcal{P}(n)$, and sets A, B, C forming an edge if $A \cup B = C$. If for two sets A, B we have $A \subseteq B$ or $B \subseteq A$ then we call (A, B) a *comparable pair*. If $\mathcal{F}$ is a fixed family in $\mathcal{P}(n)$ and $A \in \mathcal{F}$ is any element of $\mathcal{F}$, then for all $i \in [n]$ we write $B_i(A) = \{B \in \mathcal{F} : B \subseteq A, |A \backslash B| = i\}$.

Lemma 4.3.4. *Let* $0 < \delta < 1/10$, $n > n_0(\delta)$ *sufficiently large,* $\mathcal{F} \subseteq \mathcal{P}(n)$, $k \in \{1, 2, \ldots, 10\}$ *and* $A \in \mathcal{F}$ *with* $n - \sqrt{n \log n} < 2|A| < n + \sqrt{n \log n}$*. Suppose* $|B_k(A)| \geq \delta n^k$*. Then* A *is* $\left(\mathcal{F}, n, \delta^2 n^2\right)$*-eligible.*

Proof. If $k = 1$ then any two sets $C_1, C_2 \in B_1(A)$ satisfy $C_1 \cup C_2 = A$. As $|B_1(A)| \leq |A| < n$, the claim follows. So now assume $k \geq 2$.

Let G be the graph on vertex set $V(G) = B_k(A)$, two sets C_1, C_2 being connected by an edge in G if $C_1 \cup C_2 = A$. Note that C_1 and C_2 are connected by an edge in G iff $(A \setminus C_1) \cap (A \setminus C_2) = \emptyset$. We want to estimate the number of edges in G using Lemma 4.3.3, hence we define the graph G' on vertex set $V(G') = \{A \setminus B : B \in \mathcal{F},\ B \subseteq A,\ |A \setminus B| = k\}$, with edges connecting two sets precisely if they are disjoint. Note that by the above remark, the graphs G and G' are isomorphic.

Let δ' be defined by $\delta n^k = \delta' \binom{|A|}{k}$, hence $\delta' \geq \delta$. Define β by

$$1 + \beta = \frac{|V(G')|}{\binom{|A|-1}{k-1}} \geq \delta' \frac{|A|}{k} \geq \delta \frac{|A|}{k}.$$

Now we can apply Lemma 4.3.3 to conclude that the number of edges in G' is at least

$$e(G') \geq \left(1 - \frac{k}{\delta|A|}\right) \frac{\binom{|A|-k}{k}|A|}{\binom{|A|}{k}(|A|-k)} \binom{|V(G')|}{2}.$$

Choosing $n \gg \delta^{-1}$ we get that $\frac{k}{\delta|A|} < 1/2$, $\binom{|A|-k}{k}/\binom{|A|}{k} \geq 1/2$ and $|A|/(|A|-k) \geq 1/2$. Hence if n is sufficiently large we get

$$e(G') \geq \frac{1}{8}\binom{|V(G')|}{2}.$$

Now let $\mathbf{X}$ be a random n-vertex subgraph of G'. The expected number of edges in $G'[\mathbf{X}]$ is at least $\binom{n}{2}/8$, hence there exists a subgraph G'' of G' with $|V(G'')| = n$ and $e(G'') \geq \binom{n}{2}/8$, and the claim follows. $\square$

At the very end of the supersaturation proof, we will make use of the following embedding lemma.

Lemma 4.3.5. *Let m be a positive integer and let G be a graph with $|V(G)| \geq m$. Let $S \subseteq G$ be the set of the m largest degree vertices in G. Suppose $e(G \setminus S) \geq m^2$. Then there exists a subgraph $H \subseteq G$ such that $\Delta(H) \leq m$ and $e(H) \geq m^2/2$.*

Proof. If $\Delta(G \setminus S) \leq m$ then $H = G \setminus S$ will satisfy the claim. Now assume $\Delta(G \setminus S) \geq m$. Then for each $v \in S$ we have $d_G(v) \geq m$, as S was the collection of the largest degree vertices. We will build H in m steps. Initially, let H be the empty graph on vertex set $V(H) = V(G)$ and let $S = \{s_1, s_2, \ldots, s_m\}$. In step i, let $N_i = N(s_i) \setminus \{s_1, s_2, \ldots, s_{i-1}\}$ and let $N_i' \subset N_i$ be any subset with $|N_i'| = m - i + 1$. For each $v \in N_i'$ add the edge $(s_i v)$ to H.

The algorithm finishes in m steps, and we have added a total of $m(m+1)/2$ edges to H. Each vertex not in S receives at most one edge each step, hence their degree never goes above m. A vertex $s_i \in S$ receives at most one edge in steps $1, 2, \ldots, i-1$, it receives $m - i + 1$ edges in step i, and none after. Hence the graph H constructed this way satisfies all conditions. $\square$

Now we are ready to prove a supersaturated version of Theorem 4.1.1. The beginning of the proof follows Kleitman's proof [72] of Theorem 4.1.1. At the point where he used the Erdős–Ko–Rado Theorem to give an upper bound on some intersecting family (that we shall define later) we will instead use the Expander Mixing Lemma to show that in our family, that is too large to be intersecting, there are many disjoint pairs. However, we will observe that solutions to the equation $A = B \cup C$ correspond to many such disjoint pairs, hence in the second half of the proof we have to show that despite overcounting we can still find many different solutions to the equation. The constant 10 in the statement of Lemma 4.3.4 and in the proof of Theorem 4.3.1 is simply a sufficiently large constant for all our estimates to work.

Proof of Theorem 4.3.1. Let $\Delta = \frac{n}{2} - \sqrt{n \log n}$, $\quad \Delta_1 = \frac{n}{2} - \frac{1}{2}\sqrt{n \log n}$ and $\Delta_2 = \frac{n}{2} + \frac{1}{2}\sqrt{n \log n}$. Note that

$$|\{A \in \mathcal{P}(n) : |A| \geq \Delta_2 \text{ or } |A| \leq \Delta_1\}| = o\left(\binom{n}{\lfloor n/2 \rfloor}\right),$$

hence setting $\mathcal{F}_1 = \{A \in \mathcal{F} : \Delta_1 \leq |A| \leq \Delta_2\}$, for n sufficiently large we have

$$|\mathcal{F}_1| \geq (1 + \varepsilon/2)\binom{n}{\lfloor n/2 \rfloor}.$$

Replace $\mathcal{F}$ by a subset of $\mathcal{F}_1$ of size $(1+\varepsilon/2)\binom{n}{\lfloor n/2 \rfloor}$, and so from now on we will work with a family of size $(1+\varepsilon/2)\binom{n}{\lfloor n/2 \rfloor}$ with all members having size between Δ_1 and Δ_2, that we still call $\mathcal{F}$.

Given a permutation $\Pi \in S_n$ and a set $A \in \mathcal{F}$, we say the pair (Π, A) is *good* if

(i) The elements of A form the first $|A|$ elements of Π;

(ii) For every $B \subset A$, if B is an initial segment of Π then $B \notin \mathcal{F}$.

We say a pair (Π, A) is *bad* if condition (i) above holds, but (ii) does not. We say a bad pair (Π, A) is *horrible* if there is a $B \in \mathcal{F}$ with $B \subset A$, B is an initial segment of Π and $|A \backslash B| \geq 11$.

Now fix a set $A \in \mathcal{F}$. We will say that Π is bad/horrible if (Π, A) is a bad/horrible pair. Let

$$\mathcal{H}_A = \{C \in \mathcal{P}(n) : |C| = \Delta,\ C \text{ is the set of the first } \Delta \text{ elements in a horrible permutation } \Pi\}.$$

So although C does not lie in $\mathcal{F}$, we have that $C \subseteq B \subsetneq A$ for some $B \in \mathcal{F}$, with $|A \backslash B| \geq 11$. Define α_A by the equation

$$|\mathcal{H}_A| = \binom{|A|}{\Delta}\alpha_A. \tag{4.3.1}$$

Note that $\alpha_A \geq 0$ for all $A \in \mathcal{F}$. Note that if for any $k \in \{1, 2, \ldots, 10\}$ we have $|B_k(A)| \geq \frac{\varepsilon}{10^{20}} n^k$ then we are done by Lemma 4.3.4, hence we may assume this is not the case.

Claim 4.3.6. *There exists an $A^* \in \mathcal{F}$ such that $\alpha_{A^*} \geq \varepsilon/20$.*

Proof of claim. Let S_A be the number of bad permutations for the set A. Then we have

$$\begin{aligned} \frac{S_A}{(n-|A|)!} &\leq \frac{\varepsilon n}{10^{20}}\left((|A|-1)! + n(|A|-2)! \cdot 2! + \ldots + n^9(|A|-10)! \cdot 10!\right) + |\mathcal{H}_A|\Delta!(|A|-\Delta)! \\ &\leq \frac{\varepsilon}{100}|A|! + |\mathcal{H}_A|\Delta!(|A|-\Delta)!. \end{aligned} \tag{4.3.2}$$

Since every permutation is in at most one good pair, we get that

$$\sum_{A\in\mathcal{F}} (|A|!(n-|A|)! - S_A) \le n!.$$

Dividing by $n!$ and using (4.3.1) and (4.3.2) we have

$$\sum_{A\in\mathcal{F}} \frac{1-\frac{\varepsilon}{100}-\alpha_A}{\binom{n}{|A|}} \le 1.$$

Now use that $\binom{n}{|A|} \le \binom{n}{\lfloor n/2\rfloor}$ and that $|\mathcal{F}| = (1+\varepsilon/2)\binom{n}{\lfloor n/2\rfloor}$ to get that

$$\frac{\varepsilon}{10}\binom{n}{\lfloor n/2\rfloor} \le \sum_{A\in\mathcal{F}} \alpha_A,$$

and the claim follows. □

Claim 4.3.7. *If a set $A^* \in \mathcal{F}$ satisfies $\alpha_{A^*} \ge \varepsilon/20$ then A^* is $(\mathcal{F}, n, n^2/2)$-eligible.*

Proof of claim. Let $A^* \in \mathcal{F}$ be such that $\alpha_{A^*} \ge \varepsilon/20$. Note that by definition of α_{A^*} we have

$$|\mathcal{H}_{A^*}| \ge \frac{\varepsilon}{20}\binom{|A^*|}{\Delta}.$$

Consider the graph G with vertex set $V(G) = \{B \in \mathcal{F} : B \subsetneq A^*, |A^* \setminus B| \ge 11\}$ and an edge connecting B_1, B_2 if $B_1 \cup B_2 = A^*$. Let S be the set of the n largest degree vertices in G. Then

$$\begin{aligned}|S_\Delta| &:= |\{C \in \mathcal{P}(n) \ :\ |C| = \Delta,\ \exists B \in S : C \subset B\}| \le n\binom{|A^*|-11}{\Delta}\\ &\le \binom{|A^*|}{\Delta}\left(\frac{|A^*|-\Delta}{|A^*|-10}\right)^{11} \le \binom{|A^*|}{\Delta}\left(\frac{2\sqrt{n\log n}}{\frac{n}{2}-2\sqrt{n\log n}}\right)^{11} \le \frac{1}{n}\binom{|A^*|}{\Delta}.\end{aligned}$$

Let $\mathcal{H}'_{A^*} = \mathcal{H}_{A^*} \setminus S_\Delta,\ \ G' = G \setminus S$ and note that

$$|\mathcal{H}'_{A^*}| = |\mathcal{H}_{A^*}| - |S_\Delta| \ge \frac{\varepsilon}{100}\binom{|A^*|}{\Delta}.$$

Let us now count the number P of pairs $C_1, C_2 \in \mathcal{H}'_{A^*}$ satisfying $C_1 \cup C_2 = A^*$. Note that $C_1 \cup C_2 = A^*$ iff $(A^* \setminus C_1) \cap (A^* \setminus C_2) = \emptyset$. Define β by

$$1+\beta = \frac{|\mathcal{H}'_{A^*}|}{\binom{|A^*|-1}{|A^*|-\Delta-1}} > 2,$$

hence we get by Lemma 4.3.3 with $N = \binom{|A^*|}{|A^*|-\Delta}$, $D = \binom{\Delta}{|A^*|-\Delta}$, $m = |A^*|$ and $k = |A^*| - \Delta$ that

$$P \geq \left(1 - \frac{\binom{|A^*|-1}{|A^*|-\Delta-1}}{\frac{\varepsilon}{100}\binom{|A^*|}{\Delta}}\right) \frac{\binom{\Delta}{|A^*|-\Delta}|A^*|}{\binom{|A^*|}{|A^*|-\Delta}\Delta} \binom{|\mathcal{H}'_{A^*}|}{2}.$$

Note that as $n - \sqrt{n \log n} \leq 2|A^*| \leq n + \sqrt{n \log n}$ and $\Delta = n/2 - \sqrt{n \log n}$ we get

$$\begin{aligned}\frac{\binom{\Delta}{|A^*|-\Delta}}{\binom{|A^*|}{|A^*|-\Delta}} &\geq \left(\frac{\Delta - (|A^*| - \Delta)}{|A^*| - (|A^*| - \Delta)}\right)^{|A^*|-\Delta} = \left(1 - \frac{|A^*| - \Delta}{\Delta}\right)^{\frac{\Delta}{|A^*|-\Delta}\frac{(|A^*|-\Delta)^2}{\Delta}} \\ &\geq e^{-1.01\frac{(3\sqrt{n \log n}/2)^2}{n/2.01}} \geq e^{-\frac{19}{4}\log n} = \frac{1}{n^{19/4}}.\end{aligned}$$

Hence we have

$$P \geq \frac{1}{n^5}\binom{|\mathcal{H}'_{A^*}|}{2}.$$

Since every vertex in G' corresponds to at most $\binom{|A^*|-11}{\Delta}$ vertices in $\mathcal{H}'_{A^*}$, the number of edges in G' is at least

$$e(G') \geq \frac{1}{10n^5}\left(\frac{|\mathcal{H}_{A^*}|}{\binom{|A^*|-11}{\Delta}}\right)^2 \geq \frac{\varepsilon^2}{10^6 n^5}\left(\frac{|A^*|}{|A^*| - \Delta}\right)^{22} \geq \frac{\varepsilon^2}{10^{20} n^5}\left(\frac{n}{\sqrt{n \log n}}\right)^{22} \geq \frac{\varepsilon^2 n^6}{10^{20}\log^{11} n} \geq n^2.$$

(Note that this last line would not have worked if we replaced the constant 10 throughout the proof by e.g. 7.) Hence in G the n largest degree vertices are in $S \subset V(G)$ and after removing S from G we still have $e(G \setminus S) = e(G') \geq n^2$ edges. By Lemma 4.3.5 there is a subgraph $G^* \subseteq G$ with $\Delta(G^*) \leq n$ and $e(G^*) \geq n^2/2$. So A^* is $(\mathcal{F}, n, n^2/2)$-eligible and the proof is completed. ☐

Theorem 4.3.1 follows from claims 4.3.6 and 4.3.7. ☐

4.4 Proof of the main result

Proof of Theorem 4.1.3. Define the 3-uniform hypergraph $\mathcal{H}$ on vertex set $\mathcal{P}(n)$ and edge set $E(\mathcal{H}) = \{(A, B, C) : A \cup B = C\}$. For an edge (A, B, C) with $A \cup B = C$ set $f(A, B, C) = C$. Hence $\mathcal{H}$ is rooted under f. Fix a constant ε with $0 < \varepsilon < 1/200$ and let n be sufficiently large. By Theorem 4.3.1, $\mathcal{H}$ is $\left(\varepsilon, \binom{n}{\lfloor n/2 \rfloor}, n, \frac{\varepsilon^2}{10^{40}}n^2\right)$-nice. Apply Theorem 4.2.9 with parameters $s = n$, $t = \frac{\varepsilon^2}{10^{40}}n^2$, $N = \binom{n}{\lfloor n/2 \rfloor}$ to obtain a family $\mathcal{C}$ of containers. Each container $C \in \mathcal{C}$ satisfies

$$|C| \le (1+100\varepsilon)\binom{n}{\lfloor n/2 \rfloor},$$

and the size of the family of containers satisfies

$$\log_2 |\mathcal{C}| \le \frac{2 \cdot 2^n}{\varepsilon}\left(H\left(2 \cdot 10^{40}/\varepsilon^2 n\right) + H\left(1/4\varepsilon n\right)\right).$$

Since $H(x) \le 2x \log(x^{-1})$ for $x < 1/2$, we get

$$\log_2 |\mathcal{C}| \le 10^{42} \frac{2^n}{\varepsilon}\left(\frac{\log\left(\varepsilon^2 n\right)}{\varepsilon^2 n} + \frac{\log(4\varepsilon n)}{4\varepsilon n}\right) \le 10^{44}\varepsilon^{-3}\frac{2^n}{n}\log n \le \varepsilon\binom{n}{\lfloor n/2 \rfloor}.$$

The number of independent sets in $\mathcal{H}$, and hence the number $\alpha(n)$ of union-free families is bounded by

$$\alpha(n) \le 2^{\varepsilon\binom{n}{\lfloor n/2 \rfloor}} 2^{(1+100\varepsilon)\binom{n}{\lfloor n/2 \rfloor}} \le 2^{(1+101\varepsilon)\binom{n}{\lfloor n/2 \rfloor}}.$$

This completes the proof of Theorem 4.1.3. □

4.5 Concluding remarks

Instead of the definition of a *rooted* hypergraph we could have defined more generally an r-**rooted** hypergraph. A 3-uniform hypergraph $\mathcal{H}$ is r-**rooted** if there exists a function $f : E(\mathcal{H}) \to V(\mathcal{H})$ such that

- for every edge $e \in E(\mathcal{H})$ we have $f(e) \in e$, and
- for any pair of vertices u, v there exist at most r edges $e \in E(\mathcal{H})$ with $u, v \in e$ and $f(e) \notin \{u, v\}$.

Similarly as before, if $\mathcal{H}$ is an r-rooted hypergraph and f is specified then we call f a **rooting function for** $\mathcal{H}$. Given this definition, essentially the same proof gives the following container theorem:

Theorem 4.5.1. *[Container theorem for r-rooted 3-uniform hypergraphs] Let $\varepsilon, s, t, r, N, M > 0$ be parameters satisfying*

$$\varepsilon \le 1/10, \qquad \frac{4s}{t} + \frac{r}{2\varepsilon s} \le \varepsilon \qquad \textit{and} \qquad M \ge (1+100\varepsilon)N.$$

Let $\mathcal{H}$ be a 3-uniform r-rooted M-vertex hypergraph $\mathcal{H}$ such that there exists a rooting function f for $\mathcal{H}$ so that $\mathcal{H}$ is (ε, N, s, t)-nice. Then there exists a family $\mathcal{C} \subseteq \mathcal{P}(V(\mathcal{H}))$ satisfying the following:

1. *For every independent set $I \subseteq V(\mathcal{H})$, there exists a $C_I \in \mathcal{C}$ such that $I \subseteq C_I$.*

2. $\log_2 |\mathcal{C}| \leq \frac{2M}{\varepsilon}(H(2s/t) + H(r/4\varepsilon s))$.

3. *Every $C \in \mathcal{C}$ satisfies $|C| \leq (1 + 100\varepsilon)N$.* □

We note that Theorem 4.5.1 can be generalised to k-uniform hypergraphs, but due to lack of applications we chose not to do so here. It is a natural question to ask how Theorem 4.5.1 compares to the vast number of container lemmas in the literature. The primary difference is that our lemma works very well if the codegrees of a hypergraph are high, but the edges can be oriented in such a way that in one direction all codegrees are small, as is the hypergraph $\mathcal{H}$ considered throughout this chapter (indeed the reason why we proved Theorem 4.5.1 in the first place was that we were not able to prove Theorem 4.1.3 using any of the already existing container lemmas). Other than this difference, the proof of Theorem 4.5.1 resembles the main theorems of [8, 98], but we put more effort into calculating the actual dependence of the various constants.

Chapter 5

An improved lower bound for Folkman's theorem

The results in this chapter are joint work with József Balogh, Sean Eberhard, Bhargav Narayanan and Andrew Treglown [7].

Folkman's theorem asserts that for each $k \in \mathbb{N}$, there exists a natural number $n = F(k)$ such that whenever the elements of $[n]$ are two-coloured, there exists a set $A \subset [n]$ of size k with the property that all the sums of the form $\sum_{x \in B} x$, where B is a nonempty subset of A, are contained in $[n]$ and have the same colour. In 1989, Erdős and Spencer showed that $F(k) \geq 2^{ck^2/\log k}$, where $c > 0$ is an absolute constant; here, we improve this bound significantly by showing that $F(k) \geq 2^{2^{k-1}/k}$ for all $k \in \mathbb{N}$.

5.1 Introduction

Recall that Ramsey's theorem [92], for two colors, states that instead of triangles we can hope to find arbitrarily large monochromatic cliques, provided the complete graph whose edges we color is large enough.

Theorem 5.1.1 (Ramsey). *For every $n \in \mathbb{N}$ there exists an $N \in \mathbb{N}$ such that whenever the edges of K_N are two-colored, there exists a monochromatic K_n.*

We denote by $R(n)$ the smallest N for which Theorem 5.1.1 holds. The upper bound

$$R(n) \leq \binom{2n-2}{n-1} \leq 4^n$$

follows from a short pigeonhole argument due to Erdős and Szekeres [42]. The lower bound

$$\sqrt{2}^n \leq R(n)$$

is due to Erdős [35] and was instrumental in his introduction of the probabilistic method. Despite a considerable effort in the past seven decades, the two constants $\sqrt{2}$ and 4 in the lower and upper bounds were not improved. The current best bounds are due to Spencer [100] and Conlon [28].

Another classical theorem in the area is Van der Waerden's theorem [107], which states that whenever $[N]$ is r-colored, there is a monochromatic arithmetic progression of length k, provided N is sufficiently large compared to k and r.

Theorem 5.1.2 (Van der Waerden)**.** *For every $r, k \in \mathbb{N}$ there exists an $N \in \mathbb{N}$ such that the following holds. Whenever $[N]$ is r-colored, there is a monochromatic arithmetic progression of length k.*

Writing $W(k, r)$ for the smallest N satisfying Theorem 5.1.2, Berlekamp [17] showed that for p prime,

$$p \cdot 2^p \leq W(p+1, 2).$$

A recent breakthrough due to Gowers [52] is the upper bound

$$W(k, r) \leq 2^{2^{r^{2^{2^{k+9}}}}}.$$

Graham conjectured [54] that $W(k, 2) < 2^{k^2}$ and offered \$1000 for a resolution of this conjecture.

Schur showed [99] that in an r-coloring of $[n]$, one of the color classes must contain a small additive structure. Given an equation on k variables and a coloring of $[n]$, we say the equation has a monochromatic solution if there exists a solution to the equation with all k variables having the same color.

Theorem 5.1.3 (Schur)**.** *For every $r \in \mathbb{N}$ there exists an integer $S(r) \in \mathbb{N}$ such that whenever $[S(r)]$ is r-colored, there exists a monochromatic solution to the equation $x + y = z$.*

Another example of a structure one can find in an r-coloring of the natural numbers is a *Hilbert cube*, or *affine cube*. For some $a, d_1, \ldots, d_k$ the k-dimensional affine cube is

$$H(a; d_1, \ldots, d_k) := \left\{ a + \sum_{i=1}^{k} c_i \cdot d_i : c_i \in \{0, 1\} \right\}.$$

Hilbert showed [63] that no matter how we r-color the integers, there always exists a monochromatic Hilbert cube. His famous Cube Lemma was perhaps the very first result in Ramsey theory.

Theorem 5.1.4 (Hilbert's Cube Lemma)**.** *For every $k, r \in \mathbb{N}$ there exists an integer $H(k, r)$ such that whenever $[H(k, r)]$ is r-colored, there is a monochromatic Hilbert cube of dimension k.*

Hilbert's upper bound $H(k) \leq r^{2.6^k}$ was improved by Gunderson–Rödl [56] who used a lemma due to Szemerédi [102] to show

$$r^{(1-o(1))\left(2^k-1\right)/k} \leq H(k,r) \leq (2r)^{2^{k-1}}.$$

Their lower bound was recently improved by Conlon–Fox–Sudakov [29] to $2^{ck^2} \leq H(k,2)$. As every arithmetic progression of length $\approx k^2/2$ is a k-dimensional Hilbert cube, any further improvement on this lower bound would improve Berlekamp's bound on the Van der Waerden numbers as well.

About fifty years ago, a wide generalisation of Schur's theorem was obtained independently by Folkman, Rado and Sanders, and this generalisation is now commonly referred to as Folkman's theorem (see [55], for example). To state Folkman's theorem, it will be convenient to have some notation. For $n \in \mathbb{N}$, we write $[n]$ for the set $\{1, 2, \ldots, n\}$, and for a finite set $A \subset \mathbb{N}$, let

$$S(A) = \left\{ \sum_{x \in B} x : B \subset A \text{ and } B \neq \emptyset \right\}$$

denote the set of all *finite sums* of A. In this language, Folkman's theorem states that for all $k, r \in \mathbb{N}$, there exists a natural number $n = F(k,r)$ such that whenever the elements of $[n]$ are r-coloured, there exists a set $A \subset [n]$ of size k such that $S(A)$ is a monochromatic subset of $[n]$; of course, it is easy to see that Folkman's theorem, in the case where $k = 2$, implies Schur's theorem.

Taylor [103] showed that $F(r,k)$ is upper bounded by a tower of height $2r(k-1)$. Erdős–Spencer showed [41] that $2^{k^2/\log k} \leq F(k,2)$. Since $H(k,r) \leq F(k,r)$ the result of Conlon–Fox–Sudakov on Hilbert cubes implies that $2^{ck^2} \leq F(k,2)$. The purpose of this chapter is to improve this lower bound.

5.2 A new lower bound on Folkman's theorem

In this chapter, we shall be concerned with lower bounds for the two-colour Folkman numbers, i.e., for the quantity $F(k) = F(k,2)$.Recall that In 1989, Erdős and Spencer [41] proved that

$$F(k) \geq 2^{ck^2/\log k} \tag{5.2.1}$$

for all $k \in \mathbb{N}$, where $c > 0$ is an absolute constant; here, and in what follows, all logarithms are to the base 2. Our primary aim in this chapter is to improve (5.2.1).

Before we state and prove our main result, let us say a few words about the proof of (5.2.1). Erdős and Spencer establish (5.2.1) by considering uniformly random two-colourings. In particular, they show that if

$[n]$ is two-coloured uniformly at random and additionally $n \leq 2^{ck^2/\log k}$ for some suitably small absolute constant $c > 0$, then with high probability, there is no k-set $A \subset [n]$ for which $S(A)$ is monochromatic. On the other hand, it is not hard to check that if $n \geq 2^{Ck^2}$ for some suitably large absolute constant $C > 0$, then a two-colouring of $[n]$ chosen uniformly at random is such that, with high probability, there exists a set $A \subset [n]$ of size k for which $S(A)$ is monochromatic; indeed, to see this, it is sufficient to restrict our attention to sets of the form $\{p, 2p, \ldots, kp\}$, where p is a prime in the interval $[n/\log^2 n, 2n/\log^2 n]$, and notice that the sets of finite sums of such sets all have size $k(k+1)/2$ and are pairwise disjoint. With perhaps this fact in mind, in their paper, Erdős and Spencer also describe some of their attempts at removing the factor of $\log k$ in the exponent in (5.2.1); nevertheless, their bound has not been improved upon since.

Our main contribution is a new, doubly exponential, lower bound for $F(k)$, significantly strengthening the bound due to Erdős and Spencer.

Theorem 5.2.1. *For all $k \in \mathbb{N}$, we have*

$$F(k) \geq 2^{2^{k-1}/k}. \tag{5.2.2}$$

5.3 Proof of the main result

In this section, we give the proof of Theorem 5.2.1.

Proof of Theorem 5.2.1. The result is easily verified when $k \leq 3$, so suppose that $k \geq 4$ and let $n = \lfloor 2^{2^{k-1}/k} \rfloor$. In the light of our earlier remarks, a uniformly random colouring of $[n]$ is a poor candidate for establishing (5.2.2). Instead, we generate a (random) red-blue colouring of $[n]$ as follows: we first red-blue colour the odd elements of $[n]$ uniformly at random, and then extend this colouring uniquely to all of $[n]$ by insisting that the colour of $2x$ be different from the colour of x for each $x \in [n]$; hence, for example, if 5 is initially coloured blue, then 10 gets coloured red, 20 gets coloured blue, and so on.

Fix a set $A \subset [n]$ of size k with $S(A) \subset [n]$. We have the following estimate for the probability that $S(A)$ is monochromatic in our colouring.

Claim 5.3.1. $\mathbb{P}(S(A)$ *is monochromatic*$) \leq 2^{1-2^{k-1}}$.

Proof. First, if $|S(A)| \leq 2^k - 2$, then it is easy to see from the pigeonhole principle that there exist two subsets $B_1, B_2 \subset A$ such that $\sum_{x \in B_1} x = \sum_{x \in B_2} x$, and by removing $B_1 \cap B_2$ from both B_1 and B_2 if necessary, these sets may further be assumed to be disjoint; in particular, this implies that $S(A)$ contains two elements one of which is twice the other. It therefore follows from the definition of our colouring that $S(A)$ cannot be monochromatic unless $|S(A)| = 2^k - 1$.

Next, suppose that $|S(A)| = 2^k - 1$. For each odd integer $m \in \mathbb{N}$, we define $G_m = \{m, 2m, 4m, \dots\} \cap [n]$, and note that these geometric progressions partition $[n]$. Observe that $S(A)$ intersects at least 2^{k-1} of these progressions. Indeed, if there is an odd integer $r \in A$ for example, then $S(A)$ contains exactly 2^{k-1} distinct odd elements and these elements must lie in different progressions. More generally, if each element of A is divisible by 2^s and s is maximal, then there exists an element r of A with $r = 2^s t$, where t is odd; it is then clear that precisely 2^{k-1} elements of $S(A)$ are divisible by 2^s but not by 2^{s+1} and these elements must necessarily lie in different progressions. With this in mind, we define B_A to be a maximal subset of $S(A)$ with the property $|B_A \cap G_m| \le 1$ for each m; for example, we may take B_A to consist of the least elements (where they exist) of the sets $S(A) \cap G_m$. Clearly, our red-blue colouring restricted to B_A is a uniformly random colouring, so the probability that B_A is monochromatic is $2^{1-|B_A|}$; it follows that the probability that $S(A)$ is monochromatic is at most $2^{1-|B_A|} \le 2^{1-2^{k-1}}$. $\square$

It is now easy to see that if X is the number of sets $A \subset [n]$ of size k for which $S(A)$ is a monochromatic subset of $[n]$ in our colouring, then

$$\mathbb{E}[X] \le \binom{n}{k} 2^{1-2^{k-1}} \le \left(\frac{en}{k}\right)^k 2^{1-2^{k-1}} \le \left(\frac{e2^{2^{k-1}/k}}{k}\right)^k \left(2^{1-2^{k-1}}\right) = 2\left(\frac{e}{k}\right)^k < 1,$$

where the last inequality holds for all $k \ge 4$. Hence, there exists a red-blue colouring of $[n]$ without any set A of size k for which $S(A)$ is a monochromatic subset of $[n]$, proving the result. $\square$

Chapter 6

Large subgraphs in rainbow-triangle free colorings

This chapter is based on [106].

Fox–Grinshpun–Pach showed that every 3-coloring of the complete graph on n vertices without a rainbow triangle contains a clique of size $\Omega\left(n^{1/3}\log^2 n\right)$ which uses at most two colors, and this bound is tight up to the constant factor. We show that if instead of looking for large cliques one only tries to find subgraphs of large chromatic number, one can do much better. We show that every such coloring contains a 2-colored subgraph with chromatic number at least $n^{2/3}$, and this is best possible. We further show that for fixed positive integers s, r with $s \leq r$, every r-coloring of the edges of the complete graph on n vertices without a rainbow triangle contains a subgraph that uses at most s colors and has chromatic number at least $n^{s/r}$, and this is best possible. Fox–Grinshpun–Pach previously showed a clique version of this result.

As a direct corollary of our result we obtain a generalisation of the celebrated theorem of Erdős-Szekeres, which states that any sequence of n numbers contains a monotone subsequence of length at least $\sqrt{n}$. We prove that if an r-coloring of the edges of an n-vertex tournament does not contain a rainbow triangle then there is an s-colored directed path on $n^{s/r}$ vertices, which is best possible. This gives a partial answer to a question of Loh.

6.1 Introduction

A Gallai-coloring of a complete graph is an edge coloring such that no triangle is colored with three distinct colors. Such colorings arise naturally in several areas including in information theory [79], in the study of partially ordered sets, as in Gallai's original paper [50], and in the study of perfect graphs [23]. Several Ramsey-type results in Gallai-colored graphs have also emerged in the literature (see e.g. [26], [46], [57], [58]), but they mostly focus on finding large monochromatic structures in such colorings. Our main result is the observation that certain proof techniques used by Fox–Grinshpun–Pach [43] for solving the multicolor Erdős–Hajnal conjecture for rainbow triangles also give a partial answer for Loh's question [82], that asks the following: what is the value of $f(n, r, s)$, the maximum number such that every r-coloring of the edges of the transitive tournament on n vertices contains a directed path with at least $f(n, r, s)$ vertices whose edges have at most s colors?

6.1.1 Erdős–Hajnal for rainbow triangles

Erdős showed [34] that a random graph on n vertices almost surely contains no clique or independent set of order $2\log n$. On the other hand, the Erdős–Hajnal conjecture [38] states that for each fixed graph H there is an $\varepsilon = \varepsilon(H) > 0$ such that every graph G on n vertices which does not contain a fixed induced subgraph H has a clique or independent set of order n^ε, much larger than in the case of general graphs. The Erdős–Hajnal conjecture is still open, but there are now several partial results on it - we refer the reader to the introduction of [43] and the recent survey [25]. In the present chapter we will be only interested in a special case of the multicolor generalisation of the Erdős–Hajnal conjecture. Hajnal [59] conjectured that there is an $\varepsilon > 0$ such that every 3-coloring of the edges of the complete graph on n vertices without a rainbow triangle (that is, a triangle with all its edges different colors) contains a set of order n^ε which uses at most two colors. Fox–Grinshpun–Pach proved Hajnal's conjecture and further determined a tight bound on the order of the largest guaranteed 2-colored set in any such coloring. A *Gallai r-coloring* is a coloring of the edges of a complete graph using r colors without rainbow triangles.

Theorem 6.1.1 (Fox–Grinshpun–Pach, [43])**.** *Every Gallai-3-coloring on n vertices contains a set of order $\Omega\left(n^{1/3}\log^2 n\right)$ which uses at most two colors, and this bound is tight up to a constant factor.*

Instead of looking for large complete subgraphs, it is also natural to try to find 2-colored subgraphs with large chromatic number. It is easy to show that every 3-edge-colored complete graph on n vertices contains a 2-colored subgraph with chromatic number at least $\sqrt{n}$. Indeed, the complement of the graph G_g consisting of green edges is the graph G_{rb} consisting of red and blue edges, and every graph G satisfies $\chi(G)\cdot\chi(G^c) \geq |V(G)|$. It is not hard to construct an infinite set of examples on n^2 vertices for some integer n where all three 2-colored subgraphs have chromatic number precisely n, hence this is best possible. Our first result is that in the case of Gallai-colorings, we can do much better.

Theorem 6.1.2. *Every Gallai-3-coloring on n vertices contains a 2-colored subgraph that has chromatic number at least $n^{2/3}$.*

Fox–Grinshpun–Pach further obtained a generalisation of Theorem 6.1.1 to more colors.

Theorem 6.1.3 (Fox–Grinshpun–Pach, [43]). *Let r and s be fixed positive integers with $s \leq r$. Every Gallai-r-coloring on n vertices contains a set of order $\Omega\left(n^{\binom{s}{2}/\binom{r}{2}} \log^{c_{r,s}} n\right)$ which uses at most s colors, and this bound is tight up to a constant factor. Here $c_{r,s}$ is a constant depending only on r and s.*

Moreover, the value of $c_{r,s}$ was exactly determined in [43]. We prove a corresponding theorem about subgraphs with few colors and large chromatic number.

Theorem 6.1.4. *Let r and s be fixed positive integers with $s \leq r$. Every Gallai-r-coloring on n vertices contains an s-colored subgraph that has chromatic number at least $n^{s/r}$.*

Both Theorems 6.1.2 and 6.1.4 are sharp, as seen in Construction 6.2.2. The motivation of Fox–Grinshpun–Pach for proving Theorem 6.1.3 was to get one step closer towards proving the Erdős–Hajnal conjecture. On the other hand, our main motivation for proving Theorem 6.1.4 came from a completely different direction - we tried to give an answer to Loh's question (see Section 6.1.2) - in fact, when proving our main results we were not even aware of the Fox–Grinshpun–Pach paper. As it happens, we were only able to give a partial answer to Loh's question, and quite surprisingly, our proof of Theorem 6.1.4 was very similar to their proof of a weaker version of Theorem 6.1.3: we use Gallai's structure theorem for Gallai-colored complete graphs to obtain a nice block partition of the vertex set, and then the theorem follows by induction with some more work.

6.1.2 Long subchromatic paths in tournaments and Loh's question

A celebrated theorem of Erdős and Szekeres [42] states that for any two positive integers r, s, every sequence of $rs+1$ (not necessarily distinct) numbers contains a monotone increasing subsequence of length $r+1$ or a monotone decreasing subsequence of length $s+1$. A short proof of this theorem is given by the pigeonhole principle. Assign to each number in the sequence an ordered pair (x, y) where x is the length of the longest increasing subsequence ending at this number, and y is the decreasing analogue. It is then easy to see that all these ordered pairs have to be distinct, and therefore there is an ordered pair with first element at least $r+1$ or second element at least $s+1$.

The exact same proof also gives the following extension. Consider the n-vertex transitive tournament T_n, where the edge between $i < j$ is oriented in the direction $\overrightarrow{ij}$. Then every 2-coloring of the edges of T_n has a monochromatic directed path of length at least $\sqrt{n}$ (throughout this chapter the length of a path means vertex-length, i.e. the number of vertices in the path). Moreover if we consider r-colored tournaments then the same proof shows that there exists a monochromatic tournament of length $n^{1/r}$. In fact, all the above results are sharp for infinitely many n, and they are also true for non-transitive tournaments.

Loh [82] asked about the following beautiful generalisation of the above. Determine $f(n, r, s)$, the maximum number such that every r-coloring of the edges of the transitive tournament on n vertices contains a directed path with at least $f(n, r, s)$ vertices whose edges have at most s colors. By grouping together sets of s colors, a similar argument as the above shows that

$$n^{1/\lceil r/s \rceil} \leq f(n, r, s),$$

and a standard construction shows that the upper bound

$$f(n, r, s) \leq n^{s/r}$$

holds whenever n is a perfect r-th power. However, already the $r = 3, s = 2$ case is non-trivial, since the above bounds only give $\sqrt{n} \leq f(n, 3, 2) \leq n^{2/3}(1 + o(1))$. Loh proved that the correct answer in this case is not $\sqrt{n}$. Here, $\log^*$ is the iterated logarithm, or the inverse of the tower function $T(n) = 2^{T(n-1)}$, $T(0) = 1$.

Theorem 6.1.5 (Loh, [82]). *There exists a constant $C > 0$ such that*

$$C\sqrt{n} \cdot e^{\log^* n} \leq f(n, 3, 2),$$

that is, every 3-coloring of the edges of the transitive n-vertex tournament contains a directed path of length at least $C\sqrt{n} \cdot e^{\log^* n}$ whose edges use at most 2 colors.

Recently Gowers–Long [53] improved Theorem 6.1.5 by showing that there exists $\varepsilon > 0$ such that

$$n^{0.5+\varepsilon} \leq f(n, 3, 2).$$

The question of determining the order of magnitude of $f(n, 3, 2)$, and in general that of $f(n, r, s)$, is still wide open. A direct corollary of Theorem 6.1.4 is a partial answer to Loh's question, which was the main motivation of this chapter. Recall that a *rainbow triangle* is a triangle whose edges are all different colors. A *Gallai-r-coloring* of K_n is an r-coloring of the edges of K_n without rainbow triangles.

Theorem 6.1.6. *Let r, s, n be positive integers with $s \leq r$. Every Gallai-r-coloring of an n-vertex tournament contains a directed path on at least $n^{s/r}$ vertices, whose edges use at most s colors.*

Note that every 2-coloring is rainbow-triangle free, hence Theorem 6.1.6 is a generalisation of the Erdős–Szekeres Theorem. We emphasize that our result holds for non-transitive tournaments as well, but our proof methods completely break down for colorings that contain rainbow triangles.

Nevertheless, our guess is that $f(n,r,s) = n^{s/r}$ holds whenever n is a perfect r-th power. A reason for this is as follows. Suppose $r = 3$ and $s = 2$, and u, v, w are vertices of the tournament such that $\overrightarrow{uv}$ is red, $\overrightarrow{vw}$ is blue and $\overrightarrow{uw}$ is green. Consider what happens if we recolor the $\overrightarrow{uw}$ edge to become, say, color red. Then the lengths of the red-green paths do not change, as the set of red-green edges did not change. The length of the longest red-blue paths did not change, as any path that contained the $\overrightarrow{uw}$ edge could have instead contained the $\overrightarrow{uv}$ and $\overrightarrow{vw}$ edges, giving a longer path. Finally, the length of the longest blue-green paths did not increase, as the set of blue-green edges descreased by the $\overrightarrow{uw}$ edge. Hence, in some sense, destroying this uvw rainbow triangle gave us a better coloring, with shorter two-colored paths. Unfortunately, it is not the case that by repeating moves like the above one can always transform the coloring into a Gallai-coloring, hence this paragraph is nothing more than a (more or less) convincing heuristic argument.

6.2 The proof of Theorem 6.1.6

We will deduce Theorem 6.1.6 from Theorem 6.1.4. The connection between chromatic number and longest paths in orientations of graphs is given by the Gallai-Hasse-Roy-Vitaver Theorem.

Theorem 6.2.1 (Gallai-Hasse-Roy-Vitaver, [49]). *Every orientation of the edges of a graph G has a directed path on at least $\chi(G)$ vertices.*

Proof of Theorem 6.1.6. We are given a Gallai-r-coloring of an n-vertex tournament. By Theorem 6.1.4 we can find an s-colored subgraph G that has chromatic number at least $n^{s/r}$. By Theorem 6.2.1 we can find a directed path P in G on at least $n^{s/r}$ vertices. As the edges of G use at most s colors, and P is in G, we conclude that the edges of P use at most s colors and the proof is complete. □

A folklore construction shows that Theorems 6.1.6 and 6.1.4 are sharp whenever n is a perfect r-th power.

Construction 6.2.2. *Consider the n^r numbers $\{0, 1, \ldots, n^r - 1\}$ which will be the vertices of the graph. Write every such number as a r-digit base-n number (by adding trailing zeros if necessary). Color edge $i < j$ according to the leftmost digit in which they differ and orient them as $\overrightarrow{ij}$ This r-colored transitive tournament has the property that the length of any s-colored path is at most n^s, and hence the chromatic number of the corresponding subgraph is at most n^s.*

6.3 Proof of Theorem 6.1.4

We will deduce Theorem 6.1.4 from the following stronger statement.

Theorem 6.3.1. *Let $r, s, n \in \mathbf{Z}^+$ with $s \leq r$. Given a Gallai-r-coloring of K_n using colors $[r]$ and given $S \subset [r]$ write G_S for the subgraph whose edges are colored by elements of S. Then*

$$n^{\binom{r-1}{s-1}} \leq \prod_{|S|=s} \chi(G_S),$$

where the product goes over all $S \subset [r]$ with $|S| = s$.

The proof of Theorem 6.3.1 shows a lot of similarities to the proof of Theorem 7.2. in [43], the main new idea is in Claim 6.3.5. Recall that Theorem 6.3.1 is false for general colorings. Theorem 6.1.4 is an immediate consequence of Theorem 6.3.1.

Proof of Theorem 6.1.4. Given a Gallai-r-coloring on n vertices, Theorem 6.3.1 states that the geometric mean of the $\chi(G_S)$-s is at least $n^{s/r}$. Hence in particular there exists an $S \subset [r]$ with $|S| = s$ such that $\chi(G_S) \geq n^{s/r}$, as claimed. □

To prove Theorem 6.3.1 we will need another theorem by Gallai:

Lemma 6.3.2 (Gallai, [50]). *An edge-coloring F of a complete graph on a vertex set V with $|V| \geq 2$ is a Gallai coloring if and only if V may be partitioned into nonempty sets $V_1, \dots, V_t$ with $t \geq 2$ so that each V_i has no rainbow triangles under F, at most two colors are used on the edges not internal to any V_i, and the edges between any fixed pair (V_i, V_j) use only one color. Furthermore, any such substitution of Gallai colorings for vertices of a 2-edge-coloring of a complete graph K_t yields a Gallai coloring.*

For some recent progress on generalising Lemma 6.3.2, we direct the reader to the beautiful paper of Leader–Tan [81]. We will also make use of the following observation.

Observation 6.3.3. *Let G be a graph on vertex set $V(G)$ and let $V_1 \cup V_2 \cup \ldots \cup V_m$ be a partition of $V(G)$ such that for each pair of distinct $i, j \in [m]$, either all edges of the form $\{uv : u \in V_i, v \in V_j\}$ are present in G, or none of them. For each $i \in [m]$, let G'_i be an arbitrary graph with chromatic number $\chi(G'_i) = \chi(G[V_i])$. Let H be the graph obtained from G by replacing $G[V_i]$ by G'_i for each $i \in [m]$. Then $\chi(G) = \chi(H)$.*

The following is a common generalisation of Hölder's inequality that we will find useful.

Lemma 6.3.4. *If $\mathcal{F}$ is a finite set of indices and for each $S \in \mathcal{F}$ we have that a_S is a function mapping $[m]$ to the non-negative reals, then*

$$\prod_{S\in\mathcal{F}} \sum_i a_S(i) \geq \left(\sum_i \prod_{S\in\mathcal{F}} a_S(i)^{1/|\mathcal{F}|} \right)^{|\mathcal{F}|}.$$

Proof of Theorem 6.3.1. We prove the theorem by induction on n. If $n = 1$ then every $\chi(G_S)$ is equal to 1, and $n^{\binom{r-1}{s-1}}$ is also equal to 1. If $n > 1$, we can find a pair of colors Q and some non-trivial partition of the vertices $V_1, \ldots, V_m$ such that for each pair of distinct i, j in $[m]$, there is a $q \in Q$ so that all of the edges between V_i and V_j have color q. For each $S \subset [r]$ and $i \in [m]$ let $G_{S,i}$ be the subgraph of the complete graph on V_i consisting of edges colored by colors in S. Write $\chi(S, i) := \chi(G_{S,i})$ and $\chi(S) = \chi(G_S)$. Let $Q = \{q_1, q_2\}$. If $q_1 \in S$ and $q_2 \notin S$ then let $S^* := S \cup \{q_2\} \setminus \{q_1\}$.

Claim 6.3.5. *If $S \subset [k]$ with $q_1 \in S$ and $q_2 \notin S$, then*

$$\chi(S)\chi(S^*) \geq \sum_{i=1}^{m} \chi(S, i)\chi(S^*, i).$$

Proof of claim: For each $i \in [m]$, let G'_i be the 2-colored complete graph on $\chi(S, i)\chi(S^*, i)$ vertices, obtained by taking $\chi(S, i)$ disjoint copies of $K_{\chi(S^*,i)}$, coloring all edges inside the cliques by color q_2, and edges between the cliques by color q_1. For each i, replace $G[V_i]$ by G'_i, to obtain a 2-colored complete graph G' on $\sum_{i=1}^{m} \chi(S, i)\chi(S^*, i)$ vertices - let H_1 and H_2 be the subgraphs induced by edges of color q_1 and q_2 respectively. Note that $\chi(H_1) = \chi(S)$ and $\chi(H_2) = \chi(S^*)$ by Observation 6.3.3, and since H_2 is the complement of H_1 we also have $\chi(S)\chi(S^*) = \chi(H_1)\chi(H_2) \geq |V(G')| = \sum_{i=1}^{m} \chi(S, i)\chi(S^*, i)$. □

In what follows we will always omit writing $|S| = s$ in the subscripts of products for clearer presentation. By induction, for all i we have

$$|V_i|^{\binom{r-1}{s-1}} \leq \prod_S \chi(S, i). \tag{6.3.1}$$

Note that

- if $Q \cap S = \emptyset$ then $\chi(S) = \max_i\{\chi(S, i)\}$,
- If $Q \subseteq S$ then $\chi(S) = \sum_i(\chi(S, i))$ where the sum is over all $i \in [m]$.

Hence we have, using Claim 6.3.5, that

$$\prod_S \chi(S) \geq \left(\prod_{S:Q\cap S=\emptyset} \chi(S)\right)\left(\prod_{S:q_1\in S, q_2\notin S}\left(\sum_{i=1}^m \chi(S,i)\chi(S^*,i)\right)\right)\left(\prod_{S:Q\subseteq S}\sum_{i=1}^m \chi(S,i)\right). \tag{6.3.2}$$

To simplify notation, if $q_1 \in S$ and $q_2 \notin S$ then write $\alpha(S,i) := \chi(S,i)\chi(S^*,i)$, and if $Q \subseteq S$ then write $\alpha(S,i) := \chi(S,i)$. Let $\mathcal{F} = \{S \subset [r] : |S| = s, q_1 \in S\}$. So

$$\prod_S \chi(S) \geq \left(\prod_{S\in\mathcal{F}}\sum_{i=1}^m \alpha(S,i)\right)\prod_{S:Q\cap S=\emptyset}\chi(S) \geq \left(\sum_{i=1}^m\left(\prod_{S\in\mathcal{F}}\alpha(S,i)^{1/|\mathcal{F}|}\right)\right)^{|\mathcal{F}|}\prod_{S:Q\cap S=\emptyset}\chi(S) =$$
$$\left(\sum_{i=1}^m\left(\prod_{S:Q\cap S=\emptyset}\chi(S)\prod_{S\in\mathcal{F}}\alpha(S,i)\right)^{1/|\mathcal{F}|}\right)^{|\mathcal{F}|} \geq \left(\sum_{i=1}^m\left(\prod_S \chi(S,i)^{1/|\mathcal{F}|}\right)\right)^{|\mathcal{F}|} \geq \left(\sum_{i=1}^m |V_i|\right)^{|\mathcal{F}|} = n^{\binom{r-1}{s-1}},$$

where the first inequality is by rewriting (6.3.2), the second inequality is by Lemma 6.3.4, the third inequality is by $\chi(S) \geq \chi(S,i)$, and the fourth inequality is by (6.3.1), using that $|\mathcal{F}| = \binom{r-1}{s-1}$. This completes the proof. □

References

[1] R. Ahlswede, Simple hypergraphs with maximal number of adjacent pairs of edges, J. Combin. Theory B, 28, (1980), 164–167.

[2] N. Alon and F. Chung, Explicit construction of linear sized tolerant networks, Annals Discrete Math., 38, (1988), 15–19.

[3] N. Alon, S. Das, R. Glebov and B. Sudakov, Comparable pairs in families of sets, Journal of Combinatorial Theory, Series B, 115 (2015), 164–185.

[4] N. Alon and J.H. Spencer, The Probabilistic Method, John Wiley & Sons, 2004.

[5] J. Balogh, T. Bohman and D. Mubayi, Erdős–Ko–Rado in Random Hypergraphs, Combin. Probab. Comput., 18, (2009), 629–646.

[6] J. Balogh, S. Das, M. Delcourt, H. Liu and M. Sharifzadeh, The typical structure of intersecting families of discrete structures, J. Combin. Theory A, 132, (2015), 224–245.

[7] J. Balogh, S. Eberhard, B. Narayanan, A. Treglown and A. Z. Wagner, An improved lower bound for Folkman's theorem, Bulletin of the London Mathematical Society, 2017.

[8] J. Balogh, R. Morris and W. Samotij, Independent sets in hypergraphs, J. Amer. Math. Soc., 28, (2015), 669–709.

[9] J. Balogh, R. Morris and W. Samotij, The method of hypergraph containers, arXiv preprint arXiv:1801.04584.

[10] J. Balogh, R. Mycroft and A. Treglown, A random version of Sperner's theorem, J. Combin. Theory A, 128, (2014), 104–110.

[11] J. Balogh, S. Petrickova and A. Z. Wagner, Families in posets minimizing the number of comparable pairs, arXiv preprint arXiv:1703.05427 (2017).

[12] J. Balogh and W. Samotij, The number of $K_{s,t}$-free graphs, J. Lond. Math. Soc., 83, (2011), 368–388.

[13] J. Balogh, A. Treglown and A. Z. Wagner, Applications of graph containers in the Boolean lattice, Random Structures & Algorithms 49.4 (2016): 845–872.

[14] J. Balogh and A. Z. Wagner, Further applications of the container method, Recent Trends in Combinatorics. Springer International Publishing, 2016. 191–213.

[15] J. Balogh and A. Z. Wagner, On the number of union-free families, Israel Journal of Mathematics 219.1 (2017): 431–448.

[16] J. Balogh and A. Z. Wagner, "Kleitman's conjecture about families of given size minimizing the number of k-chains", to appear in Advances in Mathematics.

[17] E. R. Berlekamp, “A construction for partitions which avoid long arithmetic progressions”, Canad. Math. Bull 11.1968 (1968): 409–414.

[18] B. Bollobás, On generalized graphs, Acta Math. Acad. Sci. Hung., 16, (1965), 447–452.

[19] B. Bollobás, Z. Füredi, I. Kantor, G.O.H. Katona, and I. Leader, A coding problem for pairs of subsets, in Geometry, Structure and Randomness in Combinatorics, CRM Series, Volume 18 2014, 47–59.

[20] G. Brightwell and G.O.H. Katona, A new type of coding problem, Studia Sci. Math. Hungar., 38, (2001), 139–147.

[21] N. de Bruijn, C. Tengbergen and D. Kruyswijk, On the set of divisors of a number, Nieuw Arch. Wiskunde 23 (1951), 191–193.

[22] G. Burosch, J. Demetrovics, G.O.H. Katona, D.J. Kleitman and A.A. Sapozhenko, On the number of databases and closure operations, Theoretical Computer Science, 78 (1991), 377–381.

[23] K. Cameron, J. Edmonds, and L. Lovász, A note on perfect graphs, Period. Math. Hungar. 17 (1986), 173–175.

[24] Y.M. Chee, H.M. Kiah, H. Zhang, and X. Zhang, Optimal Codes in the Enomoto–Katona Space, Combin. Probab. Comput., 24, (2015), 382–406.

[25] M. Chudnovsky, The Erdős-Hajnal conjecture - a survey, J. Graph Theory, 75 (2014), 178–190.

[26] F. Chung and R. L. Graham, Edge-colored complete graphs with precisely colored subgraphs, Combinatorica 3 (1983), 315–324.

[27] M. Collares Neto and R. Morris, Maximum antichains in random set, Random Structures & Algorithms, to appear.

[28] D. Conlon, “A new upper bound for diagonal Ramsey numbers”, Annals of Mathematics, 170 (2): 941960, 2009.

[29] D. Conlon, J. Fox and B. Sudakov, “Short proofs of some extremal results”, Combinatorics, Probability and Computing 23.1 (2014): 8–28.

[30] S. Das, W. Gan and B. Sudakov, Sperner’s Theorem and a Problem of Erdős, Katona and Kleitman, Combinatorics, Probability and Computing, Volume 24, Special Issue 04, July 2015, pp 585–608.

[31] R. Dedekind, Uber Zerlegungen von Zahlen durch ihre gr ossten gemeinsamen Teiler, Festschrift Hoch. Braunschweig u. ges. Werke II (1897), 103–148.

[32] A. P. Dove, J. R. Griggs, R. J. Kang and J-S. Sereni, Supersaturation in the Boolean lattice, Integers 14A:A4, May 2014.

[33] D. Duffus, P. Frankl and V. Rödl, Maximal independent sets in bipartite graphs obtained from Boolean lattices. European J. Combin. 32 (2011), no. 1, 1–9.

[34] P. Erdős, On a lemma of Littlewood and Offord, Bulletin of the American Mathematical Society 51 (1945), 898–902.

[35] P. Erdős, Some remarks on the theory of graphs, Bull. Amer. Math. Soc. 53 (1947), 292–294.

[36] P. Erdős, On a theorem of Rademacher-Turán, Illinois J. Math., 6:122–127, 1962.

[37] 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 (1986), 113–121.

[38] P. Erdős and A. Hajnal, Ramsey-type theorems, Discrete Appl. Math. 25 (1989), 37–52.

[39] P. Erdős and D. Kleitman, Extremal problems among subsets of a set, Discrete Math. 8 (1974), 281–294.

[40] P. Erdős, C. Ko and R. Rado, Intersection theorems for systems of finite sets, Quart. J. Math. Oxford Ser., 12, (1961), 313–320.

[41] P. Erdős and J. Spencer, *Monochromatic sumsets*, J. Combin. Theory Ser. A **50** (1989), 162–163.

[42] P. Erdős and G. Szekeres, A combinatorial problem in geometry, Compositio Math. 2 (1935), 463–470.

[43] J. Fox, A. Grinshpun and J. Pach, The Erdős–Hajnal conjecture for rainbow triangles, Journal of Combinatorial Theory Series B 111 (2015), 75–125.

[44] P. Frankl, On the minimum number of disjoint pairs in a family of finite sets, J. Combin. Theory A, 22, (1977), 249–251.

[45] P. Frankl, An extremal problem for two families of sets, European J. Combin., 3, (1982), 125–127.

[46] S. Fujita, C. Magnant and K. Ozeki, Rainbow generalizations of Ramsey theory: a survey, Graphs Combin. 26 (2010), 1–30.

[47] Z. Füredi. "The maximum number of edges in a minimal graph of diameter 2", Journal of Graph Theory 16.1 (1992): 81-98.

[48] Z. Füredi,, J. R. Griggs and D. J. Kleitman, "A minimal cutset of the Boolean lattice with almost all members", Graphs and Combinatorics 5.1 (1989), 327–332.

[49] T. Gallai, On directed graphs and circuits, Theory of Graphs (Proceedings of the Colloquium Tihany 1966), New York: Academic Press (1968), pp. 115–118.

[50] T. Gallai, Transitiv orientierbare Graphen, Acta Math. Acad. Sci. Hungar 18 (1967), 25–66. English translation in [85].

[51] M.M. Gauy, H. Hàn and I.C. Oliveira, Erdős–Ko–Rado for random hypergraphs: asymptotics and stability, submitted.

[52] T. Gowers, "A new proof of Szemerédi's theorem", Geom. Funct. Anal. 11 (3): 465–588, 2001.

[53] W. T. Gowers and J. Long, The length of an s-increasing sequence of r-tuples, arXiv:1609.08688 (2016).

[54] R. Graham, "Some of my favorite problems in Ramsey theory", Integers: Electronic Journal of Combinatorial Number Theory 7.2 (2007): A15.

[55] R. L. Graham, B. L. Rothschild, and J. H. Spencer, *Ramsey theory*, 2^{nd} ed., Wiley-Interscience Series in Discrete Mathematics and Optimization, John Wiley & Sons, Inc., New York, 1990.

[56] D. S. Gunderson, V. Rödl, "Extremal problems for affine cubes of integers", Combinatorics, probability and computing 7.1 (1998): 65–79.

[57] A. Gyárfás and G. Sárközy, Gallai colorings of non-complete graphs, Discrete Math. 310 (2010), 977–980.

[58] A. Gyárfás, G. Sárközy, A. Sebő and S. Selkow, Ramsey-type results for Gallai colorings. J. Graph Theory 64 (2010), 233–243.

[59] A. Hajnal, Rainbow Ramsey theorems for colorings establishing negative partition relations, Fund. Math. 198 (2008), 255–262.

[60] A. Hamm and J. Kahn, On Erdős–Ko–Rado for random hypergraphs I, submitted.

[61] A. Hamm and J. Kahn, On Erdős–Ko–Rado for random hypergraphs II, submitted.

[62] G.H. Hardy, J.E. Littlewood, G. Pólya, Inequalities, Cambridge Mathematical Library (2. ed.), Cambridge: Cambridge University Press, ISBN 0-521-05206-8, MR 0046395, Zbl 0047.05302 (1952).

[63] D. Hilbert, "Über die Irreducibilität ganzer rationaler Functionen mit ganzzahligen Coefficienten", J. reine angew. Math. 110 (1892) 104–129.

[64] L. Ilinca and J. Kahn, Counting maximal antichains and independent sets, Order, 30, (2013), 427–435.

[65] G. Kalai, Intersection patterns of convex sets, Israel J. Math., 48, (1984), 161–174.

[66] G.O.H. Katona, Intersection theorems for systems of finite sets, Acta Math. Acad. Sci. Hungar., 15, (1964), 329–337.

[67] D. Kleitman, On a lemma of Littlewood and Offord on the distribution of certain sums, Math. Z., 90, (1965), 251–259.

[68] G.O.H. Katona, On a conjecture of Erdős and a stronger form of Sperner's theorem, Stud. Sci. Math. Hungar., 1, (1966), 59–63.

[69] D. Kleitman, Collections of subsets containing no two sets and their union, Proceedings of the LA Meeting AMS (1968), pp. 153–155, Pub. AMS, 1971.

[70] D. Kleitman, A conjecture of Erdős–Katona on commensurable pairs of subsets of an n-set, in Theory of Graphs, Proceedings of the Colloquium Held at Tihany, Hungary, September 1966 (P. Erdős and G. Katona, eds.), Academic Press, New York, 1968 pp. 215–218.

[71] D. Kleitman, On Dedekinds problem: the number of monotone boolean functions, Proc. Amer. Math. Soc., 21, (1969), 677–682.

[72] D. Kleitman, Extremal properties of collections of subsets containing no two sets and their union, J. Combin. Theory Ser. A 20 (1976) 390–392.

[73] D. Kleitman and K.J. Winston, The asymptotic number of lattices, Ann. Discrete Math., 6, (1980), 243–249, Combinatorial mathematics, optimal designs and their applications (Proc. Sympos. Combin. Math. and Optimal Design, Colorado State Univ., Fort Collins, Colo., 1978)

[74] D. Kleitman and K.J. Winston, On the number of graphs without 4-cycles, Discrete Math., 6, (1982), 167–172.

[75] A. D. Korshunov, The number of monotone Boolean functions, Problemy Kibernet., 1981, 38: 5–108

[76] G. Kozma, T. Meyerovitch, R. Morris, R. Peled and W. Samotij, An almost optimal bound on the volume of the metric polytope, in preparation.

[77] G. Kozma, T. Meyerovitch, R. Peled and W. Samotij. "Random points in the metric polytope", in preparation.

[78] G. Kozma, T. Meyerovitch, R. Peled, and W. Samotij, "What does a typical metric space look like?", manuscript.

[79] J. Körner and G. Simonyi, Graph pairs and their entropies: modularity problems, Combinatorica 20 (2000), 227–240.

[80] I. Leader and E. Long, Tilted Sperner families, Discrete Appl. Math., 163, (2014), 194–198.

[81] I. Leader and T. S. Tan, Connected Colourings of Complete Graphs and Hypergraphs, Graphs and Combinatorics 32(1): 257-269 (2016).

[82] P. Loh, Directed paths: from Ramsey to Ruzsa and Szemerédi, submitted.

[83] L. Lovász, Flats in matroids and geometric graphs, Combinatorial surveys (Proc. Sixth British Combinatorial Conf., Royal Holloway Coll., Egham, 1977), pages 45–86, Academic Press London, 1977.

[84] L. Lovász, On the Shannon capacity of a graph, IEEE T. Inform. Theory 25.1 (1979), 1–7.

[85] F. Maffray and M. Preissmann, A translation of Gallai's paper: 'Transitiv Orientierbare Graphen', In: Perfect Graphs (J. L. Ramirez-Alfonsin and B. A. Reed, Eds.), Wiley, New York, 2001, pp. 2566

[86] R. Morris and D. Saxton, The number of C_{2k}-free graphs, to appear in Advances in Math.

[87] F. Mousset, R. Nenadov and A. Steger, On the number of graphs without large cliques, SIAM Journal on Discrete Mathematics 28.4 (2014): 1980–1986.

[88] D. Mubayi and C. Terry, Discrete metric spaces: structure, enumeration, and 0-1 laws, arXiv preprint arXiv:1502.01212 (2015).

[89] D. Mubayi and L. Wang, personal communication.

[90] H. Nguyen and V. Vu, *Optimal inverse Littlewood-Offord theorems*, Adv. Math. **226** (2011), 5298–5319.

[91] J. A. Noel, A. Scott and B. Sudakov, Supersaturation in posets and applications involving the container method, arXiv preprint arXiv:1610.01521 (2016).

[92] F. P. Ramsey, On a problem in formal logic, Proceedings of the London Mathematical Society, vol. 30 (1930), pp. 264–286.

[93] A. Rényi, On random subsets of a finite set, Mathematica (Cluj), 3, No. 26 (1961), 355–362.

[94] M. Saks, Kleitman and combinatorics, Discrete Math., 257, (2002), 225–247.

[95] W. Samotij, Counting independent sets in graphs, European J. Combin., 48, (2015), 5–18.

[96] W. Samotij, Subsets of posets minimising the number of chains, arXiv preprint arXiv:1708.02436 (2017).

[97] A. Sapozhenko, Systems of containers and enumeration problems, Stochastic Algorithms: Foundations and Applications, Springer Berlin Heidelberg, (2005), 1–13.

[98] D. Saxton and A. Thomason, Hypergraph containers, Inventiones mathematicae, (2015), Volume 201, Issue 3, 925–992.

[99] I. Schur, "Über Kongruenz $x^m + y^m = z^m \pmod{p}$", Jahresbericht der Deutschen Mathematiker-Vereinigung 25 (1917): 114–116.

[100] J. Spencer, "Ramsey's theorem – a new lower bound", J. Combin. Theory Ser. A, 18: 108–115, 1975.

[101] E. Sperner, Ein Satz über Untermengen einer endlichen Menge, Mathematische Zeitschrift, 27 (1928), 544–548.

[102] E. Szemerédi, "On sets of integers containing no four elements in arithmetic progression", Acta Mathematica Academiae Scientiarum Hungarica 20.1-2 (1969): 89–104.

[103] A. D. Taylor, *Bounds for the disjoint unions theorem*, J. Combin. Theory Ser. A **30** (1981), 339–344.

[104] A. Tietäväinen, On the existence of perfect codes over finite fields, SIAM J. Appl. Math., 24, (1973), 88–96.

[105] S. Ulam, A Collection of Mathematical Problems, Interscience, New York, 1960, p. 27

[106] A. Z. Wagner, Large Subgraphs in RainbowTriangle Free Colorings, Journal of Graph Theory (2017).

[107] B. L. van der Waerden, “Beweis einer Baudetschen Vermutung”, Nieuw Arch. Wiskunde 15 (1927): 212–216.

[108] C. Villani, Topics in optimal transportation, vol. 58 of Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2003.