University of California
Los Angeles

Extensions of Classic Theorems in Extremal Combinatorics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Mathematics

by

Shagnik Das

2014

Abstract of the Dissertation

Extensions of Classic Theorems in Extremal Combinatorics

by

Shagnik Das

Doctor of Philosophy in Mathematics

University of California, Los Angeles, 2014

Professor Benjamin Sudakov, Chair

Extremal combinatorics deals with the following fundamental question: "how large can a structure be without containing forbidden configurations?" The structures studied are extremely flexible, allowing for a wide range of applications to diverse fields such as theoretical computer science, operations research, discrete geometry and number theory. Moreover, tools from probability theory, algebra and analysis have proven incredibly useful, spurring the development of new techniques in combinatorics. This synergy between different fields has led to incredible growth in recent decades, inspiring numerous directions for research.

In this dissertation, we present new extensions of classic theorems in extremal combinatorics, employing probabilistic and analytic arguments to solve problems connected to number theory and coding theory. In Chapter 2, we greatly improve the bounds for the rainbow Turán problem for even cycles, a problem merging the graph theoretic disciplines of Turán theory and graph colouring. In Chapter 3, we use the analytic method of flag algebras to study a variant of Turán's theorem proposed by Erdős. We then shift to extremal set theory, and in Chapter 4 study the supersaturation problem for the Erdős–Ko–Rado Theorem. In Chapter 5 we discuss a probabilistic measure of supersaturation for intersecting families, introduced recently by Katona, Katona and Katona. These problems represent the various fields within extremal combinatorics that the author has worked in.

The dissertation of Shagnik Das is approved.

Eliezer M. Gafni

Igor Pak

Benjamin Sudakov, Committee Chair

University of California, Los Angeles

2014

To my sister

—

who so selflessly sacrificed

her love of mathematics

so that I may have

twice my fair share

—

and my parents

—

for supporting me

every step of the way

Table of Contents

List of Figures

Acknowledgments

I have encountered a truly marvellous set of people, which this page is too small to contain.

This page should do nicely, though.

First and foremost, I am forever indebted to my adviser, Benny Sudakov, for going well above and beyond the call of duty when it came to offering support and wisdom. If reading this dissertation convinces you to pursue a Ph.D. in combinatorics, then I cannot recommend working with Benny highly enough. My thanks also go to my coauthors, Wenying Gan, Hao Huang, Choongbum Lee, Jie Ma, and Humberto Naves, who not only made our projects possible, but a joy to work on. I am truly grateful to my academic family, including Jacob Fox, Roman Glebov, Dániel Korándi, Po-Shen Loh, and Pedro Vieira, for helpful and encouraging discussions, and for our mutual support of Real Madrid.

Many thanks go to the mathematics departments at UCLA and ETH Zürich for their support over these past five years. I have also appreciated the opportunity to visit and work with József Balogh and his students, Peter Keevash, and Tibor Szabó and his research group, and would like to thank them for their incredible hospitality. I should also like to apologise to József Balogh, Rachel Camina, Igor Pak, and Tibor Szabó for having burdened them with so many reference requests over the years, and am very grateful for their support.

None of this would have been possible without the constant encouragement, love and support of my family. I hope my parents and my sister enjoy reading this dissertation, and look forward to quizzing them on its contents afterwards. I have also been fortunate enough to have learnt from an incredible set of teachers throughout my education. The first two to come to mind are David Hopkins—I shall never forget the algebra lesson he taught me while I was in hospital—and Imre Leader, for inspiring my passion for combinatorics.

I owe a great deal to my friends, who have ensured the graphs were accompanied by laughs. While there have been many who have shared the good times and carried me through the bad times, whose fine reputations I would not wish to sully by naming in this dissertation, I would be committing a great injustice by not celebrating the companionship of Ashay Burungale, Neil Katuna, Sam Miner, Humberto Naves, Bregje Pauwels, Anand Rajagopalan, and Károly Virágh. Finally, special thanks go to Jacques Benatar, for refusing to ghost-write this dissertation—any errors hereinafter may be directly attributed to him.

Chapter 2 is a version of the paper *Rainbow Turán problem for even cycles*, coauthored with Choongbum Lee and Benny Sudakov, which appeared in the European Journal of Combinatorics, Volume 34, in 2013. Chapter 3 is an adaptation of the paper *A problem of Erdős on the minimum number of k-cliques*, coauthored with Hao Huang, Jie Ma, Humberto Naves, and Benny Sudakov, which appeared in the Journal of Combinatorial Theory Series B, Volume 103, in 2013. Chapter 4 is a version of the paper *The minimum number of disjoint pairs in set systems and related problems*, joint work with Wenying Gan and Benny Sudakov, which is to appear in Combinatorica. Chapter 5 has been adapted from the paper *Most probably intersecting hypergraphs*, joint work with Benny Sudakov, which has been submitted for publication and is available on arXiv with identification number arXiv:1312.0840. All these projects were carried out under the supervision of Benny Sudakov.

Vita

2008	B.A.(Hon), Mathematics with Computer Science, University of Cambridge.
2009	M.Math. with Distinction, Mathematics, University of Cambridge.
2009–2013	Teaching Assistant, Department of Mathematics, UCLA.
2010	Instructor, Johns' Hopkins Center for Talented Youth, Hong Kong.
2011	M.A., Mathematics, UCLA.
2013–2014	Teaching Assistant, Department of Mathematics, ETH Zürich.

Publications

S. Das, C. Lee and B. Sudakov, Rainbow Turán problem for even cycles, *Eur. J. Combin.* 34 (2013), 905–915.

S. Das, H. Huang, J. Ma, H. Naves and B. Sudakov, A problem of Erdős on the minimum number of k-cliques, *J. Comb. Theory B* 103 (2013), 344–373.

S. Das, W. Gan and B. Sudakov, Sperner's Theorem and a problem of Erdős, Katona and Kleitman, *Comb. Probab. Comput.*, to appear.

S. Das, W. Gan and B. Sudakov, The minimum number of disjoint pairs in set systems and related problems, *Combinatorica*, to appear.

CHAPTER 1

Introduction

Extremal combinatorics is a central area of discrete mathematics, whose origins can be traced back hundreds of years. However, it is in recent decades that the field has experienced tremendous growth, a spurt that is no doubt due to its deep connections with other disciplines. On the one hand, techniques from probability, algebra and analysis have been applied to combinatorial problems with great success, leading to the development of new areas of study within the subject. On the other hand, much research has been driven by applications to theoretical computer science, operations research, discrete geometry and number theory. This synergy between combinatorics and other areas, both within mathematics and without, continues to grow stronger, inspiring both new directions of research and extensions of classic theorems.

The typical extremal problem takes the following form: "how large can a structure be without containing some forbidden configuration?" One of the major areas of study is extremal graph theory, where the structures studied are graphs, and the configurations to be avoided are forbidden subgraphs. The fundamental result in this direction is Turán's theorem [Tur41], determining the largest number of edges in graphs without large cliques. Another central area is that of extremal set theory, where the structures are families of sets, and the forbidden configurations are restricted intersection patterns. The prototypical result is the Erdős–Ko–Rado theorem [EKR61], which gives the size of the largest uniform intersecting family of sets. As is befitting of such important theorems, countless extensions and variations have been developed. We shall discuss some of these in the following chapters, but the interested reader may wish to consult the books of Bollobás [Bol04] and Jukna [Juk01] for a detailed study of extremal graph theory, while the monograph of Babai–Frankl [BF92] and

the book of Anderson [And87] are excellent resources for extremal set theory.

In this dissertation, we strengthen these celebrated results in extremal combinatorics, sometimes through the use of probabilistic or analytic methods, and in the process answer some questions arising from fields such as number theory or coding theory. We provide below a brief overview of these results, although a more thorough discussion of the background and motivation of each problem is contained in the introductions of the relevant chapters.

Our first extension concerns the rainbow Turán problem, first introduced by Keevash, Mubayi, Sudakov and Verstraëte [KMS07]. As suggested by its name, this problem combines two key areas of extremal graph theory, namely Turán theory and graph colouring.

Turán theory asks how many edges a graph on n vertices may have if it does not contain a forbidden subgraph H. Turán [Tur41] resolved the problem in the case of H being a complete graph. This was later extended by Erdős and Stone [ES46], who showed that the Turán number of a graph H is essentially determined by its chromatic number. This result resolves the problem asymptotically except when H is bipartite. To this day, bipartite Turán problems remain widely open, with very few bounds known. One such bound was found by Bondy and Simonovits [BS74], who gave an upper bound on the number of edges in a graph without an even cycle of given length.

Graph colouring problems also enjoy a long history. We say a graph is properly edge-coloured if no two edges sharing a vertex receive the same colour. Given a proper colouring, one usually is interested in finding rainbow subgraphs, where every edge is of a different colour. The Canonical Ramsey Theorem of Erdős and Rado [ER50] implies that for any finite number of colours, every proper colouring of the edges of a sufficiently large complete graph produces a rainbow copy of a complete graph on k vertices.

The rainbow Turán problem for a given graph H fuses these two areas together by asking how many edges a properly edge-coloured n-vertex graph may have without containing a rainbow copy of H. In the original paper [KMS07], Keevash et al. resolve the problem for non-bipartite graphs. They also demonstrate the connection between a problem in number theory and the rainbow Turán problem for even cycles, highlighting the latter as the most

interesting open case of this problem. In Chapter 2 we combine probabilistic and structural arguments to greatly improve the best-known bounds on the rainbow Turán number of even cycles, and also provide better bounds on the size of graphs without any rainbow cycles.

We return now to Turán's theorem [Tur41], which determines the largest n-vertex graphs not containing a complete subgraph on ℓ vertices. While the original theorem was concerned with maximising the number of edges, or complete subgraphs on two vertices, Zykov [Zyk49] showed that the same graphs also maximise the number of complete subgraphs on any number of vertices. By considering the complementary graph, Turán's theorem states that these graphs have the fewest edges while not containing an independent set of size ℓ. Erdős [Erd62b] conjectured that, just as in Zykov's theorem, these complementary graphs also minimise the number of cliques of size k, giving rise to what we call the (k, ℓ)-problem. However, Nikiforov [Nik01] disproved the conjecture for the $(4, 3)$-problem, and further showed the conjecture could only hold for finitely many pairs (k, ℓ) when $k, \ell \geq 3$.

This appears to be a very difficult problem to solve in full generality, and a complete solution would seem to be beyond our current means. However, the method of flag algebras, recently introduced by Razborov [Raz07], allows us to solve particular cases of the (k, ℓ)-problem. This method translates extremal problems into semidefinite programming problems, which can be solved numerically. These numerical solutions provide asymptotic bounds for the extremal problem, although often much work is required to find a rational solution that is not subject to rounding errors. When these bounds match those coming from constructions, one has an asymptotic solution to the original problem. A detailed introduction to the method is included in Section 3.3. For a survey of the many results it has led to, see [Raz13].

In Chapter 3, we attack the (k, ℓ)-problem with the machinery of flag algebras. By carefully analysing the output from the semidefinite programming problems, we obtain stability results that further allow us to precisely characterise the extremal graphs. In particular, we verify the Erdős conjecture for the $(3, 4)$-problem, while showing that Nikiforov's counterexample is optimal for the $(4, 3)$-problem. Finally, we combine random and explicit constructions to provide much sharper bounds on when the conjecture of Erdős can hold.

The remaining results in this dissertation concern supersaturation problems in extremal set theory, which we now describe. Recall that the typical extremal result bounds how large a structure can be if it does not contain a forbidden configuration. Equivalently, it asserts that larger structures must contain at least one forbidden configuration. A natural strengthening is to then ask how many such forbidden configurations actually appear in larger structures, and this is known as the supersaturation problem. These problems have long guided research in extremal combinatorics – indeed, in the context of graph theory, Razborov's development of flag algebras was motivated by the much-studied supersaturation problem for the minimum number of triangles in graphs with many edges [Raz08].

Supersaturation problems for extremal set theory have also received much attention over the years. It is well-known that the largest intersecting family of subsets of an n-element universe has size 2^{n-1}. Frankl [Fra77] and, independently, Ahlswede [Ahl80] studied the problem of how many disjoint pairs of sets must appear in larger families, showing that it is optimal to take sets as large as possible. This determines the large-scale structure of the extremal families, but the exact solution depends on which sets are chosen in the final level. Ahlswede [Ahl80] explicitly asked which k-uniform set families minimise the number of disjoint pairs, a problem which may also be thought of as the supersaturation problem for the celebrated Erdős–Ko–Rado theorem [EKR61].

In Chapter 4, we study supersaturation for the Erdős–Ko–Rado theorem. Utilising shifting arguments, we are able to determine the exact solutions for a wide range of parameters, and furthermore classify the extremal families. In doing so, we partially prove a conjecture of Bollobás and Leader [BL03]. Our methods also provide similar results for some classic extensions of the Erdős–Ko–Rado theorem, allowing us to determine both the minimum number of matchings of size q and the minimum number of pairs intersecting in fewer than t elements. As a special case, the latter result provides a partial solution to a problem of Kleitman and West from coding theory.

Recently, Katona, Katona and Katona [KKK12] introduced a probabilistic notion of supersaturation. The standard supersaturation problem asks how many forbidden configurations must appear beyond the extremal threshold. In the probabilistic variant, we instead

consider random substructures of large structures, and seek to maximise the probability that these substructures avoid any forbidden configurations. Note that below the extremal threshold, it is possible for the original structure itself to be free of any forbidden configurations, and hence this probability is one. Thus, just as in the original supersaturation problem, we are only interested in structures larger than the corresponding extremal bound.

In their original paper, Katona et al. [KKK12] focus on the probabilistic supersaturation problem for non-uniform intersecting families, showing that within a certain rainge it is again optimal to take large sets. This was furthered by Russell [Rus12] and Russell and Walters [RW13], who extended the result to a much wider range of family sizes. In the latter paper, it was shown that there is no nested sequence of extremal families, thus disproving a conjecture of Katona et al.

Despite these results, very little was known in the uniform setting. In Chapter 5, we extend our methods from Chapter 4 to handle probabilistic supersaturation for uniform families. In particular, we show that for a range of parameters, the solutions to the two problems coincide. However, the probabilistic version is a much more delicate problem, and in certain ranges the answer can vary depending on the underlying probability.

Each of the subsequent chapters will contain its own introduction, both covering the background and motivation of the relevant problem in greater detail, and formally presenting our results. Any specific notation will be introduced in the individual chapters, but below we define the standard notation used throughout this dissertation.

Notation: We denote by $[n]$ the first n natural numbers $\{1, 2, \ldots, n\}$. Our set families, typically denoted $\mathcal{F}$, shall be collections of subsets of the ground set, usually $[n]$. Given a set X, $\binom{X}{k}$ represents the family of all k-subsets of X. A graph $G = (V, E)$ is an ordered pair of a set V of vertices and a set $E \subset \binom{V}{2}$ of edges. Given a subset $U \subset V$ of the vertices, we write $G[U]$ for the subgraph induced by the vertices in U. Finally, we make use of asymptotic notation throughout this paper. Given two functions $f, g : \mathbb{N} \to \mathbb{R}$, we write $f(n) \sim g(n)$ if $\lim_{n\to\infty} f(n)/g(n) = 1$. If there is a constant $C > 0$ such that $f(n) \leq Cg(n)$, we write $f(n) = O(g(n))$, and if there is $c > 0$ such that $f(n) \geq cg(n)$, we write $f(n) = \Omega(g(n))$.

CHAPTER 2

Rainbow Turán problem for even cycles

2.1 Introduction

An edge-coloured graph is *rainbow* if all its edges have distinct colours. The rainbow Turán problem, first introduced by Keevash, Mubayi, Sudakov and Verstraëte [KMS07], asks the following question: given a fixed graph H, what is the maximum number of edges in a properly edge-coloured graph G on n vertices with no rainbow copy of H? This maximum is denoted $\mathrm{ex}^*(n, H)$, and is called the rainbow Turán number of H. In this chapter, we study the rainbow Turán problem for even cycles.

2.1.1 Background

The rainbow Turán problem has a certain aesthetic appeal, as it lies at the intersection of two key areas of extremal graph theory. On the one hand we have the classical Turán problem, which, for a given graph H, asks for the maximum number of edges in an H-free graph on n vertices. This maximum, the Turán number of H, is denoted by $\mathrm{ex}(n, H)$, and determining it is one of the oldest problems in extremal combinatorics. Turán [Tur41] solved the problem for cliques by finding $\mathrm{ex}(n, K_k)$. Erdős and Stone [ES46] then found the asymptotics of $\mathrm{ex}(n, H)$ for all non-bipartite graphs H. The problem of determining the Turán numbers of bipartite graphs is still largely open. Of particular interest is the case of even cycles. Erdős conjectured that $\mathrm{ex}(n, C_{2k}) = \Theta(n^{1+\frac{1}{k}})$. Bondy and Simonovits [BS74] gave the corresponding upper bound, but as of yet a matching lower bound is only known for $k = 2, 3$, or 5.

On the other hand, there is a great deal of literature on extremal problems regarding

(not necessarily proper) edge-coloured graphs. The Canonical Ramsey Theorem of Erdős and Rado [ER50] shows, as a special case, that when n is large with respect to t, then any proper edge-colouring of K_n contains a rainbow K_t. Another variation is when one allows at most k colours to be used for edges incident to each vertex. This notion, called local k-colourings, has been first introduced by Gyárfás, Lehel, Schelp, and Tuza [GLS87], and has been studied in a series of works. More recently, Alon, Jiang, Miller and Pritikin [AJM03] studied the problem of finding a rainbow copy of a graph H in an edge-colouring of K_n where each colour appears at most m times at any vertex. The rainbow Turán problem is a Turán-type extension in the case $m = 1$. From this point on, we shall only consider proper edge-colourings.

The rainbow Turán problem for even cycles is of particular interest because of the following connection to a problem in number theory, as noted in [KMS07]. Given an Abelian group Γ, a subset A is called a B_k^*-set if it does not contain disjoint k-sets B, C with the same sum. Given a set A, we form a bipartite graph G as follows: the two parts X and Y are copies of Γ, and we have an edge from $x \in X$ to $y \in Y$ if and only if $x - y \in A$. Moreover, the edge xy is given the colour $x - y \in A$. It is easy to see that this is a proper edge-colouring of a graph with $|\Gamma||A|$ edges, and A is a B_k^*-set precisely when G has no rainbow C_{2k}. Hence bounds on B_k^*-sets give bounds on $\mathrm{ex}^*(n, C_{2k})$, and vice versa.

2.1.2 Known results

Note that we trivially have the lower bound $\mathrm{ex}(n, H) \leq \mathrm{ex}^*(n, H)$, since if a graph is H-free, then it is rainbow-H-free under any proper edge colouring. One is thus generally interested in either finding a matching upper bound, or showing that $\mathrm{ex}^*(n, H)$ is asymptotically larger than $\mathrm{ex}(n, H)$ by a multiplicative constant. In the original paper of Keevash, Sudakov, Mubayi and Verstraëte [KMS07], this problem was resolved for a wide range of graphs. In particular, it was shown that for non-bipartite H, the Rainbow Turán problem can be reduced to the Turán problem, and as a result $\mathrm{ex}^*(n, H)$ is asymptotically (and in some cases exactly) equal to $\mathrm{ex}(n, H)$. For bipartite H with a maximum degree of s in one of the parts,

they found an upper bound of $\mathrm{ex}^*(n, H) = O(n^{2-\frac{1}{s}})$. This matches the general upper bound for Turán numbers of such graphs, and in particular is tight for C_4 (where $s = 2$).

An interesting case which is not implied by the above mentioned results is the case of even cycles of length at least 6, and special attention was paid to this case, in light of the connection to B_k^*-sets discussed earlier. Using Bose and Chawla's [BC62] construction of large B_k^*-sets, the authors gave a lower bound of $\mathrm{ex}^*(n, C_{2k}) = \Omega(n^{1+\frac{1}{k}})$ - this is better than the best known bound for $\mathrm{ex}(n, C_{2k})$ for general k. A matching upper bound was obtained in the case of the six-cycle C_6, so it is known that $\mathrm{ex}^*(n, C_6) = \Theta(n^{1+\frac{1}{3}})$. However, surprisingly, $\mathrm{ex}^*(n, C_6)$ is asymptotically larger than $\mathrm{ex}(n, C_6)$ by a multiplicative constant.

Another problem considered was that of rainbow acyclicity - what is the maximum number of edges in an edge-coloured graph on n vertices with no rainbow cycle of any length? Let $f(n)$ denote this maximum. In the uncoloured setting, the answer is given by a tree, which has $n - 1$ edges. However, as described in [KMS07], colouring the d-dimensional hypercube with d colours, where parallel edges get the same colour, gives a rainbow acyclic proper edge-colouring, and hence $f(n) = \Omega(n \ln n)$. The best known upper bound to date was $f(n) = O(n^{1+\frac{1}{3}})$, which follows from the bound $\mathrm{ex}^*(n, C_6) = \Theta(n^{1+\frac{1}{3}})$.

Keevash, Mubayi, Sudakov, and Verstraëte listed the questions of determining $\mathrm{ex}^*(n, C_{2k})$ and $f(n)$ as the two most interesting open problems in the study of rainbow Turán numbers.

2.1.3 Our results

In this chapter we improve the upper bound on the rainbow Turán number of even cycles, and make progress towards the two open problems mentioned in the previous subsection. Following is the main theorem of this chapter:

Theorem 2.1.1. *For every fixed $\varepsilon > 0$ there is a constant $C(\varepsilon)$ such that any properly edge-coloured graph on n vertices with at least $C(\varepsilon)n^{1+\varepsilon}$ edges contains a rainbow copy of an even cycle of length at most $2k$, where $k = \left\lceil \frac{\ln 4 - \ln \varepsilon}{\ln(1+\varepsilon)} \right\rceil$.*

Our result easily gives an upper bound on the size of rainbow acyclic graphs.[1]

[1] As we remark in Section 2.5, one can do somewhat better than this corollary.

Corollary 2.1.2. *Let $f(n)$ denote the size of the largest properly edge-coloured graph on n vertices that contains no rainbow cycle. Then for any fixed $\varepsilon > 0$ and sufficiently large n, we have $f(n) < n^{1+\varepsilon}$.*

With a little more work, we can show that a graph satisfying the condition of Theorem 2.1.1 must contain a rainbow cycle of length exactly $2k$. Therefore inverting the relationship between k and ε gives a bound on $\mathrm{ex}^*(n, C_{2k})$.

Corollary 2.1.3. *For every fixed integer $k \geq 2$, $\mathrm{ex}^*(n, C_{2k}) = O\left(n^{1+(1+\varepsilon_k)\ln k/k}\right)$, where $\varepsilon_k \to 0$ as $k \to \infty$.*

2.1.4 Outline and notation

This chapter is organised as follows. Section 2.2 provides a couple of quick probabilistic lemmas. The proof of Theorem 2.1.1 is then given in Section 2.3, although the proof of the key proposition is deferred until Section 2.4. The final section contains some further remarks and open problems.

A graph G is given by a pair of vertex set $V(G)$ and edge set $E(G)$. For a vertex $v \in V(G)$, we use $d(v)$ to denote its degree, and for a subset of vertices X, we let $d(v, X)$ be the number of neighbors of v in the set X. We use the notation $\mathrm{Bin}(n, p)$ to denote a binomial random variable with parameters n and p. Throughout the chapter, log is used for the logarithm function of base 2, and ln is used for the natural logarithm.

2.2 Preliminary lemmas

In this section we will prove a couple of technical lemmas that will be used in our proof of Theorem 2.1.1. Both will be proven using the probabilistic method, and will rely on the following form of Hoeffding's Inequality as appears in [McD98, Theorem 2.3].

Theorem 2.2.1. *Let the random variables $X_1, X_2, \ldots, X_k$ be independent, with $0 \leq X_i \leq 1$*

for each i. Let $S = \sum_{i=1}^{k} X_i$, and $\mu = \mathbb{E}[S]$. Then for any $s \leq \frac{1}{2}\mu$ and $t \geq 2\mu$, we have

$$\mathbf{P}(S \leq s) \leq \exp\left(-\frac{s}{4}\right) \qquad \text{and} \qquad \mathbf{P}(S \geq t) \leq \exp\left(-\frac{3t}{16}\right).$$

Our first lemma asserts that for any edge-coloured graph with large minimum degree, the colours of the graph can be partitioned into disjoint classes in such a way that for every colour class, the edges using colours from that class form a subgraph with large minimum degree.

Lemma 2.2.2. *Let G be an edge-coloured graph on n vertices with minimum degree δ, and let k be a positive integer. Let $\mathcal{C}$ be the set of colours in G. If $nk \exp\left(-\frac{\delta}{8k}\right) < 1$, then there is a partition $\mathcal{C} = \bigsqcup_{i=1}^{k} \mathcal{C}_i$ such that for every vertex v and colour class $\mathcal{C}_i$, v has at least $\frac{\delta}{2k}$ edges with colours from $\mathcal{C}_i$.*

Proof. Independently and uniformly at random assign each colour $c \in \mathcal{C}$ to one of the k colour classes $\mathcal{C}_i$. We will show that the resulting partition has the desired property with positive probability.

Fix a vertex v and a colour class $\mathcal{C}_i$. Let $d(v)$ be the degree of v in G, and let $d_{v,i}$ denote the number of edges incident to v that have a colour from $\mathcal{C}_i$. Note that the colour of every edge is in $\mathcal{C}_i$ with probability $\frac{1}{k}$. Moreover, since the colouring is proper, the edges incident to v have distinct colours, and hence are in $\mathcal{C}_i$ independently of one another. Thus $d_{v,i} \sim \text{Bin}\left(d(v), \frac{1}{k}\right)$, and $\mathbb{E}[d_{v,i}] = \frac{d(v)}{k} \geq \frac{\delta}{k}$ by our assumption on the minimum degree.

By Theorem 2.2.1, we have

$$\mathbf{P}\left(d_{v,i} \leq \frac{\delta}{2k}\right) \leq \exp\left(-\frac{\delta}{8k}\right).$$

By a union bound,

$$\mathbf{P}\left(\exists v, i : d_{v,i} \leq \frac{\delta}{2k}\right) \leq nk \exp\left(-\frac{\delta}{8k}\right) < 1,$$

and hence $\mathbf{P}\left(\forall v, i : d_{v,i} > \frac{\delta}{2k}\right) > 0$. Thus the desired partition exists. □

Given a set X with a family of small subsets, the second lemma allows us to choose a subset of X of specified size while retaining control over the sizes of the subsets.

Lemma 2.2.3. *Let $\beta, \gamma \in (0,1)$ be parameters. Suppose we have a set X and a collection of subsets X_j, $1 \leq j \leq m$, such that $|X_j| \leq \beta|X|$ for each j. Provided $3m \exp\left(-\frac{1}{8}\beta\gamma|X|\right) < 1$, there exists a subset $Y \subset X$ with $\frac{1}{2}\gamma|X| \leq |Y| \leq 2\gamma|X|$ such that for every j, we have $|X_j \cap Y| \leq 4\beta|Y|$.*

Proof. Let Y be the random subset of X obtained by selecting each element independently with probability γ. Let $Y_j = X_j \cap Y$. Then we have $|Y| \sim \text{Bin}(|X|, \gamma)$, and $|Y_j| \sim \text{Bin}(|X_j|, \gamma)$.

By Theorem 2.2.1,

$$\mathbf{P}\left(|Y| \leq \frac{1}{2}\gamma|X|\right) \leq \exp\left(-\frac{1}{8}\gamma|X|\right), \qquad \text{and} \qquad \mathbf{P}\left(|Y| \geq 2\gamma|X|\right) \leq \exp\left(-\frac{3}{8}\gamma|X|\right).$$

Since $\mathbb{E}[|Y_j|] = \gamma|X_j| \leq \beta\gamma|X|$, Theorem 2.2.1 also gives

$$\mathbf{P}\left(|Y_j| \geq 2\beta\gamma|X|\right) \leq \exp\left(-\frac{3}{8}\beta\gamma|X|\right).$$

By a union bound, the probability of any of these events occurring can be bounded by

$$\exp\left(-\frac{1}{8}\gamma|X|\right) + \exp\left(-\frac{3}{8}\gamma|X|\right) + m\exp\left(-\frac{3}{8}\beta\gamma|X|\right) \leq 3m\exp\left(-\frac{1}{8}\beta\gamma|X|\right) < 1.$$

Hence, with positive probability, none of these events occur. In this case we have a subset $Y \subset X$ with $\frac{1}{2}\gamma|X| < |Y| < 2\gamma|X|$ and $|X_j \cap Y| < 2\beta\gamma|X| < 4\beta|Y|$, as required. □

2.3 Proof of the main theorem

We will restrict our attention to bipartite graphs, and prove Theorem 2.1.1 for bipartite graphs by using induction within this class. The theorem for general graphs will then easily follow since every graph contains a bipartite subgraph that contains at least half of its original edges.

Our general strategy for proving Theorem 2.1.1 is as follows. We will choose an arbitrary vertex v_0, and grow a subtree T of G rooted at v_0. This subtree will have the property that every path from v_0 in T will be rainbow. The key proposition will show that if G has no

short rainbow cycles, then the levels of the tree must grow very rapidly, and will eventually need to be larger than G, which is impossible.

In this section we formalise this argument, although the proof of the key proposition is deferred to the next section.

Proof of Theorem 2.1.1. Fix $\varepsilon > 0$. Without loss of generality, we may assume $\varepsilon < \frac{1}{2}$, as otherwise the result follows from the bound of $\mathrm{ex}^*(n, C_{2k}) = O\left(n^{2-\frac{1}{s}}\right)$ (with $s = 2$) given in [KMS07]. We wish to show there is a constant C such that any edge-coloured bipartite graph G on n vertices with at least $Cn^{1+\varepsilon}$ edges contains a rainbow cycle of length at most $2k$, where $k = \left\lceil \frac{\ln 4 - \ln \varepsilon}{\ln(1+\varepsilon)} \right\rceil$.

We will prove this by induction on n. For the base case, note that if $n \leq C$, then $Cn^{1+\varepsilon} > n^2$. Hence there is no graph on n vertices with $Cn^{1+\varepsilon}$ edges, and so the statement is vacuously true. Thus by making the constant C large, we force n to be large in the induction step below. In particular, we will require $C > 8k$ to be large enough that every $n \geq C$ satisfies the following inequalities:

$$nk \exp\left(-n^{\varepsilon}\right) < 1, \quad n^{\frac{1}{4}\varepsilon^3} > [4(k+1)]^{2+\varepsilon} \log n, \quad \text{and} \quad n^{\frac{1}{2}\varepsilon^2} > 2^{7+(3k+2)\varepsilon} k^{2+\varepsilon} (\log n)^{1+k\varepsilon}.$$

Now suppose $n > C$, and G has at least $Cn^{1+\varepsilon}$ edges. If G has a vertex of degree at most Cn^{ε}, then by removing it we have a subgraph on $n-1$ vertices with at least $Cn^{1+\varepsilon} - Cn^{\varepsilon} > C(n-1)^{1+\varepsilon}$ edges. By induction, this subgraph contains a rainbow cycle of length at most $2k$. Hence we may assume G has minimum degree at least Cn^{ε}.

We now apply Lemma 2.2.2. By our bound on C, we have $nk \exp\left(-\frac{Cn^{\varepsilon}}{8k}\right) < 1$. Hence we can split the colours into disjoint classes $\mathcal{C}_i$, $1 \leq i \leq k$, such that for each class $\mathcal{C}_i$, every vertex is incident to at least $\frac{C}{2k} n^{\varepsilon}$ edges of a colour in $\mathcal{C}_i$.

Let v_0 be an arbitrary vertex in G. We will construct a subtree T rooted at v_0, with vertices arranged in levels L_i, starting with $L_0 = \{v_0\}$. Given a level L_i, the next level L_{i+1} will be a carefully chosen subset of neighbors of L_i using just the edges with colours from $\mathcal{C}_{i+1}$. Note that this ensures that every vertex has a rainbow path back to v_0 in T. Moreover, since every vertex in L_i has a path of length i back to v_0, and G is bipartite, it follows that

L_i is an independent set in G. It is useful to parameterise the size of the levels by defining α_i such that $|L_i| = n^{\alpha_i}$.

As mentioned above, every vertex $v \in T$ has a rainbow path back to v_0. It will be important to keep track of which colours are used on this path. Hence for every colour c and level i, we define $X_{i,c}$ to be the vertices in L_i with an edge of colour c in their path back to v_0. Since the path from v to v_0 has length i, it follows that $\{X_{i,c}\}_c$ forms an i-fold cover of L_i. If we have a vertex $w \in L_{i+1}$ adjacent to $v_1, v_2 \in L_i$ with v_1 and v_2 using disjoint sets of colours on their paths back to v_0, then this gives a rainbow cycle of length $2(i+1)$. Hence there must be some overlap in the colours on their paths back to v_0. It turns out that this implies large expansion from L_i to L_{i+1}.

The key proposition below formalises the above observation and shows that the levels grow quickly. As shown below, we will need to maintain control over the sets $X_{i,c}$. To see the necessity of this, suppose that we had $X_{i,c} = L_i$ for some i and c. Then every path through L_i to v_0 would use the colour c, and we could not hope to find a rainbow cycle using our strategy. Note that in the special case where the given graph is Cn^{ε}-regular and the graph is coloured using exactly Cn^{ε} colours, for every index i, there exists a colour c such that $|X_{i,c}| \geq \frac{|L_i|}{Cn^{\varepsilon}} = \Omega(n^{\alpha_i - \varepsilon})$. This implies that we cannot hope for a upper bound on $|X_{i,c}|$ that is better than $|X_{i,c}| = O(n^{\alpha_i - \varepsilon})$. The bound we achieve in the following proposition is a poly-logarithmic factor off this 'optimal' bound.

Proposition 2.3.1. *Given $1 \leq i < k$, suppose that we are given sets $L_0, \cdots, L_i$ and sets $\{X_{i,c}\}_c$ satisfying the following:*

(i) $|L_i| \geq \frac{1}{4}|L_j|$ for $0 \leq j < i$, and $\alpha_i \leq 1 - \frac{1}{4}\varepsilon^2$, and

(ii) $|X_{i,c}| \leq (8 \log n)^i n^{\alpha_i - \varepsilon}$ for all $c \in \mathcal{C}$.

Then there is a set L_{i+1} of neighbors of L_i using colours from $\mathcal{C}_{i+1}$ such that:

1. *$\left(1 + \frac{\varepsilon}{2}\right) - \alpha_{i+1} \leq (1+\varepsilon)^{-1}\left[\left(1 + \frac{\varepsilon}{2}\right) - \alpha_i\right]$, and*
2. *for all colours c, we have $|X_{i+1,c}| \leq (8 \log n)^{i+1} n^{\alpha_{i+1} - \varepsilon}$.*

Moreover, even if we have (ii') $|X_{i,c}| \leq 4(8\log n)^i n^{\alpha_i-\varepsilon}$ *instead of* (ii)*, we can still find a set* L_{i+1} *satisfying Property 1.*

This proposition will be proven in Section 2.4. Here we show how to prove Theorem 2.1.1 using this proposition. We first show how to construct sets L_0, L_1, and $\{X_{1,c}\}_c$. For $i = 0$, as mentioned above, we have $L_0 = \{v_0\}$ and thus $\alpha_0 = 0$. Note that v_0 has at least $\frac{C}{2k}n^\varepsilon$ neighbors with edge colours from $\mathcal{C}_1$. Let L_1 be these neighbors. Then we have $|L_1| = n^{\alpha_1} \geq \frac{C}{2k}n^\varepsilon$, and so $\alpha_1 \geq \varepsilon$. Hence $\left(1+\frac{\varepsilon}{2}\right) - \alpha_1 \leq 1 - \frac{\varepsilon}{2} < (1+\varepsilon)^{-1}\left[\left(1+\frac{\varepsilon}{2}\right) - \alpha_0\right]$. Since v_0 has at most one edge of each colour, we have $|X_{1,c}| \leq 1 < (8\log n)^1 n^{\alpha_1-\varepsilon}$. Now we can iteratively apply Proposition 2.3.1 to construct sets L_i and $X_{i,c}$ for $i = 2, \cdots, k$ as long as $\alpha_{i-1} \leq 1 - \frac{1}{4}\varepsilon^2$. Note that Property 1 above ensures that Condition (i) is always satisfied with every iteration.

Suppose we successfully construct the sets $L_0, L_1, \ldots, L_k$ by repeatedly applying Proposition 3.1. Recalling that $\alpha_0 = 0$, we get

$$\left(1+\frac{\varepsilon}{2}\right) - \alpha_k \leq (1+\varepsilon)^{-1}\left[\left(1+\frac{\varepsilon}{2}\right) - \alpha_{k-1}\right] \leq \ldots \leq (1+\varepsilon)^{-i}\left[\left(1+\frac{\varepsilon}{2}\right) - \alpha_0\right],$$

and so

$$\alpha_k \geq \left(1+\frac{\varepsilon}{2}\right)\left(1-(1+\varepsilon)^{-k}\right).$$

Substituting $k = \left\lceil \frac{\ln 4 - \ln\varepsilon}{\ln(1+\varepsilon)} \right\rceil$, we have

$$\alpha_k \geq \left(1+\frac{\varepsilon}{2}\right)\left(1-\frac{1}{4}\varepsilon\right) \geq 1 + \frac{1}{8}\varepsilon,$$

and so $|L_k| = n^{\alpha_k} \geq n^{1+\frac{1}{8}\varepsilon}$. Thus $|L_k| > n$, which gives the necessary contradiction.

Hence there must be some $i < k$ such that $1 - \frac{1}{4}\varepsilon^2 < \alpha_i \leq 1$. The sizes of the sets $X_{i,c}$ satisfy $|X_{i,c}| \leq (8\log n)^i n^{\alpha_i-\varepsilon} = (8\log n)^i n^{-\varepsilon}|L_i|$. Note that the total number of colours is $m = |\mathcal{C}| < n^2$, since there cannot be more colours than edges in G. Apply Lemma 2.2.3 with $X = L_i$, subsets $X_{i,c}$ for all $c \in \mathcal{C}$, $\beta = (8\log n)^i n^{-\varepsilon}$ and $\gamma = \frac{1}{2}n^{1-\frac{1}{4}\varepsilon^2-\alpha_i}$. This is possible since

$$3m\exp\left(-\frac{1}{8}\beta\gamma|L_i|\right) < 3n^2\exp\left(-\frac{1}{16}(8\log n)^i n^{1-\varepsilon-\frac{1}{4}\varepsilon^2}\right) < 1.$$

We obtain a set $Y \subset L_i$ such that $\frac{1}{2}\gamma|L_i| \leq |Y| \leq 2\gamma|L_i|$ and $|Y \cap X_{i,c}| \leq 4\beta|Y|$ for all c. Note that $\frac{1}{4}n^{1-\frac{1}{4}\varepsilon^2} \leq |Y| \leq n^{1-\frac{1}{4}\varepsilon^2}$ and $|X_{i,c} \cap Y| \leq 4(8\log n)^i|Y|n^{-\varepsilon}$. Moreover, since we must have had $\alpha_{i-1} \leq 1 - \frac{1}{4}\varepsilon^2$, we have $|Y| \geq \frac{1}{4}|L_j|$ for all $0 \leq j < i$. Let $L'_i = Y$, $|L'_i| = n^{\alpha'_i}$, and let $X'_{i,c} = X_{i,c} \cap Y$. Then the above inequalities imply $1 - \frac{1}{3}\varepsilon^2 < 1 - \frac{1}{4}\varepsilon^2 - \frac{2}{\log n} \leq \alpha'_i \leq 1 - \frac{1}{4}\varepsilon^2$, and $|X'_{i,c}| \leq 4(8\log n)^i n^{\alpha'_i - \varepsilon}$. We can now apply Proposition 2.3.1 to the sets L'_i and $X'_{i,c}$. This gives the next level L_{i+1} with

$$\left(1 + \frac{\varepsilon}{2}\right) - \alpha_{i+1} \leq (1+\varepsilon)^{-1}\left[\left(1 + \frac{\varepsilon}{2}\right) - \alpha'_i\right] \leq (1+\varepsilon)^{-1}\left[\frac{\varepsilon}{2} + \frac{\varepsilon^2}{3}\right],$$

and so $\alpha_{i+1} \geq 1 + \frac{\varepsilon^2}{6(1+\varepsilon)}$. Again, this implies $|L_{i+1}| \geq n^{1+\frac{\varepsilon^2}{6(1+\varepsilon)}} > n$, which is a contradiction.

Thus G must have a rainbow cycle of length at most $2k$, which completes the inductive step, and hence the proof of Theorem 2.1.1. □

2.4 Proof of Proposition 2.3.1

In this section, we furnish a proof of Proposition 2.3.1. Our goal is to construct the level L_{i+1} with associated sets $X_{i+1,c}$ satisfying the following properties:

1. $\left(1 + \frac{\varepsilon}{2}\right) - \alpha_{i+1} \leq (1+\varepsilon)^{-1}\left[\left(1 + \frac{\varepsilon}{2}\right) - \alpha_i\right]$, and
2. for all colours c, we have $|X_{i+1,c}| \leq (8\log n)^{i+1} n^{\alpha_{i+1} - \varepsilon}$.

Proof of Proposition 2.3.1. Suppose that $1 \leq i \leq k-1$, and levels L_j for $j \leq i$ satisfy Properties (i) and (ii) given in Proposition 2.3.1. Recall that by the inductive hypothesis, we know that Theorem 2.1.1 is true for any graph whose number of vertices n' is less than n. Thus we may assume that all the subgraphs of G on n' vertices contain at most $C[n']^{1+\varepsilon}$ edges (otherwise we would already have a rainbow cycle of length at most $2k$). Using this, we will show how to construct the level L_{i+1} satisfying both properties.

Consider the edges of colours from $\mathcal{C}_{i+1}$ coming out of L_i. Each vertex in L_i has at least $\frac{C}{2k}n^{\varepsilon}$ such edges; importantly, we will use only $\frac{C}{2k}n^{\varepsilon}$ of them, and disregard any additional edges. The reason we expand the levels 'slowly' in such a way is to prevent some of the sets

$X_{i,c}$ from expanding too fast. Indeed, if we were to use all the edges, then some $X_{i,c}$ might expand faster than we would wish, and this eventually might violate Property 2.

Thus we have a total of $\frac{C}{2k}|L_i|n^\varepsilon$ edges. If at least half of these edges went back to vertices in $L_0 \cup L_1 \cup \ldots \cup L_{i-1}$, then the vertices in $L_0 \cup L_1 \cup \ldots \cup L_i$ would span at least $\frac{C}{4k}|L_i|n^\varepsilon$ edges. This gives us a graph on at most $4k|L_i|$ vertices with at least $\frac{C}{4k}|L_i|n^\varepsilon$ edges. By the inductive hypothesis, we have

$$\frac{C}{4k}|L_i|n^\varepsilon \leq C\left[4k|L_i|\right]^{1+\varepsilon},$$

which is equivalent to

$$\left(\frac{n}{|L_i|}\right)^\varepsilon = n^{(1-\alpha_i)\varepsilon} \leq (4k)^{2+\varepsilon}.$$

However, by the condition that $\alpha_i \leq 1 - \frac{1}{4}\varepsilon^2$, this contradicts our bound on n.

Hence we may assume that at least $\frac{C}{4k}|L_i|n^\varepsilon$ edges go to vertices not in $L_0 \cup L_1 \cup \ldots \cup L_{i-1}$; call this set of new vertices Y. Partition the vertices in Y into $\log n$ sets Y_j, $0 \leq j \leq \log n - 1$, with $y \in Y_j$ if and only if $2^j \leq d(y, L_i) < 2^{j+1}$ (here we are only considering edges of a colour from $\mathcal{C}_{i+1}$). By the pigeonhole principle, there is some j^* such that Y_{j^*} receives at least $\frac{C}{4k\log n}|L_i|n^\varepsilon$ edges from L_i. Let $L_{i+1} = Y_{j^*}$, and for convenience define $d = 2^{j^*}$. As always, we will define α_{i+1} by $|L_{i+1}| = n^{\alpha_{i+1}}$. Let $\delta_i = \alpha_{i+1} - \alpha_i$.

Every vertex $y \in L_{i+1}$ has degree between d and $2d$ in L_i. Double-counting the edges between L_i and L_{i+1}, we have

$$\frac{C}{4k\log n}|L_i|n^\varepsilon \leq e(L_i, L_{i+1}) \leq 2d|L_{i+1}|.$$

This gives

$$d \geq \frac{C}{8k\log n}\frac{|L_i|n^\varepsilon}{|L_{i+1}|} = \frac{C}{8k\log n}n^{\varepsilon-\delta_i}. \tag{2.1}$$

We will show below that the set L_{i+1} is large enough to provide the expansion required for Property 1. First, however, note that every vertex $y \in L_{i+1}$ can have many edges back to L_i. In order to make this a level in our tree T, for each vertex we need to choose one edge to add to T. The choice of edge induces a path from y back to v_0, and hence these choices determine the sets $X_{i+1,c}$. We will later show that we can choose the edges so as to satisfy Property 2 as well.

2.4.1 Property 1

We begin by providing a heuristic of the argument. Given the level L_i and the sets $X_{i,c}$, we show that L_{i+1} can be partitioned into sets W_c such that for every colour c, the number of edges between $X_{i,c}$ and W_c is $\Omega(d|W_c|)$. Suppose that there exists an index c such that $|X_{i,c}| \leq |W_c|$. On one hand, the fact that we used only $\frac{C}{2K}n^{\varepsilon}$ edges from each vertex in $X_{i,c}$ gives an upper bound on the size of $|W_c|$ in terms of δ_i. On the other hand, the fact that we have a subgraph $G[X_{i,c} \cup W_c]$ which has at most $2|W_c|$ vertices and contains at least $\Omega(d|W_c|)$ edges, will by our inductive hypothesis give a lower bound on the size of $|W_c|$ in terms of δ_i. By combining these bounds, we conclude that δ_i has to be quite large.

We will use Condition (ii') instead of (ii) in Proposition 2.3.1. Thus for all $c \in \mathcal{C}$, we have $|X_{i,c}| \leq 4(8\log n)^i n^{\alpha_i - \varepsilon}$. First we claim a rather weak bound $|L_{i+1}| > k|L_i|$. Suppose this were not the case. Then in the set $L_i \cup L_{i+1}$ of at most $(k+1)|L_i|$ vertices, we have at least $\frac{C}{4k\log n}|L_i|n^{\varepsilon}$ edges. By induction, we must have $\frac{C}{4k\log n}|L_i|n^{\varepsilon} \leq C[(k+1)|L_i|]^{1+\varepsilon}$, or, equivalently,

$$\left(\frac{n}{|L_i|}\right)^{\varepsilon} = n^{(1-\alpha_i)\varepsilon} \leq 4k(k+1)^{1+\varepsilon}\log n,$$

which contradicts our choice of n (recall that $\alpha_i \leq 1 - \frac{1}{4}\varepsilon^2$). Thus we must have $|L_{i+1}| > k|L_i|$.

Consider a fixed vertex $y \in L_{i+1}$, and recall that $d(y, L_i) \geq d$. Consider any neighbor $x \in L_i$ of y. The path from v_0 to x in T uses i different colours $\{c_j : 1 \leq j \leq i\}$. If any other neighbor $x' \in L_i$ of y has a path to v_0 that avoids the colours $\{c_j\}$, then we have a rainbow cycle of length $2(i+1) \leq 2k$. Thus for every neighbor $x' \in L_i$ of y, we must have $x' \in \cup_{j=1}^{i} X_{i,c_j}$. By the pigeonhole principle, there is some j such that $d(y, X_{i,c_j}) \geq \frac{d}{i}$. Informally, this observation asserts that every vertex $y \in L_{i+1}$ sends a large proportion of its edges to some set X_{i,c_j}.

For each colour c, let W_c be the set of vertices $y \in L_{i+1}$ such that $d(y, X_{i,c}) \geq \frac{d}{i}$, and note that $\{W_c\}$ forms a cover of L_{i+1}. Thus $\sum_c |W_c| \geq |L_{i+1}| > k|L_i|$. On the other hand, the sets $\{X_{i,c}\}_c$ form an i-fold cover of L_i, and so $\sum_c |X_{i,c}| = i|L_i| < k|L_i|$. Consequently, $\sum_c (|W_c| - |X_{i,c}|) > 0$, and so for some particular colour c we have $|W_c| > |X_{i,c}|$. As stated above, we will exploit the fact that there are at least $\frac{d}{i}|W_c|$ edges between W_c and $X_{i,c}$ in

two different ways to get two inequalities. Together, these will give the claimed inequality between α_i and α_{i+1}.

First, recall that we used at most $\frac{C}{2k}n^{\varepsilon}$ edges incident to each vertex in L_i to construct the set L_{i+1}. By double-counting the edges between W_c and $X_{i,c}$, we have

$$\frac{d}{k}|W_c| < \frac{d}{i}|W_c| \leq e(W_c, X_{i,c}) \leq \frac{C}{2k}|X_{i,c}|n^{\varepsilon},$$

which by (2.1), gives $|W_c| < \frac{C}{2d}|X_{i,c}|n^{\varepsilon} \leq 4k\log n|X_{i,c}|n^{\delta_i}$. Using Condition (ii') of Proposition 2.3.1, which says that $|X_{i,c}| \leq 4(8\log n)^i n^{\alpha_i-\varepsilon}$, we have

$$|W_c| < 4k\log n|X_{i,c}|n^{\delta_i} \leq 2k(8\log n)^{i+1}n^{\alpha_{i+1}-\varepsilon} \leq 2k(8\log n)^k n^{\alpha_{i+1}-\varepsilon}. \tag{2.2}$$

Second, since there is no rainbow cycle of length at most $2k$ between $X_{i,c}$ and W_c, by the inductive hypothesis we have

$$\frac{d}{k}|W_c| < e(W_c, X_{i,c}) < C\left[|W_c| + |X_{i,c}|\right]^{1+\varepsilon} < C\left[2|W_c|\right]^{1+\varepsilon},$$

which gives $d < 2^{1+\varepsilon}Ck|W_c|^{\varepsilon}$. Hence we have

$$\frac{C}{8k\log n}n^{\varepsilon-\delta_i} \leq d < 2^{1+\varepsilon}Ck|W_c|^{\varepsilon}. \tag{2.3}$$

Combining the inequalities (2.2) and (2.3), we get

$$\begin{aligned} n^{\varepsilon-\delta_i} &< 2^{4+\varepsilon}k^2\log n|W_c|^{\varepsilon} < 2^{4+\varepsilon}k^2\log n\left(2k(8\log n)^k n^{\alpha_{i+1}-\varepsilon}\right)^{\varepsilon} \\ &= 2^{4+(3k+2)\varepsilon}k^{2+\varepsilon}(\log n)^{1+k\varepsilon}n^{(\alpha_{i+1}-\varepsilon)\varepsilon}. \end{aligned}$$

For our choice of n, we have $2^{4+(3k+2)\varepsilon}k^{2+\varepsilon}(\log n)^{1+k\varepsilon} < n^{\frac{1}{2}\varepsilon^2}$, and so $n^{\varepsilon-\delta_i} \leq n^{\frac{1}{2}\varepsilon^2+(\alpha_{i+1}-\varepsilon)\varepsilon}$. This gives $\varepsilon - \delta_i \leq \frac{1}{2}\varepsilon^2 + (\alpha_{i+1}-\varepsilon)\varepsilon = \alpha_{i+1}\varepsilon - \frac{1}{2}\varepsilon^2$, which, using $\delta_i = \alpha_{i+1} - \alpha_i$, becomes

$$\varepsilon - \alpha_{i+1} + \alpha_i \leq \alpha_{i+1}\varepsilon - \frac{1}{2}\varepsilon^2.$$

Rearranging and adding $\left(1+\frac{\varepsilon}{2}\right)$ to both sides, we get

$$(1+\varepsilon)\left[\left(1+\frac{\varepsilon}{2}\right) - \alpha_{i+1}\right] \leq \left(1+\frac{\varepsilon}{2}\right) - \alpha_i,$$

which establishes Property 1.

2.4.2 Property 2

To obtain Property 2, we assume Condition (ii) of Proposition 2.3.1 instead of (ii'). We have shown that the next level L_{i+1} is large enough. For each of its vertices, we now need to select an edge back to L_i in such a way that the sets $X_{i+1,c}$ formed satisfy the bound in Property 2. For each $y \in L_{i+1}$, let $d_y = d(y, L_i)$. Recall that there is a parameter d such that $d \geq \frac{C}{8k \log n} n^{\varepsilon - \delta_i}$ and $d \leq d_y < 2d$ for all $y \in L_{i+1}$. Also recall that each edge back to L_i extends to a rainbow path to the root v_0 in the tree T. For each vertex, we choose one edge uniformly at random, and show that with positive probability the resulting sets $X_{i+1,c}$ are small enough.

We can represent $|X_{i+1,c}|$ as a sum of indicator variables:

$$|X_{i+1,c}| = \sum_{y \in L_{i+1}} \mathbf{1}_{\{y \in X_{i+1,c}\}}.$$

Since each vertex y chooses its path independently of the others, the indicator random variables in the summand are independent. We would first like to obtain an estimate on $\mu_c = \mathbb{E}\left[|X_{i+1,c}|\right]$.

First consider those $c \in \mathcal{C}_{i+1}$. $|X_{i+1,c}|$ counts the number of times the colour c is used between the levels L_i and L_{i+1}. Since the colouring is proper, there are at most $|L_i|$ such edges. Since all the vertices of L_{i+1} have degree at least d, each such edge is chosen with probability at most $\frac{1}{d}$. Thus $\mu_c \leq \frac{|L_i|}{d}$, and by our bound (2.1) on d,

$$\mu_c \leq \frac{|L_i|}{d} \leq \frac{8k(\log n) n^{\alpha_i}}{C n^{\varepsilon - \delta_i}} < (\log n) n^{\alpha_{i+1} - \varepsilon}.$$

Now we consider those $c \notin \mathcal{C}_{i+1}$. Note that for $y \in L_{i+1}$, we have $y \in X_{i+1,c}$ only if we choose for y an edge back to $X_{i,c}$. Thus,

$$\mu_c = \sum_y \frac{d(y, X_{i,c})}{d_y} \leq \frac{1}{d} \sum_y d(y, X_{i,c}) = \frac{1}{d} e(L_{i+1}, X_{i,c}).$$

Since all the vertices in L_i send at most $\frac{C}{2k} n^{\varepsilon}$ edges into L_{i+1}, the above is at most

$$\mu_c \leq \frac{C}{2kd} |X_{i,c}| n^{\varepsilon} \leq \frac{C(8 \log n)^i}{2kd} n^{\alpha_i}.$$

Using (2.1), this gives $\mu_c \leq \frac{1}{2}(8\log n)^{i+1}n^{\alpha_{i+1}-\varepsilon}$. Thus for $t = (8\log n)^{i+1}n^{\alpha_{i+1}-\varepsilon}$, we have $t \geq 2\mu_c$ for all colours c.

By Theorem 2.2.1, for every colour c, we have

$$\mathbf{P}\left(|X_{i+1,c}| \geq t\right) \leq \exp\left(-\frac{t}{8}\right).$$

Recalling that $\alpha_{i+1} \geq \alpha_1 \geq \varepsilon$, and $i+1 \geq 2$, we have $t = (8\log n)^{i+1}n^{\alpha_{i+1}-\varepsilon} \geq 64\log n \geq 32\ln n$. Hence $\mathbf{P}\left(|X_{i+1,c}| \geq t\right) \leq \exp\left(-4\ln n\right) = n^{-4}$. There are at most n^2 colours c, and so a union bound gives

$$\mathbf{P}\left(\exists c : |X_{i+1,c}| \geq t\right) \leq n^2 \cdot n^{-4} = n^{-2} < 1.$$

Thus there is a choice of edges such that Property 2 holds.

This completes the proof of Proposition 2.3.1. □

2.4.3 Proof of Corollary 2.1.3

We can establish Corollary 2.1.3 by slightly twisting the argument given above. The structure of the proof of Corollary 2.1.3 is the same as that of Theorem 2.1.1. The only difference lies in how we establish Property 1. In this section, we show how to use the weaker condition of the graph having no rainbow cycle of length exactly $2k$ in order to derive Property 1. We first restate Property 1 here for the reader's convenience:

$$\left(1+\frac{\varepsilon}{2}\right) - \alpha_{i+1} \leq (1+\varepsilon)^{-1}\left[\left(1+\frac{\varepsilon}{2}\right) - \alpha_i\right].$$

If $d \leq 16k$, then by (2.1), we have

$$16k \geq d \geq \frac{C}{8k\log n}n^{\varepsilon-\delta_i},$$

from which it follows that

$$n^{\delta_i} \geq \frac{C}{128k^2\log n}n^{\varepsilon} \geq n^{\varepsilon-\varepsilon^2/2},$$

and $\delta_i \geq \varepsilon - \frac{1}{2}\varepsilon^2$. Since $\delta_i = \alpha_{i+1} - \alpha_i$ and $\alpha_i \geq \alpha_1 \geq \varepsilon$, we see that

$$\left(1+\frac{\varepsilon}{2}\right) - \alpha_{i+1} \leq \left(1-\frac{\varepsilon}{2}+\frac{\varepsilon^2}{2}\right) - \alpha_i \leq (1+\varepsilon)^{-1}\left[\left(1+\frac{\varepsilon}{2}\right) - \alpha_i\right].$$

Thus Property 1 follows.

Hence it suffices to consider the case when $d > 16k$. In this case, we use the following lemma.

Lemma 2.4.1. *If $d > 16k$, then there exists a subset $L'_{i+1} \subset L_{i+1}$ of size $|L'_{i+1}| \geq \frac{1}{2}|L_{i+1}|$ such that for every $y \in L'_{i+1}$, there exists a colour c for which y has at least $\frac{d}{2k}$ neighbors in $X_{i,c}$.*

Proof. Let $L'_{i+1} \subset L_{i+1}$ be the subset of vertices y such that there exists a colour c for which y has at least $\frac{d}{2k}$ neighbors in $X_{i,c}$. We assume for contradiction that $|L'_{i+1}| < \frac{1}{2}|L_{i+1}|$. Let $L''_{i+1} = L_{i+1} \setminus L'_{i+1}$.

The number of edges in the bipartite graph induced by the sets L_i and L''_{i+1} is at least

$$d|L''_{i+1}| > 16k|L''_{i+1}| \geq 8k|L_{i+1}| \geq 4k(|L_i| + |L''_{i+1}|).$$

Hence there exists a subgraph H of minimum degree at least $4k$. Fix a vertex $v \in L_i$ which lies in this subgraph. Let P_0 be the rainbow path from v_0 to v. We can find a rainbow path P_1 of length $2k - 2i - 1$ in H starting at v by greedily extending the current path one vertex at a time, choosing any edge that avoids a previously used vertex or colour. Indeed, we need to avoid at most $2k - 2i - 1$ vertices and at most $2k - 2i - 1$ colours. Since H is properly coloured and has minimum degree at least $4k$, we always have $4k > 2(2k - 2i - 1)$ edges of different colours incident to a given vertex and therefore can extend the path. Also, having split the colours into disjoint classes, we have that the colours of P_0 are disjoint from the colours of P_1. Thus concatenating P_0 and P_1 gives a rainbow path of length $2k - i - 1$ starting at v_0. Let $w \in L''_{i+1}$ be the other endpoint of this path.

In order to avoid a rainbow cycle of length $2k$, all the neighbors of w in L_i must either lie on P_1, use a colour from P_1, or lie in the set $X_{i,c}$ for some colour c in the path P_0. There are at most $4k$ neighbors accounted for by the first two cases. Hence there exists a colour c for which the number of neighbors of w in $X_{i,c}$ is at least

$$\frac{d - 4k}{k} \geq \frac{d}{2k}.$$

However, this contradicts the definition of L''_{i+1}. Thus we have $|L'_{i+1}| \geq \frac{1}{2}|L_{i+1}|$. □

Note that in Section 2.4.1 we used the fact that for all $y \in L_{i+1}$, there exists a colour c for y has at least $\frac{d}{i}$ neighbors in $X_{i,c}$. The conclusion of Lemma 2.4.1 is slightly weaker than this statement since it asserts only that there exists a subset L'_{i+1} of size $\frac{1}{2}|L_{i+1}|$ for which for all $y \in L'_{i+1}$, there exists a colour c for which y has at least $\frac{d}{2k}$ neighbors in $X_{i,c}$.

Thus to proceed further, by slightly abusing notation, we redefine L_{i+1} as the set L'_{i+1}. This adjustment, and the fact that each vertex has only $\frac{d}{2k}$ neighbors in $X_{i,c}$, instead of $\frac{d}{i}$ neighbors, are different from Section 2.4.1, but these will only affect the constants involved in the proof. We omit the straightforward adjustments that are necessary.

2.5 Further remarks

We note that at the beginning of our argument, we used Lemma 2.2.2 to separate the colours into disjoint classes to be used between levels of the tree T. This simplifies the proof, at the cost of a worse constant $C(\varepsilon)$. It is possible to remove this step from the proof, and use most of the edges out of a vertex at each stage. While we would not gain much in our argument above, this might be important if dealing with cycles of length growing with n.

Recall that $f(n)$ denotes the maximum number of edges in a rainbow acyclic graph on n vertices. In this chapter, we showed that for any fixed $\varepsilon > 0$ and large enough n, $f(n) < n^{1+\varepsilon}$. In fact, one can use our method to obtain an upper bound of the form $f(n) < n \exp\left((\log n)^{\frac{1}{2}+\eta}\right)$ for any $\eta > 0$. On the other hand, the hypercube construction of Keevash, Mubayi, Sudakov and Verstraëte gives a lower bound of $f(n) = \Omega(n \log n)$. It would be very interesting to determine the true asymptotics of $f(n)$. The problem of determining the rainbow Turán number for even cycles also remains. It would be interesting to further narrow the gap $\Omega\left(n^{1+\frac{1}{k}}\right) \leq \mathrm{ex}^*(n, C_{2k}) \leq O\left(n^{1+\frac{(1+\varepsilon_k)\ln k}{k}}\right)$, and establish the order of magnitude of the function. We believe the lower bound to be correct.

CHAPTER 3

A problem of Erdős on the minimum number of k-cliques

3.1 Introduction

Let K_l denote a complete graph on l vertices and let $\overline{K_l}$ be its complement, i.e., an independent set of size l. One of the central results in extremal combinatorics is Turán's theorem [Tur41], which asserts that the maximum number of edges in a K_l-free graph on n vertices is attained by the Turán graph $T_{n,l-1}$, a complete $(l-1)$-partite graph with nearly-equal parts. This theorem has since been extended and generalised in many different ways. Since an edge can be thought of as a clique on 2 vertices, a natural generalisation is to ask for the maximum number of K_k in an n-vertex graph with no K_l. Zykov [Zyk49] showed that this maximum was also attained by the Turán graph $T_{n,l-1}$.

For any integers $k, l \geq 2$ and n, we define $f(n,k,l)$ to be the minimum number of copies of K_k in a $\overline{K_l}$-free graph on n vertices. If one takes the complements of the graphs in Turán's theorem, then the theorem gives the minimum number of edges in an n-vertex $\overline{K_l}$-free graph. Thus the question of determining $f(n,k,l)$ is precisely the Zykov-type generalisation of this complementary version. Fifty years ago Erdős [Erd62b] asked to determine $f(n,k,l)$ and conjectured that the minimum is given by the complement of the Turán graph, $\overline{T_{n,l-1}}$, which is the disjoint union of $l-1$ complete graphs of equal size. When $k=2$, this follows from Turán's theorem.

Note that a graph is $\overline{K_l}$-free precisely when its independence number is less than l. One can thus also view this problem as a strengthening of Ramsey's theorem, which states that

any sufficiently large graph either has a clique of size k or an independent set of size l. The (k, l)-problem asks how many cliques of size k a graph must have when its independence number is less than l.

Lorden [Lor62] proved Erdős' conjecture to be true for the $(3, 3)$-case by a simple double-counting argument. However, no further progress was made in the next forty years, until Nikiforov [Nik01] disproved the conjecture in the case $(4, 3)$ by showing the balanced blow-up of C_5, which is $\overline{K_3}$-free, contains fewer 4-cliques than the disjoint union of two cliques, $\overline{T_{n,2}}$. In a *blow-up* of a graph, we replace every vertex with a clique, and every edge with a complete bipartite graph. We say the blow-up is *balanced* if the cliques are all of the same size. In a subsequent preprint [Nik05], Nikiforov showed that his construction is optimal under the additional assumption that the graph should be nearly-regular.

Moreover, by considering blow-ups of Ramsey graphs, Nikiforov showed that the conjecture could only hold for finitely many (k, l) when $k, l \geq 3$. In particular, he conjectured that equality held only for the cases $(3, 3)$ and $(3, 4)$, the latter of which remained an open problem.

3.1.1 Our results

In this chapter, we first sharpen Nikiforov's result by showing that Erdős' conjecture is always false when $k \geq 4$ and $l \geq 3$, or when $k = 3$ and $l \geq 2074$. We obtain these results through a combination of explicit and random counterexamples.

We then solve the problem in the cases $(k, l) = (4, 3)$ and $(3, 4)$. Using the machinery of flag algebras developed by Razborov [Raz07], we are able to obtain the asymptotic values of $f(n, 4, 3)$ and $f(n, 3, 4)$. By analyzing the corresponding semi-definite programming solutions, we are then able to derive stability results for these cases, which in turn allow us to determine $f(n, 4, 3)$ and $f(n, 3, 4)$ exactly for large n, and also to characterise the extremal graphs. In particular, we show that a blow-up of C_5 is indeed optimal for the $(4, 3)$ problem, while Erdős' conjecture holds for the $(3, 4)$ problem. Our results are summarised in the following theorems.

Theorem 3.1.1. $f(n,4,3) = \frac{3}{25}\binom{n}{4} + O(n^3)$, where the minimum is achieved by a blow-up of C_5 with five parts of roughly equal sizes. Moreover, the extremal structure is unique for sufficiently large n.

We determine the exact sizes of the parts of the blow-up by solving an integer optimisation problem, the precise results of which are given in Section 3.4.

Theorem 3.1.2. $f(n,3,4) = \binom{\lfloor n/3 \rfloor}{3} + \binom{\lfloor (n+1)/3 \rfloor}{3} + \binom{\lfloor (n+2)/3 \rfloor}{3} \sim \frac{1}{9}\binom{n}{3}$, where for large n the minimum is achieved by three disjoint cliques that are as equal as possible. Moreover, any extremal graph must be spanned by three such cliques.

Note that in this case the extremal graph is not unique, as we may have partial matchings between the cliques without introducing any extra triangles.

As we remark in our concluding section, solutions of corresponding SDP problems strongly suggest that a disjoint union of cliques remains optimal for the $(3,5)$- and $(3,6)$-problems, contrary to Nikiforov's conjecture.

3.1.2 Notation and organisation

Given a graph G on vertices $V(G)$, and a vertex $v \in V(G)$, we denote by $N(v)$ the set of neighbors of v in G, and by $\overline{N(v)}$ the set of non-neighbors of v. The complement graph $\overline{G}$ shares the same vertices as G, and has an edge $\{u,v\}$ if and only if $\{u,v\}$ is not an edge of G. We denote the independence number of G by $\alpha(G)$. The complete graph on k vertices is denoted by K_k. In particular, a graph G is $\overline{K_l}$-free if and only if $\alpha(G) < l$. Some other graphs we will use are the cycles C_k, and paths P_k where in each case the subscript refers to the number of edges.

Given a fixed graph H, for any graph G we let $t_H(G)$ denoted the number of induced copies of H in G. In the case $H = K_k$, we simplify the notation to $t_k(G)$. Using this notation, we can define

$$f(n,k,l) = \min\{t_k(G) : |V(G)| = n, t_l(\overline{G}) = 0\}.$$

The rest of the chapter is organised as follows. In the next section, we construct coun-

terexamples to Erdős' conjecture in the case $k \geq 4$ and $l \geq 3$ or $k = 3$ and l large. In Section 3.3, we provide an informal introduction to our main tool, flag algebras. Sections 3.4 and 3.5 contain the proofs of our main results for the $(4, 3)$- and $(3, 4)$-problems respectively. The final section contains some concluding remarks and open problems.

Some technical details are given in the final sections: Section 3.7 provides some remarks regarding implementation of flag algebras, and Section 3.8 contains the proof of the integer optimisation result for the $(4, 3)$-problem.

3.2 Counterexamples to Erdős' conjecture

Nikiforov [Nik01] showed that not only was Erdős' conjecture not true in general, but that it held only finitely often. He used bounds on the Ramsey numbers $R(3, l)$ to show the existence of k_0 and l_0 such that whenever $k > k_0$ or $l > l_0$, blow-ups of Ramsey graphs did better than disjoint unions of cliques $\overline{T_{n,l-1}}$. In the following theorem, we use a combination of explicit and random constructions to further improve this result.

Theorem 3.2.1. $\overline{T_{n,l-1}}$ is not optimal for the (k, l)-problem when

(i) $k \geq 4$ and $l \geq 3$, or

(ii) $k = 3$ and $l \geq 2074$.

3.2.1 The (k, l)-problem with $k \geq 4$

Let us first consider the case $l = 3$. That is, we are looking to minimise the number of k-cliques in a graph with independence number at most 2. For the $(4, 3)$-problem, Nikiforov [Nik05] gave an explicit counter-example to Erdős's conjecture by showing that a blow-up of C_5 contains fewer triangles than the graph $\overline{T_{n,2}}$, which consists of two disjoint cliques. In fact, it is easy to see that this construction is better than $\overline{T_{n,2}}$ for any $k \geq 4$. Indeed, a disjoint union of two cliques contains, asymptotically, $2\binom{\frac{n}{2}}{k} \sim \frac{1}{2^{k-1}}\binom{n}{k}$ k-cliques. On the other hand, the blow-up of C_5 contains $5\left(\binom{\frac{2n}{5}}{k} - \binom{\frac{n}{5}}{k}\right) \sim \frac{2^k-1}{5^{k-1}}\binom{n}{k}$ k-cliques. For $k \geq 4$, we have $\frac{2^k-1}{5^{k-1}} < \frac{1}{2^{k-1}}$, and so $\overline{T_{n,2}}$ is asymptotically not optimal for the $(k, 3)$-problem.

For $l \geq 4$, the graph $\overline{T_{n,l-1}}$ consists of $l-1$ disjoint cliques. However, as shown above, if we replace two of these cliques with a blow-up of C_5 on the same number of vertices, we will reduce the number of k-cliques. Formally, this construction has a blow-up of C_5 on five parts of size $\frac{2n}{5(l-1)}$, and $l-3$ disjoint cliques of size $\frac{n}{l-1}$, and contains fewer k-cliques than $\overline{T_{n,l-1}}$. This shows that a disjoint union of cliques is not optimal for the (k,l)-problem for any $k \geq 4$ and $l \geq 3$.

3.2.2 The $(3,l)$-problem

The situation is quite different when $k = 3$. As we will show later, the disjoint union of cliques is optimal for the $(3,3)$- and $(3,4)$-problems. However, unlike the case $k = 2$, this construction ceases to be optimal for large values of l. We consider the random graph $G \sim G(m,p)$ on m vertices, with every edge appearing independently with probability p. For suitable parameters l, m, and p, we show that with positive probability the balanced blow-up of G has no independent set of size l and has fewer triangles than $\overline{T_{n,l-1}}$. First we count the number of triangles in a balanced blow-up of an m-vertex graph G to n vertices.

There are three ways to obtain a triangle in the blow-up. The vertices of the triangle can all come from one part, in which case there are $\frac{n}{m}$ vertices to choose from. As there are m vertices in G, there are $m\binom{\frac{n}{m}}{3} \sim \frac{1}{m^2}\binom{n}{3}$ such triangles. Alternatively, the vertices of the triangle can come from an edge in G, with two vertices from one part, and the third vertex from the other. There are two ways to split the vertices, and $e(G)$ edges, so the total number of such triangles is $2e(G)\binom{\frac{n}{m}}{2}\binom{\frac{n}{m}}{1} \sim \frac{6e(G)}{m^3}\binom{n}{3}$. Finally, the vertices of the triangle can come from a triangle in G, with one vertex from each of the three parts. There are $t_3(G)$ triangles in G, and so the number of such triangles is $t_3(G)\left(\frac{n}{m}\right)^3 \sim \frac{6t_3(G)}{m^3}\binom{n}{3}$. Thus the total number of triangles in the blow-up of G is asymptotically $\left(\frac{6(e(G)+t_3(G))}{m^3} + \frac{1}{m^2}\right)\binom{n}{3}$.

On the other hand, $\overline{T_{n,l-1}}$ has $(l-1)\binom{\frac{n}{l-1}}{3} \sim \frac{1}{(l-1)^2}\binom{n}{3}$ triangles. Thus to obtain a counter-example to Erdős's conjecture, we need to show that for some l, m and p, with positive probability the random graph $G \sim G(m,p)$ has no independent set of size l and $\frac{6(e(G)+t_3(G))}{m^3} + \frac{1}{m^2} < \frac{1}{(l-1)^2}$, or $e(G) + t_3(G) < \frac{m^3}{6(l-1)^2} - \frac{m}{6}$. Let us call such a graph 'suitable'.

Let B_1 be the event that $\alpha(G) \geq l$, where $\alpha(G)$ is the independence number of G. For some parameters s and t, let B_2 be the event $\{e(G) - \mathbb{E}[e(G)] \geq s\}$, and B_3 the event $\{t_3(G) - \mathbb{E}[t_3(G)] \geq t\}$. If $\mathbb{E}[e(G) + t_3(G)] + s + t \leq \frac{m^3}{6(l-1)^2} - \frac{m}{6}$, then $\left\{e(G) + t_3(G) \geq \frac{m^3}{6(l-1)^2} - \frac{m}{6}\right\} \subset B_2 \cup B_3$. Then we have

$$\mathbb{P}(G \text{ not suitable}) \leq \mathbb{P}(B_1 \cup B_2 \cup B_3) \leq \mathbb{P}(B_1) + \mathbb{P}(B_2 \cup B_3).$$

We use a union bound for B_1: there are $\binom{m}{l}$ sets of l vertices, and the probability that a given set has no edges is $(1-p)^{\binom{l}{2}}$. Using the bound $\binom{n}{r} \leq \left(\frac{ne}{r}\right)^r$, we have

$$\mathbb{P}(B_1) \leq \binom{m}{l}(1-p)^{\binom{l}{2}} \leq \left(\frac{me(1-p)^{\frac{l-1}{2}}}{l}\right)^l.$$

Note that the other two events are increasing; that is, they are preserved by the addition of edges. It then follows from Kleitman's Lemma (see Chapter 6 in [AS08]) that $\mathbb{P}(B_2 \cap B_3) \geq \mathbb{P}(B_2)\mathbb{P}(B_3)$, and so

$$\begin{aligned}\mathbb{P}(B_2 \cup B_3) &= \mathbb{P}(B_2) + \mathbb{P}(B_3) - \mathbb{P}(B_2 \cap B_3) \leq \mathbb{P}(B_2) + \mathbb{P}(B_3) - \mathbb{P}(B_2)\mathbb{P}(B_3) \\ &= \mathbb{P}(B_2) + \mathbb{P}(B_3)\left(1 - \mathbb{P}(B_2)\right).\end{aligned}$$

Moreover, since the right-hand side is increasing in both $\mathbb{P}(B_2)$ and $\mathbb{P}(B_3)$, we can replace the probabilities with upper bounds to obtain an upper bound on $\mathbb{P}(B_2 \cup B_3)$. To obtain these upper bounds, we use the following second moment concentration inequality from [AS08]:

Proposition 3.2.2. Let X be a random variable with expectation $\mathbb{E}[X] = \mu$ and variance σ^2. Then for all $\lambda > 0$,

$$\mathbb{P}(X - \mu \geq \lambda) \leq \frac{\sigma^2}{\lambda^2 + \sigma^2}.$$

For the event B_2, with $X = e(G)$, we have $X \sim \text{Bin}\left(\binom{m}{2}, p\right)$, and so $\mu = \binom{m}{2}p$ and $\sigma^2 = \binom{m}{2}p(1-p)$. This gives $\mathbb{P}(B_2) \leq \frac{\binom{m}{2}p(1-p)}{s^2 + \binom{m}{2}p(1-p)}$.

For the event B_3, let $X = t_3(G)$. There are $\binom{m}{3}$ possible triangles, each of which appears with probability p^3, and hence $\mu = \binom{m}{3}p^3$. To find the variance, we note that any fixed triangle T is independent of all triangles except those that share at least two vertices with

T. A quick calculation gives $\sigma^2 = \binom{m}{3}p^3\left[(1-p^3)+3(m-3)p^2(1-p)\right]$. Thus $\mathbb{P}(B_3) \leq \frac{\binom{m}{3}p^3\left[(1-p^3)+3(m-3)p^2(1-p)\right]}{t^2+\binom{m}{3}p^3[(1-p^3)+3(m-3)p^2(1-p)]}$.

Thus if we can find l, m, p, s and t such that $\binom{m}{2}p + \binom{m}{3}p^3 + s + t \leq \frac{m^3}{6(l-1)^2} - \frac{m}{6}$, and

$$\left(\frac{me(1-p)^{\frac{l-1}{2}}}{l}\right)^l + \frac{\binom{m}{2}p(1-p)}{s^2+\binom{m}{2}p(1-p)} + \frac{\binom{m}{3}p^3\left[(1-p^3)+3(m-3)p^2(1-p)\right]}{t^2+\binom{m}{3}p^3[(1-p^3)+3(m-3)p^2(1-p)]}\left[1 - \frac{\binom{m}{2}p(1-p)}{s^2+\binom{m}{2}p(1-p)}\right] < 1,$$

then we prove that there is a suitable graph, and therefore $\overline{T_{n,l-1}}$ is not optimal for the $(3,l)$-problem.

A computer search determined that $l = 2074$, $m = 164397$, $p = 0.0051707$, $s = 14000$ and $t = 35000$ are suitable values. Hence the graph with 2073 disjoint cliques is not optimal for the $(3, 2074)$-problem. Moreover, if $l > 2074$, then in $\overline{T_{n,l-1}}$ we can replace 2073 cliques by a graph with fewer triangles. Hence $\overline{T_{n,l-1}}$ is not optimal for the $(3,l)$-problem for any $l \geq 2074$.

It would be interesting to find better constructions and to determine when $\overline{T_{n,l-1}}$ stops being optimal for the $(3,l)$-problem. Our flag algebra calculations suggest that it is still optimal for at least the $(3,5)$- and $(3,6)$-problems.

3.3 Flag algebra calculus

In this section we provide a brief introduction to the technique of flag algebras. First introduced by Razborov in [Raz07], it has been applied with great success to a wide variety of problems in extremal combinatorics (see, for example, [BT11, HHK12, HHK13, FV13, Raz10, Raz08]).

We will begin with a general overview of the calculus, by introducing some key definitions and providing some intuition behind the machinery. The second subsection will show how we express extremal problems in the language of flag algebras. In Section 3.7 we discuss some practical considerations regarding implementation of the method, to explain how we obtained our results in the later sections.

It is neither our goal to be rigorous nor thorough, but rather to emphasise that the com-

binatorial arguments behind the flag algebra calculus are as old as extremal combinatorics itself. Indeed, the main tools available to us are double-counting and the Cauchy-Schwarz inequality. To highlight this fact, we will use the $(3, 3)$-problem as a running example, and indeed, the proof we obtain through flag algebras will be essentially the same as the original proof Lorden gave in 1962.

The flag algebra calculus is powerful because it provides a formalism through which the problem of finding relations between subgraph densities can be reduced to a semi-definite programming (SDP) problem. This in turn enables the use of computers to find solutions, with rigorous proofs, to problems in extremal combinatorics. For a more complete survey of the technique, we refer you to the excellent expositions in [Kee11] and [FV13], while for a technical specification of flag algebras, we refer you to the original paper of Razborov [Raz07].

3.3.1 Basic definitions and notation

The flag algebra calculus is typically used to find the extremal density of some fixed subgraph J amongst graphs that avoid some forbidden subgraph. For our example, the $(3, 3)$-problem, we wish to minimise the density of triangles K_3 in graphs that do not contain $\overline{K_3}$, the empty graph on 3 vertices. While our definitions will be general, all our examples will come from this setting.

We say that a graph is *admissible* if it contains no induced copies of the forbidden graph. A *type* σ is an admissible labeled graph on vertices $[k]$ for some non-negative integer k called the *size* of σ, denoted by $|\sigma|$. In what follows, an isomorphism between graphs must preserve any labels that are present.

Given a type σ, a *σ-flag* is an admissible graph F on a partially labeled vertex set, such that the subgraph induced by the labeled vertices is isomorphic to σ. The *underlying graph* of the flag F is the graph F with all labels removed. The *size* of a flag is the number of vertices. Note that when σ is the *trivial type* of size 0 (denoted by $\sigma = 0$), a σ-flag is just an usual unlabeled admissible graph. We shall write $\mathcal{F}_l^\sigma$ for the collection of all σ-flags of size

l. Let $\mathcal{F}^\sigma = \bigcup_{l \geq 0} \mathcal{F}^\sigma_l$. When the type σ is trivial, we shall omit the superscript from our notation.

Let us now define two fundamental concepts in our calculus, namely those of flag densities in larger flags and graphs. Let σ be a type of size k, let $m \geq 1$ be an integer and let $\{F_i\}_{i=1}^m$ be a collection of σ-flags of sizes $l_i = |F_i| \geq k$. Given a σ-flag F of order at least $l = k + \sum_{i=1}^m (l_i - k)$, let $T \subseteq V(F)$ be the set of labeled vertices of F. Now select disjoint subsets $X_i \subseteq V(F) \setminus T$ of sizes $|X_i| = l_i - k$, uniformly at random. This is possible because F has at least $\sum_i (l_i - k)$ unlabeled vertices. Denote by E_i the event that the σ-flag induced by $T \cup X_i$ is isomorphic to F_i, for $i \in [m]$. We define $p_\sigma(F_1, F_2, \ldots, F_m; F) \stackrel{def}{=} \mathbb{P}(\cap_{i=1}^m E_i)$ to be the probability that all these events occur simultaneously.

If G is just an admissible graph of order at least l, and not a σ-flag, then there is no pre-labeled set of vertices T that induces the type σ. Instead, we uniformly at random select a partial labeling $L : [k] \to V(G)$. This random labeling turns G into a σ'-flag F_L, where the type σ' is the labeled subgraph induced by the set of vertices $L([k])$. If $\sigma' = \sigma$, we can then proceed as above, otherwise we say the events E_i have probability 0. Finally, we average over all possible random labelings. Formally, let Y be the following random variable

$$Y \stackrel{def}{=} \begin{cases} p_\sigma(F_1, F_2, \ldots, F_m; F_L) & \text{if } \sigma' = \sigma \\ 0 & \text{otherwise} \end{cases}.$$

Define $d_\sigma(F_1, \ldots, F_m; G) \stackrel{def}{=} \mathbb{E}(Y)$ as the expected value of the random variable Y. The quantities $p_\sigma(F_1, F_2, \ldots, F_m; F)$ and $d_\sigma(F_1, F_2, \ldots, F_m; G)$ are called *flag densities* of $\{F_i\}_{i \in [m]}$ in F and in G, respectively. Clearly these flag densities are the same whenever $\sigma = 0$, in which case we omit the subscript from both notations.

To better illustrate these definitions, we give some examples. Let *dot* be the only type of size one. Let ρ and $\overline{\rho}$ be the two *dot*-flags of size two, and let Z_i, for $1 \leq i \leq 5$, be the five admissible *dot*-flags of size three (recall that we are forbidding $\overline{K_3}$). These flags are shown in Figure 3.1.

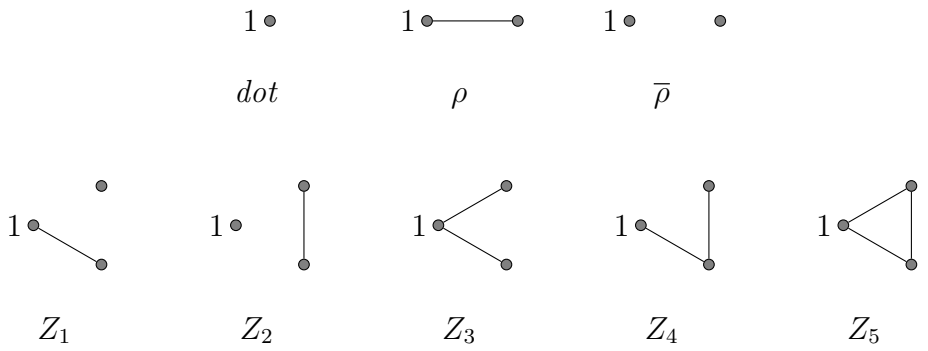


Figure 3.1: Some examples of flags of type *dot*.

We now compute the flag densities of ρ and $\overline{\rho}$ in the flags Z_i. For example, to compute $p_{dot}(\overline{\rho}; Z_1)$, note that to induce a copy of $\overline{\rho}$ we must choose an unlabeled non-neighbor of 1. As only one of the two unlabeled vertices in Z_1 is a non-neighbor of 1, we conclude that $p_{dot}(\overline{\rho}; Z_1) = \frac{1}{2}$. Similarly, $p_{dot}(\rho; Z_3) = 1$, because to induce ρ we must select a neighbor of 1, and all the unlabeled vertices in Z_3 are neighbors of 1. The other flag densities are $p_{dot}(\rho; Z_5) = p_{dot}(\overline{\rho}; Z_2) = 1$, $p_{dot}(\rho; Z_1) = p_{dot}(\rho; Z_4) = p_{dot}(\overline{\rho}; Z_1) = p_{dot}(\overline{\rho}; Z_4) = \frac{1}{2}$, and $p_{dot}(\rho; Z_2) = p_{dot}(\overline{\rho}; Z_3) = p_{dot}(\overline{\rho}; Z_5) = 0$.

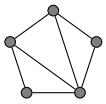


Figure 3.2: Graph W.

To see how to compute flag densities in an unlabeled graph, consider W, the graph on 5 vertices depicted in Figure 3.2. It is easy to see that $d_{dot}(\rho; W)$ and $d_{dot}(\overline{\rho}; W)$ are the edge and non-edge densities of W respectively, and so $d_{dot}(\rho; W) = \frac{7}{10}$ and $d_{dot}(\overline{\rho}; W) = \frac{3}{10}$. The computation of $d_{dot}(Z_i; W)$ is a little more involved. As an example, we explain how to compute $d_{dot}(Z_3; W)$. Note that Z_3 consists of two nonadjacent neighbors of the labeled vertex 1. Hence for every vertex $v \in V(W)$, let κ_v denote the number of nonadjacent pairs neighbors of v divided by the total number of pairs of vertices in $V(W) \setminus \{v\}$. $d_{dot}(Z_3; W)$ is then the average of κ_v over all vertices in W, which comes out to $\frac{1}{6}$. Computing the other flag densities gives $d_{dot}(Z_1; W) = \frac{2}{15}$, $d_{dot}(Z_2; W) = \frac{1}{15}$, $d_{dot}(Z_4; W) = \frac{1}{3}$, and $d_{dot}(Z_5, W) = \frac{3}{10}$.

We can also compute the joint flag densities of multiple flags. For instance, let us consider $d_{dot}(\rho, \rho; W)$. In this case, we first randomly choose a vertex v to be the labeled vertex. We must then make an *ordered* choice of two vertices in $V(W) \setminus \{v\}$, as we have two flags, each with one unlabeled vertex. If both of these vertices are neighbors of v, then we have induced two copies of the flag ρ (note that the adjacency of these two vertices is unimportant). Hence we obtain $d_{dot}(\rho, \rho; W)$ by averaging over all vertices v the ratio of the number of ordered pairs of neighbors of v to the number of ordered pairs of vertices in $V(W) \setminus \{v\}$. In this case, we have $d_{dot}(\rho, \rho; W) = \frac{7}{15}$.

Suppose as before we have a type σ of size k, a σ-flag F of size $l \geq k$, and an unlabeled graph G. To compute $d_\sigma(F; G)$, we averaged over all random partial labelings of G the probability of finding a flag isomorphic to F. A simple double-counting argument shows that we can do the averaging before the random labeling, which is the idea behind Razborov's *averaging operator*, as defined in Section 2.2 of [Raz07]. Let $F|_0$ denote the unlabeled underlying graph of F. We can compute $d_\sigma(F; G)$ by first computing $d(F|_0; G)$, the probability that l randomly chosen vertices in G form an induced copy of $F|_0$ as a subgraph. Given this copy of $F|_0$, we then randomly label k of the l vertices, and compute the probability that these k vertices are label-isomorphic to σ. This amounts to multiplying $d(F|_0; G)$ by a *normalizing factor* $q_\sigma(F)$, that is, $d_\sigma(F; G) = q_\sigma(F) d(F|_0; G) = q_\sigma(F) p(F|_0; G)$.

We can interpret the normalizing factor as $q_\sigma(F) = d_\sigma(F; F|_0)$. From our previous example, we have $q_{dot}(\rho) = q_{dot}(\overline{\rho}) = q_{dot}(Z_5) = 1$, $q_{dot}(Z_3) = q_{dot}(Z_2) = \frac{1}{3}$ and $q_{dot}(Z_4) = q_{dot}(Z_1) = \frac{2}{3}$. Since $q_{dot}(Z_5) = 1$, it follows that $d_{dot}(Z_5; G) = d(K_3; G)$ is the triangle density of G.

There are more relations involving d_σ and p_σ than the one mentioned previously. We will now state, without proof, a basic fact about flag densities that can be proved easily by double counting.

Fact 3.3.1 (Chain rule)**.** If σ is a type of size k, $m \geq 1$ is an integer, and $\{F_i\}_{i=1}^m$ is a family of σ-flags of sizes $|F_i| = l_i$, and $l \geq k + \sum_{i=1}^m (l_i - k)$ is an integer parameter, then

1. For any σ-flag F of order at least l, we have

$$p_\sigma(F_1,\dots,F_m;F)=\sum_{F'\in\mathcal{F}_l^\sigma} p_\sigma(F_1,\dots,F_m;F')p_\sigma(F';F).$$

2. For any admissible graph G of order at least l, we have

$$d_\sigma(F_1,\dots,F_m;G)=\sum_{H\in\mathcal{F}_l} d_\sigma(F_1,\dots,F_m;H)d(H;G)=\sum_{F\in\mathcal{F}_l^\sigma} p_\sigma(F_1,\dots,F_m;F)d_\sigma(F;G).$$

If we apply the chain rule for $m=1$, we have the equation

$$p_\sigma(F;F')=\sum_{F''\in\mathcal{F}_l^\sigma} p_\sigma(F;F'')p_\sigma(F'';F').$$

For instance, this gives

$$\begin{aligned}p_{dot}(\rho;F)&=p_{dot}(\rho;Z_1)p_{dot}(Z_1;F)+p_{dot}(\rho;Z_2)p_{dot}(Z_2;F)+p_{dot}(\rho;Z_3)p_{dot}(Z_3;F)+\\&\quad p_{dot}(\rho;Z_4)p_{dot}(Z_4;F)+p_{dot}(\rho;Z_5)p_{dot}(Z_5;F)\\&=\frac{1}{2}p_{dot}(Z_1;F)+p_{dot}(Z_3;F)+\frac{1}{2}p_{dot}(Z_4;F)+p_{dot}(Z_5;F).\end{aligned}$$

Similarly, we can expand $p_{dot}(\overline{\rho};F)=\frac{1}{2}p_{dot}(Z_1;F)+p_{dot}(Z_2;F)+\frac{1}{2}p_{dot}(Z_4;F)$.

For the ease of notation, we can express these two identities using the syntax of flag algebras:

$$\begin{aligned}\rho&=\frac{1}{2}Z_1+Z_3+\frac{1}{2}Z_4+Z_5,\\\overline{\rho}&=\frac{1}{2}Z_1+Z_2+\frac{1}{2}Z_4.\end{aligned}\tag{3.1}$$

In this syntax, the equation $\sum_{i\in I}\alpha_i F_i=0$ means that for all sufficiently large σ-flags F, we have $\sum_{i\in I}\alpha_i p_\sigma(F_i;F)=0$, where $\alpha_i\in\mathbb{R}$ for all $i\in I$. We call $\sum_{i\in I}\alpha_i F_i$ an *eventually zero* expression. We use $\mathcal{A}^\sigma$ to denote the set of linear combinations of flags of type σ. It is convenient to define a *product* of flags in the following way:

$$F_1\cdot F_2\stackrel{def}{=}\sum_{F\in\mathcal{F}_l^\sigma} p_\sigma(F_1,F_2;F)F,\qquad F_1\in\mathcal{F}^\sigma,F_2\in\mathcal{F}^\sigma,l\geq|F_1|+|F_2|-|\sigma|.$$

(Note that it does not matter what l we choose, as the difference will be an eventually zero expression.) For example, instead of writing $p_{dot}(\rho,\rho;F)=p_{dot}(Z_3;F)+p_{dot}(Z_5;F)$, we

could simply write $\rho^2 = \rho \cdot \rho = Z_3 + Z_5$. For the flags of our running example, involving $\overline{K_3}$-free graphs, the following equations are also easily verifiable: $\rho^2 = Z_3 + Z_5$, $\overline{\rho}^2 = Z_2$, and $\rho \cdot \overline{\rho} = \frac{1}{2}Z_4 + \frac{1}{2}Z_1$. Combining these equations, we arrive at the following equation, which we shall later require in Section 3.4:

$$4\rho^2 \cdot \overline{\rho}^2 = 4Z_2 \cdot (Z_3 + Z_5) = (Z_4 + Z_1)^2. \tag{3.2}$$

To further simplify the notation, we can extend the definitions of p_σ and d_σ to $\mathcal{A}^\sigma$ by making them linear in each coordinate. For example, $p_\sigma(F_1 + 2F_2, 4F_3; F_4 - F_5) = 4p_\sigma(F_1, F_3; F_4) - 4p_\sigma(F_1, F_3; F_5) + 8p_\sigma(F_2, F_3; F_4) - 8p_\sigma(F_2, F_3; F_5)$. The product notation simplifies these extended definitions, because $p_\sigma(f_1 \cdot f_2; f) = p_\sigma(f_1, f_2; f)$ and $d_\sigma(f_1 \cdot f_2; g) = d_\sigma(f_1, f_2; g)$, for any $f_1, f_2, f \in \mathcal{A}^\sigma$ and for any $g \in \mathcal{A}^0$.

The last piece of notation we introduce is that of the averaging operator. Recall that for any σ-flag F, we had the normalizing factors $q_\sigma(F)$ such that $d_\sigma(F; G) = q_\sigma(F)p(F|_0; G)$. In the syntax of flag algebra, this averaging operation is denoted by $[[F]]_\sigma \stackrel{def}{=} q_\sigma F|_0$. We can extend this linearly to all elements of $\mathcal{A}^\sigma$. For example

$$[[\rho]]_{dot} = K_2, \quad [[Z_5]]_{dot} = K_3, \text{ and } \quad [[Z_4 + Z_2]]_{dot} = \frac{2}{3}P_2 + \frac{1}{3}\overline{P_2},$$

where P_2 is a path of length two on three vertices, and $\overline{P_2}$ is its complement. This notation is useful, because $d_\sigma(f; g) = p([[f]]_\sigma; g)$ for any $f \in \mathcal{A}^\sigma$ and for any $g \in \mathcal{A}^0$, and hence we have a unified notation for both types of flag densities.

3.3.2 Extremal problems in the flag algebra calculus

Recall that the typical problem is to minimise the density of some fixed graph J amongst all admissible graphs G not containing a forbidden subgraph. We will show how flag algebras can be applied to this problem to reduce it to a semi-definite programming (SDP) problem, which can then be solved numerically.

For any $t \geq |J|$, the equation $d(J; G) = \sum_{H \in \mathcal{F}_t} d(J; H)d(H; G)$ follows from the chain rule. Since $\sum_{H \in \mathcal{F}_t} d(H; G) = 1$, we have

$$d(J; G) \geq \min_{H \in \mathcal{F}_t} d(J; H),$$

which is a bound that clearly does not depend on G.

This inequality is often very weak, since it only uses very local considerations about the subgraphs $H \in \mathcal{F}_t$, and does not take into account how the subgraphs fit together in the larger graph G; that is, how they intersect. For instance, returning to our example of the $(3,3)$-problem, where $J = K_3$ and $t = 3$, we obtain $d(K_3; G) \geq \min_{H \in \mathcal{F}_3} d(K_3; H) = d(K_3; P_2) = 0$, which is the most trivial bound. However, by considering how the graphs in $\mathcal{F}_3$ must intersect in G, one might hope to find inequalities of the form $\sum_{H \in \mathcal{F}_t} \alpha_H d(H; G) \geq 0$, such that when we combine them with the initial identity, we get

$$d(J; G) \geq d(J; G) - \sum_{H \in \mathcal{F}_t} \alpha_H d(H; G) = \sum_{H \in \mathcal{F}_t} (d(J; H) - \alpha_H) d(H; G) \geq \min_{H \in \mathcal{F}_t} \{d(J; H) - \alpha_H\}.$$

Since α_H can be negative for some graphs H, the hope is that this will improve the low coefficients by transferring weight from high coefficients. In order to find such inequalities, we need another property of the flag densities.

Fact 3.3.2. If σ is a type of size k, $m \geq 1$ is an integer, $\{F_i\}_{i=1}^m$ is a family of σ-flags of sizes $|F_i| = l_i$, and $l \geq k + \sum_{i=1}^m (l_i - k)$ is an integer, then for any flag F of order $n \geq l$, we have

$$p_\sigma(F_1, \ldots, F_m; F) = \left[\prod_{i=1}^m p_\sigma(F_i; F) \right] + O(1/n).$$

One can prove Fact 3.3.2 by noting that, if we drop the requirement that the sets X_i are disjoint in the definition of $p_\sigma(F_1, \ldots, F_m; F)$, the events E_i will become independent, and thus $\mathbb{P}(\cap_{i=1}^m E_i) = \prod_{i=1}^m \mathbb{P}(E_i) = \prod_{i=1}^m p_\sigma(F_i; F)$. The error introduced is the probability that these sets X_i will intersect in F, which is $O(1/n)$. It is tempting to claim a similar product formula for the unlabeled flag densities d_σ, but we cannot do so. In the above equation, it is essential that all the σ-flags F_i share the same labeled type σ, and hence we require F to be a σ-flag.

We are now ready to establish some inequalities. Let's first fix a type σ of size k. If Q is any positive semi-definite $|\mathcal{F}_l^\sigma| \times |\mathcal{F}_l^\sigma|$ matrix with rows and columns indexed by the same set $\mathcal{F}_l^\sigma$, where $l \geq k$, define

$$Q\{\mathcal{F}_l^\sigma\} \stackrel{def}{=} \sum_{F_1, F_2 \in \mathcal{F}_l^\sigma} Q_{F_1, F_2} F_1 \cdot F_2 \in \mathcal{A}^\sigma.$$

Since Q was chosen to be positive semi-definite, we have

$$p_\sigma(Q\{\mathcal{F}_l^\sigma\}; F) = \sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} p_\sigma(F_1; F) p_\sigma(F_2; F) \geq 0$$

for any σ-flags F of order at least $t = 2l - k$. When averaging, we do not necessarily have $p([[Q\{\mathcal{F}_l^\sigma\}]]_\sigma; G) \geq 0$ for an admissible graph G of order $n \geq t$, but we do have the following inequality:

$$\begin{aligned}
[[Q]]_\sigma(G) &\stackrel{def}{=} p([[Q\{\mathcal{F}_l^\sigma\}]]_\sigma; G) = \sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} d_\sigma(F_1, F_2; G) \\
&= \sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} \left(\sum_{F\in\mathcal{F}_n^\sigma} p_\sigma(F_1, F_2; F) d_\sigma(F; G) \right) \\
&= \sum_{F\in\mathcal{F}_n^\sigma} \left(\sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} p_\sigma(F_1, F_2; F) \right) d_\sigma(F; G) \\
&= \sum_{F\in\mathcal{F}_n^\sigma} \left(\sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} p_\sigma(F_1; F) p_\sigma(F_2; F) \right) d_\sigma(F; G) + O(1/n) \geq o_{n\to\infty}(1).
\end{aligned}$$

Therefore, when n is large, we have that $[[Q]]_\sigma(G)$ is asymptotically non-negative. For each admissible graph H of size exactly t, let $\alpha_H = [[Q]]_\sigma(H) = \sum_{F_1,F_2\in\mathcal{F}_t^\sigma} Q_{F_1,F_2} d_\sigma(F_1, F_2; H)$. We then have

$$[[Q]]_\sigma(G) = \sum_{H\in\mathcal{F}_t} \alpha_H d(H; G) \geq o_{n\to\infty}(1).$$

The expression in the middle of the above equation is called the *expansion* of $[[Q]]_\sigma(G)$ in graphs of size t, with α_H the coefficients of the expansion. For the sake of conciseness, we often omit the parameter G and express this asymptotic inequality (combined with the expansion in size t) in the syntax of flag algebras

$$[[Q]]_\sigma \stackrel{def}{=} [[Q\{\mathcal{F}_l^\sigma\}]]_\sigma = \left[\left[\sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} F_1 \cdot F_2 \right]\right]_\sigma = \sum_{H\in\mathcal{F}_t} \alpha_H H \geq 0. \tag{3.3}$$

(Note that all inequalities between flags stated in the language of flag algebras are asymptotic.)

For a concrete example, we return to the $(3, 3)$-problem. If we use the type $\sigma = dot$, flags

of size $l = 2$, expand in graphs of size $t = 3$, and consider

$$Q = \begin{pmatrix} +\frac{3}{4} & -\frac{3}{4} \\ -\frac{3}{4} & +\frac{3}{4} \end{pmatrix},$$

where the rows and columns are indexed by ρ and $\overline{\rho}$ (in that order), we obtain $Q\{\mathcal{F}_2^{dot}\} = \frac{3}{4}(\rho - \overline{\rho})^2 = \frac{3}{4}(-Z_1 - Z_4 + Z_2 + Z_3 + Z_5)$. This expansion is obtained by substituting the expressions for ρ^2, $\overline{\rho}^2$ and $\rho \cdot \overline{\rho}$ that are given above Equation 3.2. Averaging gives $[[Q]]_\sigma = \frac{3}{4}[[(\rho - \overline{\rho})^2]]_{dot} = \frac{3}{4}K_3 - \frac{1}{4}P_2 - \frac{1}{4}\overline{P_2}$. Recall that $K_3 + P_2 + \overline{P_2} = 1$, since we are only considering $\overline{K_3}$-free graphs. Therefore $d(K_3; G) \geq \min_{H \in \mathcal{F}_3} \{d(K_3; H) - [[Q]]_\sigma(H)\} = \frac{1}{4}$, which is the correct bound for the $(3, 3)$-problem.

In general, if we have more than one inequality available, we can combine them together, provided they are all expanded in the same size t. Suppose we have r inequalities given by the positive semi-definite matrices Q_i of the σ_i-flags of size l_i. Adding them together, we obtain

$$\sum_{i=1}^{r} [[Q_i]]_{\sigma_i} = \sum_{H \in \mathcal{F}_t} \alpha_H H \geq 0,$$

where

$$\alpha_H = \sum_{i=1}^{r} \left(\sum_{F_1, F_2 \in \mathcal{F}_{l_i}^{\sigma_i}} (Q_i)_{F_1, F_2} d_{\sigma_i}(F_1, F_2; H) \right),$$

and we want to maximise $\min_{H \in \mathcal{F}_t} \{d(J; H) - \alpha_H\}$.

Thus we have transformed the original problem of finding a maximum lower bound for $d(J; G)$ into a linear system involving the variables $(Q_i)_{F_k, F_l}$. As we have the constraint that the matrices Q_i should be positive semi-definite, this is a semi-definite programming problem. To take the minimum coefficient in the expansion, we introduce an artificial variable y, and require it to be bounded above by all the coefficients. Hence we have the following SDP problem in the variables y and $(Q_i)_{F_1, F_2}$:

Maximise y, subject to the constraints:

- $s_H = d(J; H) - \sum_{i=1}^{r} \left(\sum_{F_1, F_2 \in \mathcal{F}_{l_i}^{\sigma_i}} (Q_i)_{F_1, F_2} d_{\sigma_i}(F_1, F_2; H) \right) - y \geq 0$ for all $H \in \mathcal{F}_t$. (The variables s_H are called *surplus* variables.)

- Q_i is positive semi-definite for $i \in [r]$. (The matrices Q_i are often called the *block variables* of the SDP problem. We can assume without loss of generality that each Q_i is symmetric, as otherwise we could replace Q_i by $(Q_i + Q_i^T)/2$.)

A computer can solve this SDP problem numerically, allowing for an efficient determination of the inequalities required to prove the extremal problem. For some practical remarks on the implementation of flag algebras, please see Section 3.7. We note at this point, as shall be seen in Section 3.4, that the solution to the SDP problem need not only give the asymptotic bound, but can also provide some structural information about the extremal graphs.

3.4 The $(4, 3)$-problem

In this section we will apply the flag algebra calculus to solve the $(4, 3)$-problem. Recall in the $(4, 3)$-problem we are interested in finding the minimum number of 4-cliques in a graph with independence number less than 3. We prove that any graph on n vertices with independence number at most 2 must contain at least $\frac{3}{25}\binom{n}{4} + O(n^3)$ 4-cliques. This bound is attained by a balanced blow-up of C_5, which Nikiforov conjectured to be optimal in [Nik05].

The first subsection contains our flag algebra results, which leads to the asymptotic minimum density of 4-cliques. In the second subsection we use the structural information from the flag algebras to derive a stability result. This allows us to determine the value of $f(n, 4, 3)$ exactly for large n, and we show that a nearly-balanced blow-up of C_5 is the unique extremal graph.

3.4.1 The asymptotic result

We begin by listing the admissible graphs of size 5, the types used in the proof, and the corresponding flags. Note that the flags of size 3 and type *dot* in Figure 3.6 are those we used as examples in Section 3.3.1, Figure 3.1.

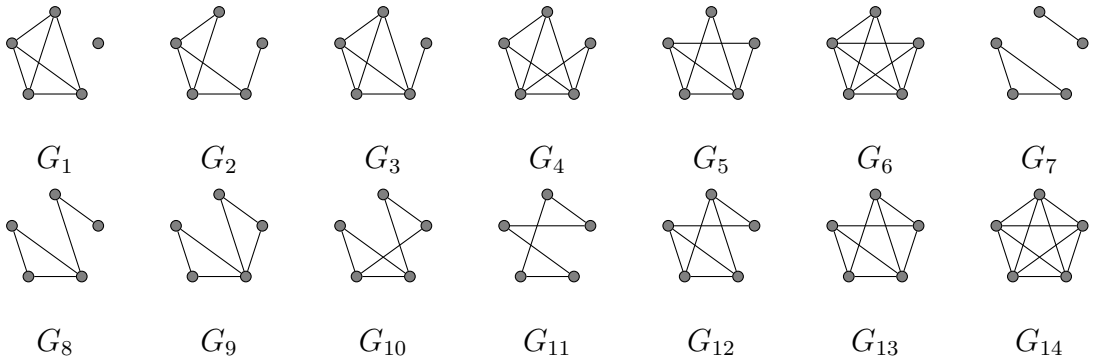


Figure 3.3: Graphs of size 5 with independence number at most 2.

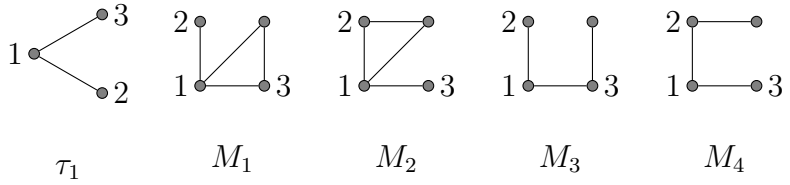


Figure 3.4: Type τ_1 and its flags of size 4.

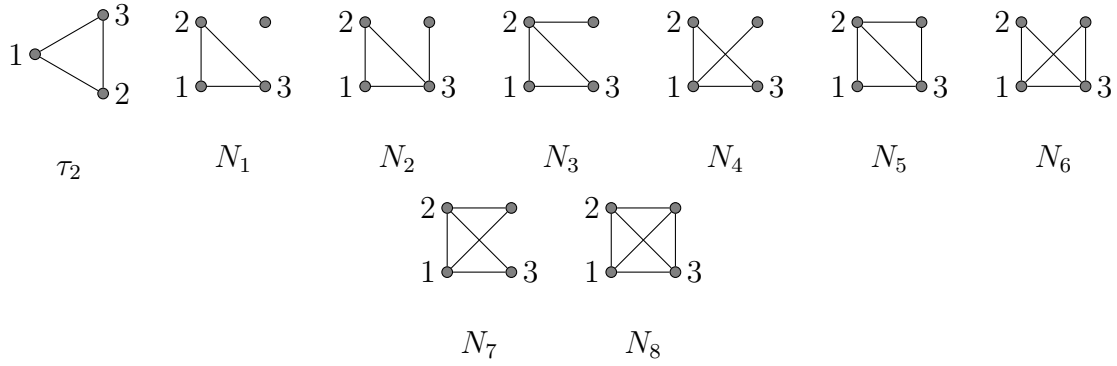


Figure 3.5: Type τ_2 and its flags of size 4.

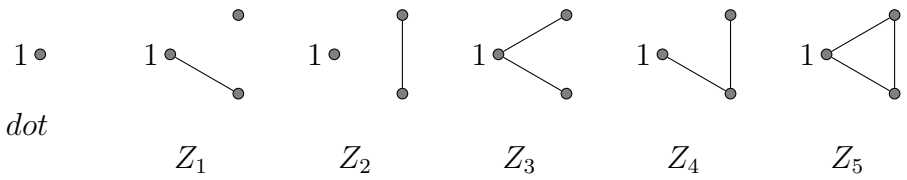


Figure 3.6: Type *dot* and its flags of size 3.

For each of the types used in the proof, we express the corresponding positive semi-definite matrices as a sum of squares. In the lemmas that follow, we give these sums of squares, their expansions into the admissible graphs of size 5, and provide sketches of combinatorial proofs (note that the lemmas were initially obtained by solving the corresponding SDP problem). We begin with the type τ_1.

Lemma 3.4.1.

$$\begin{aligned}\Delta_1 &= \left[\!\left[(M_2 + M_4 - M_1 - M_3)^2 \right]\!\right]_{\tau_1} \\ &= \frac{1}{30} \cdot (2G_2 + 3G_3 - G_5 - G_8 - 4G_9 - 2G_{10} - 5G_{11}) \geq 0.\end{aligned}$$

Sketch of proof. Let $G = (V, E)$ be a graph on n vertices. Define $\tau_1(G) = \{(x, y, z) \in V(G)^3 : \{x, y\}, \{x, z\} \in E(G) \text{ and } \{y, z\} \notin E(G)\}$. Every triple $(x, y, z) \in \tau_1(G)$ induces a copy of of the type τ_1 in G, where vertex x is labelled "1", vertex y is labelled "2" and vertex z is labelled "3". Fix some $p = (x, y, z) \in \tau_1(G)$. Note that M_2 and M_4 are flags where the unlabeled vertex is adjacent to 2 but not 3, while M_1 and M_3 are flags with the unlabeled vertex adjacent to 3 but not 2. Hence we define

$$d_p(v) \stackrel{def}{=} \begin{cases} 1, & \text{if } \{v, y\} \in E(G) \text{ but } \{v, z\} \notin E(G), \\ -1, & \text{if } \{v, z\} \in E(G) \text{ but } \{v, y\} \notin E(G), \\ 0, & \text{otherwise,} \end{cases}$$

for each $v \in V(G) \setminus \{x, y, z\}$. If we denote by F the flag induced by the labelled vertices $\{x, y, z\}$ together with the unlabelled vertex v, we have

$$d_p(v) = \begin{cases} 1, & \text{if } F = M_2 \text{ or } F = M_4, \\ -1, & \text{if } F = M_1 \text{ or } F = M_3, \\ 0, & \text{otherwise.} \end{cases}$$

Thus the combinatorial interpretation of the lemma is

$$\Delta_1(G) = \frac{1}{3!\binom{n}{3}} \cdot \left[\sum_{p=(x,y,z)\in\tau_1(G)} \frac{1}{2\binom{n-3}{2}} \left(\sum_{\substack{v,w\notin\{x,y,z\}\\ v\neq w}} d_p(v)d_p(w) \right) \right]$$
$$= \frac{1}{120\binom{n}{5}} \sum_{p=(x,y,z)\in\tau_1(G)} \sum_{\substack{v,w\notin\{x,y,z\}\\ v\neq w}} d_p(v)d_p(w) \geq o_{n\to\infty}(1).$$

The proof that this summation is asymptotically non-negative is very simple, since

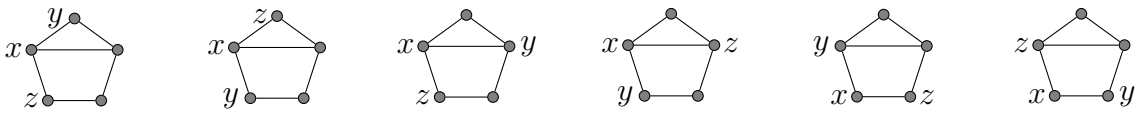
$$\sum_{\substack{v,w\notin\{x,y,z\}\\ v\neq w}} d_p(v)d_p(w) = \left(\sum_{v\notin\{x,y,z\}} d_p(v) \right)^2 - \sum_{v\notin\{x,y,z\}} d_p(v)^2,$$

and

$$\frac{1}{120\binom{n}{5}} \cdot \left[\sum_{p=(x,y,z)\in\tau_1(G)} \left(\sum_{v\notin\{x,y,z\}} d_p(v)^2 \right) \right] = O(1/n).$$

It remains to expand the products of the flags into admissible graphs of size 5, and thus show that $\Delta_1 = \frac{1}{30} \cdot (2G_2 + 3G_3 - G_5 - G_8 - 4G_9 - 2G_{10} - 5G_{11})$. For the sake of conciseness, we omit the full details of this calculation. We show how to compute the coefficient of G_{10}, that is, $\Delta_1(G_{10})$; the other coefficients follow similarly.

In this case, the set $\{x, y, z, v, w\}$ spans a copy of G_{10}.



-4 -4 0 0 0 0

Figure 3.7: Possible configurations of p inside G_{10} and corresponding contributions to $\Delta_1(G_{10})$.

We have the following cases:

1. Vertex x is one of the vertices of degree 3. There are two choices of x satisfying this condition. We have the following subcases:

(a) Vertex y is the vertex of degree 2 of the triangle containing x and z is only neighbor of x which is not adjacent to y. This configuration corresponds to the first graph in Figure 3.7. As one of the unlabeled vertices is adjacent to y and not z, and the other is adjacent to z and not y, both assignments of v and w, we have $d_p(v)d_p(w) = -1$. As there are two choices for the pair (v, w) and two choices for x, the total contribution for this configuration is -4.

(b) The same configuration as above, but with the roles of y and z swapped. This configuration corresponds to the second graph in Figure 3.7 and its contribution is -4.

(c) Vertex y is the other vertex of degree 3 and z is the only neighbor of x which is not adjacent to y. This configuration corresponds to the third graph in Figure 3.7. For any possible choice of v and w, we have $d_p(v) \cdot d_p(w) = 0$, hence the total contribution is 0.

(d) The same configuration as above, but with the roles of y and z swapped. This configuration corresponds to the fourth graph in Figure 3.7 and its contribution is 0.

2. Vertex x is one of the vertices of degree 2 not in the triangle. Again we have two choices of x satisfying this condition. We also have the following subcases:

(a) Vertex y is the only neighbor of x of degree 3 and z is the other neighbor. This configuration corresponds to the fifth graph in Figure 3.7. For any possible choice of v and w, we have $d_p(v) \cdot d_p(w) = 0$, hence the total contribution for this configuration is 0.

(b) The same configuration as above, but with the roles of y and z swapped. This configuration corresponds to the last graph in Figure 3.7 and its contribution is 0.

When we sum the contributions we get -8, and hence the coefficient of G_{10} is $\Delta_1(G_{10}) = -\frac{8}{120} = -\frac{1}{15}$. □

We now consider the type τ_2.

Lemma 3.4.2.

$$\begin{aligned}\Delta_2 &= \left[\left[\left(-3N_1 - 3N_2 - 3N_3 - 3N_4 + 2N_5 + 2N_6 + 2N_7 + 2N_8\right)^2 \right]\right]_{\tau_2} \\ &= \frac{1}{10} \cdot (-24G_1 - 12G_2 - 24G_3 - 8G_5 + 28G_6 + 9G_7 + 9G_8 + \\ &\qquad 18G_9 + 9G_{10} - 12G_{12} + 16G_{13} + 40G_{14}) \geq 0.\end{aligned}$$

Sketch of proof. Let $G = (V, E)$ be a graph on n vertices. Define $\tau_2(G) = \{(x, y, z) \in V(G)^3 : \{x, y\}, \{x, z\}, \{y, z\} \in E(G)\}$. Every triple $(x, y, z) \in \tau_2(G)$ induces a copy of of the type τ_2 in G, where vertex x is labelled "1", vertex y is labelled "2" and vertex z is labelled "3". Fix $p = (x, y, z) \in \tau_2(G)$. Note that the flags N_i for $1 \leq i \leq 4$ are those where the unlabeled vertex has at most one neighbour in the triangle τ_2, while in the flags N_i for $5 \leq i \leq 8$, the unlabeled vertex has at least two neighbours in τ_2. This motivates the definition

$$d_p(v) \stackrel{def}{=} \begin{cases} -3, & \text{if } v \text{ is connected to at most one vertex in } \{x, y, z\}, \\ 2, & \text{otherwise,} \end{cases}$$

for each $v \in V(G) \setminus \{x, y, z\}$. The combinatorial interpretation of the lemma is

$$\Delta_2(G) = \frac{1}{5!\binom{n}{5}} \left[\sum_{p=(x,y,z)\in\tau_2(G)} \left(\sum_{\substack{v,w\notin\{x,y,z\} \\ v\neq w}} d_p(v) d_p(w) \right) \right] \geq o_{n\to\infty}(1).$$

As in Lemma 3.4.1, this is easily seen to be asymptotically positive. We omit the computation of $\Delta_2(G_i)$ for $i = 1, 2, \ldots, 14$, which can be performed as in the proof of the previous lemma. □

Finally we consider the *dot* type. Note that in this case the positive semi-definite matrix takes the form of a sum of three squares.

Lemma 3.4.3.

$$\begin{aligned}\Delta_3 &= \left[\left[(Z_1 - 2Z_2)^2 + \frac{1}{16} \cdot (6Z_2 - 7Z_3 + 8Z_4 - 6Z_5)^2 + \frac{11}{80} \cdot (2Z_2 + 3Z_3 - 2Z_5)^2 \right]\right]_{dot} \\ &= \frac{1}{150} (204G_1 - 118G_2 + 54G_3 + 60G_4 - 17G_5 + 42G_6 - 144G_7 - 94G_8 + \\ &\qquad 2G_9 - 64G_{10} + 160G_{11} - 258G_{12} - 281G_{13} + 420G_{14}) \geq 0.\end{aligned}$$

Proof. We omit the proof, noting that the calculations involved are very similar to those in the previous lemmas. □

We are now in a position to combine the lemmas to obtain a bound on the minimum density of 4-cliques in admissible graphs. In what follows, K_4 represents the clique on four vertices, while C_4 denotes a cycle on four vertices.

Theorem 3.4.4.

$$\begin{aligned} K_4 - 2\Delta_1 - \frac{2}{25}\Delta_2 - \frac{1}{5}\Delta_3 &= \frac{3}{25} + \frac{1}{30}G_5 + \frac{2}{75}G_{10} + \frac{24}{75}G_{12} + \frac{19}{150}G_{13} \\ &= \frac{3}{25} + \frac{1}{30}G_5 + \frac{2}{15}C_4 + \frac{4}{15}G_{12} + \frac{1}{10}G_{13}. \end{aligned}$$

Proof. We first expand the graphs K_4 and C_4 into admissible graphs of size 5. A straightforward calculation gives $K_4 = \frac{1}{5}(G_1 + G_3 + G_4 + 2G_6 + 5G_{14})$, and $C_4 = \frac{1}{5}(G_{10} + 2G_{12} + G_{13})$. Note that the density of graphs on k vertices is measured with respect to $\binom{n}{k}$, and so the normalisation factor of $\frac{1}{5}$ appears when expanding graphs on four vertices to graphs on five vertices. Now we use Lemmas 3.4.1, 3.4.2 and 3.4.3 to expand Δ_1, Δ_2 and Δ_3 into the graphs G_i. Noting that $\sum_i G_i = 1$, we can replace $\frac{3}{25}\sum_i G_i$ with $\frac{3}{25}$, which results in the above theorem. □

We conclude this section by using the above theorem to deduce some structural information about extremal graphs. Recall that $t_4(G)$ denotes the number of 4-cliques in G, while for any graph H, $t_H(G)$ counts the number of induced copies of H in G.

Corollary 3.4.5. Suppose G is a graph on n vertices with $t_4(G) = \left(\frac{3}{25} + o(1)\right)\binom{n}{4}$. Then

(i) $t_{G_5}(G) = o(n^5)$,

(ii) $t_{C_4}(G) = o(n^4)$, and

(iii) all but $o(n)$ vertices of G have degree $(\frac{3}{5} + o(1))n$.

Proof. Applying Theorem 3.4.4 to G, we have

$$\begin{aligned} d(K_4;G) - 2\Delta_1(G) - \frac{2}{25}\Delta_2(G) - \frac{1}{5}\Delta_3(G) \\ = \frac{3}{25} + \frac{1}{30}d(G_5;G) + \frac{2}{15}d(C_4;G) + \frac{4}{15}d(G_{12};G) + \frac{1}{10}d(G_{13};G). \end{aligned}$$

In particular, using the asymptotic non-negativity of $\Delta_i(G)$, we have

$$d(K_4;G) \geq \frac{3}{25} + \frac{1}{30}d(G_5;G) + \frac{2}{15}d(C_4;G) + \frac{1}{5}\Delta_3(G) + o(1).$$

Thus if $d(K_4;G) = \frac{3}{25} + o(1)$, we must have $d(G_5;G) = d(C_4;G) = \Delta_3(G) = o(1)$. This immediately gives $t_{G_5}(G) = o(n^5)$ and $t_{C_4}(G) = o(n^4)$, and so it remains to justify (iii). We have

$$\Delta_3(G) = \left[\left[(Z_1 - 2Z_2)^2 + \frac{1}{16}(6Z_2 - 7Z_3 + 8Z_4 - 6Z_5)^2 + \frac{11}{80}(2Z_2 + 3Z_3 - 2Z_5)^2\right]\right]_{dot} = o(1).$$

For every vertex v, let F_v be the *dot*-flag obtained from G by labeling the vertex v with 1. By definition of the averaging operator, $\Delta_3(G)$ is the average over vertices v of the corresponding flag densities in F_v. The expression is a sum of squares, and thus will be asymptotically non-negative. Since the average is $o(1)$, the expression must be $o(1)$ for all but $o(n)$ vertices. In particular, for these vertices we have

$$\begin{aligned} p_{dot}(Z_1;F_v) - 2p_{dot}(Z_2;F_v) &= o(1), \\ 6p_{dot}(Z_2;F_v) - 7p_{dot}(Z_3;F_v) + 8p_{dot}(Z_4;F_v) - 6p_{dot}(Z_5;F_v) &= o(1), \text{ and} \\ 2p_{dot}(Z_2;F_v) + 3p_{dot}(Z_3;F_v) - 2p_{dot}(Z_5;F_v) &= o(1). \end{aligned}$$

Since the sum of the flag densities must be 1, we also have

$$p_{dot}(Z_1;F_v) + p_{dot}(Z_2;F_v) + p_{dot}(Z_3;F_v) + p_{dot}(Z_4;F_v) + p_{dot}(Z_5;F_v) = 1.$$

Finally, recall from Equation (3.2) in Section 3.3 that $4Z_2 \cdot (Z_3 + Z_5) - (Z_4 + Z_1)^2 = 0$. Applying this to F_v, we have

$$4p_{dot}(Z_2;F_v)\left(p_{dot}(Z_3;F_v) + p_{dot}(Z_5;F_v)\right) - \left(p_{dot}(Z_4;F_v) + p_{dot}(Z_1;F_v)\right)^2 = o(1).$$

This gives us a system of five equations in the five variables $p_{dot}(Z_i; F_v)$. The first four equations form a linear system of full rank, which we can use to express all the variables in terms of $p_{dot}(Z_5; F_v)$. Substituting these terms into the fifth equation gives a quadratic equation in $p_{dot}(Z_5; F_v)$, which results in two solutions, namely $(p_{dot}(Z_i; F_v))_{i=1}^5 = \left(\frac{8}{25}, \frac{4}{25}, \frac{2}{25}, \frac{4}{25}, \frac{7}{25}\right) + o(1)$ or $\left(\frac{1}{2}, \frac{1}{4}, 0, 0, \frac{1}{4}\right) + o(1)$.

We now show that the second solution implies a large number of 4-cliques. Indeed, suppose $v \in V$ was a vertex with $(p_{dot}(Z_i; F_v))_{i=1}^5 = \left(\frac{1}{2}, \frac{1}{4}, 0, 0, \frac{1}{4}\right) + o(1)$. Recall from Equation (3.1) in Section 3 we have $\rho = \frac{1}{2}Z_1 + Z_3 + \frac{1}{2}Z_4 + Z_5$, where ρ is the *dot*-flag of size 2 corresponding to an edge. Applying this to the flag F_v, we deduce that the degree of v is $\left(\frac{1}{2} \cdot \frac{1}{2} + \frac{1}{4} + o(1)\right) n = \frac{1}{2}n + o(n)$. Thus there are $\frac{1}{2}n + o(n)$ vertices v is not adjacent to, and since G is $\overline{K_3}$-free, these vertices must form a clique. This clique contains $\binom{\frac{1}{2}n+o(n)}{4} \sim \frac{1}{16}\binom{n}{4}$ 4-cliques. Consider now the neighborhood of v. Since $p_{dot}(Z_3; F_v) = o(1)$, it follows that the neighborhood is missing at most $o(n^2)$ edges. Hence the number of 4-cliques in the neighborhood of v is $\binom{\frac{1}{2}n+o(n)}{4} - o(n^4) \sim \frac{1}{16}\binom{n}{4}$. Thus we have $t_4(G) \geq \left(\frac{1}{8} + o(1)\right)\binom{n}{4}$, which contradicts our assumption that $t_4(G) = \left(\frac{3}{25} + o(1)\right)\binom{n}{4}$.

Hence for almost all vertices v, we have $(p_{dot}(Z_i; F_v))_{i=1}^5 = \left(\frac{8}{25}, \frac{4}{25}, \frac{2}{25}, \frac{4}{25}, \frac{7}{25}\right) + o(1)$. Applying Equation (3.1), we deduce that the degree of v is $\left(\frac{1}{2} \cdot \frac{8}{25} + \frac{2}{25} + \frac{1}{2} \cdot \frac{4}{25} + \frac{7}{25} + o(1)\right) n = \left(\frac{3}{5}n + o(1)\right) n$, as claimed.

□

3.4.2 The stability analysis

We will now use the results of the preceding section to show that, for sufficiently large n, a blow-up of C_5 is the unique extremal graph for the $(4, 3)$-problem. Recall that in a blow-up, we replace every vertex with a clique, and every edge with a complete bipartite graph. Hence a blow-up of C_5 consists of five disjoint sets of vertices V_i, with $V_i \cup V_{i+1}$ a clique for all $1 \leq i \leq 5$, and no edges between V_i and V_{i+2} for all $1 \leq i \leq 5$ (throughout this section, indices will be taken modulo 5).

Suppose G is a $\overline{K_3}$-free graph on n vertices with the minimal number of 4-cliques. Our

proof consists of three steps. We first use the results of Corollary 3.4.5 to deduce that G is close to being a blow-up of C_5 (note that this holds not only for an extremal graph, but for any family of graphs that is asymptotically optimal). In the second step we use the minimality of G to show that G must in fact be a blow-up of C_5 with asymptotically equal parts. Finally, we solve an integer optimisation problem to determine the size of the parts of G exactly.

Recall that from Corollary 3.4.5, we have that if n is sufficiently large, and G is an extremal graph on n vertices, then $t_4(G) = \frac{3}{25}\binom{n}{4} + o(n^4)$, $t_{C_4}(G) = o(n^4)$, $t_{G_5}(G) = o(n^5)$, and all but $o(n)$ vertices of G have degree $\frac{3}{5}n + o(n)$. From this we shall deduce that G is almost a blow-up of C_5. To this end, we introduce some definitions. Given subsets $A, B \subset V(G)$, we say A is an *almost clique* if all but $o(n^2)$ pairs in A are adjacent, and we say (A, B) is *almost complete* (*almost empty*) if all but $o(n^2)$ pairs in $A \times B$ are adjacent (nonadjacent). Finally, we define a triple $\{a, b, c\} \in V(G)$ to be *typical* if:

(i) $\{a, b\} \notin E(G)$, $c \in N(a) \cap N(b)$, $d(a), d(b), d(c) = \frac{3}{5}n + o(n)$,

(ii) $\{a, b\}$ is contained in $o(n^2)$ copies of C_4,

(iii) $\{a, b, c\}$ is contained in $o(n)$ copies of C_4, and

(iv) $\{a, b, c\}$ is contained in $o(n^2)$ copies of G_5.

Note that $G[\{a, b, c\}]$ is an induced path of length 2. As all but $o(n)$ vertices are of degree $\frac{3}{5}n + o(n)$, it is easy to see that there are $\Omega(n^3)$ induced paths of length 2 in G. As Corollary 3.4.5 asserts that $t_{C_4}(G) = o(n^4)$ and $t_{G_5}(G) = o(n^5)$, it follows that almost all induced paths of length 2 are typical. We will now use the neighborhoods of $\{a, b, c\}$ to define the parts corresponding to the blow-up of C_5. In particular, we define

$$V_1 = N(a) \cap N(b), V_2 = \{a\} \cup \left(N(a) \cap \overline{N(b)} \cap N(c)\right), V_3 = N(a) \cap \overline{N(b)} \cap \overline{N(c)},$$
$$V_4 = \overline{N(a)} \cap N(b) \cap \overline{N(c)}, \text{ and } V_5 = \{b\} \cup \left(\overline{N(a)} \cap N(b) \cap N(c)\right).$$

We now make some preliminary observations about the sets V_i. Clearly, by definition, the sets are disjoint. Moreover, since $\alpha(G) \leq 2$, and $\{a, b\} \notin E(G)$, we must have $N(a) \cup N(b) =$

$V(G) \setminus \{a, b\}$, and so $\cup_i V_i = V(G)$. Similarly, for any vertex $v \in V(G)$, $\overline{N(v)}$ must induce a clique, as any non-edge in $\overline{N(v)}$ forms an independent set of size three with v. Thus $V_2 \cup V_3$, $V_3 \cup V_4$, and $V_4 \cup V_5$ are (actual) cliques. Finally, note that if $u, v \in V_1$ are such that $\{u, v\} \notin E(G)$, then the set $\{a, b, u, v\}$ induces a copy of C_4. Since $\{a, b, c\}$ was chosen to be a typical triple, properties (ii) and (iii) imply that V_1 is an almost clique, and c is adjacent to all but $o(n)$ vertices in V_1.

We can also obtain some relations regarding the sizes of these parts. By property (i) of typical triples, we have $d(a), d(b), d(c) = \frac{3}{5}n + o(n)$. Since $N(a) \cup N(b) = V(G) \setminus \{a, b\}$, we have $|V_1| = |N(a) \cap N(b)| = |N(a)| + |N(b)| - |N(a) \cup N(b)| = \frac{1}{5}n + o(n)$. Moreover, as $N(a) \cup \{a\} = V_1 \cup V_2 \cup V_3$, $N(b) \cup \{b\} = V_1 \cup V_4 \cup V_5$, $\overline{N(c)} \setminus V_1 = V_3 \cup V_4$, and c has $o(n)$ non-neighbors in V_1, we deduce

$$|V_2| + |V_3| = \frac{2}{5}n + o(n), |V_3| + |V_4| = \frac{2}{5}n + o(n), \text{ and } |V_4| + |V_5| = \frac{2}{5}n + o(n),$$

which also imply $|V_2| + |V_5| = \frac{2}{5}n + o(n)$.

We are beginning to uncover the approximate C_5-blow-up structure of G. Recall that we have shown that $V_2 \cup V_3$, $V_3 \cup V_4$ and $V_4 \cup V_5$ are cliques, while V_1 is an almost clique. We will establish the relations between the remaining parts by showing:

- (V_i, V_{i+2}) is almost empty for any $1 \leq i \leq 5$, and
- (V_1, V_2) and (V_1, V_5) are almost complete.

We start by showing that (V_1, V_3) is almost empty. For any $u \in V_1 \cap N(c)$ and $v \in V_3$, if $\{u, v\} \in E(G)$, then the set $\{a, b, c, u, v\}$ induces a copy of G_5. As $\{a, b, c\}$ is a typical triple, property (iv) implies that there are at most $o(n^2)$ copies of G_5 containing $\{a, b, c\}$, and so there are at most $o(n^2)$ edges between $V_1 \cap N(c)$ and V_3. Since c is adjacent to all but $o(n)$ vertices in V_1, this shows that (V_1, V_3) is almost empty. By the symmetry between a and b (and hence V_3 and V_4), it follows that (V_1, V_4) is also almost empty.

Now consider the vertices in V_1. By Corollary 3.4.5, all but $o(n)$ of these vertices have degree $\frac{3}{5}n + o(n)$. Since $(V_1, V_3 \cup V_4)$ is almost empty, it follows that all but $o(n)$ vertices in

V_1 have $o(n)$ edges to $V_3 \cup V_4$. Hence, since $|V_1| + |V_2| + |V_5| = \frac{3}{5}n + o(n)$, it follows that V_1 is almost complete to $V_1 \cup V_2 \cup V_5$. In particular, (V_1, V_2) and (V_1, V_5) are almost complete.

Next consider the vertices in V_2. We have established that $(V_2, V_1 \cup V_2 \cup V_3)$ is almost complete. Once again, using the restriction on the degrees, and the fact that $|V_1|+|V_2|+|V_3| = \frac{3}{5}n + o(n)$, we deduce that (V_2, V_4) and (V_2, V_5) are almost empty. Symmetry implies (V_5, V_2) and (V_5, V_3) are almost empty as well, as claimed.

At this point we have determined the global structure of G, in which each part V_i corresponds approximately to the blow-up of a vertex in C_5. We now wish to show that G is an exact blow-up of C_5, with parts of size $\frac{1}{5}n + o(n)$.

In order to do so, we shall require greater control over the adjacency of individual vertices, and not just the parts V_i. With this in mind, for each $1 \leq i \leq 5$, we define a vertex $v \in V_i$ to be *bad* if v has $\Omega(n)$ non-neighbors in $V_{i-1} \cup V_i \cup V_{i+1}$ or $\Omega(n)$ neighbors in $V_{i+2} \cup V_{i+3}$. Since for each i we have that $V_i \cup V_{i+1}$ is an almost clique and (V_i, V_{i+2}) is almost empty, it follows that there are $o(n)$ bad vertices. We clean up the partition of $V(G)$ by removing bad vertices from each V_i and placing them in a set U. This results in a partition $V(G) = V_1 \cup \ldots \cup V_5 \cup U$ satisfying:

(1) for any $1 \leq i \leq 5$ and vertex $v \in V_i$, v is adjacent to all but $o(n)$ vertices in $V_{i-1} \cup V_i \cup V_{i+1}$, and v is not adjacent to all but $o(n)$ vertices in $V_{i+2} \cup V_{i+3}$, and

(2) $V_2 \cup V_3$, $V_3 \cup V_4$, $V_4 \cup V_5$ are cliques, and

(3) $|V_1| = \frac{1}{5}n + o(n)$, $|V_2 \cup V_3|, |V_3 \cup V_4|, |V_4 \cup V_5| = \frac{2}{5}n + o(n)$, and $|U| = o(n)$.

The following proposition asserts that in an asymptotically optimal graph, the above conditions imply that the almost cliques are, in fact, true cliques, and that the parts are asymptotically equal. This will in turn allow us to completely determine the structure of extremal graphs.

Proposition 3.4.6. If $V_1, V_2, \ldots, V_5$ satisfy (1), (2) and (3), then for any $1 \leq i \leq 5$, $V_i \cup V_{i+1}$ is a clique, and $|V_i| = \frac{1}{5}n + o(n)$.

Proof. We already know from (2) that many of the pairs of neighboring parts are cliques. It remains to show that $V_1 \cup V_2$ and $V_5 \cup V_1$ are both cliques. We first show that V_1 is a clique. Suppose for contradiction that there are nonadjacent vertices $u, v \in V_1$. Since $\alpha(G) \leq 2$, we must have $V_3 \cup V_4 \subset N(u) \cup N(v)$. By (3) we have $|V_3 \cup V_4| = \frac{2}{5}n + o(n)$, and so either u or v must have at least $\frac{1}{5}n + o(n)$ neighbors in $V_3 \cup V_4$. However, this contradicts (1). Thus V_1 is a clique.

We now claim that if (V_1, V_2) is not complete, we must have $|V_4| = o(n)$. Indeed, suppose $u \in V_1$ and $v \in V_2$ are not adjacent. Since $\alpha(G) \leq 2$, we must have $V_4 \subset N(u) \cup N(v)$. By (1), both u and v have $o(n)$ neighbors in V_4, which implies $|V_4| = o(n)$. By symmetry, if (V_1, V_5) is not complete, we must have $|V_3| = o(n)$.

Suppose now that one of these sets, say V_4, is of size $o(n)$. Using (3), we must have $|V_3| = |V_5| = \frac{2}{5}n + o(n)$, and $|V_2| = o(n)$. Since $|V_3| \neq o(n)$, it follows that (V_1, V_5) is complete. Thus G has two large disjoint cliques: V_3 of size $\frac{2}{5}n + o(n)$, and $V_1 \cup V_5$ of size $\frac{3}{5}n + o(n)$. This gives

$$t_4(G) \geq \binom{\frac{2}{5}n + o(n)}{4} + \binom{\frac{3}{5}n + o(n)}{4} \sim \frac{97}{625}\binom{n}{4} + o(n^4) > \frac{3}{25}\binom{n}{4},$$

contradicting the asymptotic optimality of G. Hence (V_1, V_2) and (V_1, V_5) must be complete, which implies that $V_1 \cup V_2$ and $V_1 \cup V_5$ are cliques.

Finally, we show that all parts have size $\frac{1}{5}n + o(n)$. Recall we already have $|V_1| = \frac{1}{5}n + o(n)$. Since $|V_3| + |V_4| = \frac{2}{5}n + o(n)$, we may by symmetry assume $|V_3| \geq \frac{1}{5}n + o(n)$. Corollary 3.4.5 implies there is some vertex of V_3 whose degree is $\frac{3}{5}n + o(n)$. By (1), this implies $|V_2| + |V_3| + |V_4| = \frac{3}{5}n + o(n)$. As $|V_3| + |V_4| = \frac{2}{5}n + o(n)$, this implies $|V_2| = \frac{1}{5}n + o(n)$. Combined with the equations in (3), this gives $|V_i| = \frac{1}{5}n + o(n)$ for all $2 \leq i \leq 5$. □

We now turn our attention to the set U of bad vertices. In particular, we will show that in an extremal graph, each $u \in U$ can be reintroduced into some part V_i in a way that is consistent with (1) and Proposition 3.4.6. Since $|U| = o(n)$, we can repeat this process without affecting (1) or Proposition 3.4.6, and thus we can eliminate the set U.

Proposition 3.4.7. For every $u \in U$, there is some $i = i(u)$ such that $V_{i-1} \cup V_i \cup V_{i+1} \subset N(u)$, and u has $o(n)$ neighbors in $V_{i+2} \cup V_{i+3}$.

Proof. Fix $u \in U$. We begin with a simple claim. For any $1 \leq j \leq 5$, if there is some $v \in V_j$ such that u is not adjacent to v, then u is adjacent to all but $o(n)$ vertices in $V_{j+2} \cup V_{j+3}$. Indeed, as $\alpha(G) \leq 2$, we must have $V_{j+2} \cup V_{j+3} \subset N(u) \cup N(v)$. However, v is adjacent to $o(n)$ vertices in $V_{j+2} \cup V_{j+3}$, and so the claim follows.

Now suppose there is no i such that $V_{i-1} \cup V_i \cup V_{i+1} \subset N(u)$. This implies there is an i such that u is not adjacent to some vertices in both V_{i-3} and V_{i-1}. Applying the previous claim, it follows that u is adjacent to all but $o(n)$ vertices in $V_{i-1} \cup V_i \cup V_{i+1} \cup V_{i+2}$.

In this case, remove all edges between u and V_{i+2}, and add any missing edges between u and $V_{i-1} \cup V_i \cup V_{i+1} \cup U$. It is easy to see that we still have $\alpha(G) \leq 2$. As u had $\frac{1}{5}n + o(n)$ neighbors in V_{i+2}, which is a clique, we have removed at least $\binom{\frac{1}{5}n+o(n)}{3} = \Omega(n^3)$ 4-cliques. On the other hand, we have only added $o(n)$ edges, and so created $o(n^3)$ new 4-cliques. Thus we have reduced the number of 4-cliques, which contradicts the extremality of G.

Thus there must be some $i = i(u)$ such that $V_{i-1} \cup V_i \cup V_{i+1} \subset N(u)$. It remains to show that u has $o(n)$ neighbors in $V_{i+2} \cup V_{i+3}$. Suppose for contradiction that u has $\Omega(n)$ neighbors in $V_{i+2} \cup V_{i+3}$. As $V_{i+2} \cup V_{i+3}$ is a clique, these neighbors form $\Omega(n^3)$ 4-cliques with u. Instead, we could remove all edges between u and $V_{i+2} \cup V_{i+3}$. To prevent the formation of an independent set of size 3, we add all edges between u and U. This introduces $o(n)$ new edges, and thus $o(n^3)$ new 4-cliques, while maintaining $\alpha(G) \leq 2$. Thus the number of 4-cliques is reduced, again contradicting the minimality of G. This completes the proof.

□

Given any $u \in U$, we can apply Proposition 3.4.7 to add u to $V_{i(u)}$. Repeat this process until U is empty. In this case we have a partition $V(G) = V_1 \cup \ldots \cup V_5$ such that for every $1 \leq i \leq 5$, $|V_i| = \frac{1}{5}n + o(n)$ and $V_i \cup V_{i+1}$ is a clique.

In order to conclude that G is a blow-up of C_5, it remains to show that there are no edges between V_{i-1} and V_{i+1} for any i. Suppose to the contrary there is an edge between some

$v \in V_{i-1}$ and $w \in V_{i+1}$. Note that when n is large, we must have $|V_i| = \frac{1}{5}n + o(n) \geq 2$. For any $x, y \in V_i$, $\{v, w, x, y\}$ is a 4-clique. Thus removing the edge $\{v, w\}$ reduces the number of 4-cliques without increasing the independence number. Hence in an extremal graph, there are no edges between V_{i-1} and V_{i+1} for any i, and thus G is indeed a blow-up of C_5 with parts of size $\frac{1}{5}n + o(n)$.

We now seek to determine the sizes of the sets V_i exactly. Noting that $V_i \cup V_{i+1}$ is a clique for each i, it is easily verified that

$$t_4(G) = \sum_{i=1}^{5} \binom{|V_i \cup V_{i+1}|}{4} - \sum_{i=1}^{5} \binom{|V_i|}{4}.$$

Define $y_i = |V_{2i-1} \cup V_{2i}|$ for all $1 \leq i \leq 5$. In $\sum y_i$, each vertex is counted twice, so we have $\sum y_i = 2n$. Moreover, as $|V_i| = \frac{1}{5}n + o(n)$, we have $y_i = \frac{2}{5}n + o(n)$. Finally, as $n - y_i - y_{i+1} = n - |V_{2i-1}| - |V_{2i}| - |V_{2i+1}| - |V_{2i+2}| = |V_{2i-2}|$, we can rewrite the above expression as

$$t_4(G) = \sum_{i=1}^{5} \binom{y_i}{4} - \sum_{i=1}^{5} \binom{n - y_i - y_{i+1}}{5}.$$

Thus to find the extremal graph, we must minimise the above expression over integer values of y_i subject to the conditions given earlier. The solution is given by Lemma 3.4.8, which we prove in Section 3.8.

Lemma 3.4.8. Let $\varepsilon > 0$ be sufficiently small, and n sufficiently large. Consider the function

$$g(y_1, y_2, y_3, y_4, y_5) = \sum_{i=1}^{5} \binom{y_i}{5} - \sum_{i=1}^{5} \binom{n - y_i - y_{i+1}}{4}.$$

Subject to the constraints that the y_i be integers satisfying $\sum_{i=1}^{5} y_i = 2n$ and $\left|y_i - \frac{2}{5}n\right| < \varepsilon n$, g is uniquely (up to cyclic permutation of the variables) minimised when the y_i take values $\left\lfloor \frac{2n}{5} \right\rfloor$ and $\left\lceil \frac{2n}{5} \right\rceil$ in ascending order.

From Lemma 3.4.8, we see the minimum occurs when $y_i = \left\lceil \frac{2n+i-1}{5} \right\rceil$ for $1 \leq i \leq 5$. Solving for $|V_i|$, we have that the unique extremal graph on n vertices is the blow-up of C_5 to n vertices such that:

- when $n = 5k$, $|V_i| = k$ for all i,

- when $n = 5k + 1$, $|V_1| = |V_2| = k$, $|V_3| = |V_5| = k + 1$, and $|V_4| = k - 1$,
- when $n = 5k + 2$, $|V_1| = |V_2| = |V_4| = k$, and $|V_3| = |V_5| = k + 1$,
- when $n = 5k + 3$, $|V_1| = |V_2| = |V_4| = k + 1$, and $|V_3| = |V_5| = k$, and
- when $n = 5k + 4$, $|V_1| = |V_2| = k + 1$, $|V_3| = |V_5| = k$, and $|V_4| = k + 2$.

3.5 The $(3, 4)$-problem

In this section we solve the $(3, 4)$-problem, and prove that Erdős' conjecture holds for this case. Recall that this entails showing that amongst all graphs of independence number less than four, $\overline{T_{n,3}}$, a disjoint union of three nearly-equal cliques, minimises the number of triangles.

In the first subsection we list our flag algebra results, which give the asymptotic minimum number of triangles to be $\frac{1}{9}\binom{n}{3}$. In the second subsection we use the structural information obtained to determine the value of $f(n, 3, 4)$ exactly. We also analyze the structure of extremal graphs, and show they must contain $\overline{T_{n,3}}$.

3.5.1 Getting the asymptotic result and densities

We begin by presenting the 29 admissible - that is, $\overline{K_4}$-free - graphs of size 5, followed by the three types and associated flags used in the proof.

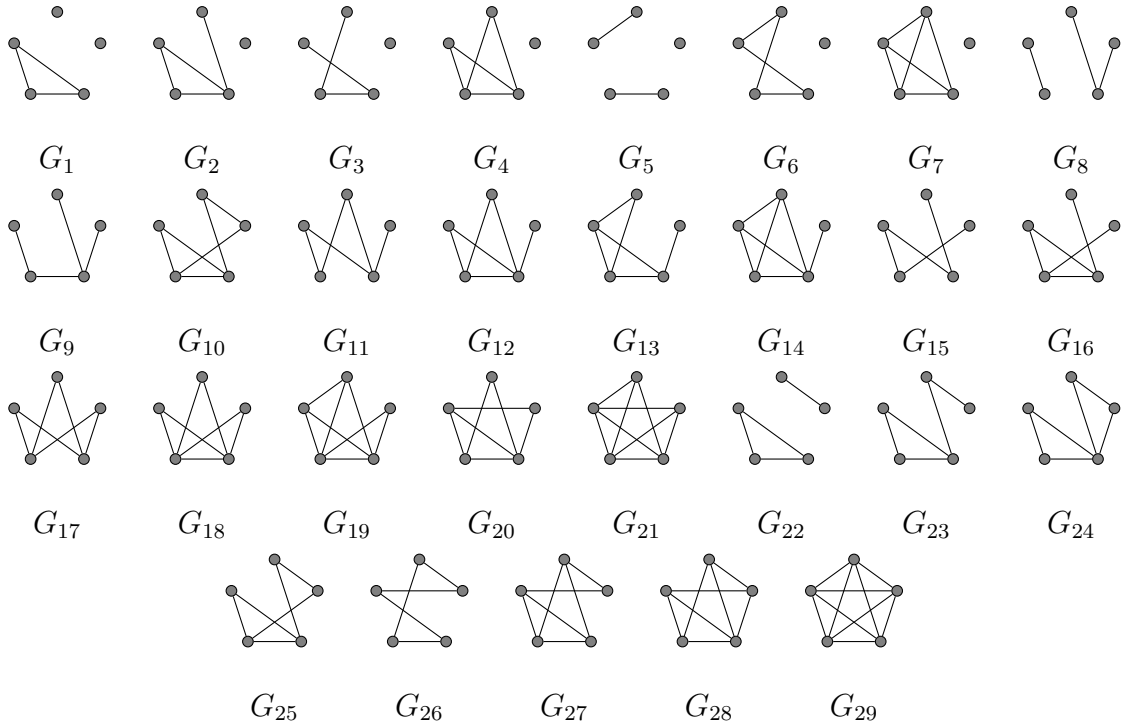


Figure 3.8: Graphs of size 5 with independence number at most 3.

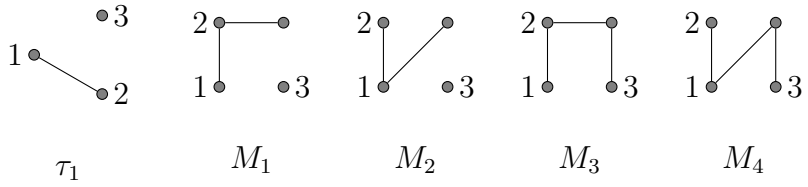


Figure 3.9: Type τ_1 and its flags of size 4.

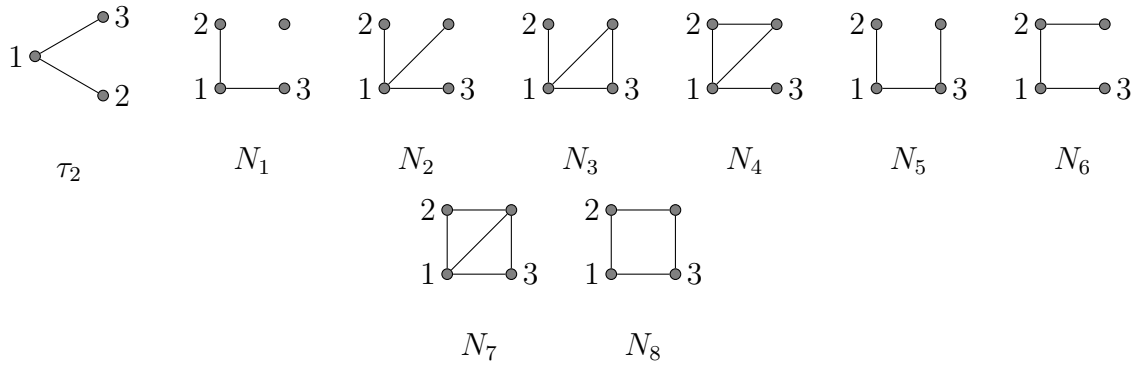


Figure 3.10: Type τ_2 and its flags of size 4.

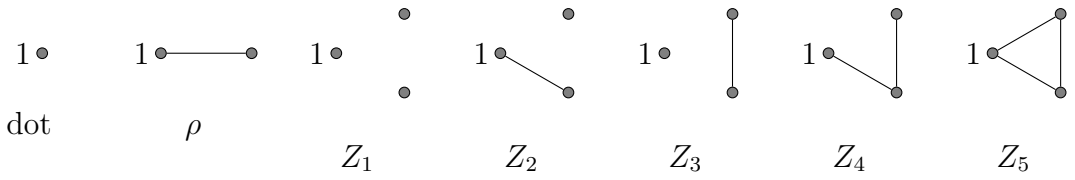


Figure 3.11: Type *dot* and its flags.

In the subsequent lemmas, for each type used in the proof, we express the corresponding positive semi-definite matrices as squares of flags, and give their expansions into graphs of size 5. The coefficients were obtained through the use of a computer program, but can easily be verified by hand, just as in the previous section.

Lemma 3.5.1. For the type τ_1, we have

$$\begin{aligned}\Delta_1 &= [[(M_1 - M_2)^2]]_{\tau_1}\\ &= \frac{1}{30}\big(G_2 - G_3 - 4G_6\big),\end{aligned}$$

$$\begin{aligned}\Delta_2 &= [[(3M_1 - 3M_2 - 10M_3 + 10M_4)^2]]_{\tau_1}\\ &= \frac{1}{30}\big(9G_2 - 9G_3 - 36G_6 - 60G_9 + 160G_{11} + 100G_{13} + 60G_{15} - 60G_{16}\\ &\quad - 100G_{25} - 500G_{26}\big).\end{aligned}$$

Lemma 3.5.2. For the type τ_2, we have

$$\begin{aligned}\Delta_3 =&[[(-3N_1 - N_2 + 3N_3 + 3N_4)^2]]_{\tau_2}\\ =&\frac{1}{30}\big(-18G_2 + 9G_8 + 3G_9 - 11G_{10} + 3G_{12} + 27G_{14} - 18G_{16} + 9G_{20} + 36G_{24}\big),\end{aligned}$$

$$\begin{aligned}\Delta_4 =&[[(-20N_1 - 20N_2 + 11N_3 + 11N_4 + 9N_5 + 9N_6)^2]]_{\tau_1}\\ =&\frac{1}{30}\big(-440G_2 - 360G_3 + 400G_8 + 121G_9 - 480G_{10} - 360G_{11} - 319G_{12} + 198G_{13}\\ &+ 363G_{14} - 279G_{15} - 242G_{16} + 121G_{20} + 279G_{23} + 484G_{24} + 198G_{25} + 405G_{26}\big),\end{aligned}$$

$$\begin{aligned}\Delta_5 =&[[(-19N_1-15N_2+15N_3+15N_4+4N_5+4N_6+15N_7)^2]]_{\tau_1}\\ =&\frac{1}{30}\big(-570G_2-152G_3-570G_4+361G_8+181G_9-675G_{10}-120G_{11}-735G_{12}\\ &+240G_{13}+675G_{14}-136G_{15}-450G_{16}-1350G_{18}+900G_{19}+795G_{20}+675G_{21}\\ &+136G_{23}+900G_{24}+120G_{25}+80G_{26}+450G_{28}\big),\end{aligned}$$

$$\begin{aligned}\Delta_6 =&[[(-6N_1-14N_2-2N_3-2N_4+8N_5+8N_6-5N_7+10N_8)^2]]_{\tau_1}\\ =&\frac{1}{30}\big(+24G_2-96G_3+60G_4-240G_6+36G_8-76G_9+308G_{10}-264G_{11}+160G_{12}\\ &-112G_{13}+12G_{14}-32G_{15}-8G_{16}-540G_{17}+420G_{18}+40G_{19}-56G_{20}+75G_{21}\\ &+32G_{23}+16G_{24}+88G_{25}+320G_{26}-80G_{27}-150G_{28}\big).\end{aligned}$$

Lemma 3.5.3. For the type *dot*, we have

$$\begin{aligned}\Delta_7 =&[[(-2Z_1+Z_2)^2]]_{dot}\\ =&\frac{1}{15}\big(6G_1+2G_2+2G_3-8G_5+4G_6-10G_8-4G_9-2G_{11}-G_{15}+G_{16}+6G_{22}+2G_{23}\\ &+G_{25}+5G_{26}\big),\end{aligned}$$

$$\begin{aligned}\Delta_8 =&[[(-2Z_1-Z_2+4Z_3)^2]]_{dot}\\ =&\frac{1}{15}\big(-42G_1-2G_2+10G_3+24G_4+24G_5+36G_6+48G_7-2G_8-4G_9+2G_{11}-4G_{13}\\ &-G_{15}-7G_{16}-18G_{22}-6G_{23}+G_{25}+5G_{26}\big),\end{aligned}$$

$$\begin{aligned}\Delta_9 =&[[(7Z_1-4Z_2+Z_3+3Z_4)^2]]_{dot}\\ =&\frac{1}{15}\big(138G_1+61G_2+43G_3-39G_4-141G_5-45G_6+3G_7-146G_8-19G_9+42G_{10}\\ &-52G_{11}+21G_{12}-22G_{13}+9G_{14}-25G_{15}+65G_{16}+54G_{17}+54G_{18}+18G_{19}-6G_{20}\\ &+72G_{22}-21G_{23}-96G_{24}+19G_{25}+125G_{26}+18G_{27}\big),\end{aligned}$$

$$\begin{aligned}\Delta_{10} &= [[(8Z_1 - 2Z_2 - 9Z_3 + 10Z_5)^2]]_{dot} \\ &= \frac{1}{15}\big(- 168G_1 + 103G_2 + 85G_3 + 170G_4 + 153G_5 + 226G_6 + 3G_7 - 16G_8 + 4G_9 \\ &+ 160G_{10} - 16G_{11} + 120G_{12} - 132G_{13} - 120G_{14} + 16G_{15} + 80G_{16} + 240G_{18} + 70G_{19} \\ &- 120G_{20} + 600G_{21} - 138G_{22} - 136G_{23} - 260G_{24} - 86G_{25} + 20G_{26} + 200G_{28} + 1500G_{29}\big),\end{aligned}$$

$$\begin{aligned}\Delta_{11} &= \left[\left[\left(\rho - \frac{1}{3}\right)^2\right]\right]_{dot} \\ &= \frac{1}{90}\big(G_1 + G_2 - 2G_3 + 4G_4 - 2G_5 - 2G_6 + 10G_7 - 5G_8 - 2G_9 + 4G_{10} - 2G_{11} + 7G_{12} \\ &+ 4G_{13} + 13G_{14} - 5G_{15} + G_{16} + G_{17} + 13G_{18} + 19G_{19} + 10G_{20} + 28G_{21} - 5G_{22} \\ &- 2G_{23} + 4G_{24} + G_{25} - 5G_{26} + 7G_{27} + 16G_{28} + 40G_{29}\big).\end{aligned}$$

We can now combine these lemmas to obtain an asymptotic lower bound on the density of triangles, K_3, in any $\overline{K_4}$-free graph.

Theorem 3.5.4. We have

$$K_3 - \sum_{i=1}^{11} c_i \Delta_i \geq \frac{1}{9} \sum_{j=1}^{29} G_j = \frac{1}{9},$$

where

$$\begin{aligned}\mathbf{c} &= (c_i)_{i=1}^{11} \\ &= \frac{1}{2^5 \cdot 3 \cdot 1009}\left(263984, 4720, 4432, \frac{412192}{371}, \frac{72789}{112}, \frac{4655105}{3392}, 1185, 8437, 3440, 856, 1128\right).\end{aligned}$$

Proof. We begin by expanding K_3 into graphs of size 5. A straightforward calculation gives

$$\begin{aligned}K_3 = \frac{1}{10}\Big(&G_1 + G_2 + 2G_4 + 4G_7 + G_{10} + 2G_{12} + 2G_{13} + 4G_{14} + G_{16} + 3G_{18} + 5G_{19} \\ &+ 3G_{20} + 7G_{21} + G_{22} + G_{23} + 2G_{24} + G_{25} + 2G_{27} + 4G_{28} + 10G_{29}\Big).\end{aligned}$$

We now use the lemmas to expand the squares Δ_i into the graphs G_j. After summing the coefficients in the linear combination, it can easily be verified that they are all at least $\frac{1}{9}$. Since the densities must sum to 1, we have $\sum_{j=1}^{29} G_j = 1$, which gives the final equality. □

Corollary 3.5.5. Any n-vertex graph G with $\alpha(G) \leq 3$ satisfies

$$\frac{t_3(G)}{\binom{n}{3}} - \frac{47}{4036n} \sum_v \left(\frac{d(v)}{n-1} - \frac{1}{3} \right)^2 \geq \frac{1}{9} - o_{n\to\infty}(1).$$

Proof. Since the Δ_i are squares of flags, they are asymptotically non-negative. Hence discarding the terms for Δ_i, $1 \leq i \leq 10$, maintains the inequality. This gives $K_3 - \frac{47}{4036}\left[\left[\left(\rho - \frac{1}{3}\right)^2\right]\right]_{dot} \geq \frac{1}{9} - o_{n\to\infty}(1)$. Interpreting these terms combinatorially gives the corollary. □

3.5.2 The stability analysis

In order to derive a stability result for the $(3,4)$-problem, we use the following well-known result of Andrásfai, Erdős and Sós [AES74].

Theorem 3.5.6. *(Andrásfai, Erdös, Sós)* A K_r-free graph on n vertices that has minimum degree larger than $\frac{3r-7}{3r-4}n$ must be $(r-1)$-partite.

Applying this to the complement of a graph with $r = 4$, we find that a graph G on n vertices with $\alpha(G) \leq 3$ and maximum degree less than $\frac{3}{8}n$ must be spanned by three cliques. The following stability result follows.

Proposition 3.5.7. Suppose $0 < \varepsilon < \frac{1}{30}$. There exists $n_0 = n_0(\varepsilon)$ such that any graph G on $n \geq n_0$ vertices with $\alpha(G) \leq 3$ and $t_3(G) < \left(\frac{1}{9} + \varepsilon^5\right)\binom{n}{3}$ contains an induced subgraph $G' \subset G$ on at least $(1 - 100\varepsilon^3)\, n$ vertices that is spanned by three cliques of size between $\left(\frac{1}{3} - 3\varepsilon\right) n$ and $\left(\frac{1}{3} + \varepsilon\right) n$. Moreover, every vertex in G' sends at most $4\varepsilon n$ edges outside its clique.

Proof. We have from Corollary 3.5.5 that for any graph G on n vertices with $\alpha(G) \leq 3$,

$$\frac{t_3(G)}{\binom{n}{3}} - \frac{47}{4036n} \sum_v \left(\frac{d(v)}{n-1} - \frac{1}{3} \right)^2 \geq \frac{1}{9} - o_{n\to\infty}(1).$$

In particular, if $t_3(G) < \left(\frac{1}{9} + \varepsilon^5\right)\binom{n}{3}$, and n is large enough, then

$$\sum_v \left(\frac{d(v)}{n-1} - \frac{1}{3} \right)^2 < 100\varepsilon^5 n.$$

Let $B = \left\{v : d(v) \geq \left(\frac{1}{3} + \varepsilon\right) n\right\}$. Then $|B|\varepsilon^2 < \sum_v \left(\frac{d(v)}{n-1} - \frac{1}{3}\right)^2 < 100\varepsilon^5 n$, and so $|B| < 100\varepsilon^3 n$.

Let G' be the induced subgraph on $V(G) \setminus B$. As claimed, G' has $n' \geq (1 - 100\varepsilon^3)n$ vertices. Moreover, since $\varepsilon < \frac{1}{30}$ the maximum degree $\Delta(G')$ is bounded by

$$\Delta(G') < \left(\frac{1}{3} + \varepsilon\right) n \leq \frac{\frac{1}{3} + \varepsilon}{1 - 100\varepsilon^3} n' < \frac{3}{8} n'.$$

Hence we can apply Theorem 3.5.6 in its complementary form to deduce that G' is spanned by three cliques.

Since $\Delta(G') < \left(\frac{1}{3} + \varepsilon\right) n$, we deduce that the largest clique in G' has size at most $\left(\frac{1}{3} + \varepsilon\right) n$. This implies that the smallest clique has size at least $(1 - 100\varepsilon^3)\, n - 2\left(\frac{1}{3} + \varepsilon\right) n > \left(\frac{1}{3} - 3\varepsilon\right) n$ (using the bound $\varepsilon < \frac{1}{30}$). This implies that every vertex in G' can send at most $\left(\frac{1}{3} + \varepsilon\right) n - \left(\frac{1}{3} - 3\varepsilon\right) n = 4\varepsilon n$ edges outside its own clique.

Finally, consider the vertices in B. If any vertex $v \in B$ is adjacent to all vertices in one of the cliques C_i, and does not have more than $4\varepsilon n$ edges outside C_i, then we can add v to C_i without affecting any of the previous bounds. Thus the only vertices left in B are either those adjacent to one clique, but with too many neighbors outside the clique, or those with a non-neighbor in each of the three cliques. □

This stability result allows us to, for large values of n, deduce the exact value of the $(3, 4)$-problem, and also to characterise all extremal graphs. Recall that we define $f(n, k, l)$ to be the minimum of $t_k(G)$ over all graphs G on n vertices with $\alpha(G) \leq l - 1$.

Theorem 3.5.8. There exists n_0 such that for every $n \geq n_0$, $f(n, 3, 4) = \binom{\lfloor n/3 \rfloor}{3} + \binom{\lfloor (n+1)/3 \rfloor}{3} + \binom{\lfloor (n+2)/3 \rfloor}{3}$. Moreover, if G is a graph on $n \geq n_0$ vertices with $t_3(G) = f(n, 3, 4)$, then G contains $\overline{T_{n,3}}$, a disjoint union of three nearly-equal cliques.

Proof. First note that $G = \overline{T_{n,3}}$ has $\alpha(G) \leq 3$ and so we have the upper bound $f(n, 3, 4) \leq t_3(\overline{T_{n,3}}) = \binom{\lfloor n/3 \rfloor}{3} + \binom{\lfloor (n+1)/3 \rfloor}{3} + \binom{\lfloor (n+2)/3 \rfloor}{3} \sim \frac{1}{9}\binom{n}{3}$ - note that this upper bound holds for all n.

To obtain a matching lower bound, we apply the stability result from Proposition 3.5.7. Take $\varepsilon = \frac{1}{100}$, and let $n \geq n_0(\varepsilon)$ be sufficiently large. Suppose G is an extremal graph on $n \geq n_0$ vertices. In particular, we have $t_3(G) < \left(\frac{1}{9} + \varepsilon^5\right)\binom{n}{3}$ for n large enough. From the proof of the proposition, we know that there is a set B of at most $100\varepsilon^3 n$ 'bad' vertices, and the remainder of the vertices are in three cliques, with at most $4\varepsilon n$ edges to the other cliques. Label the cliques in order of size, say $|C_1| \geq |C_2| \geq |C_3|$. We will show that an extremal graph cannot have any bad vertices, so G is spanned by the three cliques. We begin with a simple observation.

Claim: Every vertex $v \in V(G)$ is in at most $\binom{|C_3|+|B|}{2}$ triangles.

Proof: Suppose some vertex v were in more triangles. Delete v, and add a new vertex v' with $N(v') = C_3 \cup B$. This does not increase the independence number, and v' is in at most $\binom{|C_3|+|B|}{2}$ triangles. Hence we have decreased the number of triangles in G, which contradicts the minimality of G. Note that $\binom{|C_3|+|B|}{2} \leq \binom{|C_3|}{2} + |B|n \leq \binom{|C_3|}{2} + 100\varepsilon^3 n^2 = \binom{|C_3|}{2} + \varepsilon^2 n^2$.

Now consider a potential bad vertex $v \in B$. There are two reasons v could be bad:

Case 1: v is adjacent to all vertices of one of the cliques C_i, but has more than $4\varepsilon n$ neighbors in the other cliques.

If v has more than $4\varepsilon n$ neighbors in the other cliques, it must have at least $2\varepsilon n$ neighbors in one of them. Note that every pair of these neighbors creates a triangle with v. Thus v is in at least $\binom{|C_i|}{2} + \binom{2\varepsilon n}{2} > \binom{|C_3|}{2} + \varepsilon^2 n^2$ triangles, which contradicts our earlier claim. Hence this case cannot occur.

Case 2: v has a non-neighbor in each of the three cliques.

Let $\overline{d}_i = |C_i \setminus N(v)|$ be the number of non-neighbors of v in the ith clique. Consider the cliques in increasing order of these values, that is, suppose $\overline{d}_{i_1} \leq \overline{d}_{i_2} \leq \overline{d}_{i_3}$. Let x be a non-neighbor of v in C_{i_1}.

Case 2a: Every vertex $y \in C_{i_2}$ is adjacent to one of $\{v, x\}$.

Since x has at most $4\varepsilon n$ neighbors in C_{i_2}, it follows that $\overline{d}_{i_1} \leq \overline{d}_{i_2} \leq 4\varepsilon n$. Counting only the neighbors of v in the cliques C_{i_1} and C_{i_2}, we see that v is in at least $\binom{|C_{i_1}|-4\varepsilon n}{2} +$

$\binom{|C_{i_2}|-4\varepsilon n}{2} \geq 2\binom{|C_3|-4\varepsilon n}{2}$ triangles. We have $2\binom{|C_3|-4\varepsilon n}{2} \approx |C_3|^2 - 8\varepsilon|C_3|n + 16\varepsilon^2 n^2$. Since $|C_3| \geq \left(\frac{1}{3} - 3\varepsilon\right) n$ and $\varepsilon = \frac{1}{100}$, this is greater than $\binom{|C_3|}{2} + \varepsilon^2 n^2 < \frac{1}{2}|C_3|^2 + \varepsilon^2 n^2$, which contradicts the earlier claim.

Case 2b: v and x have a common non-neighbor in C_{i_2}, say y.

In this case, as $\alpha(G) \leq 3$, every vertex in C_{i_3} must be adjacent to one of $\{v, x, y\}$. Since x and y have at most $4\varepsilon n$ neighbors in C_{i_3}, it follows that $\overline{d}_{i_1} \leq \overline{d}_{i_2} \leq \overline{d}_{i_3} \leq 8\varepsilon n$. Thus v is in at least $\binom{|C_{i_1}|-8\varepsilon n}{2} + \binom{|C_{i_2}|-8\varepsilon n}{2} + \binom{|C_{i_3}|-8\varepsilon n}{2} \geq 3\binom{|C_3|-8\varepsilon n}{2} \approx \frac{3}{2}|C_3|^2 - 24\varepsilon|C_3|n + 96\varepsilon^2 n^2$ triangles. Again, given our bounds on $|C_3|$ and ε, this is greater than $\binom{|C_3|}{2} + \varepsilon^2 n^2$, which gives a contradiction.

Thus we have shown that in an extremal graph, there are no bad vertices, and so the three cliques span all n vertices and $|B| = 0$. Now note that any vertex in C_1 is in $\binom{|C_1|-1}{2}$ triangles from within C_1 alone. By the earlier claim, we must have $\binom{|C_1|-1}{2} \leq \binom{|C_3|+|B|}{2} = \binom{|C_3|}{2}$, from which it follows that $|C_1| - 1 \leq |C_3|$. Thus $|C_3| \leq |C_2| \leq |C_1| \leq |C_3| + 1$, which shows that the cliques must be nearly equal in size.

This implies that $\overline{T_{n,3}} \subset G$, and so it follows that for any graph G on n vertices with $\alpha(G) \leq 3$, we must have $t_3(G) \geq t_3(\overline{T_{n,3}})$. Thus $f(n, 3, 4) = t_3(\overline{T_{n,3}})$. Moreover, if G is an extremal graph, then since we have equality, there can be no triangles with vertices from different cliques. This means that each vertex can have at most one neighbor in each of the two other cliques; in other words, the bipartite graphs between cliques are (partial) matchings. These matchings must be such that there is no triangle with one vertex from each clique. However, the extremal graph is not unique, as there are many possibilities for the matchings.

□

3.6 Further remarks

In this chapter, we applied the techniques of flag algebras, combined with stability arguments, to solve the Erdős problem for the cases $(k, l) = (4, 3)$ and $(3, 4)$. In particular, we showed

that Nikiforov's construction of a blow-up of C_5 is optimal for the $(4,3)$-problem, while Erdős' conjecture still holds for the $(3,4)$-problem.

We have also run the SDP problem for larger cases, and our calculations suggests that Erdős' conjecture remains valid for the $(3,5)$- and $(3,6)$-problems. Moreover, it would appear that a blow-up of C_5 is also optimal for the $(5,3)$-problem. These results, and more, were obtained by Pikhurko and Vaughan [PV13], who were independently studying the problem.

Note that the extremal graphs we have found are all blow-ups of small graphs. In particular, the graphs are Ramsey graphs. The construction of $l-1$ cliques is a blow-up of an independent set of size $l-1$, which is the $R(2,l)$ Ramsey graph. On the other hand, C_5 is the $R(3,3)$ Ramsey graph. One may therefore ask if, for large n, the solution of the (k,l)-problem is always a blow-up of an $R(s,t)$ Ramsey graph, where s and t depend only on k and l. Solving this problem in general appears to be quite difficult.

A simpler question, first asked by Nikiforov, is to determine the extremal graphs for the (k,l)-problem as one parameter is fixed and the other grows. In particular, it remains to determine for which values of l a disjoint union of $l-1$ cliques remains optimal for the $(3,l)$-problem. In light of the above results, one could also study for which values of k the blow-up of C_5 is optimal for the $(k,3)$-problem. Proofs by flag algebras are infeasible for large values of k and l, as the search space and running time grow exponentially in these parameters. It would be of great interest to develop new techniques to attack this problem.

3.7 Implementation of flag algebras

In Section 3.3, we covered the basics of the theory behind flag algebras; here we discuss the actual implementation of the method. In particular, we will discuss how to set up the SDP problem, and then find a verifiable proof. The main steps are:

1. Identifying the types σ_i to use, and finding a suitable size t for the expansion of the positive semi-definite matrices.

2. Finding a verifiable (e.g. rational) solution that leads to a proof.

3. *(Optional)* Writing the positive semi-definite matrix as a sum of squares.

We shall address each of these steps in turn.

Identifying types:

The process of identifying the necessary types σ_i and finding a suitable size t essentially comes down to trial-and-error. Note that whatever choice of types and size we make will result in an SDP problem as outlined above, which can then be solved to provide *some* bound for the extremal problem. In order to determine whether or not this is the *right* bound, we need a conjecture on what the bound should be - this typically comes from a construction. We then seek to keep improving the flag algebra results until they match the conjectured bound.

To produce the flag algebra results, we start with the initial size t to be the size of the subgraph J, the density of which we are trying to bound. Given t, we produce a list of all admissible graphs G of size t. We then consider all possible types of size suitable for expansion into graphs of size t. Recall that if we have a type of size k, and use flags of size $l \geq k+1$, then to compute a product of two flags, we must expand into graphs of size at least $2l-k \geq k+2$. This restricts the size of types and flags we can use - our types can be of size at most $t-2$, and given a type of size k, we choose the largest possible size of flags l that satisfies $2l-k \leq t$.

For each of our types σ_i, with its associated list of flags $\mathcal{F}_{l_i}^{\sigma_i}$, we compute the product of each pair of flags, which gives the corresponding block in the SDP problem. This provides the formulation of the SDP problem, which can then be solved numerically.

If the numerical bound is less than the conjecture, then we do not have enough types to solve the problem. Thus we increase the size t, which allows the use of larger types, and repeat the process. If the numerical bound matches the conjecture, we then have enough types to solve the problem, and can proceed to finding a verifiable proof.

At this stage, we have the block variable matrices Q_i for the SDP problem. However, as they were computed numerically, they are subject to rounding error, and thus we cannot

be certain that they are truly positive semi-definite matrices, nor that the bound for the extremal problem they provide is exactly equal to the conjectured bound. To have a rigorous proof, it is necessary to find solution matrices Q_i whose entries are known exactly - they will ideally be rational. It can then be independently verified that these matrices satisfy the conditions necessary to prove the desired result. We now outline some of the steps that can be taken to find such a solution.

Finding a verifiable solution:

Typically, the space of solutions will be a high-dimensional space, with many degrees of freedom for the entries of the matrices Q_i. To try to force the solution towards rational entries, we seek to reduce the dimension of the search space. There are three methods we can apply: reducing the size of the block variables, identifying natural eigenvectors, and changing the basis to introduce zero-entries.

Recall that for each type σ_i we have the associated block variable Q_i. In identifying which types to use, we added all possible types until we obtained the right bound. However, it is possible, and even likely, that some of the types are unnecessary. Given a type σ, we remove it from the SDP problem, and run the SDP solver again. If we still obtain the correct bound, then we know the type σ was unnecessary. If instead this results in a worse bound, then we keep σ, and try removing a different type. In this way we arrive at a minimal set of necessary types, thus reducing the number of block variables in the SDP problem.

Given a set of minimal types, there is a further reduction possible. Every type σ has the natural group Γ_σ of automorphisms of the underlying graph σ_0. The group Γ_σ acts on the algebra $\mathcal{A}^\sigma$ by relabeling the flags according to the automorphism. We can then decompose $\mathcal{A}^\sigma = \mathcal{A}^\sigma_+ \oplus \mathcal{A}^\sigma_-$ into a positive and negative part, where $\mathcal{A}^\sigma_+$ consists of all elements invariant under Γ_σ, while $\mathcal{A}^\sigma_- \stackrel{def}{=} \left\{ f \in \mathcal{A}^\sigma : \sum_{\gamma\in\Gamma_\sigma} \gamma f = 0 \right\}$. For example, given the type and flags of Figure 3, both labelings of the vertices of σ give rise to automorphisms, and so Γ_σ is the symmetric group on two elements. One can verify that $F_3 \in \mathcal{A}^\sigma_+$, $F_1 + F_2 \in \mathcal{A}^\sigma_+$, and $F_1 - F_2 \in \mathcal{A}^\sigma_-$.

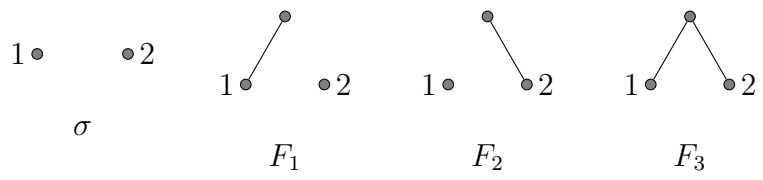


Figure 3.12: Decomposition into positive and negative parts.

This decomposition is useful because whenever we have $f \in \mathcal{A}^\sigma_+$ and $g \in \mathcal{A}^\sigma_-$, we have $[[f \cdot g]]_\sigma = 0$. Hence given the semi-definite matrix Q for the type σ, we can split it into its

'invariant' part Q^+ and 'anti-invariant' part Q^-. While this increases the number of block variables, they are now of smaller size, and hence have fewer degrees of freedom, reducing the dimension of the search space. Moreover, it may be that not all of these parts are necessary, so we can proceed as before to remove any unnecessary block variables.

The second technique we use is that of identifying natural eigenvectors. For this, we require an extremal construction that attains the conjectured bound; let G_n represent an extremal graph on n vertices, and let $\{G_n\}_{n\in\mathbb{N}}$. Given a type σ, fix a position of σ in G_n. This turns G_n into a σ-flag F_n. The family $\{F_n\}_{n\in\mathbb{N}}$ represents a way to consistently label the type σ in G_n.

Recall that in the flag algebra calculations, we used the bound $[[Q]]_\sigma(G_n) \geq o_{n\to\infty}(1)$. If G_n is an extremal graph, then the bounds are tight, and so $[[Q]]_\sigma(G_n) = o_{n\to\infty}(1)$. Hence we must have $p_\sigma(Q\{\mathcal{F}_l^\sigma\}; F_n) = \sum_{F_1,F_2\in\mathcal{F}_l^\sigma} Q_{F_1,F_2} p_\sigma(F_1; F_n)p_\sigma(F_2; F_n) + O(1/n) = o_{n\to\infty}(1)$. Taking the limit as $n \to \infty$, this implies that if we have a vector v defined by $v_F = \lim_{n\to\infty} p_\sigma(F; F_n)$ for $F \in \mathcal{F}_l^\sigma$, then v_F must be a zero-eigenvector of Q. Repeating this for different embeddings of the type σ in the extremal family of graphs $\{G_n\}$ can give rise to several eigenvectors. This procedure is formally defined using the apparatus of *ensembles of random homomorphisms* in Section 3.2 of [Raz07].

Having fixed this eigenvectors, we can then reduce the size of the block variables. Note that if we are able to remove all zero-eigenvectors this way, then we are left with positive definite matrices as our block variables. This leaves a little room for error, so we can replace the entries with simple rational entries and hope to still have a positive semi-definite matrix.

Our final method for reducing the dimension of the search space is to change the basis to introduce zero entries. Ideally the new set of variables will be a rational linear combination of the previous set, which will lead to a solution with rational entries. Moreover, we introduce zeros in such a way as to split the block variables into smaller blocks. More formally, consider the general SDP problem of the following form:

maximise $\mathrm{tr}(CX)$, subject to

- $\mathrm{tr}(A_i X) = a_i$ for $i = 1, 2, \ldots, m$
- $X \succeq 0$ (that is, X is positive semi-definite)

where X and A_i are symmetric $n \times n$ matrices for $i = 1, 2, \ldots, m$.

Suppose we had a rational $n \times n$ matrix M such that all entries of the first row (and hence column, by symmetry) of MXM^T, except possibly the first, were zero. We can then change variables to modify the SDP problem into an equivalent one, as below:

maximise $\mathrm{tr}(\tilde{C}Y)$, subject to

- $\mathrm{tr}(\tilde{A}_i Y) = a_i$ for $i = 1, 2, \ldots, m$
- $Y \succeq 0$

where $\tilde{C} = (M^{-1})^T C M^{-1}$ and $\tilde{A}_i = (M^{-1})^T A_i M^{-1}$ for $i = 1, 2, \ldots, m$.

The solutions of both problems are related by the equation $Y = MXM^T$. We can now reduce the dimension of the solution space by forcing all the non-principle entries of the first row/column of $\tilde{C}$ and $\tilde{A}_i$ to be zero for $i = 1, 2, \ldots, m$. This is possible because we already have the existence of a solution Y with $Y_{1,j} = Y_{j,1} = 0$ for $j = 2, 3, \ldots, n$, and hence this restricted solution space contains a solution to the original problem. This operation splits the block variable Y into a one-dimensional block and an $(n-1)$-dimensional block. We can now iterate the procedure.

We find such a matrix M by inspecting the numerical solution to the original SDP problem, and using a rational approximation to an eigenvector v for the first row. We then fill in the remaining rows with independent vectors orthogonal to v. Note that if the solution is initially positive definite, there is a little room for error, so we may hope to choose a simple rational approximation without worsening the solution to the SDP problem.

Expressing the solution as a sum of squares:

If we are able to repeatedly iterate the change of basis procedure outlined above, then we will eventually reach a problem whose solution is a diagonal matrix. This is advantageous for two reasons. First, the semi-definite programming problem reduces to a linear programming (LP) problem. This can be solved by only taking rational linear combinations of the entries of the variables at every step, and so the solution will be a rational combinations of the input to the LP problem. Hence the solution can be specified exactly, resulting in a verifiable proof. Second, we can write the positive semi-definite matrix as a sum of squares, which is easier to understand. This can lead to combinatorial interpretations of the proof, as we demonstrated in Section 3.4.1. Thus while this step is not necessary for solving problems with the machinery of flag algebras, it makes the resulting proofs much more understandable.

3.8 Integer optimisation problem

In this section, we prove Lemma 3.4.8 from Section 3.4.2, in which we solve the integer optimisation problem required to determine the size of the parts in the blow-up of C_5 that minimises the number of 4-cliques.

Lemma 3.8.1. Let $\varepsilon > 0$ be sufficiently small, and n sufficiently large. Consider the function

$$g(y_1, y_2, y_3, y_4, y_5) = \sum_{i=1}^{5} \binom{y_i}{5} - \sum_{i=1}^{5} \binom{n - y_i - y_{i+1}}{4}.$$

Subject to the constraints that the y_i be integers satisfying $\sum_{i=1}^{5} y_i = 2n$ and $\left|y_i - \frac{2}{5}n\right| < \varepsilon n$, g is uniquely (up to cyclic permutation of the variables) minimised when the y_i take values $\left\lfloor \frac{2n}{5} \right\rfloor$ and $\left\lceil \frac{2n}{5} \right\rceil$ in ascending order.

Proof. First we will show that if $(y_1, y_2, y_3, y_4, y_5)$ is optimal, the y_i should be as equal as possible. Suppose towards contradiction that this was not the case. Then there are i, j with $y_i - y_j \geq 2$; let i, j be such that this difference is maximal over all such pairs. There are two cases:

Case 1: i and j are consecutive.

Without loss of generality, suppose $i = 2$ and $j = 3$, so we have $y_2 - y_3 \geq 2$, with this difference being maximal. We will show that $g(y_1, y_2 - 1, y_3 + 1, y_4, y_5) < g(y_1, y_2, y_3, y_4, y_5)$, which contradicts our assumption of optimality. Indeed, we have

$$\begin{aligned}
\Delta g &= g(y_1, y_2 - 1, y_3 + 1, y_4, y_5) - g(y_1, y_2, y_3, y_4, y_5) \\
&= \binom{y_2 - 1}{4} + \binom{y_3 + 1}{4} - \binom{n - y_1 - y_2 + 1}{4} + \binom{n - y_3 - y_4 - 1}{4} \\
&\quad - \left[\binom{y_2}{4} + \binom{y_3}{4} - \binom{n - y_1 - y_2}{4} - \binom{n - y_3 - y_4}{4} \right] \\
&= \binom{y_3}{3} - \binom{y_2 - 1}{3} + \binom{n - y_3 - y_4 - 1}{3} - \binom{n - y_1 - y_2}{3}.
\end{aligned}$$

Now let $s = y_2 - y_3 - 1 \geq 1$, and let $t = (n - y_3 - y_4 - 1) - (n - y_1 - y_2) = y_1 - y_4 + y_2 - y_3 - 1 = y_1 - y_4 + s$. If $t \leq 0$, then clearly the above expression is negative, which shows $(y_1, y_2, y_3, y_4, y_5)$ is not optimal. Hence we must have $t \geq 1$. In this case, we can rewrite the above as

$$\Delta g = \left[\binom{t}{3} + \binom{t}{2}(n - y_1 - y_2) + t\binom{n - y_1 - y_2}{2} \right] - \left[\binom{s}{3} + \binom{s}{2} y_3 + s\binom{y_3}{2} \right].$$

From our constraints on the variables y_i, we have that $y_3 = \left(\frac{2}{5} + O(\varepsilon)\right) n$, $n - y_1 - y_2 = \left(\frac{1}{5} + O(\varepsilon)\right) n$, $s \leq 2\varepsilon n$ and $t \leq 4\varepsilon n$. These bounds imply that the main terms are those linear in s and t. We have

$$\Delta g = \frac{1}{50} \left[(1 + O(\varepsilon))\, t - (4 + O(\varepsilon))\, s \right] n^2 + O((s^2 + t^2)n).$$

In particular, for large n, this can only be non-negative if $t \geq (4 - O(\varepsilon))\, s$. However, we have $t = y_1 - y_4 + s$, and by our assumption of maximality of $y_2 - y_3$, we have $y_1 - y_4 \leq y_2 - y_3 = s + 1$. Hence $t \leq 2s + 1$, and we have a contradiction.

Case 2: i and j are not consecutive.

Without loss of generality, suppose $i = 2$ and $j = 4$, with $y_2 - y_4 \geq 2$ being the maximal difference. Let

$$\Delta g = g(y_1, y_2 - 1, y_3, y_4 + 1, y_5) - g(y_1, y_2, y_3, y_4, y_5).$$

By similar calculations to those in Case 1, we have

$$\begin{aligned}\Delta g =& \binom{y_4}{3} - \binom{y_2-1}{3} + \binom{n-y_5-y_4-1}{3} + \binom{n-y_4-y_3-1}{3} \\ & - \binom{n-y_2-y_3}{3} - \binom{n-y_1-y_2}{3}.\end{aligned}$$

We define $s = y_2 - y_4 - 1$, and $t = (n - y_5 - y_4 - 1) - (n - y_1 - y_2) = y_1 - y_5 + s$. If $t \leq 0$, then $\Delta g < 0$, which contradicts the optimality of $(y_1, y_2, y_3, y_4, y_5)$. Hence we may assume $t \geq 1$, and rewrite Δg in terms of s and t as before. In this case we find

$$\Delta g = \frac{1}{50}\left[(1 + O(\varepsilon))\, t - (3 + O(\varepsilon))\, s\right] n^2 + O((s^2 + t^2)n).$$

Hence for $\Delta g \geq 0$, we must have $t \geq (3 - O(\varepsilon))\, s$. However, by maximality of $y_2 - y_4$, we have $t = y_1 - y_5 + s \leq 2s + 1$. The only way these equations can be satisfied is if $s = 1$ and $y_1 - y_5 = 2$. But in this case y_1 and y_5 are two consecutive variables with a maximal difference, and so we reduce to Case 1, which leads to a contradiction.

Hence we have shown that subject to the above conditions, g is only minimised when the variables y_i take values $\lfloor \frac{2n}{5} \rfloor$ or $\lceil \frac{2n}{5} \rceil$. If $n \equiv 0, 1, 4 \pmod 5$, there is only one way (up to cyclic rotation) that these values can be distributed, so the minimum is uniquely determined. If $n \equiv 2, 3 \pmod 5$, then there are two possible distributions of the values. In each case, an easy calculation shows g is minimised when the values are in decreasing order. This completes the proof of the lemma. □

Note that we assume $|y_i - \frac{2}{5}n| < \varepsilon n$ only to simplify the proof. Even without this condition, we can prove that for any $n \geq 12$, the above result holds. However, as the flag algebra results are asymptotic in nature, we can only determine the unique extremal graph for the $(4, 3)$-problem when n is large.

CHAPTER 4

Set families with few disjoint pairs

4.1 Introduction

A set family $\mathcal{F}$ is said to be *intersecting* if $F_1 \cap F_2 \neq \emptyset$ for all $F_1, F_2 \in \mathcal{F}$. The Erdős–Ko–Rado Theorem is a classic result in extremal set theory, determining how large an intersecting k-uniform set family can be. This gives rise to the natural question of how many disjoint pairs must appear in larger set families.

We consider this problem, first asked by Ahslwede in 1980. Given a k-uniform set family $\mathcal{F}$ on $[n]$ with s sets, how many disjoint pairs must $\mathcal{F}$ contain? We denote the minimum by $\mathrm{dp}(n, k, s)$, and determine its value for a range of family sizes s, thus confirming a conjecture of Bollobás and Leader in these cases. This results in a quantitative strengthening of the Erdős–Ko–Rado Theorem. We also provide similar results regarding some well-known extensions of the Erdős–Ko–Rado Theorem, which in particular allow us to partially resolve a problem of Kleitman and West.

We now discuss the Erdős–Ko–Rado Theorem and the history of this problem in greater detail, before presenting our new results.

4.1.1 Intersecting families

Extremal set theory is one of the most rapidly developing areas of combinatorics, having enjoyed tremendous growth in recent years. The field is built on the study of very robust structures, which allow for numerous applications to other branches of mathematics and computer science, including discrete geometry, functional analysis, number theory and complexity.

One such structure that has attracted a great deal of attention over the years is the intersecting set family; that is, a collection $\mathcal{F}$ of subsets of $[n]$ that is pairwise-intersecting. The most fundamental question one may ask is how large such a family can be. Observe that we must have $|\mathcal{F}| \leq 2^{n-1}$, since for every set $F \subset [n]$, we can have at most one of F and $[n] \setminus F$ in $\mathcal{F}$. This bound is easily seen to be tight, and there are in fact numerous extremal families. For example, one could take the set family consisting of all sets containing some fixed $i \in [n]$. Another construction is to take all sets $F \subset [n]$ of size $|F| > \frac{n}{2}$. If n is odd, this consists of precisely 2^{n-1} sets. If n is even, then we must add an intersecting family of sets of size $\frac{n}{2}$; for instance, $\{F \subset [n] : |F| = \frac{n}{2}, 1 \in F\}$ would suffice.

In some sense, having large sets makes it easier for the family to be intersecting. This leads to the classic theorem of Erdős–Ko–Rado [EKR61], a central result in extremal set theory, which bounds the size of an intersecting set family with all sets restricted to have size k. Here we use $\binom{[n]}{k}$ to denote all subsets of $[n]$ of size k.

Theorem 4.1.1 (Erdős–Ko–Rado [EKR61], 1961)**.** *If $n \geq 2k$, and $\mathcal{F} \subset \binom{[n]}{k}$ is an intersecting set family, then $|\mathcal{F}| \leq \binom{n-1}{k-1}$.*

This is again tight, as we may take all sets containing some fixed element $i \in [n]$, a family we call a *(full) star with centre i*.

As is befitting of such an important theorem, there have been numerous extensions to many different settings, some of which are discussed in Anderson's book [And87]. We are particularly interested in two, namely t-intersecting families and q-matching-free families.

A pair of sets F_1, F_2 is said to be *t-intersecting* if $|F_1 \cap F_2| \geq t$, and *t-disjoint* otherwise. A set family $\mathcal{F}$ is *t-intersecting* if every pair of sets in the family is. When $t = 1$, we simply have an intersecting family. A natural construction of a t-intersecting family is to fix some t-set $X \in \binom{[n]}{t}$, and take all k-sets containing X; we call this a *(full) t-star with centre X*. In their original paper, Erdős–Ko–Rado showed that, provided n was sufficiently large, this was best possible.

Theorem 4.1.2 (Erdős–Ko–Rado [EKR61], 1961)**.** *If $n \geq n_0(k,t)$, and $\mathcal{F} \subset \binom{[n]}{k}$ is a t-intersecting set family, then $|\mathcal{F}| \leq \binom{n-t}{k-t}$.*

There was much work done on determining the correct value of $n_0(k,t)$, and how large t-intersecting families can be when n is small. This problem was completely resolved by the celebrated Complete Intersection Theorem of Ahlswede and Khachatrian [AK97] in 1997.

The second extension we shall consider concerns matchings. A *q-matching* is a collection of q pairwise-disjoint sets. A set family is therefore intersecting if and only if it does not contain a 2-matching. As an extension of the Erdős–Ko–Rado theorem, Erdős asked how large a q-matching-free k-uniform set family could be, and in [Erd65] showed that when n is large, the best construction consists of taking all sets meeting $[q-1]$. He further conjectured what the solution should be for small n, and this remains an open problem of great interest. For recent results on this conjecture, see, e.g., [Fra13, FRR12, HLS12, LM12].

4.1.2 Beyond the thresholds

The preceding results are all examples of the typical extremal problem, which asks how large a structure can be without containing a forbidden configuration. In this chapter, we study their Erdős–Rademacher variants, a name we now explain.

Arguably the most well-known result in extremal combinatorics is a theorem of Mantel [Man07] from 1907, which states that an n-vertex triangle-free graph can have at most $\left\lfloor \frac{n^2}{4} \right\rfloor$ edges. In an unpublished result, Rademacher strengthened this theorem by showing that any graph with $\left\lfloor \frac{n^2}{4} \right\rfloor + 1$ edges must contain at least $\left\lfloor \frac{n}{2} \right\rfloor$ triangles. In [Erd62a] and [Erd62c], Erdős extended this first to graphs with a linear number of extra edges, and then to cliques larger than triangles. More generally, for any extremal problem, the corresponding Erdős–Rademacher problem asks how many copies of the forbidden configuration must appear in a structure larger than the extremal bound.

In the context of intersecting families, the Erdős–Rademacher question was first investigated by Frankl [Fra77] and, independently, Ahlswede [Ahl80] some forty years ago, who

showed that the number of disjoint pairs of sets in a set family is minimised by taking the sets to be as large as possible.

Theorem 4.1.3 (Frankl [Fra77], 1977; Ahlswede [Ahl80], 1980)**.** *If* $\sum_{i=k+1}^{n}\binom{n}{i} \leq s \leq \sum_{i=k}^{n}\binom{n}{i}$*, then the minimum number of disjoint pairs in a set family of size* s *is attained by some family* $\mathcal{F}$ *with* $\cup_{i>k}\binom{[n]}{i} \subseteq \mathcal{F} \subseteq \cup_{i\geq k}\binom{[n]}{i}$*.*

Note that while this theorem provides the large-scale structure of extremal families, it does not determine exactly which families are optimal. Since we have $\cup_{i>k}\binom{[n]}{i} \subset \mathcal{F}$, each set of size k contributes the same number of disjoint pairs with larger sets. Hence the total number of disjoint pairs is minimised by minimising the number of disjoint pairs between the sets of size k, a problem raised by Ahlswede.

Question 4.1.4 (Ahlswede [Ahl80], 1980)**.** *Given* $0 \leq s \leq \binom{n}{k}$*, which* k*-uniform set families* $\mathcal{F} \subset \binom{[n]}{k}$ *with* $|\mathcal{F}| = s$ *minimise the number of disjoint pairs?*

By the Erdős–Ko–Rado Theorem, we know that when $s \leq \binom{n-1}{k-1}$, we need not have any disjoint pairs, while for $s > \binom{n-1}{k-1}$, there must be at least one disjoint pair. This question can thus be thought of as the Erdős–Rademacher problem for the Erdős–Ko–Rado Theorem.

This question is also deeply connected to the Kneser graph. The Kneser graph $K(n,k)$ has vertices $V = \binom{[n]}{k}$, with vertices X and Y adjacent if and only if the sets are disjoint. An intersecting set family corresponds to an independent set in the Kneser graph. Question 4.1.4 is thus asking which s vertices of the Kneser graph induce the smallest number of edges. Since the Kneser graph is regular, this is equivalent to the edge-isoperimetric problem of finding the largest bipartite subgraph of $K(n,r)$ with one part of size s. Kneser graphs have been extensively studied, and the problem of determining their largest bipartite subgraphs was first raised by Poljak and Tuza in [PT87].

In 2003, Bollobás and Leader [BL03] presented a new proof of Theorem 4.1.3, by relaxing the problem to a continuous version and analyzing fractional set families. They further considered Question 4.1.4, and conjectured that for small families, the initial segment of the lexicographical ordering on $\binom{[n]}{k}$ should be optimal. In the lexicographical ordering, we say

$A < B$ if $\min(A \Delta B) \in A$; that is, we prefer sets with smaller elements. More generally, Bollobás and Leader conjectured that all the extremal families should take the form of what they named ℓ-balls, as defined below. Note that a 1-ball is an initial segment of the lexicographical ordering.

Conjecture 4.1.5 (Bollobás-Leader [BL03], 2003)**.** *One of the families* $\mathcal{A}_{r,l} = \{A \in \binom{[n]}{k} : |A \cap [r]| \geq \ell\}$ *minimises the number of disjoint pairs.*

When $k = 2$, we can think of a set family in $\binom{[n]}{2}$ as a graph on n vertices, and are then asking for which graphs of a given size minimise the number of disjoint pairs of edges. This problem was solved by Ahslwede and Katona [AK78] in 1978, who showed that the extremal graphs were always either the union of stars (a collection of vertices connected to all the other vertices), or their complement.

Theorem 4.1.6 (Ahlswede–Katona [AK78], 1978)**.** *Over all graphs on* n *vertices with* m *edges, either* $\mathcal{L}_{n,2}(m)$ *or* $\mathcal{C}_{n,2}(m)$ *minimises the number of disjoint pairs of edges. Moreover, if* $m < \frac{1}{2}\binom{n}{2} - \frac{n}{2}$*, then* $\mathcal{L}_{n,2}(m)$ *is optimal, while if* $m > \frac{1}{2}\binom{n}{2} + \frac{n}{2}$*, then* $\mathcal{C}_{n,2}(m)$ *is optimal.*

In that paper, they asked for a different generalisation, namely which k-uniform set families minimise the number of $(k-1)$-disjoint pairs. This is the Erdős–Rademacher problem for t-intersecting families when $t = k - 1$, and is known as the Kleitman–West problem, which shares some connections to information theory. An exact solution appears difficult to obtain, and a natural conjecture of Kleitman for this problem has been proven to be untrue by Ahlswede and Cai [AC99]. They later solved a continuous relaxation of the $k = 3$ case in [AC06], marking the furthest progress made on this problem.

4.1.3 Our results

Our main result verifies Conjecture 4.1.5 for small families, showing that initial segments of the lexicographical ordering minimise the number of disjoint pairs. We denote by $\mathcal{L}_{n,k}(s)$ the first s sets in the lexicographical ordering on $\binom{[n]}{k}$. Note that the size of ℓ full stars, say with centres $\{1, 2, \ldots, \ell\}$, is $\binom{n}{k} - \binom{n-\ell}{k}$. The following theorem shows that provided n is large enough with respect to k and ℓ, it is optimal to take sets from the first ℓ stars.

Theorem 4.1.7. *Provided* $n > 108k^2\ell(k+\ell)$ *and* $0 \leq s \leq \binom{n}{k} - \binom{n-\ell}{k}$, $\mathcal{L}_{n,k}(s)$ *minimises the number of disjoint pairs among all families of* s *sets in* $\binom{[n]}{k}$.

As a by-product of our proof, we shall obtain a characterisation of the extremal families in this range, which we provide in Proposition 4.2.8. Corollary 4.2.10 shows that we can also use Theorem 4.1.7 to determine which families are optimal when s is very close to $\binom{n}{k}$.

We further show that $\mathcal{L}_{n,k}(s)$ also minimises the number of q-matchings.

Theorem 4.1.8. *Provided* $n > n_1(k,q,\ell)$ *and* $0 \leq s \leq \binom{n}{k} - \binom{n-\ell}{k}$, $\mathcal{L}_{n,k}(s)$ *minimises the number of* q*-matchings among all families of* s *sets in* $\binom{[n]}{k}$.

Finally, we extend our methods to determine which families minimise the number of t-disjoint pairs. When $t = k-1$, this provides a partial solution to the problem of Kleitman and West. When n is large with respect to k, t and ℓ, an extremal family is contained in the union of ℓ full t-stars. As we shall discuss in Section 4.4, not all such unions are isomorphic, and once again it is the lexicographical ordering that is optimal.

Theorem 4.1.9. *Provided* $n \geq n_2(k,t,\ell)$ *and* $0 \leq s \leq \binom{n-t+1}{k-t+1} - \binom{n-t-\ell+1}{k-t+1}$, $\mathcal{L}_{n,k}(s)$ *minimises the number of* t*-disjoint pairs among all families of* s *sets in* $\binom{[n]}{k}$.

We again characterise all extremal families in Corollary 4.4.5, while Corollary 4.4.6 provides the solution when s very close to $\binom{n}{k}$.

4.1.4 Outline and notation

The remainder of the chapter is organised as follows. In Section 4.2 we study disjoint pairs, and prove Theorem 4.1.7. In Section 4.3, we consider the number of q-matchings, and prove Theorem 4.1.8. In Section 4.4, we extend our results to t-disjoint pairs, proving Theorem 4.1.9. In the final section we present some further remarks and open problems.

We denote by $[n]$ the set of the first n positive integers, and use this as the ground set for our set families. Given a set X, $\binom{X}{k}$ is the family of all k-subsets of X. The number of disjoint pairs between two families $\mathcal{F}$ and $\mathcal{G}$ is given by $\mathrm{dp}(\mathcal{F},\mathcal{G}) = |\{(F,G) \in \mathcal{F}\times\mathcal{G} : F\cap G = \emptyset\}|$, and the number of disjoint pairs within a family $\mathcal{F}$ is denoted by $\mathrm{dp}(\mathcal{F}) = \frac{1}{2}\mathrm{dp}(\mathcal{F},\mathcal{F})$.

For given n, k and s, we let $\mathrm{dp}(n,k,s)$ denote the minimum of $\mathrm{dp}(\mathcal{F})$ over all k-uniform set families on $[n]$ of size s. We define $\mathrm{dp}^{(q)}(\mathcal{F})$ and $\mathrm{dp}^{(q)}(n,k,s)$ similarly for the number of q-matchings in set families, and $\mathrm{dp}_t(\mathcal{F}), \mathrm{dp}_t(\mathcal{F},\mathcal{G})$, and $\mathrm{dp}_t(n,k,s)$ for the number of t-disjoint pairs.

Given any set family $\mathcal{F}$, and a set $X \subset [n]$, we let $\mathcal{F}(X) = \{F \in \mathcal{F} : X \subset F\}$ be those sets in the family containing X. If X is a singleton, we shall drop the set notation, and write $\mathcal{F}(x)$. Finally, we define a *cover* of a family to be a set X with $\cup_{x\in X}\mathcal{F}(x) = \mathcal{F}$; that is, a set of elements that touches every set. A *t-cover* is a collection of t-sets such that every set in the family contains one of the t-sets.

4.2 Disjoint pairs

In this section we will show that for small families, initial segments of the lexicographical ordering, $\mathcal{L}_{n,k}(s)$, minimise the number of disjoint pairs. Note that when $s \leq \binom{n-1}{k-1}$, $\mathcal{L}_{n,k}(s)$ is a star, which is an intersecting family and thus clearly optimal. The following result of Katona et al [KKK12] shows that if we add one set to a full star, the resulting family will also be optimal.

Proposition 4.2.1. *Suppose $n \geq 2k$. Any family $\mathcal{F} \subset \binom{[n]}{k}$ with $|\mathcal{F}| = \binom{n-1}{k-1} + 1$ contains at least $\binom{n-k-1}{k-1}$ disjoint pairs.*

Our first theorem shows that as we add sets to the family, we should try to cover our family with as few stars as possible, as is achieved by $\mathcal{L}_{n,k}(s)$. Later, in Proposition 4.2.8, we shall precisely characterise all extremal families. We begin by recalling the statement of the theorem.

Theorem 4.1.6. *Provided $n > 108k^2\ell(k+\ell)$ and $0 \leq s \leq \binom{n}{k} - \binom{n-\ell}{k}$, $\mathcal{L}_{n,k}(s)$ minimises the number of disjoint pairs among all families of s sets in $\binom{[n]}{k}$.*

In our notation, the above theorem gives $\mathrm{dp}(n,k,s) = \mathrm{dp}(\mathcal{L}_{n,k}(s))$ for such values of s. Let $1 \leq r \leq \ell$ be such that $\binom{n}{k} - \binom{n-r+1}{k} < s \leq \binom{n}{k} - \binom{n-r}{k}$. Since $\mathcal{L}_{n,k}(s)$ contains all sets

meeting $\{1, 2, \ldots, r-1\}$, and all the remaining sets contain r, we can provide an explicit formula for the minimum number of disjoint pairs.

$$\begin{aligned}\mathrm{dp}(n,k,s) = \mathrm{dp}(\mathcal{L}_{n,k}(s)) &= \sum_{i=1}^{r} \sum_{\substack{F \in \mathcal{L}_{n,k}(s) \\ \min F = i}} |\{G \in \mathcal{L}_{n,k}(s) : \min G < i, F \cap G = \emptyset\}| \\ &= \sum_{i=2}^{r-1} \binom{n-i}{k-1} \sum_{j=1}^{i-1} \binom{n-j-k}{k-1} + \left(s - \sum_{i=1}^{r-1} \binom{n-i}{k-1}\right) \sum_{j=1}^{r-1} \binom{n-j-k}{k-1}.\end{aligned}$$

It will be useful to have a simpler upper bound on $\mathrm{dp}(n,k,s)$. Note that we can assign each set in $\mathcal{L}_{n,k}(s)$ to an element of $[r]$ it contains. It can then only be disjoint from sets assigned to different elements. In the worst case, an equal number of sets is assigned to each element, giving the bound

$$\mathrm{dp}(n,k,s) \leq \binom{r}{2} \left(\frac{s}{r}\right)^2 = \frac{1}{2}\left(1 - \frac{1}{r}\right) s^2. \tag{4.1}$$

We shall often require bounds on $\binom{n-2}{k-2}$ in terms of s. Since $\mathcal{L}_{n,k}(s)$ contains all sets meeting $[r-1]$ and n is large, the Bonferroni Inequalities give

$$\begin{aligned}s = |\mathcal{L}_{n,k}(s)| &\geq (r-1)\binom{n-1}{k-1} - \binom{r-1}{2}\binom{n-2}{k-2} \\ &= \left(\frac{(r-1)(n-1)}{k-1} - \binom{r-1}{2}\right)\binom{n-2}{k-2} \geq \frac{rn}{3k}\binom{n-2}{k-2}.\end{aligned} \tag{4.2}$$

Our proof of Theorem 4.1.7 will proceed according to the following steps. First we shall argue that if a family $\mathcal{F}$ has at most $\frac{1}{2}\left(1 - \frac{1}{r}\right) s^2$ disjoint pairs, then it must contain a popular element; that is, some $x \in [n]$ contained in many sets of $\mathcal{F}$. The second step consists of a series of arguments to show that $\mathcal{F}$ can be covered by r elements. The final step will then show that among all families that can be covered by r elements, $\mathcal{L}_{n,k}(s)$ minimises the number of disjoint pairs.

Proof of Theorem 4.1.7. We prove the theorem by induction on n and s. For the base case, suppose $0 \leq s \leq \binom{n}{k} - \binom{n-1}{k} = \binom{n-1}{k-1}$. In this range, $\mathcal{L}_{n,k}(s)$ is a star, consisting only of sets containing 1, and thus obviously minimises the number of disjoint pairs.

For the induction step, let $\mathcal{F}$ be an extremal family with $|\mathcal{F}| = s > \binom{n-1}{k-1}$, and so $r \geq 2$. Suppose first that $\mathcal{F}$ contains a full star. Without loss of generality, we may assume $\mathcal{F}$ has

all sets containing 1, so $\mathcal{F}(1) = \left\{F \in \binom{[n]}{k} : 1 \in F\right\}$. Since $\mathcal{F}(1)$ is an intersecting family, we have $\mathrm{dp}(\mathcal{F}) = \mathrm{dp}\left(\mathcal{F}(1), \mathcal{F} \setminus \mathcal{F}(1)\right) + \mathrm{dp}\left(\mathcal{F} \setminus \mathcal{F}(1)\right)$. Now any set $F \in \mathcal{F}$ with $1 \notin F$ is disjoint from exactly $\binom{n-k-1}{k-1}$ sets in $\mathcal{F}(1)$, giving $\mathrm{dp}\left(\mathcal{F}(1), \mathcal{F} \setminus \mathcal{F}(1)\right) = |\mathcal{F} \setminus \mathcal{F}(1)| \binom{n-k-1}{k-1} = \left(s - \binom{n-1}{k-1}\right)\binom{n-k-1}{k-1}$, regardless of the structure of $\mathcal{F} \setminus \mathcal{F}(1)$. Since $\mathcal{F} \setminus \mathcal{F}(1)$ is a family of $s - \binom{n-1}{k-1}$ sets in $[n] \setminus \{1\}$, our induction hypothesis implies $\mathrm{dp}\left(\mathcal{F} \setminus \mathcal{F}(1)\right)$ is minimised by the initial segment of the lexicographical order. Since $\mathcal{L}_{n,k}(s)$ consists of all sets containing 1, and the initial segment of the lexicographical order on $[n] \setminus \{1\}$, it follows that $\mathcal{L}_{n,k}(s)$ is optimal, as claimed.

Hence we may assume that $\mathcal{F}$ does not contain any full star. Consequently, given any $F \in \mathcal{F}$ and $x \in [n]$, we may replace F by a set containing x.

Step 1: Show there exists some $x \in [n]$ with $|\mathcal{F}(x)| \geq \frac{s}{3r}$.

We begin by showing there cannot be too many moderately popular elements.

Claim 4.2.2. $\left|\{x : |\mathcal{F}(x)| \geq \frac{s}{3kr}\}\right| < 6kr$.

Proof. Suppose not, and consider $X \subset \{x : |\mathcal{F}(x)| \geq \frac{s}{3kr}\}$ with $|X| = 6kr$. Using (4.2), we have

$$\begin{aligned} s = |\mathcal{F}| &\geq |\cup_{x \in X} \mathcal{F}(x)| \geq \sum_{x \in X} |\mathcal{F}(x)| - \sum_{x,y \in X} |\mathcal{F}(x) \cap \mathcal{F}(y)| \\ &\geq |X| \cdot \frac{s}{3kr} - \binom{|X|}{2}\binom{n-2}{k-2} \geq 2s - 18k^2r^2 \cdot \frac{3k}{nr}s = \left(2 - \frac{54k^3r}{n}\right)s. \end{aligned}$$

Since $n > 54k^3r$, we reach a contradiction. □

We now show the existence of a popular element.

Claim 4.2.3. *There is some $x \in [n]$ with $|\mathcal{F}(x)| > \frac{s}{3r}$.*

Proof. We must have $\frac{1}{2}\left(1 - \frac{1}{r}\right)s^2 \geq \mathrm{dp}(\mathcal{L}_{n,k}(s)) \geq \mathrm{dp}(\mathcal{F}) = \frac{1}{2}\sum_{F \in \mathcal{F}} \mathrm{dp}(F, \mathcal{F})$ by the extremality of $\mathcal{F}$.

Now $\mathrm{dp}(F,\mathcal{F}) = s - |\cup_{x\in F}\mathcal{F}(x)| \geq s - \sum_{x\in F}|\mathcal{F}(x)|$, and so we have

$$\left(1-\frac{1}{r}\right)s^2 \geq \sum_{F\in\mathcal{F}}\left(s-\sum_{x\in F}|\mathcal{F}(x)|\right) = s^2 - \sum_{F\in\mathcal{F}}\sum_{x\in F}|\mathcal{F}(x)| = s^2 - \sum_x |\mathcal{F}(x)|^2 .$$

Let $X = \{x : |\mathcal{F}(x)| \geq \frac{s}{3kr}\}$, and note that by the previous claim, $|X| < 6kr$. Moreover, without loss of generality, suppose 1 is the most popular element, so $|\mathcal{F}(x)| \leq |\mathcal{F}(1)|$ for all x. We split the above sum into those $x \in X$ and those $x \notin X$, giving

$$\frac{s^2}{r} \leq \sum_{x\in X}|\mathcal{F}(x)|^2 + \sum_{x\notin X}|\mathcal{F}(x)|^2 \leq |\mathcal{F}(1)|\sum_{x\in X}|\mathcal{F}(x)| + \frac{s}{3kr}\sum_{x\notin X}|\mathcal{F}(x)|. \tag{4.3}$$

We bound the first sum by noting that

$$\begin{aligned}\sum_{x\in X}|\mathcal{F}(x)| &\leq |\cup_{x\in X}\mathcal{F}(x)| + \sum_{\{x,y\}\subset X}|\mathcal{F}(x)\cap\mathcal{F}(y)| \\ &\leq s + \binom{|X|}{2}\binom{n-2}{k-2} \leq \left(1+\frac{54k^3r}{n}\right)s \leq 2s,\end{aligned}$$

using (4.2) and our bound on n. The second sum is bounded by $\sum_{x\notin X}|\mathcal{F}(x)| \leq \sum_x |\mathcal{F}(x)| = ks$. Substituting these bounds in (4.3) gives $\frac{s^2}{r} \leq 2|\mathcal{F}(1)|s + \frac{s^2}{3r}$, and so $|\mathcal{F}(1)| \geq \frac{s}{3r}$, as required. □

This concludes Step 1.

Step 2: Show there is a cover of size r.

We begin by using the existence of a popular element to argue that there is a reasonably small cover, and then provide a number of claims that together imply an extremal family must in fact be covered by r elements.

Claim 4.2.4. *$X = \{x : |\mathcal{F}(x)| \geq \frac{s}{3kr}\}$ is a cover for $\mathcal{F}$.*

Proof. Suppose for contradiction X is not a cover. Then there must be some set $F \in \mathcal{F}$ with $F \cap X = \emptyset$, and so $|\mathcal{F}(x)| < \frac{s}{3kr}$ for all $x \in F$. Hence $\mathrm{dp}(F,\mathcal{F}) = s - |\cup_{x\in F}\mathcal{F}(x)| \geq s - \sum_{x\in F}|\mathcal{F}(x)| > s - \frac{s}{3r}$. On the other hand, by Claim 4.2.3, we may assume $|\mathcal{F}(1)| \geq \frac{s}{3r}$. Thus if G is any set containing 1, we have $\mathrm{dp}(G,\mathcal{F}) \leq s - |\mathcal{F}(1)| = s - \frac{s}{3r}$. Hence replacing F

with such a set G, which is possible since $\mathcal{F}(1)$ is not a full star, would decrease the number of disjoint pairs in $\mathcal{F}$, contradicting its optimality.

Hence X must be a cover for $\mathcal{F}$, as claimed. □

By Claim 4.2.2, we have $|X| \leq 6kr$. Take a minimal subcover of X containing 1; without loss of generality, we may assume this subcover is $[m]$, for some $r \leq m \leq 6kr$. We shall now proceed to show that an extremal family must have $m = r$, giving rise to the smallest possible cover.

Rather than working with the subfamilies $\mathcal{F}(i)$, $i \in [m]$, we shall avoid double-counting by instead considering the subfamilies $\mathcal{F}^*(i) = \{F \in \mathcal{F} : \min F = i\}$. Note that the families $\mathcal{F}^*(i)$ partition $\mathcal{F}$.

Claim 4.2.5. *For every $i, j \in [m]$, we have $|\mathcal{F}^*(i)| \geq |\mathcal{F}^*(j)| - \frac{3mk^2}{rn}s$.*

Proof. First we claim that there is some $F \in \mathcal{F}$ with $F \cap [m] = \{i\}$, which in particular implies $F \in \mathcal{F}^*(i)$. Indeed, the number of sets in $\mathcal{F}(i)$ intersecting another element in $[m]$ is less than $m\binom{n-2}{k-2} \leq \frac{3mk}{rn}s \leq \frac{18k^2}{n}s < \frac{s}{3kr}$. However, since $[m]$ is a subcover of X from Claim 4.2.4, it follows that $|\mathcal{F}(i)| \geq \frac{s}{3kr}$, and thus we must have our desired set $F \in \mathcal{F}^*(i)$.

For any $j \in [m] \setminus \{i\}$, F can intersect at most $k\binom{n-2}{k-2}$ sets in $\mathcal{F}(j)$, since each of these sets must contain both j and one element from F. Summing over all j and using (4.2) gives

$$\begin{aligned}\mathrm{dp}(F, \mathcal{F}) &\geq \sum_{j \neq i} \mathrm{dp}(F, \mathcal{F}^*(j)) \geq \sum_{j \neq i} \left[|\mathcal{F}^*(j)| - k\binom{n-2}{k-2} \right] \\ &= s - |\mathcal{F}^*(i)| - (m-1)k\binom{n-2}{k-2} \geq s - |\mathcal{F}^*(i)| - \frac{3mk^2}{rn}s.\end{aligned}$$

On the other hand, if we were to replace F with a set containing j, it would intersect at least those sets in $\mathcal{F}^*(j)$, and so introduce at most $s - |\mathcal{F}^*(j)|$ disjoint pairs. Since $\mathcal{F}$ is an extremal family, we must have $s - |\mathcal{F}^*(j)| \geq s - |\mathcal{F}^*(i)| - \frac{3mk^2}{rn}s$, or $|\mathcal{F}^*(i)| \geq |\mathcal{F}^*(j)| - \frac{3mk^2}{rn}s$, as required. □

Claim 4.2.6. $m \leq 6r$.

Proof. We shall now bound $|\mathcal{F}^*(i)|$ by taking $j = 1$ in Claim 4.2.5. Recall that by Claim 4.2.3 we have $|\mathcal{F}(1)| = |\mathcal{F}^*(1)| \geq \frac{s}{3r}$, and from Claim 4.2.2 it follows that $m \leq 6kr$. Since $n > 108k^3r$, these bounds give

$$|\mathcal{F}^*(i)| \geq |\mathcal{F}^*(1)| - \frac{3mk^2}{rn}s \geq \frac{s}{3r} - \frac{18k^3}{n}s \geq \frac{s}{6r}.$$

Since $s = |\cup_{i=1}^m \mathcal{F}^*(i)| = \sum_{i=1}^m |\mathcal{F}^*(i)| \geq m \cdot \frac{s}{6r}$, we must have $m \leq 6r$. □

With this tighter bound on m, we are now able to better estimate the number of disjoint pairs in $\mathcal{F}$, and in doing so show that we must actually have $m = r$ if $\mathcal{F}$ is extremal.

Claim 4.2.7. *If $\mathcal{F}$ minimises the number of disjoint pairs, then $\mathcal{F}$ can be covered by r elements.*

Proof. We have $\mathrm{dp}(\mathcal{F}) = \sum_{i<j} \mathrm{dp}(\mathcal{F}^*(i), \mathcal{F}^*(j))$, since $\{\mathcal{F}^*(i)\}$ partitions $\mathcal{F}$ into intersecting families. For $i < j$, note that every set $F \in \mathcal{F}^*(j)$ can intersect at most $k\binom{n-2}{k-2}$ sets in $\mathcal{F}^*(i)$, since those sets would have to contain one element from F as well as i. This shows that $\mathrm{dp}(\mathcal{F}^*(i), \mathcal{F}^*(j)) \geq \left(|\mathcal{F}^*(i)| - k\binom{n-2}{k-2}\right)|\mathcal{F}^*(j)|$. Moreover, note that Claim 4.2.5 implies the bound $|\mathcal{F}^*(1)| \leq \frac{s}{m} + \frac{3mk^2}{rn}s$, since we must have some $i \in [m]$ with $|\mathcal{F}^*(i)| \leq \frac{1}{m}\sum_{j=1}^m |\mathcal{F}^*(j)| = \frac{s}{m}$, and $|\mathcal{F}^*(i)| \geq |\mathcal{F}^*(1)| - \frac{3mk^2}{rn}s$. Thus

$$\begin{aligned}
\mathrm{dp}(\mathcal{F}) &\geq \sum_{i<j} \left(|\mathcal{F}^*(i)| - k\binom{n-2}{k-2}\right)|\mathcal{F}^*(j)| \\
&= \sum_{i<j} |\mathcal{F}^*(i)||\mathcal{F}^*(j)| - k\binom{n-2}{k-2}\sum_j (j-1)|\mathcal{F}^*(j)| \\
&\geq \frac{1}{2}\left(\left(\sum_i |\mathcal{F}^*(i)|\right)^2 - \sum_i |\mathcal{F}^*(i)|^2\right) - mk\binom{n-2}{k-2}\sum_j |\mathcal{F}^*(j)| \\
&\geq \frac{1}{2}\left(s^2 - |\mathcal{F}^*(1)|\sum_i |\mathcal{F}^*(i)|\right) - mk\binom{n-2}{k-2}\sum_j |\mathcal{F}^*(j)| \\
&\geq \frac{1}{2}\left(s^2 - \left(\frac{s}{m} + \frac{3mk^2}{rn}s\right)s\right) - \frac{3mk^2}{rn}s^2 = \frac{1}{2}\left(1 - \frac{1}{m} - \frac{9mk^2}{rn}\right)s^2.
\end{aligned}$$

On the other hand, since $\mathcal{F}$ is extremal, we have $\mathrm{dp}(\mathcal{F}) \leq \mathrm{dp}(\mathcal{L}_{n,k}(s)) \leq \frac{1}{2}\left(1 - \frac{1}{r}\right)s^2$, and so we must have $\frac{1}{r} \leq \frac{1}{m} + \frac{9mk^2}{rn} \leq \frac{1}{m} + \frac{54k^2}{n}$. Since $n > 54k^2r(k+r) \geq 54k^2r(r+1)$, we have $\frac{54k^2}{n} < \frac{1}{r} - \frac{1}{r+1}$, and hence we require $m \leq r$. Thus $\mathcal{F}$ can be covered by r elements. □

This completes Step 2.

Step 3: Show that $\mathcal{L}_{n,k}(s)$ is optimal.

We will now complete the induction argument by showing that $\mathcal{L}_{n,k}(s)$ is indeed an extremal family. From the preceding steps we know $\mathcal{F}$ must be covered by r elements, which we may assume to be $[r]$. We shall now use a complementarity argument to deduce the optimality of $\mathcal{L}_{n,k}(s)$.

Let $\mathcal{A} = \left\{A \in \binom{[n]}{k} : A \cap [r] \neq \emptyset\right\}$ be all sets meeting $[r]$, so we have $\mathcal{F} \subset \mathcal{A}$. Let $\mathcal{G} = \mathcal{A} \setminus \mathcal{F}$. We have

$$\mathrm{dp}(\mathcal{F}) = \mathrm{dp}(\mathcal{A}) - \mathrm{dp}(\mathcal{G}, \mathcal{A}) + \mathrm{dp}(\mathcal{G}),$$

since only disjoint pairs contained in $\mathcal{F}$ survive on the right-hand side.

Since $\mathrm{dp}(\mathcal{A})$ is determined solely by r, and hence s, but is independent of the structure of $\mathcal{F}$, we may treat that term as a constant.

We have $\mathrm{dp}(\mathcal{G}, \mathcal{A}) = \sum_{G \in \mathcal{G}} \mathrm{dp}(G, \mathcal{A})$. For any G, $\mathrm{dp}(G, \mathcal{A})$ is determined by $|G \cap [r]|$, and is maximised when $|G \cap [r]| = 1$. For $\mathcal{F} = \mathcal{L}_{n,k}(s)$, we have $G \cap [r] = \{r\}$ for all $G \in \mathcal{G}$, and so $\mathcal{L}_{n,k}(s)$ maximises $\mathrm{dp}(\mathcal{G}, \mathcal{A})$.

Finally, we obviously have $\mathrm{dp}(\mathcal{G}) \geq 0$, with equality in the case of $\mathcal{F} = \mathcal{L}_{n,k}(s)$.

Hence it follows that $\mathcal{L}_{n,k}(s)$ minimises the number of disjoint pairs, completing the proof. □

This proof also allows us to characterise all extremal families.

Proposition 4.2.8. *Provided $n > 108k^2r(k+r)$ and $\binom{n}{k} - \binom{n-r+1}{k} \leq s \leq \binom{n}{k} - \binom{n-r}{k}$, then a set family $\mathcal{F} \subset \binom{[n]}{k}$ of size s minimises the number of disjoint pairs if and only if it has one of the two following structures:*

(i) $\mathcal{F}$ contains $r-1$ full stars, with the remaining sets forming an intersecting family, or

(ii) $\mathcal{F}$ has a cover X of size r, and if $\mathcal{G} = \left\{G \in \binom{[n]}{k} \setminus \mathcal{F} : G \cap X \neq \emptyset\right\}$, then $\mathcal{G}$ is intersecting, and $|G \cap X| = 1$ for all $G \in \mathcal{G}$.

Proof. We prove the proposition by induction on n and s. If $0 \leq s \leq \binom{n-1}{k-1}$, then clearly a family is extremal if and only if it is intersecting, as there need not be any disjoint pairs. Since $r = 1$ for this value of s, this is covered by case (i).

For the induction step, note that if $\mathcal{F}$ is extremal and contains a full star, say $\mathcal{F}(1)$, then $\mathcal{F} \setminus \mathcal{F}(1)$ must also be extremal. Applying the induction hypothesis gives the result, since adding a full star to either (i) or (ii) preserves the structure.

Hence we may assume there is no full star. Claim 4.2.7 then shows that $\mathcal{F}$ has a cover of size r, while the complementarity argument from Step 3 gives the above characterisation of the family $\mathcal{G}$. □

Finally, we use Theorem 4.1.7 to determine which large families minimise the number of disjoint pairs. Recall that, as explained in Section 4.1, we can view this problem as an edge-isoperimetric for the Kneser graph, which is regular. The following lemma links the edge-isoperimetric problem for small and large vertex sets in regular graphs.

Lemma 4.2.9. *Let $G = (V, E)$ be a regular graph on n vertices. Then $S \subset V$ minimises the number of edges $e(S)$ over all sets of $|S|$ vertices if and only if $V \setminus S$ minimises the number of edges over all sets of $n - |S|$ vertices.*

Proof. Let $s = |S|$, and suppose G is d-regular. Summing the degrees of vertices in S, we have

$$2e(S) = ds - e(S, V \setminus S) = ds - (d(n-s) - 2e(V \setminus S)) = d(2s - n) + 2e(V \setminus S),$$

and so $e(S)$ is minimised if and only if $e(V \setminus S)$ is. □

The following corollary, which is a direct consequence of Theorem 4.1.7 and Lemma 4.2.9, shows that the complement of the lexicographical initial segments are optimal when s is close to $\binom{n}{k}$.

Corollary 4.2.10. *Provided $n > 108k^2\ell(k+\ell)$ and $\binom{n-\ell}{k} \leq s \leq \binom{n}{k}$, $\binom{[n]}{k} \setminus \mathcal{L}_{n,k}\left(\binom{n}{k} - s\right)$ minimises the number of disjoint pairs.*

4.3 q-matchings

In this section, we determine which set families minimise the number of q-matchings. This extends Theorem 4.1.7, which is the case $q = 2$. Note that when $|\mathcal{F}| = s \leq \binom{n}{k} - \binom{n-q+1}{k}$, the lexicographical initial segment does not contain any q-matchings, as all sets meet $[q-1]$. Indeed, this is known to be the largest such family when $n > (2q-1)k-q$, as proven by Frankl [Fra13]. We shall show that, provided n is suitably large, $\mathcal{L}_{n,k}(s)$ continues to be optimal for families of size up to $\binom{n}{k} - \binom{n-\ell}{k}$. Unlike for Theorem 4.1.7, we have made no attempt to optimise the dependence of n on the other parameters. We provide our calculations in asymptotic notation for ease of presentation, where we fix the parameters k, ℓ and q to be constant and let $n \to \infty$. However, our result should certainly hold for $n > C\ell^2k^5(\ell^2+k^2)e^{3q}$.

Our proof strategy will be very similar to before: we will first find a popular element, deduce the existence of a smallest possible cover, and then use a complementarity argument to show that the initial segment of the lexicographical order is optimal. The main difference is in the definition of *popular* - rather than considering how many sets contain the element x, we shall be concerned with how many $(q-1)$-matchings have a set containing x. To this end, we introduce some new notation. Given a set family $\mathcal{F}$, and a set F, let $\mathcal{F}^{(q)}(F)$ denote the number of q-matchings $\{F_1, F_2, \ldots, F_q\}$ in $\mathcal{F}$ with $\cup_{i=1}^{q} F_i \cap F \neq \emptyset$. Similarly, for some $x \in [n]$, we let $\mathcal{F}^{(q)}(x) = \mathcal{F}^{(q)}(\{x\})$ be the number of q-matchings with $x \in \cup_{i=1}^{q} F_i$.

Theorem 4.1.7. *Provided $n > n_1(k, q, \ell)$ and $0 \leq s \leq \binom{n}{k} - \binom{n-\ell}{k}$, $\mathcal{L}_{n,k}(s)$ minimises the number of q-matchings among all families of s sets in $\binom{[n]}{k}$.*

As before, we start with some estimates on $\mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s))$. Let r be such that $\binom{n}{k} - \binom{n-r+1}{k} < s \leq \binom{n}{k} - \binom{n-r}{k}$. We may assign each set in $\mathcal{L}_{n,k}(s)$ to one of its elements in $[r]$. Note that a q-matching cannot contain two sets assigned to the same element, and so to obtain a q-matching, we must choose sets from different elements in $[r]$. By convexity, the worst case is when the sets are equally distributed over $[r]$, giving the upper bound

$$\mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s)) \leq \binom{r}{q}\left(\frac{s}{r}\right)^q. \tag{4.4}$$

In this case we shall also require a lower bound. Note that $\mathcal{L}_{n,k}(s)$ contains all sets meeting $[r-1]$, with the remaining sets containing $\{r\}$; suppose there are $\alpha\binom{n-1}{k-1}$ such sets. Note that we have $s = \binom{n}{k} - \binom{n-r+1}{k} + \alpha\binom{n-1}{k-1} \leq (r-1)\binom{n-1}{k-1} + \alpha\binom{n-1}{k-1}$, so $\binom{n-1}{k-1} \geq \frac{s}{r-1+\alpha}$.

We shall consider two types of q-matchings - those with one of the $\alpha\binom{n-1}{k-1}$ sets that only meet $[r]$ at r, and those without. For the first type, we have $\alpha\binom{n-1}{k-1}$ choices for the set containing r. For the remaining sets in the q-matching, we will avoid any overcounting by restricting ourselves to sets that only contain one element from $[r-1]$. We can then make one of $\binom{r-1}{q-1}$ choices for how the remaining $q-1$ sets will meet $[r-1]$. For each such set, we must avoid all other elements in $[r]$ and all previously used elements, leaving us with at least $\binom{n-kq-r}{k-1} = (1-o(1))\binom{n-1}{k-1}$ options.

For the second type of q-matchings, there are $\binom{r-1}{q}$ ways to choose how the sets meet $[r-1]$, and then at least $\binom{n-kq-r}{k-1}$ choices for each set. Hence in total we have

$$\begin{aligned}\mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s)) &\geq (1-o(1))\left(\alpha\binom{r-1}{q-1} + \binom{r-1}{q}\right)\binom{n-1}{k-1}^q \\ &\geq (1-o(1))\left(\frac{\alpha\binom{r-1}{q-1} + \binom{r-1}{q}}{(r-1+\alpha)^q}\right)s^q.\end{aligned}$$

For any $s > 0$, this function of α is monotone increasing when $0 \leq \alpha \leq 1$, and so the right-hand side is minimised when $\alpha = 0$. This gives the lower bound

$$\mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s)) \geq (1-o(1))\binom{r-1}{q}\left(\frac{s}{r-1}\right)^q. \tag{4.5}$$

Having established these bounds, we now prove Theorem 4.1.8.

Proof of Theorem 4.1.8. Our proof is by induction, on n, q and s. The base case for $q = 2$ is given by Theorem 4.1.7[1]. As noted earlier, if $s \leq \binom{n}{k} - \binom{n-q+1}{k}$, then $\mathcal{L}_{n,k}(s)$ does not contain any q-matchings, and hence is clearly optimal. Hence we may proceed to the induction step, with $q \geq 3$ and $\binom{n}{k} - \binom{n-q+1}{k} < s \leq \binom{n}{k} - \binom{n-\ell}{k}$. In particular, we have $q \leq r \leq \ell$ and $s = \Omega(n^{k-1})$.

[1]Alternatively, we may use the trivial base case of $q = 1$, where we merely count the number of sets.

Let $\mathcal{F}$ be an extremal family of size s. We again first consider the case where $\mathcal{F}$ contains a full star, which we shall assume to be all sets containing 1. We split our q-matchings based on whether or not they meet 1, giving $\mathrm{dp}^{(q)}(\mathcal{F}) = |\mathcal{F}^{(q)}(1)| + \mathrm{dp}^{(q)}(\mathcal{F} \setminus \mathcal{F}(1))$.

Note that every $(q-1)$-matching not meeting 1 can be extended to a q-matching by exactly $\binom{n-k(q-1)-1}{k-1}$ sets containing 1, so $|\mathcal{F}^{(q)}(1)| = \mathrm{dp}^{(q-1)}(\mathcal{F} \setminus \mathcal{F}(1))\binom{n-k(q-1)-1}{k-1}$. By the induction hypothesis, $\mathrm{dp}^{(q-1)}(\mathcal{F} \setminus \mathcal{F}(1))$ is minimised by the lexicographical order. Similarly, $\mathrm{dp}^{(q)}(\mathcal{F} \setminus \mathcal{F}(1))$ is also minimised by the lexicographical order, and hence we deduce that $\mathrm{dp}^{(q)}(\mathcal{F}) \geq \mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s))$.

Thus we may assume that $\mathcal{F}$ does not contain any full stars. Hence, for any $x \in [n]$ and any $F \in \mathcal{F}$, we may replace F by a set containing x.

Step 1: Show there is a popular element $x \in [n]$, with $|\mathcal{F}^{(q-1)}(x)| = \Omega(s^{q-1})$.

A $(q-1)$-matching in $\mathcal{F}$ can be extended to a q-matching by a set $F \in \mathcal{F}$ precisely when the other $q-1$ sets do not meet F. Thus F is in $\mathrm{dp}^{(q-1)}(\mathcal{F}) - |\mathcal{F}^{(q-1)}(F)|$ q-matchings. Summing over all F gives

$$q \cdot \mathrm{dp}^{(q)}(\mathcal{F}) = \sum_{F \in \mathcal{F}} \Big(\mathrm{dp}^{(q-1)}(\mathcal{F}) - |\mathcal{F}^{(q-1)}(F)| \Big) = s \cdot \mathrm{dp}^{(q-1)}(\mathcal{F}) - \sum_{F \in \mathcal{F}} |\mathcal{F}^{(q-1)}(F)|.$$

By the induction hypothesis, $\mathrm{dp}^{(q-1)}(\mathcal{F}) \geq \mathrm{dp}^{(q-1)}(\mathcal{L}_{n,k}(s))$, and since $\mathcal{F}$ is extremal, we must have $\mathrm{dp}^{(q)}(\mathcal{F}) \leq \mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s))$. Combining these facts with the bounds from (4.4) and (4.5), we get

$$\begin{aligned}\sum_{F \in \mathcal{F}} |\mathcal{F}^{(q-1)}(F)| &= s \cdot \mathrm{dp}^{(q-1)}(\mathcal{F}) - q \cdot \mathrm{dp}^{(q)}(\mathcal{F}) \geq s \cdot \mathrm{dp}^{(q-1)}(\mathcal{L}_{n,k}(s)) - q \cdot \mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s)) \\ &\geq (1 - o(1)) \binom{r-1}{q-1} \frac{s^q}{(r-1)^{q-1}} - q\binom{r}{q}\frac{s^q}{r^q} = \Omega(s^q).\end{aligned}$$

Averaging over the s sets in $\mathcal{F}$, we must have $|\mathcal{F}^{(q-1)}(F)| = \Omega(s^{q-1})$ for some $F \in \mathcal{F}$. Since $\mathcal{F}^{(q-1)}(F) = \cup_{x \in F} \mathcal{F}^{(q-1)}(x)$, by averaging over the k elements in F we have $|\mathcal{F}^{(q-1)}(x)| = \Omega(s^{q-1})$ for some $x \in F$.

This completes Step 1.

Step 2: Show there is a cover of size r.

From Step 1, we know there is some popular element, which we may assume to be 1. We start by showing the existence of a reasonably small cover.

Claim 4.3.1. *$X = \{x : |\mathcal{F}^{(q-1)}(x)| \geq \frac{1}{k}|\mathcal{F}^{(q-1)}(1)|\}$ is a cover for $\mathcal{F}$.*

Proof. Suppose for contradiction that X was not a cover for $\mathcal{F}$. Then there is some set $F \in \mathcal{F}$ such that $F \cap X = \emptyset$, and so $\left|\mathcal{F}^{(q-1)}(x)\right| < \frac{1}{k}\left|\mathcal{F}^{(q-1)}(1)\right|$ for all $x \in F$. Since $\mathcal{F}^{(q-1)}(F) = \cup_{x\in F}\mathcal{F}^{(q-1)}(x)$, the number of q-matchings F is contained in is given by

$$\mathrm{dp}^{(q-1)}(\mathcal{F}) - \left|\mathcal{F}^{(q-1)}(F)\right| \geq \mathrm{dp}^{(q-1)}(\mathcal{F}) - \sum_{x\in F}\left|\mathcal{F}^{(q-1)}(x)\right| > \mathrm{dp}^{(q-1)}(\mathcal{F}) - \left|\mathcal{F}^{(q-1)}(1)\right|.$$

On the other hand, a set containing 1 can be in at most $\mathrm{dp}^{(q-1)}(\mathcal{F}) - \left|\mathcal{F}^{(q-1)}(1)\right|$ q-matchings. Since $\mathcal{F}(1)$ is not a full star, we may replace F with a set containing 1, which would decrease the number of q-matchings in $\mathcal{F}$. This contradicts the optimality of $\mathcal{F}$, and it follows that X is a cover. □

Having shown that this set X is a cover, we now show that X is not too big; its size is bounded by a function of k, q and ℓ.

Claim 4.3.2. *$|X| = O(1)$.*

Proof. As there can be at most s^{q-1} $(q-1)$-matchings in $\mathcal{F}$, we have

$$\frac{1}{k}|\mathcal{F}^{(q-1)}(1)||X| \leq \sum_{x\in X}|\mathcal{F}^{(q-1)}(x)| \leq \sum_{x\in[n]}|\mathcal{F}^{(q-1)}(x)| = k(q-1)\mathrm{dp}^{(q-1)}(\mathcal{F}) \leq k(q-1)s^{q-1}.$$

Since $|\mathcal{F}^{(q-1)}(1)| = \Omega(s^{q-1})$, this gives $|X| = O(1)$, as required. □

Now take a minimal subcover of X, which we may assume to be $[m]$, where $m = O(1)$. We shall shift our focus from $(q-1)$-matchings to the individual sets themselves. For each $i \in [m]$, we shall let $\mathcal{F}^{-}(i) = \{F \in \mathcal{F} : F \cap [m] = \{i\}\}$ be those *sets* in $\mathcal{F}$ that meet $[m]$ precisely at i; by the minimality of the cover, these subfamilies are non-empty. Since any set in $\mathcal{F}(i) \setminus \mathcal{F}^{-}(i)$ must contain not just i but also some other element in $[m]$, we have $|\mathcal{F}^{-}(i)| \geq |\mathcal{F}(i)| - m\binom{n-2}{k-2} = |\mathcal{F}(i)| - o(s)$.

We will now show that for an extremal family, we must have $m = r$. We first require the following claim.

Claim 4.3.3. *For any $i, j \in [m]$, we have $|\mathcal{F}(i)| = |\mathcal{F}(j)| + o(s)$.*

Proof. Recall that set $F \in \mathcal{F}$ contributes $\mathrm{dp}^{(q-1)}(\mathcal{F}) - |\mathcal{F}^{(q-1)}(F)|$ q-matchings to $\mathcal{F}$. By estimating $|\mathcal{F}^{(q-1)}(F)|$ for sets containing i or j, we shall show that if $|\mathcal{F}(i)|$ and $|\mathcal{F}(j)|$ are very different, then we can decrease the number of q-matchings by shifting sets.

Consider a set $F \in \mathcal{F}^-(i)$. We wish to bound $\left|\mathcal{F}^{(q-1)}(F)\right|$.

For every $(q-1)$-matching in $\mathcal{F}^{(q-1)}(F)$, we must have at least one of the sets in the $(q-1)$-matching meeting F. Either this set can contain i, in which case there are $|\mathcal{F}(i)|$ possibilities, or it contains some element in $F \setminus \{i\}$, as well as some element in $[m]$. However, the number of options in the latter case is at most $mk\binom{n-2}{k-2} = o(s)$. We can then count the number of possibilities for the other sets in the matching just as we did when establishing the inequalities (4.4) and (4.5). First we choose representatives $A \subset [m] \setminus \{i\}$ for the other $q-2$ sets, and then we choose sets corresponding to the given elements; that is, $H \in \mathcal{F}(a)$ for all $a \in A$. This provides an overestimate for $\left|\mathcal{F}^{(q-1)}(F)\right|$, as some of these collections of $q-1$ sets may not be disjoint, while some are counted multiple times. However, we do obtain the upper bound

$$|\mathcal{F}^{(q-1)}(F)| \leq (1+o(1))\,|\mathcal{F}(i)| \sum_{A \in \binom{[m]\setminus\{i\}}{q-2}} \prod_{a \in A} |\mathcal{F}(a)|. \tag{4.6}$$

We now consider replacing F by some set G containing j, and determine how many new q-matchings would be formed. The number of q-matchings G contributes is $\mathrm{dp}^{(q-1)}(\mathcal{F}) - \left|\mathcal{F}^{(q-1)}(G)\right| \leq \mathrm{dp}^{(q-1)}(\mathcal{F}) - \left|\mathcal{F}^{(q-1)}(j)\right|$, since $j \in G$.

To bound $\left|\mathcal{F}^{(q-1)}(j)\right|$, note that we can form $(q-1)$-matchings containing j by first choosing a set from $\mathcal{F}^-(j)$, then choosing a set of $q-2$ other representatives $A \subset [m] \setminus \{j\}$, and choosing disjoint sets $H \in \mathcal{F}^-(a)$, $a \in A$. To ensure the sets we choose are disjoint, we must avoid any elements we have already used. There can be at most $k(q-1)$ such elements, and so we have to avoid at most $\binom{n-1}{k-1} - \binom{n-k(q-1)-1}{k-1} \leq k(q-1)\binom{n-2}{k-2} = o(s)$ sets each time. By choosing the sets from $\mathcal{F}^-(a)$, and not $\mathcal{F}(a)$, we ensure there is no overcounting, as each

such $(q-1)$-matching has a unique set of representatives in $[m]$. Thus we have the bound

$$\left|\mathcal{F}^{(q-1)}(j)\right| \geq \left|\mathcal{F}^{-}(j)\right| \sum_{A \in \binom{[m]\setminus\{j\}}{q-2}} \prod_{a\in A} \left(\left|\mathcal{F}^{-}(a)\right| - o(s)\right) = (1-o(1))\left|\mathcal{F}(j)\right| \sum_{A \in \binom{[m]\setminus\{j\}}{q-2}} \prod_{a\in A} \left|\mathcal{F}(a)\right| \tag{4.7}$$

since $|\mathcal{F}^{-}(a)| = |\mathcal{F}(a)| - o(s)$ for all $a \in [m]$.

Since $\mathcal{F}$ is optimal, we must have $\left|\mathcal{F}^{(q-1)}(F)\right| \geq \left|\mathcal{F}^{(q-1)}(G)\right|$. Comparing (4.6) and (4.7), we find

$$(1+o(1))\left|\mathcal{F}(i)\right| \sum_{A \in \binom{[m]\setminus\{i\}}{q-2}} \prod_{a\in A} \left|\mathcal{F}(a)\right| \geq (1-o(1))\left|\mathcal{F}(j)\right| \sum_{A \in \binom{[m]\setminus\{j\}}{q-2}} \prod_{a\in A} \left|\mathcal{F}(a)\right|.$$

Some terms appear on both sides of the inequality, and so taking the difference gives

$$\left(|\mathcal{F}(i)| - |\mathcal{F}(j)|\right) \sum_{A \in \binom{[m]\setminus\{i,j\}}{q-2}} \prod_{a\in A} |\mathcal{F}(a)| \geq o(s^{q-1}).$$

This implies $|\mathcal{F}(i)| \geq |\mathcal{F}(j)| + o(s)$. By symmetry, the reverse inequality also holds, and thus $|\mathcal{F}(i)| = |\mathcal{F}(j)| + o(s)$, as required. □

Note that we have $s = |\mathcal{F}| = \left|\cup_{i\in[m]}\mathcal{F}(i)\right| \geq \sum_{i=1}^{m} |\mathcal{F}(i)| - \sum_{i<j} |\mathcal{F}(i)\cap\mathcal{F}(j)|$. Since $|\mathcal{F}(i)\cap\mathcal{F}(j)| \leq \binom{n-2}{k-2} = o(s)$ for all i, j, it follows that $\sum_{i=1}^{m} |\mathcal{F}(i)| = s + o(s)$. Claim 4.3.3 shows that all the stars have approximately the same size, and so $|\mathcal{F}(i)| = \frac{s}{m} + o(s)$ for each $1 \leq i \leq m$. We can now show that we have a smallest possible cover.

Claim 4.3.4. *If $\mathcal{F}$ is extremal, then $\mathcal{F}$ can be covered by r elements.*

Proof. Now that we have control over the sizes of the subfamilies $\mathcal{F}(i)$, we can estimate the number of q-matchings the family contains. As in our calculations for Claim 4.3.3, we can obtain a q-matching by choosing a collection A of q elements in $[m]$, and then choosing sets from the corresponding subfamilies $\mathcal{F}(a)$, $a \in A$. In order for this choice of sets to form a q-matching, each set we choose should avoid the elements of the previously chosen sets, of which there can be at most $k(q-1)$. Moreover, to avoid overcounting, we shall choose sets from $\mathcal{F}^{-}(a)$, and so shall avoid the other $m-1$ elements of $[m]$. Thus, for a given $a \in A$, the

forbidden sets are those containing a, and one of at most $k(q-1)+m-1$ other elements, and so we forbid at most $(k(q-1)+m-1)\binom{n-2}{k-2} = o(s)$ sets. Thus we have

$$\mathrm{dp}^{(q)}(\mathcal{F}) \geq \sum_{A \in \binom{[m]}{q}} \prod_{a \in A} (|\mathcal{F}(a)| - o(s)) = (1-o(1)) \binom{m}{q} \left(\frac{s}{m}\right)^q.$$

On the other hand, since $\mathcal{F}$ is extremal, we must have $\mathrm{dp}^{(q)}(\mathcal{F}) \leq \mathrm{dp}^{(q)}(\mathcal{L}_{n,k}(s)) \leq \binom{r}{q}\left(\frac{s}{r}\right)^q$. As $\binom{m}{q}\left(\frac{s}{m}\right)^q$ is increasing in m, these bounds imply we must have $m = r$. □

This concludes Step 2.

Step 3: Show that $\mathcal{L}_{n,k}(s)$ is optimal.

We complete the induction by showing that $\mathcal{L}_{n,k}(s)$ does indeed minimise the number of q-matchings. From the previous steps, we may assume that an extremal family $\mathcal{F}$ is covered by $[r]$. As before, we shall let $\mathcal{A} = \left\{A \in \binom{[n]}{k} : A \cap [r] \neq \emptyset\right\}$, so $\mathcal{F} \subset \mathcal{A}$, and we let $\mathcal{G} = \mathcal{A} \backslash \mathcal{F}$. Note that for every $G \in \mathcal{G}$, $\mathrm{dp}^{(q-1)}(\mathcal{A}) - |\mathcal{A}^{(q-1)}(G)|$ counts the number of q-matchings in $\mathcal{A}$ containing G. Hence

$$\begin{aligned}\mathrm{dp}^{(q)}(\mathcal{F}) &\geq \mathrm{dp}^{(q)}(\mathcal{A}) - \sum_{G \in \mathcal{G}} \left(\mathrm{dp}^{(q-1)}(\mathcal{A}) - |\mathcal{A}^{(q-1)}(G)|\right) \\ &= \mathrm{dp}^{(q)}(\mathcal{A}) - |\mathcal{G}|\mathrm{dp}^{(q-1)}(\mathcal{A}) + \sum_{G \in \mathcal{G}} |\mathcal{A}^{(q-1)}(G)|.\end{aligned}$$

Now the first two terms are independent of the structure of $\mathcal{F}$. We claim that $\left|\mathcal{A}^{(q-1)}(G)\right|$ is minimised when $|G \cap [r]| = 1$. Indeed, fix some $G \in \mathcal{G}$. Note that the number of $(q-1)$-matchings in $\mathcal{A}$ that only meet G outside $[r]$ is at most $kr\binom{n-2}{k-2}s^{q-2} = o(s^{q-1})$, since we must choose one of k elements of G and one of r elements of $[r]$ for the set to contain, and then there are at most s^{q-2} choices for the remaining $q-2$ sets. Hence almost all the $(q-1)$-matchings in $\mathcal{A}^{(q-1)}(G)$ meet G in $G \cap [r]$, and thus $\left|\mathcal{A}^{(q-1)}(G)\right|$ is obviously minimised when $|G \cap [r]| = 1$.

When $\mathcal{F} = \mathcal{L}_{n,k}(s)$, we have $G \cap [r] = \{r\}$ for all $G \in \mathcal{G}$, and so the right-hand side is minimised. Moreover, because $\mathcal{G}$ is an intersecting family, it follows that every $(q-1)$-matching in $\mathcal{A}$ can contain at most 1 set from $\mathcal{G}$, and so the above inequality is in fact an equality. This shows that $\mathcal{L}_{n,k}(s)$ minimises the number of q-matchings.

This completes the induction step, and thus the proof of Theorem 4.1.8. □

4.4 t-disjoint pairs

We now seek a different extension of Theorem 4.1.7. Recall that we call a pair of sets F_1, F_2 t-intersecting if $|F_1 \cap F_2| \geq t$, and t-disjoint otherwise. As shown by Wilson [Wil84], provided $n \geq (k-t+1)(t+1)$, the largest t-intersecting family consists of $\binom{n-t}{k-t}$ sets that share a common t-set $X \in \binom{[n]}{t}$; we call such a family a *(full) t-star with centre X*. Note that $\mathcal{L}_{n,k}(\binom{n-t}{k-t})$ is itself a t-star with centre $[t]$. In the following theorem, we show that when n is sufficiently large, the minimum number of t-disjoint pairs is attained by taking full t-stars. In this setting, not all unions of t-stars are isomorphic, as the structure depends on how the centres intersect. We show that it is optimal to have the centres be the first few sets in the lexicographical ordering on $\binom{[n]}{t}$, which is the case for $\mathcal{L}_{n,k}(s)$.

Theorem 4.1.8. *Provided $n \geq n_2(k,t,\ell)$ and $0 \leq s \leq \binom{n-t+1}{k-t+1} - \binom{n-t-\ell+1}{k-t+1}$, $\mathcal{L}_{n,k}(s)$ minimises the number of t-disjoint pairs among all families of s sets in $\binom{[n]}{k}$.*

It shall sometimes be helpful to count the number of t-intersecting pairs instead of t-disjoint pairs. Thus we introduce the notation $\mathrm{int}_t(\mathcal{F})$ to represent the number of t-intersecting pairs of sets in $\mathcal{F}$, and $\mathrm{int}_t(\mathcal{F},\mathcal{G}) = |\{(F,G) \in \mathcal{F} \times \mathcal{G} : |F \cap G| \geq t\}|$ to count the number of cross-t-intersections between $\mathcal{F}$ and $\mathcal{G}$. Note that a set F is t-intersecting with itself, since $|F \cap F| = k > t$. Since $\sum_{F \in \mathcal{F}} \mathrm{int}_t(F,\mathcal{F})$ counts the t-intersecting pairs between distinct sets twice, and those with the same set only once, we obtain the identity $\sum_{F \in \mathcal{F}} \mathrm{int}_t(F,\mathcal{F}) = 2\mathrm{int}_t(\mathcal{F}) - |\mathcal{F}|$.

We begin with a heuristic calculation that suggests why it is optimal to have full t-stars. Let $\mathcal{F}$ be a full t-star, say with centre $X \in \binom{[n]}{t}$, and let F be a set not containing X. For a set G in $\mathcal{F}$ to be t-intersecting with F, G must contain the t elements of X, as well as some $t - |F \cap X|$ elements from F. The number of such sets G is maximised when $|F \cap X| = t-1$, giving

$$\mathrm{int}_t(F,\mathcal{F}) \leq (k-t+1)\binom{n-t-1}{k-t-1} = O(n^{k-t-1}) = o(|\mathcal{F}|). \tag{4.8}$$

Hence if a t-star does not contain a set F, F is t-disjoint from almost all its members. It should thus be optimal to take full t-stars, as that is where the t-intersections come from. Indeed, this turns out to be the case. As we shall see, for a set family $\mathcal{F}$, the leading term in $\mathrm{dp}_t(\mathcal{F})$ is determined by the number of t-stars in $\mathcal{F}$. While unions of t-stars may be non-isomorphic, the differences only affect the lower order terms of $\mathrm{dp}_t(\mathcal{F})$.

In order to prove Theorem 4.1.9, we shall require a few preliminary results. Proposition 4.4.1 can be thought of as a rough characterisation of extremal families, as it shows that the extremal families should be supported on the right number of t-stars. To this end, it will be useful to define an *almost full t-star* to be a t-star in $\mathcal{F}$ containing $(1-o(1))\binom{n-t}{k-t}$ sets. Formally, this means that for all fixed k, t and ℓ, there is some $\varepsilon = \varepsilon(k,t,\ell) > 0$ such that a t-star will be almost full if it contains $(1-\varepsilon)\binom{n-t}{k-t}$ sets.

Proposition 4.4.1. *Suppose $n \geq n_2(k,\ell,t)$, and $\binom{n-t+1}{k-t+1} - \binom{n-t-r+2}{k-t+1} < s \leq \binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}$. If $\mathcal{F} \subset \binom{[n]}{k}$ has the minimum number of t-disjoint pairs over all families of s sets, then either:*

(i) $\mathcal{F}$ contains $r-1$ full t-stars,

(ii) $\mathcal{F}$ consists of r almost full t-stars, or

(iii) $\mathcal{F}$ consists of $r-1$ almost full t-stars.

Once we have determined the large-scale structure of the extremal families, the following lemmas allow us to analyze the lower-order terms and determine that the lexicographical ordering is indeed optimal.

Lemma 4.4.2 shows that of all unions of r full t-stars, the lexicographical ordering contains the fewest sets. This may seem to contradict the lexicographical ordering being optimal, given that the heuristic given by (4.8) suggests that it is optimal to take as few t-stars as possible, and hence we might try to make the union of these stars accommodate as many sets as possible. However, it is because there is more overlap between the lexicographical t-stars that there are fewer t-disjoint pairs between stars.

Lemma 4.4.2. *Suppose $n \geq n_2(k,t,r)$, and let $\mathcal{F}$ be the union of r full t-stars in $\binom{[n]}{k}$. Then $|\mathcal{F}| \geq \binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1} = s$, with equality if and only if $\mathcal{F}$ is isomorphic to $\mathcal{L}_{n,k}(s)$.*

The next lemma shows that if we have r full t-stars, and add a new set to the family, we minimise the number of new t-disjoint pairs created when we have the lexicographical initial segment.

Lemma 4.4.3. *Suppose $n \geq n_2(k,t,r)$, let $\mathcal{L} = \mathcal{L}_{n,k}\left(\binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}\right)$ be the first r full t-stars in the lexicographical order, and let L be a set containing $\{1,2,\ldots,t-1\}$ that is not in $\mathcal{L}$. Let $\mathcal{F}$ be the union of r full t-stars with centres $\{X_1, X_2, \ldots, X_r\}$, and let F be any k-set not in $\mathcal{F}$. Then $\mathrm{dp}_t(F,\mathcal{F}) \geq \mathrm{dp}_t(L,\mathcal{L})$, which equality if and only if $\mathcal{F} \cup \{F\}$ is isomorphic to $\mathcal{L} \cup \{L\}$.*

However, the comparison in Lemma 4.4.3 is not entirely fair, as Lemma 4.4.2 shows that $\mathcal{L}$ will have fewer sets than $\mathcal{F}$, while we ought to be comparing families of the same size. We do this in our final lemma, in the cleanest case when the family $\mathcal{F}$ is a union of full t-stars.

Lemma 4.4.4. *Suppose $n \geq n_2(k,t,r)$, let $\mathcal{F}$ be the union of r full t-stars with centres X_i, $1 \leq i \leq r$, and let $\mathcal{L} = \mathcal{L}_{n,k}(|\mathcal{F}|)$. Then $\mathrm{dp}_t(\mathcal{F}) \geq \mathrm{dp}_t(\mathcal{L})$, with equality if and only if $\mathcal{F}$ is isomorphic to $\mathcal{L}$.*

Armed with Proposition 4.4.1 and these three lemmas, whose proofs we defer until later in this section, we now show how to deduce Theorem 4.1.9.

Proof of Theorem 4.1.9. Let r be such that $\binom{n-t+1}{k-t+1} - \binom{n-t-r+2}{k-t+1} < s \leq \binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}$. In this range $\mathcal{L}_{n,k}(s)$ consists of $r-1$ full t-stars, with the remaining sets forming a partial rth t-star. If $r = 1$, then $\mathcal{L}_{n,k}(s)$ is t-intersecting, and therefore clearly optimal. Hence we may assume $r \geq 2$, and in particular this implies $s = \Omega(n^{k-t})$.

Suppose $\mathcal{F}$ is an optimal family of size s. By analyzing the three cases in Proposition 4.4.1 in turn, we shall show that $\mathrm{dp}_t(\mathcal{F}) \geq \mathrm{dp}_t(\mathcal{L}_{n,k}(s))$, thus completing the proof of Theorem 4.1.9.

Case (i): Suppose $\mathcal{F}$ contains $r-1$ full t-stars, whose union we shall denote by $\mathcal{F}_1$, and $s_2 = s - |\mathcal{F}_1|$ other sets, denoted by $\mathcal{F}_2$. We then have

$$\begin{aligned}\mathrm{dp}_t(\mathcal{F}) &= \mathrm{dp}_t(\mathcal{F}_1) + \mathrm{dp}_t(\mathcal{F}_1, \mathcal{F}_2) + \mathrm{dp}_t(\mathcal{F}_2) \geq \mathrm{dp}_t(\mathcal{F}_1) + \mathrm{dp}_t(\mathcal{F}_1, \mathcal{F}_2) \\ &= \mathrm{dp}_t(\mathcal{F}_1) + \sum_{F \in \mathcal{F}_2} \mathrm{dp}_t(F, \mathcal{F}_1) \geq \mathrm{dp}_t(\mathcal{F}_1) + s_2 \cdot \mathrm{dp}_t(F_0, \mathcal{F}_1),\end{aligned}$$

where $F_0 \in \mathcal{F}_2$ minimises $\mathrm{dp}_t(F, \mathcal{F}_1)$.

Let $\mathcal{L} = \mathcal{L}_{n,k}(s)$ be the corresponding lexicographical initial segment, $\mathcal{L}_1 = \mathcal{L}_{n,k}(|\mathcal{F}_1|)$ be the first $|\mathcal{F}_1|$ sets in the lexicographical ordering, and let $\mathcal{L}_2 = \mathcal{L} \setminus \mathcal{L}_1$ be the next s_2 sets. By Lemma 4.4.2, it follows that $\mathcal{L}_1$ consists of at least $r-1$ full t-stars, and so $\mathcal{L}_2$ lies entirely within the rth lexicographical t-star, and is thus t-intersecting. Hence

$$\begin{aligned}\mathrm{dp}_t(\mathcal{L}) &= \mathrm{dp}_t(\mathcal{L}_1) + \mathrm{dp}_t(\mathcal{L}_1, \mathcal{L}_2) + \mathrm{dp}_t(\mathcal{L}_2) = \mathrm{dp}_t(\mathcal{L}_1) + \mathrm{dp}_t(\mathcal{L}_1, \mathcal{L}_2) \\ &= \mathrm{dp}_t(\mathcal{L}_1) + \sum_{L \in \mathcal{L}_2} \mathrm{dp}_t(L, \mathcal{L}_1) \leq \mathrm{dp}_t(\mathcal{L}_1) + s_2 \cdot \mathrm{dp}_t(L_0, \mathcal{L}_1),\end{aligned}$$

where $L_0 \in \mathcal{L}_2$ maximises $\mathrm{dp}_t(L, \mathcal{L}_1)$ (in fact, by symmetry, this is equal for all $L \in \mathcal{L}_2$).

Note that L_0 will belong to the rth t-star of $\mathcal{L}$, and hence $\mathrm{dp}_t(L_0, \mathcal{L}_1)$ will only count t-disjoint pairs between L_0 and the union of the first $r-1$ t-stars of $\mathcal{L}_1$. By Lemma 4.4.3, we have $\mathrm{dp}_t(F_0, \mathcal{F}_1) \geq \mathrm{dp}_t(L_0, \mathcal{L}_1)$, and by Lemma 4.4.4, we have $\mathrm{dp}_t(\mathcal{F}_1) \geq \mathrm{dp}_t(\mathcal{L}_1)$, from which we deduce $\mathrm{dp}_t(\mathcal{F}) \geq \mathrm{dp}_t(\mathcal{L})$, as required.

Case (ii): In this case we have r almost full t-stars. Using a complementarity argument, we shall reduce this to case (i).

Suppose $\mathcal{F}$ is the union of r almost full t-stars with centres $\{X_1, X_2, \ldots, X_r\}$, let $\mathcal{A} = \cup_{i=1}^{r} \{A \in \binom{[n]}{k} : X_i \subset A\}$ be the family of all sets containing some X_i, and let $\mathcal{G} = \mathcal{A} \setminus \mathcal{F}$. On account of the t-stars being almost full, we have $|\mathcal{G}| = o(n^{k-t})$.

Running the same complementarity argument as in the proof of Theorem 4.1.7, we have

$$\mathrm{dp}_t(\mathcal{F}) = \mathrm{dp}_t(\mathcal{A}) - \mathrm{dp}_t(\mathcal{G}, \mathcal{A}) + \mathrm{dp}_t(\mathcal{G}) = \mathrm{dp}_t(\mathcal{A}) - \sum_{G \in \mathcal{G}} \mathrm{dp}_t(G, \mathcal{A}) + \mathrm{dp}_t(\mathcal{G}). \tag{4.9}$$

To minimise $\mathrm{dp}_t(\mathcal{F})$, we seek to maximise $\sum_{G \in \mathcal{G}} \mathrm{dp}_t(G, \mathcal{A})$ while minimising $\mathrm{dp}_t(\mathcal{G})$. We shall obtain these extrema by shifting the family so that the missing sets, $\mathcal{G}$, will all belong

to one of the t-stars $\mathcal{A}(X_i)$. In this case, the shifted family, $\mathcal{F}'$, will contain $r-1$ full t-stars. Hence we will have reduced the problem to case (i), and so $\mathrm{dp}_t(\mathcal{F}) \geq \mathrm{dp}_t(\mathcal{F}') \geq \mathrm{dp}_t(\mathcal{L}_{n,k}(s))$, as desired.

Note that when $\mathcal{G}$ is a subset of one of the t-stars, $\mathcal{G}$ is t-intersecting, and so $\mathrm{dp}_t(\mathcal{G}) = 0$ is minimised. We now show how to choose which t-star $\mathcal{G}$ should belong to in order to maximise $\sum_{G\in\mathcal{G}} \mathrm{dp}_t(G, \mathcal{A})$.

Since $\mathcal{A}$ is of fixed size, maximising $\mathrm{dp}_t(G, \mathcal{A})$ is equivalent to minimising $\mathrm{int}_t(G, \mathcal{A})$. For $G \in \mathcal{G}$, $\mathrm{int}_t(G, \mathcal{A})$ is determined by the intersections $\{G \cap X_i : 1 \leq i \leq r\}$. There are only a bounded number of possibilities for these intersections, and so we may choose one which minimises $\mathrm{int}_t(G, \mathcal{A})$, under the restriction that $X_i \subset G$ for some i, since $G \in \mathcal{A}$. By (4.8), the number of t-intersecting pairs between G and a t-star it is not in is $o(s)$, and so this minimum occurs when G contains some X_i and no other elements from $\cup_j X_j \setminus X_i$. The number of choices for the set G is then at least $\binom{n-rt}{k-t}$, since after choosing the t elements of X_i, we wish to avoid the remaining elements in $\cup_j X_j$, of which there are at most $(r-1)t$. Since $\binom{n-rt}{k-t} \geq |\mathcal{G}| = o(n^{k-t})$, we may choose all $G \in \mathcal{G}$ to come from the t-star with centre X_i in order to minimise the right hand side of (4.9). We have thus resolved case (ii).

Case (iii): In this case we have $r-1$ almost full t-stars. Since the size of this family is at most $(r-1)\binom{n-t}{k-t}$, while the size of the first $r-1$ t-stars in $\mathcal{L}_{n,k}(s)$ is $\binom{n-t+1}{k-t+1} - \binom{n-t-r+2}{k-t+1} = (r-1)\binom{n-t}{k-t} + o(n^{k-t})$, we can conclude that rth partial t-star in $\mathcal{L}_{n,k}(s)$ has only $o(n^{k-t})$ sets.

Given the family $\mathcal{F}$, we shall construct a larger family $\mathcal{F}'$ by filling the $r-1$ almost full t-stars. Suppose we have to add s_1 sets in order to do so. Note that since the t-stars were almost full, we have $s_1 = o(n^{k-t})$. Since each of the s_1 sets is added to an almost full t-star, it contributes at least $(1-o(1))\binom{n-t}{k-t}$ t-intersecting pairs. Hence $\mathrm{int}_t(\mathcal{F}') \geq \mathrm{int}_t(\mathcal{F}) + (1-o(1))s_1\binom{n-t}{k-t}$.

On the other hand, consider adding the same number of sets to the lexicographical initial segment. The sets in $\mathcal{L}_{n,k}(s+s_1) \setminus \mathcal{L}_{n,k}(s)$ all belong only to the rth t-star, which has only $o(n^{k-t})$ sets. Our calculation in (4.8) shows that each such set also only gains $o(n^{k-t})$

t-intersections from the other stars, and so we have $\mathrm{int}_t(\mathcal{L}_{n,k}(s+s_1)) \leq \mathrm{int}_t(\mathcal{L}_{n,k}(s)) + s_1 \cdot o(n^{k-t})$.

Now $\mathcal{F}'$ consists of $r-1$ full t-stars, and so by Lemma 4.4.4, we have $\mathrm{dp}_t(\mathcal{F}') \geq \mathrm{dp}_t(\mathcal{L}_{n,k}(s+s_1))$, or, equivalently, $\mathrm{int}_t(\mathcal{F}') \leq \mathrm{int}_t(\mathcal{L}_{n,k}(s+s_1))$. Thus $\mathrm{int}_t(\mathcal{F}) + (1-o(1))s_1\binom{n-t}{k-t} \leq \mathrm{int}_t(\mathcal{L}_{n,k}(s)) + s_1 \cdot o(n^{k-t})$, and so $\mathrm{int}_t(\mathcal{F}) \leq \mathrm{int}_t(\mathcal{L}_{n,k}(s))$, with a strict inequality unless $s_1 = 0$. This implies $\mathrm{dp}_t(\mathcal{F}) \geq \mathrm{dp}_t(\mathcal{L}_{n,k}(s))$, as required.

Hence we may conclude that for any family $\mathcal{F}$ with s sets, we have $\mathrm{dp}_t(\mathcal{F}) \geq \mathrm{dp}_t(\mathcal{L}_{n,k}(s))$, proving Theorem 4.1.9. □

By analyzing the cases when we have equality, and using the fact that in Lemmas 4.4.2, 4.4.3 and 4.4.4 we only have equality when the families are isomorphic to the lexicographical ordering, we can characterise all extremal families.

Corollary 4.4.5. *Suppose $n \geq n_2(k,\ell,t)$, and $0 \leq s \leq \binom{n-t+1}{k-t+1} - \binom{n-t-\ell+1}{k-t+1}$, and $\mathcal{F} \subset \binom{[n]}{k}$ minimises the number of t-disjoint pairs over all families of s sets. Then all sets $F \in \mathcal{F}$ share some common $(t-1)$-set X, and $\mathcal{F}' = \{F \setminus X : F \in \mathcal{F}\}$ minimises the number of disjoint pairs over all families of s sets in $\binom{[n]\setminus X}{k-t+1}$.*

Moreover, note that this problem can again be thought of as an isoperimetric inequality in an appropriate graph. Consider the generalised Kneser graph $K(n,k,t)$, with $V = \binom{[n]}{k}$ and an edge between the sets X and Y if $|X \cap Y| < t$. This is a regular graph, with every vertex having degree $\sum_{i=0}^{t-1}\binom{k}{i}\binom{n-k}{k-i}$. Thus, combining Theorem 4.1.9 with Lemma 4.2.9, we can determine which large families minimise the number of t-disjoint pairs.

Corollary 4.4.6. *Provided $n \geq n_2(k,t,\ell)$ and $0 \leq s \leq \binom{n-t+1}{k-t+1} - \binom{n-t-\ell+1}{k-t+1}$, $\binom{[n]}{k} \setminus \mathcal{L}_{n,k}(s)$ minimises the number of t-disjoint pairs among all families of $\binom{n}{k} - s$ sets in $\binom{[n]}{k}$.*

It remains to prove the proposition and lemmas. We begin with a proof of Proposition 4.4.1. The strategy will be very similar to that of Theorem 4.1.7; assuming the extremal family $\mathcal{F}$ does not have $r-1$ full t-stars, we shall show there is some popular element (that is, an element contained in many sets of $\mathcal{F}$). From this we will deduce the existence of a small cover, and shall show that either case (ii) or case (iii) must hold.

Proof of Proposition 4.4.1. We may assume that $r \geq 2$, since if $r = 1$, then case (i) is trivially satisfied. We first estimate the number of t-intersecting pairs in $\mathcal{L}_{n,k}(s)$, so that we have a lower bound on $\mathrm{int}_t(\mathcal{F})$ for any extremal family $\mathcal{F}$.

Note that $\mathcal{L}_{n,k}(s)$ consists of $r-1$ full t-stars, with the remaining sets forming a partial t-star; suppose there are $\alpha\binom{n-t}{k-t}$ such sets. Since there are $\binom{n-t-1}{k-t-1} = o(n^{k-t})$ sets common to any two t-stars, it follows that $s = (r-1+\alpha)\binom{n-t}{k-t} + o(n^{k-t})$.

Now any two sets in the same t-star are t-intersecting, while (4.8) shows that a set is t-intersecting with $o(n^{k-t})$ sets from the other t-stars. Hence for any extremal family $\mathcal{F}$ we have the bound

$$\begin{aligned}\mathrm{int}_t(\mathcal{F}) \geq \mathrm{int}_t(\mathcal{L}_{n,k}(s)) &= (r-1)\binom{\binom{n-t}{k-t}}{2} + \binom{\alpha\binom{n-t}{k-t}}{2} + o(n^{2(k-t)}) \\ &= \frac{r-1+\alpha^2}{2}\binom{n-t}{k-t}^2 + o(n^{2(k-t)}).\end{aligned}$$

Suppose $\mathcal{F}$ contains p full t-stars. If $p = r-1$, then case (i) holds, and we are done. Hence we may assume $0 \leq p \leq r-2$. Let $\mathcal{F}_1$ be the union of the p full t-stars, and let $\mathcal{F}_2 = \mathcal{F} \setminus \mathcal{F}_1$ be the remaining sets.

By the same reasoning as above, we must have $|\mathcal{F}_1| = p\binom{n-t}{k-t} + o(n^{k-t})$, and $\mathrm{int}_t(\mathcal{F}_1) = \frac{1}{2}p\binom{n-t}{k-t}^2 + o(n^{2(k-t)})$. No set $F \in \mathcal{F}_2$ is in any of the t-stars of $\mathcal{F}_1$, and so (4.8) gives $\mathrm{int}_t(\mathcal{F}_1, \mathcal{F}_2) = |\mathcal{F}_2| \cdot o(n^{k-t}) = o(n^{2(k-t)})$. Thus $\mathrm{int}_t(\mathcal{F}) = \mathrm{int}_t(\mathcal{F}_1) + \mathrm{int}_t(\mathcal{F}_1, \mathcal{F}_2) + \mathrm{int}_t(\mathcal{F}_2) = \frac{1}{2}p\binom{n-t}{k-t}^2 + \mathrm{int}_t(\mathcal{F}_2) + o(n^{2(k-t)})$, and hence we must have

$$\mathrm{int}_t(\mathcal{F}_2) \geq \frac{r-p-1+\alpha^2}{2}\binom{n-t}{k-t}^2 + o(n^{2(k-t)}) = \Omega(n^{2(k-t)}).$$

We shall now deduce the existence of a t-cover of size $r-p-1$ or $r-p$ for $\mathcal{F}_2$, and then show that we must fall into case (ii) or (iii). The first step is to find a t-set that is in many members of $\mathcal{F}_2$. Note that none of the t-stars in $\mathcal{F}_2$ are full, and hence we may shift sets in $\mathcal{F}_2$.

Claim 4.4.7. *There is some set $X_1 \in \binom{[n]}{t}$ with $|\mathcal{F}_2(X_1)| = \Omega(n^{k-t})$.*

Proof. Let $X_1 \in \binom{[n]}{t}$ be the set maximising $|\mathcal{F}_2(X)|$. We have

$$\begin{aligned}\mathrm{int}_t(\mathcal{F}_2) - \frac{1}{2}|\mathcal{F}| = \frac{1}{2}\sum_{F\in\mathcal{F}_2}\mathrm{int}_t(F,\mathcal{F}_2) = \frac{1}{2}\sum_{F\in\mathcal{F}_2}\left|\cup_{X\in\binom{F}{t}}\mathcal{F}_2(X)\right| \leq \frac{1}{2}\sum_{F\in\mathcal{F}_2}\sum_{X\in\binom{F}{t}}|\mathcal{F}_2(X)| \\ \leq \frac{1}{2}\sum_{F\in\mathcal{F}_2}\binom{k}{t}|\mathcal{F}_2(X_1)| = \frac{1}{2}\binom{k}{t}|\mathcal{F}_2|\,|\mathcal{F}_2(X_1)|\,.\end{aligned}$$

Since $|\mathcal{F}_2| = (r-p-1+\alpha)\binom{n-t}{k-t} = O(n^{k-t})$, and $\mathrm{int}_t(\mathcal{F}_2) = \Omega(n^{2(k-t)})$, it follows that $|\mathcal{F}_2(X_1)| = \Omega(n^{k-t})$, as desired. □

This allows us to find a small t-cover.

Claim 4.4.8. $\mathcal{X} = \left\{X \in \binom{[n]}{t} : |\mathcal{F}_2(X)| \geq \frac{1}{2\binom{k}{t}}|\mathcal{F}_2(X_1)|\right\}$ *is a* t*-cover for* $\mathcal{F}_2$*.*

Proof. Suppose not. Then there is some $F \in \mathcal{F}$ such that for all $X \in \binom{F}{t}$, $|\mathcal{F}_2(X)| < \frac{1}{2\binom{k}{t}}|\mathcal{F}_2(X_1)|$. Thus $\mathrm{int}_t(F,\mathcal{F}_2) \leq \sum_{X\in\binom{F}{t}}|\mathcal{F}_2(X)| < \frac{1}{2}|\mathcal{F}_2(X_1)|$. Since F has $o(n^{k-t})$ t-intersecting pairs in $\mathcal{F}_1$, it follows that $\mathrm{int}_t(F,\mathcal{F}) \leq \frac{1}{2}|\mathcal{F}_2(X_1)| + o(n^{k-t})$.

If we were to replace F with some set G containing X_1, which is possible as $\mathcal{F}(X_1)$ is not a full t-star, then we would create at least $|\mathcal{F}_2(X_1)|$ t-intersecting pairs. Since $|\mathcal{F}_2(X_1)| = \Omega(n^{k-t})$, it follows that $\mathrm{int}_t(G,\mathcal{F}) > \mathrm{int}_t(F,\mathcal{F})$, which contradicts $\mathcal{F}$ being optimal.

Hence $\mathcal{X}$ must be a t-cover for $\mathcal{F}_2$, as claimed. □

Claim 4.4.9. $|\mathcal{X}| = O(1)$.

Proof. We have

$$\binom{k}{t}|\mathcal{F}_2| = \sum_{F\in\mathcal{F}_2}\left|\binom{F}{t}\right| = \sum_{X\in\binom{[n]}{t}}|\mathcal{F}_2(X)| \geq \sum_{X\in\mathcal{X}}|\mathcal{F}_2(X)| \geq \frac{1}{2\binom{k}{t}}|\mathcal{F}_2(X_1)|\,|\mathcal{X}|\,.$$

Since $|\mathcal{F}_2| = O(n^{k-t})$ and $|\mathcal{F}_2(X_1)| = \Omega(n^{k-t})$, it follows that $|\mathcal{X}| = O(1)$, as claimed. □

Hence we can write $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$, where $m = O(1)$. Note that there are at most $\binom{n-t-1}{k-t-1} = o(n^{k-t})$ sets in common between any two stars, while the number of sets each t-star contains is at least $\frac{1}{2\binom{k}{t}}|\mathcal{F}_2(X_1)| = \Omega(n^{k-t})$. Thus in what follows, we consider only those sets in exactly one t-star $\mathcal{F}_2(X_i)$, and shall only lose $o(n^{2(k-t)})$ t-intersecting pairs.

Claim 4.4.10. *For all* $1 \le i < j \le m$, $|\mathcal{F}_2(X_i)| = |\mathcal{F}_2(X_j)| + o(n^{k-t})$.

Proof. Consider a set $F \in \mathcal{F}_2(X_i)$. F is t-intersecting with all sets in $\mathcal{F}_2(X_i)$, and, by (4.8), t-disjoint from almost all other sets. Thus $\mathrm{int}_t(F, \mathcal{F}) = |\mathcal{F}_2(X_i)| + o(n^{k-t})$. If we were instead to replace F with a set G containing X_j, which is possible as $\mathcal{F}(X_j)$ is not a full t-star, then we would create at least $|\mathcal{F}_2(X_j)|$ new t-intersecting pairs. Since $\mathcal{F}$ is optimal, we must have $|\mathcal{F}_2(X_i)| + o(n^{k-t}) \ge |\mathcal{F}_2(X_j)|$.

By symmetry, it follows that $|\mathcal{F}_2(X_i)| = |\mathcal{F}_2(X_j)| + o(n^{k-t})$. □

Recall that we had $|\mathcal{F}_2| = (r-p-1+\alpha)\binom{n-t}{k-t} + o(n^{k-t})$. By Claim 4.4.10, it follows that these sets are almost equally distributed between the m t-stars in the t-cover $\mathcal{X}$, and so $|\mathcal{F}_2(X_i)| = \frac{r-p-1+\alpha}{m}\binom{n-t}{k-t} + o(n^{k-t})$ for each $1 \le i \le m$. Moreover, since $|\mathcal{F}_2(X_i)| \le \binom{n-t}{k-t}$, we must have $m \ge r-p-1$ if $\alpha = o(1)$, or $m \ge r-p$ if $\alpha = \Omega(1)$.

We can now estimate $\mathrm{int}_t(\mathcal{F}_2)$. We know every set belonging only to the t-star $\mathcal{F}_2(X_i)$ contributes $|\mathcal{F}_2(X_i)| + o(n^{k-t})$ t-intersecting pairs, while there are only $o(n^{2(k-t)})$ t-intersecting pairs from sets in multiple t-stars. Thus

$$\begin{aligned}
\mathrm{int}_t(\mathcal{F}_2) &= \frac{1}{2}\sum_{F\in\mathcal{F}_2} \mathrm{int}_t(F,\mathcal{F}_2) + \frac{1}{2}|\mathcal{F}| = \frac{1}{2}\sum_{i=1}^{m}\sum_{F\in\mathcal{F}_2(X_i)} \mathrm{int}_t(F,\mathcal{F}_2) + o(n^{2(k-t)}) \\
&= \frac{1}{2}\sum_{i=1}^{m} |\mathcal{F}_2(X_i)|\left(|\mathcal{F}_2(X_i)| + o(n^{k-t})\right) + o(n^{2(k-t)}) = \frac{1}{2}\sum_{i=1}^{m}|\mathcal{F}_2(X_i)|^2 + o(n^{2(k-t)}) \\
&= \frac{(r-p-1+\alpha)^2}{2m}\binom{n-t}{k-t}^2 + o(n^{2(k-t)})
\end{aligned}$$

On the other hand, we had the bound

$$\mathrm{int}_t(\mathcal{F}_2) \ge \frac{r-p-1+\alpha^2}{2}\binom{n-t}{k-t}^2 + o(n^{2(k-t)}).$$

Comparing the two, we must have

$$\frac{(r-p-1+\alpha)^2}{2m} \ge \frac{r-p-1+\alpha^2}{2} + o(1). \tag{4.10}$$

Note that we can write $\frac{r-p-1+\alpha^2}{2} = \frac{1}{2}\sum_{i=1}^{m} x_i^2$, where

$$x_i = \begin{cases} 1 & 1 \le i \le r-p-1 \\ \alpha & i = r-p \\ 0 & r-p+1 \le i \le m \end{cases}.$$

Let $\overline{x} = \frac{1}{m}\sum_{i=1}^{m} x_i = \frac{r-p-1+\alpha}{m}$. With this definition, we then have $\frac{(r-p-1+\alpha)^2}{2m} = \frac{1}{2}m\overline{x}^2$. Since

$$\sum_{i=1}^{m} x_i^2 = m\overline{x}^2 + \sum_{i=1}^{m} (x_i - \overline{x})^2,$$

for (4.10) to hold, we must have $\sum_{i=1}^{m}(x_i - \overline{x})^2 = o(1)$, and thus $x_i = \overline{x} + o(1)$ for all $1 \le i \le m$.

Since $x_1 = 1$, $x_{r-p} = \alpha$, and $x_{r-p+1} = 0$, we must have $m \le r-p$. Recalling our earlier bound $m \ge r-p-1$, there are only two possibilities. We could have $m = r-p$ and $\alpha = 1 - o(1)$. In this case, each of the $r-p$ t-stars in $\mathcal{F}_2$ has size $\frac{r-1-p+\alpha}{m}\binom{n-t}{k-t} + o(n^{k-t}) = (1-o(1))\binom{n-t}{k-t}$. Combined with the p full t-stars in $\mathcal{F}_1$, we see that $\mathcal{F}$ consists of r almost full t-stars, and so we are in case (ii).

The other possible solution is to have $m = r-p-1$, with $\alpha = o(1)$. This implies $\mathcal{F}_2$ consists of $r-1-p$ almost full t-stars, which, combined with the p full t-stars of $\mathcal{F}_1$, means $\mathcal{F}$ falls under case (iii). This completes the proof of Proposition 4.4.1. □

We complete this section by proving the three lemmas. First we show that unions of lexicographical stars contain the fewest sets.

Proof of Lemma 4.4.2. Note that the first r t-stars in the lexicographical ordering have centres $Y_i = \{1, 2, \ldots, t-1, t+i-1\}$, $1 \le i \le r$, and their union has size $s = \binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}$. Letting $\mathcal{L} = \mathcal{L}_{n,k}\left(\binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}\right)$, note that for any set $I \subset [r]$, since $|\cup_{i\in I} Y_i| = t + |I| - 1$, we have $|\cap_{i\in I} \mathcal{L}(Y_i)| = \binom{n-t-|I|+1}{k-t-|I|+1}$. Thus, by Inclusion-Exclusion,

$$\begin{aligned} |\mathcal{L}| = |\cup_{i=1}^{r} \mathcal{L}(Y_i)| &= \sum_i |\mathcal{L}(Y_i)| - \sum_{i_1<i_2} |\mathcal{L}(Y_{i_1}) \cap \mathcal{L}(Y_{i_2})| + O(n^{k-t-2}) \\ &= r\binom{n-t}{k-t} - \binom{r}{2}\binom{n-t-1}{k-t-1} + O(n^{k-t-2}). \end{aligned}$$

Now we consider the size of $\mathcal{F}$. Suppose $\mathcal{F}$ is the union of the r full t-stars with centres $\{X_1, \dots, X_r\}$. We have

$$|\mathcal{F}| = |\cup_{i=1}^{r} \mathcal{F}(X_i)| \geq \sum_{i=1}^{r} |\mathcal{F}(X_i)| - \sum_{i_1<i_2} |\mathcal{F}(X_{i_1}) \cap \mathcal{F}(X_{i_2})| = r\binom{n-t}{k-t} - \sum_{i_1<i_2} |\mathcal{F}(X_{i_1}) \cap \mathcal{F}(X_{i_2})|.$$

For every $i_1 < i_2$ we have $|\mathcal{F}(X_{i_1}) \cap \mathcal{F}(X_{i_2})| = \binom{n-|X_{i_1} \cup X_{i_2}|}{k-|X_{i_1} \cup X_{i_2}|}$. If $|X_{i_1} \cap X_{i_2}| \leq t-2$, then $|X_{i_1} \cup X_{i_2}| \geq t+2$. Hence $|\mathcal{F}(X_{i_1}) \cap \mathcal{F}(X_{i_2})| = O(n^{k-t-2})$, and so

$$|\mathcal{F}| \geq r\binom{n-t}{k-t} - \binom{r}{2}\binom{n-t-1}{k-t-1} + O(n^{k-t-2}) > |\mathcal{L}|.$$

Hence we must have $|X_{i_1} \cap X_{i_2}| = t-1$ for all $i_1 < i_2$.

Now, by Inclusion-Exclusion, we have

$$|\mathcal{F}| - r\binom{n-t}{k-t} + \binom{r}{2}\binom{n-t-1}{k-t-1} = \sum_{\substack{I \subset [r] \\ |I| \geq 3}} (-1)^{|I|+1} |\cap_{i \in I} \mathcal{F}(X_i)|.$$

For any set F containing $a \geq 3$ sets X_i, the contribution to the right-hand side is

$$\sum_{b=3}^{a} (-1)^{b+1} \binom{a}{b} = (1-1)^a + 1 - a + \binom{a}{2} = 1 - a + \binom{a}{2} \geq 1.$$

If we have some $i_1 < i_2 < i_3$ with $|X_{i_1} \cup X_{i_2} \cup X_{i_3}| = t+1$, then we would have $\binom{n-t-1}{k-t-1}$ sets containing X_{i_1}, X_{i_2} and X_{i_3}. By the preceding equation, we then have

$$|\mathcal{F}| \geq r\binom{n-t}{k-t} - \binom{r}{2}\binom{n-t-1}{k-t-1} + \binom{n-t-1}{k-t-1} > |\mathcal{L}|.$$

Hence we may assume $|X_{i_1} \cup X_{i_2} \cup X_{i_3}| \geq t+2$ for all $i_1 < i_2 < i_3$. Since we must have $|X_{i_1} \cap X_{i_2}| = t-1$ for all $i_1 < i_2$, this implies all of the sets X_i share a common $(t-1)$-set, and hence $\mathcal{F}$ is isomorphic to $\mathcal{L}$, as desired. □

The next lemma showed that when adding a set to r full t-stars, the lexicographical stars minimise the number of new t-disjoint pairs.

Proof of Lemma 4.4.3. $\mathcal{L}$ is the union of the t-stars with centres $\{Y_1, Y_2, \dots, Y_r\}$, as in Lemma 4.4.2. Since all these sets, and L, contain $[t-1]$, it is easy to bound $\mathrm{dp}_t(L, \mathcal{L})$

from above by

$$\sum_{i=1}^{r}\mathrm{dp}_t(L,\mathcal{L}(Y_i)) - \sum_{i_1<i_2}\mathrm{dp}_t(L,\mathcal{L}(Y_{i_1})\cap\mathcal{L}(Y_{i_2})) + \sum_{i_1<i_2<i_3}\mathrm{dp}_t(L,\mathcal{L}(Y_{i_1})\cap\mathcal{L}(Y_{i_2})\cap\mathcal{L}(Y_{i_3}))$$
$$= r\binom{n-k-1}{k-t} - \binom{r}{2}\binom{n-k-2}{k-t-1} + O(n^{k-t-2}).$$

On the other hand, we have

$$\mathrm{dp}_t(F,\mathcal{F}) \geq \sum_{i=1}^{r}\mathrm{dp}_t(F,\mathcal{F}(X_i)) - \sum_{i_1<i_2}\mathrm{dp}_t(F,\mathcal{F}(X_{i_1})\cap\mathcal{F}(X_{i_2})).$$

The first term can be evaluated as follows. Since

$$\mathrm{dp}_t(F,\mathcal{F}(X_i)) = \sum_{a=0}^{t-1-|F\cap X_i|}\binom{k-|F\cap X_i|}{a}\binom{n-k-t+|F\cap X_i|}{k-t-a},$$

if $|F\cap X_i| = t-1$ we have $\mathrm{dp}_t(F,\mathcal{F}(X_i)) = \binom{n-k-1}{k-t}$, while $\mathrm{dp}_t(F,\mathcal{F}(X_i)) \geq \binom{n-k-2}{k-t} + (k-t+2)\binom{n-k-2}{k-t-1} = \binom{n-k-1}{k-t} + (k-t+1)\binom{n-k-2}{k-t-1}$ otherwise. Moreover, for every $i_1 < i_2$ we have the bound $\mathrm{dp}_t(F,\mathcal{F}(X_{i_1})\cap\mathcal{F}(X_{i_2})) \leq |\mathcal{F}(X_{i_1})\cap\mathcal{F}(X_{i_2})| \leq \binom{n-t-1}{k-t-1}$. Hence, if $|F\cap X_i| \leq t-2$ for some i,

$$\mathrm{dp}_t(F,\mathcal{F}) \geq r\binom{n-k-1}{k-t} + (k-t+1)\binom{n-k-2}{k-t-1} - \binom{r}{2}\binom{n-t-1}{k-t-1}$$
$$= r\binom{n-k-1}{k-t} - \left(\binom{r}{2} - (k-t+1)\right)\binom{n-k-2}{k-t-1} + O(n^{k-t-2}) > \mathrm{dp}_t(L,\mathcal{L}).$$

Thus we may assume $|F\cap X_i| = t-1$ for all i. Given this condition, it follows that

$$\mathrm{dp}_t(F,\mathcal{F}(X_{i_1})\cap\mathcal{F}(X_{i_2})) = \begin{cases} \binom{n-k-2}{k-t-1} & \text{if } F\cap X_{i_1} = F\cap X_{i_2} \\ 0 & \text{otherwise, since } |F\cap(X_{i_1}\cup X_{i_2})| \geq t \end{cases}.$$

Hence, in order to have $\mathrm{dp}_t(F,\mathcal{F}) \leq \mathrm{dp}_t(L,\mathcal{L}) = r\binom{n-k-1}{k-t} - \binom{r}{2}\binom{n-k-2}{k-t-1} + O(n^{k-t-2})$, we must have $F\cap X_{i_1} = F\cap X_{i_2}$ for all $i_1 < i_2$. This implies that F shares a common $(t-1)$-set with all the sets X_i, and thus $\mathcal{F}\cup\{F\}$ is isomorphic to $\mathcal{L}\cup\{L\}$, as required. □

The final lemma showed that the union of any r full t-stars contains at least as many disjoint pairs as the initial segment of the lexicographical ordering with the same number of sets.

Proof of Lemma 4.4.4. We shall find it more convenient to count the number of t-intersecting pairs. Suppose $\mathcal{F}$ is the union of the full t-stars with centres $\{X_1, X_2, \ldots, X_r\} \subset \binom{[n]}{t}$. By Lemma 4.4.2, it follows that $\mathcal{L} = \mathcal{L}_{n,k}(|\mathcal{F}|)$ consists of the full t-stars with centres $\{Y_1, Y_2, \ldots, Y_r\}$, possibly with some additional sets in an $(r+1)$st t-star with centre Y_{r+1}, where $Y_i = \{1, 2, \ldots, t-1, t-1+i\}$. Note that in this setting we have $|\mathcal{F}| = |\mathcal{L}|$.

We first show that if $|X_i \cap X_j| \leq t-2$ for some $1 \leq i < j \leq r$, then the r full t-stars of $\mathcal{L}$ alone contain more t-intersecting pairs than $\mathcal{F}$. We have

$$\begin{aligned}\mathrm{int}_t(\mathcal{F}) &\leq \sum_{i=1}^{r} \mathrm{int}_t(\mathcal{F}(X_i)) + \sum_{i<j} \mathrm{int}_t(\mathcal{F}(X_i) \setminus \mathcal{F}(X_j), \mathcal{F}(X_j) \setminus \mathcal{F}(X_i)) \\ &= r\binom{\binom{n-t}{k-t}}{2} + r\binom{n-t}{k-t} + \sum_{i<j} \mathrm{int}_t(\mathcal{F}(X_i) \setminus \mathcal{F}(X_j), \mathcal{F}(X_j) \setminus \mathcal{F}(X_i)), \qquad (4.11)\end{aligned}$$

where the inequality is due to the fact that t-intersecting pairs involving sets in multiple t-stars are overcounted.

First suppose $|X_i \cap X_j| = t-1$. Given a set $F \in \mathcal{F}(X_i) \setminus \mathcal{F}(X_j)$, we wish to bound how many sets $G \in \mathcal{F}(X_j) \setminus \mathcal{F}(X_i)$ can be t-intersecting with F. Since $X_i \cap X_j \subset F \cap G$, we require G to contain one additional element of F. However, this element cannot be from X_i, as then we would have $G \in \mathcal{F}(X_i)$. Thus there are $k-t$ choices for this additional element. Given that G already contains X_j, there are $\binom{n-t-1}{k-t-1}$ ways to choose the remaining elements of G. Hence there can be at most $(k-t)\binom{n-t-1}{k-t-1}$ such sets G, giving

$$\begin{aligned}\mathrm{int}_t(\mathcal{F}(X_i) \setminus \mathcal{F}(X_j), \mathcal{F}(X_j) \setminus \mathcal{F}(X_i)) &\leq (k-t)\binom{n-t-1}{k-t-1} |\mathcal{F}(X_i) \setminus \mathcal{F}(X_j)| \\ &\leq (k-t)\binom{n-t}{k-t}\binom{n-t-1}{k-t-1}.\end{aligned}$$

Now suppose $|X_i \cap X_j| \leq t-2$. There are two types of $F \in \mathcal{F}(X_i) \setminus \mathcal{F}(X_j)$: those with $F \cap X_j = X_i \cap X_j$, and those with $(F \setminus X_i) \cap X_j \neq \emptyset$. In the first case, note that for $G \in \mathcal{F}(X_j) \setminus \mathcal{F}(X_i)$ to be t-intersecting with F, G must contain at least 2 elements from F in addition to X_j. Hence there are at most $\binom{k}{2}\binom{n-t-2}{k-t-2}$ such sets. In the second case, note that there are at most $t\binom{n-t-1}{k-t-1}$ such sets F, as we can choose at most t elements from $X_j \setminus X_i$ for F to contain, and then there are $\binom{n-t-1}{k-t-1}$ ways to choose the remaining elements for F. For

each such F, in order for $G \in \mathcal{F}(X_j) \setminus \mathcal{F}(X_i)$ to be t-intersecting with F, G must contain some element of F in addition to $F \cap X_j$. There are at most k choices for this element, with $\binom{n-t-1}{k-t-1}$ ways to complete G. Thus there are at most $kt\binom{n-t-1}{k-t-1}^2$ t-intersections of this type. Hence we upper-bound $\mathrm{int}_t(\mathcal{F}(X_i) \setminus \mathcal{F}(X_j), \mathcal{F}(X_j) \setminus \mathcal{F}(X_i))$ by

$$\binom{k}{2}\binom{n-t}{k-t}\binom{n-t-2}{k-t-2} + kt\binom{n-t-1}{k-t-1}^2 = O(n^{2k-2t-2}).$$

Substituting these bounds into (4.11), if we have d pairs $\{i,j\}$ with $|X_i \cap X_j| \leq t-2$, we have

$$\mathrm{int}_t(\mathcal{F}) \leq r\binom{\binom{n-t}{k-t}}{2} + \left(\binom{r}{2} - d\right)(k-t)\binom{n-t}{k-t}\binom{n-t-1}{k-t-1} + O(n^{2k-2t-2}). \qquad (4.12)$$

We now provide a lower bound for $\mathrm{int}_t(\mathcal{L})$, considering only the r full t-stars. There are two types of t-intersecting pairs: those from within a single t-star, and those between two t-stars. Within each of the r t-stars, every pair of sets is t-intersecting, and hence there are at least $r\binom{\binom{n-t}{k-t}}{2}$ pairs of the first type.

Now we count the second type of pairs, and consider two t-stars with centres Y_i and Y_j. Since $Y_i \cap Y_j = [t-1]$, any sets $G_i \in \mathcal{L}(Y_i)$ and $G_j \in \mathcal{L}(Y_j)$ must have at least $t-1$ elements in common. They will therefore be t-intersecting if they have one more common element. Moreover, this common element should not be from $Y_i \cup Y_j$, as otherwise the two sets are in fact in the same t-star, making this pair of the first type. For each fixed G_i, there are therefore $k-t$ choices for the final intersection, and $\binom{n-t-1}{k-t-1}$ choices for corresponding sets G_j.

There is some slight error in this calculation, as, for example, pairs involving sets contained in multiple t-stars are overcounted. However, it is easy to check that the error is of order $n^{2k-2t-2}$, and thus a lower-order term. This gives

$$\mathrm{int}_t(\mathcal{L}) \geq r\binom{\binom{n-t}{k-t}}{2} + (k-t)\binom{r}{2}\binom{n-t}{k-t}\binom{n-t-1}{k-t-1} + O(n^{2k-2t-2}).$$

Comparing this to (4.12), we find that unless $d = 0$, we must have $\mathrm{int}_t(\mathcal{L}) > \mathrm{int}_t(\mathcal{F})$, as desired. It remains to consider the case when $|X_i \cap X_j| = t-1$ for all $1 \leq i < j \leq r$.

There are only two possibilities. In the first, all the sets X_i share $t-1$ elements in common, in which case $\mathcal{F}$ is isomorphic to $\mathcal{L}$. The second case, up to isomorphism, is when $r \leq t+1$, and $X_i \in \binom{[t+1]}{t}$. Note that if $1 \leq r \leq 2$, the two constructions are isomorphic, so we may assume $r \geq 3$.

In this case, as we know the exact structure of both constructions, we are able to compute the number of intersecting pairs rather more precisely. We begin with $\mathcal{F}$, the union of r full t-stars with centres from $\binom{[t+1]}{t}$.

$\mathcal{F}$ contains all $\binom{n-t-1}{k-t-1}$ sets containing $[t+1]$, and then $r\binom{n-t-1}{k-t}$ sets that meet $[t+1]$ in t elements. The sets containing $[t+1]$ are t-intersecting with all other sets in $\mathcal{F}$.

On the other hand, if $F \in \mathcal{F}(X_i)$ is such that $F \cap [t+1] = X_i$, then there are three types of sets in $\mathcal{F}$ that can be t-intersecting with F:

(i) a set containing $[t+1]$,

(ii) a set whose intersection with $[t+1]$ is precisely X_i, or

(iii) a set whose intersection with $[t+1]$ is X_j for some $j \neq i$.

There are $\binom{n-t-1}{k-t-1}$ sets of type (i) and $\binom{n-t-1}{k-t}$ sets of type (ii). For a set to be of type (iii), it must contain some X_j, not contain X_i, and then meet F in some element of $F \setminus X_i$. For each choice of j, the set should contain the t elements of X_j, not the single element in $X_i \setminus X_j$, and should not avoid the remaining $k-t$ elements of F. Hence there are $\binom{n-t-1}{k-t} - \binom{n-k-1}{k-t}$ such sets.

Putting this all together, we find

$$2\mathrm{int}_t(\mathcal{F}) - |\mathcal{F}| = \sum_{F \in \mathcal{F}} \mathrm{int}_t(F, \mathcal{F}) = I_1 + I_2, \tag{4.13}$$

where

$$I_1 = \binom{n-t-1}{k-t-1}\left[\binom{n-t-1}{k-t-1} + r\binom{n-t-1}{k-t}\right],$$

and

$$I_2 = r\binom{n-t-1}{k-t}\left[\binom{n-t-1}{k-t-1} + \binom{n-t-1}{k-t} + (r-1)\left(\binom{n-t-1}{k-t} - \binom{n-k-1}{k-t}\right)\right].$$

We now turn our attention to $\mathcal{L}$.

First observe that we have r full stars, with centres $\{Y_1, Y_2, \ldots, Y_r\}$. The remaining sets fall into an $(r+1)$st star with centre Y_{r+1}. To avoid overcounting, we shall partition $\mathcal{L}$ into the subfamilies $\mathcal{L}^*(i) = \{L \in \mathcal{L} : \min(L \setminus [t-1]) = t-1+i\}$, $1 \leq i \leq r+1$; that is, $L \in \mathcal{L}^*(i)$ if $\mathcal{L}(Y_i)$ is the first t-star L is in.

For $1 \leq i \leq r$, $\mathcal{L}^*(i)$ consists of all sets containing $[t-1] \cup \{t-1+i\}$, but disjoint from the interval $[t, t-2+i]$. Hence we have $|\mathcal{L}^*(i)| = \binom{n-t-i+1}{k-t}$. Summing up the telescoping binomial coefficients, we find the first r t-stars contain $\binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}$ sets. $\mathcal{L}^*(r+1)$ then contains enough sets to make $\mathcal{L}$ the right size, and so $|\mathcal{L}^*(r+1)|$ is equal to

$$|\mathcal{F}| - |\cup_{i=1}^{r} \mathcal{L}^*(i)| = \left[\binom{n-t-1}{k-t-1} + r\binom{n-t-1}{k-t}\right] - \left[\binom{n-t+1}{k-t+1} - \binom{n-t-r+1}{k-t+1}\right].$$

Note that all the subfamilies $\mathcal{L}^*(i)$ are t-intersecting. Moreover, if $j < i$, and $L \in \mathcal{L}^*(i)$, then for a set $K \in \mathcal{L}^*(j)$ to be t-intersecting with L, it must contain $[t-1] \cup \{t-1+j\}$, be disjoint from the interval $[t, t-2+j]$, and contain one of the $k-t+1$ elements in $L \setminus [t-1+j]$. Hence we have $\mathrm{int}_t(L, \mathcal{L}^*(j)) = \binom{n-t-j+1}{k-t} - \binom{n-k-j}{k-t}$. We can now count the number of t-intersecting pairs in $\mathcal{L}$:

$$\begin{aligned} 2\mathrm{int}_t(\mathcal{L}) - |\mathcal{L}| &= \sum_{L \in \mathcal{L}} \mathrm{int}_t(L, \mathcal{L}) = \sum_{i=1}^{r+1} \left(\mathrm{int}_t(\mathcal{L}^*(i), \mathcal{L}^*(i)) + 2\sum_{j<i} \mathrm{int}_t(\mathcal{L}^*(i), \mathcal{L}^*(j)) \right) \\ &= \sum_{i=1}^{r+1} |\mathcal{L}^*(i)|^2 + 2\sum_{j<i} |\mathcal{L}^*(i)| \left[\binom{n-t-j+1}{k-t} - \binom{n-k-j}{k-t}\right]. \end{aligned} \tag{4.14}$$

We now wish to show $\mathrm{int}_t(\mathcal{L}) \geq \mathrm{int}_t(\mathcal{F})$; that is, to show the quantity in (4.14) is greater than that in (4.13). To make this task easier, we shall rewrite all products of binomial coefficients in the form $\binom{n-t}{k-t}^2$, $\binom{n-t}{k-t}\binom{n-t}{k-t-1}$, or $\binom{n-t}{k-t-1}^2$, using the identities

$$\begin{aligned} \binom{m-a}{r} &= \binom{m}{r} - a\binom{m}{r-1} + \binom{a+1}{2}\binom{m}{r-2} + O(m^{r-3}) \text{ and} \\ \binom{n-t}{k-t}\binom{n-t}{k-t-2} &= \frac{n-k+1}{n-k+2} \cdot \frac{k-t-1}{k-t} \cdot \binom{n-t}{k-t-1}^2 \\ &= \frac{k-t-1}{k-t}\binom{n-t}{k-t-1}^2 + O(n^{2k-2t-3}). \end{aligned}$$

After performing the routine but tedious calculations, we find,

$$2\mathrm{int}_t(\mathcal{F}) - |\mathcal{F}| + O(n^{2k-2t-3}) = r\binom{n-t}{k-t}^2 + r(r-1)(k-t)\binom{n-t}{k-t}\binom{n-t}{k-t-1} - \left[\frac{1}{2}r(r-1)(k-t)^2 + 2r(r-1)(k-t) - \left(\frac{3r}{2}-1\right)(r-1)\right]\binom{n-t}{k-t-1}^2,$$

$$\text{and } 2\mathrm{int}_t(\mathcal{L}) - |\mathcal{L}| + O(n^{2k-2t-3}) = r\binom{n-t}{k-t}^2 + r(r-1)(k-t)\binom{n-t}{k-t}\binom{n-t}{k-t-1} - \left[\frac{1}{2}r(r-1)(k-t)^2 + 2r(r-1)(k-t) - \frac{1}{4}(r-1)^2(r^2+4)\right]\binom{n-t}{k-t-1}^2.$$

The coefficient of the leading term of the difference between the two constructions is thus

$$\frac{2\mathrm{int}_t(\mathcal{L}) - 2\mathrm{int}_t(\mathcal{F})}{\binom{n-t}{k-t-1}^2} - O\left(\frac{1}{n}\right) = \frac{1}{4}(r-1)^2(r^2+4) - \left(\frac{3r}{2}-1\right)(r-1) = \frac{1}{4}(r+1)r(r-1)(r-2),$$

which is at least 6 for $r \geq 3$. Hence we indeed find $\mathrm{int}_t(\mathcal{L}) > \mathrm{int}_t(\mathcal{F})$, as required. □

4.5 Further remarks and open problems

In this chapter, we have provided a partial solution to a problem of Ahlswede on the minimum number of disjoint pairs in set families. For small families, we verified Bollobás and Leader's conjecture by showing that the initial segment of the lexicographical ordering is optimal. By considering the complementary set families, this also resolves the problem for very large set families. However, it remains to determine which families are optimal in between.

When $k = 2$, Ahlswede and Katona showed that the optimal family was always either a union of stars or its complement. For $k \geq 3$, Bollobás and Leader suggest a larger family of possible extremal families. We note that for families of size $s = \frac{1}{2}\binom{n}{k}$, the lexicographical family is at least near-optimal. A straightforward calculation shows $\mathrm{dp}(n,k,s) \leq \mathrm{dp}(\mathcal{L}_{n,k}(s)) \leq \frac{1}{2}\left(1 - \frac{2^{1/k}k^2}{n} + O(n^{-2})\right)s^2$. On the other hand, exploiting the connection to the Kneser graph, we can use spectral techniques to obtain the bound $\mathrm{dp}(n,k,s) \geq \frac{1}{2}\left(1 - \frac{k(k+2)}{n} + O(n^{-2})\right)s^2$.

While our focus has been showing that a family with more than $\binom{n-1}{k-1}$ sets must contain many disjoint pairs, a closely related problem is to determine whether such a family must

have any sets disjoint from many other sets. This type of question has been studied before in other settings. For example, when one is considering the number of triangles in a graph, Erdős showed in [Erd62a] that any graph with $\left\lfloor \frac{n^2}{4} \right\rfloor + 1$ edges must contain an edge in at least $\frac{n}{6} + o(n)$ triangles. It is well-known and easy to see that the hypercube, a graph whose vertices are subsets of $[n]$, with two vertices adjacent if they are comparable and differ in exactly one element, has independence number 2^{n-1}. In [CFG88], it is proved that any induced subgraph on $2^{n-1}+1$ vertices contains a vertex of degree at least $(\frac{1}{2} + o(1)) \log_2 n$. It is an open problem to determine whether or not this bound is tight (the corresponding upper bound is $O(\sqrt{n})$), and the answer to this question has ramifications in theoretical computer science.

In the context of the Erdős–Ko–Rado Theorem, it is trivial to show that in a family of $\binom{n-1}{k-1} + 1$ sets, there must be a set disjoint from at least $\frac{1}{2}\left(1 - \frac{k^3}{n}\right)\binom{n-1}{k-1}$ other sets. Indeed, by the Erdős–Ko–Rado theorem, there exists a pair F_1, F_2 of disjoint sets. At most $k^2\binom{n-2}{k-2} < \frac{k^3}{n}\binom{n-1}{k-1}$ sets can intersect both F_1 and F_2, and so either F_1 or F_2 must be disjoint from at least half of the remaining sets, resulting in the above bound. Furthermore, this is easily seen to be asymptotically tight, as one may take all sets containing $\{1, 2\}$, and then take half the remaining sets to contain 1, and half to contain 2. It may be of interest to obtain sharper estimates for this problem, especially as the aforementioned construction shows that this is closely related to the original problem when $s \approx \frac{1}{2}\binom{n}{k}$, since one should choose the sets containing 1 or 2 optimally. A related problem, that of determining the largest possible family where no set is disjoint from more than ℓ other sets, was studied in [GLP12].

We find most exciting the prospect of studying Erdős–Rademacher-type problems in other settings. In an earlier paper [DGS14b], we presented an Erdős–Rademacher-type strengthening of Sperner's Theorem, a problem that was also studied in [DGK14]. However, as one can investigate similar extensions for any extremal result, there is truly no end to the number of directions in which this project can be continued. We hope that further work of this nature will lead to many interesting results and a greater understanding of classical theorems in extremal combinatorics.

CHAPTER 5

Most probably intersecting hypergraphs

5.1 Introduction

A family of sets $\mathcal{F}$ is said to be *intersecting* if $F_1 \cap F_2 \neq \emptyset$ for all $F_1, F_2 \in \mathcal{F}$. A central result in extremal set theory is the Erdős-Ko-Rado theorem, which determines the largest size of an intersecting k-uniform family over $[n]$. Given this extremal result, one may then investigate the appearance of disjoint pairs in larger families of sets.

Recently Katona, Katona and Katona introduced a probabilistic version of this supersaturation problem. Given a set family $\mathcal{F}$, let $\mathcal{F}_p$ denote the random subfamily obtained by keeping each set independently with probability p. They asked, for a given p, n and m, which set families on $[n]$ with m sets maximise the probability of $\mathcal{F}_p$ forming an intersecting family. We study this problem for k-uniform set families. In the case $k = 2$, we determine the optimal graphs when they are not too dense. In the hypergraph setting, we provide an approximate structural result, and are able to determine the extremal hypergraphs exactly for some ranges of values of m. These mark the first general results for the probabilistic supersaturation problem for k-uniform set families.

Recall from Chapter 4 that the initial segments of the lexicographic order, where $A < B$ if $\min(A \Delta B) \in A$, minimise the number of disjoint pairs in small k-uniform set families. We write $\mathcal{L}_{n,k}(m)$ for the first m sets in $\binom{[n]}{k}$ under the lexicographic order. The complement of $\mathcal{L}_{n,k}(m)$ is isomorphic to the corresponding initial segment of the colexicographic order, where $A < B$ if $\max(A \Delta B) \in B$. We write $\mathcal{C}_{n,k}(m)$ for an initial segment of this order.

We now introduce the probabilistic supersaturation problem due to Katona, Katona and Katona, and then present our new results.

5.1.1 Probabilistic supersaturation

In 2012, Katona, Katona and Katona [KKK12] introduced a probabilistic measure of supersaturation for large families. Rather than minimising the total number of disjoint pairs in large families, they sought to maximise the probability of a random subfamily being intersecting. More formally, given a (not necessarily uniform) family $\mathcal{F}$ of sets, and some $p \in [0,1]$, let $\mathcal{F}_p$ denote the random subfamily of $\mathcal{F}$, where each set is retained independently with probability p. For a given $0 \leq m \leq 2^n$, they asked for the families $\mathcal{F}$ of m subsets of $[n]$ maximising $\mathbb{P}(\mathcal{F}_p \text{ is intersecting})$.

Clearly, if $\mathcal{F}$ is intersecting, then $\mathcal{F}_p$ must also be intersecting, and hence one should take an intersecting family if possible. Thus, as in the case of the counting supersaturation problem, one is interested in families larger than the extremal bound.

We observe here that the probabilistic problem is in fact stronger than the counting version described before. Indeed, note that by conditioning on the number of sets in $\mathcal{F}_p$, we have

$$\begin{aligned}\mathbb{P}(\mathcal{F}_p \text{ is intersecting}) &= \sum_{t=0}^{m} \mathbb{P}\left(\mathcal{F}_p \text{ is intersecting} \,||\mathcal{F}_p| = t\right) \mathbb{P}\left(|\mathcal{F}_p| = t\right) \\ &= \sum_{t=0}^{m} \operatorname{int}(\mathcal{F}, t) p^t (1-p)^{m-t}, \end{aligned} \tag{5.1}$$

where $\operatorname{int}(\mathcal{F}, t)$ denotes the number of intersecting subfamilies of $\mathcal{F}$ of size t. In particular, it follows that $\operatorname{int}(\mathcal{F}, t) < 2^m$ for all t. If we take $p = o(2^{-m})$, then $2^m p^3 = o(p^2)$, and so expanding the first few terms of the sum on the right-hand side gives

$$\mathbb{P}(\mathcal{F}_p \text{ is intersecting}) = (1-p)^m + mp(1-p)^{m-1} + \operatorname{int}(\mathcal{F}, 2) p^2 (1-p)^{m-2} + o(p^2).$$

This quantity is maximised if and only if the number of intersecting pairs of sets in $\mathcal{F}$ is maximised, and thus the number of disjoint pairs must be minimised. Hence a solution to the probabilistic problem for all values of p provides a solution to the counting problem as well.

Katona, Katona and Katona [KKK12] determined the extremal families for $m \leq 2^{n-1} + \binom{n-1}{\lceil (n-3)/2 \rceil}$. In particular, they showed that for all $0 \leq p \leq 1$, it is optimal to take all sets

of size larger than $\frac{n}{2}$, with the remaining sets of size $\lfloor \frac{n}{2} \rfloor$ chosen to minimise the number of disjoint pairs. They further conjectured the existence of a nested sequence $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_{2^n}$ of families such that $\mathcal{F}_m$ is the most probably intersecting family of size m.

In the same year, Russell [Rus12] provided some evidence towards this conjecture, by proving a result similar to Theorem 4.1.3, showing that there is a most probably intersecting family consisting of sets that are as large as possible. However, in a later paper with Walters [RW13], they used the non-nestedness of the extremal graphs in Theorem 4.1.6 to show that the most probably intersecting families are not nested for $\sum_{i=3}^{n} \binom{n}{i} \leq m \leq \sum_{i=2}^{n} \binom{n}{i}$.

While the above results hold for non-uniform families, much less was known in the uniform setting. By the Erdős-Ko-Rado theorem [EKR61], when $n \geq 2k$, the largest k-uniform intersecting family has size $\binom{n-1}{k-1}$, a bound attained when we take all sets containing some fixed element. We call such a structure a *star*; note that for $m \leq \binom{n-1}{k-1}$, $\mathcal{L}_{n,k}(m)$ is a star consisting of m sets containing 1.

Hence it follows that for $m \leq \binom{n-1}{k-1}$, $\mathcal{L}_{n,k}(m)$ is an intersecting family, and thus a most probably intersecting family. Once we have $m > \binom{n-1}{k-1}$, we can no longer take an intersecting family. Katona, Katona and Katona showed in [KKK12] that for $m = \binom{n-1}{k-1} + 1$, it is optimal to add any set to a full star, and thus $\mathcal{L}_{n,k}(m)$ is again optimal. By applying i,j-compressions, Russell and Walters [RW13] were able to show that for any m, there is a left-compressed most probably intersecting family, but were unable to show which compressed family is optimal.

5.1.2 Our results

We apply the shifting arguments developed in [DGS14a] to this probabilistic supersaturation for k-uniform set families. In the case $k = 2$, we show that the lexicographic order provides the most probably intersecting graphs for all sizes up to $c\binom{n}{2}$, with c approximately $\frac{1}{17}$.

Theorem 5.1.1. *For n, ℓ and m satisfying $n \geq 32\ell$ and $0 \leq m \leq \binom{n}{2} - \binom{n-\ell}{2}$, the lexicographic graph $\mathcal{L}_{n,2}(m)$ is the most probably intersecting graph on $[n]$ with m edges.*

When $k \geq 3$, the situation is rather more intricate. In Proposition 5.3.3, we provide

a rough structural description of most probably intersecting hypergraphs that are not too dense. While this approximation holds for all small families, it is only for some particular sizes that this result allows us to determine the most probably intersecting hypergraphs exactly, as given below.

Theorem 5.1.2. *Let k and ℓ be integers, and suppose $n \geq n_0(k,\ell)$ and $\binom{n}{k} - \binom{n-\ell}{k} \leq m \leq \binom{n}{k} - \binom{n-\ell}{k} + n - \ell - k + 1$. For this range of parameters, $\mathcal{L}_{n,k}(m)$ is the most probably intersecting k-uniform hypergraph on $[n]$ with m sets.*

We note that in both cases we actually prove something stronger, showing that $\mathcal{L}_{n,k}(m)$ simultaneously maximises the number of intersecting subfamilies of size t for all t, as stated in Propositions 5.2.1 and 5.3.4. Our proofs also extend to show that in these ranges, the most probably intersecting graph is essentially unique.

5.1.3 Outline and notation

The remainder of this chapter is organised as follows. In Section 5.2, we study the most probably intersecting graphs, proving Theorem 5.1.1. In Section 5.3, we extend these methods to hypergraphs, and prove Theorem 5.1.2. Finally, in Section 5.4, we provide some further remarks and open questions.

Our notation is fairly standard. We denote by $[n]$ the first n natural numbers, and for any set X, we write $\binom{X}{k}$ for the subsets of X of size k. $\mathcal{L}_{n,k}(m)$ represents the first m sets in $\binom{[n]}{k}$ in the lexicographic order, while $\mathcal{C}_{n,k}(m)$ is the corresponding initial segment of the colexicographic order; see the paragraph preceding Theorem 4.1.6 for a description of these orders.

If $\mathcal{F}$ is a k-uniform family of subsets of $[n]$, then for any vertex $i \in [n]$, we write d_i for its degree; that is, the number of sets containing i. A subset $X \subset [n]$ of elements *covers* $\mathcal{F}$ if for every set $F \in \mathcal{F}$, we have $F \cap X \neq \emptyset$. We let $\mathrm{int}(\mathcal{F}, t)$ denote the number of intersecting subfamilies of $\mathcal{F}$ of size t. We say that an intersecting family $\mathcal{G}$ is *trivially intersecting* if $\cap_{G\in\mathcal{G}} G \neq \emptyset$, and we call such a family a *star* with centre $\cap_{G\in\mathcal{G}} G$. Finally, we say a star with

centre i is *full* if it contains all $\binom{n-1}{k-1}$ sets containing i, and *almost-full* if it has $(1-o(1))\binom{n-1}{k-1}$ sets.

5.2 Intersecting graphs

In this section we prove Theorem 5.1.1, thus showing the initial segment of the lexicographic order is the most probably intersecting graph when the graphs in question are not too dense. We recall the statement below.

Theorem 5.1.3. *For n, ℓ and m satisfying $n \geq 32\ell$ and $0 \leq m \leq \binom{n}{2} - \binom{n-\ell}{2}$, the lexicographic graph $\mathcal{L}_{n,2}(m)$ is the most probably intersecting graph on $[n]$ with m edges.*

In order to prove this theorem, we use (5.1) to convert the problem into one of counting intersecting subgraphs of a given size. At the heart of the proof, therefore, is the following proposition, which shows that in this range of densities, $\mathcal{L}_{n,2}(m)$ maximises the number of intersecting subgraphs of size t for any t. Proposition 5.2.1 can be viewed as an extension of Theorem 4.1.6 to larger intersecting subgraphs.

Proposition 5.2.1. *Suppose $t \geq 0$, and n and ℓ satisfy $n \geq 2^{2+6/(t-1)}\ell$. Then, for any $0 \leq m \leq \binom{n}{2} - \binom{n-\ell}{2}$, the lexicographic graph $\mathcal{L}_{n,2}(m)$ maximises $\mathrm{int}(G,t)$ over all graphs G on $[n]$ with m edges.*

Note that in the case of graphs, there are only two possible intersecting structures: the star and, when $t = 3$, the triangle. We will show that there are relatively few triangles, and hence the number of intersecting subgraphs is essentially determined by the number of stars. By considering the central vertex of a star, we find that, for $t \geq 2$, the number of stars in a graph G is given by $\sum_{i \in V(G)} \binom{d_i}{t}$. As it is cleaner to first count only the stars, we separate this (main) case into the following proposition.

Proposition 5.2.2. *Suppose $t \geq 0$, and n and ℓ satisfy $n \geq 2^{2+6/(t-1)}\ell$. Then, for any $0 \leq m \leq \binom{n}{2} - \binom{n-\ell}{2}$, the lexicographic graph $\mathcal{L}_{n,2}(m)$ maximises $f(G,t) = \sum_i \binom{d_i}{t}$ over all graphs G on $[n]$ with m edges.*

We now begin by showing how Theorem 5.1.1 follows easily from Proposition 5.2.1.

Proof of Theorem 5.1.1. We wish to find a graph G on $[n]$ with m edges that maximises $\mathbb{P}(G_p \text{ is intersecting})$. Recall Equation (5.1):

$$\mathbb{P}(G_p \text{ is intersecting}) = \sum_{t=0}^{m} \text{int}(G,t)p^t(1-p)^{m-t}.$$

By Theorem 4.1.6 for $t = 2$ and Proposition 5.2.1 otherwise, among all graphs on $[n]$ with m edges, $\text{int}(G,t)$ is maximised by $\mathcal{L}_{n,2}(m)$ for all $t \geq 0$. Thus we have

$$\begin{aligned}\mathbb{P}(G_p \text{ is intersecting}) &= \sum_{t=0}^{m} \text{int}(G,t)p^t(1-p)^{m-t} \\ &\leq \sum_{t=0}^{m} \text{int}(\mathcal{L}_{n,2}(m),t)p^t(1-p)^{m-t} \\ &= \mathbb{P}(\mathcal{L}_{n,2}(m)_p \text{ is intersecting}),\end{aligned}$$

and so $\mathcal{L}_{n,2}(m)$ is the most probably intersecting graph, as claimed. □

In the remainder of this section, we seek to prove Proposition 5.2.1. We begin by dealing with the cleaner case of counting stars, namely Proposition 5.2.2.

Proof of Proposition 5.2.2. Our proof is by induction on $m + t$. Note that when $t = 0$, the statement is obvious, and for $t = 1$, $f(G,1) = \sum_i \binom{d_i}{1} = \sum_i d_i = 2m$, and is thus maximised by $\mathcal{L}_{n,2}(m)$ and, indeed, by any other graph with m edges.

For the case $t = 2$, note that since $\ell \leq 2^{-2-6/(t-1)}n \leq \frac{1}{4}n$, we have at most $\binom{n}{2} - \binom{\frac{3n}{4}}{2} < \frac{1}{2}\binom{n}{2} - \frac{1}{2}n$ edges. Hence, by Theorem 4.1.6, it is known that $\mathcal{L}_{n,2}(m)$ maximises the number of intersecting pairs of edges, which is precisely the quantity $f(G,2)$.

Moreover, when $m \leq n-1$, it is easy to see that $\mathcal{L}_{n,2}(m)$ is again optimal. Indeed, $f(G,t) = \sum_i \binom{d_i}{t}$ counts the number of t-edge stars in G. For $m \leq n-1$, $\mathcal{L}_{n,2}(m)$ is itself a star, and thus all subgraphs of t edges are stars. Clearly, $f(\mathcal{L}_{n,2}(m),t) = \binom{m}{t}$ is optimal.

Hence we may assume $t \geq 3$ and $m \geq n$. Suppose first that G is an extremal graph containing a full star; without loss of generality, we may assume it has all edges containing

the vertex 1. Let $\tilde{G}$ be the induced subgraph of G on the vertices $[n] \setminus \{1\}$. Note that for all $2 \leq i \leq n$, the degrees in $\tilde{G}$ are given by $\tilde{d}_i = d_i - 1$, as we lose the edge to 1. Thus we have

$$\begin{aligned} f(G,t) &= \sum_i \binom{d_i}{t} = \binom{n-1}{t} + \sum_{i=2}^{n} \binom{\tilde{d}_i + 1}{t} \\ &= \binom{n-1}{t} + \sum_{i=2}^{n} \left(\binom{\tilde{d}_i}{t} + \binom{\tilde{d}_i}{t-1} \right) = \binom{n-1}{t} + f(\tilde{G}, t) + f(\tilde{G}, t-1). \end{aligned}$$

By the induction hypothesis, both $f(\tilde{G}, t)$ and $f(\tilde{G}, t-1)$ are maximised by $\tilde{G} = \mathcal{L}_{n-1,2}(m-(n-1))$. Adding to this the full star with centre 1, we obtain $\mathcal{L}_{n,2}(m)$, thus proving its optimality.

Now suppose G is an extremal graph with the largest possible maximum degree Δ, and that $\Delta \leq n-2$. This means for any edge e and vertex i, we can replace e by an edge containing i. This shifting operation, coupled with the assumption of optimality, will allow us to determine the structure of G, and eventually derive a contradiction.

To begin with, we establish a lower bound for $f(G,t)$. Let $1 \leq r \leq \ell - 1$ be such that $\binom{n}{2} - \binom{n-r}{2} < m \leq \binom{n}{2} - \binom{n-r-1}{2}$. In this range, $\mathcal{L}_{n,2}(m)$ consists of r full stars and a partial star. Thus if G is extremal, we must have $f(G,t) \geq f(\mathcal{L}_{n,2}(m), t) \geq r\binom{n-1}{t} + (n-r)\binom{r}{t} > r\binom{n-1}{t}$.

We shall now double-count to deduce the existence of a high-degree vertex. Since every star we count in $f(G,t)$ contains t edges, and the number of stars an edge is in is determined by the degrees of its endpoints, we have

$$tf(G,t) = \sum_{e=\{i,j\} \in E(G)} \left(\binom{d_i - 1}{t-1} + \binom{d_j - 1}{t-1} \right) \leq 2m \binom{\Delta - 1}{t-1},$$

where Δ is the maximum degree in G. Applying the previous lower bound on $f(G,t)$ gives $\binom{\Delta-1}{t-1} \geq \frac{rt}{2m}\binom{n-1}{t} = \frac{r(n-1)}{2m}\binom{n-2}{t-1} > \frac{1}{4}\binom{n-2}{t-1}$, since $m < (r+1)(n-1)$. Since $\binom{\alpha p}{q} \leq \alpha^q \binom{p}{q}$ for $0 \leq \alpha \leq 1$ (see Lemma 5.3.2), it follows that $\Delta > 4^{-1/(t-1)}(n-2) + 1 \geq 4^{-1/(t-1)}n - 1$. Without loss of generality, suppose 1 is a vertex of maximum degree.

Consider any edge $e = \{i,j\} \in E(G)$, and suppose without loss of generality $d_i \geq d_j$. e is in $\binom{d_i-1}{t-1} + \binom{d_j-1}{t-1} \leq 2\binom{d_i-1}{t-1}$ stars of size t. On the other hand, if we replace e with an edge

containing 1, we would create at least $\binom{\Delta}{t-1}$ new stars. Hence, by the extremality of G, we must have $2\binom{d_i-1}{t-1} \geq \binom{\Delta}{t-1}$, so $d_i - 1 \geq 2^{-1/(t-1)}\Delta \geq 2^{-3/(t-1)}n - 1$.

This implies that $X = \{x \in [n] : d_x \geq 2^{-3/(t-1)}n\}$ forms a vertex cover of G. This cover cannot be too large, as we have the bound

$$2(r+1)(n-1) > 2m = \sum_i d_i \geq \sum_{x \in X} d_x \geq 2^{-3/(t-1)}n\,|X|\,,$$

and so $s = |X| < 2^{1+3/(t-1)}(r+1)$.

Moreover, let j be any vertex not adjacent to 1. We claim that j must in fact be isolated. Suppose to the contrary there were some vertex $i \neq 1$ with the edge $\{i,j\} \in E(G)$. This edge is contained in $\binom{d_i-1}{t-1} + \binom{d_j-1}{t-1}$ stars of t edges. If we were to replace $\{i,j\}$ with the edge $\{1,j\}$, we would create $\binom{\Delta}{t-1} + \binom{d_j-1}{t-1} > \binom{d_i-1}{t-1} + \binom{d_j-1}{t-1}$ stars, contradicting the optimality of G.

Thus it follows that all the edges of G are supported on the $\Delta + 1$ vertices in the closed neighbourhood of 1, and that G has a cover X of size $s < 2^{1+3/(t-1)}(r+1)$ vertices, all of which have degree at least $2^{-3/(t-1)}n$. Note that the vertices outside the cover have degree at most s, as they can only be adjacent to vertices in X.

To complete the argument, we shall show by shifting some edges that a graph with isolated vertices cannot be optimal. Without loss of generality, let $X = [s]$ be the cover mentioned above, and further assume that v has the lowest degree in X. Note that v has at least $2^{-3/(t-1)}n - (s-1) \geq s - 1$ neighbours outside X, since $s < 2^{1+3/(t-1)}(r+1)$. Let G' be the graph obtained from G by removing $s-1$ edges from v to neighbours $N \subset X^c$, and replacing them with $s-1$ edges from a previously isolated vertex w to the other $s-1$ vertices in X. Note that these vertices all have degree at least d_v.

Comparing degrees in G' to those in G, we find that the $s-1$ vertices in $X \setminus \{v\}$ have degree one larger, the degree of v has decreased by $s-1$, the degrees of the $s-1$ vertices in N, which were previously at most s, have decreased by 1, and w now has degree $s-1$. The change in the number of intersecting subgraphs is thus

$$
\begin{aligned}
&f(G',t) - f(G,t)\\
&= \sum_i \binom{d_i'}{t} - \sum_i \binom{d_i}{t}\\
&= \sum_{i\in X\setminus\{v\}} \left(\binom{d_i+1}{t} - \binom{d_i}{t}\right) + \left(\binom{d_v-s+1}{t} - \binom{d_v}{t}\right)\\
&\quad + \sum_{i\in N} \left(\binom{d_i-1}{t} - \binom{d_i}{t}\right) + \binom{d_w'}{t}\\
&= \sum_{i\in X\setminus\{v\}} \binom{d_i}{t-1} - \sum_{j=1}^{s-1} \left(\binom{d_v-j+1}{t} - \binom{d_v-j}{t}\right) - \sum_{i\in N} \binom{d_i-1}{t-1} + \binom{s-1}{t}\\
&\geq \sum_{i\in X\setminus\{v\}} \binom{d_v}{t-1} - \sum_{j=1}^{s-1} \binom{d_v-j}{t-1} - (s-1)\binom{s-1}{t-1}\\
&= \sum_{j=1}^{s-1} \left(\binom{d_v}{t-1} - \binom{d_v-j}{t-1}\right) - (s-1)\binom{s-1}{t-1}\\
&\geq \sum_{j=1}^{s-1} j\binom{d_v-j}{t-2} - (s-1)\binom{s-1}{t-1} \qquad (5.2)\\
&\geq \sum_{j=1}^{s-1} j\binom{d_v-s+1}{t-2} - (s-1)\binom{s-1}{t-1} = \binom{s}{2}\binom{d_v-s+1}{t-2} - (s-1)\binom{s-1}{t-1}\\
&\geq \binom{s}{2}\binom{s+1}{t-2} - (s-1)\binom{s-1}{t-1} \qquad [\text{since } d_v \geq 2^{-3/(t-1)}n \geq 2^{2+3/(t-1)}(r+1) > 2s]\\
&= \binom{s}{2}\binom{s+1}{t-2} - \frac{(s-1)^2}{t-1}\binom{s-2}{t-2} \geq \binom{s}{2}\left(\binom{s+1}{t-2} - \binom{s-2}{t-2}\right) \geq 0,
\end{aligned}
$$

since $t \geq 3$. Hence, by shifting edges, we can increase the maximum degree of G without decreasing the objective function. This contradicts the assumption that G was optimal with the largest maximum degree. □

Finally, we show how to deduce the general case of Proposition 5.2.1 from this result. This requires only minor modifications of the above proof, which we highlight below.

Proof of Proposition 5.2.1. Note that $f(G,t)$ counts precisely the number of stars of t edges in the graph G (except when $t = 0$, when the empty graph is counted n times, and $t = 1$, in which case the single edges are counted twice). When $t \neq 3$, these stars are the only

intersecting graphs of t edges, and thus Proposition 5.2.1 follows directly from Proposition 5.2.2.

When $t = 3$, we must augment the proof of Proposition 5.2.2 to also account for the triangles in the graph. However, the number of possible triangles is a lower order term that can be taken care of by slightly altering the argument.

We begin by observing that the inductive argument still holds. The theorem holds for $m \leq n-1$, as every 3-edge subgraph is intersecting, which is clearly the best possible. Moreover, suppose G contains a full star, and let G' denote the subgraph with the full star removed. Then each edge in G' induces one triangle with edges from the full star. Thus we can again write the number of intersecting subgraphs of 3 edges as a constant term, independent of the structure of G', plus the corresponding terms from G', and can then apply the inductive hypothesis.

We next need a lower bound on the maximum degree Δ. Note that an edge $\{i,j\}$ can be involved in at most $\min\{d_i-1, d_j-1\}$ triangles, and thus in at most $\binom{d_i-1}{2} + \binom{d_j-1}{2} + \min\{d_i-1, d_j-1\}$ intersecting subgraphs of three edges in total. Hence we have

$$\begin{aligned} 3\text{int}(G,3) &\leq \sum_{\{i,j\}\in E(G)} \left(\binom{d_i-1}{2} + \binom{d_j-1}{2} + \min\{d_i-1, d_j-1\} \right) \\ &\leq 2m \left(\binom{\Delta-1}{2} + \Delta - 1 \right) = 2m\binom{\Delta}{2}. \end{aligned}$$

On the other hand, we have $\text{int}(G,3) \geq \text{int}(\mathcal{L}_{n,2}(m)) \geq r\binom{n-1}{3}$. From these inequalities, we can deduce $\Delta(\Delta-1) \geq \frac{1}{4}(n-2)(n-3)$. Again, assume that 1 is a vertex of maximum degree.

Now if we have the edge $e = \{i,j\}$ with $d_i \geq d_j$, then e is contained in at most $\binom{d_i-1}{2} + \binom{d_j-1}{2} + d_j - 1 \leq \binom{d_i-1}{2} + \binom{d_j}{2} \leq \binom{d_i-1}{2} + \binom{d_i}{2}$ intersecting families of three edges. Since replacing e with an edge containing 1 would create at least $\binom{\Delta}{2}$ new stars of three edges, we must have $\binom{d_i-1}{2} + \binom{d_i}{2} > \binom{\Delta}{2}$, which, given our above bound on Δ, shows $X = \{i : d_i \geq \frac{1}{2\sqrt{2}}n\}$ is a cover for G. In fact, these shifting arguments also show that X must be a clique. As before, we can also show that if v is not adjacent to 1, then v must in fact be an isolated vertex.

To complete the argument, we show that graphs with isolated vertices cannot be optimal by shifting $s-1$ edges to an isolated vertex, where $|X| = s \geq 2$. In the proof of Proposition 5.2.2, we saw that such a shift results in a gain of at least $\binom{s}{2}\left(\binom{s+1}{t-2} - \binom{s-2}{t-2}\right) = 3\binom{s}{2}$ stars of three edges. On the other hand, as $V(G) \setminus X$ is an independent set, we lose at most $(s-1)^2$ triangles, since every edge removed can only form a triangle with another vertex from X. However, by adding $s-1$ edges from a clique to a new vertex, we create $\binom{s-1}{2}$ new triangles. Hence we incur a net loss of at most $\binom{s}{2}$ triangles. For $s \geq 3$ we have $3\binom{s-1}{2} \geq \binom{s}{2}$, and so shifting the edges increases the maximum degree without decreasing $\mathrm{int}(G,3)$, contradicting our choice of G. If $s = 2$, then we are shifting one edge from the vertex of second-highest degree, say 2, to the vertex of maximum degree. By performing the preceding calculations more carefully, we find that we gain at least $d_2 - 2 > 0$ intersecting subgraphs of three edges, again contradicting the optimality of G. □

This completes the proof of Theorem 5.1.1, showing that the initial segment of the lexicographic order is the most probably intersecting graph up to moderate densities. Note that, as in all previously obtained results in [KKK12] and [RW13], these graphs actually simultaneously maximise the number of intersecting subgraphs of all sizes, and hence the most probably intersecting graphs do not depend on p. This phenomenon fails to hold for denser graphs, but we defer this discussion until Section 5.4.

We conclude with some remarks on the uniqueness of the extremal graphs. To have equality, we must in particular have equality in (5.2), namely that $\binom{d_v}{t-1} - \binom{d_v-j}{t-1} = j\binom{d_v-j}{t-1}$ for all $1 \leq j \leq s-1$. There are only three possible cases: $t \leq 2$, $t \geq d_v + 2$ or $s = 2$. In the first case, if $t = 0$ or $t = 1$ it is trivially that there is no uniqueness, as any graph with m edges will be extremal. When $t = 2$, this reduces to the question of uniqueness in Theorem 4.1.6. In this case, the extremal graphs are completely characterised by Ábrego et al [AFN09], where it is shown that they are closely related to $\mathcal{L}_{n,2}(m)$.

If $s = 2$ and $t \leq d_v$, it is easy to see that shifting an edge to the vertex of highest degree increases the number of intersecting subgraphs. For $t \geq d_v$ (and $t \leq \binom{n-1}{k-1}$), the edges meeting the cover X only at v are not contained in any intersecting subgraphs of size t,

and hence we may remove them to complete a star and increase the number of intersecting subgraphs. Thus in these cases it follows that the extremal graph must contain r full stars, and $\mathcal{L}_{n,2}(m)$ is uniquely extremal if the number of additional edges is at least $t - r$.

5.3 Intersecting hypergraphs

We now seek to extend these results to the hypergraph setting and prove Theorem 5.1.2, which we recall below.

Theorem 5.1.4. *Let k and ℓ be integers, and suppose $n \geq n_0(k, \ell)$ and $\binom{n}{k} - \binom{n-\ell}{k} \leq m \leq \binom{n}{k} - \binom{n-\ell}{k} + n - \ell - k + 1$. For this range of parameters, $\mathcal{L}_{n,k}(m)$ is the most probably intersecting k-uniform hypergraph on $[n]$ with m sets.*

The general proof strategy will follow that of Section 5.2, in that we shall deduce the probabilistic result by counting the number of intersecting subfamilies. However, in contrast to the graph case, there is a rich variety of non-isomorphic intersecting structures we shall have to account for. We call intersecting families that are not stars *non-trivially intersecting*. Despite the wide range of non-trivially intersecting families, these are very small families, as the Hilton-Milner theorem [HM67] shows that the largest non-trivially intersecting family has size $\binom{n-1}{k-1} - \binom{n-k-1}{k-1} + 1 = o\left(\binom{n-1}{k-1}\right)$. It remains the case that most intersecting subfamilies are stars, as we show in the following lemma.

Lemma 5.3.1. *For $F \in \mathcal{F}$, the number of non-trivially intersecting families of size t in $\mathcal{F}$ containing F is $O\left(n^{-t/4k\binom{n-1}{k-1}\binom{n-1}{t-1}}\right)$, and the total number of such families in $\mathcal{F}$ is $O\left(n^{-t/4k\binom{n-1}{k-1}\binom{n-1}{t}}\right)$.*

While the bounds required on n can be explicitly calculated, we have chosen to simplify the presentation through the use of asymptotic notation, where we fix k and ℓ and let n tend to infinity. Note, however, that we make no assumption on the relative magnitudes of n and t; t may be as large as $\binom{n-1}{k-1}$.

The proof of Lemma 5.3.1 is slightly technical, and so we defer it until the end of this

section. However, throughout this section we shall require some estimates on binomial coefficients, which we collect below.

Lemma 5.3.2. *Suppose we have integers $0 \le a \le b \le c$ and $0 < M \le S$. Then*

(i) $\binom{b}{r} \le \left(\frac{b}{c}\right)^r \binom{c}{r}$,

(ii) for $r \ge 1$, if $\sum_i n_i = S$ and $0 \le n_i \le M$ for all i, then $\sum_i \binom{n_i}{r} \le \frac{S}{M}\binom{M}{r}$, and

(iii) for $r \ge 2$, $\left[\binom{b-a}{r} + \binom{c+a}{r}\right] - \left[\binom{b}{r} + \binom{c}{r}\right] \ge \left(1 - \frac{b-a}{c}\right) \frac{ar}{c-r+1}\binom{c}{r}$.

Proof of Lemma 5.3.2. (i) By definition, we have

$$\binom{b}{r} = \frac{1}{r!}\prod_{j=0}^{r-1}(b-j) \le \frac{1}{r!}\prod_{j=0}^{r-1}\frac{b}{c}(c-j) = \left(\frac{b}{c}\right)^r \binom{c}{r}.$$

(ii) Suppose we had i and j such that $0 < n_j \le n_i < M$. Fixing the other variables, we have

$$\begin{aligned}\binom{n_j - 1}{r} + \binom{n_i + 1}{r} &= \binom{n_j}{r} - \binom{n_j - 1}{r-1} + \binom{n_i}{r} + \binom{n_i}{r-1} \\ &= \binom{n_j}{r} + \binom{n_i}{r} + \binom{n_i}{r-1} - \binom{n_j - 1}{r-1} \ge \binom{n_j}{r} + \binom{n_i}{r}.\end{aligned}$$

This shows we may assume there is at most one i for which $0 < n_i < M$. Since $\sum_i n_i = S$, this implies we have $m = \left\lfloor \frac{S}{M} \right\rfloor$ variables $n_j = M$, with one variable equal to $S - mM$. Hence, using (i),

$$\begin{aligned}\sum_i \binom{n_i}{r} &\le m\binom{M}{r} + \binom{S - mM}{r} \le m\binom{M}{r} + \left(\frac{S - mM}{M}\right)^r \binom{M}{r} \\ &\le \left(m + \frac{S - mM}{M}\right)\binom{M}{r} = \frac{S}{M}\binom{M}{r}.\end{aligned}$$

(iii) We rearrange and telescope the sums

$$\begin{aligned}&\left[\binom{b-a}{r} + \binom{c+a}{r}\right] - \left[\binom{b}{r} + \binom{c}{r}\right] = \left[\binom{c+a}{r} - \binom{c}{r}\right] - \left[\binom{b}{r} - \binom{b-a}{r}\right] \\ &= \sum_{j=1}^{a}\left(\left[\binom{c+j}{r} - \binom{c+j-1}{r}\right] - \left[\binom{b-a+j}{r} - \binom{b-a+j-1}{r}\right]\right) \\ &= \sum_{j=1}^{a}\left[\binom{c+j-1}{r-1} - \binom{b-a+j-1}{r-1}\right].\end{aligned}$$

Using (i), we can estimate these differences

$$\begin{aligned}\binom{c+j-1}{r-1} - \binom{b-a+j-1}{r-1} &\geq \binom{c+j-1}{r-1} - \left(\frac{b-a+j-1}{c+j-1}\right)^{r-1}\binom{c+j-1}{r-1} \\ &\geq \left(1 - \frac{b-a+j-1}{c+j-1}\right)\binom{c+j-1}{r-1} \geq \left(1-\frac{b-a}{c}\right)\binom{c}{r-1}.\end{aligned}$$

Thus we have

$$\begin{aligned}\left[\binom{b-a}{r} + \binom{c+a}{r}\right] - \left[\binom{b}{r} + \binom{c}{r}\right] &\geq \left(1-\frac{b-a}{c}\right)\sum_{j=1}^{a}\binom{c}{r-1} \\ &= \left(1-\frac{b-a}{c}\right)\frac{ar}{c-r+1}\binom{c}{r}.\end{aligned}$$

□

Armed with these lemmas, we may now proceed to deduce our counting result. In particular, Lemma 5.3.1 implies that when counting intersecting subfamilies of size t, the non-trivially intersecting families are a lower order term, and so we may focus on the number of stars with t edges. Applying similar shifting arguments to those in Section 5.2, we shall deduce a rough structural characterisation of optimal families. Recall that we say a star with centre i is *full* if $\mathcal{F}$ contains all $\binom{n-1}{k-1}$ sets containing i, and *almost-full* if it has $(1-o(1))\binom{n-1}{k-1}$ such sets.

Proposition 5.3.3. *Let k, ℓ and $t \geq 2$ be integers, and suppose $n \geq n_0(k,\ell)$ and $\binom{n}{k} - \binom{n-\ell}{k} \leq m \leq \binom{n}{k} - \binom{n-\ell-1}{k}$. If $\mathcal{F}$ is a k-uniform set family on $[n]$ of size m maximising the number of intersecting subfamilies of size t, then either*

(i) $\mathcal{F}$ contains ℓ full stars, or

(ii) $\mathcal{F}$ consists of $\ell+1$ almost-full stars.

Before we begin to prove Proposition 5.3.3, we first analyse the initial segment of the lexicographic order to obtain a lower bound on the number of intersecting subfamilies in an optimal family. Note that, for m in the above range, $\mathcal{L}_{n,k}(m)$ consists of all sets intersecting $[\ell]$, with $m - \binom{n}{k} + \binom{n-\ell}{k}$ additional sets all containing $\ell+1$. Hence $\mathcal{L}_{n,k}(m)$ falls under case (i) above.

When counting the intersecting subfamilies in $\mathcal{L}_{n,k}(m)$, we consider only the stars with centre i for some $1 \leq i \leq \ell$. There are ℓ choices for the centre of the star, and then for each star we must choose t of the $\binom{n-1}{k-1}$ possible sets. A star is overcounted only if all its sets contain at least two elements from $[\ell]$, giving at most $\binom{n-2}{k-2}$ sets for each choice of elements from $[\ell]$. By the Bonferroni Inequalities and Lemma 5.3.2, we have

$$\begin{aligned}\operatorname{int}(\mathcal{L}_{n,k}(m), t) &\geq \ell\binom{\binom{n-1}{k-1}}{t} - \binom{\ell}{2}\binom{\binom{n-2}{k-2}}{t} \geq \left[\ell - \binom{\ell}{2}\left(\frac{k-1}{n-1}\right)^t\right]\binom{\binom{n-1}{k-1}}{t} \\ &= (\ell - o(1))\binom{\binom{n-1}{k-1}}{t}.\end{aligned}$$

This gives us a lower bound on $\operatorname{int}(\mathcal{F}, t)$ for any optimal family $\mathcal{F}$. We now proceed with the proof of Proposition 5.3.3.

Proof of Proposition 5.3.3. Suppose $\mathcal{F}$ is optimal for the given parameters. Note that we may assume $\ell \geq 1$, as (i) is trivially satisfied for $\ell = 0$.

Let d_i denote the degree of vertex i. Our goal is to show that either $d_i = \binom{n-1}{k-1}$ for ℓ vertices i, or $d_i = (1 - o(1))\binom{n-1}{k-1}$ for $\ell+1$ vertices that cover $\mathcal{F}$. Suppose $\mathcal{F}$ has p full stars, which we may assume have centres $1 \leq i \leq p$. If $p = \ell$ we are done, so assume $p \leq \ell - 1$.

Note that for $i > p$, none of the vertices have full degree, and so we may replace any set in $\mathcal{F}$ with a set containing i. In order to fully utilise this shifting, we will first show there is a vertex of relatively large degree. From this, we shall deduce the existence of a small set of vertices covering all the edges. Finally, we shall shift sets in this small cover to obtain the desired result.

To begin, note that by optimality we must have

$$\operatorname{int}(\mathcal{F}, t) \geq \operatorname{int}(\mathcal{L}_{n,k}(m), t) \geq (\ell - o(1))\binom{\binom{n-1}{k-1}}{t}.$$

By Lemma 5.3.1, it follows that almost all of these intersecting subfamilies should be stars. Let d_i denote the degree of vertex i. Then, counting over the centres of the stars, we have

$$\operatorname{int}(\mathcal{F}, t) \leq \sum_i \binom{d_i}{t} + o\left(\binom{\binom{n-1}{k-1}}{t}\right) = (p + o(1))\binom{\binom{n-1}{k-1}}{t} + \sum_{i>p}\binom{d_i}{t},$$

and so

$$\sum_{i>p} \binom{d_i}{t} \geq (\ell - p - o(1)) \binom{\binom{n-1}{k-1}}{t}.$$

Note that by double-counting the edges, we have $\sum_{i>p} d_i \leq \sum_i d_i = km \leq k(\ell+1)\binom{n-1}{k-1}$. Suppose we had $d_i \leq M = c\binom{n-1}{k-1}$ for all $i > p$. By Lemma 5.3.2, we have

$$\sum_{i>p} \binom{d_i}{t} \leq \frac{k(\ell+1)\binom{n-1}{k-1}}{M} \binom{M}{t} \leq k(\ell+1)c^{t-1} \binom{\binom{n-1}{k-1}}{t}.$$

Comparing this to the lower bound, we must have $k(\ell+1)c^{t-1} \geq \ell - p - o(1)$, which implies $c = \Omega(1)$. Hence we have some vertex, which we may assume to be $i = p+1$, with $d_{p+1} \geq c\binom{n-1}{k-1}$.

We shall now show that there is a small cover of vertices of large degree. Let $X = \left\{i : d_i \geq \frac{c}{k}\binom{n-1}{k-1}\right\}$, and suppose for contradiction we have $F \in \mathcal{F}$ with $F \cap X = \emptyset$.

We have $\{G \in \mathcal{F} : G \cap F \neq \emptyset\} = \cup_{i \in F}\{G \in \mathcal{F} : i \in G\}$, and so, since $F \cap X = \emptyset$, there are at most $\sum_{i \in F} d_i < c\binom{n-1}{k-1}$ sets in $\mathcal{F}$ intersecting F. Since any intersecting subfamily containing F must consist only of sets intersecting F, there are fewer than $\binom{c\binom{n-1}{k-1}}{t-1}$ such subfamilies.

On the other hand, if we replace F with a set containing $p+1$, the new set would be in $\binom{d_{p+1}}{t-1} \geq \binom{c\binom{n-1}{k-1}}{t-1}$ stars in $\mathcal{F}$. This shift would thus increase the number of intersecting subfamilies in $\mathcal{F}$, contradicting the optimality of $\mathcal{F}$. Hence we must have $F \cap X \neq \emptyset$ for all $F \in \mathcal{F}$; that is, X covers $\mathcal{F}$.

We now show this cover is small. Indeed, we have

$$k(\ell+1)\binom{n-1}{k-1} \geq km = \sum_i d_i \geq \sum_{i \in X} d_i \geq \frac{c}{k}\binom{n-1}{k-1}|X|,$$

and so $|X| \leq \frac{k^2(\ell+1)}{c} = O(1)$, as desired.

Now take a minimal subcover in X, which we may assume to be $[r]$. Thus $r \leq |X| = O(1)$. Since $m \geq \binom{n}{k} - \binom{n-\ell}{k}$, we must have $r \geq \ell+1$ (we cannot have $r = \ell$, as we have assumed $\mathcal{F}$ only has $p < \ell$ full stars). Note that every vertex in $[r]$ has degree at least $\frac{c}{k}\binom{n-1}{k-1}$. Moreover, for any vertex $i \notin [r]$, all sets containing i must also meet $[r]$, and so we have $d_i \leq r\binom{n-2}{k-2}$.

We shall employ shifting arguments to show that all vertices in $[r]$ that are not of full degree should have approximately equal degrees. Indeed, let i and j be two such vertices. By the minimality of the cover, there must be some set F with $F \cap [r] = i$. From the preceding remarks, it follows that the number of sets intersecting F is at most $\sum_{v \in F} d_v \leq d_i + r(k-1)\binom{n-2}{k-2}$. Hence F is in at most $\binom{d_i + r(k-1)\binom{n-2}{k-2}}{t-1}$ intersecting subfamilies.

On the other hand, if we were to add a new set containing j, it would be in at least $\binom{d_j}{t-1}$ stars of sets containing j. By optimality, it cannot be desirable to shift F to a set containing j, and so we must have $\binom{d_i + r(k-1)\binom{n-2}{k-2}}{t-1} \geq \binom{d_j}{t-1}$, and hence $d_j \leq d_i + r(k-1)\binom{n-2}{k-2} = d_i + o\left(\binom{n-1}{k-1}\right)$. By symmetry, we have $d_j = d_i + o\left(\binom{n-1}{k-1}\right)$ for all such vertices i, j.

Let us now review what we have revealed of the structure of $\mathcal{F}$. There are p vertices $[p]$ of degree $\binom{n-1}{k-1}$, and a further $r-p$ vertices $[r] \setminus [p]$ of almost-equal degree that cover the remaining edges. Let $\alpha \in [0,1]$ be such that $m = \binom{n}{k} - \binom{n-\ell}{k} + \alpha\binom{n-1}{k-1} = (\ell + \alpha - o(1))\binom{n-1}{k-1}$. Since the first p vertices cover $(p - o(1))\binom{n-1}{k-1}$ edges, the degrees of the remaining $r-p$ vertices must be $\frac{\ell - p + \alpha + o(1)}{r-p}\binom{n-1}{k-1}$. Let us assume they are listed in order of decreasing degrees, so $d_{p+1} \geq d_r$.

Suppose for some fixed $0 < \varepsilon < \frac{c}{k}$ we had $\frac{\ell - p + \alpha + o(1)}{r-p} < 1 - \varepsilon$. Since $d_r \geq \frac{c}{k}\binom{n-1}{k-1}$, and there are $o\left(\binom{n-1}{k-1}\right)$ sets containing r that also contain another element of $[r]$, we can find a set of $\varepsilon\binom{n-1}{k-1}$ edges that only meet $[r]$ at r. We shall shift these edges to the vertex $p+1$.

By Lemma 5.3.1, the number of non-trivially intersecting subfamilies created or destroyed is a lower-order term, while the degrees of vertices outside $[r]$ are so small that by Lemma 5.3.2 we may ignore the number of stars with centres outside $[r]$. Hence the only intersecting subfamilies we need to consider are the stars with centres $p+1$ or r.

Before the shift, we had $\binom{d_{p+1}}{t} + \binom{d_r}{t}$ such stars, and after the shift, there are $\binom{d_{p+1} + \varepsilon\binom{n-1}{k-1}}{t} + \binom{d_r - \varepsilon\binom{n-1}{k-1}}{t}$ stars. Applying Lemma 5.3.2, we gain at least

$$\left(1 - \frac{d_r - \varepsilon\binom{n-1}{k-1}}{d_{p+1}}\right) \frac{\varepsilon t \binom{n-1}{k-1}}{d_{p+1} - t + 1} \binom{d_{p+1}}{t} > \varepsilon^2 \binom{d_{p+1}}{t}$$

stars. This is strictly positive unless $t > d_{p+1} \geq \frac{c}{k}\binom{n-1}{k-1}$. In this case, it follows by the Hilton-Milner theorem [HM67] that the only intersecting families of size t are stars. Since

no set meeting the cover X only in $p+1$ is contained in a star of size t (as $t > d_{p+1} \geq d_i$ for any vertex i in such a set), we may shift sets containing $p+1$ to other vertices in the cover. We can repeat this process until we obtain a full star, which will strictly increase the number of t-stars, contradicting the optimality of $\mathcal{F}$.

This contradicts the optimality of $\mathcal{F}$. Hence we must have $\frac{\ell-p+\alpha+o(1)}{r-p} = 1-o(1)$. Since $r \geq \ell+1$, this is only possible when $r = \ell+1$ (and $\alpha = 1-o(1)$), and so it follows that $\mathcal{F}$ consists of $\ell+1$ almost-full stars, and thus we are in case (ii).

This completes the proof of Proposition 5.3.3. □

This result provides us with the approximate structure of the extremal families. In particular, when α is not $1-o(1)$, we know that any extremal family contains ℓ full stars, and hence is close to $\mathcal{L}_{n,k}(m)$ in structure. In order to show that $\mathcal{L}_{n,k}(m)$ is in fact optimal, it remains to determine the structure of the sets outside the ℓ full stars. In some special cases, we are able to do this exactly, as given by the following proposition.

Proposition 5.3.4. *Let k, ℓ and t be integers, and suppose $n \geq n_0(k,\ell)$ and $\binom{n}{k} - \binom{n-\ell}{k} \leq m \leq \binom{n}{k} - \binom{n-\ell}{k} + n - \ell - k + 1$. If $\mathcal{F}$ is a k-uniform set family on $[n]$ with m edges, then $\mathrm{int}(\mathcal{F}, t) \leq \mathrm{int}(\mathcal{L}_{n,k}(m), t)$.*

Proof of Proposition 5.3.4. If $t = 0$ or $t = 1$, then there is nothing to prove, as $\mathrm{int}(\mathcal{F}, 0) = 1$ and $\mathrm{int}(\mathcal{F}, 1) = m$ for all such families $\mathcal{F}$. Hence we may assume $t \geq 2$, and thus apply Proposition 5.3.3. It follows that $\mathcal{F}$ must contain ℓ full stars. Let us write $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, where $\mathcal{F}_0$ is the union of the ℓ full stars, and $\mathcal{F}_1$ consists of the remaining sets. Let $m_1 = |\mathcal{F}_1| = m - \binom{n}{k} + \binom{n-\ell}{k}$ denote the number of additional sets $\mathcal{F}$ contains. If $m_1 = 0$ then we are done, as all edges are accounted for. If $m_1 = 1$, then by symmetry it does not matter which set we add outside the ℓ stars, and so it again follows that $\mathcal{L}_{n,k}(m)$ is optimal. Hence we may assume $m_1 \geq 2$.

We will now show that $\mathrm{int}(\mathcal{F}, t)$ is maximised when $|\cap_{F \in \mathcal{F}_1} F| = k-1$; that is, when the sets in $\mathcal{F}_1$ have the maximum possible intersection. In our case, since $m_1 \leq n - \ell - k + 1$, the additional sets in $\mathcal{L}_{n,k}(m)$ all share the elements $\{\ell+1, \ell+2, \ldots, \ell+k-1\}$, and hence

it will follow that $\mathcal{L}_{n,k}(m)$ is optimal.

We count the intersecting subfamilies of $\mathcal{F}$ based on their intersection with $\mathcal{F}_1$. Given some $\mathcal{H} \subset \mathcal{F}_1$ with h sets, let $\mathrm{ext}(\mathcal{H})$ denote the number of extensions of $\mathcal{H}$ to an intersecting subfamily of $\mathcal{F}$ of size t. In other words, it is the number of intersecting subfamilies in $\mathcal{F}_0$ of size $t-h$ that intersect all sets in $\mathcal{H}$. We then have

$$\mathrm{int}(\mathcal{F}, t) = \sum_{h=0}^{t} \sum_{\mathcal{H} \in \binom{\mathcal{F}_1}{h}} \mathrm{ext}(\mathcal{H}).$$

When $h = 0$, we simply obtain the number of intersecting subfamilies of size t in $\mathcal{F}_0$, which is independent of $\mathcal{F}_1$. If $h = 1$, then by symmetry it does not matter which set we choose for $\mathcal{H}$. Hence we may assume $h \geq 2$. Suppose we have $|\cap_{H \in \mathcal{H}} H| = a$. The number of sets $F \in \mathcal{F}_0$ that intersect $\mathcal{H}$ without containing one of the a common elements is very small. Indeed, fix any set $H \in \mathcal{H}$. Since $F \cap H \neq \emptyset$, there are k options for this intersection x. As we are not selecting one of the a common elements of $\mathcal{H}$, there must be some other set $H' \in \mathcal{H}$ not containing x. Hence we must again select an element of H', giving a further k options at the most. Finally, since F belongs to $\mathcal{F}_0$, we must choose one of the ℓ centres of the stars. There are then a further $k-3$ elements to choose for F. Thus there are at most $\ell k^2 \binom{n-3}{k-3} < \frac{\ell k^3}{n} \binom{n-2}{k-2}$ such sets F. This will be a lower order term, which we may disregard. In particular, this implies that we should have $a \geq 1$ for $\mathcal{H}$ to have a significant number of extensions.

We shall now estimate $\mathrm{ext}(\mathcal{H})$. Calculations similar to those in the proof of Lemma 5.3.1 show that the number of extensions that are not themselves stars is a lower-order term, and hence we need only consider trivially intersecting extensions. There are three cases to consider.

The centre of the star could be one of the centres of the ℓ full stars in $\mathcal{F}_0$. There are thus ℓ choices for the centre, and then the sets chosen must intersect $\mathcal{H}$. In light of our previous remarks, the number of such sets is dominated by those containing one of the a common elements, giving $(a + o(1))\binom{n-2}{k-2}$ options. We double-count very few extensions, as then the sets from $\mathcal{F}_0$ must all contain two of the ℓ centres of the stars, giving at most $\binom{\ell}{2} a \binom{n-3}{k-3}$ such

sets. Thus the number of extensions of this type is $(\ell - o(1))\binom{(a+o(1))\binom{n-2}{k-2}}{t-h}$.

The second type of trivially intersecting extension is that where the centre is one of the a common elements of $\mathcal{H}$. These sets must then contain any one of the ℓ centres of the stars in $\mathcal{F}_0$, and thus the number of extensions is $(a - o(1))\binom{(\ell-o(1))\binom{n-2}{k-2}}{t-h}$.

The final type is that where the centre of the star x is neither one of the ℓ centres from $\mathcal{F}_0$ nor one of the a common elements from $\mathcal{F}_1$. These sets must then contain x, one of the ℓ centres, and some elements from $\mathcal{H}$, and thus there are very few such sets.

We thus conclude that $\text{ext}(\mathcal{H}) = (\ell - o(1))\binom{(a+o(1))\binom{n-2}{k-2}}{t-h} + (a - o(1))\binom{(\ell-o(1))\binom{n-2}{k-2}}{t-h}$. This is increasing in a, and so to maximise $\text{ext}(\mathcal{H})$ we must have $a = k-1$. However, all subfamilies $\mathcal{H}$ with $a = k-1$ are isomorphic, as they consist of h distinct vertices attached to a common core of $k-1$ vertices. Hence in this case $\text{ext}(\mathcal{H})$ does not depend on which sets we choose, and thus $\text{ext}(\mathcal{H})$ is maximised if and only if $a = k-1$.

This completes the proof of Proposition 5.3.4.

□

Note that for a family $\mathcal{F}$ to be extremal, it should maximise $\text{ext}(\mathcal{H})$ for all $\mathcal{H} \subset \mathcal{F}_1$. In particular, provided t is not too large, this implies that $\mathcal{F}$ is extremal if and only if it contains ℓ full stars and, for $2 \leq h \leq t-1$, any collection of h sets in $\mathcal{F}$ outside the full stars have $k-1$ vertices in common. When t is large, we will have $\text{ext}(\mathcal{H}) = 0$ for all $\mathcal{H}$, as it will be impossible to find t sets that intersect $\mathcal{H}$ and meet the centres of the sets $[\ell]$. Hence in this case $\mathcal{F}_1$ may be chosen arbitrarily, and $\mathcal{F}$ is extremal if and only if it contains ℓ full stars.

Unfortunately, in contrast to the graph case, this gives a rather narrow range of family sizes for which we are able to determine the extremal families exactly. However, it is necessary to have a somewhat more restricted range, as we shall show in Section 5.4 that even for $\binom{n-1}{k-1} < m < 2\binom{n-1}{k-1}$, $\mathcal{L}_{n,k}(m)$ is not always optimal.

Finally, note that the exact counting result in Proposition 5.3.4 implies that for these ranges of family sizes, $\mathcal{L}_{n,k}(m)$ is a most probably intersecting family, thus giving Theorem

5.1.2. The proof is exactly the same as the derivation of Theorem 5.1.1 from Proposition 5.2.1, and so we do not repeat it here.

To complete this section, we now furnish a proof of Lemma 5.3.1, bounding the number of non-trivially intersecting families.

Proof of Lemma 5.3.1. We begin by bounding the total number of non-trivially intersecting families of size t in $\binom{[n]}{k}$. Given such a family $\mathcal{F}$, we write $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, where $\mathcal{F}_0$ is the largest star in $\mathcal{F}$. Note that we must have $\mathcal{F}_1 \neq \emptyset$, as $\mathcal{F}$ is non-trivially intersecting. Let $S = \cap_{F \in \mathcal{F}_0} F$ be the centre of $\mathcal{F}_0$, and let $\mathcal{M} \subset \{F \setminus S : F \in \mathcal{F}_0\}$ be the largest matching in the sets of the star after the centre is removed. We denote the sizes of these sets as follows: $|\mathcal{F}_0| = t_0$, $|S| = s$ and $|\mathcal{M}| = b$.

Let us first provide some bounds on these parameters. Clearly, $s \leq k$, as S is a subset of each set in the star $\mathcal{F}_0$. Moreover, we claim $b \leq k$ as well. Indeed, every set F in $\mathcal{F}_1$ must be disjoint from S, as otherwise $\mathcal{F}_1 \cup \{F\}$ would form a larger star. However, it must intersect the sets $\{S \cup M : M \in \mathcal{M}\} \subset \mathcal{F}_0$, and thus it must contain one element from each of the b disjoint sets in $\mathcal{M}$. Since $|F| \leq k$, we must have $b \leq k$. An easy lower bound on t_0 is $t_0 \geq 2$, since any pair of sets in $\mathcal{F}$ forms a star. We in fact claim $t_0 \geq \frac{t}{k}$. Taking any set $F \in \mathcal{F}$, note that all the other sets in $\mathcal{F}$ must intersect F. By the pigeonhole principle, there is some element of F contained in at least a $\frac{1}{k}$-proportion of the other sets, giving a star of size at least $\frac{t}{k}$, as desired.

We now construct the intersecting family $\mathcal{F}$. There are $\binom{n}{s}$ choices for the centre S. We then have to select b sets of size $k-s$ for the matching $\mathcal{M}$. There are $\binom{n-s}{k-s}$ options for each set, giving $\binom{\binom{n-s}{k-s}}{b}$ possible matchings $\mathcal{M}$. By the maximality of $\mathcal{M}$, each of the remaining sets in $\mathcal{F}_0$ must meet the $(k-s)b$ elements covered by the matching $\mathcal{M}$. Hence there are at most $(k-s)b\binom{n-s-1}{k-s-1}$ choices for each set, providing $\binom{(k-s)b\binom{n-s-1}{k-s-1}}{t_0-b}$ ways to completing $\mathcal{F}_0$. As mentioned earlier, each set in $\mathcal{F}_1$ must avoid S and contain at least one element from each set in $\mathcal{M}$. This leaves at most $(k-s)^b\binom{n-s-b}{k-b}$ sets, from which we have to choose $t-t_0$. Thus the number of non-trivially intersecting families with these parameters is bounded above by

$$\binom{n}{s}\binom{\binom{n-s}{k-s}}{b}\binom{(k-s)b\binom{n-s-1}{k-s-1}}{t_0-b}\binom{(k-s)^b\binom{n-s-b}{k-b}}{t-t_0}.$$

Applying the estimates in part (i) of Lemma 5.3.2, this can be further bounded by

$$n^s \left[\left(\frac{k}{n}\right)^{b(s-1)} \binom{n-1}{k-1}\binom{n-1}{b} \right] \left[\left(\frac{k^{s+1}b}{n^s}\right)^{t_0-b} \binom{n-1}{k-1}\binom{n-1}{t_0-b} \right] \left[\left(\frac{k^{2b-1}}{n^{b-1}}\right)^{t-t_0} \binom{n-1}{k-1}\binom{n-1}{t-t_0} \right]$$
$$= \frac{b^{t_0-b}k^{2b(t-t_0-1)+(s+2)t_0-t}}{n^{(b-1)(t-t_0-1)+s(t_0-1)-1}} \binom{n-1}{k-1}\binom{n-1}{b}\binom{n-1}{t_0-b}\binom{n-1}{t-t_0}$$

We now simplify this expression. Since $b, s \leq k$ and $t_0 \leq t$, we can easily bound the numerator above by k^{4kt}. For the denominator, note that $t_0 \leq t-1$, as $\mathcal{F}_1 \neq \emptyset$, $s \geq 1$ and $t_0 - 1 \geq \frac{t}{2k}$, as $1 \leq \frac{t_0}{2}$ and $t_0 \geq \frac{t}{k}$, giving a lower bound of $n^{t/2k-1}$. Thus the number of non-trivially intersecting families with parameters s, b and t_0 is at most $n\left(\frac{k^{4k}}{n^{1/2k}}\right)^t \binom{n-1}{k-1}\binom{n-1}{b}\binom{n-1}{t_0-b}\binom{n-1}{t-t_0}$.

For the total number of non-trivially intersecting families, we now sum over all s, b and t_0, obtaining a bound of

$$\sum_{s=1}^{k}\sum_{b=1}^{k}\sum_{t_0=t/k}^{t-1} n\left(\frac{k^{4k}}{n^{1/2k}}\right)^t \binom{n-1}{k-1}\binom{n-1}{b}\binom{n-1}{t_0-b}\binom{n-1}{t-t_0}$$
$$\leq kn\left(\frac{k^{4k}}{n^{1/2k}}\right)^t \sum_{0\leq b\leq t_0\leq t} \binom{n-1}{k-1}\binom{n-1}{b}\binom{n-1}{t_0-b}\binom{n-1}{t-t_0}$$
$$\leq kn\left(\frac{8k^{4k}}{n^{1/2k}}\right)^t \binom{\binom{n-1}{k-1}}{t}.$$

To obtain the last inequality, we interpret the sum of the products of the three binomial coefficients as selecting, with repetition, from a collection of $\binom{n-1}{k-1}$ objects three sets A, B and C whose sizes sum to t. We could instead first select t elements from this collection, and then for each element decide which sets among A,B and C the elements should belong to. As the selection was with repetition, an element could belong to several of the sets, and hence there are 2^3 choices for each element.

By symmetry, every set in $\binom{[n]}{k}$ is in the same number of non-trivially intersecting families of size t. Hence, averaging over all sets, we find that each set $F \in \mathcal{F}$ can be in at most

$$tkn\left(\frac{8k^{4k}}{n^{1/2k}}\right)^t \binom{\binom{n-1}{k-1}}{t} \Big/ \binom{n}{k} = k^2\left(\frac{8k^{4k}}{n^{1/2k}}\right)^t \binom{\binom{n-1}{k-1}-1}{t-1} < n^{-t/4k}\binom{\binom{n-1}{k-1}}{t-1}$$

for sufficiently large n.

Summing over the m sets $F \in \mathcal{F}$, the number of non-trivially intersecting families of size t in $\mathcal{F}$ is no larger than

$$mn^{-t/4k}\binom{\binom{n-1}{k-1}-1}{t-1}/t \leq (\ell+1)n^{-t/4k}\binom{n-1}{k-1}\binom{\binom{n-1}{k-1}-1}{t-1}/t = (\ell+1)n^{-t/4k}\binom{\binom{n-1}{k-1}}{t},$$

thus giving the desired bounds.

□

5.4 Further remarks

In this chapter, we have extended the shifting arguments of [DGS14a] to determine which uniform families of sets are most probably intersecting. To derive the probabilistic result, we studied the counting version of the problem, finding families with the maximum number of intersecting subfamilies of any given size.

In particular, for graphs we showed that, provided the graphs are not too dense, the initial segment of the lexicographic order $\mathcal{L}_{n,2}(m)$ maximises the number of intersecting subgraphs with t edges. This leaves open the question for denser graphs, on which we provide some remarks.

In the case $t \geq \frac{n}{2}$, it is easy to show by shifting that $\mathcal{L}_{n,2}(m)$ is optimal for any m. Indeed, suppose we have a graph with vertices x, y, z of degrees $d_x \leq d_y \leq d_z < n-1$, and suppose $\{x, y\}$ is an edge of the graph. The number of stars this edge is contained in is $\binom{d_x-1}{t-1} + \binom{d_y-1}{t-1}$. On the other hand, if we were to add an edge containing z, it would be contained in at least $\binom{d_z}{t-1}$ stars. Since $t \geq \frac{n}{2}$, we have $t-1 \geq \frac{n-2}{2} > \frac{d_x-1}{2}$, and so

$$\binom{d_x-1}{t-1}+\binom{d_y-1}{t-1} \leq \binom{d_x-1}{t-2}+\binom{d_y-1}{t-1} \leq \binom{d_y-1}{t-2}+\binom{d_y-1}{t-1} = \binom{d_y}{t-1} \leq \binom{d_z}{t-1}.$$

Hence we may always shift edges to the vertex of highest degree until that star is filled. Repeating the process for the remaining vertices, we obtain a graph isomorphic to $\mathcal{L}_{n,2}(m)$, and hence $\mathcal{L}_{n,2}(m)$ maximises $\mathrm{int}(G, t)$ over all graphs G with m edges.

By the theorem of Ahlswede-Katona [AK78], we know for $m \geq \frac{1}{2}\binom{n}{2} + \frac{n}{2}$, the number of intersecting pairs of edges is maximised not by $\mathcal{L}_{n,2}(m)$, but by its complement, $\mathcal{C}_{n,2}(m)$.

Hence for such m we cannot hope to have one graph G that simultaneously maximises the number of intersecting subgraphs of all given orders. Referring to Equation (5.1), it follows that in this regime the most probably intersecting graph depends on the probability p. For very small values of p, $\mathcal{C}_{n,2}(m)$ is optimal, while for very large values of p, $\mathcal{L}_{n,2}(m)$ is better.

However, the convexity of the binomial coefficients (see, for instance, Lemma 5.3.2), suggests that if $\mathcal{L}_{n,2}(m)$ maximises $\mathrm{int}(G,t)$, then it should maximise $\mathrm{int}(G,t')$ for all $t' \geq t$. In particular, we believe that the result in Theorem 5.1.1 should extend to $m \leq \frac{1}{2}\binom{n}{2} - \frac{n}{2}$.

In the case of hypergraphs, the situation is even more intricate. We showed that when $\binom{n}{k} - \binom{n-\ell}{k} \leq m \leq \binom{n}{k} - \binom{n-\ell}{k} + n - \ell - k + 1$, $\mathcal{L}_{n,k}(m)$ maximises $\mathrm{int}(\mathcal{F},t)$. Thus we are able to determine the extremal families for the counting problem for a number of isolated ranges of family sizes. One might hope that, as in the graph case, $\mathcal{L}_{n,k}(m)$ remains optimal between these ranges as well. However, we show now that this is not the case.

Suppose, for simplicity, that we are counting the number of intersecting subfamilies of size three in 3-uniform hypergraphs, whose number of edges is between one and two full stars. Then $m = \binom{n-1}{2} + m'$, where $0 \leq m' \leq \binom{n-2}{2}$. Provided we do not have two almost-full stars, Proposition 5.3.3 shows that any extremal family is of the form $\mathcal{F} = \mathcal{F}_0 \cup \mathcal{F}_1$, where $\mathcal{F}_0$ is a full star, and $\mathcal{F}_1$ consists of the remaining m' sets.

There are four types of intersecting subfamilies of three sets: those with 0, 1, 2 and 3 sets from $\mathcal{F}_1$ respectively. To maximise the number of subfamilies with 3 sets from $\mathcal{F}_1$, it suffices to take $\mathcal{F}_1$ to be intersecting. The number of subfamilies with 0 and 1 sets from $\mathcal{F}_1$ is independent of the structure of $\mathcal{F}_1$. Finally, to maximise the number of subfamilies with two sets from $\mathcal{F}_1$, it follows from the calculations in Proposition 5.3.4 that we should seek to maximise the number of pairs of sets in $\mathcal{F}_1$ that intersect in two elements.

Note that in $\mathcal{L}_{n,3}(m)$, the sets in $\mathcal{F}_1$ all share a common element. If we remove this common element, $\mathcal{F}_1$ will be the lexicographic graph with m' edges. Since we have removed a common element from each set, we are trying to maximise the number of pairs of intersecting edges. By the result in [AK78], if $m' > \frac{1}{2}\binom{n-2}{2} + \frac{n-2}{2}$, this maximum is attained by the colexicographic graph instead, and hence it follows that $\mathcal{L}_{n,3}(m)$ does not maximise $\mathrm{int}(\mathcal{F},t)$.

This phenomenon holds in general, and shows that determining the exact optimal k-uniform families for all $\binom{n}{k} - \binom{n-\ell+1}{k} \leq m \leq \binom{n}{k} - \binom{n-\ell}{k}$ may require a complete solution to the counting problem for the number of t-intersecting subfamilies of a $(k-1)$-uniform set family. Indeed, it further suggests that even in this initial range, there may not be one set family that simultaneously maximises the number of intersecting subfamilies of any given size, and thus the optimal families may depend on the probability p.

Finally, as with the results in [KKK12], [Rus12] and [RW13], the extremal families we obtain here are simultaneously optimal for the counting problems as well, and thus we use Equation (5.1) to resolve the probabilistic problem. It would be very interesting to develop techniques to attack the probabilistic problem directly, as one might then find a complete solution even in the regime where the optimal family depends on the underlying probability p.

References

[AC99] R. Ahlswede and N. Cai. "A counterexample to Kleitman's conjecture concerning an edge-isoperimetric problem." *Comb. Probab. Comput.*, **8**(4):301–305, 1999.

[AC06] R. Ahlswede and N. Cai. "Appendix: on edge-isoperimetric theorems for uniform hypergraphs." In *General Theory of Information Transfer and Combinatorics*, pp. 979–1005. Springer, 2006.

[AES74] B. Andrásfai, P. Erdős, and V. Sós. "On the connection between chromatic number, maximal clique and minimal degree of a graph." *Discrete Math.*, **8**:205–218, 1974.

[AFN09] S. Ábrego, S. Fernández-Merchant, M. G. Neubauer, and W. Watkins. "Sum of squares of degrees in a graph." *J. Inequal. Pure Appl. Math.*, **10**(3):1–34, 2009.

[Ahl80] R. Ahlswede. "Simple hypergraphs with maximal number of adjacent pairs of edges." *J. Comb. Theory B*, **28**:164–167, 1980.

[AJM03] N. Alon, T. Jiang, Z. Miller, and D. Pritikin. "Properly colored subgraphs and rainbow subgraphs in edge-colorings with local constraints." *Random Struct. Algor.*, **23**:409–433, 2003.

[AK78] R. Ahlswede and G. O. H. Katona. "Graphs with maximal number of adjacent pairs of edges." *Acta Math. Hung.*, **32**(1):97–120, 1978.

[AK97] R. Ahlswede and L. H. Khachatrian. "The complete intersection theorem for systems of finite sets." *Eur. J. Combin.*, **18**(2):125–136, 1997.

[And87] I. Anderson. *Combinatorics of Finite Sets*. Courier Dover Publications, 1987.

[AS08] N. Alon and J. Spencer. *The Probabilistic Method*. John Wiley & Sons, 2008.

[BC62] R. C. Bose and S. Chowla. "Theorems in the additive theory of numbers." *Comment. Math. Helv.*, **37**:141–147, 1962.

[BF92] L. Babai and P. Frankl. *Linear Algebra Methods in Combinatorics with Applications to Geometry and Computer Science*. Department of Computer Science, The University of Chicago, 1992.

[BL03] B. Bollobás and I. Leader. "Set systems with few disjoint pairs." *Combinatorica*, **23**(4):559–570, 2003.

[Bol04] B. Bollobás. *Extremal Graph Theory*. Courier Dover Publications, 2004.

[BS74] J. A. Bondy and M. Simonovits. "Cycles of even length in graphs." *J. Comb. Theory B*, **16**:97–105, 1974.

[BT11] R. Baber and J. Talbot. "Hypergraphs do jump." *Comb. Probab. Comput.*, **20**(2):161–171, 2011.

[CFG88] F. R. K. Chung, Z. Füredi, R. L. Graham, and P. Seymour. "On induced subgraphs of the cube." *J. Comb. Theory A*, **49**:180–187, 1988.

[DGK14] A. P. Dove, J. R. Griggs, R. J. Kang, and J. S. Sereni. "Supersaturation in the Boolean lattice." *Integers*, 2014. to appear.

[DGS14a] S. Das, W. Gan, and B. Sudakov. "The minimum number of disjoint pairs in set systems and related problems." arXiv: 1305.6715, 2014.

[DGS14b] S. Das, W. Gan, and B. Sudakov. "Sperner's theorem and a problem of Erdős, Katona and Kleitman." *Comb. Probab. Comput.*, 2014. to appear.

[EKR61] P. Erdős, C. Ko, and R. Rado. "Intersection theorems for systems of finite sets." *Q. J. Math.*, **12**(1):313–320, 1961.

[ER50] P. Erdős and R. Rado. "A combinatorial theorem." *J. London Math. Soc.*, **25**:249–255, 1950.

[Erd62a] P. Erdős. "On a theorem of Rademacher–Turán." *Illinois J. Math.*, **6**:122–127, 1962.

[Erd62b] P. Erdős. "On the number of complete subgraphs contained in certain graphs." *Publ. Math. Inst. Hung. Acad. Sci.*, **7**:459–464, 1962.

[Erd62c] P. Erdős. "On the number of complete subgraphs contained in certain graphs." *Magy. Tud. Acad. Mat. Kut. Int. Közl.*, **7**:459–474, 1962.

[Erd65] P. Erdős. "A problem on independent r-tuples." *Ann. Univ. Sci. Budapest Eötvös Sect. Math.*, **8**:93–95, 1965.

[ES46] P. Erdős and A. H. Stone. "On the structure of linear graphs." *B. Am. Math. Soc.*, **52**:1087–1091, 1946.

[Fra77] P. Frankl. "On the minimum number of disjoint pairs in a family of finite sets." *J. Comb. Theory A*, **22**:249–251, 1977.

[Fra13] P. Frankl. "Improved bounds for Erdős' matching conjecture." *J. Comb. Theory A*, **120**:1068–1072, 2013.

[FRR12] P. Frankl, V. Rödl, and A. Ruciński. "On the maximum number of edges in a triple system not containing a disjoint family of a given size." *Comb. Probab. Comput.*, **21**:141–148, 2012.

[FV13] V. Falgas-Ravry and E. R. Vaughan. "Applications of the semi-definite method to the Turán density problem for 3-graphs." *Comb. Probab. Comput.*, **22**(1):21–54, 2013.

[GLP12] D. Gerbner, N. Lemons, C. Palmer, B. Patkós, and V. Szécsi. "Almost intersecting families of sets." *SIAM J. Discrete Math.*, **26**:1657–1669, 2012.

[GLS87] A Gyárfás, J. Lehel, R. Schelp, and Z. Tuza. "Ramsey numbers for local colorings." *Graph. Combinator.*, **3**:267–277, 1987.

[HHK12] H. Hatami, J. Hladký, D. Král', S. Norine, and A. Razborov. "Non-three-colorable common graphs exist." *Comb. Probab. Comput.*, **21**(5):734–742, 2012.

[HHK13] H. Hatami, J. Hladký, D. Král', S. Norine, and A. Razborov. "On the number of pentagons in triangle-free graphs." *J. Comb. Theory A*, **120**(3):722–732, 2013.

[HLS12] H. Huang, P. Loh, and B. Sudakov. "The size of a hypergraph and its matching number." *Comb. Probab. Comput.*, **21**:442–450, 2012.

[HM67] A. J. W. Hilton and E. C. Milner. "Some intersection theorems for systems of finite sets." *Q. J. Math.*, **18**(1):369–384, 1967.

[Juk01] S. Jukna. *Extremal Combinatorics (Vol. 2).* Springer, 2001.

[Kee11] P. Keevash. "Hypergraph Turán problems." In *Surveys in Combinatorics*, pp. 83–140. Cambridge University Press, 2011.

[KKK12] G. O. H. Katona, G. Y. Katona, and Z. Katona. "Most probably intersecting familes of subsets." *Comb. Probab. Comput.*, **21**:219–227, 2012.

[KMS07] P. Keevash, D. Mubayi, B. Sudakov, and J. Verstraëte. "Rainbow Turán problems." *Comb. Probab. Comput.*, **16**:109–126, 2007.

[LM12] T. Łuczak and K. Mieczkowska. "On Erdős' extremal problem on matchings in hypergraphs." arXiv: 1202.4196, 2012.

[Lor62] G. Lorden. "Blue-empty chromatic graphs." *Am. Math. Mon.*, **69**:114–120, 1962.

[Man07] W. Mantel. "Problem 28." *Winkundige Opgaven*, **10**:60–61, 1907.

[McD98] Colin McDiarmid. "Concentration." In *Probabilistic methods for algorithmic discrete mathematics*, pp. 195–248. Springer, 1998.

[Nik01] V. Nikiforov. "On the minimum number of k-cliques in graphs with restricted independence number." *Comb. Probab. Comput.*, **10**:361–366, 2001.

[Nik05] V. Nikiforov. "The minimum number of 4-cliques in a graph with triangle-free complement." Preprint, 2005.

[PT87] S. Poljak and Z. Tuza. "Maximum bipartite subgraphs of Kneser graphs." *Graph. Combinator.*, **3**:191–199, 1987.

[PV13] O. Pikhurko and E. R. Vaughan. "Minimum Number of k-Cliques in Graphs with Bounded Independence Number." *Comb. Probab. Comput.*, **22**(6):910–934, 2013.

[Raz07] A. Razborov. "Flag algebras." *J. Symbolic Logic*, **72**(4):1239–1282, 2007.

[Raz08] A. Razborov. "On the minimum density of triangles in graphs." *Comb. Probab. Comput.*, **17**(4):603–618, 2008.

[Raz10] A. Razborov. "On 3-hypergraphs with forbidden 4-vertex configurations." *SIAM J. Discrete Math.*, **24**:946–963, 2010.

[Raz13] A. Razborov. "Flag algebras: an interim report." In *The Mathematics of Paul Erdős II*, pp. 207–232. Springer, 2013.

[Rus12] P. A. Russell. "Compressions and probably intersecting families." *Comb. Probab. Comput.*, **21**(1–2):301–313, 2012.

[RW13] P. A. Russell and M. Walters. "Probably intersecting families are not nested." *Comb. Probab. Comput.*, **22**(1):146–160, 2013.

[Tur41] P. Turán. "On an extremal problem in graph theory." *Matematikai és Fizikai Lapok*, **48**:436–452, 1941.

[Wil84] R. M. Wilson. "The exact bound on the Erdős–Ko–Rado theorem." *Combinatorica*, **4**:247–257, 1984.

[Zyk49] A. Zykov. "On some properties of linear complexes." *Mat. Sbornik N. S.*, **24**(66):163–188, 1949.