Comparable pairs in families of sets

Noga Alon [*]   Shagnik Das [†]   Roman Glebov [‡]   Benny Sudakov [§]

December 13, 2015

Abstract

Given a family $\mathcal{F}$ of subsets of $[n]$, we say two sets $A, B \in \mathcal{F}$ are *comparable* if $A \subset B$ or $B \subset A$. Sperner's celebrated theorem gives the size of the largest family without any comparable pairs. This result was later generalised by Kleitman, who gave the minimum number of comparable pairs appearing in families of a given size.

In this paper we study a complementary problem posed by Erdős and Daykin and Frankl in the early '80s. They asked for the maximum number of comparable pairs that can appear in a family of m subsets of $[n]$, a quantity we denote by $c(n, m)$. We first resolve an old conjecture of Alon and Frankl, showing that $c(n, m) = o(m^2)$ when $m = n^{\omega(1)} 2^{n/2}$. We also obtain more accurate bounds for $c(n, m)$ for sparse and dense families, characterise the extremal constructions for certain values of m, and sharpen some other known results.

1 Introduction

Extremal set theory, with its connections and applications to numerous other fields, has enjoyed tremendous growth in the last few decades. However, its origins date back much further, with many considering the classic theorem of Sperner [18] to be the starting point of the field. An *antichain* is a family of pairwise incomparable sets, and Sperner's theorem states that the largest antichain over $[n]$ has $\binom{n}{\lfloor n/2 \rfloor}$ sets. This bound is attained by the middle levels of the Boolean hypercube; that is, the families of sets of size $\lfloor n/2 \rfloor$ or of sets of size $\lceil n/2 \rceil$.

1.1 Comparable pairs

Given this extremal result, it is natural to ask how many comparable pairs may appear in set families over $[n]$ of a given size. For the minimisation problem, Sperner's theorem shows that there need not be any comparable pairs in families of size $m \leq \binom{n}{\lfloor n/2 \rfloor}$. Kleitman [14] later completely resolved the problem for larger families, showing there is a nested sequence of extremal families consisting of sets as close to the middle levels as possible. He further conjectured that the same families also minimise the number of *k-chains*, which are collections of k pairwise-comparable sets $F_1 \subset F_2 \subset \ldots \subset F_k$, for every k. Erdős [8] had earlier generalised Sperner's theorem to show that

the largest k-chain-free family consisted of the $k-1$ middle levels of the hypercube. Das, Gan and Sudakov [4] verified Kleitman's conjecture for families of size up to the $k+1$ middle levels of the hypercube (Dove, Griggs, Kang and Sereni [6] independently obtained the same result for k middle levels), but the conjecture otherwise remains open.

Our focus is the corresponding maximisation question, a problem which has also attracted a great deal of attention. To this end, we denote by $c(\mathcal{F})$ the number of comparable pairs in the set family $\mathcal{F}$, and define $c(n, m)$ to be the maximum number of comparable pairs in a family of m subsets of $[n]$. We clearly have $c(n, m) \le \binom{m}{2}$, and can only have equality for an m-chain, which requires $m \le n+1$. Daykin and Frankl [5] showed that one may have much larger families where *almost all* pairs are comparable. More precisely, they proved $c(n, m) = (1 - o(1))\binom{m}{2}$ if and only if $m = 2^{o(n)}$.

Their lower bound came from a construction known as a *tower of cubes*, which generalises chains. A *subcube* of the hypercube $2^{[n]}$ is, for some $F_1 \subset F_2$, the family of sets $\{F \subset [n] : F_1 \subset F \subset F_2\}$. We say the dimension of the subcube is $|F_2| - |F_1|$. To define a tower of cubes, assume for simplicity n is divisible by k, and let $\ell = n/k$. Let $X_i = [i\ell]$ for $1 \le i \le k$, and consider the subcube $\mathcal{F}_i = \{F \subset [n] : X_{i-1} \subseteq F \subseteq X_i\}$. A tower of k cubes is the set family $\mathcal{F} = \bigcup_{i=1}^{k} \mathcal{F}_i$. We have $m = |\mathcal{F}| = k2^{n/k} - k + 1$. Moreover, two sets from different subcubes must be comparable, and so we have $c(\mathcal{F}) \ge \left(1 - \frac{1}{k}\right)\binom{m}{2}$. When $k = \omega(1)$, we have $c(\mathcal{F}) = (1 - o(1))\binom{m}{2}$, as in the theorem of Daykin and Frankl.

Alon and Frankl [1] proved the towers of cubes are asymptotically optimal even when k is constant, as shown by their theorem below.

Theorem 1.1. *For every positive integer k there exists a positive $\beta = \beta(k)$ such that if $m = 2^{(1/(k+1)+\delta)n}$ for $\delta > 0$, then*

$$c(n, m) < \left(1 - \frac{1}{k}\right)\binom{m}{2} + O\left(m^{2 - \beta\delta^{k+1}}\right).$$

The case $k = 1$ is of particular interest. In [1], Alon and Frankl prove that a family $\mathcal{F}$ of size $m = 2^{(1/2+\delta)n}$ must have $c(\mathcal{F}) < 4m^{2-\delta^2/2} = o(m^2)$, thus proving a conjecture of Daykin and Erdős [11]. Erdős [10] had made a finer conjecture, asking whether $m = \omega(2^{n/2})$ implies $c(n, m) = o(m^2)$. Alon and Frankl disproved this conjecture, exhibiting for any $d \ge 1$ a family $\mathcal{F}$ of size $\Omega(n^d 2^{n/2})$ with $c(\mathcal{F}) \ge 2^{-2d-1}\binom{m}{2}$. They in turn conjectured that this construction, which we describe in detail in Section 2, is essentially the best possible.

Conjecture 1.2. *If $m = n^{\omega(1)}2^{n/2}$, then $c(n, m) = o(m^2)$.*

1.2 Our results

In this paper, we further the study of the maximum number of comparable pairs in a set family. We begin by proving Conjecture 1.2. We shall in fact prove the slightly more general two-family version below. Given two set families $\mathcal{A}$ and $\mathcal{B}$, we write $c(\mathcal{A}, \mathcal{B})$ for the number of pairs $(A, B) \in \mathcal{A} \times \mathcal{B}$ with $A \subset B$.

Theorem 1.3. *If $\mathcal{A}$ and $\mathcal{B}$ are set families over $[n]$ with $|\mathcal{A}||\mathcal{B}| = n^d 2^n$, then $c(\mathcal{A}, \mathcal{B}) \le 2^{-d/300}|\mathcal{A}||\mathcal{B}|$.*

Conjecture 1.2 follows easily, since for any set family $\mathcal{F}$ with $|\mathcal{F}| = m = n^{\omega(1)}2^{n/2}$, Theorem 1.3 implies $c(\mathcal{F}) = c(\mathcal{F}, \mathcal{F}) \le 2^{-\omega(1)}|\mathcal{F}|^2 = o(m^2)$.

We next strengthen Theorem 1.1 by proving a stability result, showing families with close to $(1 - 1/k)\binom{m}{2}$ comparable pairs must be close in structure to a tower of k cubes.

Theorem 1.4. *For every $\varepsilon > 0$ and integer $k \geq 2$, there is an $\eta > 0$ such that for sufficiently large n, if a set family $\mathcal{F}$ over $[n]$ of size $m = |\mathcal{F}| \geq (1 - \eta)k2^{n/k}$ has at least $\left(1 - \frac{1}{k} - \eta\right)\binom{m}{2}$ comparable pairs, then all but at most εm sets in $\mathcal{F}$ are contained inside a tower of k cubes of dimension n/k.*

As a consequence, we are able to show that the towers of cubes are (uniquely) extremal.

Corollary 1.5. *Given $k \geq 2$ and n sufficiently large, if $k|n$ and $\mathcal{F}$ is a set family over $[n]$ of size $m = |\mathcal{F}| = k2^{n/k} - k + 1$ maximising the number of comparable pairs, then $\mathcal{F}$ is a tower of k cubes of dimension n/k.*

For sparser families, when $m = 2^{o(n)}$, we refine the theorem of Daykin and Frankl by using Theorem 1.1 to determine the value of $c(n, m)$ more precisely. Since almost all pairs are comparable in this regime, we instead count the number of incomparable pairs in a family $\mathcal{F}$, which we denote by $i(\mathcal{F})$. We denote by $i(n, m)$ the minimum number of incomparable pairs in a family $\mathcal{F}$ of m subsets of $[n]$. Daykin and Frankl showed we have $i(n, m) = o(m^2)$ in this range, and we provide asymptotic lower bounds. In this setting, it is more convenient to parameterise $m = |\mathcal{F}| = n\ell$.

Theorem 1.6. *Given $\varepsilon > 0$, for ℓ and n sufficiently large we have $i(n, n\ell) \geq (1/2 - \varepsilon)\, n\ell^2 \log \ell$.*

The above theorem shows that in the sparse regime, the towers of cubes are again asymptotically optimal in this finer sense. Indeed, a tower of k cubes has $|\mathcal{F}| = n\ell = k2^{n/k} - k + 1$ and $i(\mathcal{F}) \approx \frac{1}{k}\binom{n\ell}{2}$, since almost all pairs of sets from the same subcube are incomparable. This shows that Theorem 1.6 is asymptotically tight, as $n\ell^2 \log \ell / 2 \approx \frac{\log \ell}{n}\binom{n\ell}{2} \approx \frac{1}{k}\binom{n\ell}{2}$.

The above results all hold for very sparse families, where $m = |\mathcal{F}|$ is much smaller than 2^n, the total number of subsets of $[n]$. We finally turn our attention to dense families, where we show the extremal families are of a very different nature. Note that, within the complete hypercube $2^{[n]}$, the sets of size $n/2$ are in the fewest comparable pairs, while the sets of extreme size, namely $\emptyset$ and $[n]$, are in the most. Intuitively, we might expect that families maximising the number of comparable pairs should avoid sets near the middle levels, and using shifting arguments we are able to prove this is the case. Given $0 \leq k \leq n/2$, define $\mathcal{H}_k = \{F \subset [n] : |F| \leq k\} \cup \{F \subset [n] : |F| \geq n - k\}$, and let $M_k = |\mathcal{H}_k| = 2\sum_{i \leq k}\binom{n}{i}$.

Theorem 1.7. *If $M_{k-1} \leq m \leq M_k$ for some k with $n/3 + \sqrt{2n\ln 2} \leq k \leq n/2$, then every family $\mathcal{F}$ of m sets over $[n]$ maximising the number of comparable pairs satisfies $\mathcal{H}_{k-1} \subset \mathcal{F} \subset \mathcal{H}_k$.*

The entropy bound (see Lemma 2.3) gives the estimate $\sum_{i \leq pn}\binom{n}{i} \leq 2^{H(p)n}$, where $H(p) = -p \log p - (1 - p)\log(1 - p)$. It follows that Theorem 1.7 applies when $m \geq 2^{0.92n}$, giving the large-scale structure of extremal families in this range. To determine $c(n, m)$ precisely, one must prescribe which sets of size k and $n - k$ should be chosen. We are able to do so for some special values of m, which we discuss further in Section 4.

1.3 Notation and organisation

For a family of sets $\mathcal{F}$ and $x \in [n]$, we define $\mathcal{F}(x) = \{F \in \mathcal{F} : x \in F\}$. We say an element x is *η-dense* in $\mathcal{F}$ if $|\mathcal{F}(x)| \geq \eta|\mathcal{F}|$; that is, it is contained in at least an η-proportion of sets in $\mathcal{F}$. We say x is *covered* by $\mathcal{F}$ if it is contained in some set in $\mathcal{F}$, i.e. $|\mathcal{F}(x)| \geq 1$. The remainder of our set notation is standard. All asymptotics are taken as $n \to \infty$. We use log for the binary logarithm and ln for the natural logarithm.

In Section 2, we prove Theorem 1.3, settling Conjecture 1.2. We next extend the known results for sparse families in Section 3, proving Theorems 1.4 and 1.6. In Section 4 we turn to dense families, proving Theorem 1.7. Finally, Section 5 contains some concluding remarks and open questions.

2 The Alon–Frankl Conjecture

In this section we will prove Theorem 1.3, thus resolving Conjecture 1.2. We begin, however, by describing the construction given in [1], which motivates some of the ideas behind the proof.

Recall that our goal is to construct, for every constant d, a family $\mathcal{F}$ of size $\Omega\left(n^d 2^{n/2}\right)$ with a positive density of comparable pairs. For simplicity, we shall assume n is even. When $d = 0$, the natural construction is to take a tower of two cubes, namely all subsets of $[n/2]$ and all sets containing $[n/2]$. Since any set of the first type is comparable to any set of the second, this gives a family $\mathcal{F}$ of $2 \cdot 2^{n/2}$ sets with at least $\frac{1}{2}\binom{|\mathcal{F}|}{2}$ comparable pairs.

To obtain larger families, we 'fatten' the construction. Let $\mathcal{F} = \mathcal{F}_1 \cup \mathcal{F}_2$, where $\mathcal{F}_1 = \{F \subset [n] : |F \setminus [n/2]| \leq d\}$ and $\mathcal{F}_2 = \{F \subset [n] : |[n/2] \setminus F| \leq d\}$, so that sets of the first type are now allowed to have at most d elements outside $[n/2]$, while sets of the second type can miss up to d elements from $[n/2]$. This clearly gives a family of size $|F| = \Omega(n^d 2^{n/2})$, and it is not difficult to check that $c(\mathcal{F}_1, \mathcal{F}_2) \geq 2^{-2d-1}\binom{|F|}{2}$.

In this construction, the elements of $[n]$ are of two types. Those in $[n/2]$ are in half the sets from $\mathcal{F}_1$ and almost all the sets from $\mathcal{F}_2$, while those in $[n] \setminus [n/2]$ are in very few of the sets from $\mathcal{F}_1$ and in half the sets from $\mathcal{F}_2$. Our proof below, for the two-family version of the problem, is guided by an attempt to recover this partition of the ground set in an extremal construction; this motivation is obvious in the proof of Lemma 2.1, but such a partition is also a by-product of the calculations in Lemma 2.2.

Proof of Theorem 1.3. Our proof is by induction on n. It is easy to see the statement holds for $n = 2$, by checking it for all choices of $\mathcal{A}$ and $\mathcal{B}$.[1]

We now proceed to the induction step with $n \geq 3$. Note that the statement is trivial for $d \leq 0$, since we always have $c(\mathcal{A}, \mathcal{B}) \leq |\mathcal{A}||\mathcal{B}|$. Now consider the case $d \in (0, 1]$, and let $\mathcal{A}$ and $\mathcal{B}$ be two families with $c(\mathcal{A}, \mathcal{B}) \geq 2^{-d/300}|\mathcal{A}||\mathcal{B}|$. We have $2^{-d/300} \geq 2^{-1/300} > 0.99$, so the comparable pairs are very dense indeed. In this case, the following lemma, to be proven later, gives the required bound on $|\mathcal{A}||\mathcal{B}|$.

Lemma 2.1. *For $n \geq 2$ and $d \in (0, 1]$, let $\mathcal{A}$ and $\mathcal{B}$ be two families of subsets of $[n]$ satisfying $c(\mathcal{A}, \mathcal{B}) \geq 2^{-d/300}|\mathcal{A}||\mathcal{B}|$. Then $|\mathcal{A}||\mathcal{B}| < n^d 2^n$.*

Hence we may assume $d \geq 1$. In this case, we shall project the families onto $[n-1]$ and apply induction. Let $\mathcal{A}_0 = \{A \in \mathcal{A} : n \notin A\}$ and $\mathcal{A}_1 = \{A \subset [n-1] : A \cup \{n\} \in \mathcal{A}\}$, and define $\mathcal{B}_0$ and $\mathcal{B}_1$ similarly; note that these families are supported on $[n-1]$. We have

$$c(\mathcal{A}, \mathcal{B}) = c(\mathcal{A}_0, \mathcal{B}_0) + c(\mathcal{A}_0, \mathcal{B}_1) + c(\mathcal{A}_1, \mathcal{B}_1). \tag{1}$$

Now define $p = |\mathcal{A}_0|/|\mathcal{A}|$ to be the probability that n is *not* in a random set of $\mathcal{A}$, and $q = |\mathcal{B}_1|/|\mathcal{B}|$ to be the probability that n *is* in a random set of $\mathcal{B}$. We then have, for instance, $|\mathcal{A}_0||\mathcal{B}_0| = p(1-q)|\mathcal{A}||\mathcal{B}| = p(1-q)n^d 2^n$. However, to apply the induction hypothesis, we wish to rewrite this in terms of $n-1$. Define d_{00} to be such that $|\mathcal{A}_0||\mathcal{B}_0| = p(1-q)n^d 2^n = (n-1)^{d_{00}} 2^{n-1}$. A simple calculation reveals $d_{00} = d + \log_{n-1}\left(2p(1-q)\left(1 + \frac{1}{n-1}\right)^d\right)$. Applying the induction hypothesis, we deduce

$$c(\mathcal{A}_0, \mathcal{B}_0) \leq 2^{-\frac{d_{00}}{300}}|\mathcal{A}_0||\mathcal{B}_0| = p(1-q)\left(2p(1-q)\left(1 + \frac{1}{n-1}\right)^d\right)^{-\frac{1}{300\log(n-1)}} 2^{-\frac{d}{300}}|\mathcal{A}||\mathcal{B}|.$$

---

[1]Indeed, we may assume $\mathcal{A}$ is a left-compressed down-set and $\mathcal{B}$ is a left-compressed up-set, and so we only need to check the families $\mathcal{A} \in \{\{\emptyset\}, \{\emptyset, \{1\}\}, \{\emptyset, \{1\}, \{2\}\}, \{\emptyset, \{1\}, \{2\}, \{1,2\}\}\}$ and $\mathcal{B} \in \{\{\{1,2\}\}, \{\{1\}, \{1,2\}\}, \{\{1\}, \{2\}, \{1,2\}\}, \{\emptyset, \{1\}, \{2\}, \{1,2\}\}\}$.

Performing similar calculations for the other two terms and then substituting into (1), we have

$$c(\mathcal{A}, \mathcal{B}) \le \left[(p(1-q))^{1-\alpha} + (pq)^{1-\alpha} + ((1-p)q)^{1-\alpha} \right] \left(2 \left(1 + \frac{1}{n-1} \right)^d \right)^{-\alpha} 2^{-\frac{d}{300}} |\mathcal{A}| |\mathcal{B}|,$$

where $\alpha = 1/(300 \log(n-1))$. Note that $n - 1 = 2^{1/(300\alpha)}$, and, since $d \ge 1$, we have

$$c(\mathcal{A}, \mathcal{B}) \le \left[(p(1-q))^{1-\alpha} + (pq)^{1-\alpha} + ((1-p)q)^{1-\alpha} \right] \left(2 + 2^{1-\frac{1}{300\alpha}} \right)^{-\alpha} 2^{-\frac{d}{300}} |\mathcal{A}| |\mathcal{B}|. \tag{2}$$

We shall later prove the following analytic inequality.

Lemma 2.2. *For $p, q \in [0, 1]$ and $0 < \alpha \le 1/300$, we have*

$$(p(1-q))^{1-\alpha} + (pq)^{1-\alpha} + ((1-p)q)^{1-\alpha} \le \left(2 + 2^{1-\frac{1}{300\alpha}} \right)^{\alpha}. \tag{3}$$

This completes the induction, since substituting (3) into (2) gives $c(\mathcal{A}, \mathcal{B}) \le 2^{-d/300} |\mathcal{A}| |\mathcal{B}|$, as required. □

We now give the proofs of the two lemmas.

Proof of Lemma 2.1. Let $\varepsilon = 1 - 2^{-d/300}$, so that we have $c(\mathcal{A}, \mathcal{B}) \ge (1 - \varepsilon) |\mathcal{A}| |\mathcal{B}|$. We now pass to subfamilies with every set in many comparable pairs. Let $\mathcal{A}' = \left\{ A \in \mathcal{A} : c(\{A\}, \mathcal{B}) \ge \frac{19}{20} |\mathcal{B}| \right\}$ and $\mathcal{B}' = \left\{ B \in \mathcal{B} : c(\mathcal{A}, \{B\}) \ge \frac{19}{20} |\mathcal{A}| \right\}$. Given the number of comparable pairs between $\mathcal{A}$ and $\mathcal{B}$, it follows from Markov's Inequality that $|\mathcal{A}'| \ge (1 - 20\varepsilon) |\mathcal{A}|$ and $|\mathcal{B}'| \ge (1 - 20\varepsilon) |\mathcal{B}|$. Since

$$\frac{19}{20} - 20\varepsilon \ge \frac{19}{20} - 20 \left(1 - 2^{-\frac{1}{300}} \right) > \frac{9}{10},$$

for every $A \in \mathcal{A}'$ we have $c(\{A\}, \mathcal{B}') \ge c(\{A\}, \mathcal{B}) - |\mathcal{B} \setminus \mathcal{B}'| \ge \left(\frac{19}{20} - 20\varepsilon \right) |\mathcal{B}| > \frac{9}{10} |\mathcal{B}'|$, and similarly for every $B \in \mathcal{B}'$.

We now partition the elements of $[n]$ according to how often they appear in $\mathcal{A}'$. Let $[n] = R \cup S \cup T$, where R is the set of elements that are $(1/10)$-dense in $\mathcal{A}'$, S is the set of elements in $[n] \setminus R$ covered by $\mathcal{A}'$, and $T = [n] \setminus (R \cup S)$ is the set of elements not covered by $\mathcal{A}'$. We use the following result, which appears as Corollary 15.7.3 in [2], to bound the size of $\mathcal{A}'$.

Lemma 2.3. *Let $\mathcal{F}$ be a family of subsets of $\{1, 2, \ldots, n\}$ and let p_i denote the fraction of sets in $\mathcal{F}$ that contain i. Then*

$$|\mathcal{F}| \le 2^{\sum_{i=1}^{n} H(p_i)},$$

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

Note that the binary entropy $H(p)$ increases from 0 to 1 as p ranges from 0 to $1/2$, and then decreases back to 0 as p increases to 1. In our setting, we have $p_i \le 1/10$ for $i \in S$ and $p_i = 0$ for $i \in T$, while we could have $p_i = 1/2$ for $i \in R$. Thus $|\mathcal{A}'| \le 2^{|R| + H(1/10)|S|}$.

We now seek to bound $|\mathcal{B}'|$. By our minimum degree condition, every set in $\mathcal{B}'$ contains at least $9|\mathcal{A}'|/10$ sets in $\mathcal{A}'$, and hence must contain any element that is $(1/10)$-dense in $\mathcal{A}'$. Thus $R \subset B$ for all $B \in \mathcal{B}'$. Since every set in $\mathcal{A}'$ is contained in at least $9|\mathcal{B}'|/10$ sets in $\mathcal{B}'$, it follows that any element covered by $\mathcal{A}'$, and in particular those in S, must be $(9/10)$-dense in $\mathcal{B}'$. Thus, applying Lemma 2.3 once more, we have $|\mathcal{B}'| \le 2^{H(9/10)|S| + |T|}$.

Thus we have $|\mathcal{A}'| |\mathcal{B}'| \le 2^{|R| + (H(1/10) + H(9/10))|S| + |T|} = 2^{n - (1 - 2H(1/10))|S|} \le 2^n$, since $H(1/10) < 1/2$. Hence we conclude

$$|\mathcal{A}| |\mathcal{B}| \le (1 - 20\varepsilon)^{-2} |\mathcal{A}'| |\mathcal{B}'| < (1 - \varepsilon)^{-300} 2^n = 2^d 2^n \le n^d 2^n,$$

as claimed. □

We conclude this section with the proof of the second lemma. Here the aforementioned partition of the ground set $[n]$ is not immediately evident. However, the following calculations show that the left-hand side of (3) is maximised when one of p or q is close to $1/2$ while the other is close to 1. This implies that in an optimal construction, elements should essentially be fixed in one of $\mathcal{A}$ or $\mathcal{B}$ and free in the other, giving a partition as in the construction of Alon and Frankl.

Proof of Lemma 2.2. The inequality clearly holds on the boundary of the square, as its left-hand side takes a maximum of 1 when $p = 0$ or $q = 0$ and a maximum of 2^α when $p = 1$ or $q = 1$. Hence we only need to check the local maxima within the square.

Taking the derivative of the left-hand side of (3) with respect to p, we see that a local maximum can only be obtained at

$$p_0 = p_0(q) = 1 - \frac{q^{\frac{1}{\alpha}-1}}{(q^{1-\alpha} + (1-q)^{1-\alpha})^{\frac{1}{\alpha}} + q^{\frac{1}{\alpha}-1}}, \tag{4}$$

and the value of the left-hand side there is

$$\left(\left(q^{1-\alpha} + (1-q)^{1-\alpha} \right)^{\frac{1}{\alpha}} + q^{\frac{1}{\alpha}-1} \right)^\alpha.$$

Thus the inequality in the lemma reduces to showing

$$f(q) := \left(q^{1-\alpha} + (1-q)^{1-\alpha} \right)^{\frac{1}{\alpha}} + q^{\frac{1}{\alpha}-1} < 2 + 2^{1-\frac{1}{300\alpha}}. \tag{5}$$

We first limit the range of q. As the left-hand side of (3) is symmetric in p and q, we may restrict ourselves to local maxima (p, q) with $q \leq p = p_0(q)$. Suppose we had $q^{\frac{1}{\alpha}-1} > 1/10$. Since $\alpha \leq 1/300$, we have $1/10 < q^{\frac{1}{\alpha}-1} \leq q^{299}$, which implies $q > 0.99$. On the other hand, since $(q^{1-\alpha} + (1-q)^{1-\alpha})^{\frac{1}{\alpha}} \leq 2$ and $q^{\frac{1}{\alpha}-1} \leq 1$, (4) implies $p_0 < 1 - 1/30 < 0.99 < q$, a contradiction. Hence we may assume $q^{\frac{1}{\alpha}-1} \leq 1/10$.

Define $g(q) = q^{1-\alpha} + (1-q)^{1-\alpha}$ and $h(q) = g(q)^{\frac{1}{\alpha}} = \left(q^{1-\alpha} + (1-q)^{1-\alpha} \right)^{\frac{1}{\alpha}} \leq 2$. Therefore $f(q) \leq 2 + q^{\frac{1}{\alpha}-1}$ and the desired inequality holds unless $q^{\frac{1}{\alpha}-1} > 2^{1-\frac{1}{300\alpha}}$, which comfortably implies $q \geq 0.99$. Thus we may assume $0.99 \leq q \leq p_0$. To complete the proof we show that in this range $f(q) < 2$. Given our bound on $q^{\frac{1}{\alpha}-1}$, it suffices to show that $h(q) < 1.9$ in this range. As $g'(q) = (1-\alpha)\left(q^{-\alpha} - (1-q)^{-\alpha} \right)$, g is decreasing in $[1/2, 1]$, and hence so is $h(q)$. Thus it suffices to show that $h(0.99) < 1.9$. By the Mean Value Theorem, $h(0.99) = h(0.75) + 0.24h'(q_1)$ for some $q_1 \in [0.75, 0.99]$. However, since $g(q) \geq 1$ and $g'(q)$ is negative in this range,

$$h'(q_1) = \frac{1}{\alpha}g(q_1)^{\frac{1}{\alpha}-1}g'(q_1) \leq \frac{1}{\alpha}(1-\alpha)(q_1^{-\alpha} - (1-q_1)^{-\alpha})$$

$$= \frac{1}{\alpha}(1-\alpha)(-\alpha)q_2^{-1-\alpha}(2q_1 - 1) \leq -(1-\alpha)(2 \cdot 0.75 - 1) < -0.49,$$

where we have again used the Mean Value Theorem, this time for the function $z^{-\alpha}$, with $q_2 \in [1-q_1, q_1]$. We thus have $h(0.99) = h(0.75) + 0.24h'(q_1) \leq 2 - 0.24 \cdot 0.49 < 1.9$, implying the desired result. $\square$

3 Sparse families

In this section, we study sparse families further, extending the known results in this regime. In the first subsection, we prove Theorem 1.4, a stability result for Theorem 1.1 of Alon and Frankl [1]. The second subsection contains the proof of Theorem 1.6, which sharpens the result of Daykin and Frankl [5] by giving the asymptotics of the number of incomparable pairs in even sparser families.

3.1 The proof of Theorem 1.4

Following the original proof of Alon and Frankl, we shall determine the structure of a large set family with many comparable pairs by studying its comparability graph. The *comparability graph* $G_{\mathcal{F}}$ of a set family $\mathcal{F}$ has vertices $\mathcal{F}$ and an edge $\{A, B\}$ if A and B are comparable. Thus the number of edges in $G_{\mathcal{F}}$ is precisely the number of comparable pairs $c(\mathcal{F})$.

The key to our proof of Theorem 1.4 is the following lemma, which shows the comparability graph of such a large family cannot have many cliques of size $k+1$. This lemma essentially appears in [1] and [7], and we include its proof here for the sake of completeness.

Lemma 3.1. *For every $k \geq 2$ and $\gamma > 0$, there is an $n_0 = n_0(k, \gamma)$ such that for every $n \geq n_0$, if $\mathcal{F}$ is a set family over $[n]$ of size $m = |\mathcal{F}| \geq 2^{n/k}$, then $G_{\mathcal{F}}$ contains at most $\gamma\binom{m}{k+1}$ copies of K_{k+1}.*

To prove this lemma, we shall need a supersaturation result on complete multipartite graphs. Let $K_r(\ell)$ denote the complete r-partite graph with r parts of size ℓ. Given a graph G with many r-cliques K_r, form an auxiliary r-uniform hypergraph H over the same vertex set by placing an r-edge for each copy of K_r. The complete multipartite $K_r(\ell)$ in G then corresponds to a complete r-partite r-graph with parts of size ℓ in H. Erdős [9] showed such an r-partite r-graph must be contained in any sufficiently large r-graph of positive density, and a standard averaging argument gives the following proposition.

Proposition 3.2. *For integers $r \geq 2$, $\ell \geq 1$ and any real $\gamma > 0$, there is a $\delta_1 = \delta_1(r, \ell, \gamma)$ such that if n is sufficiently large and G is a graph on n vertices with at least $\gamma\binom{n}{r}$ copies of K_r, then G contains at least $\delta_1\binom{n}{r\ell}$ copies of $K_r(\ell)$.*

We may now proceed with the proof of Lemma 3.1. Our strategy is to show that a positive density of $(k+1)$-cliques implies a random sample of a large number of sets from $\mathcal{F}$ must contain many sets from a low-dimensional subcube with positive probability. On the other hand, a random sample from a family of this size gives rise to such a dense subcube with very low probability, giving the necessary contradiction.

Proof of Lemma 3.1. Suppose for contradiction we have a family $\mathcal{F}$ of size $m \geq 2^{n/k}$ with more than $\gamma\binom{m}{k+1}$ copies of K_{k+1}. While we wish to have a positive density of cliques, it would be convenient to know there are not many larger cliques. To this end, we define the recursive sequence

$$\gamma_{k+1} = \gamma, \text{ and } \gamma_r = \frac{\delta_1\left(r-1, 3k(k+1), \gamma_{r-1}\right)}{2\binom{3k(k+1)(r-1)}{r}} \text{ for } k+2 \leq r \leq 12k^2,$$

where δ_1 is as in Proposition 3.2. Note that $\gamma_r > 0$ is independent of n for each $k+1 \leq r \leq 12k^2$. Let t be the largest integer such that $t \leq 12k^2$ and $G_{\mathcal{F}}$ has more than $\gamma_t\binom{m}{t}$ copies of K_t. Note that $t \geq k+1$ by assumption and $t \leq 12k^2$ by definition. If $t = 12k^2$, $G_{\mathcal{F}}$ has more than $\gamma_t\binom{m}{t}$ copies of K_t. Otherwise, if $t < 12k^2$, $G_{\mathcal{F}}$ has more than $\gamma_t\binom{m}{t}$ copies of K_t and at most $\gamma_{t+1}\binom{m}{t+1}$ copies of K_{t+1}.

<u>Case 1:</u> $t = 12k^2$

Let $S = \{S_1, S_2, \ldots, S_t\}$ be a random sample of t distinct vertices of $G_{\mathcal{F}}$. Since $G_{\mathcal{F}}$ has more than $\gamma_t\binom{m}{t}$ cliques K_t, it follows that $G_{\mathcal{F}}[S]$ is a clique on t vertices with probability at least γ_t. If $G_{\mathcal{F}}[S]$ is complete, the corresponding sets in $\mathcal{F}$ must form a t-chain $F_1 \subset \ldots \subset F_t \subset [n]$. As $t = 12k^2$, we can find some j such that $|F_{j+6k}| - |F_j| \leq n/(2k)$, which implies the $6k+1$ sets $F_j \subset F_{j+1} \subset \ldots \subset F_{j+6k}$ lie in a subcube of dimension $n/(2k)$.

Now there are $2^{n-n/(2k)}\binom{n}{n/(2k)}$ such subcubes, each containing $2^{n/(2k)}$ sets, and so the probability that a random sample of t sets contains $6k$ sets from a single subcube of dimension $n/(2k)$ can be bounded above by

$$\binom{t}{6k} \cdot 2^{n-\frac{n}{2k}}\binom{n}{\frac{n}{2k}}\left(\frac{2^{\frac{n}{2k}}}{m}\right)^{6k} \leq 2^{t} \cdot 2^{2n}\left(2^{-\frac{n}{2k}}\right)^{6k} = 2^{t} \cdot 2^{-n},$$

since $m \geq 2^{n/k}$. Since every sample producing a t-clique must give rise to such a dense subcube, we deduce $\gamma_t \leq 2^t \cdot 2^{-n}$, which gives a contradiction when n is suitably large.

<u>Case 2:</u> $t < 12k^2$

We now deal with the case when $t < 12k^2$, where we have more than $\gamma_t\binom{m}{t}$ copies of K_t, but at most $\gamma_{t+1}\binom{m}{t+1}$ copies of K_{t+1}. Since the cliques we have may be relatively small, they will not be enough to find the dense subcubes we require. However, we will find large complete multipartite graphs which, coupled with the lack of larger cliques, will give us a large number of sets contained in a tower of small cubes.

Indeed, letting $\ell = 3k(k+1)$, Proposition 3.2 implies the density of $K_t(\ell)$ is at least $\delta_1 = \delta_1(t, \ell, \gamma_t)$. A random sample of $t\ell$ vertices of $G_{\mathcal{F}}$ therefore gives a copy of $K_t(\ell)$ with probability at least δ_1. On the other hand, the density of K_{t+1} is at most γ_{t+1}. Thus the probability that our $t\ell$ random vertices contain some $t+1$ vertices inducing a K_{t+1} is at most $\binom{t\ell}{t+1}\gamma_{t+1} \leq \delta_1/2$. Hence with probability at least $\delta_1/2$, a random sample of $t\ell$ vertices gives a K_{t+1}-free copy of $K_t(\ell)$.

If we consider the sets in $\mathcal{F}$ corresponding to this copy of $K_t(\ell)$ in $G_{\mathcal{F}}$, we find there are t parts $\mathcal{F}_i$, each consisting of ℓ sets $F_{i,j}$, such that for every $i \neq i'$ and j, j', $F_{i,j} \subset F_{i',j'}$ or $F_{i',j'} \subset F_{i,j}$. If we had some $i \neq i'$ and j, j', j'' such that $F_{i,j} \subset F_{i',j'} \subset F_{i,j''}$, then $F_{i,j}$ and $F_{i,j''}$ form a comparable pair within the part $\mathcal{F}_i$, which would give a copy of K_{t+1} in the corresponding subgraph of $G_{\mathcal{F}}$, a contradiction. Hence we may assume for each pair $i < i'$ and j, j', we have $F_{i,j} \subset F_{i',j'}$. Thus the sets are contained in the tower of t cubes determined by the chain $\left\{\bigcup_{i \leq p, 1 \leq j \leq \ell} F_{i,j} : 0 \leq p \leq t\right\}$. By considering the smallest cube, we find there must be ℓ sets from our sample in a subcube of dimension n/t.

Running the same probabilistic argument as before, the probability of finding ℓ sets from our sample within a subcube of dimension n/t can be bounded above, giving the inequality

$$\frac{1}{2}\delta_1 \leq \binom{t\ell}{\ell} \cdot 2^{n-\frac{n}{t}}\binom{n}{\frac{n}{t}}\left(\frac{2^{n/t}}{m}\right)^{\ell} \leq 2^{t\ell} \cdot 2^{2n}\left(2^{-\frac{n}{k(k+1)}}\right)^{\ell} = 2^{t\ell} \cdot 2^{-n},$$

since $t \geq k+1$ and $\ell = 3k(k+1)$. This again provides a contradiction for n sufficiently large, thus proving the lemma. $\qquad\square$

This lemma shows set families of size $m \geq 2^{n/k}$ cannot have many copies of K_{k+1} in their comparability graphs. We shall now employ some well-known graph theoretic results to derive further structural information for families with $c(\mathcal{F}) \geq (1 - (1+\eta)/k)\binom{m}{2}$. The first is the graph removal lemma, a generalisation of the triangle removal lemma of Ruzsa–Szemerédi [16], which shows that we can make the comparability graph K_{k+1}-free by removing very few edges. For a detailed account of the history of the removal lemma, see [3].

Theorem 3.3. *For every fixed graph H on h vertices and $\beta > 0$ there is a $\gamma > 0$ such that if G is a graph on m vertices with at most $\gamma\binom{m}{h}$ copies of H, then G can be made H-free by removing at most $\beta\binom{m}{2}$ edges.*

8

In light of Lemma 3.1, this shows $G_{\mathcal{F}}$ can be made K_{k+1}-free by removing at most $\beta\binom{m}{2}$ edges. We would thus arrive at a K_{k+1}-free graph with at least $(1 - 1/k - \eta - \beta)\binom{m}{2}$ edges, which is very close to the maximum possible. Using stability for Turán's theorem, we will deduce $G_{\mathcal{F}}$ is very close in structure to the extremal Turán graph $T_{m,k}$; that is, the balanced k-partite graphs. Such a stability result was first proved by Simonovits [17]. Here we use the following quantitative form of stability, whose short proof can be found in [13].

Theorem 3.4. *Suppose G is a K_{k+1}-free graph on m vertices with at least $(1 - 1/k - \nu)\binom{m}{2}$ edges and $\nu < 1/(4k)^4$. Then there is a partition of the vertex set of G as $V(G) = V_1 \cup V_2 \cup \ldots \cup V_k$ with $\sum e(V_i) < (2k+1)\nu^{1/2}m^2$.*

Thus, after removing few edges from $G_{\mathcal{F}}$ to make it K_{k+1}-free, we must be left with a graph that is nearly k-partite. To prove Theorem 1.4, it remains to transfer this structural information about $G_{\mathcal{F}}$ to $\mathcal{F}$ itself.

Proof of Theorem 1.4. To simplify the presentation, we shall adopt asymptotic notation. Given $\eta = o(k^{-10})$, we shall show that for any family $\mathcal{F}$ over $[n]$ of size $m = |\mathcal{F}| \geq (1 - \eta)k2^{n/k}$ with at least $(1 - 1/k - \eta)\binom{m}{2}$ comparable pairs, there is some tower of k cubes of dimension n/k containing all but $o(m)$ sets of $\mathcal{F}$.

Let $\beta = \eta$, and let γ be as in Theorem 3.3. By Lemma 3.1, if n is sufficiently large, $G_{\mathcal{F}}$ contains at most $\gamma\binom{m}{k+1}$ copies of K_{k+1}. By Theorem 3.3, $G_{\mathcal{F}}$ can be made K_{k+1}-free by removing $\beta\binom{m}{2}$ edges. Setting $\nu = \eta + \beta$, we then have a K_{k+1}-free graph on m vertices with at least $(1 - 1/k - \nu)\binom{m}{2}$ edges. By Theorem 3.4, removing a further $O(k\nu^{1/2}m^2) = o(m^2k^{-4})$ edges results in a k-partite subgraph with parts $V_1, V_2, \ldots, V_k$. Such a graph can have at most

$$\binom{m}{2} - \sum_i \binom{|V_i|}{2} = \left(1 - \frac{1}{k}\right)\binom{m}{2} - \frac{1}{2}\sum_i\left(|V_i| - \frac{m}{k}\right)^2 - O(m)$$

edges, from which it follows that $|V_i| = (1 - o(k^{-1}))mk^{-1}$ for $1 \leq i \leq k$. Since there are $\Omega(m^2k^{-2})$ possible edges between two parts V_i and V_j, and only $o(m^2k^{-4})$ edges missing from the k-partite subgraph, it follows that almost every edge is present between every pair of parts.

In other words, we can partition the sets of $\mathcal{F}$ into k families $\mathcal{F}_1, \ldots, \mathcal{F}_k$ such that for every $\{i, j\} \in \binom{[k]}{2}$ and almost every pair $(F_i, F_j) \in \mathcal{F}_i \times \mathcal{F}_j$, either $F_i \subset F_j$ or $F_j \subset F_i$. Moreover, since we removed only $o(m^2k^{-4})$ edges to find the k-partite subgraph of $G_{\mathcal{F}}$, there can be at most $o(m^2k^{-4})$ comparable pairs within the subfamilies $\mathcal{F}_i$ of $\mathcal{F}$.

Now we claim that for every pair of indices $\{i, j\} \in \binom{[k]}{2}$, either almost every set from $\mathcal{F}_i$ is contained in almost every set from $\mathcal{F}_j$ or vice versa. Suppose towards contradiction we have $\Omega(m^2k^{-2})$ pairs $(F_i, F_j) \in \mathcal{F}_i \times \mathcal{F}_j$ with $F_i \subset F_j$ and $\Omega(m^2k^{-2})$ pairs $(F_i, F_j) \in \mathcal{F}_i \times \mathcal{F}_j$ with $F_j \subset F_i$. Let $\mathcal{A} \subset \mathcal{F}_i$ be those sets contained in $\Omega(mk^{-1})$ sets in $\mathcal{F}_j$, and $\mathcal{B} \subset \mathcal{F}_j$ be those sets contained in $\Omega(mk^{-1})$ sets in $\mathcal{F}_i$; we must have $|\mathcal{A}|, |\mathcal{B}| = \Omega(mk^{-1})$. Consider the pairs in $\mathcal{A} \times \mathcal{B}$. Since there are $\Omega(m^2k^{-2})$ such pairs, and only $o(m^2k^{-4})$ non-edges of $G_{\mathcal{F}}$ between $\mathcal{F}_i$ and $\mathcal{F}_j$, there are $\Omega(m^2k^{-2})$ comparable pairs between $\mathcal{A}$ and $\mathcal{B}$. However, if $A \subset B$ for some $A \in \mathcal{A}$ and $B \in \mathcal{B}$, then A is contained in each of the $\Omega(mk^{-1})$ sets of $\mathcal{F}_i$ that B is contained in, and similarly if $B \subset A$. Since there are $\Omega(m^2k^{-2})$ such pairs to count over, and each comparable pair within $\mathcal{F}_i$ and $\mathcal{F}_j$ gets counted $O(mk^{-1})$ times, it follows that there are $\Omega(m^2k^{-2})$ comparable pairs within $\mathcal{F}_i$ and $\mathcal{F}_j$, contradicting our earlier observation.

Hence we may assume from now on that for every $i < j$, almost all sets from $\mathcal{F}_i$ are contained in almost all sets from $\mathcal{F}_j$. We "clean up" the families by discarding any atypical sets, of which there can be at most $o(m)$, and so we may assume there is some $\delta = o(k^{-1})$ such that for every $i \in [k]$

9

we have subfamilies $\mathcal{F}_i' \subseteq \mathcal{F}_i$ of sizes at least $(1-\delta)2^{n/k}$ such that every set $F_i \in \mathcal{F}_i'$ is contained in at least $(1-\delta)|\mathcal{F}_{i+1}'|$ sets from $\mathcal{F}_{i+1}'$ (for $i < k$) and contains at least $(1-\delta)|\mathcal{F}_{i-1}'|$ sets from $\mathcal{F}_{i-1}'$ (for $i > 1$).

We now generalise the proof of Lemma 2.1, seeking to discover a vertex partition corresponding to a tower of cubes containing $\bigcup_i \mathcal{F}_i'$. Recursively define the following partition of $[n]$: set $C_0 = R_0 = \emptyset$, and, for $i \in [k]$, let C_i consist of all those elements from $[n] \setminus \bigcup_{j<i}(C_j \cup R_j)$ that appear in more than $\delta|\mathcal{F}_i'|$ sets in $\mathcal{F}_i'$, and R_i to be the set of all those elements from $[n] \setminus \bigcup_{j<i}(C_j \cup R_j)$ that appear in at least one and at most $\delta|\mathcal{F}_i'|$ sets from $\mathcal{F}_i'$. Let $R_{k+1} = [n] \setminus \bigcup_{j \leq k}(C_j \cup R_j)$ complete the partition.

Now, for an arbitrary $i \in [k]$, let us take a closer look at the structure of the sets $F_i \in \mathcal{F}_i'$. Note that if $x \in C_i \cup R_i$, there is some set $F_i \in \mathcal{F}_i'$ such that $x \in F_i$. Since F_i is a subset of almost every set in $\mathcal{F}_{i+1}'$, it follows that x is contained in at least $(1-\delta)|\mathcal{F}_{i+1}'|$ sets in $\mathcal{F}_{i+1}'$. Moreover, if x appears in at least $\delta|\mathcal{F}_i'|$ sets in $\mathcal{F}_i'$, then it must appear in every set in $\mathcal{F}_{i+1}'$, since these sets are comparable to at least $(1-\delta)|\mathcal{F}_i'|$ sets in $\mathcal{F}_i'$.

From this, we can easily deduce that every set $F_i \in \mathcal{F}_i'$ must contain $C_{i-1} \cup \bigcup_{j \leq i-2}(C_j \cup R_j)$ as a subset. Furthermore, every $x \in R_{i-1}$ must be contained in at least $(1-\delta)|\mathcal{F}_i'|$ sets in $\mathcal{F}_i'$. By definition of R_i, each element in R_i is contained in at most $\delta|\mathcal{F}_i'|$ sets in $\mathcal{F}_i'$. Using Lemma 2.3 we have

$$|\mathcal{F}_i'| \leq 2^{|C_i|+H(\delta)(|R_{i-1}|+|R_i|)}.$$

Hence, we observe that, provided δ is sufficiently small,

$$2^{n-0.5} < (1-\delta)^k 2^n \leq \prod_{i \in [k]} |\mathcal{F}_i'| \leq 2^{\sum_{i \in [k]}(|C_i|+2H(\delta)|R_i|)} = 2^{n-(1-2H(\delta))\sum_i |R_i|-|R_{k+1}|} \leq 2^{n-\sum_{i=1}^{k+1}|R_i|/2},$$

implying that for every $i \in [k+1]$, $R_i = \emptyset$. Moreover, since $|\mathcal{F}_i'| \geq (1-\delta)2^{n/k}$, we must have $|C_i| \geq n/k$. Since $\sum_{i \in [k]} |C_i| = n$, we in fact have equality. In conclusion, the family $\mathcal{F}' = \cup_i \mathcal{F}_i'$, which has all but $o(m)$ sets of $\mathcal{F}$, is contained in the tower of cubes of dimension n/k given by the chain $\left\{\bigcup_{j=0}^i C_j\right\}_{0 \leq i \leq k}$. $\qquad \square$

We conclude this subsection by deducing the exact result given in Corollary 1.5, showing that when n is large, $k|n$ and $m = k2^{n/k} - k + 1$, every extremal m-set family $\mathcal{F}$ with $c(\mathcal{F}) = c(n,m)$ is a tower of k cubes of dimension n/k.

Proof of Corollary 1.5. Let $\ell = n/k$ be the dimensions of the cubes. From Theorem 1.4, we know that all but $\varepsilon|\mathcal{F}|$ sets of $\mathcal{F}$ are contained in a tower of cubes; without loss of generality, suppose these cubes correspond to the chain $\{[i\ell] : 0 \leq i \leq k\}$. We show that if we have sets outside this tower of cubes, we can increase the number of comparable pairs by shifting them inside.

Suppose G is a set in $\mathcal{F}$ outside the tower of cubes, and let i be such that $i\ell \leq |G| \leq (i+1)\ell$. Since G is not in the tower of cubes, we cannot have $[i\ell] \subset G \subset [(i+1)\ell]$. Consider how many sets from the subcubes between $[(i-1)\ell], [i\ell], [(i+1)\ell]$ and $[(i+2)\ell]$ G can be comparable to.

In the subcube $\{F : [(i-1)\ell] \subset F \subset [i\ell]\}$, the number of sets with $[(i-1)\ell] \subset F \subset [i\ell] \cap G$ is at most $2^{\ell-|[i\ell]\setminus G|}$. Similarly, in the subcube $\{F : [(i+1)\ell] \subset F \subset [(i+2)\ell]\}$, there are at most $2^{\ell-|G\setminus[(i+1)\ell]|}$ sets with $G \cup [(i+1)\ell] \subset F \subset [(i+2)\ell]$.

Finally we consider the subcube $\{F : [i\ell] \subset F \subset [(i+1)\ell]\}$. If $[i\ell] \subset G$, then there are at most $2^{\ell-|[(i+1)\ell]\setminus G|}$ sets F with $[i\ell] \subset F \subset G \cap [(i+1)\ell]$; otherwise there are no such sets. Similarly, if $G \subset [(i+1)\ell]$, there are at most $2^{\ell-|G\setminus[i\ell]|}$ sets F with $[i\ell] \cup G \subset F \subset [(i+1)\ell]$; otherwise there are none.

10

Suppose $[i\ell] \not\subset G$ (the case $G \not\subset [(i+1)\ell]$ is similar). G can be comparable to at most $2^\ell + 2^{\ell - |[i\ell]\setminus G|} + 2^{\ell - |G\setminus[i\ell]|}$ sets from these three subcubes. Since $|G| \geq i\ell$ and $[i\ell] \not\subset G$, it follows that $|G \setminus [i\ell]|, |[i\ell] \setminus G| \geq 1$. If either one of these set differences has size at least 2, then it follows that G is comparable to at most $7 \cdot 2^\ell/4$ sets in these three subcubes. Adding the remaining subcubes, and the small fraction of sets outside the tower of cubes, it follows that G is in at most $(k - 5/4 + \varepsilon) 2^\ell$ comparable pairs. However, any set inside the tower of cubes would be in at least $(k - 1 - \varepsilon) 2^\ell$ comparable pairs, and so we would increase the number of comparable pairs by replacing G with a set missing from the tower of cubes.

Hence we may assume that $|G \setminus [i\ell]| = |[i\ell] \setminus G| = 1$ for every set G outside the tower of cubes. However, for each i, there are at most n^2 such sets, and hence we can have at most n^3 sets outside the tower of cubes. Each such set is in at most $(k - 1)2^\ell + n^3$ comparable pairs. However, any set in the tower of cubes is comparable to $(k-1)2^\ell$ sets from the other subcubes, and at least a further $2^{\ell/2+1} = 2^{n/(2k)+1} \gg n^3$ sets from its own subcube, and hence is in $(k - 1)2^\ell + 2^{\ell/2+1}$ comparable pairs. Thus it is again beneficial to replace sets outside the tower of cubes with sets within the tower.

This shows the extremal families must be towers of cubes, completing the proof. □

3.2   The proof of Theorem 1.6

We now turn to sparser families, where $|\mathcal{F}| = m$ is subexponential (but superlinear). Recall that in this setting we have $c(n, m) = (1 - o(1))\binom{m}{2}$, so we instead count the number of incomparable pairs $i(n, m) = \binom{m}{2} - c(n, m)$. We also parameterise the size of families by setting $m = n\ell$.

The key to proving Theorem 1.6 is the following proposition, which gives the desired bound on $i(\mathcal{F})$ provided there are no sets in $\mathcal{F}$ incomparable to many other sets.

Proposition 3.5. *Given $\varepsilon > 0$, for ℓ and n sufficiently large we have $i(\mathcal{F}) \geq (1/2 - \varepsilon) n\ell^2 \log \ell$ for any family $\mathcal{F}$ of $n\ell$ sets over $[n]$ such that every set $F \in \mathcal{F}$ is incomparable to at most $4\ell \log \ell$ other sets.*

Proof. We shall obtain the desired result by finding dense subcubes of $\mathcal{F}$, and applying Theorem 1.1 to these subcubes. We first define some parameters we shall need. Set $k = \max\{81, 4\varepsilon^{-2}\}$, $s = k \log \ell$ and $r = n/s$. Partition $[n]$ into r intervals $I_1, I_2, \ldots, I_r$, each of length s, and partition $\mathcal{F} = \cup_{j=1}^r \mathcal{F}_j$, where $\mathcal{F}_j = \{F \in \mathcal{F} : |F| \in I_j\}$.

We shall remove few sets from the subfamilies $\mathcal{F}_j$ to obtain cleaner $\mathcal{F}_j' \subset \mathcal{F}_j$ with the property that $\mathcal{F}_j'$ is either empty or contained in a small subcube. Call an interval I_j *light* if $|\mathcal{F}_j| < \ell\sqrt{k} \log \ell$, and *heavy* otherwise. If I_j is light, we remove all the subsets from $\mathcal{F}_j$ and set $\mathcal{F}_j' = \emptyset$, thus losing at most $\ell\sqrt{k} \log \ell$ sets.

On the other hand, if I_j is heavy, let $F_{j,0}$ and $F_{j,1}$ be sets of minimum and maximum size in $\mathcal{F}_j$ respectively. By assumption, there are at most $8\ell \log \ell$ sets in $\mathcal{F}_j$ that are incomparable to one of $F_{j,0}$ or $F_{j,1}$. Let $\mathcal{F}_j'$ be the remaining subfamily, and observe we must have $F_{j,0} \subseteq F \subseteq F_{j,1}$ for any $F \in \mathcal{F}_j'$. Since I_j is heavy, we have $\left|\mathcal{F}_j'\right| \geq \ell\sqrt{k} \log \ell - 8\ell \log \ell \geq \ell$, since $\sqrt{k} \geq 9$. Moreover, $\mathcal{F}_j'$ is contained in the subcube spanned by $F_{j,0}$ and $F_{j,1}$, which has dimension $|F_{j,1}| - |F_{j,0}| \leq |I_j| = s$.

Since $\mathcal{F}_j'$ lies in a subcube of dimension s, we shall use Theorem 1.1 to estimate $i(\mathcal{F}_j')$. Setting $\delta = 1/(k(k+1))$, we have $2^{(1/(k+1)+\delta)s} = 2^{s/k} = \ell \leq \left|\mathcal{F}_j'\right|$. Hence, by Theorem 1.1, we have, for some $\beta = \beta_k$ and constant C,

$$i(\mathcal{F}_j') \geq \frac{1}{k}\binom{\left|\mathcal{F}_j'\right|}{2} - C\left|\mathcal{F}_j'\right|^{2 - \beta\delta^{k+1}} \geq \left(1 - \frac{\varepsilon}{2}\right)\frac{1}{k}\binom{\left|\mathcal{F}_j'\right|}{2},$$

11

provided we take $\ell \geq (4Ck\varepsilon^{-1})^{(k(k+1))^{k+1}/\beta}$.

Finally, in passing to these cleaner subfamilies $\mathcal{F}_j'$, we lost at most $\ell\sqrt{k}\log\ell$ sets from each $\mathcal{F}_j$. Hence we have $\sum_{j=1}^r \left|\mathcal{F}_j'\right| \geq |\mathcal{F}| - r\ell\sqrt{k}\log\ell = n\ell - n\ell\sqrt{k}\log\ell/s = (1 - k^{-1/2})n\ell \geq (1 - \varepsilon/2)n\ell$. With these estimates in place, we can now lower bound $i(\mathcal{F})$ by summing up $i(\mathcal{F}_j')$ over each interval. This gives

$$i(\mathcal{F}) \geq \sum_{j=1}^r i(\mathcal{F}_j') \geq \left(1 - \frac{\varepsilon}{2}\right)\frac{1}{k}\sum_{j=1}^r \binom{\left|\mathcal{F}_j'\right|}{2} \geq (1 - \varepsilon)\frac{1}{2k}\sum_{j=1}^r |\mathcal{F}_j'|^2$$

$$\geq (1 - \varepsilon)\frac{1}{2rk}\left(\sum_{j=1}^r |\mathcal{F}_j'|\right)^2 \geq (1 - \varepsilon)\left(1 - \frac{\varepsilon}{2}\right)^2\frac{n^2\ell^2}{2rk} \geq \left(\frac{1}{2} - \varepsilon\right)n\ell^2\log\ell,$$

where we use the Cauchy-Schwarz inequality in the second line. Hence we obtain the claimed bound on $i(\mathcal{F})$. $\qquad\square$

We now deduce Theorem 1.6 from the above proposition.

Proof of Theorem 1.6. Given $\varepsilon > 0$, let ℓ be twice as large as needed in Proposition 3.5. We shall show that if $\mathcal{F}$ is a family of $n\ell$ sets over $[n]$, then $i(\mathcal{F}) \geq (1/2 - \varepsilon)\,n\ell^2\log\ell$. To do so, we successively remove sets from $\mathcal{F}$ until no set is incomparable to many other sets, and then apply the proposition.

Start with $\mathcal{F}_0 = \mathcal{F}$ and $\ell_0 = \ell$. Now, given $\mathcal{F}_j$ and ℓ_j, if there is a set $F_j \in \mathcal{F}_j$ incomparable to at least $4\ell_j\log\ell_j$ other sets in $\mathcal{F}_j$, then define $\mathcal{F}_{j+1} = \mathcal{F}_j \setminus \{F_j\}$ and set $\ell_{j+1} = \ell_j - 1/n$, so that we have $|\mathcal{F}_{j+1}| = n\ell_{j+1}$. Otherwise, stop the process and set $t = j$.

Since the set F_j accounts for at least $4\ell_j\log\ell_j$ distinct incomparable pairs, we have

$$i(\mathcal{F}) \geq 4\sum_{j=0}^{t-1} \ell_j\log\ell_j + i(\mathcal{F}_t).$$

Observe that $\ell_j \geq \ell/2$ for $j \leq n\ell/2$. Hence, if $t \geq n\ell/2$, we have

$$i(\mathcal{F}) \geq 4\sum_{j=0}^{n\ell/2} \ell_j\log\ell_j \geq n\ell^2\log\frac{\ell}{2} > \frac{1}{2}n\ell^2\log\ell,$$

as required. Hence we may assume $t \leq n\ell/2$, and thus $\ell_t \geq \ell/2$. By our bound on ℓ, this implies that we may apply Proposition 3.5 to $\mathcal{F}_t$, giving

$$i(\mathcal{F}) \geq 4\sum_{j=0}^{t-1} \ell_j\log\ell_j + i(\mathcal{F}_t) \geq 4\sum_{j=0}^{t-1} \ell_j\log\ell_j + \left(\frac{1}{2} - \varepsilon\right)n\ell_t^2\log\ell_t. \tag{6}$$

Since $\ell_{j+1} = \ell_j - 1/n$ and $-\log(1 - x) \leq \log(1 + 2x) \leq 4x$ for $0 \leq x \leq 1/2$, we have

$$\left(\frac{1}{2} - \varepsilon\right)n\ell_j^2\log\ell_j - \left(\frac{1}{2} - \varepsilon\right)n\ell_{j+1}^2\log\ell_{j+1}$$

$$= \left(\frac{1}{2} - \varepsilon\right)\left(n\ell_j^2\log\ell_j - n\left(\ell_j - \frac{1}{n}\right)^2\log\ell_j - n\left(\ell_j - \frac{1}{n}\right)^2\log\left(1 - \frac{1}{n\ell_j}\right)\right)$$

$$\leq \left(\frac{1}{2} - \varepsilon\right)(2\ell_j\log\ell_j + 4\ell_j) < 4\ell_j\log\ell_j,$$

12

which we may use to telescope the sum in (6) and deduce that

$$i(\mathcal{F}) \geq 4\sum_{j=0}^{t-1} \ell_j \log \ell_j + \left(\frac{1}{2} - \varepsilon\right) n\ell_t^2 \log \ell_t \geq \left(\frac{1}{2} - \varepsilon\right) n\ell_0^2 \log \ell_0 = \left(\frac{1}{2} - \varepsilon\right) n\ell^2 \log \ell.$$

Since the family $\mathcal{F}$ was arbitrary, it follows that $i(n, n\ell) \geq (1/2 - \varepsilon)n\ell^2 \log \ell$. $\qquad\square$

4   Dense families

In this section we consider much denser families, for which $m \geq 2^{0.92n}$. Theorem 1.1 shows that such families have $o(m^2)$ comparable pairs, and so the situation is qualitatively different from the preceding sections. The extremal families are of a different nature as well. We shall prove Theorem 1.7, which shows that the extremal families contain sets as far from the middle levels as possible. To this end, it is more convenient to work with the complementary set family $\mathcal{G} = 2^{[n]} \setminus \mathcal{F}$. Note that to maximise $c(\mathcal{F})$, we must minimise the number of comparable pairs containing at least one set from $\mathcal{G}$. Since sets in the middle levels are contained in the fewest comparable pairs, it is intuitive that $\mathcal{G}$ should be taken from the middle levels, and we will use shifting arguments to make this intuition precise.

Proof of Theorem 1.7. Let $\mathcal{F} \subset 2^{[n]}$ be a family of m sets maximising $c(\mathcal{F})$. We consider the complementary family $\mathcal{G} = 2^{[n]} \setminus \mathcal{F}$, which minimises the number of comparable pairs containing a set from $\mathcal{G}$. We need to show that $\mathcal{G}$ contains all sets F with $k + 1 \leq |F| \leq n - k - 1$ and does not contain any set F with $|F| \leq k - 1$ or $|F| \geq n - k + 1$.

To this end, we partition $2^{[n]}$ into three families. Let $\mathcal{A}_0 = 2^{[n]} \setminus \mathcal{H}_{k-1}$ be the family of all sets of sizes between k and $n - k$, $\mathcal{A}_1 = \mathcal{H}_{k-1} \setminus \mathcal{H}_{k-2}$ the family of all sets of size $k - 1$ or $n - k + 1$, and $\mathcal{A}_2 = \mathcal{H}_{k-2}$ the family of all sets of size at most $k - 2$ or at least $n - k + 2$. For $0 \leq i \leq 2$, define $\mathcal{G}_i = \mathcal{G} \cap \mathcal{A}_i$.

We start by showing that $\mathcal{G} \subset \mathcal{A}_0$, or, equivalently, $\mathcal{G}_1 \cup \mathcal{G}_2 = \emptyset$. Suppose we had $m' = |\mathcal{G}_1| + |\mathcal{G}_2| > 0$. By shifting the sets in $\mathcal{G}_1 \cup \mathcal{G}_2$ into $\mathcal{A}_0$, we shall decrease the number of comparable pairs involving sets in $\mathcal{G}$, contradicting the optimality of $\mathcal{G}$ (and hence of $\mathcal{F}$).

First note that since $|\mathcal{F}| \geq M_{k-1} = |\mathcal{H}_{k-1}|$, $|\mathcal{G}| \leq |\mathcal{A}_0|$, and hence there must be at least m' available sets in $\mathcal{A}_0 \setminus \mathcal{G}$ to shift to. Each such set is comparable to at most $2^{n-k} + 2^k - 2 = \left(1 + 2^{-(n-2k)}\right) 2^{n-k} - 2$ other sets, and hence the number of new comparable pairs the shift could introduce is at most $\left[\left(1 + 2^{-(n-2k)}\right) 2^{n-k} - 2\right] m'$.

On the other hand, we bound from below the number of comparable pairs we would lose through this shift. We would lose all the pairs containing sets in $\mathcal{G}_1 \cup \mathcal{G}_2$ except those that also contain a set from $\mathcal{G}_0$. To avoid counting the latter pairs, for every set $G \in \mathcal{G}_1 \cup \mathcal{G}_2$ we shall count only the sets in $\mathcal{A}_2$ comparable to G. In this way we do not count any pairs containing a set from $\mathcal{G}_0$. Furthermore, a pair containing a set from $\mathcal{G}_1$ is counted only once. However, comparable pairs with both sets from $\mathcal{G}_2$ are counted twice, which we shall have to account for.

The following lemma, proven later, will be required for our calculations.

Lemma 4.1. *If $\lambda \geq 0$ and $\mathcal{A} = \{A \subset [r] : |A| \geq r/2 + \lambda\sqrt{r}\}$, then $|\mathcal{A}| \leq e^{-2\lambda^2} 2^r$.*

We first consider sets $G \in \mathcal{G}_1$. By symmetry, we may assume $|G| = n - k + 1$. G is contained in $2^{k-1} - 1$ larger sets in $\mathcal{A}_2$. The number of subsets of G in $\mathcal{A}_2$ is given by $\sum_{j=0}^{k-2} \binom{n-k+1}{j} = 2^{n-k+1} - \sum_{j=k-1}^{n-k+1} \binom{n-k+1}{j}$. By our bound on k from the statement of Theorem 1.7, it follows that

13

$2(k-1) > (n-k+1) + 2\sqrt{2n \ln 2}$, and so we may apply Lemma 4.1 with $r = n - k + 1$ and $\lambda = \sqrt{2 \ln 2}$ to deduce

$$2^{n-k+1} - \sum_{j=k-1}^{n-k+1} \binom{n-k+1}{j} \geq \left(1 - e^{-4\ln 2}\right) 2^{n-k+1} = \frac{15}{16} \cdot 2^{n-k+1}.$$

Hence, when shifting, we lose at least $15 \cdot 2^{n-k+1}/16 + 2^{k-1} - 1 = \left(15/8 + 2^{-(n-2k+1)}\right)2^{n-k} - 1$ comparable pairs for each $G \in \mathcal{G}_1$.

We now perform a similar analysis for $G \in \mathcal{G}_2$. Again, by symmetry, we may assume $|G| \geq n - k + 2$. Such a set is contained in $2^{n-|G|} - 1$ larger sets in $\mathcal{A}_2$. The number of subsets of G in $\mathcal{A}_2$ is at least $\sum_{j=0}^{k-2} \binom{|G|}{j}$. The following lemma, whose proof we defer for the moment, shows that this quantity is minimised when $|G|$ is as small as possible.

Lemma 4.2. *Given integers n and s satisfying $n/3 < s \leq n/2$, the quantity*

$$2^{n-t} + \sum_{j \leq s} \binom{t}{j}$$

is minimised over $n - s \leq t \leq n$ when $t = n - s$.

Hence, for the worst-case scenario, we may assume $|G| = n - k + 2$. Using Lemma 4.1 with $r = n - k + 2$ and $\lambda = \sqrt{2\ln 2}$, we may similarly conclude that each set $G \in \mathcal{G}_2$ contributes at least $15 \cdot 2^{n-k+2}/16 + 2^{k-2} - 1 = (15/4 + 2^{-(n-2k+2)})2^{n-k} - 1$ comparable pairs.

Recalling that comparable pairs between sets in $\mathcal{G}_2$ are double-counted, it follows that by shifting the sets in $\mathcal{G}_1 \cup \mathcal{G}_2$ to $\mathcal{A}_0$, we lose at least

$$\left[\left(\frac{15}{8} + 2^{-(n-2k+1)}\right)2^{n-k} - 1\right]|\mathcal{G}_1| + \frac{1}{2}\left[\left(\frac{15}{4} + 2^{-(n-2k+2)}\right)2^{n-k} - 1\right]|\mathcal{G}_2|$$

$$> \left[\left(\frac{15}{8} + 2^{-(n-2k+3)}\right)2^{n-k} - 1\right]m'$$

comparable pairs, while we gain at most $\left[\left(1 + 2^{-(n-2k)}\right)2^{n-k} - 2\right]m'$ new pairs. Since $(15/8 + 2^{-(n-2k+3)}) - (1 + 2^{-(n-2k)}) = 7(1 - 2^{-(n-2k)})/8 \geq 0$, it follows that shifting the sets from $\mathcal{G}_1 \cup \mathcal{G}_2$ decreases the number of comparable pairs involving sets in $\mathcal{G}$. This gives the desired contradiction, and hence we must have $\mathcal{G} \subset \mathcal{A}_0$.

It remains to show that $\mathcal{G}$ contains all sets of sizes between $k+1$ and $n-k-1$. Note that this is trivial unless $n \geq 2k+2$. Suppose for contradiction there is some set $G_0 \notin \mathcal{G}$ with $k+1 \leq |G_0| \leq n-k-1$. Given the size of $\mathcal{G}$, there must be some $G_1 \in \mathcal{G}$ with $|G_1| \in \{k, n-k\}$. We use the same shifting arguments as before to deduce that we should replace G_1 with G_0.

Indeed, adding G_0 to $\mathcal{G}$ can introduce at most $2^{n-k-1} + 2^{k+1} - 2 = \left(1 + 2^{-(n-2k-2)}\right)2^{n-k-1} - 2$ new comparable pairs. On the other hand, since $\mathcal{G} \subset \mathcal{A}_0$, by removing G_1 we would lose all comparable pairs between G_1 and $\mathcal{A}_1 \cup \mathcal{A}_2$. As above, we apply Lemma 4.1, this time with $r = n-k$ and $\lambda = \sqrt{2\ln 2}$, to find this gives at least $\left(15/8 + 2^{-(n-2k-1)}\right)2^{n-k-1} - 1$ comparable pairs. Since $n \geq 2k+2$, $15/8 + 2^{-(n-2k-1)} > 1 + 2^{-(n-2k-2)}$, and so switching G_1 for G_0 decreases the number of comparable pairs containing a set from $\mathcal{G}$, contradicting the optimality of $\mathcal{G}$.

By taking complements, it follows that any optimal family $\mathcal{F}$ must satisfy $\mathcal{H}_{k-1} \subset \mathcal{F} \subset \mathcal{H}_k$, as claimed. $\qquad\square$

We now prove the two lemmas used in the proof above.

Proof of Lemma 4.1. For every $i \in [r]$, we have $|\mathcal{A}(i)| \geq (1/2 + \lambda/\sqrt{r})\,|\mathcal{A}|$. Applying Lemma 2.3 gives $|\mathcal{A}| \leq 2^{rH(1/2+\lambda r^{-1/2})}$. The second-order Taylor series expansion of $H(x) = -x \log x - (1-x) \log(1-x)$ gives $H(1/2 + q) = H(1/2) + qH'(1/2) + q^2 H''(z)/2$ for some z between $1/2$ and $1/2 + q$. Since $H(1/2) = 1$, $H'(1/2) = 0$, and $H''(z) = -(z^{-1} + (1-z)^{-1})/\ln 2 \leq -4/\ln 2$, we have the bound $H(1/2 + q) \leq 1 - 2q^2/\ln 2$, and so

$$|\mathcal{A}| \leq 2^{rH(\frac{1}{2}+\lambda r^{-1/2})} \leq 2^{r(1-2\lambda^2/(r \ln 2))} = e^{-2\lambda^2} 2^r.$$

$\square$

Proof of Lemma 4.2. We show that the quantity $x_t = 2^{n-t} + \sum_{j \leq s} \binom{t}{j}$ is increasing in t. Recalling that $t \geq n - s$, we have

$$
\begin{aligned}
x_{t+1} - x_t &= \left(2^{n-(t+1)} + \sum_{j \leq s} \binom{t+1}{j} \right) - \left(2^{n-t} + \sum_{j \leq s} \binom{t}{j} \right) \\
&= 2^{n-t-1} - 2^{n-t} + \sum_{j \leq s} \left(\binom{t+1}{j} - \binom{t}{j} \right) = \sum_{j \leq s} \binom{t}{j-1} - 2^{n-t-1} \\
&= \sum_{j \leq s-1} \binom{t}{j} - 2^{n-t-1} \geq \sum_{j \leq s-1} \binom{n-s}{j} - 2^{s-1} = 2^{n-s} - \sum_{j=s}^{n-s} \binom{n-s}{j} - 2^{s-1}.
\end{aligned}
$$

Since $s > n/3$, we have $s > (n-s)/2$, and so the sum of the binomial coefficients is at most 2^{n-s-1}. Thus we have $x_{t+1} - x_t \geq 2^{n-s-1} - 2^{s-1} \geq 0$, since $s \leq n/2$. Hence x_t is minimised when $t = n - s$. $\square$

This completes the proof of Theorem 1.7, which gives the large-scale structure of dense extremal families. If $M_{k-1} \leq m \leq M_k$, and $\mathcal{F}$ is a family of size m with the maximum number of comparable pairs, then we know $\mathcal{H}_{k-1} \subset \mathcal{F} \subset \mathcal{H}_k$. This is enough to determine $c(n, m)$ asymptotically, but if we wish to find the exact value of $c(n, m)$, we must determine which sets of sizes k or $n - k$ should be contained in $\mathcal{F}$.

As the following corollary shows, this is easy for some particular ranges of m. To define the optimal construction, we shall assume for simplicity that n is even. Fix an arbitrary subset X of $n/2$ elements from $[n]$. Given $m = M_{k-1} + m'$, with $0 \leq m' \leq 2\binom{n/2}{k}$, let $\mathcal{F}_m^* = \mathcal{H}_{k-1} \cup \mathcal{A} \cup \mathcal{B}$, where $\mathcal{A}$ is a set of $\lfloor m'/2 \rfloor$ k-subsets of X, while $\mathcal{B}$ is a set of $\lceil m'/2 \rceil$ $(n-k)$-sets containing X.

Corollary 4.3. *Suppose $n/3 + \sqrt{2n \ln 2} \leq k < n/2$ and $m = M_{k-1} + m'$, where $0 \leq m' \leq 2\binom{n/2}{k}$. Then $c(n, m) = c(\mathcal{F}_m^*)$.*

Proof. By Theorem 1.7, we know that we can partition $\mathcal{F} = \mathcal{H}_{k-1} \cup \mathcal{A} \cup \mathcal{B}$, where $\mathcal{A}$ is a family of k-sets and $\mathcal{B}$ is a family of $(n-k)$-sets. Let $a = |\mathcal{A}|$ and $b = |\mathcal{B}|$; we must have $a + b = m'$.

By symmetry, each set in $\mathcal{A} \cup \mathcal{B}$ has the same number of comparable pairs with sets in $\mathcal{H}_{k-1}$, and so $c(\mathcal{F})$ is maximised precisely when $c(\mathcal{A}, \mathcal{B})$ is maximised. We trivially have $c(\mathcal{A}, \mathcal{B}) \leq ab \leq \lfloor (m')^2/4 \rfloor$. On the other hand, in our family $\mathcal{F}_m^*$ every $A \in \mathcal{A}$ and $B \in \mathcal{B}$ satisfy $A \subset X \subset B$, and so we indeed have $c(\mathcal{A}, \mathcal{B}) = ab = \lfloor (m')^2/4 \rfloor$. Thus $\mathcal{F}_m^*$ maximises the number of comparable pairs. $\square$

5 Concluding remarks

In this paper we studied the maximum number of comparable pairs present in families of a given size, continuing the work of Alon, Daykin, Frankl and others. One feature that sets this problem apart from a number of others in extremal set theory is the absence of a nested sequence of extremal families. We have shown that when $m = k2^{n/k} - k + 1$, a tower of k cubes is optimal, while when $m \geq 2^{0.92n}$, the extremal families avoid the middle layers, and hence there cannot be a sequence of families $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \ldots \subset \mathcal{F}_{2^n}$ such that $|\mathcal{F}_m| = m$ and $c(\mathcal{F}_m) = c(n, m)$. This precludes the global use of many standard techniques in the field, thus partially explaining the different arguments needed for the sparse and dense regimes. It also leaves a number of open problems, some of which we discuss below.

The Alon–Frankl conjecture

In Section 2, we settled a conjecture of Alon and Frankl, showing $m = n^{\omega(1)}2^{n/2}$ implies $c(n, m) = o(m^2)$. More precisely, we showed that if $\mathcal{A}$ and $\mathcal{B}$ are two families over $[n]$ with $|\mathcal{A}| |\mathcal{B}| = n^d 2^n$, then $c(\mathcal{A}, \mathcal{B}) \leq 2^{-d/300} |\mathcal{A}| |\mathcal{B}|$. On the other hand, the construction of Alon and Frankl shows that we can have $|\mathcal{A}| |\mathcal{B}| = \Omega(n^d 2^n)$ and $c(\mathcal{A}, \mathcal{B}) \geq 2^{-d} |\mathcal{A}| |\mathcal{B}|$. It would be very interesting to close this gap, and determine the true constant in the exponent. An exhaustive search for $n \leq 7$ suggests that the answer lies much closer to 1, and that the construction may be near-optimal.

Sparse families

We have shown that towers of finitely many cubes are extremal families, and that they are asymptotically extremal when the number of cubes tends to infinity (sublinearly in n). However, these towers of cubes can only exist for certain family sizes, and it remains to understand what the extremal families are in between. For instance, how do the extremal families transition from a tower of two cubes to a tower of three cubes when m ranges from $2 \cdot 2^{n/2} - 1$ to $3 \cdot 2^{n/3} - 2$?

Dense families

Finally, in the dense setting we determined the approximate structure of the extremal families. Thus the problem of determining $c(n, m)$ in this range reduces to maximising $c(\mathcal{A}, \mathcal{B})$ over $\mathcal{A} \subset \binom{[n]}{k}$ and $\mathcal{B} \subset \binom{[n]}{n-k}$ for fixed $m' = |\mathcal{A}| + |\mathcal{B}|$.

We showed in Corollary 4.3 that one can achieve the trivial upper bound $c(\mathcal{A}, \mathcal{B}) = \lfloor (m')^2/4 \rfloor$ when $m' \leq 2\binom{n/2}{k}$. This is in fact the largest m' for which this upper bound can be attained, as it is only possible when $|\mathcal{A}|, |\mathcal{B}| \geq \lfloor m'/2 \rfloor$ and $c(\mathcal{A}, \mathcal{B}) = |\mathcal{A}| |\mathcal{B}|$. The latter implies there is some $X \subset [n]$ with $A \subset X \subset B$ for all $A \in \mathcal{A}$ and $B \in \mathcal{B}$, which requires one of the two families to have size at most $\binom{n/2}{k}$.

One might slightly generalise the above two-level problem to obtain the following question.

Question 5.1. *Given $0 \leq k_1 \leq k_2 \leq n$, $0 \leq a \leq \binom{n}{k_1}$ and $0 \leq b \leq \binom{n}{k_2}$, which families $\mathcal{A} \subset \binom{[n]}{k_1}$ and $\mathcal{B} \subset \binom{[n]}{k_2}$ maximise $c(\mathcal{A}, \mathcal{B})$?*

While this question may seem a mild generalisation of our original problem, it has some subtle complexities. In particular, it contains as a special case the famous Kruskal–Katona theorem [12, 15]. Let $k_1 = k_2 - 1$. Note that each set in $\binom{[n]}{k_2}$ contains k_2 sets in $\binom{[n]}{k_1}$. The minimum size of a lower shadow of b sets of size k_2 is then the minimum a such that the solution to the above problem is bk_2. One can similarly model the upper shadows.

By Kruskal–Katona, we know the lower shadows are minimised when $\mathcal{A}$ and $\mathcal{B}$ are initial segments of the colexicographic ordering, while the upper shadows are minimised by initial segments of the lexicographic ordering. On the other hand, in Corollary 4.3 the extremal families were mixed: initial segments of the lexicographic ordering for $\mathcal{A}$ and colexicographic ordering for $\mathcal{B}$. This shows that, once again, there will not be a nested sequence of solutions, and the nature of the extremal families depends on the range of parameters.

The phase transition

As we have mentioned before, the extremal families are very different in nature for small and large m. Our results give information about $c(n, m)$ when either $m \leq n^d 2^{n/2}$ or $m \geq 2^{0.92n}$. It would be of great interest to determine what happens in between, and to examine how the extremal families transition from tower-of-cubes-type of constructions to those consisting of sets of size far from $n/2$.

References

[1] N. Alon and P. Frankl, *The maximum number of disjoint pairs in a family of subsets*, Graph. Combinator. **1** (1985), 13–21.

[2] N. Alon and J. Spencer, *The Probabilistic Method*, Third Edition, John Wiley Inc., New York (2008).

[3] D. Conlon and J. Fox, *Graph removal lemmas*, Surveys in Combinatorics, Cambridge University Press (2013), 1–50.

[4] S. Das, W. Gan and B. Sudakov, *Sperner's theorem and a problem of Erdős–Katona–Kleitman*, Combin. Probab. Comput., to appear.

[5] D.E. Daykin and P. Frankl, *On Kruskal's cascades and counting containments in a set of subsets*, Mathematika **30** (1983), 133–141.

[6] A.P. Dove, J.R. Griggs, R.J. Kang and J.-S. Sereni, *Supersaturation in the Boolean lattice*, Integers **14A** (2014), 1–7.

[7] D. Ellis and B. Sudakov, *Generating all subsets of a finite set with disjoint unions*, J. Comb. Theory A **118** (2011), 2319–2345.

[8] P. Erdős, *On a lemma of Littlewood and Offord*, Bull. Am. Math. Soc. **51** (1945), 898–902.

[9] P. Erdős, *On extremal problems of graphs and generalized graphs*, Israel J. Math. **2.3** (1964), 183–190.

[10] P. Erdős, *Problem sessions*, In: Ordered Sets (Proc. NATO Adv. Study), edited by I. Rival, Dordrecht: Reidel (1981), 860–861.

[11] R.K. Guy, *A miscellany of Erdős problems*, Am. Math. Mon. **90** (1983), 118–120.

[12] G.O.H. Katona, *A theorem of finite sets*, in: Theory of Graphs, ed: P. Erdős, G. Katona, Akademia Kiadó, Budapest, 1966, 187–207.

[13] P. Keevash and B. Sudakov, *On a hypergraph Turán problem of Frankl*, Combinatorica **25.6** (2005), 673–706.

[14] D. Kleitman, *A conjecture of Erdős–Katona on commensurable pairs among subsets of an n-set*, Theory of Graphs, Proc. Colloq., Tihany (1966), 215–218.

[15] J.B. Kruskal, *The number of simplices in a complex*, in: Mathematical Optimization Techniques, ed: R. Bellman, University of California Press, 1963, 251–278.

[16] I.Z. Ruzsa and E. Szemerédi, *Triple systems with no six points carrying three triangles*, in Combinatorics (Keszthely, 1976), Coll. Math. Soc. J. Bolyai **18**, Volume II, 939–945.

[17] M. Simonovits, *A method for solving extremal problems in graph theory, stability problems;* in Theory of Graphs (Proc. Colloq. Tihany, 1966), Academic Press, New York, and Akad. Kiadó, Budapest (1968), 279–319.

[18] E. Sperner, *Ein Satz über Untermengen einer endlichen Menge*, Mathematische Zeitschrift **27** (1928), 544–548.