Thresholds versus fractional expectation-thresholds

Keith Frankston, Jeff Kahn, Bhargav Narayanan, and Jinyoung Park

Abstract. Proving a conjecture of Talagrand from 2010, a fractional version of the 'expectation-threshold conjecture' of the second author and Kalai, we show that for any increasing family $\mathcal{F}$ on a finite set X, we have $p_c(\mathcal{F}) = O(q_f(\mathcal{F}) \log \ell(\mathcal{F}))$, where $p_c(\mathcal{F})$ and $q_f(\mathcal{F})$ are the threshold and 'fractional expectation-threshold' of $\mathcal{F}$, and $\ell(\mathcal{F})$ is the maximum size of a minimal member of $\mathcal{F}$. This easily implies several heretofore difficult results and conjectures in probabilistic combinatorics: thresholds for perfect hypergraph matchings (Johansson–Kahn–Vu), bounded-degree spanning trees (Montgomery), and bounded-degree spanning graphs (new), amongst others. We also give optimal bounds on the extrema of 'hypergraph-indexed' stochastic processes; these resolve, and vastly extend, the random multi-dimensional assignment problem (earlier considered by Martin–Mézard–Rivoire and Frieze–Sorkin). Our approach to both results builds on a recent breakthrough of Alweiss–Lovett–Wu–Zhang on the Erdős–Rado 'sunflower conjecture'.

1. Introduction

Our most important contribution here is the proof of a conjecture of Talagrand [30] that is a fractional version of the 'expectation-threshold' conjecture of the second author and Kalai [17]. With definitions following shortly, for an increasing family $\mathcal{F}$ on a finite set X, we write $p_c(\mathcal{F})$, $q_f(\mathcal{F})$ and $\ell(\mathcal{F})$ for the threshold, fractional expectation-threshold, and size of a largest minimal element of $\mathcal{F}$. In this language, our main result is the following.

Theorem 1.1. *There is a universal $K > 0$ such that for every finite X and increasing $\mathcal{F} \subset 2^X$,*

$$p_c(\mathcal{F}) \leq K q_f(\mathcal{F}) \log \ell(\mathcal{F}).$$

As we shall see, $q_f(\mathcal{F})$ is a more or less trivial lower bound on $p_c(\mathcal{F})$, and Theorem 1.1 says this bound is never far from the truth; furthermore, apart from the constant K, this upper bound is tight in many of the most interesting cases.

Thresholds have been a — maybe *the* — central concern of the study of random discrete structures (random graphs and hypergraphs, for example) since its initiation by Erdős and Rényi [7], with much work around locating thresholds of specific properties

Date: 11 October, 2019.

2010 *Mathematics Subject Classification.* Primary 05C80; Secondary 60C05, 82B26, 06E30.

of interest (see [3, 14]), though it was not observed until [4] that *every* increasing family admits a threshold (in the Erdős–Rényi sense). See also [11] for developments, since [10], on the very interesting question of *sharpness* of thresholds, though it is perhaps worth mentioning here that establishing sharpness typically does not require locating the threshold in question.

Our second main result is Theorem 1.7 below, which was motivated by work of Frieze and Sorkin [12] on the random multi-dimensional assignment problem; the statement of this result is postponed until we have filled in some background, to which we now turn.

Thresholds. For a given finite set X and $p \in [0,1]$, μ_p is the product measure on the power set 2^X of X given by

$$\mu_p(S) = p^{|S|}(1-p)^{|X\setminus S|} \quad \forall S \subset X.$$

A family $\mathcal{F} \subset 2^X$ is *increasing* if it is closed under taking supersets, and if this is true (and $\mathcal{F} \neq 2^X, \emptyset$), then $\mu_p(\mathcal{F}) = \sum_{S\in\mathcal{F}} \mu_p(S)$ is strictly increasing in p, and the *threshold* $p_c(\mathcal{F})$ of $\mathcal{F}$ is the unique p for which $\mu_p(\mathcal{F}) = 1/2$. This is finer than the original Erdős–Rényi notion, according to which $p^* = p^*(n)$ is *a* threshold for the sequence $\mathcal{F} = \mathcal{F}_n$ if $\mu_p(\mathcal{F}) \to 0$ if $p \ll p^*$ and $\mu_p(\mathcal{F}) \to 1$ if $p \gg p^*$; that $p_c(\mathcal{F})$ *is* always an Erdős–Rényi threshold for $\mathcal{F}$ follows from [4].

Following Talagrand [27, 28, 30], we say an increasing $\mathcal{F} \subset 2^X$ is *p-small* if there is a $\mathcal{G} \subset 2^X$ such that $\mathcal{F} \subset \langle\mathcal{G}\rangle$, where $\langle\mathcal{G}\rangle$ is the increasing family generated by $\mathcal{G}$, and

$$\textstyle\sum_{S\in\mathcal{G}} p^{|S|} \leq 1/2. \tag{1}$$

Then $q(\mathcal{F}) = \max\{p : \mathcal{F} \text{ is } p\text{-small}\}$, which we call the *expectation-threshold* of $\mathcal{F}$ (note that this term is used slightly differently in [17]), is a trivial lower bound on $p_c(\mathcal{F})$, since for $\mathcal{G}$ as above and T drawn from μ_p,

$$\mu_p(\mathcal{F}) \leq \mu_p(\langle\mathcal{G}\rangle) \leq \textstyle\sum_{S\in\mathcal{G}} \mu_p(T \supset S) = \sum_{S\in\mathcal{G}} p^{|S|}. \tag{2}$$

The following statement, the main conjecture of [17], says that for *any* increasing $\mathcal{F}$, this trivial lower bound on $p_c(\mathcal{F})$ is close to the truth.

Conjecture 1.2. *There is a universal $K > 0$ such that for every finite X and increasing $\mathcal{F} \subset 2^X$,*

$$p_c(\mathcal{F}) \leq K q(\mathcal{F}) \log |X|.$$

We should emphasise that this conjecture is very strong; indeed, quoting [17], "It would probably be more sensible to conjecture that it is *not* true." For example, it easily implies — and was largely motivated by — Erdős–Rényi thresholds for the appearance of a perfect matching in a random r-uniform hypergraph, and the appearance of a given bounded-degree spanning tree in a random graph. These have since been resolved: the first — *Shamir's problem*, circa 1980 — in [15], and the second — a mid-90's suggestion

of the second author — in [24]. Both arguments are difficult and specific to the problems they address; they are utterly unrelated either to each other or to what we do here.

Talagrand [27, 30] suggests relaxing the notion of p-small by replacing the set system $\mathcal{G}$ above by what we may think of as a *fractional* set system g. We say an increasing $\mathcal{F} \subset 2^X$ is *weakly p-small* if there is a $g : 2^X \to \mathbb{R}^+$ such that

$$\sum_{T \subset S} g(T) \geq 1 \ \ \forall S \in \mathcal{F} \ \text{ and } \ \sum_{T \subset X} g(T) p^{|T|} \leq 1/2.$$

Then $q_f(\mathcal{F}) = \max\{p : \mathcal{F} \text{ is weakly } p\text{-small}\}$, the *fractional expectation-threshold* of $\mathcal{F}$, satisfies

$$q(\mathcal{F}) \leq q_f(\mathcal{F}) \leq p_c(\mathcal{F}); \tag{3}$$

here, the first inequality is trivial and the second is similar to (2). Talagrand [30] proposes a 'linear programming relaxation' of Conjecture 1.2, and then a strengthening thereof. The first of these, the following, replaces q by q_f in Conjecture 1.2; the second, which suggests the replacement of $|X|$ by the smaller $\ell(\mathcal{F})$, is our Theorem 1.1.

Conjecture 1.3. *There is a universal $K > 0$ such that for every finite X and increasing $\mathcal{F} \subset 2^X$,*

$$p_c(\mathcal{F}) \leq K q_f(\mathcal{F}) \log |X|.$$

Talagrand further suggests the following 'very nice problem of combinatorics', which implies *equivalence of* Conjectures 1.2 and 1.3, as well as of Theorem 1.1 and the corresponding strengthening of Conjecture 1.2.

Conjecture 1.4. *There is a universal $K > 0$ such that, for any increasing $\mathcal{F}$ on a finite set X, we have $q(\mathcal{F}) \geq q_f(\mathcal{F})/K$.*

Note the interest here is in Conjecture 1.4 for its own sake, and as the most likely route to Conjecture 1.2. The equivalence predicted by this conjecture is not necessary for all the applications of Conjecture 1.2 that we are aware of; they follow just as easily from Theorem 1.1.

Spread hypergraphs and spread measures. In this paper, a *hypergraph* $\mathcal{H}$ on a finite set X (the *vertices* of $\mathcal{H}$) is a collection of subsets of X (the *edges* of $\mathcal{H}$), with *reptitions allowed*. For $S \subset X$, we use $\langle S \rangle$ for $\{T \subset X : T \supset S\}$, and for a hypergraph $\mathcal{H}$ on X, we write $\langle \mathcal{H} \rangle$ for $\cup_{S \in \mathcal{H}} \langle S \rangle$, the increasing family generated by $\mathcal{H}$. We say $\mathcal{H}$ is *ℓ-bounded* (respectively, *ℓ-uniform* or an *ℓ-graph*) if each of its members has size at most (respectively, exactly) ℓ, and *κ-spread* if

$$|\mathcal{H} \cap \langle S \rangle| \leq \kappa^{-|S|} |\mathcal{H}| \ \ \forall S \subset X; \tag{4}$$

note that edges are counted with multiplicities on both sides of (4). For example, a reasonably 'generic' ℓ-graph $\mathcal{H}$ on n vertices might — and in some of the more

interesting cases *will* — have spread like n/ℓ, since a set of s vertices would naturally lie in about an $(\ell/n)^s$-fraction of the edges of $\mathcal{H}$.

A major advantage of the fractional versions (i.e., Conjecture 1.3 and Theorem 1.1) over Conjecture 1.2 — and the source of the present relevance of [2] — is that they admit, via linear programming duality, reformulations in which the specification of $q_f(\mathcal{F})$ gives us a usable starting point. Following [30], we say that a probability measure ν on 2^X is *q-spread* if

$$\nu(\langle S\rangle) \leq q^{|S|} \quad \forall S \subset X.$$

Thus a hypergraph $\mathcal{H}$ is κ-spread if and only if the uniform measure on $\mathcal{H}$ is q-spread with $q = \kappa^{-1}$. As observed by Talagrand [30], the following is an easy consequence of duality.

Proposition 1.5. *For an increasing $\mathcal{F}$ on a finite X, if $q_f(\mathcal{F}) \leq q$, then there is a $(2q)$-spread probability measure on 2^X supported on $\mathcal{F}$.* □

This allows us to reduce Theorem 1.1 to the following alternate (actually, equivalent) statement. In this paper, *with high probability* means with probability tending to 1 as $\ell \to \infty$.

Theorem 1.6. *There is a universal $K > 0$ such that for any ℓ-bounded, κ-spread hypergraph $\mathcal{H}$ on a finite X, a uniformly random $((K\kappa^{-1}\log\ell)|X|)$-element subset of X belongs to $\langle\mathcal{H}\rangle$ with high probability.*

Minima of hypergraph-indexed stochastic processes. Our second main result provides upper bounds on the minima of a large class of hypergraph-based stochastic processes, somewhat in the spirit of [29] (see also [28, 31]), saying that in 'smoother' settings, the logarithmic corrections of Conjectures 1.2 and 1.3 and Theorem 1.1 are not needed. For a hypergraph $\mathcal{H}$ on a finite set X, let $(\xi_x)_{x\in X}$ be independent random variables, each uniform from $[0, 1]$, and set

$$\xi_{\mathcal{H}} = \min_{S\in\mathcal{H}} \sum_{x\in S} \xi_x \quad \text{and} \quad Z_{\mathcal{H}} = \mathbb{E}[\xi_{\mathcal{H}}]. \tag{5}$$

In this language, our second main theorem is as follows.

Theorem 1.7. *There is a universal $K > 0$ such that for any ℓ-bounded, κ-spread hypergraph $\mathcal{H}$, we have $Z_{\mathcal{H}} \leq K\ell/\kappa$, and $\xi_{\mathcal{H}} \leq K\ell/\kappa$ with high probability.*

These bounds are again tight up to the value of the constant K. Furthermore, the distribution of the ξ_x's is not very important; for example, it is not hard to adapt the present argument to show that the same statement holds for Exp(1) random variables, as in the next example.

Theorem 1.7 was motivated by the work of Martin, Mézard and Rivoire [23] and Frieze and Sorkin [12] on the *random d-dimensional assignment problem.* This asks, for fixed d and large n, for the estimation of

$$Z_d^A(n) = \mathbb{E}\left[\min_{S\in\mathcal{A}}\sum_{x\in S}\xi_x\right], \tag{6}$$

where $X = [n]^d$, the ξ_x's are independent Exp(1) weights for $x \in X$, and $\mathcal{A}$ is the family of 'axial assignments', meaning $S \in \mathcal{A}$ meets each *axis-parallel hyperplane* ($\{x \in X : x_i = a\}$ for some $i \in [d]$ and $a \in [n]$) exactly once. For $d = 2$, this is classical; see [12] for its rather glorious history. For $d = 3$, the deterministic version was one of Karp's [18] original NP-complete problems. Progress on the random version in higher dimensions has been limited; see [12] for a guide to the literature.

Frieze and Sorkin show for suitable $c_1, c_2 > 0$ that

$$c_1 n^{-(d-2)} < Z_d^A(n) < c_2 n^{-(d-2)}\log n. \tag{7}$$

Here, the lower bound is easy, and the upper bound follows from the bounds of [15] on Shamir's problem. In present language, $Z_d^A(n)$ is essentially (that is, apart from the difference in the distributions of the ξ_x's) $Z_{\mathcal{H}}$, with $\mathcal{H}$ the family of perfect matchings of the complete, balanced d-uniform d-partite hypergraph on dn vertices. This is easily seen to be κ-spread with $\kappa = (n/e)^{d-1}$ (apart from the nearly irrelevant d-particity, it is the same $\mathcal{H}$ in Shamir's problem), so the correct bound, as heuristically predicted by the 'cavity method' in [23], is an instance of Theorem 1.7.

Corollary 1.8. *For all fixed $d \in \mathbb{N}$, we have*

$$Z_d^A(n) = \Theta(n^{-(d-2)}).$$

Frieze and Sorkin also consider another version of the problem, in which $\mathcal{A}$ in (6) consists of those S that meet each *axis-parallel line* ($\{x \in X : x_j = y_j\ \forall j \neq i\}$ for some $i \in [d]$ and $y \in X$) exactly once, and one may of course generalise from hyperplanes/lines to k-dimensional 'subspaces' for any given $k \in [d-1]$. It is easy to see what to expect here, and one may hope Theorem 1.7 will eventually apply, but we at present lack the technology to say that the relevant hypergraphs are suitably spread.

Organisation. Section 2 includes minor preliminaries and the derivation of Theorem 1.1 from Theorem 1.6. The main lemma at the heart of our argument is proved in Section 3. Our approach here strengthens that of the recent breakthrough of Alweiss, Lovett, Wu and Zhang [2] on the Erdős–Rado sunflower conjecture [6]; see also [25, 32] for other views of the argument of [2]. The proofs of Theorems 1.6 and 1.7 follow in Section 4. Finally, Section 5 outlines a few applications already alluded to here, and Section 6 discusses unresolved questions.

2. Preliminaries

As is usual, we use $[n]$ for $\{1, 2, \ldots, , n\}$, 2^X for the power set of X, $\binom{X}{r}$ for the family of r-element subsets of X, and $[S, T]$ for $\{R : S \subset R \subset T\}$. Our default universe is X, with $|X| = n$.

In what follows, we assume ℓ and n are somewhat large (and when there is an ℓ, it will be at most n), as we may do since smaller values can by handled by adjusting the constants in Theorems 1.6 and 1.7. Asymptotic notation referring to some parameter λ (usually ℓ) is used in the natural way: implied constants in $O(\cdot)$ and $\Omega(\cdot)$ are independent of λ, and we use $f = o(g)$ and $f \ll g$ synonymously. Following a standard abuse, we usually pretend large numbers are integers.

For $p \in [0, 1]$ and $m \in [n]$, X_p and X_m are (respectively) a p-random subset of X (drawn from μ_p) and a uniformly random m-element subset of X. The latter is not entirely kosher, since we will also see *sequences* X_i; however, we will never see both interpretations in close proximity, and the overlap should cause no confusion.

In a couple places it will be helpful to assume uniformity, which we will justify using the next little point.

Observation 2.1. *If $\mathcal{H}$ is ℓ-bounded and κ-spread, and we replace each $S \in \mathcal{H}$ by M new edges, each consisting of S plus $\ell - |S|$ new vertices (each used just once), then for large enough M, the resulting ℓ-graph $\mathcal{G}$ is again κ-spread.*

We close this section with the reduction promised earlier.

Derivation of Theorem 1.1 from Theorem 1.6. Let $\mathcal{F}$ be as in Theorem 1.1 with $\mathcal{G}$ its set of minimal elements, let ℓ with $\ell(\mathcal{F}) \le \ell = O(\ell(\mathcal{F}))$ be large enough that the exceptional probability in Theorem 1.6 is less than $1/4$, and let ν be the $(2q)$-spread probability measure promised by Proposition 1.5, where $q = q_f(\mathcal{F})$. We may assume ν is supported on $\mathcal{G}$ (since transferring weight from S to $T \subset S$ does not destroy the spread condition) and that ν takes values in $\mathbb{Q}$ (where we should really relax to $((2 + \varepsilon)q)$-spread, but we ignore this immaterial difference). We may then replace $\mathcal{G}$ by $\mathcal{H}$ whose edges are copies of edges of $\mathcal{G}$, and ν by the uniform measure on $\mathcal{H}$. Setting $m = ((2Kq \log \ell)n)$ and $p = 2m/n$ (with $n = |X|$ and K as in Theorem 1.6), we then see (using Theorem 1.6 with $\kappa = 1/(2q)$) that

$$\mu_p(\mathcal{F}) \ge \mathbb{P}(X_p \in \langle\mathcal{H}\rangle) \ge \mathbb{P}(|X_p| \ge m)\mathbb{P}(X_m \in \langle\mathcal{H}\rangle) \ge 3\mathbb{P}(|X_p| \ge m)/4 > 1/2,$$

implying that $p_c(\mathcal{F}) < p = 4Kq \log \ell$. Note that $\mathcal{H}$ being q-spread with $\emptyset \notin \mathcal{H}$ implies that $q \ge 1/n$, so that m is somewhat large and $\mathbb{P}(|X_p| \ge m) > 2/3$, with room to spare. $\square$

Remark 2.2. *This was done fussily to cover smaller ℓ in Theorem 1.1; if $\ell \to \infty$, then the above reduction gives $\mathbb{P}(X_p \in \langle\mathcal{H}\rangle) \to 1$.*

3. Main lemma

Let γ be a small constant ($\gamma = 0.1$ certainly suffices), and let C_0 be a constant large enough to support the estimates that follow. Let $\mathcal{H}$ be an r-bounded, κ-spread hypergraph on a set X of size n, with $r, \kappa \geq C_0^2$. Set $p = C/\kappa$ with $C_0 \leq C \leq \kappa/C_0$ so that $p \leq 1/C_0$, $r' = (1-\gamma)r$ and $N = \binom{n}{np}$. Finally, fix $\psi : \langle\mathcal{H}\rangle \to \mathcal{H}$ satisfying $\psi(Z) \subset Z$ for all $Z \in \langle\mathcal{H}\rangle$; set, for $W \subset X$ and $S \in \mathcal{H}$,

$$\chi(S,W) = \psi(S \cup W) \setminus W,$$

and say the pair (S,W) is *bad* if $|\chi(S,W)| > r'$ and *good* otherwise.

The heart of our argument is an improvement of the main lemma of [2], regarding which a little of orientation may be helpful. We will (in Theorems 1.6 and 1.7) be choosing a random subset of X in small increments and would like to say we are likely to be making good progress toward containing some $S \in \mathcal{H}$. Of course, such progress is not to be expected for a *typical* S, but this is not the goal: having chosen some portion W of our eventual set, we just need the remainder to contain *some* $S \setminus W$, and may focus on those that are more likely (meaning small). The key idea (introduced in [2] and refined here) is that a general $S \setminus W$, while not itself small, will in consequence of the spread assumption, typically *contain* some small $S' \setminus W$. In fact $\chi(S,W)$ will usually be one of these: an $S' \setminus W$ contained in $S \setminus W$ will *typically* be small, so we do not need to steer this choice. We then replace each 'good' $S \setminus W$ by $\chi(S,W)$ and iterate, a second nice feature of the spread condition being that it is not much affected by this substitution.

With this outline in place, we are now ready to state and prove our main lemma.

Lemma 3.1. *For $\mathcal{H}$ as above, and W chosen uniformly from $\binom{X}{np}$,*

$$\mathbb{E}[|\{S \in \mathcal{H} : (S,W) \text{ is bad}\}|] \leq |\mathcal{H}|C^{-r/3}.$$

Proof. It is enough to show, for $s \in (r', r]$, that

$$\mathbb{E}\left[|\{S \in \mathcal{H} : (S,W) \text{ is bad and } |S| = s\}|\right] \leq (\gamma r)^{-1}|\mathcal{H}|C^{-r/3}, \tag{8}$$

or, equivalently, that

$$|\{(S,W) : (S,W) \text{ is bad and } |S| = s\}| \leq (\gamma r)^{-1}N|\mathcal{H}|C^{-r/3}. \tag{9}$$

Note $\gamma r = r - r'$ bounds the number of s for which the set in question can be nonempty, whence the negligible factor $(\gamma r)^{-1}$.

We now use $\mathcal{H}_s = \{S \in \mathcal{H} : |S| = s\}$. Let $B = \sqrt{C}$ and for $Z \supset S \in \mathcal{H}_s$, say (S,Z) is *pathological* if there is $T \subset S$ with $t = |T| > r'$ and

$$|\{S' \in \mathcal{H}_s : S' \in [T,Z]\}| > B^r|\mathcal{H}|\kappa^{-t}p^{s-t}. \tag{10}$$

From now on, we will always take $Z = W \cup S$ (with W as in Lemma 3.1); thus $|Z|$ is typically roughly np and, since $\mathcal{H}$ is κ-spread, $|\mathcal{H}|\kappa^{-t}p^{s-t}$ is a natural upper bound on what one might expect for the left-hand side of (10).

Note that in proving (9), we may assume $s \le n/2$: we may of course assume $|\mathcal{H}_s|$ is at least the right-hand side of (8), but then for an $S \in \mathcal{H}_s$ of the largest multiplicity, say M, we have

$$M \le \kappa^{-s}|\mathcal{H}| \le \kappa^{-s}\gamma r C^{r/3}|\mathcal{H}_s| \le \kappa^{-s}\gamma r C^{r/3} M 2^n,$$

which is less than M if $s > n/2$ (since $\kappa > C$).

We bound the nonpathological and pathological parts of (9) separately; this, along with the introduction of the notion of 'pathological', is the source of our improvement over [2].

Nonpathological contributions. We first bound the number of (S, W) in (9) with (S, Z) nonpathological. This basically follows [2], but 'nonpathological' allows us to bound the number of possibilities in Step 3 below by the right-hand side of (10) (where [2] settles for something like $|\mathcal{H}|\kappa^{-t}$).

(1) There are at most

$$\sum_{i=0}^{s} \binom{n}{np+i} \le \binom{n+s}{np+s} \le Np^{-s} \tag{11}$$

choices for $Z = W \cup S$.

(2) Given Z, let $S' = \psi(Z)$. Choose $T = S \cap S'$, for which there are at most $2^{|S'|} \le 2^r$ possibilities, and set $t = |T| > r'$. If $t \le r'$ then, as $\chi(S, W) = S' \setminus W \subset T$, (S, W) cannot be bad.

(3) Since we are only interested in nonpathological choices, the number of possibilities for S is now at most

$$B^r|\mathcal{H}|\kappa^{-t}p^{s-t}.$$

(4) Complete the specification of (S, W) by choosing $W \cap S$, the number of possibilities for which is at most 2^s.

In sum, since $s \le r$ and $t > r' = (1-\gamma)r$, the number of nonpathological possibilities is at most

$$2^{r+s}N|\mathcal{H}|B^r(p\kappa)^{-t} \le N|\mathcal{H}|(4B)^r C^{-t} < N|\mathcal{H}| \left(4BC^{-(1-\gamma)}\right)^r. \tag{12}$$

Pathological contributions. We next bound the number of (S, W) as in (9) with (S, Z) pathological. The main point here is Step 4 below.

(1) There are at most $|\mathcal{H}|$ possibilities for S.

(2) Choose $T \subset S$ witnessing the pathology of (S, Z), i.e., for which (10) holds; there are at most 2^s possibilities for T.

(3) Choose $U \in [T,S]$ for which

$$|\mathcal{H}_s \cap [U,(Z\setminus S)\cup U]| > 2^{-(s-t)}B^r|\mathcal{H}|\kappa^{-t}p^{s-t}. \tag{13}$$

Here the left-hand side counts members of $\mathcal{H}_s$ in Z whose intersection with S is precisely U; of course, the existence of U as in (13) follows from (10). The number of possibilities for this choice is clearly at most 2^{s-t}.

(4) Choose $Z\setminus S$, the number of choices for which is less than $N(2/B)^r$. To see this, write Φ for the right-hand side of (13). Noting that $Z\setminus S$ must belong to $\binom{X\setminus S}{np}\cup\binom{X\setminus S}{np-1}\cup\cdots\cup\binom{X\setminus S}{np-s}$, we consider, for Y drawn uniformly from this set,

$$\mathbb{P}(|\mathcal{H}_s\cap[U,Y\cup U]|>\Phi). \tag{14}$$

Set $|U|=u$. We have

$$|\mathcal{H}_s\cap\langle U\rangle|\le|\mathcal{H}\cap\langle U\rangle|\le|\mathcal{H}|\kappa^{-u},$$

while, for any $S'\in\mathcal{H}_s\cap\langle U\rangle$,

$$\mathbb{P}(Y\supset S'\setminus U)\le\left(\frac{np}{n-s}\right)^{s-u},$$

though, of course, if $S'\cap S\ne U$, then the probability in question is zero. It follows that

$$\vartheta=\mathbb{E}\left[|\mathcal{H}_s\cap[U,Y\cup U]|\right]\le|\mathcal{H}|\kappa^{-u}\left(\frac{np}{n-s}\right)^{s-u}\le|\mathcal{H}|\kappa^{-u}(2p)^{s-u},$$

where the last inequality holds since $n-s\ge n/2$. Markov's inequality then bounds the probability in (14) by ϑ/Φ, and this bounds the number of possibilities for $Z\setminus S$ by $N(\vartheta/\Phi)$ (see (11)), which is easily seen to be less than $N(2/B)^r$.

(5) Complete the specification of (S,W) by choosing $S\cap W$, which can be done in at most 2^s ways.

Combining (and slightly simplifying), we find that the number of pathological possibilities is at most

$$|\mathcal{H}|N(16/B)^r. \tag{15}$$

Finally, the sum of the bounds in (12) and (15) is at most $(\gamma r)^{-1}N|\mathcal{H}|C^{-r/3}$, as required in (9), proving the lemma. □

As in [2], we handle small uniformities by a simple application of Janson's inequality.

Lemma 3.2. *For an r-bounded, κ-spread $\mathcal{G}$ on Y, and $\alpha\in(0,1)$,*

$$\mathbb{P}(Y_\alpha\notin\langle\mathcal{G}\rangle)\le\exp\left(-\left(\sum_{t=1}^{r}\binom{r}{t}(\alpha\kappa)^{-t}\right)^{-1}\right). \tag{16}$$

Proof. We may assume $\mathcal{G}$ is r-uniform, since modifying it according to Observation 2.1 does not decrease the probability in (16). Denote members of $\mathcal{G}$ by $S_1, S_2, \ldots, S_{|\mathcal{G}|}$ and set $\zeta_i = \mathbb{1}_{\{Y_\alpha \supset S_i\}}$. Then

$$\mu = \sum_{i=1}^{|\mathcal{G}|} \mathbb{E}[\zeta_i] = |\mathcal{G}|\alpha^r$$

and

$$\Lambda = \sum_{i=1}^{|\mathcal{G}|}\sum_{j=1}^{|\mathcal{G}|} \mathbb{E}[\zeta_i\zeta_j \mathbb{1}_{\{S_i\cap S_j\neq\emptyset\}}] \leq |\mathcal{G}|\sum_{t=1}^{r}\binom{r}{t}\kappa^{-t}|\mathcal{G}|\alpha^{2r-t} = \mu^2\sum_{t=1}^{r}\binom{r}{t}(\alpha\kappa)^{-t},$$

where the inequality holds because $\mathcal{G}$ is κ-spread, and Janson's inequality (see [14], for example) bounds the probability in (16) by $\exp(-\mu^2/\Lambda)$. □

Corollary 3.3. *Let $\mathcal{G}$ be as in Lemma 3.2, let $t = \alpha|Y|$ be an integer with $\alpha\kappa \geq 2r$, and let $W = Y_t$. Then*

$$\mathbb{P}(W \notin \langle\mathcal{G}\rangle) \leq 2\exp(-\alpha\kappa/(2r)).$$

Proof. Lemma 3.2 gives

$$\exp(-\alpha\kappa/(2r)) \geq \mathbb{P}(Y_\alpha \notin \langle\mathcal{G}\rangle) \geq \mathbb{P}(|Y_\alpha| \leq t)\mathbb{P}(W \notin \langle\mathcal{G}\rangle) \geq \mathbb{P}(W \notin \langle\mathcal{G}\rangle)/2,$$

where we use the fact (see [22]) that any binomial ξ with $\mathbb{E}[\xi] \in \mathbb{N}$ satisfies $\mathbb{P}(\xi \leq \mathbb{E}[\xi]) \geq 1/2$. □

4. Proofs of the main results

We now give the proofs of our two main results. We start with Theorem 1.6.

Proof of Theorem 1.6. It will be (very slightly) convenient to prove the theorem assuming $\mathcal{H}$ is (2κ)-spread. Let γ and C_0 be as in Section 3 and $\mathcal{H}$ as in the statement of Theorem 1.6, and recall that asymptotics refer to ℓ. We may of course assume that $\kappa \geq 2\gamma^{-1}C_0\log\ell$ (or the result is trivial with a suitably adjusted K).

Fix an arbitrary ordering $\prec$ of $\mathcal{H}$. In what follows, we will have a sequence of hypergraphs $\mathcal{H}_i$, with $\mathcal{H}_0 = \mathcal{H}$ and

$$\mathcal{H}_i \subset \{\chi_i(S, W_i) : S \in \mathcal{H}_{i-1}\},$$

where W_i and χ_i will be defined below (with χ_i a version of the χ of Section 3). We then order $\mathcal{H}_i$ by setting

$$\chi_i(S, W_i) \prec_i \chi_i(S', W_i) \Leftrightarrow S \prec_{i-1} S'.$$

In other words, each member of $\mathcal{H}_i$ ultimately inherits its position in $\prec_i$ from some member of $\mathcal{H}$. This is not very important; we will be applying Lemma 3.1 repeatedly, and the present convention just provides a concrete ψ for each stage of the iteration.

Set $C = C_0$ and $p = C/\kappa$, define m by $(1-\gamma)^m = \sqrt{\log \ell}/\ell$, and set $q = \log \ell/\kappa$. Then $m \leq \gamma^{-1}\log \ell$ and Theorem 1.6 will follow from the next assertion.

Claim 4.1. *If W is a uniform $((mp+q)n)$-subset of X, then $W \in \langle \mathcal{H} \rangle$ with high probability.*

Proof. Set $\delta = 1/(2m)$. Let $r_0 = \ell$ and $r_i = (1-\gamma)r_{i-1} = (1-\gamma)^i r_0$ for $i \in [m]$. Let $X_0 = X$ and, for $i = 1, 2, \ldots, m$, let W_i be uniform from $\binom{X_{i-1}}{np}$ and set $X_i = X_{i-1} \setminus W_i$. Finally, note the assumption that $\kappa \geq 2\gamma^{-1}C_0 \log \ell$ ensures $|X_m| \geq n/2$.

For $S \in \mathcal{H}_{i-1}$, let $\chi_i(S, W_i) = S' \setminus W_i$, where S' is the first member of $\mathcal{H}_{i-1}$ contained in $W_i \cup S$ (with $\mathcal{H}_{i-1}$ ordered by $\prec_{i-1}$). Say S is *good* if $|\chi_i(S, W_i)| \leq r_i$ and *bad* otherwise, and set

$$\mathcal{H}_i = \{\chi_i(S, W_i) : S \in \mathcal{H}_{i-1} \text{ is good}\}.$$

Thus, $\mathcal{H}_i$ is an r_i-bounded collection of subsets of X_i and inherits the ordering $\prec_i$ as described above.

Finally, choose W_{m+1} uniformly from $\binom{X_m}{nq}$. Then $W = W_1 \cup \cdots \cup W_{m+1}$ is as in Claim 4.1. Note also that $W \in \langle \mathcal{H} \rangle$ whenever $W_{m+1} \in \langle \mathcal{H}_m \rangle$, and more generally, $W_1 \cup \cdots \cup W_i \cup Y \in \langle \mathcal{H} \rangle$ whenever $Y \subset X_i$ lies in $\langle \mathcal{H}_i \rangle$.

To prove the claim, it suffices to show that

$$\mathbb{P}(W_{m+1} \in \langle \mathcal{H}_m \rangle) = 1 - o(1), \tag{17}$$

where $\mathbb{P}$ now refers to the entire sequence $W_1, W_2, \ldots, W_{m+1}$.

For $i \in [m]$, call W_i *successful* if $|\mathcal{H}_i| \geq (1-\delta)|\mathcal{H}_{i-1}|$, call W_{m+1} successful if it lies in $\langle \mathcal{H}_m \rangle$, and say a sequence of W_i's is successful if each of its entries is. We show a little more than (17) and prove that

$$\mathbb{P}(W_1, W_2, \ldots, W_{m+1} \text{ is successful}) = 1 - \exp\left(-\Omega(\sqrt{\log \ell})\right). \tag{18}$$

For $i \in [m]$, Lemma 3.1 and Markov's inequality give

$$\mathbb{P}(W_i \text{ is not successful} \,|\, W_1, W_2, \ldots, W_{i-1} \text{ is successful}) < \delta^{-1} C^{-r_{i-1}/3},$$

since $W_1, W_2, \ldots, W_{i-1}$ being successful implies that $|\mathcal{H}_{i-1}| > (1-\delta)^m |\mathcal{H}| > |\mathcal{H}|/2$, which ensures that $\mathcal{H}_{i-1}$ is κ-spread (since $\mathcal{H}$ is (2κ)-spread). Thus

$$\mathbb{P}(W_1, W_2, \ldots, W_m \text{ is successful}) > 1 - \delta^{-1} \sum_{i=1}^{m} C^{-r_{i-1}/3} > 1 - \exp\left(-\sqrt{\log \ell}\right), \tag{19}$$

where we use the fact that $r_m = \sqrt{\log \ell}$, $\delta^{-1} = 2m$, and the fact that $C = C_0$ is somewhat large.

Finally, if $W_1, W_2, \ldots, W_m$ is successful, then Corollary 3.3 (applied with $\mathcal{G} = \mathcal{H}_m$, $Y = X_m$, $\alpha = nq/|Y| \geq q$, $r = r_m$, and $W = W_{m+1}$) gives

$$\mathbb{P}(W_{m+1} \notin \langle \mathcal{H}_m \rangle) \leq 2\exp\left(-\sqrt{\log \ell}/2\right); \tag{20}$$

this yields (18) and the claim. □

The above claim implies the result; this completes the proof. □

Next, we give the proof of Theorem 1.7.

Proof of Theorem 1.7. We assume the setup of Theorem 1.7 with γ and C_0 as in Section 3 and $\kappa \geq C_0^2$ (or there is nothing to prove). We may assume $\mathcal{H}$ is ℓ-uniform, since the construction of Observation 2.1 produces an ℓ-uniform, κ-spread $\mathcal{G}$ with $\xi_{\mathcal{G}} \geq \xi_{\mathcal{H}}$. In particular, this gives

$$|\mathcal{H}|\ell = \sum_{x \in X} |\mathcal{H} \cap \langle x \rangle| \leq n\kappa^{-1}|\mathcal{H}|. \tag{21}$$

We first assume κ is somewhat large, precisely

$$\kappa \geq (\log \ell)^3; \tag{22}$$

the similar (but easier) argument for smaller values will be given at the end. It is worth mentioning while the bound in (22) provides a convenient demarcation, there is nothing delicate about this choice.

Claim 4.2. *For κ as in (22) and $C_0 \leq C \leq \gamma\kappa/(4\log \ell)$,*

$$\mathbb{P}(\xi_{\mathcal{H}} > (3C/\gamma)\ell/\kappa) < \exp(-(\log \ell \log C)/4).$$

Claim 4.2 is easily to seen to handle all κ as in (22). The 'with high probability' statement about $\xi_{\mathcal{H}}$ is immediate (take $C = C_0$). For the expectation $Z_{\mathcal{H}}$, set $t = (3C_0/\gamma)\ell/\kappa$ and $T = 3\ell/(4\log \ell)$. By Claim 4.2 we have, for all $x \in [t, T]$,

$$\mathbb{P}(\xi_{\mathcal{H}} > x) \leq f(x) = \exp\left(-\log \ell \log(\gamma\kappa x/3\ell)/4\right) = (bx)^a = b^a x^a,$$

where $a = -(\log \ell)/4$ and $b = \gamma\kappa/3\ell$. Noting that $\xi_{\mathcal{H}} \leq \ell$, we then have

$$Z_{\mathcal{H}} \leq t + \int_t^T \mathbb{P}(\xi_{\mathcal{H}} > x)dx + \ell\mathbb{P}(\xi_{\mathcal{H}} > T) \leq t + \int_t^T f(x)dx + \ell f(T) = O(\ell/\kappa).$$

Here, $t = O(\ell/\kappa)$ and the other terms are much smaller; the integral is less than $-1/(a+1)b^a t^{a+1} = O(C_0^a t/\log \ell)$, while (22) easily implies that $f(T) = (\gamma\kappa/(4\log \ell))^a$ is $o(1/\kappa)$.

Proof of Claim 4.2. Terms not defined here beginning with $p = C/\kappa$ and W_i (note C is now as in Claim 4.2, rather than set to C_0) are as in the proof of Theorem 1.6, but we now define m by $(1-\gamma)^m = \log \ell/\ell$ and set $q = \log C(\log \ell)^2/\kappa$, noting that (21) gives $p \geq C\ell/n$.

It is now convenient to generate the W_i's using the ξ_x's in the natural way: let

$$a_i = \begin{cases} (ip)n & \text{if } i \in \{0,1,\ldots,m\}, \\ (mp+q)n & \text{if } i = m+1, \end{cases}$$

and let W_i consist of the x's in positions $a_{i-1}+1, a_{i-1}+2, \ldots, a_i$ when X is ordered according to the ξ_x's.

Proposition 4.3. *With probability* $1 - \exp(-\Omega(C\ell))$, *for all* $i \in \{0,1,\ldots,m+1\}$ *and* $x \in W_i$, *we have*

$$\xi_x \le \varepsilon_i = \begin{cases} 2ip & \text{if } i \in \{0,1,\ldots,m\} \\ 2(mp+q) & \text{if } i = m+1. \end{cases} \tag{23}$$

Proof. Failure at $i \ge 1$ implies that

$$|\xi^{-1}[0,\varepsilon_i]| < a_i. \tag{24}$$

But $|\xi^{-1}[0,\varepsilon_i]|$ is binomial with mean $\varepsilon_i n = 2a_i \ge 2C\ell$, so the probability that (24) occurs for some i is less than $\exp(-\Omega(C\ell))$; see [14], for example. $\square$

Now writing $\overline{W}_i$ for $W_1 \cup \cdots \cup W_i$, we have the following fact.

Proposition 4.4. *If* $W_{m+1} \in \langle \mathcal{H}_m \rangle$, *then* W *contains some* $S \in \mathcal{H}$ *with*

$$|S \setminus \overline{W}_i| \le r_i \quad \forall i \in [m].$$

Proof. Suppose $W \supset S_m \in \mathcal{H}_m$. By the construction of the $\mathcal{H}_i$'s, there exist

$$S_{m-1}, S_{m-2}, \ldots, S_1, S_0 = S$$

with $S_i \in \mathcal{H}_i$ and $S_i = S_{i-1} \setminus W_i$, whence $S_i = S \setminus \overline{W}_i$ for $i \in [m]$; then $S_i \in \mathcal{H}_i$ gives the proposition. $\square$

We now define 'success' for $\{\xi_x : x \in X\}$ to mean that $W_1, W_2, \ldots, W_{m+1}$ is successful in our earlier sense *and* (23) holds. Notice that with our current values of m and q (and $r_m = \ell(1-\gamma)^m = \log \ell$), we can replace the error terms in (19) and (20) by $\delta^{-1}C^{-\log \ell/3}$ and $e^{-(\log C \log \ell)/2}$, which with Proposition 4.3 bounds the probability that $\{\xi_x : x \in X\}$ is *not* successful by, say, $\exp(-(\log \ell \log C)/4)$. We finish with the following observation.

Proposition 4.5. *If* $\{\xi_x : x \in X\}$ *is successful, then* $\xi_{\mathcal{H}} \le (3C/\gamma)\ell/\kappa$.

Proof. For S as in Proposition 4.4, we have (with $W_0 = \emptyset$ and $\varepsilon_0 = 0$)

$$\begin{aligned} \xi_{\mathcal{H}} &\le \sum_{i=1}^{m+1} \varepsilon_i |S \cap W_i| = \sum_{i=1}^{m+1} (\varepsilon_i - \varepsilon_{i-1}) |S \setminus \overline{W}_{i-1}| \le 2 \left(\sum_{i=1}^{m} (1-\gamma)^{i-1} p + (1-\gamma)^m q \right) \ell \\ &\le 2(C/(\gamma\kappa) + (\log \ell/\ell)(\log C (\log \ell)^2/\kappa))\ell < (3C/\gamma)\ell/\kappa. \end{aligned}$$

$\square$

This completes the proof of the claim, and also the theorem, when κ satisfies (22). $\square$

Finally, for κ below the bound in (22), a subset of the preceding argument suffices. We proceed as before, but now with $C = C_0$ (so $p = C_0/\kappa$), stopping at m defined by $(1-\gamma)^m = 1/\kappa$ (so $m = \Theta(\gamma^{-1}\log\kappa)$). Here, the main difference is that there is no 'Janson' phase: $W_1, W_2, \ldots, W_m$ is successful with probability $1 - \exp(-\Omega(\ell/\kappa))$, and when it *is* successful, we have (as in the proof of Proposition 4.5, but now taking $W_{m+1} = X \setminus \overline{W}_m$)

$$\xi_{\mathcal{H}} \leq \sum_{i=1}^{m} (\varepsilon_i - \varepsilon_{i-1})|S \setminus \overline{W}_{i-1}| + |S \cap W_{m+1}| < 2(C_0/(\gamma\kappa))\ell + \ell/\kappa;$$

of course, we also get $Z_{\mathcal{H}} \leq O(\ell/\kappa) + \exp(-\Omega(\ell/\kappa))\ell = O(\ell/\kappa)$. □

5. Applications

Much of the significance of Theorem 1.1 — and of the skepticism with which Conjecture 1.2 was viewed in [17] — derives from the strength of its consequences, a few of which we discuss briefly here. For this discussion, $\mathcal{K}_n^r = \binom{V}{r}$ is the complete r-graph on $V = [n]$, and $\mathcal{H}_{n,p}^r$ is the binomial random r-graph, i.e., the r-uniform counterpart of the usual binomial random graph $G_{n,p}$. Given $r, n \in \mathbb{N}$ and an r-graph H, we use $\mathcal{G}_H$ for the collection of unlabelled copies of H in $\mathcal{K} = \mathcal{K}_n^r$ and $\mathcal{F}_H$ for $\langle \mathcal{G}_H \rangle$, and as usual, write $\Delta(H)$ for the maximum degree of H.

As noted earlier, Conjecture 1.2 was motivated especially by Shamir's problem, since resolved in [15], and the conjecture that became Montgomery's theorem [24]. A brief summary is as follows: for fixed $r \in \mathbb{N}$ and n running over multiples of r, Shamir's problem asks for the estimation of $p_c(\mathcal{F}_H)$ when H is a perfect matching (i.e., n/r disjoint edges), and [15] proves the natural conjecture that this threshold is $\Theta(n^{-(r-1)}\log n)$; next, for fixed $d \in \mathbb{N}$, [24] shows that the threshold for $G_{n,p}$ to contain a given n-vertex tree with maximum degree d is $\Theta(n^{-1}\log n)$, where the implied constant in the upper bound depends on d (though it probably should not). In both of these — and in most of the other examples mentioned below following Theorem 5.1 (with the exception of the one from [20]) — the lower bounds derive from the 'coupon-collector' requirement that the edges cover the vertices, and it is the upper bounds that are of interest.

In fact, Theorem 1.1 gives not just Montgomery's theorem, but its natural extension to r-graphs and more. Say an r-graph F is a *forest* if it contains no *cycle*, meaning distinct vertices $v_1, v_2, \ldots, v_k$ and distinct edges $e_1, e_2, \ldots, e_k$ such that $v_{i-1}, v_i \in e_i$ for all i (with subscripts mod k). A *spanning tree* is then a forest of size $(n-1)/(r-1)$. For an r-graph F, let $\rho(F)$ be the maximum size of a forest in F and set

$$\varphi(F) = \max\{1 - \rho(F')/|F'| : \emptyset \neq F' \subset F\}.$$

Theorem 5.1. *For each $r \in \mathbb{N}$ and $c > 0$, there is a $K > 0$ such that if H is an r-graph on $[n]$ with $\Delta(H) \leq d$ and $\varphi(H) \leq c/\log n$, then*

$$p_c(\mathcal{F}_H) < Kdn^{-(r-1)} \log |H|. \qquad \square$$

For example, this gives $p_c(\mathcal{F}_H) = \Theta(n^{-(r-1)} \log n)$ if H is a perfect matching (as in Shamir's problem) or a 'loose' Hamiltonian cycle (a result of [5], to which we refer for the history of the problem). This result also gives $p_c(\mathcal{F}_H) < Kdn^{-(r-1)} \log n$ if H is a spanning tree with $\Delta(H) \leq d$; for $d = O(1)$, this is the aforementioned r-graph generalisation of [24], for $d = n^{\Omega(1)}$, it is a result of Krivelevich [20], and this bound is tight up to value of the constant K in both regimes.

The last application we discuss here concerns bounded-degree spanning graphs. Writing $c_d = (d!)^{2/(d(d+1))}$ and $p^*(d, n) = c_d n^{-2/(d+1)}(\log n)^{2/(d(d+1))}$, we have the following.

Theorem 5.2. *For fixed $d \in \mathbb{N}$ and any graph H on $[n]$ with $\Delta(H) \leq d$,*

$$p_c(\mathcal{F}_H) < (1 + o(1))p^*(d, n). \tag{25}$$

When $(d+1) \mid n$ and H is a K_{d+1}-factor (i.e., $n/(d+1)$ disjoint K_{d+1}'s), $p^*(d, n)$ is the asymptotic value of $p_c(\mathcal{F}_H)$; in this case, (25) with $O(1)$ in place of $1 + o(1)$ was proved in [15], while the precise asymptotics are given by the combination of [16] and [26, 13].

Theorem 5.3. *For fixed $d \in \mathbb{N}$ and $\varepsilon > 0$, and n ranging over multiples of $d + 1$, if $p > (1 + \varepsilon)p^*(d, n)$, then $G_{n,p}$ contains a K_{d+1}-factor with high probability.* $\square$

Interest in $p_c(\mathcal{F}_H)$ for H as in Theorem 5.2 dates to at least the early-90s, when Alon and Füredi [1] showed an upper bound of $O(n^{-1/d}(\log n)^{1/d})$, and has intensified since [15], motivated by the idea that K_{d+1}-factors should be hardest such graphs to find. See [8, 9] for both the history and the most recent results; Theorem 5.2 is conjectured in [9] (with $O(1)$ in place of $1 + o(1)$), and in a stronger 'universal' form in [8].

Remark 5.4. *Theorem 5.2 likely extends to r-graphs and d of the form $\binom{s-1}{r-1}$ with $s \in \mathbb{N}$. This just needs the extension of the main result of [26] to r-graphs (suggested at the end of [26]), which (with [16]) would give asymptotics of the threshold for $\mathcal{H}^r_{n,p}$ to contain a $\mathcal{K}^r_s$-factor.*

Each of Theorems 5.1 and 5.2 begins with the following easy observations. The first, an approximate converse of Proposition 1.5, is the trivial direction of linear programming duality.

Observation 5.5. *If an increasing $\mathcal{F}$ supports a q-spread measure, then $q_f(\mathcal{F}) < q$.*

The second allows us to compute the spread of the hypergraphs $\mathcal{G}_H$; recall that $\mathcal{G}_H$ is the collection of unlabelled copies of H in $\mathcal{K}^r_n$.

Observation 5.6. *The uniform measure on $\mathcal{G}_H$ is q-spread if and only if for each $S \subset \mathcal{K}_n^r$ isomorphic to a subhypergraph of H, σ a uniformly random permutation of V, and $H_0 \subset \mathcal{K}_n^r$ a given copy of H, we have*

$$\mathbb{P}(\sigma(S) \subset H_0) \leq q^{|S|}. \tag{26}$$

Proving Theorem 5.1 is now just a matter of verifying (26) with $q = O(dn^{-(r-1)})$, which we leave to the reader; the calculation is similar to (28) below. Theorem 5.2 requires a short proof, to which we now turn.

Proof of Theorem 5.2. As one might expect, we use Theorem 5.3 for embedding the copies of K_{d+1} and Theorem 1.1 for the rest of H (where we have more room), ordering these two steps so that the second is still looking at a suitably large number of vertices.The next assertion is the main thing we need to check here.

Lemma 5.7. *There is an $\varepsilon = \varepsilon_d > 0$ such that if H is as in Theorem 5.2 and has no component isomorphic to K_{d+1}, then*

$$q_f(\mathcal{F}_H) \leq n^{-(2/(d+1)+\varepsilon)}. \tag{27}$$

Proof. We just need to show (26) for $q = n^{-(2/(d+1)+\varepsilon)}$ and S, H_0 as in Observation 5.6, say with $W = V(S)$, $s = |S|$, and f the size of a spanning forest of S. We may of course assume S has no isolated vertices, so $w = |W| \leq 2f$. We show that

$$\mathbb{P}(\sigma(S) \subset H_0) < (e^2 d/n)^f, \tag{28}$$

and that

$$\frac{f}{s} \geq \frac{2(d+1)}{(d+2)d} = \frac{2}{d+1} + \varepsilon_0, \tag{29}$$

where $\varepsilon_0 = 1/((d+2)(d+1)d)$, implying, for any fixed $\varepsilon < \varepsilon_0$, that (26) holds for large enough n.

Proof of (28). Let $\alpha, \beta : W \to V$ be, respectively, a uniformly random injection and a uniformly random map. Then

$$\begin{aligned}(d/n)^f &\geq \mathbb{P}(\beta(S) \subset H_0) \geq \mathbb{P}(\beta \text{ is injective})\mathbb{P}(\beta(S) \subset H_0 \,|\, \beta \text{ is injective}) \\ &= \left(n^{-w}\prod_{i=0}^{w-1}(n-i)\right)\mathbb{P}(\alpha(S) \subset H_0) > e^{-2f}\mathbb{P}(\sigma(S) \subset H_0). \qquad \square\end{aligned}$$

Proof of (29). We may of course assume S is connected, in which case we have $f = w-1$ and the following upper bounds on s: $\binom{w}{2}$ if $w \leq d$, $\binom{d+1}{2} - 1$ if $w = d+1$, and $wd/2$ if $w \geq d+2$. The corresponding lower bounds on f/s are $2/d$, $2d/((d+2)(d+1)-2)$ and $2(d+1)/((d+2)d)$, the smallest of which is the last. $\square$

This completes the proof of Lemma 5.7. $\square$

We are now ready for Theorem 5.2. Let $\varsigma = \varsigma(n)$ be a function going slowly to 0 as n grows (for example, $1/\log n$ suffices with room to spare). By Theorem 5.3, there is $p_1 \sim p^*(d,n)$ such that if $m > (1-\varsigma)n$ and $(d+1) \mid m$, then G_{m,p_1} contains a K_{d+1}-factor with high probability, while by Lemma 5.7 and Theorem 1.1 (or, more precisely, Remark 2.2), there is p_2 with $p^*(d,n) \gg p_2 \gg n^{-(2/(d+1)+\varepsilon_d)}$ such that if $m \geq \varsigma n$, then for any given m-vertex K_{d+1}-free H' with $\Delta(H') \leq d$, G_{m,p_2} contains a copy of H' with high probability.

The above two facts allow us to finish with a standard two-round exposure argument. Let H_1 be the union of the copies of K_{d+1} in H (each of which must be a component of H), $H_2 = H - H_1$, and $n_i = |V(H_i)|$ for $i = 1,2$ so that $n_1 + n_2 = n$. Let $G_1 \sim G_{n,p_1}$ and $G_2 \sim G_{n,p_2}$ be independent on the common vertex set $V = [n]$ and $G = G_1 \cup G_2$. Then $G \sim G_{n,p}$ with $p = 1-(1-p_1)(1-p_2) \sim p^*(d,n)$ and we just need to show G contains a copy of H with high probability. In fact, we find each H_i in the corresponding G_i, in order depending on n_2: if $n_2 \geq \varsigma n$, then with high probability, G_1 contains H_1, say on vertex set V_1, and with high probability $G_2[V \setminus V_1]$ contains H_2; if $n_2 < \varsigma n$, then with high probability G_2 contains H_2 on some V_2, and with high probability $G_1[V \setminus V_2]$ contains H_1. □

6. Concluding remarks

A number of open problems remain, perhaps the most basic of which is Conjecture 1.4; settling this conjecture would now imply Conjecture 1.2. We briefly mention a few other unresolved issues related to the present work below.

It would be interesting to understand whether, in Shamir's and related problems, the $\log \ell$ emerging from our argument somehow reflects the coupon-collector requirement (edges must cover vertices) that drives the lower bounds. Partly as a way of testing this, one might try to see if the present machinery can be extended to apply directly (rather than via [26, 13]) to questions where coupon-collector considerations (correctly) predict a smaller gap, as in the fractional powers of $\log n$ in Theorem 5.3.

The arguments of [24] and [9] give stronger 'universality' results; for example, [24] says that the appropriate random graph with high probability contains *every* tree respecting the degree bound. Whether this can be proved along present lines remains unclear; if so, it would seem to be more a question of managing some understanding of the class of universal graphs (with, of course, a view to the spread) than of extending Theorem 1.1.

As mentioned following Corollary 1.8, what prevents us from extending our results on the assignment problem to other 'k-dimensional' variants is inadequate control of the spread. The difficulty is the same for the related problem of locating thresholds for the existence of designs. We unfortunately do not have anything to suggest in the way

of a remedy and just indicate one issue, for simplicity sticking to Steiner triple systems (see [33] for some background). When $X = \mathcal{K}_n^3$ (with $n \equiv 1$ or $3 \pmod 6$) and $\mathcal{H}$ is the hypergraph of all Steinter triple systems on X, for the spread κ of $\mathcal{H}$ (which in theory should be $\Theta(1/n)$), we may take

$$\kappa = \min_{S \subset X} \left(|\mathcal{H}|/|\mathcal{H} \cap \langle S \rangle|\right)^{1/|S|}. \tag{30}$$

Results of Linial and Luria [21] (upper bound) and Keevash [19] (lower bound) give

$$|\mathcal{H}| = \left((1/e^2 + o(1))n\right)^{n^2/6}. \tag{31}$$

Viewed enumeratively, this is very satisfactory, having been an old conjecture of Wilson [34]. But for present purposes, even ignoring our weaker understanding of $|\mathcal{H} \cap \langle S \rangle|$ (i.e., the number of completions of a partial Steiner triple system S), it is not enough: even if this quantity is, as one expects, roughly $(n/e^2)^{n^2/6 - |S|}$, the bounds of (30) can be dominated by the 'error' term $(1 + o(1))^{n^2/(6|S|)}$ if S is small and the $o(1)$ in (31) is negative.

Finally, we recall a related conjecture from [17] (stated there only for graphs, but this should not matter). For $\mathcal{F} = \mathcal{F}_H$ as in Section 5, let $p_e(\mathcal{F})$ to be the least p such that for every $H' \subset H$, the expected number of unlabeled copies of H' in $\mathcal{H}_{n,p}^r$ is at least 1. Then $p_e(\mathcal{F})/2$ is again a trivial lower bound on $p_c(\mathcal{F})$ — and, where it makes sense, probably more intuitive than $q(\mathcal{F})$ or $q_f(\mathcal{F})$ — and we have the following from [17].

Conjecture 6.1. *There is a universal $K > 0$ such that for every $\mathcal{F} = \mathcal{F}_H$ as above,*

$$p_c(\mathcal{F}) \le K p_e(\mathcal{F}) \log |X|.$$

Again, we can presumably replace $\log |X|$ by $\log |H|$, as would now follow from a positive answer to the obvious question: do we always have $q_f(\mathcal{F}) = O(p_e(\mathcal{F}))$?

Acknowledgements

The first, second and fourth authors were supported by NSF grant DMS-1501962 and BSF Grant 2014290, and the third author was supported by NSF grant DMS-1800521. We would like to thank Simon Griffiths and Rob Morris for suggesting the application of our results to bounded-degree spanning graphs.

References

1. N. Alon and Z. Füredi, *Spanning subgraphs of random graphs*, Graphs Combin. **8** (1992), 91–94. 15
2. R. Alweiss, S. Lovett, K. Wu, and J. Zhang, *Improved bounds for the sunflower lemma*, Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing, STOC 2020, Association for Computing Machinery, 2020, p. 624–630. 4, 5, 7, 8, 9

3. B. Bollobás, *Random graphs*, Cambridge Studies in Advanced Mathematics, vol. 73, Cambridge University Press, Cambridge, 2001. 2
4. B. Bollobás and A. Thomason, *Threshold functions*, Combinatorica **7** (1987), 35–38. 2
5. A. Dudek, A. Frieze, P. Loh, and S. Speiss, *Optimal divisibility conditions for loose Hamilton cycles in random hypergraphs*, Electron. J. Combin. **19** (2012), Paper 44, 17. 15
6. P. Erdős and R. Rado, *Intersection theorems for systems of sets*, J. London Math. Soc. **35** (1960), 85–90. 5
7. P. Erdős and A. Rényi, *On the evolution of random graphs*, Magyar Tud. Akad. Mat. Kutató Int. Közl. **5** (1960), 17–61. 1
8. A. Ferber, G. Kronenberg, and K. Luh, *Optimal threshold for a random graph to be 2-universal*, Trans. Amer. Math. Soc. **372** (2019), 4239–4262. 15
9. A. Ferber, K. Luh, and O. Nguyen, *Embedding large graphs into a random graph*, Bull. Lond. Math. Soc. **49** (2017), 784–797. 15, 17
10. E. Friedgut, *Sharp thresholds of graph properties, and the k-sat problem*, J. Amer. Math. Soc. **12** (1999), 1017–1054, With an appendix by J. Bourgain. 2
11. ———, *Hunting for sharp thresholds*, Random Structures Algorithms **26** (2005), 37–51. 2
12. A. Frieze and G. B. Sorkin, *Efficient algorithms for three-dimensional axial and planar random assignment problems*, Random Structures Algorithms **46** (2015), 160–196. 2, 5
13. A. Heckel, *Random triangles in random graphs*, Preprint, arXiv:1802.08472. 15, 17
14. S. Janson, T. Łuczak, and A. Rucinski, *Random graphs*, Wiley-Interscience Series in Discrete Mathematics and Optimization, Wiley-Interscience, New York, 2000. 2, 10, 13
15. A. Johansson, J. Kahn, and V. Vu, *Factors in random graphs*, Random Structures Algorithms **33** (2008), 1–28. 2, 5, 14, 15
16. J. Kahn, *Asymptotics for Shamir's problem*, Preprint, arXiv:1909.06834. 15
17. J. Kahn and G. Kalai, *Thresholds and expectation thresholds*, Combin. Probab. Comput. **16** (2007), 495–502. 1, 2, 14, 18
18. R. M. Karp, *Reducibility among combinatorial problems*, Complexity of computer computations (Proc. Sympos., IBM Thomas J. Watson Res. Center, Yorktown Heights, New York.), 1972, pp. 85–103. 5
19. P. Keevash, *Counting designs*, J. Eur. Math. Soc. (JEMS) **20** (2018), no. 4, 903–927. 18
20. M. Krivelevich, *Embedding spanning trees in random graphs*, SIAM J. Discrete Math. **24** (2010), 1495–1500. 14, 15
21. N. Linial and Z. Luria, *An upper bound on the number of Steiner triple systems*, Random Structures Algorithms **43** (2013), 399–406. 18

22. N. Lord, *Binomial averages when the mean is an integer*, Math. Gaz. **94** (2010), 331–332. 10
23. O. C. Martin, M. Mézard, and O. Rivoire, *Random multi-index matching problems*, J. Stat. Mech. Theory Exp. (2005), P09006, 36. 5
24. R. Montgomery, *Spanning trees in random graphs*, Adv. Math. **356** (2019), 106793, 92. 3, 14, 15, 17
25. A. Rao, *Coding for sunflowers*, Discrete Anal. (2020), Paper No. 2, 8. 5
26. O. Riordan, *Random cliques in random graphs*, Preprint, arXiv:1802.01948. 15, 17
27. M. Talagrand, *Are all sets of positive measure essentially convex?*, Geometric aspects of functional analysis (Israel, 1992–1994), Oper. Theory Adv. Appl., vol. 77, Birkhäuser, Basel, 1995, pp. 295–310. 2, 3
28. ———, *The generic chaining*, Springer Monographs in Mathematics, Springer-Verlag, Berlin, 2005. 2, 4
29. ———, *Selector processes on classes of sets*, Probab. Theory Related Fields **135** (2006), 471–486. 4
30. ———, *Are many small sets explicitly small?*, Proceedings of the 2010 ACM International Symposium on Theory of Computing, 2010, pp. 13–35. 1, 2, 3, 4
31. ———, *Upper and lower bounds for stochastic processes*, A Series of Modern Surveys in Mathematics, vol. 60, Springer, Heidelberg, 2014. 4
32. T. Tao, *The sunflower lemma via Shannon entropy*, What's New, Blog Post at `terrytao.wordpress.com/2020/07/20/`. 5
33. J. H. van Lint and R. M. Wilson, *A course in combinatorics*, second ed., Cambridge University Press, Cambridge, 2001. 18
34. R. M. Wilson, *Nonisomorphic Steiner triple systems*, Math. Z. **135** (1973/74), 303–313. 18

Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA
Email address: `keith.frankston@math.rutgers.edu`

Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA
Email address: `jkahn@math.rutgers.edu`

Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA
Email address: `narayanan@math.rutgers.edu`

Department of Mathematics, Rutgers University, Piscataway, NJ 08854, USA
Email address: `jp1324@math.rutgers.edu`