LONGEST PATHS IN RANDOM HYPERGRAPHS

OLIVER COOLEY*, FREDERIK GARBE†, ENG KEAT HNG‡, MIHYUN KANG*, NICOLÁS SANHUEZA-MATAMALA§, JULIAN ZALLA*

Abstract. Given integers k, j with $1 \leq j \leq k-1$, we consider the length of the longest j-tight path in the binomial random k-uniform hypergraph $H^k(n,p)$. We show that this length undergoes a phase transition from logarithmic length to linear and determine the critical threshold, as well as proving upper and lower bounds on the length in the subcritical and supercritical ranges.

In particular, for the supercritical case we introduce the `Pathfinder` algorithm, a depth-first search algorithm which discovers j-tight paths in a k-uniform hypergraph. We prove that, in the supercritical case, with high probability this algorithm will find a long j-tight path.

1. Introduction

The celebrated phase transition result of Erdős and Rényi [10] for random graphs states, in modern terminology, that the binomial random graph[1] $G(n,p)$ displays a dramatic change in the order of the largest component when p is approximately $1/n$. If p is slightly smaller than $1/n$, then whp[2] all components are at most of logarithmic order, while if p is slightly larger than $1/n$, then there is a unique "giant" component of linear order and all other components are again of logarithmic order.

1.1. Paths in random graphs.

While by definition any two vertices in a component are connected by a path, there is not necessarily a correlation between the order of the component and the lengths of such paths. Of course, if a component is small, then it can only contain short paths, but if a component is large, this does not guarantee the existence of a long path. Nevertheless, Ajtai, Komlós and Szemerédi [1] showed that if p is larger than $1/n$, then whp $G(n,p)$ does indeed contain a path of linear length.

Incorporating various extensions of the results of Erdős and Rényi and of Ajtai, Komlós and Szemerédi by Pittel [19], by Łuczak [16], and by Kemkes and Wormald [13], gives the following.

Theorem 1. *Let L denote the length of the longest path in $G(n,p)$.*

(i) *If $0 < \varepsilon < 1$ is a constant and $p = \frac{1-\varepsilon}{n}$, then for any $\omega = \omega(n)$ such that $\omega \xrightarrow{n\to\infty} \infty$, whp*

$$\frac{\ln n - \omega}{-\ln(1-\varepsilon)} \leq L \leq \frac{\ln n + \omega}{-\ln(1-\varepsilon)}.$$

(ii) *If $0 < \varepsilon = \varepsilon(n) = o(1)$ satisfies $\varepsilon^5 n \to \infty$ and $p = \frac{1+\varepsilon}{n}$, then whp*

$$\left(\frac{4}{3} + o(1)\right)\varepsilon^2 n \leq L \leq (1.7395 + o(1))\varepsilon^2 n.$$

*Supported by Austrian Science Fund (FWF): I3747, W1230, `{cooley,kang,zalla}@math.tugraz.at`, Institute of Discrete Mathematics, Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria.

†Supported by GAČR project 18-01472Y and RVO: 67985840, `garbe@fi.muni.cz`, Masaryk University, Faculty of Informatics, Botanická 68A, 602 00, Brno, Czech Republic.

‡Supported by an LSE PhD Studentship, `e.hng@lse.ac.uk`, Department of Mathematics, London School of Economics, Houghton Street, London, WC2A 2AE, United Kingdom.

§Supported by the Czech Science Foundation, grant number GA19-08740S with institutional support RVO: 6798580. `nicolas@sanhueza.net`, The Czech Academy of Sciences, Institute of Computer Science, Pod Vodárenskou věží 2, 182 07 Prague, Czech Republic.

[1] $G(n,p)$ is the random graph on vertex set $[n] := \{1,\dots,n\}$, in which each pair of vertices is connected by an edge with probability p independently.

[2] Short for "with high probability", meaning with probability tending to 1 as n tends to infinity.

Let us also note that very recently, Anastos and Frieze [3] determined L asymptotically in the range when $p = c/n$ for a sufficiently large constant c (in particular, c is much larger than 1).

In fact, the bounds in the supercritical case followed from results about the length of the longest *cycle*. These original results also hold under the weaker assumption that $\varepsilon^3 n \to \infty$, and in particular the lower bound for paths is still valid even with this weaker assumption. For the upper bound, however, the standard sprinkling argument to show that the longest cycle is not significantly shorter than the longest path breaks down when $\varepsilon = O(n^{-5})$, and so we would no longer obtain the upper bound on L in the supercritical case.

Our goal in this paper is to generalise Theorem 1 for various notions of paths in random *hypergraphs*.

1.2. **Main result: paths in hypergraphs.** Given a natural number k, a *k-uniform hypergraph* consists of a vertex set V and an edge set E, where each edge consists of precisely k distinct vertices. Thus a 2-uniform hypergraph is simply a graph. Let $H^k(n,p)$ denote the binomial random k-uniform hypergraph on vertex set $[n]$ in which each set of k distinct vertices forms an edge with probability p independently. Thus in particular $H^2(n,p) = G(n,p)$.

There are several different ways of generalising the concept of paths in k-uniform hypergraphs. One important concept leads to a whole family of different types of paths which have been extensively studied. Each path type is defined by a parameter $j \in [k-1]$, which is a measure of how tightly connected the path is. Formally, we have the following definition.

Definition 2. Let $k, j \in \mathbb{N}$ satisfy $1 \le j \le k-1$ and let $\ell \in \mathbb{N}$. A *j-tight path of length ℓ* in a k-uniform hypergraph consists of a sequence of distinct vertices $v_1, \dots, v_{\ell(k-j)+j}$ and a sequence of edges $e_1, \dots, e_\ell$, where $e_i = \{v_{(i-1)(k-j)+1}, \dots, v_{(i-1)(k-j)+k}\}$ for $i = 1, \dots, \ell$, see Figure 1.

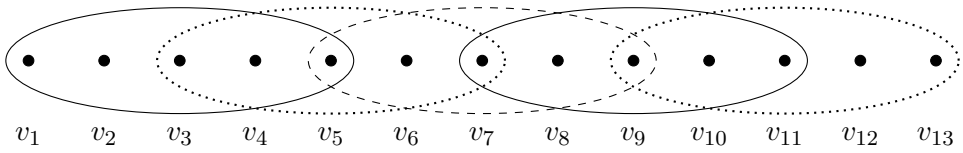

FIGURE 1. A 3-tight path of length 5 in a 5-uniform hypergraph

Note that the case $k = 2$ and $j = 1$ simply defines a path in a graph. For $k \ge 3$, the case $j = 1$ is often called a *loose path*, while the case $j = k-1$ is often called a *tight path*.

The main result of this paper is a phase transition result for j-tight paths similar to Theorem 1.

Definition 3. We use the notation $f \ll g$ to mean that $f \le g/C$ for some sufficiently large constant C, and similarly $f \gg g$ to mean that $f \ge Cg$ for some sufficiently large constant C.

Theorem 4. *Let $k, j \in \mathbb{N}$ satisfy $1 \le j \le k-1$. Let $a \in \mathbb{N}$ be the unique integer satisfying $1 \le a \le k-j$ and $a \equiv k \bmod (k-j)$. Let $\varepsilon = \varepsilon(n) \ll 1$ satisfy $\varepsilon^3 n \xrightarrow{n\to\infty} \infty$ and let*

$$p_0 = p_0(n;k,j) := \frac{1}{\binom{k-j}{a}\binom{n-j}{k-j}}.$$

Let L be the length of the longest j-tight path in $H^k(n,p)$.

(i) *If $p = (1-\varepsilon)p_0$, then whp*

$$\frac{j\ln n - \omega + 3\ln\varepsilon}{-\ln(1-\varepsilon)} \le L \le \frac{j\ln n + \omega}{-\ln(1-\varepsilon)},$$

for any $\omega = \omega(n)$ such that $\omega \xrightarrow{n\to\infty} \infty$.

(ii) *If $p = (1+\varepsilon)p_0$ and $j \ge 2$, then for any δ satisfying $\delta \gg \max\{\varepsilon, \frac{\ln n}{\varepsilon^2 n}\}$, whp*

$$(1-\delta)\frac{\varepsilon n}{(k-j)^2} \le L \le (1+\delta)\frac{2\varepsilon n}{(k-j)^2}.$$

(iii) *If* $p = (1+\varepsilon)p_0$ *and* $j = 1$, *then for all* $\delta \gg \varepsilon$ *satisfying* $\delta^2\varepsilon^3 n \xrightarrow{n\to\infty} \infty$, *whp*

$$(1-\delta)\frac{\varepsilon^2 n}{4(k-1)^2} \le L \le (1+\delta)\frac{2\varepsilon n}{(k-1)^2}.$$

In other words, we have a phase transition at threshold p_0.

The upper bounds in all three cases can be proved using the first moment method. The lower bound in the subcritical case, i.e. in (i), will be proved using the second moment method—while the strategy is standard, there are significant technical complications to be overcome. However, the second moment method is not strong enough in the *supercritical* cases, and therefore we will prove the lower bounds in (ii) and (iii) by introducing the `Pathfinder` search algorithm which explores j-tight paths in k-uniform hypergraphs, and which is the main contribution of this paper. The algorithm is based on a depth-first search process, but it is a rather delicate task to design it in such a way that it both correctly constructs j-tight paths and also admits reasonable probabilistic analysis. We will analyse the likely evolution of this algorithm and prove that whp it discovers a j-tight path of the appropriate length.

To help interpret Theorem 4, let us first observe that the results become stronger for smaller δ, so δ may be thought of as an error term. Furthermore, in all cases of the theorem we may choose δ to be no larger than an arbitrarily small constant, while in some cases we may even have $\delta \to 0$. In the subcritical regime (Theorem 4(i)), note that $-\ln(1-\varepsilon) = \varepsilon + O(\varepsilon^2)$ and that the term $3\ln\varepsilon$ in the lower bound becomes negligible (and in particular could be incorporated into ω) if ε is constant. For smaller ε, however, it represents a gap between the lower and upper bounds. In the supercritical case for $j \ge 2$ (Theorem 4(ii)), the length L is certainly of order $\Theta(\varepsilon n)$, but the lower and upper bounds differ by approximately a multiplicative factor of 2. In the supercritical case for $j = 1$ (Theorem 4(iii)), the lower and upper bounds differ by a multiplicative factor of $\Theta(\varepsilon)$. This has subsequently been improved by Cooley, Kang and Zalla [7], who lowered the upper bound to within a constant of the lower bound by analysing a structure similar to the 2-core in random hypergraphs. We will discuss all of these bounds and how they might be improved in more detail in Section 9.

Remark 5. In fact, the statement of Theorem 4 has been slightly weakened compared to what we actually prove in order to improve the clarity. More precisely, the full strength of the assumption on δ in (iii) is only required for the lower bound; the upper bound would in fact hold for any $\delta \gg \max\{\varepsilon, \frac{\ln n}{\varepsilon^2 n}\}$ as in (ii) (c.f. Lemma 35 in arXiv:2003.14143). Furthermore, the assumption that $\delta \gg \frac{\ln n}{\varepsilon^2 n}$ in (ii) is only needed for the upper bound; the lower bound holds with just the assumption that $\delta \gg \varepsilon$ (c.f. Lemma 30).

1.3. **Related work.** The study of j-tight paths (and the corresponding notion of j-tight cycles) has been a central theme in hypergraph theory, with many generalisations of classical graph results, including Dirac-type and Ramsey-type (see [15, 17, 21] for surveys), as well as Erdős-Gallai-type results [2, 11].

There has also been some work on j-tight cycles in random hypergraphs. Dudek and Frieze [8, 9] determined the thresholds for the appearance of both loose and tight Hamilton cycles in $H^k(n,p)$, as well as determining the threshold for a j-tight Hamilton cycle up to a multiplicative constant. Recently, Narayanan and Schacht [18] pinpointed the precise value of the sharp threshold for the appearance of j-tight Hamilton cycles in k-uniform hypergraphs, provided that $k > j > 1$.

Theorem 4 addresses a range when p is significantly smaller than the threshold for a j-tight Hamilton cycle, and consequently the longest j-tight paths are far shorter. Recently Cooley [4] has extended the lower bound in Theorem 4(ii) to the range when $p = cp_0$ for some constant $c > 1$, and shown that with a much more difficult version of the common "sprinkling" argument, one can also find a j-tight *cycle* of approximately the same length.

Recall that for random graphs, the phase transition thresholds for the length of the longest path and the order of the largest component are both $1/n$. It is therefore natural to wonder whether something similar holds for j-tight paths in random hypergraphs, since for each $1 \le j \le k-1$, there is a notion of connectedness that is closely related to j-tight paths: two j-tuples J_1, J_2 of vertices are j-tuple-connected if there is a sequence of edges $e_1, \dots, e_\ell$ such that $J_1 \subset e_1$ and

$J_2 \subset e_\ell$, and furthermore any two consecutive edges e_i, e_{i+1} intersect in at least j vertices. A j-tuple component is a maximal collection of pairwise j-tuple-connected j-sets.

The threshold for the emergence of the giant j-tuple component in $H^k(n,p)$ is known to be

$$p_g = p_g(n;k,j) = \frac{1}{\left(\binom{k}{j}-1\right)\binom{n-j}{k-j}}.$$

The case $k=2$ and $j=1$ is the classical graph result of Erdős and Rényi. The case $j=1$ for general k was first proved by Schmidt-Pruzan and Shamir [20]. The case of general k and j was first proved by Cooley, Kang, and Person [6].

One might expect the threshold for the emergence of a j-tight path of linear length to have the same threshold. However, it turns out that this is only true in the case when $j=1$. More precisely, in the case $j=1$, the probability threshold of $\frac{1}{(k-1)\binom{n-j}{k-j}}$ given by Theorem 4 matches the threshold for the emergence of the giant (vertex-)component. However, for $j \geq 2$, the two thresholds do not match. A heuristic explanation for this is that when exploring a j-tuple component via a (breadth-first or depth-first) search process, each time we find an edge we may continue exploring a *j-tuple component* from any of the $\binom{k}{j}-1$ new j-sets within this edge (all are new except the j-set from which we first found the edge). However, when exploring a *j-tight path*, the restrictions on the structure mean that not all j-sets within the edge may form the last j vertices of the path. For a as defined in Theorem 4, it will turn out that we only have $\binom{k-j}{a}$ choices for the j-set from which to continue the path (this will be explained in more detail in Section 4.1).

1.4. **Paper overview.** The remainder of the paper is arranged as follows.

In Section 2, we will analyse the structure of j-tight paths and prove some preliminary results concerning the number of automorphisms, which will be needed later. We also collect some standard probabilistic results which we will use.

Subsequently, Section 3 will be devoted to a second moment calculation, which will be used to prove the lower bound on L in the subcritical case of Theorem 4. This is in essence a very standard method, although this particular application presents considerable technical challenges.

The second moment method breaks down when the paths become too long, and in particular it is too weak to prove the lower bounds in the supercritical case. Therefore the main contribution of this paper is an alternative strategy, inspired by previous proofs of phase transition results regarding the order of the giant component. These proofs, due to Krivelevich and Sudakov [14] as well as Cooley, Kang, and Person [6] and Cooley, Kang, and Koch [5], are based on an analysis of search processes which explore components.

We therefore introduce the `Pathfinder` algorithm, which is in essence a depth-first search process for paths, in Section 4. In Section 5, we observe some basic facts about the `Pathfinder` algorithm, which we subsequently use in Section 6 ($j=1$) and Section 7 ($j \geq 2$) to prove that whp the `Pathfinder` algorithm finds a j-tight path of the appropriate length, proving the lower bounds on L in the supercritical case of Theorem 4.

We collect together all of the previous results to complete the proof of Theorem 4 in Section 8. Finally in Section 9 we discuss some open problems, including possible strengthenings of Theorem 4.

2. Preliminaries

We first gather some notation and terminology which we will use throughout the paper.

Throughout the paper, k and j are fixed integers with $1 \leq j \leq k-1$. All asymptotics are with respect to n, and we use the standard Landau notations $o(\cdot), O(\cdot), \Theta(\cdot), \Omega(\cdot)$ with respect to these asymptotics. In particular, any value which is bounded by a function of k and j is $O(1)$. If S is a set and $m \in \mathbb{N}_0$, then $\binom{S}{m}$ denotes the set of m-element subsets of S. For $m, i \in \mathbb{N}$, we use $(m)_i := m(m-1)\dots(m-i+1)$ to denote the i-th falling factorial.

Recall that for $\ell \in \mathbb{N}$, a j-tight path of length ℓ in a k-uniform hypergraph contains ℓ edges and $(k-j)\ell + j$ vertices. Throughout the paper, whenever j, k, ℓ are clear from the context, we will denote by

$$v = v_{j,k}(\ell) := (k-j)\ell + j \tag{1}$$

the number of vertices in such a path. Furthermore, for the rest of the paper we fix a as in Theorem 4, i.e. a is the unique integer such that

$$1 \leq a \leq k-j \qquad \text{and} \qquad a \equiv k \pmod{k-j} \tag{2}$$

and we set

$$b := k - j - a. \tag{3}$$

Throughout the paper we ignore floors and ceilings whenever these do not significantly affect the argument. For the sake of clarity and readability, we omit many proofs of auxiliary results, particularly those that are applications of standard ideas or involve lengthy technical details. The interested reader may refer to arXiv:2003.14143 for the full details.

2.1. **Structure of j-tight paths.** For $\ell \in \mathbb{N}$, let $\mathcal{P}_\ell$ be the set of all j-tight paths of length ℓ in the complete k-uniform hypergraph on $[n]$, denoted by $K_n^{(k)}$. Thus $\mathcal{P}_\ell$ is the set of *potential* j-tight paths of length ℓ in $H^k(n,p)$.

It is important to observe that, depending on the values of k and j, the presence of one j-tight path $P \in \mathcal{P}_\ell$ in $H^k(n,p)$ may instantly imply the presence of many more with exactly the same edge set. In the graph case, there are only two paths with exactly the same edge set (we obtain the second by reversing the orientation), but for general k and j there may be more.

Let us demonstrate this with the following example for the case $k = 5$ and $j = 2$ (see Figure 2).

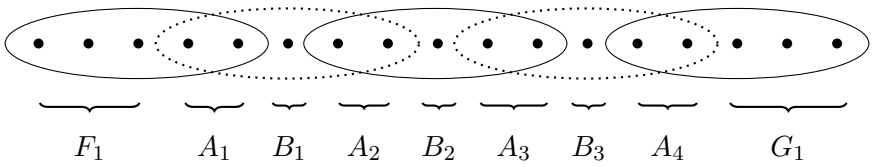

FIGURE 2. A 2-tight path of length 5 in a 5-uniform hypergraph, with a natural partition of vertices.

Observe that we have partitioned the vertices into sets $(F_1, A_1, \ldots)$ according to which edges they are in—each set of the partition is maximal with the property that every vertex in that set is in exactly the same edges of the j-tight path. Therefore we can re-order the vertices arbitrarily within any of these sets and obtain another j-tight path with the same edge set, and therefore also the same length. Similarly as for graphs, we can also reverse the orientation of the vertices (and also the edges) to obtain another j-tight path with the same edge set.

It will often be convenient to consider such paths as being the same, even though the order of vertices is different. Therefore we define an equivalence relation $\sim_\ell$ on $\mathcal{P}_\ell$ as follows. For any $A, B \in \mathcal{P}_\ell$, we say that $A \sim_\ell B$ if they have exactly the same edges.

We will be interested in the equivalence classes of this relation. Let $z_\ell = z_\ell(k,j)$ denote the size of each equivalence class of $\sim_\ell$ (note that, by symmetry, each equivalence class has the same size and so z_ℓ is well-defined). Further, let $\hat{\mathcal{P}}_\ell$ be the set of equivalence classes of $\sim_\ell$. Observe that if some $P \in \mathcal{P}_\ell$ is in $H^k(n,p)$, then so is every path in its equivalence class $\hat{P} \in \hat{\mathcal{P}}_\ell$. We abuse terminology slightly by saying that the equivalence class $\hat{P}$ lies in $H^k(n,p)$, and write $\hat{P} \subset H^k(n,p)$. We define $\hat{X}_\ell$ to be the number of equivalence classes for which this is the case. Then

$$\mathbb{E}(\hat{X}_\ell) = \sum_{\hat{P} \in \hat{\mathcal{P}}_\ell} \mathbb{P}\left(\hat{P} \subset H^k(n,p)\right) = |\hat{\mathcal{P}}_\ell| p^\ell = \frac{(n)_v}{z_\ell} p^\ell, \tag{4}$$

where $v = (k-j)\ell + j$ is the number of vertices in a j-tight path with ℓ edges (as defined in (1)).

We therefore need to estimate z_ℓ. To do so, we will analyse the structure of j-tight paths, inspired by the example in Figure 2. This analysis leads to the following lemma.

Lemma 6. *Let* $s = s(j,k) := \left\lceil \frac{k}{k-j} \right\rceil - 1$. *Then*

$$z_\ell = \begin{cases} \Theta(1) & \text{if } \ell \le s+1; \\ \frac{2}{b!}(a!b!)^{\ell-s}((k-j)!)^{2s} & \text{if } \ell \ge s+2. \end{cases}$$

In particular,

$$z_\ell = \Theta\left((a!b!)^\ell\right). \tag{5}$$

Proof. Let us first observe that if $\ell \le s+1$, then a j-tight path with ℓ edges has v vertices, where

$$v = (k-j)\ell + j \le k(\ell+1) \le k(s+2) = O(1),$$

and therefore $1 \le z_\ell \le v! = O(1)$, and the statement of the lemma follows for this case. We therefore assume that $\ell \ge s+2$.

We aim to determine the natural partition of the vertices of a j-tight path according to which edges they are in, as we did in the example in Figure 2.

Denote the edges of the j-tight path $P \in \mathcal{P}_\ell$ by $(e_1, \dots, e_\ell)$, in the natural order. Recall that $s = \lceil \frac{k}{k-j} \rceil - 1$, and observe that s is the largest integer such that $(k-j)s < k$, and therefore the largest integer such that $e_i \cap e_{i+s} \neq \emptyset$. We define

$$\begin{aligned} F_i &:= e_i \setminus e_{i+1} && \text{for } 1 \le i \le s; \\ G_i &:= e_{\ell-s+i} \setminus e_{\ell-s+i-1} && \text{for } 1 \le i \le s. \end{aligned}$$

We also define

$$\begin{aligned} A_i &:= e_i \cap e_{i+s} && \text{for } 1 \le i \le \ell - s, \\ B_i &:= e_{i+s} \setminus (e_{i+s+1} \cup e_i) && \text{for } 1 \le i \le \ell - s - 1. \end{aligned}$$

Observe that $A_i \cup B_i = e_{i+s} \setminus e_{i+s+1}$. Furthermore, since s is the largest integer such that e_{i+s+1} intersects e_{i+1}, we have that $(e_{i+s} \setminus e_{i+s+1}) \subset e_{i+1}$ and that $A_{i+1} \subseteq (e_{i+1} \setminus e_i)$, and therefore $A_{i+1} \cup B_i = e_{i+1} \setminus e_i$. Since we also have $A_i \cap B_i = A_{i+1} \cap B_i = \emptyset$, the vertices of the path P are now partitioned into parts

$$(F_1, \dots, F_s, A_1, B_1, A_2, B_2, \dots, A_{\ell-s-1}, B_{\ell-s-1}, A_{\ell-s}, G_1, \dots, G_s)$$

(in the natural order along P). Observe further that the parts are of maximal size such that the vertices within each part are in exactly the same edges. We refer to $\bigcup_{i=1}^s F_i = e_1 \setminus e_{s+1}$ as the *head* of the path P and to $\bigcup_{i=1}^s G_i = e_\ell \setminus e_{\ell-s}$ as the *tail*. Note that the vertices within each part can be rearranged to obtain a new j-tight path with exactly the same edges. We can also change the orientation of the path (i.e. reverse the order of the edges) to obtain a new path with the same edge set. (If $\ell = 0, 1$, this reorientation would already have been counted, but recall that we have assumed that $\ell \ge s+2$.) Thus we have

$$z_\ell = 2\left(\prod_{i=1}^s |F_i|!|G_i|!\right)\left(\prod_{i=1}^{\ell-s} |A_i|!\right)\left(\prod_{i=1}^{\ell-s-1} |B_i|!\right). \tag{6}$$

It therefore remains to determine the sizes of the F_i, G_i, A_i, B_i.

Claim 7.

$$\begin{aligned} |F_i| = |G_i| &= k-j && \text{for } 1 \le i \le s; \\ |A_i| &= a && \text{for } 1 \le i \le \ell - s; \\ |B_i| &= b && \text{for } 1 \le i \le \ell - s - 1. \end{aligned}$$

Substituting these values into (6), we obtain precisely the statement of Lemma 6. Thus the proof is complete up to verifying Claim 7. We omit the proof of this claim, which consists of an elementary checking of the definitions, but it can be found in Appendix A.1 of arXiv:2003.14143. □

Equation (4) and Lemma 6 together give the following immediate corollary.

Corollary 8.

$$\mathbb{E}(\hat{X}_\ell) = \Theta(1)\frac{(n)_v}{(a!b!)^\ell}p^\ell.$$

2.2. **Large deviation bounds.** In this section we collect some standard results which will be needed later.

We will use the following Chernoff bound, (see e.g. [12, Theorem 2.1]). We use $\mathrm{Bin}(N,p)$ to denote the binomial distribution with parameters $N \in \mathbb{N}$ and $p \in [0,1]$.

Lemma 9. *If $X \sim \mathrm{Bin}(N,p)$, then for any $\xi \geq 0$*

$$\mathbb{P}(X \geq Np + \xi) \leq \exp\left(-\frac{\xi^2}{2(Np + \frac{\xi}{3})}\right), \tag{7}$$

and

$$\mathbb{P}(X \leq Np - \xi) \leq \exp\left(-\frac{\xi^2}{2Np}\right).$$

It will often be more convenient to use the following one-sided form, which follows directly from Lemma 9. We omit the elementary proof, but it can be found in Appendix A.2 of arXiv:2003.14143.

Lemma 10. *Let $X \sim \mathrm{Bin}(N,p)$ and let $\alpha > 0$ be some arbitrarily small constant. Then with probability at least $1 - \exp(-\Theta(n^\alpha))$ we have $X \leq 2Np + n^\alpha$.*

3. Second moment method: lower bound

In this section we prove the lower bound in statement (i) of Theorem 4. The general basis of the argument is a completely standard second moment method— however, applying the method to this particular problem is rather tricky and so the argument is lengthy.

For technical reasons that become apparent during the proof, we need to handle the case when $2 \leq j = k-1$ slightly differently. We therefore distinguish two cases:

- Case 1: Either $j \leq k-2$ or $j = k-1 = 1$.
- Case 2: $2 \leq j = k-1$.

Correspondingly, we split the lower bound we aim to prove into two lemmas. In Case 1, we need to prove the following.

Lemma 11. *Let $k, j \in \mathbb{N}$ satisfy $1 \leq j \leq k-1$, and additionally either $j \leq k-2$ or $j = k-1 = 1$. Let $a \in \mathbb{N}$ be the unique integer satisfying $1 \leq a \leq k-j$ and $a \equiv k \bmod (k-j)$. Let $\varepsilon = \varepsilon(n) \ll 1$ satisfy $\varepsilon^3 n \xrightarrow{n\to\infty} \infty$ and let*

$$p = \frac{1-\varepsilon}{\binom{k-j}{a}\binom{n-j}{k-j}}.$$

Let L be the length of the longest j-tight path in $H^k(n,p)$. Then whp

$$L \geq \frac{j\ln n - \omega + 3\ln\varepsilon}{-\ln(1-\varepsilon)},$$

for any $\omega = \omega(n)$ such that $\omega \xrightarrow{n\to\infty} \infty$.

On the other hand, in Case 2 we have $k - j = 1$, and therefore the parameter a from Theorem 4 is simply 1. Thus also $\binom{k-j}{a} = 1$ and $p_0 = \frac{1}{n-k+1}$, and so the lower bound in Theorem 4 (i) simplifies to the following.

Lemma 12. *Let $k, j \in \mathbb{N}$ satisfy $2 \leq j = k-1$. Let $\varepsilon = \varepsilon(n) \ll 1$ satisfy $\varepsilon^3 n \xrightarrow{n\to\infty} \infty$ and let*

$$p = \frac{1-\varepsilon}{n-k+1}.$$

Let L be the length of the longest j-tight path in $H^k(n,p)$. Then whp

$$L \geq \frac{j\ln n - \omega + 3\ln\varepsilon}{-\ln(1-\varepsilon)},$$

for any $\omega = \omega(n)$ such that $\omega \xrightarrow{n\to\infty} \infty$.

Since the main ideas in the proofs of these two lemmas are essentially identical, we will treat only Case 1 (i.e. Lemma 11) here and omit the proof of Case 2 (i.e. Lemma 12). The interested reader can find the details in Appendix C of arXiv:2003.14143.

3.1. **Case 1: Either $j \le k-2$ or $j = k-1 = 1$.** We will prove Lemma 11 with the help of various auxiliary results. Since these results are rather technical in nature, we will omit their proofs in this paper, but these proofs may be found in Appendix B of arXiv:2003.14143.

Let us set $\ell = \frac{j \ln n - \omega + 3 \ln \varepsilon}{-\ln(1-\varepsilon)}$.

Recall that $\mathcal{P}_\ell$ is the set of all j-tight paths of length ℓ in $K_n^{(k)}$, and therefore

$$\mathbb{E}(X_\ell^2) = \sum_{A,B \in \mathcal{P}_\ell} \mathbb{P}(A, B \subset H^k(n,p)).$$

The probability term in the sum is fundamentally dependent on how many edges the paths A and B share, so we will need to calculate the number of pairs of possible paths with given intersections.

For any $A, B \in \mathcal{P}_\ell$, let $Q(A,B)$ be the set of common edges of A and B and define $q(A,B) := |Q(A,B)|$. Observe that there is a natural partition of $Q(A,B)$ into *intervals*, where each interval is a maximal set of edges in $Q(A,B)$ which are consecutive along both A and B. Let $r(A,B)$ be the number of intervals in this natural partition of $Q(A,B)$. Set $\mathbf{c}(A,B) := (c_1, \dots, c_r)$, where $c_1 \ge \cdots \ge c_r \ge 1$, to be the lengths (i.e. the number of edges) of these intervals. Given non-negative integers q, r and an r-tuple $\mathbf{c} = (c_1, \dots, c_r)$ such that $c_1 \ge \cdots \ge c_r \ge 1$ and $c_1 + \cdots + c_r = q$, define

$$\begin{aligned} \mathcal{P}_\ell^2(q, r, \mathbf{c}) := \{(A,B) \in \mathcal{P}_\ell^2 :\ & q(A,B) = q, \\ & r(A,B) = r, \\ & \mathbf{c}(A,B) = \mathbf{c}\}. \end{aligned}$$

For any $q, r, \mathbf{c}$ not satisfying these conditions, $\mathcal{P}_\ell^2(q,r,\mathbf{c})$ is empty. Recall from (1) that $v = (k-j)\ell + j$ is the number of vertices in a j-tight path of length ℓ.

Claim 13.

$$\mathbb{E}(X_\ell^2) \le ((n)_v)^2 \, p^{2\ell} + \sum_{q \ge 1} \sum_{r \ge 1} \sum_{\mathbf{c}} |\mathcal{P}_\ell^2(q,r,\mathbf{c})| p^{2\ell - q}. \tag{8}$$

Thus we need to estimate $|\mathcal{P}_\ell^2(q,r,\mathbf{c})|$ for $q, r \ge 1$. Given $q, r \ge 1$, we define the parameter

$$T(r) = T_q(r) := (k-j)q + j + (r-1)\min\{j, k-j\}.$$

This slightly arbitrary-looking expression is in fact a lower bound on the number of vertices in $Q(A,B)$. We obtain the following.

Proposition 14. *There exists a constant $C > 0$ such that for any $q \ge 1$ we have*

$$|\mathcal{P}_\ell^2(q,r,\mathbf{c})| \le ((n)_v)^2 \, \frac{(\ell - q + 1)^2 \ell^{2(r-1)} (a!b!)^q C^r}{(n-v)^{T(r)}}.$$

Proposition 14 together with (8) gives the following immediate corollary.

Corollary 15. *There exists a constant $C > 0$ such that*

$$\mathbb{E}(X_\ell^2) \le ((n)_v)^2 \, p^{2\ell} \left(1 + \sum_{q=1}^{\ell} \sum_{r=1}^{q} \sum_{\substack{c_1 + \cdots + c_r = q \\ c_1 \ge \cdots \ge c_r \ge 1}} \frac{(\ell - q + 1)^2 \ell^{2(r-1)} (a!b!)^q C^r}{p^q (n-v)^{T(r)}} \right). \tag{9}$$

We bound the triple-sum using the following two results.

Proposition 16.

$$\sum_{r=1}^{q} \sum_{\substack{c_1 + \cdots + c_r = q \\ c_1 \ge \cdots \ge c_r \ge 1}} \frac{(\ell - q + 1)^2 \ell^{2(r-1)} (a!b!)^q C^r}{p^q (n-v)^{T(r)}} = O\left(n^{-j}\right) \frac{(\ell - q + 1)^2}{(1-\varepsilon)^q}. \tag{10}$$

Claim 17.

$$\sum_{q=1}^{\ell} \frac{(\ell - q + 1)^2}{(1-\varepsilon)^q} = \frac{2(1-\varepsilon)^{-\ell}}{\varepsilon^3}. \tag{11}$$

Substituting (10) and (11) into (9), using the fact that $\ell = \frac{j\ln n - \omega + 3\ln\varepsilon}{-\ln(1-\varepsilon)}$ and performing some elementary approximations leads to the following.

Claim 18. $\mathbb{E}(X_\ell^2) = ((n)_v)^2\, p^{2\ell}(1+o(1))$.

We can now use these auxiliary results to prove our lower bound.

Proof of Lemma 11. Recalling that $\mathcal{P}_\ell$ is the set of all possible j-tight paths of length ℓ in $H^k(n,p)$, clearly $\mathbb{E}(X_\ell) = |\mathcal{P}_\ell|p^\ell = (n)_v p^\ell$. Therefore by Claim 18, we have

$$\mathbb{E}(X_\ell^2) = \mathbb{E}(X_\ell)^2(1+o(1)),$$

and a standard application of Chebyshev's inequality shows that whp $X_\ell \geq 1$, i.e. whp

$$L(G(n,p)) \geq \ell = \frac{j\ln n - \omega + 3\ln\varepsilon}{-\ln(1-\varepsilon)}$$

as claimed. □

It would be tempting to try to generalise this proof to also prove a lower bound in the supercritical case. However, this strategy fails because as the paths A and B become longer, there are many more ways in which they can intersect each other, and therefore the terms which, in the subcritical case, were negligible lower order terms (i.e. $q \geq 1$) become more significant. We will therefore use an entirely different strategy for the supercritical case.

4. The `Pathfinder` algorithm

The proof strategy for the lower bound in the supercritical case is to define a depth-first search algorithm, which we call `Pathfinder` and which discovers j-tight paths in a k-uniform hypergraph, and to show that whp this algorithm, when applied to $H^k(n,p)$, will find a path of the appropriate length.

4.1. Hypergraph exploration using DFS.

In this section, we will describe the `Pathfinder` algorithm to find j-tight paths in k-uniform hypergraphs in full generality. The reader who is not completely familiar with j-tight paths in hypergraphs, or who prefers a more gentle introduction, may refer to Section 4.1 in arXiv:2003.14143, where the algorithm is first described in the special case when $k=3$ and $j=2$. We will use the following notation: if $\mathcal{F}$ is a family of sets and X is a set, we write $\mathcal{F}+X$ and $\mathcal{F}-X$ to mean $\mathcal{F}\cup\{X\}$ and $\mathcal{F}\setminus\{X\}$ respectively.

Recall from (2) that $a \in [k-j]$ is such that $a \equiv k \bmod (k-j)$, and from the statement of Lemma 6 that $s = \lceil \frac{k}{k-j} \rceil - 1 = \lceil \frac{j}{k-j} \rceil$. Let us define $r := s-1 = \lceil \frac{j}{k-j} \rceil - 1$, so that $j = a + (k-j)r$.

Definition 19. Given a set J of j vertices, an *extendable partition of* J is an ordered partition $(C_0, C_1, \dots, C_r)$ of J such that $|C_0| = a$ and $|C_i| = k-j$ for all $i \in [r]$.

Note that if we have constructed a reasonably long j-tight path (i.e. of length at least s), the final j-vertices naturally come with an extendable partition $(C_0, C_1, \dots, C_r)$ according to which edges of the path they lie in, similar to the partition of all vertices of the path described in Section 2.1. The vertices within each part of the extendable partition could be re-ordered arbitrarily to obtain a new path with the same edge set. Therefore if we find a further edge from the final j-set to extend the path, there is more than one possibility for the final j-set of the extended path—it must contain $C_2, \dots, C_r$ and a further a vertices from C_1, which may be chosen arbitrarily. Thus an extendable partition provides a convenient way to describe the j-sets from which we might further extend the path.

Although for paths of length shorter than s the j-sets come only with a coarser (and therefore less restrictive) partition, it is convenient for a unified description of the algorithm for them to be given an extendable partition. In particular, we will start our search process from a j-set which we artificially endow with an extendable partition; this additional restriction is permissible for a lower bound on the longest path length.

We begin by giving an informal overview of the algorithm— the formal description follows.

At any given point, the algorithm will maintain a j-tight path P and a partition of the j-sets of $V(H)$ into *neutral*, *active* or *explored* sets. Initially, P is empty and every j-set is neutral. During

Algorithm: `Pathfinder`

Input: Integers k, j such that $1 \le j \le k-1$.
Input: H, a k-uniform hypergraph.
Let $a \in [k-j]$ be such that $a \equiv k \bmod (k-j)$
Let $r = \lceil \frac{j}{k-j} \rceil - 1$
For $i \in \{j, k\}$, let σ_i be a permutation of the i-sets of $V(H)$, chosen uniformly at random
$N \leftarrow \binom{V(H)}{j}$ // neutral j-sets
$A, E \leftarrow \emptyset$ // active, explored j-sets
$P \leftarrow \emptyset$ // current j-tight path
$\ell \leftarrow 0$ // index tracking the current length of P
$t \leftarrow 0$ // "time", number of queries made so far
while $N \neq \emptyset$ **do**
  Let J be the smallest j-set in N, according to σ_j // "new start"
  Choose an arbitrary extendable partition $\mathcal{P}_J$ of J
  $B_0 = \{J\}$
  $A \leftarrow \{J\}$
  while $A \neq \emptyset$ **do**
    Let J be the last j-set in A
    Let $\mathcal{K}$ be the set of k-sets $K \subset V(H)$ such that $K \supset J$, K was not queried from J before, $K \setminus J$ is vertex-disjoint from P, and K does not contain any $J' \in E$
    if $\mathcal{K} \neq \emptyset$ **then**
      Let K be the first k-set in $\mathcal{K}$ according to σ_k
      $t \leftarrow t+1$ // a new query is made
      if $K \in H$ **then** // "query K"
        $e_\ell \leftarrow K$
        $P \leftarrow P + e_\ell$ // P is extended by adding $K = e_\ell$
        $\ell \leftarrow \ell + 1$ // length of P increases by one
        Let $\mathcal{P}_J = (C_0, C_1, \dots, C_r)$ be the extendable partition of J
        for *each* $Z \in \binom{C_1}{a}$ **do**
          $J_Z \leftarrow Z \cup C_2 \cup \cdots \cup C_r \cup (K \setminus J)$ // j-set to be added
          $\mathcal{P}_{J_Z} \leftarrow (Z, C_2, \dots, C_r, K \setminus J)$ // extendable partition
          $i(J_Z) \leftarrow \ell$
          $A \leftarrow A + J_Z$ // j-set becomes active
        $\mathcal{B}_\ell \leftarrow \{J_Z : Z \in \binom{C_1}{a}\}$
      $(A_t, E_t, P_t) \leftarrow (A, E, P)$ // update "snapshot" at time t
    else if $\mathcal{K} = \emptyset$ **then** // all extensions from J were queried
      $A \leftarrow A - J$ // J becomes explored
      $E \leftarrow E + J$
      if $\mathcal{B}_\ell \subset E$ **then** // the current batch is fully explored
        $\mathcal{B}_\ell \leftarrow \emptyset$ // empty this batch
        $P \leftarrow P - e_\ell$ // last edge of P is removed
        $\ell \leftarrow \ell - 1$ // length of P decreases by one

the algorithm every j-set can change its status from neutral to active and from active to explored. The j-sets which are active or explored will be referred to as *discovered*.

The edges of P will be $e_1, \dots, e_\ell$ (in this order), and every active j-set will be contained inside some edge of P. Whenever a new edge $e_{\ell+1}$ is added to the end of P, a batch $\mathcal{B}_{\ell+1}$ of neutral j-sets within that edge will become active: these are the j-sets from which we could potentially extend the current path. A j-set J becomes explored once all possibilities to extend P from J have been queried. Once all of the j-sets in the batch $\mathcal{B}_\ell$ corresponding to e_ℓ have been declared explored, e_ℓ will be removed from P.

The active sets will be stored in a "stack" structure (last in, first out). Each active j-set J will have an associated extendable partition $\mathcal{P}_J$ of J, and an index $i(J) \in \{0, \dots, \ell\}$, where ℓ is the current length of P. The extendable partition will keep track of the ways in which we can

extend P from J in a consistent manner, as described in Section 4.1. The index $i(J)$ will indicate that J belongs to the batch $\mathcal{B}_{i(J)}$ which was added when the edge $e_{i(J)}$ was added to P. Thus the algorithm will maintain a collection of batches $\mathcal{B}_0, \dots, \mathcal{B}_\ell$, all of which consist of discovered j-sets which are inside $V(P)$. It will hold that $|\mathcal{B}_0| = 1$ and $|\mathcal{B}_i| = \binom{k-j}{a}$ for all $i \geq 1$, and all the batches will be disjoint.

All the j-sets from a single batch will change their status from neutral to active in a single step, and they will be added to the stack according to some fixed order which is chosen uniformly at random during the initialisation of the algorithm.

An iteration of the algorithm can be described as follows. Suppose J is the last active j-set in the stack. We will query k-sets K, to check whether K is an edge in H or not. We only query a k-set K subject to the following conditions:

(Q1) K contains J;
(Q2) $K \setminus J$ is disjoint from the current path P;
(Q3) K was not queried from J before;
(Q4) K does not contain any explored j-set.

Condition (Q1) ensures that we only query k-sets with which we might sensibly continue the path in a j-tight manner. Condition (Q2) ensures that we do not re-use vertices that are already in P. Together, these two conditions guarantee that P will indeed always be a j-tight path. Moreover, Condition (Q3) ensures that we never query a k-set more than once from the same j-set, thus guaranteeing that the algorithm does not get stuck in an infinite loop. Finally Condition (Q4) ensures that we never query a k-set a second time from a *different* j-set (note that the possibility that K could have been queried from another *active* j-set is already excluded by Condition (Q2), since such an active j-set would lie within P). Note that Condition (Q4) is not actually necessary for the correctness of the algorithm, but it does ensure independence of queries and is therefore necessary for our *analysis* of the algorithm.

If no such k-set K can be found in the graph H, then we declare J explored and move on to the previous active j-set in the stack. Moreover, if at this point all of the j-sets in the batch $\mathcal{B}_{i(J)}$ of J have been declared explored, the last edge e_ℓ of the current path is removed and ℓ is replaced by $\ell - 1$. If the set of active j-sets is now empty, we choose a new j-set J from which to start, declare J active and choose an extendable partition of J.

On the other hand, if we can find a suitable set K for J, we query K, and if it forms an edge, then according to the extendable partition of J, the set K will yield a new batch of j-sets (which previously were neutral and now become active). More precisely, if the extendable partition of J is $(C_0, C_1, \dots, C_r)$, then the batch consists of all j-sets which contain $K \setminus J$ and $C_2, \dots, C_r$, as well as a vertices of C_1. Thus the batch consists of $\binom{k-j}{a}$ many j-sets.

Finally, we keep track of a "time" parameter t, which counts the number of queries the algorithm has made. Initially, $t = 0$ and t increases by one each time we query a k-set.

During the analysis we will make reference to certain objects or families which are implicit in the algorithm at each time t even if the algorithm does not formally track them. These include the sets of neutral, active and discovered j-sets N_t, A_t, E_t and the current path P_t, which are simply the sets N, A, E and the path P at time t. We say that (A_t, E_t, P_t) is the *snapshot of H at time t*. We also refer to certain families of j-sets, including D_t (the discovered j-sets), R_t (the "new starts") and S_t (the "standard j-sets"), as well as families $F_t^{(1)}, F_t^{(2)}, F_t$ of $(k-j)$-sets (the "forbidden subsets"). The precise definitions of all of these families will be given when they become relevant.

4.2. **Proof strategy.** Our aim is to analyse the `Pathfinder` algorithm and show that whp it finds a path of length at least $\frac{(1-\delta)\varepsilon n}{(k-j)^2}$, or at least $\frac{(1-\delta)\varepsilon^2 n}{4(k-j)^2}$ if $j = 1$. The overall strategy can be described rather simply: suppose that by some time t, which is reasonably large, we have not discovered a path of the appropriate length. Then whp (and disregarding some small error terms), the following holds:

(A) We have discovered at least $pt\binom{k-j}{a}$ many j-sets;
(B) Very few j-sets are active, therefore at least $pt\binom{k-j}{a}$ are explored;

(C) From each explored j-set, we queried at least $\binom{n'}{k-j}$ many k-sets, where $n' = \left(1 - \frac{(1-\delta)\varepsilon}{k-j}\right) n$.

(D) Thus the number of queries made is at least

$$pt\binom{k-j}{a}\binom{n'}{k-j} = t\frac{(1+\varepsilon)}{\binom{n}{k-j}}\binom{\left(1-\frac{(1-\delta)\varepsilon}{k-j}\right)n}{k-j} \approx t(1+\varepsilon)(1-(1-\delta)\varepsilon) > t.$$

This yields a contradiction since the number of queries made is exactly t by definition. The proof consists of making these four steps more precise, and the main difficulty is to prove Step (C).

5. Basic properties of the algorithm

Before analysing the likely evolution of the `Pathfinder` algorithm, we first collect some basic properties which will be useful later.

Note that there are two ways in which a j-set J can be discovered up to time t. First, it could have been included as a *new start* when the set of active j-sets was empty and we chose a j-set J from which to start exploring a new path (Line 10). Second, J could have been declared active if it was part of a batch of j-sets activated when we discovered an edge, which we refer to as a *standard activation* (Lines 20-30), and we refer to the j-sets which were discovered in this way as *standard* j-sets.

For any $t \geq 0$, let $\ell_t := |E(P_t)|$ be the length (i.e. number of edges) of the path found by the algorithm at time t.

Proposition 20. *At any time t, the number $|A_t|$ of active j-sets is at most*

$$|A_t| \leq 1 + \binom{k-j}{a}\ell_t. \tag{12}$$

Proof. Recall that by construction, every active j-set in A_t is contained in some edge of P_t. Moreover, every time an edge is added to the current path, exactly $\binom{k-j}{a}$ many j-sets are added via a standard activation. There is also exactly one further active j-set which was added as a new start, which gives the desired inequality. □

Note that equality does *not* necessarily hold, because some j-sets which once were active may already be explored.

For every t, let R_t be the set of all discovered j-sets at time t which were new starts, and let S_t be the discovered j-sets up to time t which are standard. Thus, for all t,

$$R_t \cup S_t = A_t \cup E_t.$$

Note that if the query at time t is answered positively, then $|S_t| = |S_{t-1}| + \binom{k-j}{a}$, and otherwise $|S_t| = |S_{t-1}|$. Thus, if $X_1, X_2, \dots$ are the indicator variables that track which queries are answered positively, i.e. X_i is 1 if the i-th k-tuple queried forms an edge and 0 otherwise, then we have

$$|S_t| = \binom{k-j}{a}\sum_{i=1}^{t} X_i. \tag{13}$$

Note that with input hypergraph $H = H^k(n,p)$, the $X_1, X_2, \dots$ are simply i.i.d. Bernoulli random variables with probability p. In particular, using Chernoff bounds, we can approximate $|S_t|$ when t is large. We omit the proof, which is an elementary application of Lemma 9 – the details can be found in Appendix A.4 of arXiv:2003.14143.

Proposition 21. *Let $p = \frac{1+\varepsilon}{\binom{k-j}{a}\binom{n-j}{k-j}}$, let $t = t(n) \in \mathbb{N}$, and let $0 \leq \gamma = \gamma(n) = O(1)$. Then when* `Pathfinder` *is run with input k, j and $H = H^k(n,p)$, with probability at least $1 - \exp(-\Theta(\gamma^2 pt))$ we have*

$$(1-\gamma)\frac{(1+\varepsilon)t}{\binom{n-j}{k-j}} \leq |S_t| \leq (1+\gamma)\frac{(1+\varepsilon)t}{\binom{n-j}{k-j}}.$$

In particular, if $\gamma^2 t n^{-(k-j)} \to \infty$, then these inequalities hold whp.

Note that this proposition gives a lower bound on the number of discovered j-sets, but it does not immediately give an upper bound, since it says nothing about the number of new starts that have been made. (Later the number of new starts will be bounded by Proposition 33 in the case $j \geq 2$, ; we will not need such an upper bound in the case $j = 1$.)

How many queries are made from a given j-set J before it is declared explored? Clearly $\binom{n-j}{k-j}$ is an upper bound, since this is the number of k-sets that contain J, but some of these are excluded in the algorithm, and we will need a lower bound. In what follows, for convenience we slightly abuse terminology by referring to querying not a k-set $K \supset J$, but rather the $(k-j)$-set $K \setminus J$. (If J is already determined, this is clearly equivalent.)

There are two reasons why a $(k-j)$-set disjoint from the current j-set J may never be queried—either it contains a vertex of the current path, or it contains an explored j-set.

Definition 22. Consider an exploration of a k-uniform hypergraph H using `Pathfinder`. Given t, let J be the last active set in the stack of A_t. We call a $(k-j)$-set $X \subset V(H) \setminus J$ *forbidden at time* t, if

(1) $X \cap V(P_t) \neq \emptyset$, or
(2) there exists an explored j-set $J' \in E_t$ such that $J' \subset (J \cup X)$.

If X satisfies (1) we say X is a *forbidden set of type 1*; if it satisfies (2) we say it is a *forbidden set of type 2*. Let $F^{(1)} = F_t^{(1)}$ and $F^{(2)} = F_t^{(2)}$ denote the corresponding sets of forbidden $(k-j)$-sets at time t, and let $F = F_t := F_t^{(1)} \cup F_t^{(2)}$ be the set of all forbidden $(k-j)$-sets at time t.

Observe that a $(k-j)$-set might be a forbidden set of both types, i.e. may lie in both $F^{(1)}$ and $F^{(2)}$. The following consequence of the definition of forbidden $(k-j)$-sets is crucial: if J is declared explored at time t and a $(k-j)$-set X disjoint from J is not in F_t, then X was queried from J by the algorithm (at some time $t' \leq t$). Thus, if the number of forbidden sets at time t is "small", then a "large" number of queries were required to declare J explored.

Our aim is to bound the size of $F_t = F_t^{(1)} \cup F_t^{(2)}$. If the `Pathfinder` algorithm has not found a long path, then $F_t^{(1)}$ is small. More precisely, we obtain the following bound.

Proposition 23. *For all times $t \geq 0$,*

$$|F_t^{(1)}| \leq \ell_t \cdot (k-j)\binom{n-j-1}{k-j-1}.$$

Proof. Let J be the current active j-set in A_t. A $(k-j)$-set X is in $F_t^{(1)}$ if and only if $X \cap J = \emptyset$ and $X \cap V(P_t) \neq \emptyset$; thus $|F_t^{(1)}| \leq |V(P_t) \setminus J|\binom{n-j-1}{k-j-1}$. Since $J \subset V(P_t)$ and P_t has ℓ_t edges, we have $|V(P_t) \setminus J| = \ell_t \cdot (k-j)$, and the desired bound follows. □

It remains to estimate the number of forbidden sets of type 2. To achieve this, in the next section we will give more precise estimates on the evolution of the algorithm run with input $H^k(n,p)$ (and in particular the evolution of discovered j-sets, which certainly includes all explored j-sets).

We will need to treat the case $j = 1$ separately from the case $j \geq 2$. We begin with the case $j = 1$, since this is significantly easier but introduces some of the ideas that will be used in the more complex case $j \geq 2$.

6. Algorithm analysis: loose case ($j = 1$)

Our aim in this section is to prove the lower bound in Theorem 4 (iii). For convenience, we restate the result we are aiming to prove as a lemma.

Lemma 24. *Let $k \in \mathbb{N}$ and let $\varepsilon = \varepsilon(n)$ satisfy $\varepsilon^3 n \xrightarrow{n\to\infty} \infty$. Let*

$$p = (1+\varepsilon)p_0 = \frac{1+\varepsilon}{(k-1)\binom{n-1}{k-1}}$$

and let L be the length of the longest loose path in $H^k(n,p)$. Then for all $\delta \gg \varepsilon$ satisfying $\delta^2\varepsilon^3 n \xrightarrow{n\to\infty} \infty$, whp

$$L \geq (1-\delta)\frac{\varepsilon^2 n}{4(k-1)^2}.$$

We define

$$\ell_0 := \frac{(1-\delta)\varepsilon^2 n}{4(k-1)^2},$$

so our goal is to show that whp the `Pathfinder` algorithm discovers a path of length at least ℓ_0. We also define

$$T_0 := \frac{\varepsilon n\binom{n-1}{k-1}}{2(k-1)} = \frac{\varepsilon n}{2(k-1)^2 p_0}.$$

We will show that whp at some time $t \leq T_0$, we have $\ell_t \geq \ell_0$, as required. We begin with the following proposition, which is a simple application of Proposition 21. The proof can be found in Appendix A.4 of arXiv:2003.14143.

Proposition 25. *At time $t = T_0$, whp we have*

$$|A_t \cup E_t| \geq (1-o(\delta\varepsilon))(k-1)pt.$$

Let T_1 denote the first time t at which

$$|A_t \cup E_t| = \left(1-\frac{\delta\varepsilon}{3}\right)(k-1)pT_0 = \left(1-\frac{\delta\varepsilon}{3}\right)(1+\varepsilon)\frac{\varepsilon n}{2(k-1)} \tag{14}$$

(recall that we ignore floors and ceilings). Then from Proposition 25, we immediately obtain the following.

Corollary 26. *Whp $T_1 \leq T_0$.*

We claim furthermore that this inequality implies that we must have a long loose path.

Proposition 27. *If $T_1 \leq T_0$, then at time $t = T_1$ we have $\ell_t \geq \ell_0$.*

Proof. Suppose for a contradiction that $T_1 \leq T_0$, but that at time $t = T_1$ we have $\ell_t < \ell_0$. Then by (14) and (12)

$$\begin{aligned}
|E_t| &= |A_t \cup E_t| - |A_t| \\
&\geq \left(1-\frac{\delta\varepsilon}{3}\right)(k-1)pT_0 - ((k-1)\ell_0 + 1) \\
&= \left(1+\varepsilon-\frac{\delta\varepsilon}{3} - O(\delta\varepsilon^2)\right)(k-1)p_0T_0 - \frac{(1-\delta)\varepsilon^2 n}{4(k-1)} - o(\delta\varepsilon^2 n) \\
&= \left(1+\varepsilon-\frac{\delta\varepsilon}{3} - \frac{(1-\delta)\varepsilon}{2} - O(\delta\varepsilon^2) - o(\delta\varepsilon)\right)(k-1)p_0T_0 \\
&\geq \left(1+\frac{\varepsilon}{2}+\frac{\delta\varepsilon}{7}\right)(k-1)p_0T_0,
\end{aligned}$$

where we have used the fact that $(k-1)p_0T_0 = \frac{\varepsilon n}{2(k-1)} = \Theta(\varepsilon n)$, and that $\delta\varepsilon^2 n \geq \varepsilon^3 n \to \infty$.

On the other hand, N_t, the set of neutral vertices, satisfies

$$\begin{aligned}
|N_t| = n - |A_t \cup E_t| &\overset{(14)}{=} n - (1-o(\delta\varepsilon))(1+\varepsilon)\frac{\varepsilon n}{2(k-1)} \\
&= \left(1 - \frac{\varepsilon}{2(k-1)} + o(\delta\varepsilon) + O(\varepsilon^2)\right) n.
\end{aligned}$$

Note that no vertex of N_t can possibly have been forbidden at any time $t' \leq t$. This implies, since the vertices of E_t are fully explored, that from each explored vertex we certainly queried any k-set

containing the vertex and $k-1$ vertices of N_t. Thus the number of queries t that we have made so far certainly satisfies

$$\begin{aligned} t &\geq |E_t|\binom{|N_t|}{k-1} \\ &\geq \left(1+\frac{\varepsilon}{2}+\frac{\delta\varepsilon}{7}\right)(k-1)p_0T_0 \\ &\qquad\qquad \cdot\left(1+O\left(\frac{1}{n}\right)\right)\frac{\left(1-\frac{\varepsilon}{2(k-1)}+o(\delta\varepsilon)+O(\varepsilon^2)\right)^{k-1}n^{k-1}}{(k-1)!} \\ &= \left(1+\frac{\varepsilon}{2}+\frac{\delta\varepsilon}{7}\right)T_0\cdot\left(1+O\left(\frac{1}{n}\right)\right)\left(1-\frac{\varepsilon}{2}+o(\delta\varepsilon)+O(\varepsilon^2)\right) \\ &= \left(1+\frac{\delta\varepsilon}{7}+o(\delta\varepsilon)+O(\varepsilon^2)\right)T_0 \\ &> T_0, \end{aligned}$$

which gives the required contradiction since we assumed that $t = T_1 \leq T_0$. □

Proof of Lemma 24. The statement of Lemma 24 follows directly from Corollary 26 and Proposition 27. □

Let us note that although we proved that whp $\ell_t \geq \ell_0$ at *some* time $t \leq T_0$, with a small amount of extra work we could actually prove that this even holds at exactly $t = T_0$: we would need a corresponding upper bound in Proposition 25, which follows from a Chernoff bound on the number of edges discovered so far and an upper bound on the number of new starts we have made by time T_0.

7. Algorithm analysis: high-order case ($j \geq 2$)

In the case $j \geq 2$, we will use the `Pathfinder` algorithm to study j-tight paths in $H^k(n,p)$ by running the algorithm up to a certain *stopping time* T_{stop}, i.e. until we have made T_{stop} queries. In order to define T_{stop}, we need some additional definitions.

Given some time $t \geq 0$ let D_t denote the set of all j-sets which are discovered by time t. With a slight abuse of notation, we will sometimes also use D_t to denote the j-uniform hypergraph on vertex set $[n]$ with edge set D_t. Note that a j-set J lies in D_t if and only if there exists $t' \leq t$ such that $J \in A_{t'}$, or in other words, every j-set which is discovered at time t was active at some time $t' \leq t$. Also, note that for every $t_1 \leq t_2$, $D_{t_1} \subseteq D_{t_2}$, i.e. the sequence of discovered j-sets is always increasing (although the sequence of active sets A_t is not).

Suppose that $0 \leq i \leq j$ and that I is an i-set. Then define $d(I) = d_t(I) = \deg_{D_t}(I)$ to be the number of j-sets of D_t that contain I.

Definition 28. Let $\varepsilon \ll \delta \leq 1$ be as in Theorem 4(ii),[3] and recall that $|R_t|$ is the number of new starts made by time t. Let

$$C_{k,j,j-1} \gg C_{k,j,j-2} \gg \cdots \geq C_{k,j,0} \gg 1$$

be some sufficiently large constants and let $0 < \beta \ll 1$ be a sufficiently small constant. Define

$$T_0 := \frac{n^{k-j+1}}{\varepsilon}.$$

We define T_{stop} to be the smallest time t such that one of the following stopping conditions hold:

- **(S1)** `Pathfinder` found a path of length at least $(1-\delta)\frac{\varepsilon n}{(k-j)^2}$;
- **(S2)** $t = T_0$;
- **(S3)** $|R_t| \geq 2(k-j)!\sqrt{\frac{tn^\beta}{n^{k-j}}} + \frac{n^\beta}{2}$;
- **(S4)** There exists some $0 \leq i \leq j-1$ and an i-set I with $d_t(I) \geq \frac{C_{k,j,i}t}{n^{k-j+i}} + n^\beta$.

[3] Recall from Remark 5 that we will not actually use the additional condition $\delta \gg \frac{\ln n}{\varepsilon^2 n}$ for the proof of the lower bound, c.f. Lemma 30.

We first observe that T_{stop} is well-defined.

Claim 29. *If* `Pathfinder` *is run on inputs k, j and any k-uniform hypergraph H on $[n]$, then one of the four stopping conditions is always applied.*

Proof. If none of the stopping conditions is applied, the algorithm will continue until all j-sets are explored (since a new start is always possible from any neutral j-set). If this occurs at time $t \geq T_0$, then **(S2)** would already have been applied (if none of the other stopping conditions were applied first). On the other hand, if this occurs at time $t \leq T_0$, then **(S4)** is certainly satisfied with $i = 0$ and $I = \emptyset$. □

We will often use the fact that for $t \leq T_{\text{stop}}$, the (non-strict) inequalites in stopping conditions **(S1)**, **(S3)** and **(S4)** are reversed. For example, for $t \leq T_{\text{stop}}$ we have $|R_t| \leq 2(k-j)!\sqrt{\frac{tn^\beta}{n^{k-j}}} + n^\beta$. This is because

$$|R_t| \leq |R_{t-1}| + 1 < 2(k-j)!\sqrt{\frac{(t-1)n^\beta}{n^{k-j}}} + n^\beta + 1,$$

where the second inequality holds because we did not apply **(S3)** by time $t-1$ (and recall that we ignore floors and ceilings). In such a situation, we will slightly abuse terminology by saying that "by **(S3)**" we have $|R_t| \leq 2(k-j)!\sqrt{\frac{tn^\beta}{n^{k-j}}} + n^\beta$.

Our main goal is to show that whp it is **(S1)** which is applied first, i.e. the algorithm has indeed discovered a path of the appropriate length.

Lemma 30. *Let $k, j \in \mathbb{N}$ satisfy $2 \leq j \leq k-1$. Let $a \in \mathbb{N}$ be the unique integer satisfying $1 \leq a \leq k-j$ and $a \equiv k \bmod (k-j)$. Let $\varepsilon = \varepsilon(n) \ll 1$ satisfy $\varepsilon^3 n \xrightarrow{n\to\infty} \infty$ and let*

$$p_0 = p_0(n; k, j) := \frac{1}{\binom{k-j}{a}\binom{n-j}{k-j}}.$$

Let L be the length of the longest j-tight path in $H^k(n,p)$, and let $\delta \gg \varepsilon$.

Suppose `Pathfinder` *is run with input k, j and $H = H^k(n,p)$. Then whp **(S1)** is applied. In particular, whp*

$$L \geq \ell_{T_{\text{stop}}} = (1-\delta)\frac{\varepsilon n}{(k-j)^2}.$$

For the rest of this section, we will assume that all parameters are as defined in Lemma 30.

We first prove an auxiliary lemma which gives an upper bound on the number of forbidden $(k-j)$-sets up to time T_{stop}. Recall that $F_t^{(1)}$ and $F_t^{(2)}$ denote the sets of forbidden $(k-j)$-sets at time t of types 1 and 2, respectively. Let $f^{(i)} = f_t^{(i)} := |F_t^{(i)}|$ for $i = 1, 2$.

Lemma 31. *Let $t \leq T_{\text{stop}}$. Then*

$$f^{(1)} + f^{(2)} \leq (1-\delta/2)\varepsilon\binom{n-j}{k-j}.$$

In particular, from every explored j-set we made at least

$$(1-\varepsilon+\delta\varepsilon/2)\binom{n-j}{k-j}$$

queries.

Proof. Due to condition **(S1)**, the length ℓ_t of the path P_t at any time t is at most $\frac{(1-\delta)\varepsilon n}{(k-j)^2}$. Thus by Proposition 23 we have that

$$f^{(1)} \leq \frac{(1-\delta)\varepsilon n}{(k-j)} \cdot \binom{n-j-1}{k-j-1} \leq \left(1 - \frac{2\delta}{3}\right)\varepsilon\binom{n-j}{k-j}. \tag{15}$$

By condition **(S2)**, we have $T_{\text{stop}} \le \frac{n^{k-j+1}}{\varepsilon}$. Furthermore, by condition **(S4)**, for any $0 \le i \le j-1$ and any i-set I we have

$$d_t(I) \le d_{T_{\text{stop}}}(I) \le \frac{C_{k,j,i}T_{\text{stop}}}{n^{k-j+i}} + n^{\beta} \le \frac{C_{k,j,i}}{\varepsilon n^{i-1}} + n^{\beta}.$$

Observe that if J is the current j-set, any forbidden $(k-j)$-set of type 2 can be identified by:

- choosing an integer $i = 0, \dots, j-1$;
- choosing a proper subset $I \subset J$ of size i (there are $\binom{j}{i}$ possibilities);
- choosing an explored (and therefore discovered) j-set $J' \supset I$ such that $(J' \setminus I) \cap J = \emptyset$, (at most $d_t(I)$ possibilities);
- choosing a k-set K containing both J and J' (there are $\binom{n-2j+i}{k-2j+i}$ possibilities).

Then the forbidden $(k-j)$-set is $K \setminus J$. Note that if $j > k/2$, then $k-2j+i$ may be negative for some values of i. In this case we interpret $\binom{n-2j+i}{k-2j+i}$ to be zero.

Therefore we obtain

$$\begin{aligned} f^{(2)} &\le \sum_{i=0}^{j-1} \binom{j}{i} \cdot \left(\max_{|I|=i} d_t(I) \right) \cdot \binom{n-2j+i}{k-2j+i} \\ &\le \sum_{i=0}^{j-1} 2^j \cdot \left(\frac{C_{k,j,i}}{\varepsilon n^{i-1}} + n^{\beta} \right) \cdot O(n^{-j+i}) \binom{n-j}{k-j} \\ &= O\left(\frac{1}{\delta \varepsilon^2 n^{j-1}} + \frac{n^{\beta}}{\delta \varepsilon n} \right) \delta\varepsilon \binom{n-j}{k-j}. \end{aligned}$$

Now recall that $\delta \gg \varepsilon$ and that we are considering the case $j \ge 2$, which means that $\delta\varepsilon^2 n^{j-1} \ge \varepsilon^3 n \to \infty$. Furthermore $\beta \ll 1$, which implies that $\delta\varepsilon n^{1-\beta} \ge \varepsilon^2 n^{2/3} \to \infty$, so we obtain

$$f^{(2)} = o(1)\delta\varepsilon \binom{n-j}{k-j}.$$

Together with (15), this leads to

$$f^{(1)} + f^{(2)} \le \left(1 - \frac{2\delta}{3} + o(\delta)\right) \varepsilon \binom{n-j}{k-j} \le (1 - \delta/2)\varepsilon \binom{n-j}{k-j}$$

as claimed. □

Our aim now is to prove Lemma 30, i.e. that whp stopping condition **(S1)** is applied. Our strategy is to show that whp each of the other three stopping conditions is *not* applied. The arguments for **(S2)** and **(S3)** are almost identical, so it is convenient to handle them together. We begin with the following proposition.

Proposition 32. *There exists an event $\mathcal{A}$ such that:*

(i) $\mathbb{P}(\mathcal{A}) = 1 - o(1)$;

(ii) *if $\mathcal{A}$ holds and either* **(S2)** *or* **(S3)** *is applied at time $t = T_{\text{stop}}$, then*

$$|E_t| \ge \frac{(1 - 2\delta\varepsilon/5)(1+\varepsilon)t}{\binom{n-j}{k-j}}.$$

Proof. We first define the event $\mathcal{A}$ explicitly. For any time $t > 0$ we define

$$\gamma_t := \begin{cases} \sqrt{\frac{n^{k-j+\beta}}{t}} & \text{if } t < T_0, \\ \frac{\delta\varepsilon}{3} & \text{otherwise,} \end{cases}$$

and let

$$\mathcal{A}_t := \left\{ |S_t| \ge (1 - \gamma_t) \frac{(1+\varepsilon)t}{\binom{n-j}{k-j}} \right\}.$$

Now we define

$$\mathcal{A} := \bigcap_{\frac{n^{k-j+\beta}}{4(k-j)!} \le t \le T_0} \mathcal{A}_t.$$

We now need to show that the two properties of the proposition are satisfied for this choice of $\mathcal{A}$. First observe that for $\frac{n^{k-j+\beta}}{4(k-j)!} \le t < T_0$, Proposition 21 (applied with $\gamma = \gamma_t$) implies that

$$\mathbb{P}(\mathcal{A}_t) \ge 1 - \exp\left(-\Theta\left(\gamma_t^2 pt\right)\right) \ge 1 - \exp\left(-\Theta\left(\gamma_t^2 \frac{t}{n^{k-j}}\right)\right) \ge 1 - \exp\left(-\Theta\left(n^\beta\right)\right).$$

On the other hand, for $t = T_0$ again Proposition 21 implies that

$$\mathbb{P}(\mathcal{A}_{T_0}) \ge 1 - \exp\left(-\Theta\left(\gamma_t^2 pT_0\right)\right) = 1 - \exp\left(-\Theta\left(\delta^2 \varepsilon n\right)\right) = 1 - o(1),$$

where the convergence holds because $\delta^2 \varepsilon n \ge \varepsilon^3 n \to \infty$. Therefore by applying a union bound,

$$\mathbb{P}(\mathcal{A}) \ge 1 - T_0 \exp\left(-\Theta\left(n^\beta\right)\right) - o(1) = 1 - o(1),$$

as required.

We now aim to prove the second statement, so let us assume that $\mathcal{A}$ holds, and we make a case distinction according to which of **(S3)** and **(S2)** is applied.

Case 1: **(S3)** *is applied.* By applying Lemma 31 we can bound the number of queries made from each explored j-set at any time $t \le T_{\text{stop}}$ from below by

$$(1 - \varepsilon + \delta\varepsilon/2)\binom{n-j}{k-j} \ge \frac{3n^{k-j}}{4(k-j)!}.$$

In particular, since **(S3)** is applied, we must have made at least $n^\beta/2$ new starts, and therefore at least $n^\beta/2 - 1 \ge n^\beta/3$ many j-sets are explored. Thus we have made at least $\frac{n^\beta}{3} \cdot \frac{3n^{k-j}}{4(k-j)!}$ queries, and therefore we may assume that $T_{\text{stop}} \ge \frac{n^{k-j+\beta}}{4(k-j)!}$. (Note that this in particular motivates why the definition of $\mathcal{A}$ did not include any $\mathcal{A}_t$ for $t < \frac{n^{k-j+\beta}}{4(k-j)!}$.)

Furthermore, since **(S2)** is *not* applied, we have $T_{\text{stop}} < T_0$. Therefore, the fact that $\mathcal{A}$ holds tells us that for $t = T_{\text{stop}}$,

$$|D_t| \ge |S_t| + |R_t| \ge (1 - \gamma_t)(1 + \varepsilon)\frac{t}{\binom{n-j}{k-j}} + |R_t|. \tag{16}$$

Since **(S3)** is applied at $t = T_{\text{stop}}$, we further have

$$|R_t| \ge 2(k-j)!\sqrt{\frac{tn^\beta}{n^{k-j}}} \ge \frac{3\gamma_t t}{2\binom{n-j}{k-j}}.$$

Substituting this inequality into (16), we obtain

$$|D_t| \ge (1 - \gamma_t)(1 + \varepsilon)\frac{t}{\binom{n-j}{k-j}} + \frac{3\gamma_t t}{2\binom{n-j}{k-j}} \ge (1 + \varepsilon)\frac{t}{\binom{n-j}{k-j}}.$$

Furthermore, since **(S3)** is applied at $t = T_{\text{stop}}$, a new start must have been made at time t. This implies that the set of active sets at time A_t was empty, i.e. $|A_t| = 0$. This means that

$$|E_t| = |D_t| \ge (1 + \varepsilon)\frac{t}{\binom{n-j}{k-j}} \ge \frac{(1 - 2\delta\varepsilon/5)(1 + \varepsilon)t}{\binom{n-j}{k-j}},$$

as claimed.

Case 2: **(S2)** *is applied.* We will use the trivial bound $|R_t| \geq 0$, and therefore $\mathcal{A}$ tells us that at time $t = T_0 = T_{\text{stop}}$ we have

$$|D_t| = |S_t| + |R_t| \geq \left(1 - \frac{\delta\varepsilon}{3}\right)(1+\varepsilon)\frac{t}{\binom{n-j}{k-j}}.$$

Furthermore, by **(S1)**,

$$\ell_t = O(\varepsilon n),$$

and therefore by (12)

$$|A_t| \leq 1 + \binom{k-j}{a}\ell_t = O(\varepsilon n) = O\left(\frac{\varepsilon^2 T_0}{n^{k-j}}\right).$$

Thus the number of explored sets at time T_0 satisfies

$$|E_{T_0}| = |D_{T_0}| - |A_{T_0}| \geq \frac{\left((1-\delta\varepsilon/3)(1+\varepsilon) - O\left(\varepsilon^2\right)\right)T_0}{\binom{n-j}{k-j}} \geq \frac{(1-2\delta\varepsilon/5)(1+\varepsilon)T_0}{\binom{n-j}{k-j}},$$

where in the last step we have used the fact that $\delta \gg \varepsilon$. □

The previous result enables us to prove the following.

Proposition 33. *Whp neither* **(S2)** *nor* **(S3)** *is applied.*

Proof. For any time $t \geq 0$, let us define the event

$$\mathcal{E}_t := \left\{ |E_t| \geq \frac{(1-2\delta\varepsilon/5)(1+\varepsilon)t}{\binom{n-j}{k-j}} \right\},$$

i.e. that the bound on $|E_t|$ from Proposition 32 holds. We will show that in fact it is not possible that $\mathcal{E}_t$ holds for any $t \leq T_{\text{stop}}$. Therefore, Proposition 32 implies that the probability that one of **(S3)** and **(S2)** is applied is at most $1 - \mathbb{P}(\mathcal{A}) = o(1)$. So suppose for a contradiction that $\mathcal{E}_t$ holds for some $t \leq T_{\text{stop}}$.

As in the proof of Proposition 32, an application of Lemma 31 implies that from each explored j-set at any time $t \leq T_{\text{stop}}$ we made at least

$$(1 - \varepsilon + \delta\varepsilon/2)\binom{n-j}{k-j} \geq \frac{3n^{k-j}}{4(k-j)!}$$

queries. Therefore, by Proposition 32, the total number t of queries made satisfies

$$t \geq |E_t| \cdot (1 - \varepsilon + \delta\varepsilon/2)\binom{n-j}{k-j} \geq (1 - 2\delta\varepsilon/5 + \delta\varepsilon/2 + O(\varepsilon^2))t > t,$$

yielding the desired contradiction. □

We next prove that whp **(S4)** is not applied. This may be seen as a form of *bounded degree lemma*. Both the result and the proof are inspired by similar results in [5, 6].

The intuition behind this stopping condition is that the *average* degree of an i-set should be of order $\frac{tp}{n^i} \sim \frac{t}{n^{k-j+i}}$, and **(S4)** guarantees that, for $t \leq T_{\text{stop}}$, no i-set exceeds this by more than a constant factor. The n^β-term can be interpreted as an error term which takes over when the average i-degree (i.e. the average degree over all i-sets) is too small to guarantee an appropriate concentration result.

Note, however, that due to the choice of T_0, the average i-degree is actually much smaller than n^β for any $i \geq 2$ (and possibly even for $i = 1$ if $\varepsilon = \Omega(n^{-\beta})$). Meanwhile, the statement for $i = 0$ is simply a statement about the number of discovered j-sets, which follows from a simple Chernoff bound on the number of edges discovered, together with **(S3)** to bound the number of new starts. Thus the strongest and most interesting case of the statement is when $i = 1$; nevertheless, our proof strategy is strong enough to cover all i and would even work for any $t > T_0$, provided **(S3)** has not yet been applied.

Lemma 34. *Whp* **(S4)** *is not applied.*

Proof. We will prove that the probability that **(S4)** is applied at a particular time $t \le T_{\text{stop}}$, i.e. before any other stopping condition has been applied, is at most $\exp\left(-\Theta\left(n^{\beta/2}\right)\right) = o(n^{-k})$, and then a union bound over all possible t completes the argument.

We will prove the lemma by induction on i. For $i = 0$ the statement is just that the number of discovered j-sets is at most $C_{k,j,0}t/n^{k-j} + n^\beta$, which follows from Lemma 10 and **(S3)**. More precisely, using (13) and applying Lemma 10 with $\alpha = \beta/2$, we have that

$$\mathbb{P}\left(\frac{|S_t|}{\binom{k-j}{a}} \ge 2tp + n^{\beta/2}\right) \le \exp\left(-\Theta\left(n^{\beta/2}\right)\right).$$

Furthermore, by **(S3)**, we have

$$\begin{aligned}
|R_t| &\le 2(k-j)!\sqrt{\frac{tn^\beta}{n^{k-j}}} + \frac{n^\beta}{2} \\
&\le \begin{cases} \frac{3n^\beta}{4} & \text{if } t \le \frac{n^{k-j+\beta}}{64((k-j)!)^2}, \\ 16((k-j)!)^2\frac{t}{n^{k-j}} + \frac{n^\beta}{2} & \text{if } t \ge \frac{n^{k-j+\beta}}{64((k-j)!)^2} \end{cases} \\
&\le 16((k-j)!)^2\frac{t}{n^{k-j}} + \frac{3n^\beta}{4}.
\end{aligned}$$

Thus with probability at least $1 - \exp\left(-\Theta\left(n^{\beta/2}\right)\right)$ we have

$$\begin{aligned}
|D_t| = |S_t| + |R_t| &\le \binom{k-j}{a}\left(2tp + n^{\beta/2}\right) + 16((k-j)!)^2\frac{t}{n^{k-j}} + \frac{3n^\beta}{4} \\
&\le \left(3(k-j)! + 16((k-j)!)^2\right)\cdot\frac{t}{n^{k-j}} + n^\beta \\
&\le \frac{20((k-j)!)^2 t}{n^{k-j}} + n^\beta,
\end{aligned}$$

and since we chose $C_{k,j,0} \gg 1$, and in particular $C_{k,j,0} > 20((k-j)!)^2$, this shows that whp **(S4)** is not applied because of $I = \emptyset$ (i.e. with $i = 0$). So we will assume that $i \ge 1$ and that **(S4)** is not applied for $0, 1, \dots, i-1$.

Given $1 \le i \le j-1$ and an i-set I, let us consider the possible ways in which some j-sets containing I may become active.

- A *new start* at I occurs when there are no active j-sets and we make a new start at a j-set which happens to contain I. In this case $d(I)$ increases by 1;
- A *jump* to I occurs when we query a k-set containing I from a j-set not containing I and discover an edge. In this case $d(I)$ increases by at most $\binom{k-j}{a}$ (the number of new j-sets which become active in a batch, each of which may or may not contain I);
- A *pivot* at I occurs when we query a k-set from a j-set containing I and discover an edge. In this case $d(I)$ increases by at most $\binom{k-j}{a}$.

Each possibility makes a contribution to the degree of I according to how many j-sets containing I become active as a result of each type of event. We bound the three contributions separately.

New starts: Whenever we make a new start, we choose the starting j-set according to some (previously fixed) random ordering σ_j (recall that σ_j was a permutation of the j-sets chosen uniformly at random during the initialisation of the algorithm). By **(S3)**, at time $t \le T_{\text{stop}}$ the number of new starts we have made is

$$|R_t| \le 2(k-j)!\sqrt{\frac{tn^\beta}{n^{k-j}}} + \frac{n^\beta}{2}.$$

Observe that

$$\sqrt{\frac{tn^\beta}{n^{k-j}}} \le \begin{cases} n^\beta & \text{if } t \le n^{k-j+\beta}, \\ \frac{t}{n^{k-j}} & \text{if } t \ge n^{k-j+\beta}, \end{cases}$$

which means that the number of new starts satisfies

$$|R_t| \le 2(k-j)!tn^{j-k} + 3(k-j)!n^\beta =: N^*.$$

Since the new starts are distributed randomly, the probability that a j-set chosen for a new start at time $t' \le t$ contains I is precisely the proportion of neutral j-sets at time t' which contain I. Since **(S4)** has not yet been applied, in particular with $i = 0$, the total number of non-neutral j-sets (which cannot be chosen for a new start) at time $t' \le t$ is at most

$$d_{t'}(\emptyset) \le d_t(\emptyset) \le \frac{C_{k,j,0}t}{n^{k-j}} + n^\beta \le \frac{C_{k,j,0}n}{\varepsilon} + n^\beta \le n^{4/3} = o(n^j).$$

Thus the probability that the j-set chosen contains I is at most

$$\frac{\binom{n-i}{j-i}}{\binom{n}{j} - o(n^j)} \le 2j!n^{-i}.$$

Therefore the number of new starts containing I is dominated by $\mathrm{Bin}(N^*, 2j!n^{-i})$, which has expectation at most $4k!tn^{j-k-i} + 1$ (since $n^{\beta-i} = o(1)$). By Lemma 10, with probability at least $1 - \exp(-\Theta(n^{\beta/2}))$ the number of new starts at I is at most

$$8k!tn^{j-k-i} + 2 + n^{\beta/2} \le 8k!tn^{j-k-i} + n^{2\beta/3}.$$

Taking a union bound over all possible i-sets I, with probability at least

$$1 - \binom{n}{i}\exp(-\Theta(n^{\beta/2})) = 1 - \exp(-\Theta(n^{\beta/2})),$$

every i-set is contained in at most

$$8k!n^{j-k-i}t + n^{2\beta/3} \tag{17}$$

new starts.

Jumps: From each j-set J which became active in the search process, but which did not contain I, if we queried a k-set containing I and this k-set was an edge, then the degree of I may increase by up to $\binom{k-i}{j-i}$. To bound the number of such jumps, we distinguish according to the intersection $Z = J \cap I$, and denote $z := |Z|$. Observe that $0 \le z \le i-1$, and for each of the $\binom{i}{z}$ many z-sets $Z \subset I$, by the fact that **(S4)** has not been previously applied for this set Z, there are at most $d_t(Z) \le \frac{C_{k,j,z}t}{n^{k-j+z}} + n^\beta$ many j-sets in D_t which intersect I in Z. For each such j-set J, there are at most $\binom{n}{k-j-i+z} \le n^{k-j-i+z}$ many k-sets containing both J and I, i.e. which we might have queried from J and which would result in jumps to I.

Thus in total, the number of k-sets which we may have queried and which might have resulted in a jump to I is at most

$$\begin{aligned}\sum_{z=0}^{i-1}\binom{i}{z}\left(\frac{C_{k,j,z}t}{n^{k-j+z}} + n^\beta\right)n^{k-j-i+z} &= \sum_{z=0}^{i-1}\binom{i}{z}\left(\frac{C_{k,j,z}t}{n^i} + n^{k-j-i+z+\beta}\right)\\ &\le 2^i\left(\max_{0\le z\le i-1} C_{k,j,z}\frac{t}{n^i} + n^{k-j-1+\beta}\right)\\ &= 2^i\left(C_{k,j,i-1}\frac{t}{n^i} + n^{k-j-1+\beta}\right) =: N,\end{aligned}$$

since we chose $C_{k,j,j-1} \gg C_{k,j,j-2} \gg \ldots \gg C_{k,j,0}$. Then the number of edges that we discover which result in jumps to I is dominated by $\mathrm{Bin}(N,p)$. By Lemma 10, with probability at least $1 - \exp(-\Theta(n^{\beta/2}))$ this random variable is at most

$$\begin{aligned}2Np + n^{\beta/2} &\le \frac{(k-j)!}{\binom{k-j}{a}}2^{i+2}C_{k,j,i-1}\frac{t}{n^{k-j+i}} + O\left(n^{-1+\beta}\right) + n^{\beta/2}\\ &\le \frac{(k-j)!}{\binom{k-j}{a}}2^{i+2}C_{k,j,i-1}\frac{t}{n^{k-j+i}} + 2n^{\beta/2},\end{aligned}$$

and so the contribution to the degree of I made by jumps to I is at most

$$(k-j)!2^{i+2}C_{k,j,i-1}\frac{t}{n^{k-j+i}} + n^{2\beta/3}. \tag{18}$$

Pivots: Whenever we have a jump to I or a new start at I, some j-sets containing I become active. From these j-sets we may query further k-sets, potentially resulting in some more j-sets containing I becoming active. However, the number of such j-sets containing I that become active due to such a pivot is certainly at most $\binom{k-j}{a}$. Thus the number of further j-sets that become active due to pivots from some j-set J is at most $\binom{k-j}{a} \cdot \mathrm{Bin}\left(\binom{n-j}{k-j}, p\right)$, which has expectation $1+\varepsilon$.

Furthermore, the number of such sequential pivots that we may make before leaving I in the j-tight path is $\lfloor \frac{k-i}{k-j} \rfloor \leq k-i$. Thus the number of pivots arising from a single j-set containing I may be upper coupled with a branching process in which vertices in the first $(k-i)$ generations produce $\binom{k-j}{a} \cdot \mathrm{Bin}\left(\binom{n-j}{k-j}, p\right)$ children, and thereafter no more children are produced.

We bound the total size of all such branching processes together. Suppose the contribution to the degree of I made by jumps and new starts is x. Then we have x vertices in total in the first generation, and by the arguments above, with probability $1-\exp(-\Omega(n^{\beta/2}))$ we have, by (17) and (18), that

$$x \leq \left(8k! + (k-j)!2^{i+2}C_{k,j,i-1}\right)\frac{t}{n^{k-j+i}} + 2n^{2\beta/3} \leq 2^{i+3}k!C_{k,j,i-1}\frac{t}{n^{k-j+i}} + 2n^{2\beta/3}.$$

For convenience, we will assume (for an upper bound) that in fact $x \geq n^{\beta}$. The number of children in the second generation is dominated by $\binom{k-j}{a} \cdot \mathrm{Bin}\left(x\binom{n-j}{k-j}, p\right)$, which has expectation $(1+\varepsilon)x$, and so by Lemma 10, with probability $1-\exp(-\Omega(n^{\beta/2}))$, the number of children is at most $2(1+\varepsilon)x + n^{\beta/2} \leq 4x$. Similarly, with probability $1-\exp(-\Omega(n^{\beta/2}))$, the number of vertices in the third generation is at most $16x$, and inductively the number of vertices in the m-th generation is at most $2^{2(m-1)}x$ for $1 \leq m \leq k-i+1$. Thus in total, with probability at least $1-\exp(-\Theta(n^{\beta/2}))$, the number of vertices in total in all these branching processes is at most

$$\sum_{m=1}^{k-i} 2^{2(m-1)}x \leq 2^{2k}x \leq 2^{3k+3}k!C_{k,j,i-1}\frac{t}{n^{k-j+i}} + n^{\beta}.$$

However, the vertices in the branching process exactly represent (an upper coupling on) the j-sets which can be discovered due to jumps to or new starts at I and the pivots arising from them, which are all of the j-sets containing I which we discover in the `Pathfinder` algorithm. Thus with probability at least $1-\exp(-\Theta(n^{\beta/2}))$, the number of j-sets containing I which became active is at most

$$2^{3k+3}k!C_{k,j,i-1}\frac{t}{n^{k-j+i}} + n^{\beta} \leq C_{k,j,i}\frac{t}{n^{k-j+i}} + n^{\beta},$$

since we chose $C_{k,j,i} \gg C_{k,j,i-1}$. Taking a union bound over all $\binom{n}{i}$ many i-sets I, and observing that $\binom{n}{i}\exp(-\Theta(n^{\beta/2})) = o(1)$, the result follows. □

Proof of Lemma 30. The statement of Lemma 30 follows directly from Proposition 33 and Lemma 34. □

8. Longest paths: proof of Theorem 4

The various statements contained in Theorem 4 have now all been proved, with the exception of the upper bounds, whose completely standard proofs we omit.

- The upper bounds of statements (i), (ii) and (iii) of Theorem 4 can be proved using a basic first moment method. The details can be found in Appendix A.3 of arXiv:2003.14143.
- The lower bound of statement (i) follows directly from Lemmas 11 and 12.
- The lower bound of statement (ii) is implied by Lemma 30, which is identical except that it omits the assumption that $\delta \gg \frac{\ln n}{\varepsilon^2 n}$.
- The lower bound of statement (iii) is precisely Lemma 24.

9. Concluding remarks

Theorem 4 provides various bounds on the length L of the longest j-tight path, but these bounds may not be best possible. Let us examine each of the three cases in turn.

9.1. **Subcritical case.** Here we proved the bounds

$$\frac{j\ln n-\omega+3\ln\varepsilon}{-\ln(1-\varepsilon)}\le L\le\frac{j\ln n+\omega}{-\ln(1-\varepsilon)}.$$

A more careful version of the first moment calculation implies that if $\ell=\frac{j\ln n+c}{-\ln(1-\varepsilon)}$ for some constant $c\in\mathbb{R}$, then the expected number of paths of length ℓ is asymptotically $d\cdot e^c$, where $d=\frac{(a!b!)^\ell}{z_\ell}=\frac{b!(a!b!)^s}{2((k-j)!)^{2s}}$. This suggests heuristically that in this range, the probability that $X_\ell=0$, i.e. that there are no paths of length ℓ, is a constant bounded away from both 0 and 1, and that in fact the bounds on L are best possible up to the $3\ln\varepsilon$ term in the lower bound. This term is negligible (and can be incorporated into ω) if ε is constant, but as ε decreases, it becomes more significant. The term arises because as ε decreases, the paths become longer, meaning that there are many more pairs of possible paths whose existences in $H^k(n,p)$ are heavily dependent on one another, and the second moment method breaks down. Thus to remove the $3\ln\varepsilon$ term in the lower bound requires some new ideas.

9.2. **Supercritical case for $j\ge 2$.** In this case, we had the bounds

$$(1-\delta)\frac{\varepsilon n}{(k-j)^2}\le L\le(1+\delta)\frac{2\varepsilon n}{(k-j)^2}.$$

Since in particular we may assume that $\delta\ll 1$, the upper bound (provided by the first moment method) and the lower bound (provided by the analysis of the `Pathfinder` algorithm) differ by approximately a factor of 2.

One possible explanation for this discrepancy comes from the fact that we do not query a k-set if it contains some explored j-set. As previously explained, this condition is not necessary to guarantee the correct running of the algorithm, but it is fundamentally necessary for our *analysis* of the algorithm, since it ensures that no k-set is queried twice and therefore each query is independent.

Removing this condition would allow us to try out many different paths with the same end (i.e. different ways of reaching the same destination), which could potentially lead to a longer final path since different sets of vertices are used in the current path and are therefore forbidden for the continuation.

It is not hard to prove that the length ℓ of the current path in the modified algorithm would very quickly reach almost $\frac{\varepsilon n}{(k-j)^2}$ (i.e. our lower bound). For each possible way of reaching this, it is extremely unlikely that the path can be extended significantly, and in particular to length $\frac{2\varepsilon n}{(k-j)^2}$. However, since there will be very many of these paths, it is plausible that at least one of them may go on to reach a larger size, and therefore our lower bound may not be best possible.

On the other hand, it could be that our upper bound is not best possible, i.e. that the first moment heuristic does not give the correct threshold path length. This could be because if there is one very long path, there are likely to be many more (which can be obtained by minor modifications), and so we may not have concentration around the expectation.

Therefore further study is required to determine the asymptotic value of L more precisely.

9.3. **Supercritical case for $j=1$.** For loose paths, we proved the bounds

$$(1-\delta)\frac{\varepsilon^2 n}{4(k-1)^2}\le L\le(1+\delta)\frac{2\varepsilon n}{(k-1)^2},$$

which differ by a factor of $\Theta(\varepsilon)$. In view of the supercritical case for $j\ge 2$, when the longest path is of length $\Theta(\varepsilon n)$ one might naively expect this to be the case for $j=1$ as well, and that the lower bound is incorrect simply because the proof method is too weak for $j=1$.

However, this is not the case for graphs, i.e. when $k=2$ and $j=1$, when the longest path is indeed of length $\Theta(\varepsilon^2 n)$. The analogous result for general k and $j=1$ was recently achieved by

Cooley, Kang and Zalla [7], who proved an upper bound of approximately $\frac{2\varepsilon^2 n}{(k-1)^2}$ by bounding the length of the longest loose *cycle* (via consideration of an appropriate 2-core-like structure) and using a sprinkling argument. Nevertheless, this leaves a multiplicative factor of 8 between the upper and lower bounds, which it would be interesting to close.

9.4. **Critical window.** One might also ask what happens when ε is smaller than allowed here, i.e. when $\varepsilon^3 n \nrightarrow \infty$. In the case $j = 1$, the lower bounds in the subcritical and supercritical case, of orders $\frac{\ln(\varepsilon^3 n)}{\varepsilon}$ and $\varepsilon^2 n$ respectively, would both be $\Theta(n^{1/3})$ when $\varepsilon^3 n = \Theta(1)$, which suggests that this may indeed be the correct critical window when $j = 1$. However, for $j \geq 2$, the bounds differ by approximately a factor of $n^{1/3}$ when $\varepsilon^3 n = \Theta(1)$. It would therefore be interesting to examine whether the statement of Theorem 4 remains true for $j \geq 2$ even for smaller ε.

Acknowledgement

The collaboration leading to this paper was made possible partially by the support of the Heilbronn Institute for Mathematical Research and EPSRC grant EP/P032125/1 during the workshop "Structure and randomness in hypergraphs", 17-21 December 2018, LSE London.

References

[1] M. Ajtai, J. Komlós, and E. Szemerédi. The longest path in a random graph. *Combinatorica*, 1:1–12, 03 1981.
[2] P. Allen, J. Böttcher, O. Cooley, and R. Mycroft. Tight cycles and regular slices in dense hypergraphs. *J. Combin. Theory Ser. A*, 149:30–100, 2017.
[3] M. Anastos and A. Frieze. A scaling limit for the length of the longest cycle in a sparse random graph. *J. Combin. Theory Ser. B*, 148:184–208, 2021.
[4] O. Cooley. Paths, cycles and sprinkling in random hypergraphs, 2021. ArXiv: 2103.16527.
[5] O. Cooley, M. Kang, and C. Koch. The size of the giant high-order component in random hypergraphs. *Random. Struct. Algor.*, 53(2):238–288, 2018.
[6] O. Cooley, M. Kang, and Y. Person. Largest components in random hypergraphs. *Combin. Probab. Comput.*, 27(5):741–762, 2018.
[7] O. Cooley, M. Kang, and J. Zalla. Loose cores and cycles in random hypergraphs, 2021. ArXiv: 2101.05008.
[8] A. Dudek and A. Frieze. Loose Hamilton cycles in random uniform hypergraphs. *Electron. J. Combin.*, 18(1):Paper 48, 14, 2011.
[9] A. Dudek and A. Frieze. Tight Hamilton cycles in random uniform hypergraphs. *Random. Struct. Algor.*, 42(3):374–385, 2013.
[10] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.
[11] E. Győri, G. Y. Katona, and N. Lemons. Hypergraph extensions of the Erdős-Gallai theorem. *European J. Combin.*, 58:238–246, 2016.
[12] S. Janson, T. Łuczak, and A. Ruciński. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000.
[13] G. Kemkes and N. Wormald. An improved upper bound on the length of the longest cycle of a supercritical random graph. *SIAM J. Discrete Math.*, 27(1):342–362, 2013.
[14] M. Krivelevich and B. Sudakov. The phase transition in random graphs: A simple proof. *Random. Struct. Algor.*, 43(2):131–138, 2013.
[15] D. Kühn and D. Osthus. Hamilton cycles in graphs and hypergraphs: an extremal perspective. In *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. IV*, pages 381–406. Kyung Moon Sa, Seoul, 2014.
[16] T. Łuczak. Cycles in a random graph near the critical point. *Random. Struct. Algor.*, 2(4):421–439, 1991.
[17] D. Mubayi and A. Suk. A survey of hypergraph Ramsey problems. In *Discrete Math. Appl*, volume 165 of *Springer Optim. Appl.*, pages 405–428. Springer, Cham, [2020] ©2020.
[18] B. Narayanan and M. Schacht. Sharp thresholds for nonlinear Hamiltonian cycles in hypergraphs. *Random Structures Algorithms*, 57(1):244–255, 2020.
[19] B. Pittel. A random graph with a subcritical number of edges. *Trans. Amer. Math. Soc.*, 309(1):51–75, 1988.
[20] J. Schmidt-Pruzan and E. Shamir. Component structure in the evolution of random hypergraphs. *Combinatorica*, 5(1):81–94, 1985.
[21] Y. Zhao. Recent advances on Dirac-type problems for hypergraphs. In *Recent trends in combinatorics*, volume 159 of *IMA Vol. Math. Appl.*, pages 145–165. Springer, 2016.