Combinatorics, Probability and Computing (2018) 27, 741–762.
doi:10.1017/S096354831800010X

Largest Components in Random Hypergraphs

OLIVER COOLEY[1†], MIHYUN KANG[1‡] and YURY PERSON[2†]

[1] Institute of Discrete Mathematics, Graz University of Technology, Steyrergasse 30, 8010 Graz, Austria
(e-mail: cooley@math.tugraz.at, kang@math.tugraz.at)

[2] Goethe-Universität, Institute of Mathematics, Robert-Mayer-Strasse 10, 60325 Frankfurt, Germany
(e-mail: person@math.uni-frankfurt.de)

Received 17 December 2014; revised 24 November 2017; first published online 4 April 2018

In this paper we consider j-tuple-connected components in random k-uniform hypergraphs (the j-tuple-connectedness relation can be defined by letting two j-sets be connected if they lie in a common edge and considering the transitive closure; the case $j = 1$ corresponds to the common notion of vertex-connectedness). We show that the existence of a j-tuple-connected component containing $\Theta(n^j)$ j-sets undergoes a phase transition and show that the threshold occurs at edge probability

$$\frac{(k-j)!}{\binom{k}{j}-1}n^{j-k}.$$

Our proof extends the recent short proof for the graph case by Krivelevich and Sudakov, which makes use of a depth-first search to reveal the edges of a random graph.

Our main original contribution is a *bounded degree lemma*, which controls the structure of the component grown in the search process.

2010 *Mathematics subject classification*: Primary 05C65
Secondary 05C80

1. Introduction

1.1. Phase transition in random graphs

The Erdős–Rényi random graph [11] $G(n,p)$ (resp. $G(n,M)$) is one of the most intensely studied in the theory of random graphs. It is well known, and it has been studied in great detail (see *e.g.* [8, 12]) how the structure of the components changes as p (resp. M) grows. In the seminal paper

† The first and third authors were supported by short visit grants 5639 and 5472, respectively, from the European Science Foundation (ESF) within the 'Random Geometry of Large Interacting Systems and Statistical Physics' (RGLIS) programme.

‡ The second author is supported by Austrian Science Fund (FWF): P26826, W1230, Doctoral Programme 'Discrete Mathematics'.

[11] entitled 'On the evolution of random graphs', Erdős and Rényi discovered among other things that the Erdős–Rényi random graph undergoes a drastic change of the size and structure of largest components, which happens when the number of edges is around $n/2$. In terms of the binomial model $G(n,p)$, this phenomenon can be stated as follows. Consider $G(n,p)$ with $p=c/n$ for a constant $c>0$. If $c<1$, then asymptotically almost surely (a.a.s. for short, meaning with probability tending to one as n tends to ∞) all the components in $G(n,p)$ have $O(\log n)$ vertices, whereas if $c>1$, then a.a.s. there is a unique component with $\rho n+o(n)$ vertices, the so-called giant component, where ρ is the unique positive solution of the equation

$$1-\rho=\exp(-c\rho).$$

In 1984, Bollobás [7] made a breakthrough in the study of the so-called critical phenomenon associated with the phase transition by studying the case $c\to 1$ in more detail. His result was improved by Łuczak in 1990 [17]. Let $\lambda=\lambda(n)$ be such that

$$p=\frac{1}{n}+\frac{\lambda}{n^{4/3}}. \tag{1.1}$$

If $\lambda\to-\infty$, then a.a.s. all the components have order $o(n^{2/3})$. If $\lambda\to+\infty$, then there is a.a.s. a unique component of order $\gg n^{2/3}$, while all other components have order $o(n^{2/3})$. If λ is a constant, then the size of the largest component is $\Theta_p(n^{2/3})$.

1.2. Phase transition in random hypergraphs

A k-uniform hypergraph H is a tuple (V,E), where V is the vertex set of H and E is its edge set with $E\subseteq\binom{V}{k}$. The random k-uniform hypergraph $H^k(n,p)$ is defined similarly to $G(n,p)$: each of the $\binom{n}{k}$ possible edges is included independently of the others with probability p.

Similar phase transition phenomena were discovered in random hypergraphs. In particular, a straightforward generalisation of the giant component was studied in [4, 5, 13, 21], where the following concept of 'component', called *vertex-component*, was studied: two vertices u and v are connected in a k-uniform hypergraph H if there is a sequence $e_0,\dots,e_\ell$ of edges of H such that $u\in e_0$ and $v\in e_\ell$ and $e_i\cap e_{i+1}\neq\emptyset$.

The threshold for a giant vertex-component in $H^k(n,p)$ was first determined by Schmidt-Pruzan and Shamir [21]. More precisely, let $p=c/\binom{n-1}{k-1}$. If $c<(k-1)^{-1}-\varepsilon$ for an arbitrarily small but fixed $\varepsilon>0$, then a.a.s. the number of vertices of the largest component is $O(\log n)$. But if $c>(k-1)^{-1}+\varepsilon$, then a.a.s. there is a unique component containing a linear number of vertices, which is called the *giant component*; more precisely, the number of vertices of the giant component is $\rho n+o(n)$, where ρ is the unique positive solution to the equation

$$1-\rho=\exp(c((1-\rho)^{k-1}-1)).$$

This result was subsequently strengthened in various ways by Karoński and Łuczak [13], Ravelomanana and Rijamamy [20], Behrisch, Coja-Oghlan and Kang [4], and Bollobás and Riordan [9].

While in the graph case two vertices are connected if there is a path (or walk) between them, in hypergraphs the notion of a path (or walk) is ambiguous, and in fact there are several possible definitions. An *s-tight path* of length m in a k-uniform hypergraph H is a sequence $e_0,\dots,e_{m-1}$ of edges of H such that $e_i=\{v_{i(k-s)+1},\dots,v_{i(k-s)+k}\}$ for some distinct vertices v_j. In the case $s=1$ we call an s-tight path a *loose path*, and for $s=k-1$ simply a *tight* path.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Note that when $p = (k-2)!/n^{k-1}$, the edges in $H^k(n,p)$ typically intersect in at most one vertex; thus, 'morally' if two vertices are connected then they are connected by loose paths (mostly). The result of Schmidt-Pruzan and Shamir, incorporating some of the improvements mentioned above, can be restated as follows.

Theorem 1.1 ([9, 13, 21]). *Let $k \geqslant 2$ and $n^{-1/3} \ll \varepsilon = \varepsilon(n) < 1$ be given, and let ρ be the unique positive solution to the equation*

$$1-\rho = \exp\left(\frac{1+\varepsilon}{k-1}((1-\rho)^{k-1}-1)\right).$$

Then a.a.s. the (vertex) size of the giant component in the random k-uniform hypergraph $H^k(n,p)$ is $(1+o(1))\rho n$ if $p = (1+\varepsilon)(k-2)!/n^{k-1}$ and $O(\log n/\varepsilon^2)$ if $p = (1-\varepsilon)(k-2)!/n^{k-1}$.

In this paper we study the following notion of j-tuple-connectedness in k-uniform hypergraphs, which generalises the notion mentioned above (if $j = 1$ we simply speak about vertex-connectedness). We say two j-sets (j-tuples of vertices) J_0 and J_s ($j \in \{1,\dots,k-1\}$) are *j-tuple-connected* in the k-uniform hypergraph H if there is an alternating sequence of j- and k-element subsets of $V(H)$: $J_0, e_0, J_1, e_1, \dots, J_s$ such that $J_i \cup J_{i+1} \subseteq e_i$ and $e_i \in E(H)$ for $i = 0,\dots,s-1$. The components then consist of j-element subsets of the vertex set of H. Again, one might wonder when a j-tuple-connected giant component of size (*i.e.* number of j-sets) $\Theta(n^j)$ emerges in the random k-uniform hypergraph.

For the rest of the paper we will regard k and j as fixed constants. In particular, this means that any parameter which is a function only of k and j is also a fixed constant.

1.3. Intuition: where to locate the thresholds?

The intuition (as in the case of random graphs) comes from the branching processes, which can be described for the general case of j-tuple-connectedness as follows. Initially we start with a single j-element set J_0. We expect that there are $m := p\binom{n-j}{k-j} \sim pn^{k-j}/(k-j)!$ edges $e_1,\dots,e_m$ containing J_0 in $H^k(n,p)$. The range of p is typically such that these edges intersect pairwise only in J_0, which leads to $\binom{k}{j}-1$ offspring for each edge e_i. From the theory of branching processes, the process survives forever with positive probability if $\left(\binom{k}{j}-1\right)m > 1$ (if everything is independent). This suggests that the threshold should be

$$p_{k,j} = p_{k,j}(n) := \frac{(k-j)!}{\binom{k}{j}-1} n^{j-k}. \tag{1.2}$$

For $j = 1$ we obtain $p = (k-2)!/n^{k-1}$, which is exactly the threshold: see [11, 21]. For $j = k-1$ the conjectured threshold is $1/((k-1)n)$. Our main theorem shows that $p_{k,j}$ is the correct threshold for all k, j, as was suggested recently by Bollobás and Riordan [9].

Our approach builds on the recent proof strategy of Krivelevich and Sudakov [14] who used a search algorithm in graphs to give a simple and short proof of the phase transition in $G(n,p)$. More precisely, we first adapt their proof strategy for $j = 1$, thus deriving an alternative proof of Theorem 1.1. Moreover, the approach via depth-first search allows us to study the largest component in the early supercritical phase, *i.e.* when $p = (1+\varepsilon)(k-2)!/n^{k-1}$ with $\varepsilon = \varepsilon(n) \gg n^{-1/3}$, which gives the lower bound $\Omega(\varepsilon n)$ for the size of the largest component in random

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

hypergraphs. This range of ε matches that considered by Bollobás and Riordan [9] and is essentially best possible.

Then we turn to the case of general k, j, which requires some additional work, most notably Lemma 4.5. We obtain the following theorem, thus confirming the threshold for $p_{k,j}$ mentioned above. By $\omega(f(n))$ we denote any function $g(n)$ such that $g(n)/f(n) \to \infty$ as $n \to \infty$.

Theorem 1.2. *Let $0 < \varepsilon = \varepsilon(n) < 1$ and $1 \leqslant j \leqslant k-1$ be given. Then a.a.s. the size of the largest j-tuple-connected component in the random k-uniform hypergraph $H^k(n,p)$ is $O(\varepsilon^{-2}\log n)$ if $p = (1-\varepsilon)p_{k,j}$.*

Let $\delta \in (0,1)$ be any constant. If furthermore $\varepsilon = \omega(n^{\delta-1} + n^{-j/3})$, then a.a.s. the largest j-tuple-connected component in $H^k(n,p)$ has size $\Omega(\varepsilon n^j)$ if $p = (1+\varepsilon)p_{k,j}$.

Note in particular that Theorem 1.1 is an immediate corollary, although we will prove Theorem 1.1 first, since the proof of the special case is substantially simpler.

Note also that there is no lower bound on the size of ε in the first part of the theorem. However, for very small ε, the bound on the largest component is not best possible, and may even be greater than $\binom{n}{j}$ and therefore useless as a bound. We discuss the critical window in more detail in Section 5.

While preparing this paper we discovered that independently Lu and Peng [16] have claimed to have a proof of a similar result, although only for constant ε.

Our main contribution to the proof of Theorem 1.2 is Lemma 4.5, which will be formally stated in Section 4. Briefly, it states that for some $\alpha \in (0,1)$ which is a function of ε, with high probability the set of j-sets which have been discovered by time αn^k is 'smooth' in the following sense: for any $1 \leqslant \ell \leqslant j-1$, any ℓ-set is contained in $O(\alpha n^{j-\ell})$ such j-sets. (This is best possible up to a constant factor.)

In the proof of Theorem 1.2 (and the special case $j = 1$, which is Theorem 4.3), for the case $p = (1+\varepsilon)p_{k,j}$ we will implicitly assume that ε is less than some small constant, say ε_0, which is dependent on k, j. This is permissible since if $\varepsilon > \varepsilon_0$, then our aim is to prove that there is a component of size $\Omega(\varepsilon n^j) = \Omega(\varepsilon_0 n^j) = \Omega(n^j)$, and thus the result for $\varepsilon > \varepsilon_0$ is implied by the result for $\varepsilon = \varepsilon_0$.

1.4. Motivation from random simplicial complexes

A parallel development was initiated by Linial and Meshulam, who studied homological connectivity of random simplicial complexes [15]. Furthermore, motivated by finding thresholds for various algebraic notions of cycles in $H^k(n,p)$, questions such as collapsibility and vanishing of the top homology have been investigated in [1, 2, 3]. A k-uniform hypergraph H is *collapsible* if one can delete all of its edges one by one, such that in each step we remove some edge e containing a $(k-1)$-element set J if e does not intersect any other edge in J. It has been shown in [3] that the first emerging cycle in the $(k-1)$th homology group of $H^k(n,p)$ is either K^k_{k+1} or contains $\Omega(n^{k-1})$ edges. Our Theorem 1.2 in the case $j = k-1$ may be seen as the study of the acyclic case of $H^k(n,p)$, where we have a sharp threshold for the emergence of a tightly connected 'hypertree' with $\Theta(n^{k-1})$ edges covering all vertices.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

2. Exploration of random hypergraphs via depth-first search

Usually we denote the vertex set of a hypergraph $\mathcal{H}$ by $[n] := \{1,2,\dots,n\}$. We also let $e(\mathcal{H})$ denote the order of the edge set $E(\mathcal{H})$.

2.1. Exploration algorithm in hypergraphs

Now we introduce the depth-first search algorithm (DFS) for hypergraphs. We are given as input two hypergraphs H and $\mathcal{H}$ on the same vertex set V with $E(H) \subseteq E(\mathcal{H})$, and we would like to discover all edges of H by *querying* the edges of $\mathcal{H}$ to determine whether they belong to H. We will choose vertices and edges to query according to certain linear orderings (see Algorithm 1).

There are three types of vertices: *neutral*, *active* and *explored* (we borrow the terminology from [19]). Additionally, vertices that are active or explored (*i.e.* any non-neutral vertices) are called *discovered*. Initially, all vertices are neutral. We start our exploration from the smallest neutral vertex, which we mark as active. Departing from some (currently considered) active vertex v we *query* any previously unqueried triples which contain v and at least one neutral vertex (but no explored vertices) in increasing order until we discover an edge e of $\mathcal{H}$. Once such an edge is found we mark all neutral vertices in e active and start the same query process from the vertex which was marked active last. If no edge e could be found, we mark v as explored and start the same querying process from the next active vertex. If no vertices are active then we have discovered some component completely, and we proceed as in the beginning of the depth-first search. Finally, once all vertices are discovered, we query all unqueried edges. Notice however that at the moment when all vertices are discovered, we know the vertex-components of H.

A complete description of the algorithm is given in Algorithm 1.

2.2. Coupling

Let $H := \mathcal{H}_p$ denote the random subhypergraph of $\mathcal{H}$ where every edge of $\mathcal{H}$ is chosen independently of the other edges with probability p. Further, let $(X_i)_{i\in[e(\mathcal{H})]}$ be a sequence of $e(\mathcal{H})$ i.i.d. Bernoulli random variables with mean p. We can associate with the ith query the random variable X_i, meaning that if $X_i = 1$ then the queried edge is in H and otherwise not. Once we have fixed the orderings of the vertices and edges and the values of the X_i, Algorithm 1 is a deterministic one and it queries every edge of $\mathcal{H}$ exactly once, thus every $\{0,1\}$-sequence of length $e(\mathcal{H})$ corresponds to a unique subgraph of $\mathcal{H}$.

In this way, for given fixed orderings σ and τ as in Algorithm 1, we couple the X_i with $\mathcal{H}_p$. For technical reasons which are not needed in the case of vertex-connectedness, we let the choice of τ be uniformly random independently of σ (and also of $\mathcal{H}_p$, since τ is chosen before any of $\mathcal{H}_p$ is revealed). In fact, we will only need that τ is chosen randomly in this way at one point in the paper (Lemma 4.6); an arbitrary ordering τ would work almost as well, but would lead to some additional technical difficulties in the range when ε is very small.

Remark 2.1. Note that the ordering τ is only used when starting a new component to determine which vertex we will continue exploring: the rest of the algorithm is independent of τ. It is easy to see that choosing τ in this way is equivalent to choosing a neutral vertex uniformly at random from which to continue exploring. It is this interpretation that we will consider in Lemma 4.6.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Algorithm 1: Hypergraph exploration DFS

Input: $\mathcal{H}$, $H = (V,E) \subseteq \mathcal{H}$: k-uniform hypergraphs.
σ: linear ordering of $E(\mathcal{H})$.
τ: linear ordering of $V(\mathcal{H})$.
Output: $\mathcal{C}$: set of vertex-connected components of H

```
1  C := ∅;
2  let S be an empty stack;
3  repeat
4      let x be the smallest neutral vertex in the ordering τ;
5      mark x as active;
6      add x to S;
7      while S ≠ ∅ do
8          Let x be the top vertex of S;
9          if ∃ the smallest unqueried edge e of H such that x ∈ e and e contains at least one
           neutral vertex then
10             if e ∈ E(H) ('query e') then
11                 add all neutral vertices of e in ascending order to the top of S;
12                 mark these vertices as active;
13         else
14             remove x from S;
15             mark x as explored;
16     let C be the set of vertices explored in the while-loop above;
17     C := C ∪ {C};
18 until all vertices are explored;
19 query all remaining edges of H in ascending order;
```

We will show the existence of a large component (Theorem 1.2) by proving that for appropriate small α, after αn^k queries Algorithm 1 (or one of its relatives, which will be defined later) will have found a large j-tuple-connected component in $H^k(n,p)$ a.a.s.

In the case of vertex-connectedness our $\mathcal{H} = \left([n], \binom{[n]}{k}\right)$ is the complete k-uniform hypergraph and H is the random k-uniform hypergraph $H^k(n,p)$.

In order to study j-tuple-connectedness we could alter Algorithm 1, in that we visit j-element sets of vertices instead of single vertices. Instead of this, we define the $\binom{k}{j}$-uniform hypergraph $\mathcal{H} = \mathcal{H}_{k,j}$ as follows.

Definition 2.2. The vertex set of $\mathcal{H}_{k,j}$ is $\binom{[n]}{j}$ and the edges are those $\binom{k}{j}$-element subsets of $V(\mathcal{H}_{k,j})$ that consist of all j-element subsets of some k-element set from $[n]$.

Thus, we have reduced a question about j-tuple-connectedness in a k-uniform hypergraph to one about vertex-connectedness in an appropriately defined auxiliary hypergraph $H \subseteq \mathcal{H}_{k,j}$. In the following we will analyse Algorithm 1 when applied to $\mathcal{H}_{k,j}$ and $H = (\mathcal{H}_{k,j})_p$.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

At several points in this paper we will want to calculate an upper bound on the number of edges found in some subset of the DFS process, *e.g.* the set of queried edges that contain a given vertex v or, generally, a given ℓ-set L. It will often be convenient to simplify such situations by allowing some additional queries which are not actually made within this subset (or possibly within the DFS process at all). Formally, we couple the subset of the DFS with a number of dummy variables which are also i.i.d. Bernoulli random variables with probability p, and which mimic these additional queries. Then the number of '1's in the subset we consider is certainly at most the number of '1's in the subset together with the dummy variables. In what follows we shall therefore assume the existence of these extra queries without mentioning the formal interpretation.

2.3. Chernoff bounds

We will use the following version of the Chernoff bound from [12, Theorem 2.1].

Theorem 2.3. *Let X be the sum of t i.i.d. Bernoulli random variables with mean p. Then for $a \geqslant 0$,*

$$\mathbb{P}[X \geqslant \mathbb{E}(X)+a] \leqslant \exp\left(-\frac{a^2}{2(tp+a/3)}\right),$$
$$\mathbb{P}[X \leqslant \mathbb{E}(X)-a] \leqslant \exp\left(-\frac{a^2}{2tp}\right).$$

3. Before the phase transition

First we prove that when p is not too large, all the components in a generalised random hypergraph on N vertices have size $O(\log N/\varepsilon^2)$. We start with an auxiliary lemma [14, Lemma 1].

Lemma 3.1. *Let $M \in \mathbb{N}$, $\varepsilon \in (0,1)$, $c \in \mathbb{R}$ and let $(X_i)_{i\in[M]}$ be i.i.d. random Bernoulli variables with mean p. If $p \leqslant (1-\varepsilon)/c$ and $t \geqslant 9c\log M/\varepsilon^2$, then with probability at least $1 - M\exp(-\varepsilon^2 t/(3c)) \geqslant 1 - 1/M^2$, the sum of any t consecutive X_i is less than $t/c - 1$.*

Proof. For $p = (1-\varepsilon)/c$ we have

$$\mathbb{E}\left(\sum_{i=t_0}^{t_0+t-1} X_i\right) = (1-\varepsilon)\frac{t}{c}.$$

We apply Theorem 2.3 to bound the probability that the sum within a fixed interval of length t is at least

$$\frac{t}{c} - 1 = \mathbb{E}\left(\sum_{i=t_0}^{t_0+t-1} X_i\right) + \varepsilon\frac{t}{c} - 1,$$

that is,

$$\mathbb{P}\left[\sum_{i=t_0}^{t_0+t-1} X_i > \frac{t}{c} - 1\right] \leqslant \exp\left(-\frac{(\varepsilon(t/c)-1)^2}{2((1-\varepsilon)t/c + \varepsilon t/(3c))}\right) < \exp\left(\frac{-\varepsilon^2 t}{3c}\right) \leqslant 1/M^3.$$

The union bound over all possible intervals gives the claim. □

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Theorem 3.2. *Let $\mathcal{H}$ be an ℓ-uniform hypergraph on N vertices with maximum degree $\Delta = \Delta(N)$, and let $\varepsilon = \varepsilon(N) > 0$. Then for $p = p(N) \leqslant (1-\varepsilon)/((\ell-1)\Delta)$, the random hypergraph $\mathcal{H}_p$ has a.a.s. (as $N \to \infty$) only vertex-connected components of size at most $9(\ell-1)\log(\Delta N/\ell)/\varepsilon^2$.*

Proof. We couple $\mathcal{H}_p$ with a sequence $(X_i)_{i\in[e(\mathcal{H})]}$ of i.i.d. Bernoulli variables with mean p, as described in Section 2.2 (observe: $e(\mathcal{H}) \leqslant \Delta N/\ell$). We run Algorithm 1 to explore the hypergraph $\mathcal{H}_p$.

If there exists a component $C \subseteq V(\mathcal{H})$ of size at least $m := 9(\ell-1)\log(\Delta N/\ell)/\varepsilon^2$, then it is found during some while-loop. Indeed, we must have found at least $(m-1)/(\ell-1)$ edges in the component after having explored at most m vertices, therefore having made at most $m\Delta$ queries during this while-loop. In other words, there exists an interval of at most $m\Delta$ queries in the process, of which at least $(m-1)/(\ell-1)$ were successful. But Lemma 3.1 applied with $t = m\Delta$, with $M = \Delta N/\ell \geqslant e(\mathcal{H})$ and with $c = (\ell-1)\Delta$ shows that the probability that there exists any such interval is at most $1/M^2 = \ell^2/(\Delta N)^2 \overset{N\to\infty}{\longrightarrow} 0$. □

From Theorem 3.2 we immediately obtain the cases of Theorems 1.1 and 1.2 when $p \leqslant (1-\varepsilon)p_{k,j}$.

Corollary 3.3. *Let $k, j \in \mathbb{N}$ with $k > j$ and $\varepsilon = \varepsilon(n) > 0$ be given. If*

$$p \leqslant (1-\varepsilon)\frac{(k-j)!}{\binom{k}{j}-1}n^{j-k},$$

then a.a.s. the j-tuple-connected components of the random k-uniform hypergraph $H^k(n,p)$ have size $O(\varepsilon^{-2}\log n)$.

Proof. We define $\mathcal{H} = \mathcal{H}_{k,j}$ as in Definition 2.2. Thus, every $S \in V(\mathcal{H})$ has degree $\deg_{\mathcal{H}}(S) = \binom{n-j}{k-j}$, implying that if

$$p \leqslant (1-\varepsilon)\frac{(k-j)!n^{j-k}}{\binom{k}{j}-1} \leqslant (1-\varepsilon)\frac{1}{\left(\binom{k}{j}-1\right)\binom{n}{k-j}},$$

then a.a.s. $\mathcal{H}_p$ has components of size at most

$$\frac{9}{\varepsilon^2}\left(\binom{k}{j}-1\right)\log\left(\frac{\binom{n-j}{k-j}\binom{n}{j}}{\binom{k}{j}-1}\right) \leqslant \frac{9}{\varepsilon^2}\binom{k}{j}\log(n^k) = O(\varepsilon^{-2}\log n).$$

Therefore, the random hypergraph $H^k(n,p)$ has, for the same p, j-tuple-connected components of size at most $O(\varepsilon^{-2}\log n)$ a.a.s. Thus, the assertions of Theorems 1.1 and 1.2 follow when $p \leqslant (1-\varepsilon)p_{k,j}$. □

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

4. After the phase transition

4.1. Algorithm 2

For the regime when $p \geqslant (1+\varepsilon)p_{k,j}$, we slightly alter our algorithm in that in the main *if*-condition during the while-loop we only consider those unqueried edges $e \in E(\mathcal{H})$ such that $e \setminus \{x\}$ consists of only *neutral* vertices. Thus, when an edge in H (during some while-loop) is found, we get $\binom{k}{j} - 1$ new active vertices. We refer to this algorithm as *Algorithm 2*. Observe that in this case we may not fully discover the j-tuple-connected components, but if we find a sufficiently large partial component in this way, then this clearly gives a lower bound on the size of the largest component.

4.2. Vertex-connectedness

First we look at the case of vertex-connectedness in the random k-uniform hypergraph. As mentioned above, Algorithm 2 gives a lower bound on the size of the largest component (indeed, it constructs a 'hypertree' in $H^k(n,p)$ of this size). Furthermore, since in $H^k(n,p)$ for $p = (1+\varepsilon)(k-2)!/n^{k-1}$ the expected number of pairs of edges sharing at least two vertices is $O(1)$, we expect that even after removing all such pairs of edges, most of the largest component remains connected via loose paths.

We will need the following martingale result, an asymmetric version of the Hoeffding inequality proved by Bohman [6].

Lemma 4.1 (Lemmas 6 and 7 from [6]). *Suppose $0 = Y_0, Y_1, \dots, Y_m$ is a martingale in which $-c \leqslant Y_i - Y_{i-1} \leqslant C$ for all $1 \leqslant i \leqslant m$ and some real numbers $c, C > 0$ with $c \leqslant C/10$. Then for every $0 < a < cm$,*

$$\mathbb{P}(|Y_m| \geqslant a) \leqslant 2\exp\left(\frac{-a^2}{3cCm}\right).$$

We use Lemma 4.1 to prove the following concentration result.

Lemma 4.2. *Let $j < k \in \mathbb{N}$ and let $X_1, \dots, X_{\alpha n^k}$ be i.i.d. Bernoulli random variables with parameter $p \leqslant k! n^{j-k}$. Suppose $\alpha = \alpha(n)$ is such that $\alpha^3 n^j \to \infty$. Then with high probability, for every $1 \leqslant t \leqslant \alpha n^k$ we have*

$$\left|\sum_{i=1}^{t} X_i - pt\right| \leqslant \alpha^2 n^j.$$

Note that the concentration given by this lemma is useless for very small t. However, we will only need to apply it for $t = \Theta(\alpha n^k)$, where it gives a better concentration than that which would be given by applying a Chernoff bound and a union bound over all t.

Proof. Let us define a martingale $Y_0, Y_1, \dots, Y_{\alpha n^k}$ as follows:

$$Y_0 := 0,$$
$$Y_{i+1} := \begin{cases} Y_i + X_{i+1} - p & \text{if } |Y_i| \leqslant \alpha^2 n^j, \\ Y_i & \text{otherwise.} \end{cases}$$

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Note that this may be seen as a martingale with a stopping time, where the stopping condition is $|Y_i| > \alpha^2 n^j$. It is easy to check that this is indeed a martingale. Furthermore, we have $-p \leqslant Y_{i+1} - Y_i \leqslant 1-p \leqslant 1$. Therefore, by Lemma 4.1 we have

$$\mathbb{P}(|Y_{\alpha n^k}| > \alpha^2 n^j) \leqslant 2\exp\left(-\frac{(\alpha^2 n^j)^2}{3p\alpha n^k}\right) \leqslant 2\exp(-\alpha^3 n^j/(3k!)) = o(1).$$

Furthermore, note that the conclusion of Lemma 4.2 holds if and only if $|Y_{\alpha n^k}| \leqslant \alpha^2 n^j$, and therefore the above calculation proves the lemma. □

With the above lemma to hand we follow the lines of [14, Theorem 2] to show the following theorem.

Theorem 4.3. *Let $k \in \mathbb{N}$, $k \geqslant 2$ and let $\varepsilon = \varepsilon(n)$ be a function satisfying $\varepsilon = \omega(n^{-1/3})$. If $p = (1+\varepsilon)(k-2)!/n^{k-1}$, then a.a.s. the random hypergraph $H^k(n,p)$ contains $\Omega(\varepsilon n)$ vertices that are pairwise connected by loose paths.*

In particular, $H^k(n,p)$ has a vertex-connected component of size $\Omega(\varepsilon n)$.

Proof. We consider a sequence of i.i.d. Bernoulli random variables coupled with $H^k(n,p)$, as explained in Section 2.2. Our $\mathcal{H}$ is the complete k-uniform hypergraph K_n^k with n vertices.

We choose $\alpha := \varepsilon/(8k!)$.

We will show that between the query $\alpha n^k/2$ and the query αn^k the stack S of active vertices a.a.s. has not been empty, meaning that during this time Algorithm 2 is discovering a single (large) component. Furthermore, Lemma 4.2 implies that a.a.s. the number of X_i that are answered as 1 between the query $\alpha n^k/2$ and αn^k is at least

$$(p\alpha n^k - \alpha^2 n) - (p\alpha n^k/2 + \alpha^2 n) \geqslant (k-2)!\alpha n/2 - 2\alpha^2 n \geqslant \frac{\varepsilon}{16k^2} n.$$

Together, these two facts yield the assertion of Theorem 4.3, since there are still some unexplored vertices, and therefore Algorithm 2 has found at least $\varepsilon n/(16k^2)$ edges in some component and each such edge makes $k-1$ previously neutral vertices active, which results in a component of size $\Omega(\varepsilon n)$.

So let us assume for a contradiction that the high probability event of Lemma 4.2 holds, but that after some t queries where $t \in \{\alpha n^k/2, \dots, \alpha n^k\}$, the stack S is empty. Then Algorithm 2 has discovered $pt \pm \alpha^2 n$ edges in $H^k(n,p)$. Since with each explored edge, $k-1$ vertices become active, and after emptying the stack S all active vertices are explored, this implies that if s edges have been found, then at least $s(k-1)+1$ vertices are explored. Observe that when the stack S is empty there are only explored and neutral vertices. Further, if s' vertices are explored then Algorithm 2 must have made (at least) $s'\binom{n-s'}{k-1}$ queries. Further observe that this function is increasing for $s' \leqslant n/(k+2) - 1$. We estimate how many edges have been queried at time $t \leqslant \alpha n^k$. This number is at least

$$\begin{aligned}(pt - \alpha^2 n)(k-1)&\binom{n - (pt-\alpha^2 n)(k-1)}{k-1} \\ &\geqslant \frac{pt-\alpha^2 n}{(k-2)!}(n - pt(k-1))^{k-1}\end{aligned}$$

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

$$\geqslant \frac{pt-\alpha^2 n}{(k-2)!}n^{k-1}(1-(1+\varepsilon)(k-1)!\alpha)^{k-1}$$
$$\geqslant (1+\varepsilon/2)t(1-\varepsilon/8) > t$$

for $\alpha \leqslant \varepsilon/(8k!)$. However, this is a contradiction since we assumed that only t queries have been made so far.

Therefore, for large enough n, the probability that the stack becomes empty between queries $\alpha n^k/2$ and αn^k is at most the error probability in Lemma 4.2. □

Remark 4.4. Similarly to the results in [14] we can show that the large component contains a loose path of length $\Omega_k(\varepsilon^2 n)$ in $H^k(n,p)$ for $p=(1+\varepsilon)(k-2)!/n^{k-1}$. Roughly speaking, the argument is as follows. We have already shown that the stack of active vertices does not become empty between times $\alpha n^k/2$ and αn^k. On the other hand, if the set of active vertices is small enough ($\Theta(\varepsilon^2 n)$ will do), then this will not affect the previous calculations significantly. We can therefore deduce that the set of active vertices never becomes smaller than $\Theta(\varepsilon^2 n)$ in this time interval. But because we are exploring via a depth-first search process, the set of active vertices automatically lies in a loose path (possibly with some explored vertices to complete the edges).

4.3. j-tuple-connectedness

Our aim in this section is to prove Theorem 1.2. For the remainder of this section we therefore fix δ and $\varepsilon=\varepsilon(n)$ as in Theorem 1.2. We will also assume that $n \geqslant n_0$ for some sufficiently large constant n_0 which we do not determine explicitly (but which is implicitly dependent on k, j and δ). Let $\alpha=\alpha(n)$ satisfy

$$\frac{\varepsilon}{32k!2^j C} \geqslant \alpha = \omega(n^{\delta-1}+n^{-j/3}),$$

where C is a constant depending only on k, j which we determine implicitly later. (We note that for the purposes of this paper, setting $\alpha=\varepsilon/(32k!2^jC)$ would be sufficient. However, in [10] we will need to quote Lemma 4.5 for a wider range of α.) We first give an outline of the main ideas of the proof.

Proof sketch of Theorem 1.2. We consider $\mathcal{H}=\mathcal{H}_{k,j}$ as defined in Definition 2.2. As explained in Section 2.2, the random $\binom{k}{j}$-uniform hypergraph $\mathcal{H}_p$ we consider has a natural correspondence with $H^k(n,p)$. We perform Algorithm 2 as described above, that is, only when all vertices but one are neutral do we query an edge in $\mathcal{H}$. As in Theorem 4.3 we shall estimate the number of queries made given that the stack S is emptied between $\alpha n^k/2$ and αn^k queries.

This time, however, we need to take account of the fact that (since $\mathcal{H}$ is clearly *not the complete* hypergraph) not every explored vertex in $\mathcal{H}$ forms an already queried edge with any $\binom{k}{j}-1$ neutral vertices in $\mathcal{H}$. This is because vertices of $\mathcal{H}$ are j-element subsets of $[n]$ and edges correspond to only those $\binom{k}{j}$ j-sets whose union gives a k-element set. Therefore, we will need to keep track of the already discovered j-sets of $H^k(n,p)$. More precisely, let G_j be the j-uniform hypergraph on vertex set $[n]$ whose edges are the j-sets which have been discovered by Algorithm 2 up to time αn^k (recall that the j-sets are vertices in Algorithm 2). We need to bound the degrees of sets of vertices in G_j. Suppose for the moment that we are able to show Lemma 4.5 below, stating that a.a.s. G_j has small maximum degrees depending on α. Then from

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

each j-set we have made at least $\binom{n-j}{k-j}(1-O(\alpha))$ queries. Furthermore, we know that we have found approximately pt edges, from each of which we discovered $\binom{k}{j}-1$ new j-sets. Thus the number of queries is at least

$$pt\left(\binom{k}{j}-1\right)\frac{n^{k-j}}{(k-j)!}(1-O(\alpha))=(1+\varepsilon)(1-O(\alpha))t>t$$

for α sufficiently small compared to ε. But this is a contradiction since at time t we have made exactly t queries. This argument will be given in more detail at the end of this section.

Bounding the maximum degree of G_j. Let $G_j(t)$ denote the j-uniform hypergraph on vertex set $[n]$ whose edges are the discovered j-sets at time t (so $G_j=G_j(\alpha n^k)$). For each $1\leqslant \ell<j$, let $\Delta_\ell(G_j(t))$ denote the maximum ℓ-degree of this hypergraph, *i.e.* the maximum over all ℓ-sets of the number of edges of $G_j(t)$ containing this ℓ-set. For convenience, we sometimes use $\Delta_0(G_j(t))$ to denote the number of edges in $G_j(t)$ (*i.e.* the natural generalisation for $\ell=0$). The aim of this section is to prove that a.a.s. $G_j(\alpha n^k)$ does not have too large a maximum ℓ-degree for any $0\leqslant \ell\leqslant j-1$.

In fact we prove a slightly stronger statement which also applies to a breadth-first search process. We first define two new breadth-first search algorithms.

- BFS1 is the breadth-first search analogue of Algorithm 1: any edge containing a neutral vertex (which corresponds to a neutral j-set) may be queried. Formally, we change line 8 in the algorithm to 'Let x be the bottom vertex of S'.
- BFS2 is the breadth-first search analogue of Algorithm 2: only edges containing $\binom{k}{j}-1$ neutral vertices (which correspond to neutral j-sets) may be queried.

We analyse the maximum degrees given by each of the algorithms (Algorithm 1, Algorithm 2, BFS1 and BFS2). Since we will never use specific information about which algorithm we are considering, we will go through all the proofs together and simply refer to the 'search algorithm', which may be any one of these four. We still use $G_j(t)$ to refer to the hypergraph that has been found by time t using any one of the algorithms.

Lemma 4.5 (bounded degree lemma). *For some constant C, using any one of Algorithm 1, Algorithm 2, BFS1 or BFS2, with probability at least* $1-\exp(-n^{\delta/2})$,

$$\Delta_\ell(G_j(\alpha n^k))\leqslant C\alpha n^{j-\ell}$$

for all $0\leqslant \ell\leqslant j-1$.

Since we will be considering the structure of $G_j(t)$, from now on we will think of the exploration process as one on j-sets in $H^k(n,p)$, rather than on vertices in $\mathcal{H}_p$ (there is of course a natural correspondence between the two).

In fact, we will prove that $\Delta_\ell(G_j(\alpha n^k))\leqslant C_\ell\alpha n^{j-\ell}$ for each ℓ, for constants C_ℓ which we will determine later, and then we may set $C:=\max_\ell\{C_\ell\}$. Note that by a simple application of the Chernoff bound, the lemma is true for $\ell=0$ if

$$C_0\geqslant 2\frac{(k-j)!}{\binom{k}{j}-1}.$$

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

For $\ell \geqslant 1$, we pick an ℓ-set L and note that there are three ways in which the degree of L in $G_j(t)$ may grow as t increases during the search process.

(1) A *new start* at L occurs when there are no active j-sets (all discovered j-sets have been explored) and the search algorithm picks a new j-set from which to start. If this j-set contains L, then the degree of L in $G_j(t)$ has grown by one. Recall that the search algorithm chooses a j-set uniformly at random among all neutral j-sets: see Algorithm 1.
(2) A *jump* to L occurs when the search process queries a k-set K containing L from a j-set J not containing L (though possibly intersecting L) and the edge K is present. Then for each $A \in \binom{K\setminus L}{j-\ell}$, the j-set $A \cup L$ becomes active (if it wasn't already) and the degree of L in $G_j(t)$ grows by at most $\binom{k-\ell}{j-\ell}$ (this is exact for Algorithm 2 or BFS2).
(3) From an active j-set J containing L we may query a k-set K also containing L. If this forms an edge then for each $A \in \binom{K\setminus L}{j-\ell}$, the j-set $A \cup L$ becomes active (if it wasn't already) and the degree of L in $G_j(t)$ grows by at most $\binom{k-\ell}{j-\ell} - 1$ (this is exact for Algorithm 2 or BFS2). We call this a *branching* at L.

We will bound the contributions to the ℓ-degree $d_L(G_j(t))$ (defined as $|\{J \in E(G_j(t)) : J \supseteq L\}|$) made by each of these possibilities individually. However we must take care to avoid a circular argument, since the bounds are interdependent.

Let $E(t)$ be the event that $\Delta_\ell(G_j(t)) \leqslant C_\ell \alpha n^{j-\ell}$ for all $0 \leqslant \ell < j$. We aim to show that with high probability $E(\alpha n^k)$ holds, which we do by showing that with high probability $E(t-1) \Rightarrow E(t)$ for every $t \leqslant \alpha n^k$. More precisely, we will first prove some probabilistic lemmas, saying that with high probability, various very likely events will hold throughout the search process. The second part of the proof will be deterministic, showing when these good events hold, $E(t-1) \Rightarrow E(t)$ for any $t \leqslant \alpha n^k$, and since $E(0)$ automatically holds, by induction $E(\alpha n^k)$ holds.

Probabilistic lemmas. Let us first consider where the new starts are made. Since we select the j-set for our new start uniformly at random (it corresponds to choosing a new vertex of $\mathcal{H}$, and the ordering of $V(\mathcal{H})$ was chosen randomly), we expect the new starts to be, in some sense, evenly distributed. The next lemma makes this more precise.

Set $m = 2\alpha(k-j)!n^j$ and let m_0 be the minimum of m and the number of new starts made during the first αn^k queries. Let A' be the event that for every $1 \leqslant \ell \leqslant j-1$, every ℓ-set is contained in at most

$$\max\left\{\frac{4mj!}{(j-\ell)!n^\ell}, n^\delta\right\}$$

many j-sets that were chosen to be a new start during the first m_0 new starts.

Let $A^{(1)}$ be the intersection of the event A' and the event

$$\left\{\sum_{i=1}^{\alpha n^k} X_i \leqslant 2p\alpha n^k\right\}.$$

Lemma 4.6. $\mathbb{P}(A^{(1)}) \geqslant 1 - \exp(n^{-\delta/2})$.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Proof. By the Chernoff bound (Theorem 2.3) we have

$$\mathbb{P}\left(\sum_{i=1}^{\alpha n^k} X_i \geqslant 2p\alpha n^k\right) \leqslant \exp\left(-\frac{p\alpha n^k}{3p\alpha n^k}\right) = \exp(-\Theta(\alpha n^j)) \leqslant \exp(-n^{2/3}).$$

Thus we may assume that we have discovered at most $2p\alpha n^k = O(\alpha n^j)$ edges so far, and therefore the number of j-sets which are discovered is at most $m + \left(\binom{k}{j} - 1\right)2p\alpha n^k = O(\alpha n^j)$. Thus, whenever we made a new start so far, we always had at least $\frac{1}{2}\binom{n}{j}$ j-sets available to choose from, and so the probability of picking any one of these was certainly at most $2/\binom{n}{j}$.

Now given any ℓ-set L, the number of j-sets in which L lies is less than $\binom{n}{j-\ell}$. Therefore the number of new starts at a j-set containing L has distribution dominated by

$$\mathrm{Bi}\left(m, 2\binom{n}{j-\ell}\Big/\binom{n}{j}\right),$$

which in turn is dominated by the binomial distribution

$$\mathrm{Bi}\left(\max\{m, n^{\ell+\delta/2}\}, \frac{3j!}{(j-\ell)!n^\ell}\right).$$

By the Chernoff bound, the probability that this is greater than

$$\max\left\{\frac{4mj!}{(j-\ell)!n^\ell}, n^\delta\right\}$$

is at most $\exp(-n^{2\delta/3})$, and a union bound over all ℓ and L gives the lemma. □

We next state an auxiliary lemma, which states that we may 'pick out' certain (random) subsequences of queries and treat them as an interval in the search process. Recall that our sequence of queries gives a sequence of independent Bernoulli random variables $X_1, X_2, \ldots, X_{\binom{n}{k}}$. We will be considering a random subsequence $t_1, t_2, \ldots, t_s$ from $[\binom{n}{k}]$. We say 't_i is determined by the values of $X_1, \ldots, X_{t_i-1}$' to mean the following: for any j, whether the event $\{t_i = j\}$ holds is determined by the values of $X_1, \ldots, X_{j-1}$. In particular this means that t_i is chosen before X_{t_i} is revealed.

Lemma 4.7. *Let $S = (t_1, t_2, \ldots, t_s)$ be a (random, ordered) index set chosen according to some criterion such that:*

- *t_i is determined by the values of $X_1, \ldots, X_{t_i-1}$,*
- *with probability 1 we have $1 \leqslant t_1 < t_2 < \ldots < t_s \leqslant \binom{n}{k}$.*

Then $(X_{t_1}, \ldots, X_{t_s}) \sim (Y_1, \ldots, Y_s)$, where $Y_1, \ldots, Y_s$ are independent $Be(p)$ variables. In particular, we may apply a Chernoff bound to $\sum_{i \in S} X_i$.

We omit the proof of this simple and intuitively obvious result. We will apply Lemma 4.7 to prove two further probabilistic lemmas.

For any $x \in \mathbb{N}$, any $1 \leqslant \ell \leqslant j-1$ and any ℓ-set L, let $S(x, L)$ be the set of the first x times at which we make a query which could result in a jump to L.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Let $A_L^{(2)}(x)$ be the event that these queries result in at most $2px$ edges (*i.e.* $\sum_{i\in S(x,L)} X_i \leqslant 2px$). Further, let $A^{(2)}$ be the the intersection of all the events $A_L^{(2)}(x)$ over all choices of ℓ, L and $x \geqslant n^{k-j+\delta}$.

Lemma 4.8. *For any $n^{k-j+\delta} \leqslant x \in \mathbb{N}$, for any $1 \leqslant \ell \leqslant j-1$ and any ℓ-set L, we have*

$$\mathbb{P}(A_L^{(2)}(x)) \geqslant 1 - \exp(-n^{2\delta/3}).$$

Furthermore, $\mathbb{P}(A^{(2)}) \geqslant 1 - \exp(-n^{\delta/2})$.

Proof. We apply Lemma 4.7 to bound the number of jumps to L within $S(x,L)$. Thus

$$\begin{aligned}\mathbb{P}(A_L^{(2)}(x)) &\geqslant 1 - \exp\left(-\frac{(px)^2}{3px}\right)\\ &\geqslant 1 - \exp\left(-\frac{n^\delta (k-j)!}{3\left(\binom{k}{j}-1\right)}\right)\\ &\geqslant 1 - \exp(-n^{2\delta/3}).\end{aligned}$$

For the last statement, we take a union bound over all $\sum_{\ell=1}^{j-1}\binom{n}{\ell} \leqslant n^j$ possible choices of L and all choices of x (observing that x is certainly at most $\binom{n}{j} \leqslant n^j$). We therefore obtain

$$\mathbb{P}(A^{(2)}) \geqslant 1 - n^{2j}\exp(-n^{2\delta/3}) \geqslant 1 - \exp(-n^{\delta/2})$$

as required. □

We now aim to prove something similar for the number of branchings at a set L of size ℓ. Fix L and consider a neighbourhood branching process at L. More precisely, given a j-set J containing L, we make a number of queries in the search process and whenever we discover an edge, at most further $\binom{k-\ell}{j-\ell} - 1$ j-sets containing L become active (these are considered children of the original j-set). For an upper bound we assume exactly $\binom{k-\ell}{j-\ell} - 1$ j-sets become active.

By deleting L from each of the sets we consider, we may view this as a search process on $(j-\ell)$-sets starting at $J \setminus L$ in a $(k-\ell)$-uniform hypergraph. This may not correspond to a simple time interval in the branching process, but we pick out only those queries which are made from a j-set containing L (this is permissible by Lemma 4.7). The hypergraph in which this search process takes place has $n-\ell$ vertices, but for an upper bound we replace this by n. Furthermore, we ignore the fact that some j-sets may already have been discovered some other way, and are therefore not neutral within this search process. If we further assume that from any $(j-\ell)$-set in the process we may still query $\binom{n}{k-j}$ many $(k-j)$-sets (effectively ignoring the fact that we may have seen some before), then we may consider the process no longer as a hypergraph process, but as an abstract branching process in which the number of children has distribution $r\cdot \mathrm{Bi}\left(\binom{n}{k-j}, p\right)$, where $r = r(k,j,\ell) = \binom{k-\ell}{j-\ell} - 1$. (By the notation $a\cdot X$, for a real number a and real-valued probability distribution X, we mean the probability distribution given by $\mathbb{P}(a\cdot X = ai) = \mathbb{P}(X = i)$ for any real number i.)

For a probability distribution Q, let T_Q be the tree of a branching process starting at a single vertex in which each vertex has number of children with distribution Q independently. Given an

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

integer x, define $T_Q(x)$ to be the union of x independent copies of T_Q (or equivalently, the forest of a branching process starting with x vertices with offspring distribution Q).

Let us define, for each $1 \leqslant \ell \leqslant j-1$,

$$c_\ell := \frac{1}{2} + \frac{1}{2}\frac{\binom{k-\ell}{j-\ell} - 1}{\binom{k}{j} - 1} < 1$$

and observe that

$$\max_{1 \leqslant \ell \leqslant j-1} c_\ell = c_1.$$

Let

$$C^\dagger = C^\dagger(k,j) := \frac{8c_1}{(1-c_1)^2} \geqslant \frac{4}{1-c_1}.$$

For any $n^\delta \leqslant x \in \mathbb{N}$, for any $1 \leqslant \ell \leqslant j-1$ and for any ℓ-set L, let $A_L^{(3)}(x)$ be the event that the first x neighbourhood branching processes started at L result in at most $C^\dagger x$ branchings.

Let $A^{(3)}$ be the the intersection of all the events $A_L^{(3)}(x)$ over all choices of L and $x \geqslant n^\delta$.

Lemma 4.9. *For any $n^\delta \leqslant x \in \mathbb{N}$, for any $1 \leqslant \ell \leqslant j-1$ and for any ℓ-set L, with probability at least $1-\exp(-x) \geqslant 1-\exp(-n^\delta)$, the event $A_L^{(3)}(x)$ holds. Furthermore, with probability at least $1-\exp(-n^{\delta/2})$, the event $A^{(3)}$ holds.*

Proof. For an upper bound, we may model the x neighbourhood branching processes as

$$T \sim T_{r\cdot \mathrm{Bi}(\binom{n}{k-j},p)}(x).$$

We consider exploring this branching process via a search process (either depth- or breadth-first search will do here), and can thus couple the branching process with a (possibly infinite) sequence of independent Bernoulli(p) variables $Y_1, Y_2, Y_3, \ldots$, such that each variable which takes the value 1 corresponds to a set of r children being discovered.

In order for T to have total size at least $C^\dagger x$, the first $C^\dagger x$ vertices which we explore in the search process must have at least $(C^\dagger - 1)x$ children in total. Thus the first $C^\dagger x\binom{n}{k-j}$ of the Y_i would have to contain at least $(C^\dagger - 1)x/r$ many 1s. Let us observe that the expected number of 1s in this interval is

$$C^\dagger x\binom{n}{k-j}p \leqslant C^\dagger x\frac{1+\varepsilon}{\binom{k}{j}-1} \leqslant C^\dagger x\frac{c_\ell}{r}$$

for ε small enough (recalling that $r = \binom{k-\ell}{j-\ell} - 1$). Thus, using the Chernoff bound (Theorem 2.3) with

$$a = (C^\dagger - 1)\frac{x}{r} - C^\dagger x\binom{n}{k-j}p \geqslant \frac{x}{r}(C^\dagger - 1 - C^\dagger c_\ell) \geqslant \frac{C^\dagger x(1-c_\ell)}{2r},$$

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

we obtain the probability bound

$$\begin{aligned}\mathbb{P}(|T| \geqslant C^\dagger x) &\leqslant \mathbb{P}\left(\mathrm{Bi}\left(C^\dagger x\binom{n}{k-j}, p\right) \geqslant (C^\dagger - 1)\frac{x}{r}\right)\\ &\leqslant \exp\left(-\left(\frac{C^\dagger x(1-c_\ell)}{2r}\right)^2 \Big/ 2\left(\frac{C^\dagger x c_\ell}{r} + \frac{C^\dagger x(1-c_\ell)}{6r}\right)\right)\\ &= \exp\left(-\frac{C^\dagger x}{8r}\cdot\frac{(1-c_\ell)^2}{(5/6)c_\ell + 1/6}\right)\\ &\leqslant \exp\left(-\frac{C^\dagger x(1-c_1)^2}{8c_1}\right)\\ &\leqslant \exp(-x),\end{aligned}$$

where the last line holds since $C^\dagger \geqslant 8c_1/(1-c_1)^2$.

This proves the first part of the lemma, and for the second part we simply take a union bound over all choices of ℓ, L and x (of which there are certainly at most n^{2j} in total). □

Finally, let

$$A^{(\mathrm{all})} := A^{(1)} \wedge A^{(2)} \wedge A^{(3)}.$$

The following is an immediate corollary of Lemmas 4.6, 4.8 and 4.9.

Corollary 4.10. $\mathbb{P}(A^{(all)}) = 1 - 3\exp(-n^{\delta/2})$. □

Inductive proof. Recall that $E(t)$ is the event that $\Delta_\ell(G_j(t)) \leqslant C_\ell \alpha n^{j-\ell}$ for all $0 \leqslant \ell < j$. In this section we will show that $A^{(\mathrm{all})} \Rightarrow E(\alpha n^k)$. More precisely, we prove that $A^{(\mathrm{all})} \Rightarrow E(t)$ for all $t \leqslant \alpha n^k$ by induction on t.

- Let $d_L^{(1)}(t)$ be the number of new starts at L by time t and let $D_\ell^{(1)}(t) := \max d_L^{(1)}(t)$, where the maximum is over all sets L of size ℓ.
- Let $d_L^{(2)}(t)$ be the number of jumps to L by time t and let $D_\ell^{(2)}(t) := \max d_L^{(2)}(t)$, where the maximum is over all sets L of size ℓ.
- Let $d_L^{(3)}(t)$ be the number of branchings at L up to time t and let $D_\ell^{(3)}(t) := \max d_L^{(3)}(t)$, where the maximum is over all sets L of size ℓ.

Let $\hat{C}_0 := 2$, $C_0^* := 2$ and recursively define

$$\begin{aligned}C_\ell &:= \max\left\{\hat{C}_\ell + C_\ell^* + \frac{8j!(k-j)!}{(j-\ell)!}, C_{\ell-1}\right\},\\ \hat{C}_{\ell+1} &:= \max\left\{2^{\ell+2}\frac{(k-j)!}{\binom{k}{j}-1}C_\ell, 8\right\},\\ C_{\ell+1}^* &:= 2k!\hat{C}_{\ell+1}C^\dagger,\end{aligned}$$

for $\ell \geqslant 0$, where $C^\dagger$ is the constant from Lemma 4.9. (For the sake of the definition of C_0, we adopt the convention that $C_{-1} = 0$.)

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

- Let $E^{(1)}(t)$ be the event that for each $0 \leqslant \ell < j$,
$$D_\ell^{(1)}(t) \leqslant \frac{8j!(k-j)!}{(j-\ell)!}\alpha n^{j-\ell}.$$
- Let $E^{(2)}(t)$ be the event that for each $0 \leqslant \ell < j$, $D_\ell^{(2)}(t) \leqslant \hat{C}_\ell \alpha n^{j-\ell}$.
- Let $E^{(3)}(t)$ be the event that for each $0 \leqslant \ell < j$, $D_\ell^{(3)}(t) \leqslant C_\ell^* \alpha n^{j-\ell}$.

We further define $E^*(t) := E^{(1)}(t) \wedge E^{(2)}(t) \wedge E^{(3)}(t)$. Note that since

$$C_\ell \geqslant \hat{C}_\ell + C_\ell^* + \frac{8j!(k-j)!}{(j-\ell)!}$$

we have $E^*(t) \Rightarrow E(t)$. We will actually prove that $A^{(\mathrm{all})} \Rightarrow E^*(t)$ for all $t \leqslant \alpha n^k$ by induction on t. The base case is trivial, since $E^*(0)$ holds with probability 1.

We aim to show that if $A^{(\mathrm{all})}$ holds, then none of $E^{(1)}(t), E^{(2)}(t), E^{(3)}(t)$ can be the first to fail before time αn^k. However, we must be careful with the time steps since it may be that two of these events become false simultaneously.

Lemma 4.11. *$A^{(1)} \wedge E(t) \Rightarrow A^{(1)} \wedge E^{(1)}(t+1)$ for $t \leqslant \alpha n^k$.*

Proof. That $A^{(1)} \wedge E(t) \Rightarrow A^{(1)}$ is immediate, so we only need to show that $A^{(1)} \wedge E(t) \Rightarrow E^{(1)}(t+1)$. Note that by $E(t)$, we have $\Delta_\ell(G_j(t)) \leqslant C_\ell \alpha n^{j-\ell}$ for all $0 \leqslant \ell \leqslant j-1$. Thus, for each j-set we have made at least

$$\binom{n-j}{k-j} - \sum_{\ell=0}^{j-1} \binom{j}{\ell} \Delta_\ell(G_j(t)) \binom{n-2j+\ell}{k-2j+\ell} = (1-O(\alpha)) \binom{n}{k-j}$$

queries. Thus, the number of new starts we have made is certainly at most

$$\frac{\alpha n^k}{(1-O(\alpha))\binom{n}{k-j}} \leqslant 2\alpha(k-j)!n^j,$$

which is precisely the m in the definition of $A^{(1)}$. Therefore by $A^{(1)}$, for any ℓ-set L we have made at most $(8j!(k-j)!/(j-\ell)!)\alpha n^{j-\ell}$ new starts at L, as required. □

Lemma 4.12. *$A^{(2)} \wedge E(t) \Rightarrow A^{(2)} \wedge E^{(2)}(t+1)$ for $t \leqslant \alpha n^k$.*

Proof. Similarly to Lemma 4.11, it is enough to show that $A^{(2)} \wedge E(t) \Rightarrow E^{(2)}(t+1)$

Given an ℓ-set L, we consider the number of jumps to L by time t. For each $0 \leqslant i < \ell$, the number of queries to L from j-sets which intersect L in a set I of i vertices is certainly at most $\Delta_i(G_j(t)) \leqslant C_i \alpha n^{j-i}$ (since $E(t)$ holds, we can bound the number of j-sets which have been active and contain I). We have $\binom{\ell}{i}$ such sets I, and for each of these, if we are to jump to L we have already chosen $j+\ell-i$ vertices, and therefore have at most $\binom{n}{k-j-\ell+i}$ choices for the remaining vertices. Thus the total number of queries by time t which may have resulted in jumps to L is at most

$$\sum_{i=0}^{\ell-1} \binom{\ell}{i} C_i \alpha n^{j-i} \binom{n}{k-j-\ell+i} \leqslant 2^\ell C_{\ell-1} \alpha n^{k-\ell}.$$

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

Thus by $A^{(2)}$, the number of jumps to L is at most

$$2\cdot 2^{\ell}C_{\ell-1}\alpha n^{k-\ell}p = (1+\varepsilon)2^{\ell+1}C_{\ell-1}\alpha n^{j-\ell}\frac{(k-j)!}{\binom{k}{j}-1}$$
$$\leqslant \hat{C}_{\ell}\alpha n^{j-\ell}.$$

Since L was chosen arbitrarily, this holds for all L, and therefore $E^{(2)}(t+1)$ is satisfied, as required. □

Lemma 4.13. *$A^{(3)}\wedge E^{*}(t)\Rightarrow A^{(3)}\wedge E^{(3)}(t+1)$ for $t\leqslant \alpha n^{k}$.*

Proof. Since we assume that $E^{*}(t)$ holds, the number of neighbourhood branching processes which we start at a set L is at most the number of new starts at L plus $\binom{k-\ell}{j-\ell}$ times the number of jumps to L, or at most

$$\left(\frac{8j!(k-j)!}{(j-\ell)!}+\binom{k-\ell}{j-\ell}\hat{C}_{\ell}\right)\alpha n^{j-\ell}\leqslant 2k!\hat{C}_{\ell}\alpha n^{j-\ell}.$$

For an upper bound, we will assume that we have exactly $2k!\hat{C}_{\ell}\alpha n^{j-\ell}\geqslant n^{\delta}$ neighbourhood branching processes. Then by $A_L^{(3)}(2k!\hat{C}_{\ell}\alpha n^{j-\ell})$, the total number of vertices in all of these branching processes is at most $2k!C^{\dagger}\hat{C}_{\ell}\alpha n^{j-\ell}$ as required.

Since L was chosen arbitrarily, this holds for any L, and thus $E^{(3)}(t+1)$ holds. □

Now combining Lemmas 4.11, 4.12 and 4.13, we have that for $t\leqslant \alpha n^{k}$

$$A^{(\mathrm{all})}\wedge E^{*}(t)\Rightarrow A^{(\mathrm{all})}\wedge E^{(1)}(t+1)\wedge E^{(2)}(t+1)\wedge E^{(3)}(t+1)$$
$$\Leftrightarrow A^{(\mathrm{all})}\wedge E^{*}(t+1).$$

Since $E^{*}(0)$ holds trivially, by induction we may deduce that $A^{(\mathrm{all})}\Rightarrow E^{*}(\alpha n^{k})\Rightarrow E(\alpha n^{k})$, and therefore

$$\mathbb{P}(E(\alpha n^{k}))\geqslant \mathbb{P}(A^{(\mathrm{all})})\geqslant 1-3\exp(-n^{\delta/2})$$

as required. This completes the proof of Lemma 4.5. □

Proof of Theorem 1.2. We now complete the proof by filling in the details of the argument sketched earlier. We now choose $\alpha=\varepsilon/(32k!2^{j}C)$.

We assume that the stack S is empty at some time $t\in\{\alpha n^{k}/2,\dots,\alpha n^{k}\}$ (and thus there is a new while-loop between queries $\alpha n^{k}/2$ and αn^{k}). Thus we can estimate, using Lemma 4.2, that a.a.s. at least $tp-\alpha^{2}n^{j}$ edges have been found by time t. Recall that since we run Algorithm 2, whenever an edge appears, we discover $\binom{k}{j}-1$ new j-sets of vertices, so

$$e(G_j(t))\geqslant \left(\binom{k}{j}-1\right)(tp-\alpha^{2}n) \text{ a.a.s.}$$

We note that

$$\frac{\alpha^{2}n^{j}}{tp}\leqslant \frac{2\left(\binom{k}{j}-1\right)\alpha}{(k-j)!}\leqslant k!\alpha\leqslant \varepsilon/4$$

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

and so

$$\left(\binom{k}{j}-1\right)(tp-\alpha^2 n^j) \geqslant \left(\binom{k}{j}-1\right)tp(1-\varepsilon/4).$$

Furthermore, we know from Lemma 4.5 that $\Delta_\ell(G_j(t)) \leqslant C\alpha n^{j-\ell}$ a.a.s. Since every k-subset of $[n]$ which contains exactly one j-set which is an edge of $G_j(t)$ and $\binom{k}{j}-1$ not in $G_j(t)$ (the stack S is empty) must have been queried at this time, we infer that at time t (a.a.s.) at least

$$\begin{aligned}
&\left(\binom{k}{j}-1\right)tp(1-\varepsilon/4)\left(\binom{n-j}{k-j}-\sum_{\ell=0}^{j-1}\binom{j}{\ell}\Delta_\ell(G_j(t))\binom{n-2j+\ell}{k-2j+\ell}\right)\\
&\quad > t(1+3\varepsilon/5)\frac{(k-j)!}{n^{k-j}}\left(\binom{n-j}{k-j}-2^j C\alpha n^{k-j}\right)\\
&\quad > (1+\varepsilon/2)(1-2^j(k-j)!C\alpha)t
\end{aligned}$$

queries were made. This is larger than t if $\alpha \leqslant \varepsilon/(2^{j+2}(k-j)!C)$ and therefore we obtain a contradiction (since up to this time only t queries have been made). Thus, between $\alpha n^k/2$ and αn^k the stack remains non-empty, which again implies by Lemma 4.2 that at least $\alpha pn^k/2 - 2\alpha^2 n^j$ edges are in some j-tuple-connected component, which therefore contains at least

$$\left(\binom{k}{j}-1\right)(\alpha pn^k/2 - 2\alpha^2 n^j) = \Omega(\varepsilon n^j)$$

j-sets. This completes the proof of Theorem 1.2. □

Remark 4.14. As we did for vertex-connectedness, we could modify our calculations to prove that with high probability the set of active j-sets does not become small between times $\alpha n^k/2$ and αn^k. However, for $j > 1$ the set of active j-sets does *not* automatically form a j-tight path since they could, for example, all contain one vertex. We would, however, obtain a long j-tight walk which is non-repeating in the sense that a j-set is only visited once in the walk.

Remark 4.15. The most difficult part of the proof, the bounded degree lemma (Lemma 4.5), explicitly allowed the search process to be a breadth-first search rather than a depth-first search. In fact, the rest of the proof would also work equally well for a breadth-first search. The only point at which we actually need a depth-first search process is in Remarks 4.4 and 4.14, where we note that the set of active vertices forms either a path or a j-tight walk. The breadth-first search algorithm is used in [10].

5. Concluding remarks

For $p = (1+\varepsilon)p_{k,j}$, a natural conjecture is that a *unique* largest component of size $\Omega(\varepsilon n^j)$ should exist with high probability for any ε such that $\varepsilon^3 n^j \to \infty$. In this paper, we have the additional condition that $\varepsilon \gg n^{\delta-1}$ (for some $\delta > 0$). For $j = 1, 2$, this condition is already implied by $\varepsilon^3 n^j \to \infty$, so in these cases our range of ε is best possible. However, once $j \geqslant 3$, the condition $\varepsilon \gg n^{\delta-1}$ takes over.

The extra condition arises because of our proof method: in the bounded degree lemma, we wish to show that degrees which we expect to have size $\Theta(\varepsilon n^{j-\ell})$ do not exceed their expected

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press

size by more than a constant factor (a.a.s.). For this to be plausible, we certainly need $\Theta(\varepsilon n^{j-\ell})$ to be large, which for $\ell = j-1$ leads to the extra condition on ε. If one were to attempt to remove this condition while still using this proof method, presumably some information about the distribution of degrees (which may now be small) would be required.

We have shown here that the largest component has size $\Omega(\varepsilon n^j)$, which for constant ε is certainly the correct order of magnitude. In [10], the asymptotic size of the largest component is determined and its uniqueness (*i.e.* that all other components are much smaller) proved, although the range of ε is slightly more restrictive than that allowed here. The argument in that paper makes fundamental use of the bounded degree lemma from this paper. Independently Lu and Peng [16] also claim to have proved the asymptotic size and uniqueness of the largest component, though only for constant ε.

It would also be interesting to know about the structure of the components and in particular whether there is a simple generalisation of the well-known fact that for graphs all small components (*i.e.* any except the giant component, if it exists) are either trees or unicyclic graphs a.a.s. For the case $j = 1$, results in this direction were obtained in [13, 20].

Finally, one could also study the emergence of the *s-cores* of a random hypergraph. For $1 \leqslant \ell < k$ we have defined the degree of a set of ℓ vertices, and so we have a well-defined notion of minimum ℓ-degree. We can therefore ask when a.a.s. there exists a non-empty subhypergraph of $H^k(n,p)$ with minimum ℓ-degree at least s, which is called the s-core. This has already been studied in the case $\ell = 1$ by Molloy [18], but for other values of ℓ this question remains wide open.

Acknowledgements

We would like to thank an anonymous referee, whose helpful suggestions significantly improved the clarity of the paper.

References

[1] Aronshtam, L. and Linial, N. (2015) When does the top homology of a random simplicial complex vanish? *Random Struct. Alg.* **46** 26–35.
[2] Aronshtam, L. and Linial, N. (2016) The threshold for d-collapsibility in random complexes. *Random Struct. Alg.* **48** 260–269.
[3] Aronshtam, L., Linial, N., Łuczak, T. and Meshulam, R. (2013) Collapsibility and vanishing of top homology in random simplicial complexes. *Discrete Comput. Geom.* 49, no. 2, 317–334.
[4] Behrisch, M., Coja-Oghlan, A. and Kang, M. (2010) The order of the giant component of random hypergraphs. *Random Struct. Alg.* **36** 149–184.
[5] Behrisch, M., Coja-Oghlan, A. and Kang, M. (2014) Local limit theorems for the giant component of random hypergraphs. *Combin. Probab. Comput.* **23** 331–366.
[6] Bohman, T. (2009) The triangle-free process. *Adv. Math.* **221** 1653–1677.
[7] Bollobás, B. (1984) The evolution of random graphs. *Trans. Amer. Math. Soc.* **286** 257–274.
[8] Bollobás, B. (2001) *Random Graphs*, second edition, Cambridge University Press.
[9] Bollobás, B. and Riordan, O. (2012) Asymptotic normality of the size of the giant component in a random hypergraph. *Random Struct. Alg.* **41** 441–450.
[10] Cooley, O., Kang, M. and Koch, K. The size of the giant component in random hypergraphs. *Random Struct. Alg.*, DOI: 10.1002/rsa.20761.

[11] Erdős, P. and Rényi, A. (1960) On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** 17–61.
[12] Janson, S., Łuczak, T. and A. Ruciński, A. (2000) *Random Graphs*, Wiley.
[13] Karoński, M. and Łuczak, T. (2002) The phase transition in a random hypergraph. *J. Comput. Appl. Math.* **142** 125–135.
[14] Krivelevich, M. and Sudakov, B. (2013) The phase transition in random graphs: A simple proof. *Random Struct. Alg.* **43** 131–138.
[15] Linial, N. and Meshulam, R. (2006) Homological connectivity of random 2-complexes. *Combinatorica* **26** 475–487.
[16] Lu, L. and Peng, X. High-order phase transition in random hypergraphs. arXiv:1409.1174
[17] Łuczak, T. (1990) Component behavior near the critical point of the random graph process. *Random Struct. Alg.* **1** 287–310.
[18] Molloy, M. (2005) Cores in random hypergraphs and boolean formulas. *Random Struct. Alg.* **27** 124–135.
[19] Nachmias, A. and Peres, Y. (2010) The critical random graph, with martingales. *Israel J. Math.* **176** 29–41.
[20] Ravelomanana and Rijamamy (2006) Creation and growth of components in a random hypergraph process. In *COCOON 2006: Computing and Combinatorics*, Springer, pp. 350–359.
[21] Schmidt-Pruzan, J. and E. Shamir, E. (1985) Component structure in the evolution of random hypergraphs. *Combinatorica* **5** 81–94.

https://doi.org/10.1017/S096354831800010X Published online by Cambridge University Press