Subcritical random hypergraphs, high-order components, and hypertrees

Oliver Cooley*, Wenjie Fang†, Nicola Del Giudice‡, Mihyun Kang *†‡
Graz University of Technology
Institute of Discrete Mathematics
8010 Graz, Austria
{cooley,fang,delgiudice,kang}@math.tugraz.at

Abstract

One of the central topics in the theory of random graphs deals with the phase transition in the order of the largest components. In the binomial random graph $\mathcal{G}(n,p)$, the threshold for the appearance of the unique largest component (also known as the giant component) is $p_g = n^{-1}$. More precisely, when p changes from $(1-\varepsilon)p_g$ (subcritical case) to p_g and then to $(1+\varepsilon)p_g$ (supercritical case) for $\varepsilon > 0$, with high probability the order of the largest component increases smoothly from $O(\varepsilon^{-2}\log(\varepsilon^3 n))$ to $\Theta(n^{2/3})$ and then to $(1\pm o(1))2\varepsilon n$. Furthermore, in the supercritical case, with high probability the largest components except the giant component are trees of order $O(\varepsilon^{-2}\log(\varepsilon^3 n))$, exhibiting a structural symmetry between the subcritical random graph and the graph obtained from the supercritical random graph by deleting its giant component.

As a natural generalisation of random graphs and connectedness, we consider the binomial random k-uniform hypergraph $\mathcal{H}^k(n,p)$ (where each k-tuple of vertices is present as a hyperedge with probability p independently) and the following notion of high-order connectedness. Given an integer $1 \le j \le k-1$, two sets of j vertices are called *j-connected* if there is a walk of hyperedges between them such that any two consecutive hyperedges intersect in at least j vertices. A j-connected component is a maximal collection of pairwise j-connected j-tuples of vertices. Recently, the threshold for the appearance of the giant j-connected component in $\mathcal{H}^k(n,p)$ and its order were determined. In this article, we take a closer look at the subcritical random hypergraph. We determine the structure and size (*i.e.* number of hyperedges) of the largest j-connected components, with the help of a certain class of "hypertrees" and related objects. In our proofs, we combine various probabilistic and enumerative techniques, such as generating functions and couplings with branching processes. Our study will pave the way to establishing a symmetry between the subcritical random hypergraph and the hypergraph obtained from the supercritical random hypergraph by deleting its giant j-connected component.

1 Introduction

1.1 Motivation One of the most prominent results on random graphs is the so-called *phase transition* in the order of the largest components, first discussed by Erdős and Rényi in their seminal work [8]: a small change in the edge density around the critical value drastically alters the structure and order of the largest components. Their result was improved for example by Bollobás [2] and Łuczak [11] and is often stated for the binomial random graph $\mathcal{G}(n,p)$, a graph with vertex set $[n] := \{1,\dots,n\}$ in which each pair of vertices is present as an edge with probability p independently. Throughout the paper, log denotes the natural logarithm and we say that an event holds *with high probability* (*whp* for short) if the probability that it holds tends to 1 as $n \to \infty$.

THEOREM 1.1. ([2, 3, 8, 11]) *Let $0 < \varepsilon < 1$ be a constant or a function in n satisfying $\varepsilon \to 0$ and $\varepsilon^3 n \to \infty$. For each $i \in \mathbb{N}$, let $C_i = C_i(\mathcal{G}(n,p))$ denote the number of vertices in the i-th largest component in $\mathcal{G}(n,p)$.*

(1) If $p = (1-\varepsilon)n^{-1}$, then whp every component is either a tree or unicyclic, and for every $i \ge 1$, the i-th largest component is a tree. Furthermore, for any function $\omega = \omega(n) \to \infty$, whp

$$\left| C_1 - \alpha^{-1}\left(\log(\varepsilon^3 n) - \frac{5}{2}\log\log(\varepsilon^3 n)\right)\right| \le \omega\varepsilon^{-2},$$

where $\alpha = -\varepsilon - \log(1-\varepsilon) = \varepsilon^2/2 + O(\varepsilon^3)$ and

*Supported by Austrian Science Fund (FWF) & German Research Foundation (DFG): I3747
†Supported by Austrian Science Fund (FWF): I2309 & P27290
‡Supported by Austrian Science Fund (FWF): W1230II

$C_i = (2 + o(1))\varepsilon^{-2}\log(\varepsilon^3 n)$ *for* $i \geq 2$.

(2) If $p = (1+\varepsilon)n^{-1}$*, then whp the largest component contains at least two cycles and*

$$C_1 = (1 + o(1))2\varepsilon n.$$

Furthermore, for every $i \geq 2$*, whp the* i*-th largest component is a tree with* $C_i = (2 + o(1))(\varepsilon^{-2}\log(\varepsilon^3 n))$ *and in particular, for any function* $\omega = \omega(n) \to \infty$*, whp*

$$\left| C_2 - \beta^{-1}\left(\log(\varepsilon^3 n) - \frac{5}{2}\log\log(\varepsilon^3 n)\right)\right| \leq \omega\varepsilon^{-2},$$

where $\beta = \varepsilon - \log(1+\varepsilon) = \varepsilon^2/2 + O(\varepsilon^3)$.

Note that in the supercritical random graph (*i.e.* when $p = (1+\varepsilon)n^{-1}$), the largest component is substantially larger than the second-largest component and therefore it is also called the *giant* component. Theorem 1.1 shows a *symmetry* between the structure of the subcritical random graph and the supercritical random graph with the giant component removed.

Since this seminal work, various models of random graphs have been introduced and analysed. The most commonly studied higher-dimensional analogue of $\mathcal{G}(n,p)$ is the binomial random k-uniform hypergraph $\mathcal{H}^k(n,p)$ defined below. Amongst other properties, vertex-connectedness [1, 4, 5, 10, 12] and high-order connectedness (also known as j-tuple-connectedness) [6, 7] of $\mathcal{H}^k(n,p)$ have been extensively studied. Before stating the relevant results, we introduce the necessary concepts.

Let $k \geq 2$ and $1 \leq j \leq k-1$ be integers. A *k-uniform hypergraph* H is a pair $H = (V, E)$, where V is the set of vertices, and $E \subseteq \binom{V}{k}$, a collection of k-element subsets of V. Each $K \in E$ is called a *hyperedge*. An ℓ-element subset of V is called an *ℓ-set of V* (or *ℓ-set* for short). A pair $\{J_1, J_2\}$ of j-sets are called *j-tuple-connected* (*j-connected* for short) in H if there is a sequence of hyperedges $K_1, \ldots, K_m$ such that $J_1 \subset K_1$, $J_2 \subset K_m$ and $|K_i \cap K_{i+1}| \geq j$ for all $1 \leq i \leq m-1$. Additionally, any j-set is always j-connected to itself. The *j-connected components* (*j-components* for short) of H are equivalence classes of the relation $\sim_j$ defined by $J_1 \sim_j J_2$ if and only if J_1 and J_2 are j-connected. In other words, a j-component is a maximal collection of j-sets that are pairwise j-connected. Given a j-component $\mathcal{J}$, let $\mathcal{K}_{\mathcal{J}}$ be the set of hyperedges containing any j-set in $\mathcal{J}$. In a slight abuse of terminology, we say that the j-component $\mathcal{J}$ also contains the hyperedges $\mathcal{K}_{\mathcal{J}}$. The *order* of a j-component denotes the number of j-sets it contains, and the *size* of a j-component denotes the number of hyperedges it contains. A *hypertree j-component* (a *hypertree* for short) is a j-component that contains as many j-sets as possible *given its size*, *i.e.* if it has size s and order t, then $t = 1 + \left(\binom{k}{j} - 1\right)s$. The case $k = 2$ and $j = 1$ corresponds to the classical concepts in a graph.

We denote by $\mathcal{H}^k(n,p)$ the *random k-uniform hypergraph* with vertex set $[n]$, in which every hyperedge is present independently with probability p. When parameters are clear from the context, we use $\mathcal{H}$ as a shorthand for $\mathcal{H}^k(n,p)$. The following higher-dimensional analogue of the random graph phase transition for $\mathcal{H}$ and j-connectedness was obtained in [6, 7].

THEOREM 1.2. ([6, 7]) *Given integers* $k \geq 2$ *and* $1 \leq j \leq k-1$*, let* $\varepsilon = \varepsilon(n) > 0$ *satisfy* $\varepsilon \to 0$*,* $\varepsilon^3 n^j \to \infty$ *and* $\varepsilon^2 n^{1-2\delta} \to \infty$*, as* $n \to \infty$*, for some constant* $\delta > 0$*. Let*

$$\bar{p}_0 := \left(\binom{k}{j} - 1\right)^{-1}\binom{n}{k-j}^{-1}.$$

(1) If $p = (1-\varepsilon)\bar{p}_0$*, then whp all j-components of* $\mathcal{H}^k(n,p)$ *have order at most* $O(\varepsilon^{-2}\log n)$.

(2) If $p = (1+\varepsilon)\bar{p}_0$*, then whp the order of the largest j-component of* $\mathcal{H}^k(n,p)$ *is* $(1 \pm o(1))\frac{2\varepsilon}{\binom{k}{j}-1}\binom{n}{j}$*, while all other j-components have order at most* $o(\varepsilon n^j)$.

The aim of this paper is to strengthen Theorem 1.2 in view of Theorem 1.1 by taking a closer look at the *subcritical* case and addressing the following natural questions.

(a) What is the precise asymptotic size of the largest j-component of $\mathcal{H}^k(n,p)$?

(b) What does the largest j-component look like? Is it whp a hypertree or some other more complex structure?

1.2 Main result In Theorems 1.1 and 1.2, the order of a j-component was studied. In this paper, we study the size of a j-component, *i.e.* the number of hyperedges it contains. Therefore, whenever we talk about the *i-th largest j-component*, the ranking is determined by the size rather than the order (*i.e.* the number of j-sets it contains). Observe that in the range of the hyperedge probability in our study, whp the order and size of the largest j-component only differ roughly by a multiplicative constant $c_0 = \binom{k}{j} - 1$.

We say that a random variable $X = X(n)$ satisfies $X = O_p(1)$ if for any function $K = K(n) \to \infty$ we have $|X| \leq K$ whp.

THEOREM 1.3. *Given integers $k \geq 2$ and $1 \leq j \leq k-1$, and $\varepsilon = \varepsilon(n)$ with $0 < \varepsilon < 1$, $\varepsilon^2 n^{k-j} (\log n)^{-3/2} \to \infty$ and $\varepsilon^4 n^j (\log n)^{-3} \to \infty$, let*

$$c_0 = \left(\binom{k}{j} - 1 \right) \text{ and } p_0 = c_0^{-1} \binom{n-j}{k-j}^{-1}.$$

Let $L_1 = L_1(\mathcal{H}^k(n,p))$ be the number of hyperedges in the largest j-component of $\mathcal{H}^k(n,p)$. If $p = (1-\varepsilon)p_0$, then whp

$$L_1 = \delta^{-1} \left(\log \lambda - \frac{5}{2} \log\log \lambda + O_p(1) \right),$$

where $\delta = -\varepsilon - \log(1-\varepsilon) = \varepsilon^2/2 + O(\varepsilon^3)$ and $\lambda = \varepsilon^3 \binom{n}{j}$. Furthermore, whp $\mathcal{H}^k(n,p)$ contains at least one hypertree of size at least $\delta^{-1} \left(\log \lambda - \frac{5}{2} \log\log \lambda - K(n) \right)$ for any $K(n) \to \infty$.

We note that the critical probability p_0 differs from $\bar{p}_0$ (defined in Theorem 1.2) by a factor of $1 + O(n^{-1})$. This is because we analyse a range closer to criticality, which requires a more precise value of p_0.

The two conditions on ε in Theorem 1.3 are necessary for the proofs. We discuss these conditions further in Section 6.

We also note that the coefficient $-5/2$ before the $\log\log\lambda$ factor in Theorem 1.3 is the same as that in Theorem 1.1 for graphs, and arises from the universal asymptotic behaviour of various families of labelled trees, *i.e.* connected acyclic graphs. More precisely, the asymptotic number of trees on t vertices in such a family has the form $c \cdot t! \gamma^t t^{-5/2}$, with c and γ depending on the precise nature of the family. The proofs of both Theorem 1.1 and Theorem 1.3 involve asymptotic counting of such families of trees, and the coefficient $-5/2$ comes from the common polynomial factor $t^{-5/2}$. In the case when the trees are rooted, which is more commonly considered, the exponent $-5/2$ would become $-3/2$ (see [9, Section VII.3]) – the extra factor of t comes from the choice of the root.

2 Largest component: Proof of Theorem 1.3

In order to prove Theorem 1.3, we bound the size of the largest j-component from above and below: we prove that whp there is no j-component of size larger than the claimed value (Lemma 2.1) and that for any size smaller than the claimed value, there is at least one j-component (and indeed a hypertree component) with larger size (Lemma 2.2). In the following we will quantify these bounds precisely.

For the upper bound, we compare the *component search process*, which explores a j-component, with a *two-type branching process* (Section 3). We observe that the two-type branching process gives an upper coupling on the search process (Lemma 3.1) to obtain the following upper bound on sizes of j-components:

LEMMA 2.1. *Let $k, j, \varepsilon, p, \delta, \lambda$ be given as in Theorem 1.3, and $K(n) \to \infty$. Whp, $\mathcal{H}^k(n,p)$ contains no j-component of size larger than*

$$\delta^{-1} \left(\log \lambda - \frac{5}{2} \log\log \lambda + K(n) \right).$$

Lemma 2.1 will be proved in Section 4 by estimating the number of possible instances of the two-type branching process (Lemma 4.1), which are the so-called *rooted labelled two-type trees*. Using the coupling argument of Lemma 3.1, we obtain an upper bound on the expected number of j-components whose size is larger than a fixed value. We conclude by applying Markov's inequality to prove that if we run (at most) $\binom{n}{j}$ component search processes (each one starting from a different j-set), whp no component of size larger than the claimed value will be discovered.

For the lower bound, in Section 5 we estimate the expected number of *hypertree components*. We introduce and count *wheels* (Lemma 5.1), which are higher-dimensional analogues of cycles in graphs. Using the enumeration result for wheels, we prove that for a certain range of s, most of the instances of our two-type branching process of size s contain no wheel, and so correspond to hypertree components (Lemma 5.2). Applying probabilistic arguments, *e.g.* Chebyshev's inequality, we obtain:

LEMMA 2.2. *Let $k, j, \varepsilon, p, \delta, \lambda$ be given as in Theorem 1.3, and $K(n) \to \infty$. Whp, $\mathcal{H}^k(n,p)$ contains at least one hypertree component of size at least*

$$\delta^{-1} \left(\log \lambda - \frac{5}{2} \log\log \lambda - K(n) \right).$$

Proof. [Proof of Theorem 1.3] Observe that Theorem 1.3 follows directly from Lemmas 2.1 and 2.2.

3 Search process and branching process

Throughout the paper, we let k, j, ε, p, δ, λ, c_0, p_0 be as given in Theorem 1.3. For positive integers $n \geq k$, we write $n_{(k)}$ for the *falling factorial* $n_{(k)} := n(n-1)(n-2)\cdots(n-k+1) = n!/(n-k)!$.

We also introduce auxiliary two-type graphs. A *two-type graph* is a *connected* bipartite graph on a set of vertices of type k and a set of vertices of type j, where each vertex of type k is connected to exactly $\binom{k}{j}$ vertices of type j. We only consider two-type graphs with at least one vertex of type k. A *labelled two-type graph* with label set $[n]$ is a two-type graph with labels on its

vertices, where vertices of type j are labelled by j-sets of $[n]$ and vertices of type k by k-sets of $[n]$ in such a way that to each of the $\binom{k}{j}$ vertices of type j connected to a given vertex of type k with label $K \in \binom{[n]}{k}$ is assigned a distinct j-set $J \subset K$ as label. There is a natural bijection between the set of labelled two-type graphs with label set $[n]$ whose vertices all have distinct labels and the set of *possible components*, *i.e.* pairs $(\mathcal{J}, \mathcal{K})$ which could potentially be the families of j-sets and k-sets in a j-component of some hypergraph on vertex set $[n]$. *Acyclic* two-type graphs are called *two-type trees*.

The *component search process* is defined as follows. We explore j-components in $\mathcal{H}$ via a breadth-first search algorithm: we use a queue to stock active (defined below) j-sets and k-sets, with a j-set J_0 in the queue when we start. When the queue is not empty, an element pops out from the queue. If the popped element is a j-set J_*, then we consider every k-set K containing J_* in arbitrary order, and if K is in $\mathcal{H}$ but not yet visited, then we call it *active* and put it in the queue; if the popped element is a k-set K_*, then we consider every j-set $J \subset K_*$ in arbitrary order, and if J has not been visited before, then we call it *active* and put it in the queue. We continue until the queue is empty, and we find the j-component containing J_0. To get all the j-components, we only need to perform the same procedure on unexplored j-sets in $\mathcal{H}$ until exhaustion. (An example of the search process is given in Figure 1.)

To prove Lemma 2.1, we provide in Lemma 3.1 an upper coupling on the search process with the following *two-type branching process*. Each vertex of the branching process is either of type j or of type k, and is labelled by a j-set or k set of $[n]$, respectively. The branching process constructs a labelled two-type tree with label set $[n]$ in the following way. Given a vertex u_0 of type j, the branching process begins with u_0 (as the root). For each vertex u of type j with label J, let $\mathcal{K}_J := \{K \in \binom{[n]}{k} \mid J \subset K\}$ be the set of possible labels of children of u. For each label $K \in \mathcal{K}_J$, independently with probability p, we generate a new vertex v of type k and assign the label K to v. For each such new vertex v, we then attach $\binom{k}{j} - 1$ many new vertices of type j as children of v, with distinct labels from $\binom{K}{j} \setminus \{J\}$ (note that different vertices may have the same label). We denote this branching process by $\mathcal{T}$. The size of $\mathcal{T}$ is defined as the number of vertices of type k that it discovers. Note that each instance of $\mathcal{T}$ corresponds to a labelled two-type tree rooted at a vertex of type j, which we call a *rooted labelled two-type tree*. An example of the two-type branching process is given in Figure 1.

LEMMA 3.1. *We can couple the component search process on j-components of $\mathcal{H}$ from above with $\binom{n}{j}$ copies of $\mathcal{T}$, which implies that for any given $s \in \mathbb{N}$, the expected number of j-components of $\mathcal{H}$ of size at least s is bounded above by the expected number of instances of $\mathcal{T}$ of size at least s.*

Proof. In the component search process, whenever a j-set J becomes active, the set of k-sets we may query is certainly contained in $\mathcal{K}_J$ (some may not be permissible since they have already been queried), and for each k-set K discovered in this way, the j-sets that become active are all in $\binom{K}{j} \setminus \{J\}$ (some may not become active if they were already discovered). Thus, in $\mathcal{T}$ we may have made some additional queries which are not made in the component search process, and we may have added some vertices of type j whose labels correspond to j-sets not added in the search process. Hence, one component search process terminated when the component is fully discovered can certainly be coupled with one instance of $\mathcal{T}$.

Since we need at most $\binom{n}{j}$ component search process to discover all j-components, we can couple with $\binom{n}{j}$ branching processes starting from vertices of type j with all possible labels $J \in \binom{[n]}{j}$. More precisely, whenever we start exploring a new j-component from a j-set J, we upper couple this portion of the component search process by the branching process starting from a vertex of type j with label J; using in total $\binom{n}{j}$ branching processes (although some of them may be superfluous) we have an upper coupling for the search process on all the j-components.

4 Upper bound on L_1: Proof of Lemma 2.1

To prove Lemma 2.1 we first bound the number of possible rooted two-type trees that can be constructed by the two-type branching process $\mathcal{T}$. Let $\mathcal{B}$ be the set of all possible instances of $\mathcal{T}$ and thus also of all rooted labelled two-type trees. For each $s \in \mathbb{N}$, let $\mathcal{B}_s$ be the set of elements in $\mathcal{B}$ of size s, which is equal to the set of all rooted labelled two-type trees with s vertices of type k. Let B_s be the cardinality of $\mathcal{B}_s$. In the following lemma we determine the order of B_s.

LEMMA 4.1. *For $s \in \mathbb{N}$ with $s \leq \varepsilon^{-2}(\log n)^{3/2}$, we have*

$$B_s = \Theta\left(\binom{n}{j}\binom{n-j}{k-j}^s \frac{(c_0 e)^s}{s^{3/2}}\right).$$

Proof. We first consider the class of rooted (unlabelled) two-type trees with a vertex J of type j as the root and at least one vertex of type k. Note that although the vertices do not receive labels from $\binom{[n]}{j}$ or $\binom{[n]}{k}$, they are considered distinguishable. The generating function of this class is $T_J = T_J(z)$, where z marks the number of

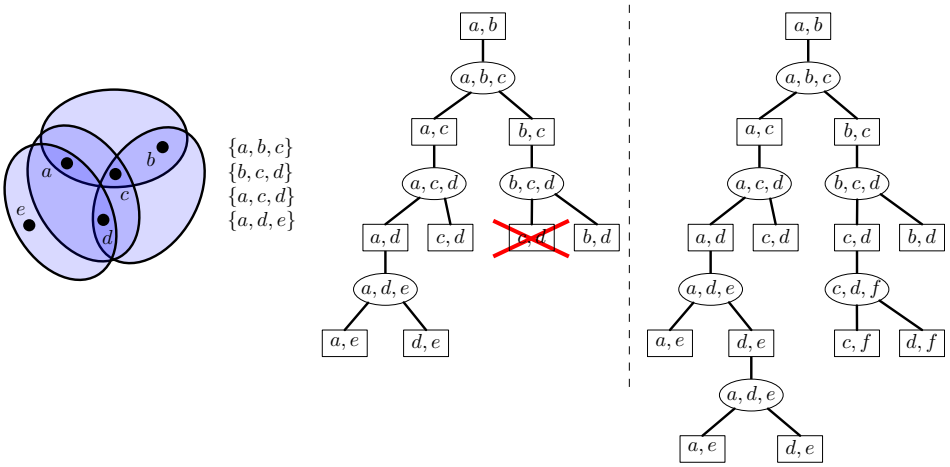

Figure 1: Examples with $k = 3, j = 2$. Left: a j-component with the two-type tree of its search process starting from $\{a, b\}$. Right: an instance of the branching process, giving a two-type tree that cannot come from the search process.

vertices of type k, satisfies the following equation:

$$T_J(z) = \exp\left(z(1 + T_J(z))^{c_0}\right) - 1. \tag{4.1}$$

Let $W(z)$ denote the Lambert W-function defined by the equation $z = W(z)\exp(W(z))$. We have

$$T_J(z) = \exp\left(-\frac{W(-c_0 z)}{c_0}\right) - 1. \tag{4.2}$$

By the Lagrange Inversion Theorem (see [9, Appendix A.6]), we have

$$W^r(z) = \sum_{i \geq r} \frac{-r(-i)^{i-r-1}}{(i-r)!} z^i. \tag{4.3}$$

Since $W(z)$ is analytic in a neighbourhood of $z = 0$ and $W(0) = 0$, for any function $F(z)$ analytic in a neighbourhood of $z = 0$, the composition $F(W(z))$ is still analytic near $z = 0$. Thus, using Taylor expansion we have

$$\begin{aligned} T_J(z) &= \exp\left(-\frac{W(-c_0 z)}{c_0}\right) - 1 \\ &= \sum_{r \geq 1} \frac{1}{r!} \left(-\frac{W(-c_0 z)}{c_0}\right)^r \\ &= \sum_{r \geq 1} \frac{1}{r!} \sum_{i \geq r} \frac{r c_0^{i-r} i^{i-r-1}}{(i-r)!} z^i \\ &= \sum_{i \geq 1} z^i \left(\sum_{r=1}^{i} \frac{c_0^{i-r} i^{i-r-1}}{(r-1)!(i-r)!}\right). \end{aligned}$$

Thus, the number F_s of rooted unlabelled two-type trees with s vertices of type k is

$$\begin{aligned} F_s = [z^s]T_J(z) &= \sum_{r=1}^{s} \frac{c_0^{s-r} s^{s-r-1}}{(r-1)!(s-r)!} \\ &= \frac{c_0^{s-1} s^{s-2}}{(s-1)!} \sum_{r=0}^{s-1} \frac{c_0^{-r} s^{-r} (s-1)_{(r)}}{r!}. \end{aligned}$$

As an upper bound, we have

$$F_s \leq \frac{c_0^{s-1} s^{s-1}}{s!} \sum_{r=0}^{s-1} \frac{c_0^{-r}}{r!} \leq \frac{c_0^{s-1} s^{s-1}}{s!} e^{1/c_0}.$$

As a lower bound, we have

$$F_s \geq \frac{c_0^{s-1} s^{s-1}}{s!}.$$

Thus, using Stirling's formula,

$$F_s = \Theta\left(\frac{c_0^{s-1} s^{s-1}}{s!}\right) = \Theta\left(\frac{(c_0 e)^s}{s^{3/2}}\right).$$

In each instance of $\mathcal{T}$, we have $\binom{n}{j}$ choices for the label of the initial vertex of type j and for each vertex of type k discovered from a vertex of type j, we must choose $k-j$ new elements among $n-j$ (excluding those already in the parent vertex of type j). Then the labels of the next c_0 vertices of type j are already determined. Therefore, we have the following relation between B_s and F_s:

$$B_s = \binom{n}{j}\binom{n-j}{k-j}^s F_s = \Theta\left(\binom{n}{j}\binom{n-j}{k-j}^s \frac{(c_0 e)^s}{s^{3/2}}\right)$$

as claimed.

Proof. [Proof of Lemma 2.1] We consider $\binom{n}{j}$ independent instances of the branching process $\mathcal{T}$. Let R_s be the random variable which counts the number of vertices of type j present in total in the instances that have size s. In each such instance, the s vertices of type k are present with probability p^s. For the number of absent vertices of type k (*i.e.* which are not selected during the process) in an instance of size s, we observe that the j-set that labels the starting vertex of type j is contained in $\binom{n-j}{k-j}$ many k-sets, and subsequently for each of the s vertices of type k we discover c_0 further vertices of type j, whose labels are each contained in $\left(\binom{n-j}{k-j}-1\right)$ further k-sets. However, we have to consider the s vertices of type k that are indeed discovered. The total number of absent vertices of type k is therefore

$$\begin{aligned}&\binom{n-j}{k-j}+c_0 s\left(\binom{n-j}{k-j}-1\right)-s\\&\qquad=(1+c_0 s)\binom{n-j}{k-j}-s(1+c_0).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}(R_s)&\le B_s p^s(1-p)^{(1+c_0 s)\binom{n-j}{k-j}-s(1+c_0)}\\&=\Theta(1)\binom{n}{j}\left(\binom{n-j}{k-j}c_0 ep(1-p)^{c_0\binom{n-j}{k-j}}\right)^s\\&\qquad\cdot(1-p)^{\binom{n-j}{k-j}-s(1+c_0)}s^{-3/2}.\end{aligned}$$

where the equality is due to Lemma 4.1. Recall that $p=(1-\varepsilon)c_0^{-1}\binom{n-j}{k-j}^{-1}$. Using $(1-p)\le e^{-p}$ we have

$$\begin{aligned}\mathbb{E}(R_s)&\le\Theta(1)\binom{n}{j}\left((1-\varepsilon)\cdot e\cdot e^{-(1-\varepsilon)}\right)^s s^{-3/2}\\&\qquad\cdot\exp\left((1-\varepsilon)(1+c_0^{-1})s\binom{n-j}{k-j}^{-1}-\frac{1-\varepsilon}{c_0}\right)\\&\le\Theta(1)\binom{n}{j}\exp\left(s(\log(1-\varepsilon)+\varepsilon)\right)s^{-3/2}\\&\qquad\cdot\exp\left(2s\binom{n-j}{k-j}^{-1}\right).\end{aligned}$$

Since $\delta=-\varepsilon-\log(1-\varepsilon)$, we have

$$\mathbb{E}(R_s)\le\Theta(1)\binom{n}{j}\exp\left(-s\delta\right)s^{-3/2}\exp\left(\frac{2s}{\binom{n-j}{k-j}}\right).$$

Now, let $D_{\ge s}$ be the random variable for the number of components of $\mathcal{H}$ of size at least s. By Lemma 3.1, $\mathbb{E}(D_{\ge s})$ is bounded above by the expected number of instances of the two-type branching process $\mathcal{T}$ of size at least s, which is bounded by $\sum_{t\ge s}\mathbb{E}(R_t)c_0^{-1}t^{-1}$.

Let $\hat{s}=\delta^{-1}(\log\lambda-\frac{5}{2}\log\log\lambda+K(n))$, and $\delta^-=\delta-2\binom{n-j}{k-j}^{-1}$. Recalling that $\delta=\varepsilon^2/2+O(\varepsilon^3)$ and $\varepsilon^2 n^{k-j}(\log n)^{-3/2}\to\infty$, it holds that $\delta=\omega(n^{j-k}(\log n)^{3/2})$, which means $\delta^-=\delta-2\binom{n-j}{k-j}^{-1}=(1-o(1))\delta$. Thus, we have $\delta^->0$ for n large enough and we obtain

$$\begin{aligned}\mathbb{E}(D_{\ge\hat{s}})&\le\sum_{t\ge\hat{s}}\mathbb{E}(R_t)c_0^{-1}t^{-1}\\&\le\Theta(1)\binom{n}{j}(\hat{s})^{-5/2}\sum_{t\ge\hat{s}}(e^{-\delta^-})^t\\&\le\Theta(1)\binom{n}{j}(\hat{s})^{-5/2}\frac{e^{-\delta^-\hat{s}}}{1-e^{-\delta^-}}\\&\le\Theta(1)\binom{n}{j}(\hat{s})^{-5/2}\frac{e^{-\delta^-\hat{s}}}{\delta^-}.\end{aligned}$$

Since $\varepsilon^2 n^{k-j}(\log n)^{-3/2}\to\infty$, we have

$$\begin{aligned}\hat{s}\binom{n-j}{k-j}^{-1}&=(1+o(1))\delta^{-1}(\log\lambda)\binom{n-j}{k-j}^{-1}\\&=O(1)\frac{\log(\varepsilon^3 n^j)}{\varepsilon^2 n^{k-j}}\\&\le O(1)(\log n)\varepsilon^{-2}n^{j-k}=o(1).\end{aligned}$$

This leads to

$$\begin{aligned}e^{-\delta^-\hat{s}}&=e^{-\delta\hat{s}}\exp\left(2\hat{s}\binom{n-j}{k-j}^{-1}\right)\\&=e^{-\delta\hat{s}}\exp(o(1))=(1+o(1))e^{-\delta\hat{s}}.\end{aligned}$$

Since $\varepsilon^4 n^j(\log n)^{-3}\to\infty$, we have $\lambda=\varepsilon^3\binom{n}{j}\to\infty$. Without loss of generality, we may assume that $K(n)=o(\log\lambda)$. We thus obtain

$$\begin{aligned}&\mathbb{E}(D_{\ge\hat{s}})\\&\le\Theta(1)\binom{n}{j}(\hat{s})^{-5/2}\frac{(1+o(1))e^{-\delta\hat{s}}}{(1+o(1))\delta}\\&=\Theta(1)\binom{n}{j}(\hat{s})^{-5/2}\frac{e^{-\delta\hat{s}}}{\delta}\\&\le\Theta(1)\exp\left(\log\binom{n}{j}-\log\lambda+\frac{5}{2}\log\log\lambda-K(n)\right.\\&\qquad\left.-\frac{5}{2}\log\left((1+o(1))\frac{\log\lambda}{\delta}\right)-\log\delta\right)\\&=O\left(\exp\left(-K(n)\right)\right).\end{aligned}$$

Since $K(n)\to\infty$, by Markov's inequality, whp we have $D_{\ge\hat{s}}=0$, meaning that there is no j-component of size larger than $\hat{s}=\delta^{-1}(\log\lambda-\frac{5}{2}\log\log\lambda+K(n))$.

5 Lower bound on L_1: Proof of Lemma 2.2

In this section we will prove that $\mathcal{H}$ contains a hypertree component larger than a certain size, which provides a lower bound on L_1 (Lemma 2.2).

Recall that a hypertree component (*i.e.* j-component that contains as many j-sets as possible given its size) corresponds to a labelled two-type tree with no repeated labels. An important structure that may appear in $\mathcal{H}$ is the so-called *wheel*. A *wheel* of length $\ell \geq 2$ is a pair of sequences, one of ℓ distinct hyperedges $K_1, K_2, \ldots, K_\ell, K_{\ell+1} = K_1$ and the other of ℓ distinct j-sets $J_0, J_1, \ldots, J_{\ell-1}, J_\ell = J_0$ such that $J_i \subset K_i \cap K_{i+1}$ for all $1 \leq i \leq \ell$. Two wheels are considered identical if they only differ by a cyclic rotation or order reversion of the elements of the sequences. Given a wheel, it lies in a single j-connected component, and the presence of a wheel is the only obstacle for a component to be a hypertree. The reason is that a component ceases to be a hypertree if and only if we encounter the same j-set or k-set at least twice in the component search process, which makes a wheel. We have the following enumeration result on wheels.

LEMMA 5.1. *Let $w_\ell = w_\ell(n)$ be the number of possible wheels of length $\ell \geq 2$, with vertices chosen from $[n]$. We have*

$$w_\ell \leq \frac{c_w n^{k-j}}{p_0^{\ell-1}\ell},$$

where

$$c_w = \frac{(k-j)^j}{j!(k-j)!} \prod_{m=1}^{j-1} \left(1 - c_0^{-1}\left(\binom{k-m}{j-m} - 1\right)\right)^{-1}.$$

We note that c_w does not depend on n or ℓ.

Proof. [Sketch] We construct a wheel with ℓ distinct hyperedges by taking hyperedges one by one. For the first hyperedge, we have $\binom{n}{k}$ choices, and for each subsequent hyperedge, we first choose the shared j-set (c_0 choices), then the other $k-j$ vertices ($\binom{n-j}{k-j}$ choices), in total p_0^{-1} choices. However, for the last hyperedge, we need it to contain at least j vertices of the first hyperedge, which gives heuristically a factor of n^j. The constant c_w roughly accounts for the precise ways for the last hyperedge to gain those j vertices.

Recall that we denote by $\mathcal{B}$ the set of possible instances of $\mathcal{T}$ and by $\mathcal{B}_s$ the elements in $\mathcal{B}$ with s vertices of type k (bounds on $B_s = |\mathcal{B}_s|$ were given in Lemma 4.1). We now consider the subset $\mathcal{B}^-$ of $\mathcal{B}$ formed by rooted labelled two-type trees in which all labels are distinct, *i.e.* that correspond to hypertrees, and we denote by $\mathcal{B}_s^-$ the set of elements in $\mathcal{B}^-$ with s vertices of type k.

LEMMA 5.2. *For s larger than a certain constant, we have*

$$\begin{aligned} B_s^- &:= |\mathcal{B}_s^-| \\ &= (1 - O(s\log(s)^{1/2}n^{j-k}) - O(s^2\log(s)^{1/2}n^{-j}))B_s. \end{aligned}$$

In particular, if $s \to \infty$ and also $s(\log s)^{1/2}n^{j-k} \to 0$ and $s^2(\log s)^{1/2}n^{-j} \to 0$, we have $B_s^- = (1 - o(1))B_s$.

Proof. [Sketch] By a combinatorial injection from rooted labelled two-type trees to two-type graphs involving wheels, we can give an upper bound on $|\mathcal{B}_s \setminus \mathcal{B}_s^-|$ using Lemma 5.1, generating functions, dominance relations, properties of the Lambert W-function, and Laplace's method for bounding summations.

Let C_s be the number of j-sets in hypertree components of size s in $\mathcal{H}$. It is clear that C_s is a lower bound for the number of j-sets in components of size s.

LEMMA 5.3. *For $s \to \infty$ with $s(\log n)^{1/2}n^{j-k} \to 0$ and $s^2(\log n)^{1/2}n^{-j} \to 0$, we have*

$$\mathbb{E}(C_s) \geq \Theta(1)\exp\left(\log\binom{n}{j} - s\delta - \frac{3}{2}\log s\right).$$

Proof. [Sketch] Given a hypertree of size s, the probability that it occurs as a component in $\mathcal{H}$ is at least $p^s(1-p)^{(1+sc_0)\binom{n-j}{k-j}}$, since each j-set implies the absence of at most $\binom{n-j}{k-j}$ hyperedges, and since the number of j-sets in a hypertree component of size s is exactly $sc_0 + 1$. By Lemma 5.2, we have $B_s^- = (1 + o(1))B_s$, therefore, using similar computations as in the proof of Lemma 2.1, we can derive from B_s^- the desired lower bound on $\mathbb{E}(C_s)$.

Proof. [Proof of Lemma 2.2] We set $s^* = \delta^{-1}\log\lambda$ and $s_* = \delta^{-1}(\log\lambda - \frac{5}{2}\log\log\lambda - K(n))$ where $K(n) \to \infty$. We assume here $K(n) = o(\log\lambda)$ without loss of generality. Furthermore, we set $s_0 = s_* + \delta^{-1}K(n)/2 = \delta^{-1}(\log\lambda - \frac{5}{2}\log\log\lambda - K(n)/2)$. Firstly, we know from Lemma 2.1 that whp there is no component larger than s^*. Let S_+ denote the number of j-sets in components of size between s_* and s^*. Since the range of ε implies that s_* and s_0 satisfy the conditions in Lemma 5.3, we have the following lower bound by counting only hypertree components:

$$\begin{aligned} \mathbb{E}(S_+) &\geq \sum_{s_* \leq s \leq s_0} \mathbb{E}(C_s) \\ &\geq (s_0 - s_*)\Theta(1)\exp\left(\log\binom{n}{j} - s_0\delta - \frac{3}{2}\log s_0\right) \\ &\geq \Theta(1)\frac{K(n)}{2\delta}\exp(K(n)/2)\log\lambda \\ &= \Theta(1)K(n)\exp(K(n)/2)s^* = \omega(s^*). \end{aligned}$$

We now show that $\mathbb{E}(S_+^2)$ is approximately $\mathbb{E}(S_+)^2$. Let q be the probability that a j-set is in a component of size between s_* and s^*, then $\mathbb{E}(S_+) = q\binom{n}{j}$. For two j-sets J_1, J_2 (not necessarily distinct), let κ_1, κ_2 be the components in which they lie respectively. Let s_1, s_2 be the sizes of κ_1 and κ_2 respectively. We have

$$\begin{aligned}\mathbb{E}(S_+^2) &= \sum_{J_1,J_2} \Pr(s_* \le s_1 \le s^*,\ s_* \le s_2 \le s^*) \\ &\le \sum_{J_1} \Pr(s_* \le s_1 \le s^*) \\ &\qquad \cdot \sum_{J_2} \Pr(s_2 \ge s_* \mid s_* \le s_1 \le s^*).\end{aligned}$$

Given that κ_1 is of size between s_* and s^*, we want to bound the probability that κ_2 is of size at least s_*. Given the j-set J_2, if $J_2 \in \kappa_1$, then $\kappa_2 = \kappa_1$ is of size at least s_*; otherwise, we start a modified search process in the hypergraph, where we ignore an existing hyperedge whenever it contains a j-set in κ_1. This modified search process can be upper coupled by the unmodified search process in which all j-sets and k-sets are still available. Therefore, we have

$$\begin{aligned}\mathbb{E}(S_+^2) &\le \sum_{J_1} \Pr(s_* \le s_1 \le s^*) \\ &\qquad \cdot \left(s_1 + \left(\binom{n}{j} - s_1\right)(1+o(1))q\right) \\ &\le \mathbb{E}(S_+)\left(s^* + (1+o(1))q\binom{n}{j}\right) \\ &= \mathbb{E}(S_+)^2\left(1 + o(1) + \frac{s^*}{\mathbb{E}(S_+)}\right) \\ &= \mathbb{E}(S_+)^2(1+o(1)).\end{aligned}$$

By Lemma 2.1 and Chebyshev's inequality, we have

$$\begin{aligned}&\Pr(\mathcal{H} \text{ contains no component of size at least } s_*) \\ &\le \Pr(S_+ = 0) + o(1) \\ &\le \frac{\mathbb{E}(S_+^2) - \mathbb{E}(S_+)^2}{\mathbb{E}(S_+)^2} + o(1) = o(1).\end{aligned}$$

Since the main contribution to S_+ comes from C_s, we know that there is at least one component of size at least s_* that is a hypertree.

6 Discussions

We notice that our proofs impose two restrictions on ε: $\varepsilon^2 n^{k-j}(\log n)^{-3/2} \to \infty$ arising from Lemmas 2.1 and 2.2 and $\varepsilon^4 n^j (\log n)^{-3} \to \infty$ from Lemma 5.2. The latter restriction is not at the conjectured critical window given by $\varepsilon^3 n^j \to \infty$, probably due to our proof techniques. The former restriction may be due to our proof techniques or it may be a real phenomenon, since it appears both in the proofs of upper and lower bounds. Further study is needed to clarify the situation.

In this paper we investigated the structure of the largest component, but it is desirable to characterise the structure of the *whole* subcritical hypergraphs. To this end, we investigated the wheels. Heuristically, these wheels occur in components (average or largest ones) at different stages when p approaches the critical threshold. It would be interesting to investigate these "layered" phase transitions. We believe that this will help us to clarify a structural symmetry between the subcritical random hypergraph and the supercritical random hypergraph with its giant j-component deleted.

References

[1] M. Behrisch, A. Coja-Oghlan, and M. Kang. Local limit theorems for the giant component of random hypergraphs. *Combinatorics, Probabability and Computing*, 23(3):331–366, 2014.

[2] B. Bollobás. The evolution of random graphs. *Trans. Amer. Math. Soc.*, 286(1):257–274, 1984.

[3] B. Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001.

[4] B. Bollobás and O. Riordan. Asymptotic normality of the size of the giant component in a random hypergraph. *Random Structures & Algorithms*, 41(4):441–450, 2012.

[5] B. Bollobás and O. Riordan. Exploring hypergraphs with martingales. *Random Structures & Algorithms*, 50(3):325–352, 2017.

[6] O. Cooley, M. Kang, and C. Koch. The size of the giant high-order component in random hypergraphs. *Random Structures & Algorithms*, 53(2):238–288, 2018.

[7] O. Cooley, M. Kang, and Y. Person. Largest components in random hypergraphs. *Combinatorics, Probability and Computing*, pages 1–22, 2018.

[8] P. Erdős and A. Rényi. On the evolution of random graphs. *Magyar Tud. Akad. Mat. Kutató Int. Közl.*, 5:17–61, 1960.

[9] P. Flajolet and R. Sedgewick. *Analytic combinatorics*. Cambridge University Press, Cambridge, 2009.

[10] M. Karoński and T. Łuczak. Random hypergraphs. In *Combinatorics, Paul Erdős is eighty, Vol. 2 (Keszthely, 1993)*, volume 2 of *Bolyai Soc. Math. Stud.*, pages 283–293. János Bolyai Math. Soc., Budapest, 1996.

[11] T. Łuczak. Component behavior near the critical point of the random graph process. *Random Structures & Algorithms*, 1(3):287–310, 1990.

[12] J. Schmidt-Pruzan and E. Shamir. Component structure in the evolution of random hypergraphs. *Combinatorica*, 5(1):81–94, 1985.