

Counting dense connected hypergraphs via the probabilistic method

Béla Bollobás^{*†} and Oliver Riordan[‡]

November 15, 2015; revised June 21, 2017

Abstract

In 1990 Bender, Canfield and McKay gave an asymptotic formula for the number of connected graphs on $[n] = \{1, 2, \dots, n\}$ with m edges, whenever $n \rightarrow \infty$ and $n - 1 \leq m = m(n) \leq \binom{n}{2}$. We give an asymptotic formula for the number $C_r(n, m)$ of connected r -uniform hypergraphs on $[n]$ with m edges, whenever $r \geq 3$ is fixed and $m = m(n)$ with $m/n \rightarrow \infty$, i.e., the average degree tends to infinity. This complements recent results of Behrisch, Coja-Oghlan and Kang (the case $m = n/(r-1) + \Theta(n)$) and the present authors (the case $m = n/(r-1) + o(n)$, i.e., ‘nullity’ or ‘excess’ $o(n)$). The proof is based on probabilistic methods, and in particular on a bivariate local limit theorem for the number of vertices and edges in the largest component of a certain random hypergraph. The arguments are much simpler than in the sparse case; in particular, we can use ‘smoothing’ techniques to directly prove the local limit theorem, without needing to first prove a central limit theorem.

1 Introduction and results

Our aim in this paper is to prove a result that can be viewed in two equivalent ways: as an asymptotic formula for the number of dense connected r -uniform hypergraphs with a given number of vertices and edges, and as a local limit theorem concerning the numbers of vertices and edges in the largest component of a certain random hypergraph. This paper is a companion to [8], where we used related (but much more complicated) methods to study the sparse case. Here we shall phrase our results in terms of the number of vertices and the

^{*}Department of Pure Mathematics and Mathematical Statistics, Wilberforce Road, Cambridge CB3 0WB, UK, Department of Mathematical Sciences, University of Memphis, Memphis TN 38152, USA, and London Institute for Mathematical Sciences, 35a South St., Mayfair, London W1K 2XF, UK. E-mail: b.bollobas@dpmmms.cam.ac.uk.

[†]Research supported in part by NSF grant DMS-1301614 and EU MULTIPLEX grant 317532.

[‡]Mathematical Institute, University of Oxford, Radcliffe Observatory Quarter, Woodstock Road, Oxford OX2 6GG, UK. E-mail: riordan@maths.ox.ac.uk.

number of edges, rather than considering the nullity as in [8]. (The latter is a more natural parameter when it grows slowly, but not here.)

Throughout the paper we consider r -uniform hypergraphs, where $r \geq 2$ is fixed; much of the time $r \geq 3$. A hypergraph is *connected* if it cannot be written as the vertex disjoint union of two strictly smaller hypergraphs. (This is not the only possible sense of connectedness when $r \geq 3$, but it is the most important one, and the only one we consider here.) A basic problem in enumerative combinatorics is to count the number of ‘irreducible’ objects of a certain type according to certain size parameters. Here we study $C_r(s, m)$, the number of connected r -uniform hypergraphs on $[s] = \{1, 2, \dots, s\}$ with precisely m edges. We write s rather than n for the number of vertices in part for notational consistency with [8], but also because in the bulk of the paper $n \neq s$ will be the number of vertices in a certain random hypergraph; see Section 1.2.

An asymptotic formula for $C_2(s, m)$ (the graph case) was proved by Bender, Canfield and McKay [6] in 1990, throughout the range $s - 1 \leq m \leq \binom{s}{2}$. For $r \geq 3$, in 1997 Karoński and Łuczak [10] proved a result covering the case $m = s/(r-1) + o(\log s / \log \log s)$: this result concerns hypergraphs that are very close to trees. This was generalized to $m = s/(r-1) + o(s^{1/3})$ in an extended abstract of Andriamampianina and Ravelomanana [1] in 2006. Recently, in [8], we proved a result covering the entire ‘sparse case’ $m = s/(r-1) + o(s)$. A formula covering the ‘middle range’ $m = s/(r-1) + \Theta(s)$ was given by Behrisch, Coja-Oghlan and Kang [4, 5]. Our main result here covers the entire remaining range, the ‘dense case’ $m/s \rightarrow \infty$. As usual in this context, the statement involves an implicit definition, and so requires a little preparation.

For $\xi \in (0, 1)$ define

$$\Phi_r(\xi) = \frac{\log(1/\xi)(1 - \xi^r)}{(1 - \xi^{r-1})(1 - \xi)}. \quad (1)$$

It is easy to check that, with $r \geq 2$ fixed, Φ_r is strictly decreasing, since each of the ratios $\log(1/\xi)/(1 - \xi)$ and $(1 - \xi^r)/(1 - \xi^{r-1})$ is. Moreover, $\Phi_r(\xi) \rightarrow r/(r-1)$ as $\xi \rightarrow 1$ and $\Phi_r(\xi) \sim \log(1/\xi) \rightarrow \infty$ as $\xi \rightarrow 0$. It is this latter limit which will be important here. Since Φ_r is continuous, it defines a bijection from $(0, 1)$ to $(r/(r-1), \infty)$.

Given $\bar{d} > r/(r-1)$ let

$$\xi = \xi(\bar{d}) = \Phi_r^{-1}(\bar{d}), \quad (2)$$

and set

$$F_r(\bar{d}) = \bar{d} \log(1 - \xi) - \frac{\bar{d}}{r} \log(1 - \xi^r) - \frac{\xi}{1 - \xi} \log \xi - \log(1 - \xi). \quad (3)$$

Theorem 1.1. *Let $r \geq 2$ be fixed, and let $m = m(s)$ satisfy $m/s \rightarrow \infty$ as $s \rightarrow \infty$. Let $\bar{d} = rm/s$ be the average degree of an m -edge r -uniform hypergraph on $[s]$, and let $H_{s,m}^r$ be such a hypergraph chosen uniformly at random. Then the probability $P_r(s, m)$ that $H_{s,m}^r$ is connected satisfies*

$$P_r(s, m) \sim \exp(-sF_r(\bar{d})) \quad (4)$$

as $s \rightarrow \infty$.

Furthermore, if in addition $m = o(s^{4/3})$, then the number $C_r(s, m)$ of connected m -edge r -uniform hypergraphs on $[s]$ satisfies

$$C_r(s, m) \sim e^{-(r-1)\bar{d}/2 - \mathbb{1}_{r=2}\bar{d}^2/4} \frac{s^{rm}}{m!r!^m} \exp(-sF_r(\bar{d})) \quad (5)$$

as $s \rightarrow \infty$, where $\mathbb{1}_A$ is the indicator function of A .

Writing $N = \binom{s}{r}$, we have $P_r(s, m) = C_r(s, m)/\binom{N}{m}$, so the formulae (4) and (5) are equivalent up to a straightforward calculation; see Lemma 6.2. (In fact, for $r \geq 3$, (5) applies for $m = o(s^{3/2})$, not just $m = o(s^{4/3})$.)

Remark 1.2. We focus on the case $r \geq 3$, since the case $r = 2$ is covered by the result of Bender, Canfield and McKay [6]. For our proof strategy, there is very little difference between the two situations; some formulae have extra terms when $r = 2$, since then certain error terms involving factors of $n^{-(r-1)}$ are not totally negligible. When convenient, we assume $r \geq 3$, commenting briefly on these extra terms.

Remark 1.3. As we shall show later (in Lemma 6.1), for $r \geq 3$ we have

$$\xi = e^{-\bar{d}} + \bar{d}e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}})$$

and

$$F_r(\bar{d}) = e^{-\bar{d}} + \frac{\bar{d}+1}{2}e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}}) \quad (6)$$

as $\bar{d} \rightarrow \infty$. (For $r = 2$ the first term in the formulae above is the same as for $r \geq 3$, but the second is different. The next term in the expansion (6) is different in the cases $r = 2$, $r = 3$ and $r \geq 4$.) The probability that a vertex of $H_{s,m}^r$ is isolated is very close to $e^{-\bar{d}}$, so the expected number of isolated vertices is approximately $\mu = se^{-\bar{d}}$, and the Poisson intuition suggests that the probability that $H_{s,m}^r$ has no such vertex should be approximately $\exp(-\mu)$, at least when μ is bounded or tends to infinity fairly slowly. In turn, in this range we expect the presence of an isolated vertex to be the main obstruction to connectivity. In the light of (6), Theorem 1.1 says that when $s\bar{d}e^{-2\bar{d}} = o(1)$ (corresponding to $\mu(s) = o(\sqrt{s/\log s})$), we have

$$P_r(s, m) \sim \exp(-se^{-\bar{d}}).$$

In other words, the intuition just described gives the right asymptotic answer in this (surprisingly large) range. As a trivial special case, in the ‘very dense’ case where $\bar{d} - \log s \rightarrow \infty$, we have $P_r(s, m) \sim 1$ and so $C_r(s, m) \sim \binom{N}{m}$ where $N = \binom{s}{r}$.

1.1 Comparison to related results

The formulae appearing in Theorem 1.1 (the ‘dense case’ $m/s \rightarrow \infty$), are superficially rather different from those in Theorem 1.1 of [8] (the ‘sparse case’

$m = s/(r-1) + o(s)$) and in (the corrected version of) Theorem 1.1 of Behrisch, Coja-Oghlan and Kang [5], covering the ‘middle range’ $m = s/(r-1) + \Theta(s)$. However, after a suitable change of notation, they are actually rather similar, despite the different ranges of applicability.

Indeed, writing $\rho = 1 - \xi$, the definition of ξ and hence of ρ given by (2) is easily seen to coincide with that in [8]. There, we set $\Psi_r(\rho) = (t-1)/s$ where $t = (r-1)m - s + 1$ is the nullity, and Ψ_r was given by a certain formula, (1.2) in [8]. Since $(t-1)/s = (r-1)m/s - 1 = (r-1)\bar{d}/r - 1$ in our present notation, and $\Psi_r(1-\xi) = (r-1)\Phi_r(\xi)/r - 1$, the quantities ξ and ρ here and in [8] are defined from the average degree \bar{d} in exactly the same way. Here we work mostly with ξ rather than ρ since in the dense case $\xi \rightarrow 0$, which makes the asymptotics more intuitive. (In the sparse case $\rho \rightarrow 0$ instead.)

With $\xi = \Phi_r^{-1}(\bar{d})$ as above, since $\bar{d} = rm/s$ we may write

$$\exp(-sF_r(\bar{d})) = ((1-\xi)^{-r}(1-\xi^r))^m (\xi^{\xi/(1-\xi)}(1-\xi))^s.$$

Equivalently, setting $\rho = 1 - \xi$,

$$\exp(-sF_r(\bar{d})) = \left(\frac{1 - (1-\rho)^r}{\rho^r} \right)^m ((1-\rho)^{(1-\rho)/\rho} \rho)^s. \quad (7)$$

Thus we see that Theorem 1.1 of [8] says exactly that in the sparse case

$$P_r(s, m) \sim e^{r/2 + \mathbf{1}_{r=2}} \sqrt{\frac{3(r-1)}{2}} \exp(-sF_r(\bar{d})). \quad (8)$$

Turning to the middle range, as noted in the appendix to [8], the quantity called $\Phi_d(r, \zeta)$ in [5] is exactly the right-hand side of (7) above, i.e., simply $\exp(-sF_r(\bar{d}))$ in our present notation. Behrisch, Coja-Oghlan and Kang [5] write ζ for $\bar{d} = rm/s$, d for the uniformity (r here) and r for what we call ξ . In our notation, their main result says that in the middle range we have

$$P_r(s, m) \sim G_r(\bar{d}) \exp(-sF_r(\bar{d})), \quad (9)$$

with

$$G_r(\bar{d}) = \frac{a_r(\bar{d})}{\sqrt{b_r(\bar{d})}} e^{g_r(\bar{d})}, \quad (10)$$

where, for $r \geq 3$,

$$\begin{aligned} a_r(\bar{d}) &= 1 - \xi^r - (1-\xi)(r-1)\bar{d}\xi^{r-1}, \\ b_r(\bar{d}) &= (1 - \xi^r + \bar{d}(r-1)(\xi - \xi^{r-1}))(1 - \xi^r) - r\bar{d}\xi(1 - \xi^{r-1})^2, \\ g_r(\bar{d}) &= \frac{(r-1)\bar{d}(\xi - 2\xi^r + \xi^{r-1})}{2(1 - \xi^r)}, \end{aligned}$$

and

$$\begin{aligned} a_2(\bar{d}) &= 1 + \xi - \bar{d}\xi, \\ b_2(\bar{d}) &= (1 + \xi)^2 - 2\bar{d}\xi, \quad \text{and} \\ g_2(\bar{d}) &= \frac{2\bar{d}\xi + \bar{d}^2\xi}{2(1 + \xi)}. \end{aligned}$$

Since $\xi \sim e^{-\bar{d}}$ as $\bar{d} \rightarrow \infty$, it is trivial to check that as $\bar{d} \rightarrow \infty$ we have $a_r(\bar{d}) \rightarrow 1$, $b_r(\bar{d}) \rightarrow 1$ and $g_r(\bar{d}) \rightarrow 0$. Hence

$$G_r(\bar{d}) \rightarrow 1 \quad \text{as } \bar{d} \rightarrow \infty, \quad (11)$$

and (9) coincides with our much simpler formula $\exp(-sF_r(\bar{d}))$ in this case.

Finally, we noted in the appendix to [8] (see equations (A.5) and (A.6)) that as $\bar{d} \rightarrow r/(r-1)$ (corresponding to the nullity t satisfying $t = o(s)$) then $\xi \rightarrow 1$ and $G_r(\bar{d}) \rightarrow e^{r/2 + \mathbf{1}_{r=2} \sqrt{\frac{3(r-1)}{2}}}$, corresponding to the pre-factor in (8). Collecting together these results, we have the following extension of the Bender–Canfield–McKay formula [6] to hypergraphs.

Theorem 1.4. *Let $r \geq 2$ be fixed, and let $m = m(s)$ satisfy $m - s/(r-1) \rightarrow \infty$ and $m \leq \binom{s}{r}$. Then the proportion $P_r(s, m)$ of m -edge r -uniform hypergraphs on $[s]$ that are connected satisfies*

$$P_r(s, m) \sim G_r(\bar{d}) \exp(-sF_r(\bar{d})) \quad (12)$$

where $\bar{d} = rm/s$ is the average degree, and $F_r(\bar{d})$ and $G_r(\bar{d})$ are defined in (3) and (10), with $\xi = \xi(r, \bar{d})$ defined in (2).

Proof. Passing to a subsequence we may assume that the average degree \bar{d} satisfies one of the conditions (i) $\bar{d} \rightarrow r/(r-1)$, (ii) $\bar{d} \rightarrow c$ with $r/(r-1) < c < \infty$, or (iii) $\bar{d} \rightarrow \infty$. Then we apply Theorem 1.1 of [8] (in the form (8) above), Theorem 1.1 of [5], or Theorem 1.1 above, recalling (11). \square

In the Appendix, we show that the $r = 2$ case of the formula (12) does indeed coincide with the (very different looking) formula in [6]. Note that while one *can* use the same formula in all cases (sparse, middle and dense), in the sparse and dense cases this does not make much sense in practice, since the formula simplifies greatly in these cases. Note also that while we have shown that the Behrisch–Coja-Oghlan–Kang formula applies in the sparse and dense ranges too, so far as we know their *proof* does not adapt to these cases. Indeed, the sparse case treated in [8] seems to require more complicated arguments despite the relative simplicity of the formula.

1.2 Probabilistic reformulation

We shall prove Theorem 1.1 (which, despite the trivial use of probability in the statement, is a purely enumerative result) by probabilistic methods. Indeed, as we shall see in Section 6, up to a (rather lengthy) calculation, Theorem 1.1 is essentially equivalent to a local limit theorem for the numbers of vertices and edges in the largest component of a certain random hypergraph. This is a similar situation to that in [8].

Turning to the details, given $r \geq 2$, $n \geq 1$ and $0 < p < 1$, let $H_{n,p}^r$ be the random r -uniform hypergraph on $[n]$ in which each of the $\binom{n}{r}$ possible edges is included with probability p , independently of the others. From now until

Section 6 (where we return to the enumerative viewpoint) we fix $r \geq 2$ and consider a function $d = d(n)$ satisfying

$$d \rightarrow \infty \quad \text{and} \quad \log n - d \rightarrow \infty, \quad (13)$$

as $n \rightarrow \infty$. We consider the random hypergraph $H_{n,p}^r$ with

$$p = p(n) = d \frac{(r-1)!}{n^{r-1}}, \quad (14)$$

noting that the expected degree of a vertex is $p \binom{n-1}{r-1} = d(1 + O(1/n))$. Writing

$$\mu_0 = ne^{-d},$$

then μ_0 is roughly the expected number of isolated vertices in $H_{n,p}^r$; the significance of the second condition in (13) is that it implies $\mu_0 \rightarrow \infty$ as $n \rightarrow \infty$.

Given $d > 1/(r-1)$, define $\xi = \xi(d) \in (0, 1)$ by

$$\xi = \exp(-d(1 - \xi^{r-1})). \quad (15)$$

Since $\log(1/\xi)/(1 - \xi^{r-1})$ is strictly decreasing, this uniquely defines ξ . In fact, ξ is the extinction probability of a certain branching process naturally associated to the neighbourhood exploration process in $H_{n,p}^r$; see Section 2. It is not hard to see that $\xi \rightarrow 0$ as $d \rightarrow \infty$. Substituting this back into (15), it follows that $\xi = e^{-d+o(d)}$. From this it follows that $d\xi^{r-1} = o(1)$ and hence, by (15) again,

$$\xi \sim e^{-d} \quad \text{as} \quad d \rightarrow \infty. \quad (16)$$

When we come to relate the probabilistic and enumerative viewpoints, the average degree parameters d and \bar{d} will not be (quite) equal (and neither will the numbers of vertices, n and s). However, for d and \bar{d} related as they will be, $\xi(\bar{d})$ as defined in the previous section and $\xi(d)$ as defined above will coincide.

For further background on the phase transition in the component structure of $H_{n,p}^r$ see, for example, Section 2 of [8]. Here, we are well above the critical edge density. As is well known (and we shall show below), when $d \rightarrow \infty$ then with very high probability $H_{n,p}^r$ has a (necessarily unique) ‘giant’ component containing almost all the vertices – in fact, all but around μ_0 vertices.

We say that a sequence $((X_n, Y_n))$ of random variables taking values in \mathbb{Z}^2 satisfies a *local limit theorem* with parameters $(\mu_X(n), \mu_Y(n))$ and $(\sigma_X^2(n), \sigma_Y^2(n))$ if for any sequence $(x_n, y_n) \in \mathbb{Z}^2$ we have

$$\mathbb{P}((X_n, Y_n) = (x_n, y_n)) = \frac{1}{2\pi\sigma_X\sigma_Y} \left(\exp\left(-\frac{(x_n - \mu_X)^2}{2\sigma_X^2} - \frac{(y_n - \mu_Y)^2}{2\sigma_Y^2}\right) + o(1) \right),$$

where we have partially suppressed the dependence on n in the notation. Note that, considering ‘almost worst case’ values of x_n and y_n , the $o(1)$ term can be taken to be uniform over all $(x_n, y_n) \in \mathbb{Z}^2$.

Let $L_1(H)$ and $M_1(H)$ denote the numbers of vertices and edges in the largest component of a hypergraph H , where ‘largest’ means with the most vertices, and we break ties arbitrarily. Here, then, is our local limit theorem for the number of vertices and edges in the giant component of $H_{n,p}^r$.

Theorem 1.5. *Let $r \geq 2$ be fixed, let $d = d(n) \rightarrow \infty$ with $\log n - d \rightarrow \infty$, and set $p = p(n) = d \frac{(r-1)!}{n^{r-1}}$. Let $L_1 = L_1(H_{n,p}^r)$ and $M_1 = M_1(H_{n,p}^r)$. Then we have*

$$\mathbb{E}[L_1] = (1 - \xi)n + o(1) \quad \text{and} \quad \mathbb{E}[M_1] = \frac{d(1 - \xi^r)}{r}n + O(d), \quad (17)$$

where $\xi = \xi(d)$ is defined in (15), and

$$\text{Var}[L_1] \sim \sigma_L^2 \quad \text{and} \quad \text{Var}[M_1] \sim \sigma_M^2 \quad (18)$$

where

$$\sigma_L^2 = \sigma_L^2(n) = ne^{-d} \quad \text{and} \quad \sigma_M^2 = \sigma_M^2(n) = \frac{dn}{r}. \quad (19)$$

Furthermore, the pair (L_1, M_1) satisfies a local limit theorem with parameters $(\mathbb{E}[L_1], \mathbb{E}[M_1])$ and (σ_L^2, σ_M^2) .

Remark 1.6. In the light of (17) and (18), it is easy to check that we may replace the parameters for the means in the local limit theorem above by $((1 - \xi)n, d(1 - \xi^r)n/r)$; see Lemma 5.7 below. The key point is that the error terms $o(1)$ and $O(d)$ in (17) are (much) smaller than σ_L and σ_M , respectively.

Although we are not aware of such accurate estimates for $\mathbb{E}[L_1]$ and $\mathbb{E}[M_1]$ in the literature, these are relatively straightforward. The main point of Theorem 1.5 is the local limit theorem. Analogous results for the sparse regime ($d = 1/(r-1) + o(1)$) and the ‘middle range’ $d = 1/(r-1) + \Theta(1)$ were proved in [8] and by Behrisch, Coja-Oghlan and Kang [4]. Indeed, recalling that the ‘nullity’ N_1 studied in [8] is defined to be $(r-1)M_1 - L_1 + 1$, after a little manipulation it is not hard to check that the quantities $\rho_{r,\lambda}$ and $\rho_{r,\lambda}^*$ appearing in [8] as approximations to $\mathbb{E}[L_1]/n$ and $\mathbb{E}[M_1]/n$ correspond to $(1 - \xi)$ and $(r-1)d(1 - \xi^r)/r - (1 - \xi)$ here. In other words, the formulae for the means match up; the asymptotics of the variances are different in the different ranges considered here and in [8].

The rest of the paper is organized as follows. In Sections 2–4 we prepare the ground for the proof of Theorem 1.5. These sections contain lemmas concerning, respectively, a certain branching process, basic properties of $H_{n,p}^r$, and the mean and variance of L_1 and M_1 . Then, in Section 5 (the heart of the paper), we use ‘smoothing’ arguments to prove Theorem 1.5. Finally, in Section 6 we deduce Theorem 1.1 via a somewhat involved calculation. In the Appendix, we compare the $r = 2$ case of our enumerative formula with that of Bender, Canfield and McKay.

2 Branching process preliminaries

Given an integer $r \geq 2$ and a real number $d > 0$, let $\mathfrak{X}_{r,d}$ be the Galton–Watson branching process defined as follows. Start in generation 0 with one individual. Each individual in generation t has a random number of *groups* of $r-1$ children, with the number of groups having a Poisson distribution $\text{Po}(d)$. These children

make up generation $t+1$. The numbers of children of all individuals in generation t are independent of each other and of the history.

The branching process $\mathfrak{X}_{r,d}$ can be naturally viewed as a (possibly infinite) r -uniform hypergraph, with a vertex for each individual, and a hyperedge for each group of children, consisting of these children together with their parent. This hypergraph is of course an r -tree, by which we simply mean an r -uniform hypergraph that is a tree. (Often, when there is no danger of confusion, we simply write ‘tree’.) We write $|\mathfrak{X}_{r,d}|$ and $e(\mathfrak{X}_{r,d})$ for the number of vertices and edges in this r -tree, noting that

$$|\mathfrak{X}_{r,d}| = 1 + (r-1)e(\mathfrak{X}_{r,d}) \leq \infty.$$

Let $\rho = \rho_{r,d} = \mathbb{P}(|\mathfrak{X}_{r,d}| = \infty)$ be the probability that the branching process $\mathfrak{X}_{r,d}$ *survives* forever, and $\xi = 1 - \rho$ its *extinction probability*. Elementary properties of branching processes (see, e.g., Athreya and Ney [2]) imply that ξ is the smallest solution in $[0, 1]$ to the equation (15). Moreover, if the branching factor $\lambda = (r-1)d$ is strictly greater than 1, then (15) has a unique solution in $[0, 1)$, and in particular $\xi < 1$; otherwise, $\xi = 1$. Indeed, ξ^{r-1} is the probability that all $r-1$ children in a given group lead to finite trees, so $1 - \xi^{r-1}$ is the probability that a given group *survives*, i.e., has infinitely many descendants. From thinning properties of Poisson distributions, the number of such surviving groups of children of the root is Poisson with mean $d(1 - \xi^{r-1})$, and ξ is exactly the probability that this number is 0. Hence ξ satisfies (15); it is not hard to see that ξ is indeed the smallest solution to this equation.

For comparison with the results in [8], in terms of $\rho = 1 - \xi$ we may write (15) as

$$1 - \rho = \exp(-d(1 - (1 - \rho)^{r-1})), \quad (20)$$

which, writing λ for $(r-1)d$, matches the definition given by (2.1) and (2.2) in [8]. Here, where d tends to infinity, we will have $\xi \rightarrow 0$, so ξ is a more natural parameter to work with than ρ . This contrasts with the situation in [8], where $\rho \rightarrow 0$.

2.1 The dual process

In the branching process $\mathfrak{X}_{r,d}$, the *groups* of children of the root may be classified into two types as above; those that survive, and those that do not. By thinning properties of Poisson distributions, the number of groups that do not survive has a Poisson distribution with mean $d\xi^{r-1}$, and is independent of the number that do survive. For $d > 1/(r-1)$ (the supercritical case) define the *dual parameter* d^* by

$$d^* = d\xi^{r-1}. \quad (21)$$

Then it follows that the conditional distribution of $\mathfrak{X}_{r,d}$ given $|\mathfrak{X}_{r,d}| < \infty$ is exactly the unconditional distribution of the *dual process* \mathfrak{X}_{r,d^*} . It is not hard to check that the dual parameter coincides with that defined in [8]; the key point is that (21) and (15) imply $d^*e^{-(r-1)d^*} = de^{-(r-1)d}$.

2.2 Point probabilities

For $0 \leq k < \infty$ define

$$\pi_k = \pi_{k,r,d} = \mathbb{P}(e(\mathfrak{X}_{r,d}) = k).$$

The next lemma is a standard calculation, specialized to the particular offspring distribution we have here. In this lemma, and much of this and the next two sections, we adopt the convention of writing $s = 1 + (r - 1)k$ for the number of vertices of an r -tree with k edges; this will make the formulae concerning ‘small’ tree components much more concise.

Lemma 2.1. *For any $r \geq 2$, $d > 0$ and $k \geq 0$ we have*

$$\pi_k = \pi_{k,r,d} = \frac{1}{s} \mathbb{P}(\text{Po}(ds) = k) = \frac{s^{k-1} d^k}{k!} e^{-ds},$$

where $s = 1 + (r - 1)k$.

Proof. Consider the following alternative way of generating a random rooted r -tree. Let (a_1, a_2, \dots) be independent and identically distributed, with $a_i \sim \text{Po}(d)$. Start with the root in generation 0. Construct a_1 groups of $r - 1$ children of the root. Then proceed through these children one-by-one, assigning each child a_i groups of $r - 1$ children of its own, for $i = 2, 3, \dots$. Continue with the new individuals (if any) in generation 2, and so on. By the definition of $\mathfrak{X}_{r,d}$ this tree has the same distribution as $\mathfrak{X}_{r,d}$. This construction stops if and only if, for some i , the first i individuals ‘explored’ have between them at most, and hence exactly, $i - 1$ children, i.e., $\frac{i-1}{r-1}$ groups of children. This can only happen for $i \equiv 1$ modulo $r - 1$. For $k \geq 0$ and $s = 1 + (r - 1)k$, let \mathcal{S}_k be the set of sequences $(a_i)_{1 \leq i \leq s}$ of non-negative integers with the properties

- (i) $\sum_{i \leq s} a_i = k$ and
- (ii) for all $1 \leq j < s$, $\sum_{i \leq j} (r - 1)a_i \geq j$.

Note that condition (i) may be written as $\sum_{i \leq s} (r - 1)a_i = s - 1$. From the construction above, $e(\mathfrak{X}_{r,d}) = k$ if and only if our random sequence (a_i) starts with a sequence in \mathcal{S}_k . By Spitzer’s Lemma, the probability of this is exactly

$$s^{-1} \mathbb{P}\left(\sum_{i \leq s} a_i = k\right).$$

Indeed, given any sequence $(a_i)_{i \leq s}$ with $\sum_{i \leq s} a_i = k$, there is a unique ‘rotation’ giving a sequence that also satisfies (ii) above, and since the a_i are i.i.d., a sequence and its rotations are equiprobable.

Since $\sum_{i \leq s} a_i$ has a Poisson distribution with mean ds , the result follows. \square

We state a trivial consequence for later reference.

Corollary 2.2. *Let $r \geq 2$ be fixed. There is a constant d_0 such that for all $d \geq d_0$ and $k \geq 0$ we have*

$$\pi_{k,r,d} \leq e^{-d(s+1)/2},$$

where $s = 1 + (r - 1)k$.

Proof. For $k \geq 1$ we have $s/k \leq r$, say. Since $k! \geq (k/e)^k$ it follows that

$$\begin{aligned}\pi_{k,r,d} &= \frac{s^{k-1}d^k}{k!}e^{-ds} = \frac{e^{-d}}{sk!}(sde^{-(r-1)d})^k \leq e^{-d} \left(\frac{esd}{k} e^{-(r-1)d} \right)^k \\ &\leq e^{-d}(erde^{-(r-1)d})^k \leq e^{-d}e^{-(r-1)dk/2} = e^{-d(s+1)/2},\end{aligned}$$

choosing d_0 so that $d \geq d_0$ implies $erde^{-(r-1)d/2} \leq 1$ for all $d \geq d_0$. Of course, the final bound holds for $k = 0$ (i.e., $s = 1$), since $\pi_{0,r,d} = e^{-d}$. \square

The key consequence of Corollary 2.2 is that the values π_k decrease rapidly to zero starting from $\pi_0 = e^{-d}$.

In the subcritical case, we have the following simple result.

Lemma 2.3. *For any $r \geq 2$ and $d \geq 0$ with $(r-1)d < 1$ we have*

$$\mathbb{E}[|\mathfrak{X}_{r,d}|^{-1}] = \sum_{k \geq 0} s^{-1} \pi_k = 1 - (r-1)d/r,$$

where, as usual, $s = 1 + (r-1)d$.

Proof. Doubtless there is a direct algebraic proof of this. We outline a different argument: with r and d fixed, consider the subcritical random hypergraph $H_{n,p}^r$, $p = d(r-1)!n^{-(r-1)}$. It is easy to check that for each fixed $k \geq 0$, the expected number of k -edge tree components is $s^{-1}\pi_k n + o(n)$. (Either directly calculate the expectation or, using a standard coupling argument, note that $\pi_k n$ approximates the expected number of vertices in components that are k -edge trees. The factor $1/s$ arises from counting components rather than vertices.) Since the hypergraph is subcritical, for any $K(n)$ tending to infinity, the expected number of components with at least $K(n)$ vertices is $o(n)$, as is the expected number of non-tree components. Since $\sum_k s^{-1}\pi_k$ converges, it follows that the expectation μ_n of the number of components satisfies $\mu_n = n \sum_k s^{-1}\pi_k + o(n)$. On the other hand, adding edges one-by-one, the expected number of edges forming cycles is small, and when an edge does not form a cycle the number of components goes down by $r-1$. There are $dn/r + o(n)$ edges, so $\mu_n = n(1 - (r-1)d/r) + o(n)$. Combining these expressions gives the result, noting that the final statement does not involve n . \square

3 Random hypergraph preliminaries

We start with a trivial lower bound on the number L_1 of vertices in the largest component of the random hypergraph $H_{n,p}^r$ defined in Section 1.2. Throughout, $r \geq 2$ is fixed. Recall that $p = p(n) = d \frac{(r-1)!}{n^{r-1}}$ where $d = d(n) \rightarrow \infty$ and $\log n - d \rightarrow \infty$. Set

$$s_1 = s_1(n) = 100 \max\{ne^{-d}, \log n\}, \quad (22)$$

ignoring the irrelevant rounding to integers. Note for later that

$$ds_1 = o(n), \quad (23)$$

since $de^{-d} \rightarrow 0$ and $d < \log n$ for n large enough.

Lemma 3.1. *Let $d = d(n) \rightarrow \infty$ with $\log n - d \rightarrow \infty$, and define p as in (14) and s_1 as in (22). Then*

$$\mathbb{P}(L_1(H_{n,p}^r) \leq n - s_1) = o(n^{-100}).$$

Proof. Noting that $s_1 \leq n/4$, say, for n large enough, it is easy to see that if $L_1 \leq n - s_1$ then there is a vertex cut $[n] = A \cup A^c$ with $s_1 \leq |A| \leq n/2$ such that no hyperedge meets both A and A^c . For a given value of $a = |A|$, considering only potential hyperedges with one vertex in A and the others in A^c , the expected number of such cuts is crudely at most

$$\nu_a = \binom{n}{a} (1-p)^{a \binom{n-a}{r-1}} \leq \left(\frac{en}{a}\right)^a \exp\left(-pa \binom{n-a}{r-1}\right).$$

Now $\binom{n-a}{r-1} = \frac{n^{r-1}}{(r-1)!} (1 + O(a/n))$. Hence there is a constant $c > 0$ such that for $a \leq cn/d$ we have $\binom{n-a}{r-1} \geq \frac{n^{r-1}}{(r-1)!} (1 - 1/d)$, say, and thus

$$\nu_a \leq \left(\frac{en}{a}\right)^a \exp\left(-a \frac{pn^{r-1}}{(r-1)!} (1 - 1/d)\right) = \left(\frac{e^2 ne^{-d}}{a}\right)^a \leq e^{-2a} \leq n^{-200},$$

using $a \geq 100ne^{-d}$ and $a \geq 100 \log n$ in the last two steps. On the other hand, since $a \leq n/2$, for $a \geq cn/d$ we still have $\binom{n-a}{r-1} \geq 2^{-r} \frac{n^{r-1}}{(r-1)!}$ and so

$$\nu_a \leq \left(\frac{en}{a} e^{-d/2^r}\right)^a \leq \left(\frac{ed}{c} e^{-d/2^r}\right)^a \leq e^{-2a}$$

if n is large enough, recalling that $d = d(n) \rightarrow \infty$. It follows easily that $\sum_{s_1 \leq a \leq n/2} \nu_a = o(n^{-100})$, so the probability that there is a cut as described is $o(n^{-100})$. \square

We have shown that, up to a negligible error probability, the total size of all components with at most $n/2$ vertices is at most s_1 . In particular, there are no individual components with more than s_1 vertices other than the unique giant component. We shall now show that with high probability the giant component is the only component containing cycles. Furthermore, it is extremely unlikely that there is a cycle in a component of size between around $1000(\log n)/d$, say, and $n/2$.

Set

$$s_0 = \left\lceil \frac{1000 \log n}{d} \right\rceil \geq 1000. \quad (24)$$

Recall that we assume that $\log n - d \rightarrow \infty$. Since de^{-d} is a decreasing function of d for $d \geq 1$, it follows that

$$\frac{nde^{-d}}{\log n} \rightarrow \infty. \quad (25)$$

Hence $s_0 = o(s_1)$ and in particular, for n large enough, $s_0 < s_1$.

Lemma 3.2. *Under the assumptions above, the probability that $H_{n,p}^r$ contains a non-tree component with at most s_0 vertices is $o(1)$. Furthermore, if X is the total number of edges in such components, then $\mathbb{E}[X^2] = o(1)$. Finally, the probability that there is a non-tree component with more than s_0 but fewer than $n/2$ vertices is $o(n^{-99})$.*

Proof. We start with the second statement, which implies the first. We aim for simplicity rather than a strong bound. If a non-tree component has k edges and v vertices, then $v \leq (r-1)k$. Hence $k \geq v/(r-1)$. Let C_v be the number of v -element subsets S of $[n]$ with the following properties: no edge of $H_{n,p}^r$ meets both S and S^c , and S spans at least $v/(r-1)$ edges of $H_{n,p}^r$. The vertex set of any non-tree component is such a set for some $v \geq r$. Hence the number of non-tree components of $H_{n,p}^r$ with at most s_0 vertices is at most

$$\sum_{v=r}^{s_0} C_v.$$

Suppose that e and f are (not necessarily distinct) edges of $H_{n,p}^r$ both in non-tree components with at most s_0 vertices, with vertex sets S_1 and S_2 , say. Then S_i spans at least $|S_i|/(r-1)$ edges so (since the S_i are equal or disjoint), $S = S_1 \cup S_2$ has the properties above. Since a set of v vertices spans (very crudely) at most v^r edges, we thus have

$$X^2 \leq \sum_{v=r}^{2s_0} v^{2r} C_v.$$

For $r \leq v \leq 2s_1$ let $k = \lceil v/(r-1) \rceil = \Theta(v)$. Very crudely,

$$\mathbb{E}[C_v] \leq \binom{n}{v} \binom{\binom{v}{r}}{k} p^k (1-p)^{v \binom{n-v}{r-1}},$$

where the last factor accounts for the fact that there are no edges consisting of one vertex in the set S and $r-1$ vertices outside. Using the very crude bound $\binom{v}{r} \leq v^r$, and the slightly more careful bound $\binom{n-v}{r-1} = \frac{n^{r-1}}{(r-1)!} (1 + O(v/n))$ together with the inequality $1-p \leq e^{-p}$, it follows that

$$\begin{aligned} \mathbb{E}[C_v] &\leq \left(\frac{en}{v}\right)^v \left(\frac{ev^r}{k}\right)^k p^k \exp\left(-\frac{pvn^{r-1}}{(r-1)!} (1 + O(v/n))\right) \\ &= \frac{e^{v+k} n^v v^{rk}}{v^v k^k} d^k n^{-(r-1)k} (r-1)!^k \exp(-dv + O(dv^2/n)). \end{aligned}$$

Since r is constant and $v = \Theta(k)$, for some constant B depending only on r we thus have

$$\mathbb{E}[C_v] \leq B^k \frac{n^{v-(r-1)k}}{k^{v-(r-1)k}} d^k \exp(-dv + O(dv^2/n)).$$

Recall (from (23)) that $v \leq 2s_1 = o(n/d) = o(n)$. For n large enough, in the range $r \leq v \leq 2s_1$ we thus have

$$\mathbb{E}[C_v] \leq B^k \frac{n^{v-(r-1)k}}{k^{v-(r-1)k}} d^k \exp(-dv/2). \quad (26)$$

Since $(n/k)^{r-1} d^{-1} \geq (n/k) d^{-1} \geq n/(s_1 d) \rightarrow \infty$, if v is not a multiple of $r-1$, then replacing $k = \lceil v/(r-1) \rceil$ by $k' = v/(r-1)$ in the right-hand side of (26) can only increase it, so

$$\mathbb{E}[C_v] \leq B^{v/(r-1)} d^{v/(r-1)} \exp(-dv/2) = ((Bd)^{1/(r-1)} e^{-d/2})^v \leq e^{-dv/4}$$

if n is large enough. Then

$$\sum_{v=r}^{2s_0} v^{2r} \mathbb{E}[C_v] \leq \sum_{v=r}^{\infty} v^{2r} e^{-dv/4} \sim r^{2r} e^{-dr/4} \rightarrow 0,$$

and the bound on $\mathbb{E}[X^2]$ follows.

For the final statement, simply note that

$$\sum_{v=s_0}^{s_1} \mathbb{E}[C_v] \leq \sum_{v=s_0}^{\infty} e^{-dv/4} \sim e^{-ds_0/4} = o(n^{-100})$$

by choice of s_0 , and apply Lemma 3.1 to deal with sizes between s_1 and $n/2$. \square

3.1 Small tree components

Let $T_k = T_k(H_{n,p}^r)$ denote the number of components of $H_{n,p}^r$ that are trees with k edges, and so $s = 1 + (r-1)k$ vertices. Then sT_k is the number of vertices in such components. We already know that with very high probability there are no components with more than s_1 vertices other than the giant component; we shall see that with very high probability there are no tree components with more than s_0 vertices. Recalling our convention of writing $s = 1 + (r-1)k$, set

$$k_0 = \frac{s_0 - 1}{r - 1},$$

ignoring the rounding to integers. Define π_k as in Section 2.2.

Lemma 3.3. *For $0 \leq k = k(n) \leq k_0$ we have*

$$\mathbb{E}[T_k] = n \frac{\pi_k}{s} (1 + O(d^2 s^2 / n)),$$

while the expected number of tree components with between s_0 and $n/2$ vertices is $o(n^{-99})$.

Proof. Let $k \geq 0$ and set $s = 1 + (r-1)k$. We assume throughout the rather weak bound $s \leq s_1$; Lemma 3.1 implies that the expected number of tree components with between s_1 and $n/2$ vertices is $o(n^{-99})$. By a result of Selivanov [11] (see also [10, 8]), there are exactly

$$n_k = s^{k-1} \frac{(s-1)!}{k!(r-1)!^k}$$

k -edge r -trees on a given set of s vertices. Clearly,

$$\mathbb{E}[T_k] = \binom{n}{s} n_k p^k (1-p)^{\binom{n}{r} - \binom{n-s}{r} + k}.$$

Since $\binom{n}{s} = \frac{n^s}{s!} \exp(O(s^2/n))$, we have

$$\binom{n}{s} n_k = \frac{n^s s^{k-2}}{k!(r-1)!^k} \exp(O(s^2/n)).$$

Recalling that $p = d(r-1)!n^{-(r-1)}$, it follows that

$$\mathbb{E}[T_k] = \frac{n^{s-(r-1)k} s^{k-2}}{k!} d^k (1-p)^{\binom{n}{r} - \binom{n-s}{r} + k} \exp(O(s^2/n)). \quad (27)$$

Now $\binom{n}{r} - \binom{n-s}{r} = sn^{r-1}/(r-1)! + O(s^2 n^{r-2})$. Since $k \leq s = O(s^2 n^{r-2})$, we thus have

$$\binom{n}{r} - \binom{n-s}{r} + k = \frac{sn^{r-1}}{(r-1)!} (1 + O(s/n)).$$

Since $p = O(d/n^{r-1}) = O(d/n)$, we have $-\log(1-p) = p + O(p^2) = p(1 + O(d/n))$. Thus

$$\begin{aligned} (1-p)^{\binom{n}{r} - \binom{n-s}{r} + k} &= \exp\left(-p \frac{sn^{r-1}}{(r-1)!} (1 + O((d+s)/n))\right) \\ &= \exp(-ds(1 + O(d+s)/n)) = \exp(-ds + O(d^2 s^2/n)), \end{aligned} \quad (28)$$

where in the second step we used again the definition $p = d(r-1)!n^{-(r-1)}$, and in the last step we bounded $d+s$ by ds just to keep the formula compact. Since $s - (r-1)k = 1$, combining (27) and (28) we obtain

$$\mathbb{E}[T_k] = n \frac{s^{k-2} d^k}{k!} \exp(-ds + O(d^2 s^2/n)) = n \frac{\pi_k}{s} \exp(O(d^2 s^2/n)), \quad (29)$$

using Lemma 2.1 in the last step.

For $s \leq s_0$ we have $d^2 s^2/n = O(d^2 s_0^2/n) = O((\log n)^2/n) = o(1)$, so we may write the $\exp(O(d^2 s^2/n))$ error term as $1 + O(d^2 s^2/n)$, giving the result.

For $s_0 < s \leq s_1$, from (29), Corollary 2.2 and the bound $ds/n \leq ds_1/n = o(1)$ (see (23)), we have

$$\mathbb{E}[T_k] \leq n \exp(-ds/2 + O(d^2 s^2/n)) \leq n \exp(-ds/4)$$

if n is large enough. Since $ds \geq ds_0 \geq 1000 \log n$, this completes the proof. \square

Corollary 3.4. *With probability $1 - o(n^{-98})$ the hypergraph $H_{n,p}^r$ consists of a ‘giant’ component with at least $n - s_1 > n/2$ vertices together with ‘small’ components, each with at most s_0 vertices.*

Proof. Immediate from Lemmas 3.1, 3.2 and 3.3. \square

4 Key parameters

As in the previous section, fix $r \geq 2$, let $d = d(n) \rightarrow \infty$ with $\log n - d \rightarrow \infty$, as in (13), and set $p = p(n) = d \frac{(r-1)!}{n^{r-1}}$. Define $\xi = \xi(d)$ by (15), recalling that $\xi \sim e^{-d}$ as $d \rightarrow \infty$. Let L_1 and M_1 denote the numbers of vertices and edges in the largest component of the random hypergraph $H_{n,p}^r$, chosen according to any rule if there is tie. (With high probability there will not be a tie.)

Lemma 4.1. *Under the assumptions above we have*

$$\mathbb{E}[L_1] = n(1 - \xi) + o(1)$$

and

$$\text{Var}[L_1] \sim \mu_0 = ne^{-d}.$$

Recall that $\mu_0 = \mu_0(n) = ne^{-d}$ is roughly the expected number of isolated vertices in $H_{n,p}^r$. Since $\xi \sim e^{-d}$, the lemma says that isolated vertices give the dominant contribution to $\mathbb{E}[n - L_1]$, and (roughly speaking) that the Poisson-type distribution of the number of isolated vertices is the dominant contribution to $\text{Var}[L_1]$. The bound $\mathbb{E}[L_1] = n - \mu_0 + o(\mu_0)$ would not be precise enough when we come to apply our local limit theorem; we need a bound with an error that is $o(\sqrt{\mu_0})$.

Proof. Let v_T and v_C denote the number of vertices in small tree and cyclic components, respectively, where ‘small’ means with at most s_0 vertices. By Corollary 3.4, with probability $1 - o(n^{-98})$ we have

$$L_1 = n - v_T - v_C. \tag{30}$$

Since all relevant quantities are bounded by n , it follows easily that

$$\mathbb{E}[L_1] = n - \mathbb{E}[v_T] - \mathbb{E}[v_C] + o(n^{-97}) = n - \mathbb{E}[v_T] + o(1), \tag{31}$$

using Lemma 3.2. Now

$$v_T = \sum_{k=0}^{k_0} s T_k$$

where T_k is the number of k -edge tree components, we write s for $1 + (r-1)k$ as usual, and $k_0 = (s_0 - 1)/(r-1)$. (We ignore the irrelevant rounding to integers.) By Lemma 3.3,

$$\mathbb{E}[v_T] = \sum_{k=0}^{k_0} n \pi_k (1 + O(d^2 s^2 / n)).$$

Now $\pi_0 = e^{-d}$ and, by Corollary 2.2, $\pi_k \leq e^{-d(s+1)/2}$. Hence

$$\sum_{k=0}^{\infty} \pi_k s^2 = e^{-d} + O(e^{-d(r+1)/2}) \sim e^{-d}. \quad (32)$$

Also $\sum_{k>k_0} \pi_k = o(n^{-99})$. It follows that

$$\mathbb{E}[v_T] = n \sum_{k=0}^{\infty} \pi_k + O(d^2 e^{-d}) = n\xi + o(1), \quad (33)$$

which, with (31) proves the first statement of the lemma.

Turning to the variance, from (30) we have

$$\text{Var}[L_1] = \text{Var}[v_T + v_C] + o(1) = \text{Var}[v_T] + \text{Var}[v_C] + \text{Cov}[v_C, v_T] + o(1). \quad (34)$$

Let $T_{k,k'}$ denote the number of ordered pairs of distinct tree components where the first has k edges and the second k' . Writing $s = 1 + (r-1)k$ and $s' = 1 + (r-1)k'$, and considering separately pairs of vertices in the same or distinct tree components, we have

$$\mathbb{E}[v_T^2] = \mathbb{E} \left[\left(\sum_{k=0}^{k_0} s T_k \right)^2 \right] = \sum_{k \leq k_0} s^2 \mathbb{E}[T_k] + \sum_{k, k' \leq k_0} s s' \mathbb{E}[T_{k,k'}]. \quad (35)$$

By Lemma 3.3 again,

$$\sum_{k \leq k_0} s^2 \mathbb{E}[T_k] = n \sum_{k=0}^{k_0} s \pi_k (1 + O(d^2 s^2/n)) \sim n \pi_0 = \mu_0, \quad (36)$$

using the rapid decrease of the π_k for the last approximation. On the other hand, writing $m_{s,s'}$ for the number of potential hyperedges that meet both a given set of s vertices and a given disjoint set of s' vertices, we have

$$\mathbb{E}[T_{k,k'}] = \mathbb{E}[T_k] \mathbb{E}[T_{k'}] (1-p)^{-m_{s,s'}}.$$

Since $m_{s,s'} \leq s s' n^{r-2}$, $p = O(d n^{-r+1})$, and $s s' d = o(n)$ for $s, s' \leq s_0$, it follows that

$$s s' \mathbb{E}[T_{k,k'}] = s \mathbb{E}[T_k] s' \mathbb{E}[T_{k'}] (1 + O(d s s' / n)) = n^2 \pi_k \pi_{k'} (1 + O(d^2 (s + s')^2 / n)),$$

by Lemma 3.3. Arguing as for (33) above, it follows that

$$\begin{aligned} \sum_{k, k' \leq k_0} s s' \mathbb{E}[T_{k,k'}] &= n^2 \sum_{k, k' \leq k_0} \pi_k \pi_{k'} + O(n d^2 e^{-2d}) \\ &= n^2 \xi^2 + o(n e^{-d}) = n^2 \xi^2 + o(n \xi). \end{aligned} \quad (37)$$

Indeed, from the rapid decay of π_k as k increases, the dominant contribution to the error term is from the case $k = k' = 0$; this contribution is $O(n^2\pi_0^2d^2/n) = O(ne^{-2d}d^2)$.

Putting the pieces together, from (35), (36), (37), (33) and the fact that $n\xi \sim ne^{-d} = \mu_0$, it follows that

$$\begin{aligned}\text{Var}[v_T^2] &= \mathbb{E}[v_T^2] - \mathbb{E}[v_T]^2 = (1 + o(1))\mu_0 + n^2\xi^2 + o(n\xi) - (n\xi + o(1))^2 \\ &= (1 + o(1))\mu_0.\end{aligned}\quad (38)$$

It remains only to note that from Lemma 3.2 we have $\text{Var}[v_C] \leq \mathbb{E}[v_C^2] = o(1)$ and hence $\text{Cov}[v_C, v_T] \leq (\text{Var}[v_C]\text{Var}[v_T])^{1/2} = o(\mu_0)$. Then, recalling (34), the result follows. \square

Lemma 4.2. *We have*

$$\mathbb{E}[M_1] = \frac{d(1 - \xi^r)}{r}n + O(d).$$

Proof. Calling a component ‘small’ if it has at most s_0 vertices, let e_T be the number of edges in small tree components and e_C the number in small cyclic components. By Corollary 3.4, with probability $1 - o(n^{-98})$ we have

$$M_1 = e(H_{n,p}^r) - e_T - e_C. \quad (39)$$

Since $\mathbb{E}[e(H_{n,p}^r)] = p\binom{n}{r} = dn/r + O(d)$, it follows that

$$\mathbb{E}[M_1] = dn/r - \mathbb{E}[e_T] - \mathbb{E}[e_C] + O(d). \quad (40)$$

Now

$$\mathbb{E}[e_C] \leq \mathbb{E}[e_C^2] = o(1) \quad (41)$$

by Lemma 3.2. On the other hand, writing $s = 1 + (r-1)k$ as usual, and setting $k_0 = (s_0 - 1)/(r-1)$,

$$\mathbb{E}[e_T] = \sum_{k=1}^{k_0} k\mathbb{E}[T_k] = \sum_{k=0}^{k_0} \frac{kn}{s}\pi_k(1 + O(d^2s^2/n)),$$

by Lemma 3.3. Using Corollary 2.2 as before to bound both the tail of the sum and the contribution from the $O(d^2s^2/n)$ term (see (32)), it follows that

$$\begin{aligned}(r-1)\mathbb{E}[e_T] &= n \sum_{k=0}^{\infty} \frac{(r-1)k}{s}\pi_k + o(1) \\ &= n \sum_{k=0}^{\infty} \frac{s-1}{s}\pi_k + o(1) = (\xi - \mathbb{E}[|\mathfrak{X}_{r,d}|^{-1}])n + o(1),\end{aligned}$$

since $\xi = \mathbb{P}(|\mathfrak{X}_{r,d}| < \infty) = \sum_k \pi_k$.

Recall from Section 2.1 that the conditional distribution of $|\mathfrak{X}_{r,d}|$ given that it is finite is exactly the distribution of $|\mathfrak{X}_{r,d^*}|$, where $d^* = d\xi^{r-1}$ is the dual parameter, as in (21). It follows by Lemma 2.3 that

$$\mathbb{E}[|\mathfrak{X}_{r,d}|^{-1}] = \mathbb{P}(|\mathfrak{X}_{r,d}| < \infty) \mathbb{E}[|\mathfrak{X}_{r,d^*}^{-1}|] = \xi(1 - (r-1)d^*/r).$$

Thus

$$\mathbb{E}[e_T] = \frac{\xi - \mathbb{E}[|\mathfrak{X}_{r,d}|^{-1}]}{r-1}n + o(1) = \frac{\xi d^*}{r}n + o(1). \quad (42)$$

From (40), (41) and (42) we have

$$\mathbb{E}[M_1] = \frac{d - \xi d^*}{r}n + O(d) = \frac{d(1 - \xi^r)}{r}n + O(d),$$

completing the proof. \square

Lemma 4.3. *We have*

$$\text{Var}[M_1] \sim \frac{dn}{r}.$$

Proof. Clearly, $\text{Var}[e(H_{n,p}^r)] = \binom{n}{r}p(1-p) \sim p\binom{n}{r} \sim \frac{dn}{r}$. Define e_T and e_C as in the proof of Lemma 4.2. We claim that

$$\text{Var}[e_T] = o(dn). \quad (43)$$

Let us first show that this implies the result. Indeed, from Lemma 3.2 we have $\text{Var}[e_C] \leq \mathbb{E}[e_C^2] = o(1) = o(dn)$. Hence $\text{Var}[e_T + e_C] = o(dn)$, and from (39) it easily follows that $\text{Var}[M_1] \sim \frac{dn}{r}$.

To establish (43) we argue as in the proof of (38), using $e_T = \sum_{k \leq k_0} kT_k$ in place of $v_T = \sum_{k \leq k_0} sT_k$. The argument is essentially identical, leading to the conclusion $\text{Var}[e_T] = \mathbb{E}[e_T] + o(n\xi)$. Since ξ and d^* are $o(1)$, from (42) we have $\mathbb{E}[e_T] = o(n)$. Hence $\text{Var}[e_T] = o(n) = o(dn)$. \square

5 Proof of Theorem 1.5

In this section we prove Theorem 1.5. This will require some further preparation.

Let $d = d(n)$ satisfy $d \rightarrow \infty$ and $\log n - d \rightarrow \infty$ as $n \rightarrow \infty$. Note that by (25), $nde^{-d}/\log n \rightarrow \infty$. Later, in various error terms we shall consider a function $\gamma(n)$ tending to zero slowly: pick $\gamma = \gamma(n)$ such that

$$\gamma \rightarrow 0, \quad \gamma d \rightarrow \infty \quad \text{and} \quad \frac{\gamma^2 nde^{-d}}{\log n} \rightarrow \infty \quad (44)$$

as $n \rightarrow \infty$.

As before, let $p = d(r-1)!n^{-r+1}$. Set $d_1 = \sqrt{d}$, and choose d_2 so that $p_1 + p_2 - p_1p_2 = p$ where $p_i = d_i(r-1)!n^{-r+1}$. Note that

$$d_1 \rightarrow \infty, \quad d_2 \sim d \quad \text{and} \quad d^2 e^{-d_1} \rightarrow 0. \quad (45)$$

Since $p_1 + p_2 = p(1 + O(dn^{-r+1})) = p(1 + O(d/n))$, we have

$$d_1 + d_2 = d + O(d^2/n) = d + o(\gamma), \quad (46)$$

since $d \leq \log n$ for n large and so (since $\gamma d \rightarrow \infty$) $\gamma \geq 1/\log n$.

Let H_1 and H_2 be independent random hypergraphs on the same vertex set $[n]$, with H_i having the distribution of H_{n,p_i}^r . Clearly, $H = H_1 \cup H_2$ has the distribution of $H_{n,p}^r$. We shall call the edges of H_1 *red* and those of H_2 *blue*. Note that there may be a red and a blue edge on the same set of r vertices.

The idea of the proof is as follows: we shall reveal the graph H_1 and some partial information about H_2 . We write the pair (L_1, M_1) as $(L, M) + (X, Y)$ where (L, M) is determined by the revealed information, and the conditional distribution of (X, Y) is with very high probability a fixed, very simple distribution. The latter distribution (essentially two independent binomial random variables) will satisfy a local limit theorem. We will also have $\text{Var}[X] \sim \text{Var}[L_1]$ and $\text{Var}[Y] \sim \text{Var}[M_1]$. This will easily imply that L and M are concentrated on the relevant scales, allowing us to transfer the local limit theorem to (L_1, M_1) . Related smoothing ideas were used in [9, 3, 8], though in a much more complicated way – there, part of the starting point was a central limit theorem. Here we do not need this, since our ‘smoothing distribution’ (X, Y) has asymptotically the entire variance of the original distribution.

Roughly speaking, the partial information will be as follows: we reveal H_1 and find with very high probability a very large connected component. We reveal all edges of H_2 except those of the following form: ones within the giant component of H_1 (‘internal edges’ below), and ones consisting of $r - 1$ vertices in this giant component and one vertex that is otherwise isolated (‘peripheral edges’). Then X and Y will be, roughly speaking, the numbers of peripheral and internal edges present. To obtain fixed distributions for X and Y we work with a subset of the giant component of H_1 of a fixed size $b = b(n)$, and a fixed number $i = i(n)$ of vertices outside on which we allow peripheral edges. We will need to show that with very high probability $L_1(H_1) \geq b$, and that there are enough of these outside vertices.

Turning to the details, note that since $d_1 = \sqrt{d} \leq \sqrt{\log n}$ we have $e^{-d_1} = n^{-o(1)}$, so

$$ne^{-d_1} = n^{1-o(1)}. \quad (47)$$

By analogy with (22), but replacing d by d_1 , define $s_{1,1} = 100ne^{-d_1}$ and set

$$b = \lceil n - s_{1,1} \rceil \sim n.$$

Let \mathcal{G}_1 be the ‘good’ event

$$\mathcal{G}_1 = \{L_1(H_1) \geq b\}.$$

By Lemma 3.1, applied with d_1 in place of d , we have

$$\mathbb{P}(\mathcal{G}_1^c) = o(n^{-100}). \quad (48)$$

Whenever \mathcal{G}_1 holds, let $B = B(H_1)$ be a set of b vertices from the largest component of H_1 , say the first b in numerical order. When \mathcal{G}_1 does not hold, we take $B = \emptyset$. (This is just so that B is always defined; we will never use B in this case.)

Let \mathcal{G}_2 be the event that $e(H_1) \leq n^{3/2}$, say. Since, crudely, $\mathbb{E}[e(H_1)] = p_1 \binom{n}{r} = O(n \log n)$, and $e(H_1)$ has a binomial distribution, we certainly have

$$\mathbb{P}(\mathcal{G}_2^c) = o(n^{-100}). \quad (49)$$

Set

$$S = \binom{b}{r} - \lceil n^{3/2} \rceil \sim \binom{n}{r}. \quad (50)$$

When $\mathcal{G}_1 \cap \mathcal{G}_2$ holds, we select a set E of r -element subsets of B , none of which is an edge of H_1 , with $|E| = S$. When $\mathcal{G}_1 \cap \mathcal{G}_2$ does not hold, set $E = \emptyset$. We call an edge e of H_2 *internal* if $e \in E$.

As a first step towards defining ‘peripheral’ edges, let us call a vertex $v \notin B$ *peripheral* if either v is isolated in H , or v is in precisely one edge e , and that edge e is blue and consists of v together with $r-1$ vertices in B . Let P_0 be the set of peripheral vertices. Let

$$i = \lceil (1 - \gamma)nd_2e^{-d} \rceil, \quad (51)$$

let \mathcal{G}_3 be the event

$$\mathcal{G}_3 = \{|P_0| \geq i\},$$

and set

$$\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3.$$

When \mathcal{G} holds, let P consist of the first i vertices in P_0 in some arbitrary order; otherwise, set $P = \emptyset$. We call an edge e of H_2 *peripheral* if it consists of one vertex in P and $r-1$ vertices in B . Given the pair (H_1, H_2) , let E_{int} be the set of internal blue edges, and let E_{per} be the set of peripheral blue edges.

Define the *reduced (blue) hypergraph* to be

$$H_2^- = H_2 - E_{\text{int}} - E_{\text{per}},$$

and let $H^- = H_1 \cup H_2^-$; see Figure 1. Note that any $v \in P$ is isolated in H_2^- . We shall condition on the pair (H_1, H_2^-) .

Remark 5.1. A key point is that we can determine B , E , P_0 and P , and hence whether \mathcal{G} holds, knowing only the reduced graph (H_1, H_2^-) , without knowing the original value of H_2 . For B and E this is immediate; they are defined in terms of the red graph H_1 . With H_1 and hence B fixed, for any possible value H_2' of H_2 , let us temporarily write $P(H_1, H_2')$ for the set of peripheral vertices in (H_1, H_2') . Thus $v \in P(H_1, H_2')$ if and only if v is isolated in H_1 , is in at most one edge of H_2' of the form $\{v, b_1, \dots, b_{r-1}\}$ with $b_i \in B$, and is in no other edges of H_2' . The key observation is that deleting internal or peripheral blue edges does not change $P(H_1, \cdot)$. Thus $P_0 = P(H_1, H_2) = P(H_1, H_2^-)$ is a function of H_1 and H_2^- as claimed. Since P is defined in terms of P_0 only, P is also a function of H_1 and H_2^- .

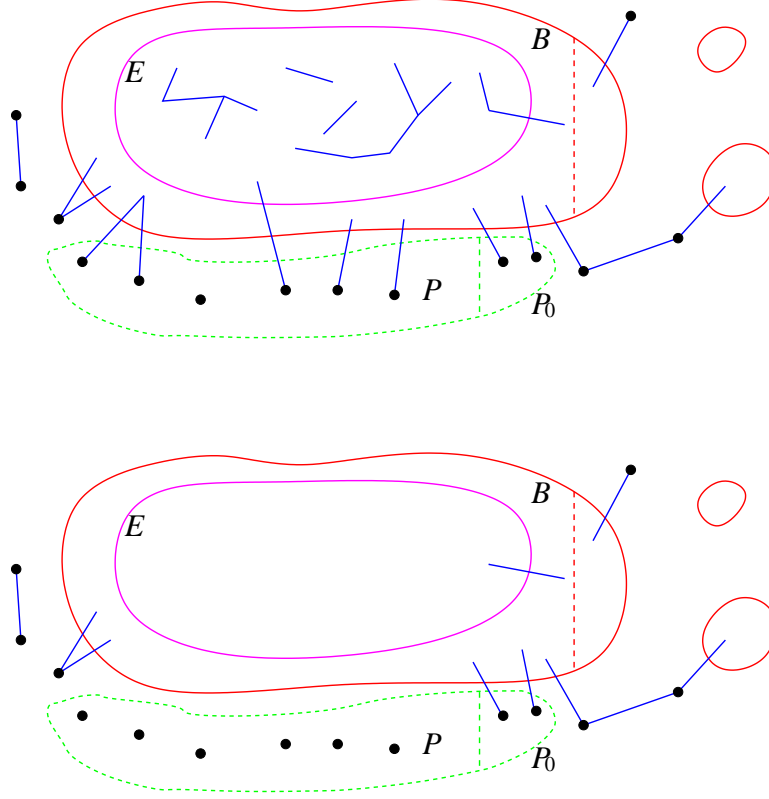


Figure 1: An example of the coloured graph (H_1, H_2) (top) and the corresponding reduced graph (H_1, H_2^-) (bottom), assuming \mathcal{G} holds. The red blobs indicate components of H_1 ; vertices isolated in H_1 are shown as black dots. One component of H_1 contains almost all vertices. B is a large subset of this component with $|B| = b$. E indicates a set of potential edges, all inside B . Edges of H_2 are blue; there are many more than are shown. P_0 consists of those vertices v that are isolated in H_1 and in at most one blue edge, with that edge consisting of v and $r - 1$ vertices in B . P is a subset of P_0 of a fixed size i . In reducing (H_1, H_2) we delete all blue edges in E , and all blue edges between P and B .

Lemma 5.2. *If n is large enough, then whenever \mathcal{G} holds we have*

$$L_1 = L + |E_{\text{per}}| \quad \text{and} \quad M_1 = M + |E_{\text{per}}| + |E_{\text{int}}|$$

for some quantities L, M that are functions of (H_1, H_2^-) .

Proof. Recall that $H = H_1 \cup H_2 = H_1 \cup H_2^- \cup E_{\text{per}} \cup E_{\text{int}}$. When \mathcal{G} holds, then B is a set of b vertices all within the same component of H_1 . Let C^- be the

component of $H^- = H_1 \cup H_2^-$ containing this component. Then $|C^-| \geq b > n/2$ if n is large enough, so C^- is certainly a subset of the largest component C of $H = H_1 \cup H_2$. Furthermore, each edge of E_{int} lies entirely within C^- , and each edge in E_{per} connects a distinct isolated vertex of H^- to C^- , so we have $|C| = |C^-| + |E_{\text{per}}|$ and $e(C) = e(C^-) + |E_{\text{per}}| + |E_{\text{int}}|$. \square

Since $b = n - s_{1,1} = n(1 - O(e^{-d_1}))$, we have

$$p_2 \binom{b}{r-1} = p_2 \frac{n^{r-1}}{(r-1)!} (1 + O(s_{1,1}/n)) = d_2(1 + O(e^{-d_1})). \quad (52)$$

This estimate will be useful in the proofs of the next two lemmas.

Recall from Remark 5.1 that knowing the reduced graph (H_1, H_2^-) determines B , E , P_0 and P , and hence also whether or not \mathcal{G} holds.

Lemma 5.3. *Whenever \mathcal{G} holds, the conditional distributions of $|E_{\text{int}}|$ and $|E_{\text{per}}|$ given (H_1, H_2^-) are independent, with*

$$|E_{\text{per}}| \sim \text{Bin}(i, \pi) \quad \text{and} \quad |E_{\text{int}}| \sim \text{Bin}(S, p_2),$$

where

$$\pi = \frac{p_2 \binom{b}{r-1}}{p_2 \binom{b}{r-1} + 1 - p_2} \sim \frac{d_2}{1 + d_2}. \quad (53)$$

Proof. Given (H_1, H_2^-) it is easy to identify the possible values of the pair $(E_{\text{per}}, E_{\text{int}})$. Indeed, as noted above the pair (H_1, H_2^-) determines B , E and hence P_0 and P . When \mathcal{G} holds, then $|B| = b$, $|E| = S$, $|P| = i$, and P and B are disjoint. Now E_{int} must be a subset of E , and any subset is possible. A peripheral blue edge must consist of a vertex of P and $r-1$ vertices of B , and any set of such edges including each vertex of P at most once is a possibility for E_{per} . In other words, to obtain a possible value of H_2 , given (H_1, H_2^-) , starting from H_2^- we must

- (a) for each edge $e \in E$, either add it or not, and
- (b) for each vertex $v \in P$, either add one of the $\binom{b}{r-1}$ edges consisting of v and $r-1$ vertices from B or not.

Since all combinations are possible, and the probability of a possible value H of H_2 is proportional to $p_2/(1-p_2)$ to the power of the number of edges, we see that in describing the conditional distribution of H_2 given (H_1, H_2^-) , all these choices are independent. Furthermore, in (a) each edge is included with probability p_2 , and in (b) the probability of including an edge is π as defined above. This implies the claimed formulae, with the asymptotic estimate for π following from (52). \square

Our ‘smoothing random variables’ will be essentially $|E_{\text{per}}|$ and $|E_{\text{int}}|$. For later, it will be convenient to ‘cook’ these random variables when \mathcal{G} does not hold, so that they *always* have the conditional distribution defined above. More precisely, let

$$X' \sim \text{Bin}(i, \pi) \quad \text{and} \quad Y' \sim \text{Bin}(S, p_2)$$

be independent of each other and of (H_1, H_2^-) . Set

$$X = \mathbb{1}_{\mathcal{G}}|E_{\text{per}}| + \mathbb{1}_{\mathcal{G}^c}X' \quad \text{and} \quad Y = \mathbb{1}_{\mathcal{G}}|E_{\text{int}}| + \mathbb{1}_{\mathcal{G}^c}Y', \quad (54)$$

where $\mathbb{1}_A$ denotes the indicator function of an event A . The only properties of (X, Y) we shall need are the following.

Lemma 5.4. *The random variables (H_1, H_2^-) and (X, Y) are independent, with (X, Y) having the distribution of a pair of independent binomial random variables $\text{Bin}(i, \pi)$ and $\text{Bin}(S, p_2)$. Moreover, if n is large enough, then whenever \mathcal{G} holds we have*

$$L_1 = L + X \quad \text{and} \quad M_1 = M + X + Y$$

for some quantities L, M that are functions of (H_1, H_2^-) .

Proof. To prove the first statement we must show exactly that the conditional distribution of (X, Y) given (H_1, H_2^-) is always that of the given independent binomial distributions. This follows immediately from the definition (54) of X and Y and Lemma 5.3 (applied only when \mathcal{G} holds). The second statement follows immediately from (54) and Lemma 5.2. \square

We have already shown that the events \mathcal{G}_1 and \mathcal{G}_2 are very likely to hold. We now show the same for the event $\mathcal{G}_3 = \{P_0 \geq i\}$ that there are at least $i = \lceil (1 - \gamma)nd_2e^{-d} \rceil$ peripheral vertices.

Lemma 5.5. *We have $\mathbb{P}(\mathcal{G}_3) = 1 - o(n^{-100})$.*

Proof. The probability that a given vertex is isolated in H_1 is $(1 - p_1)^{\binom{n-1}{r-1}}$. Recalling that $p_1 = O(n^{-r+1} \log n)$, since $\binom{n-1}{r-1} = n^{r-1}/(r-1)! + O(n^{r-2})$ and $p_1^2 n^{r-1} = O(n^{-r+1}(\log n)^2) = O((\log n)^2/n)$, we have (crudely)

$$\begin{aligned} (1 - p_1)^{\binom{n-1}{r-1}} &= \exp(-p_1 \binom{n-1}{r-1} + O((\log n)^2/n)) \\ &= \exp(-p_1 \frac{n^{r-1}}{(r-1)!} + O((\log n)^2/n)) = e^{-d_1 + O((\log n)^2/n)} = e^{-d_1} (1 + o(\gamma)), \end{aligned}$$

since $\gamma n/(\log n)^2 \geq \gamma d \rightarrow \infty$, from (44). Let I_1 be the set of isolated vertices of H_1 . Then $\mathbb{E}[|I_1|] = ne^{-d_1}(1 + o(\gamma))$. Since $\mathbb{E}[e(H_1)] = nd_1/r \leq n \log n$, say, it is not hard to see that

$$\mathbb{P}(|I_1| \leq \mathbb{E}[|I_1|] - n^{0.51}) = o(n^{-100}),$$

say. For example, one approach is to consider a variant of the model $H_1 = H_{n, p_1}^r$ in which we add uniformly random edges one-by-one, and apply the Hoeffding–Azuma inequality in this model; we omit the details. Recalling (47), we see that with probability $1 - o(n^{-100})$ we have

$$|I_1| \geq (1 - n^{-1/3})\mathbb{E}[|I_1|] \geq ne^{-d_1}(1 - o(\gamma)), \quad (55)$$

say.

For the rest of the proof we condition on H_1 , which determines I_1 and B ; we assume, as we may, that the event \mathcal{G}_1 and inequality (55) hold. Since \mathcal{G}_1 holds, the sets I_1 and B are disjoint. Call a possible (blue) edge *acceptable* if it consists of one vertex of I_1 and $r-1$ vertices of B . We call any other possible edge meeting I_1 *annoying*. It remains to show that with very high probability there are at least i vertices $v \in I_1$ such that v is in no annoying blue edges and v is in at most one acceptable blue edge. Indeed, when \mathcal{G}_1 holds, then P_0 consists precisely of the set of such $v \in I_1$.

We first consider annoying edges. There are at most

$$N = |I_1|(n - |B|)n^{r-2} = |I_1|s_{1,1}n^{r-2}$$

possible edges that are annoying. Each is present in H_2 independently with probability p_2 . By (45), $d_2e^{-d_1} \leq de^{-d_1} \rightarrow 0$, so if n is large enough we have

$$Np_2 = |I_1|100ne^{-d_1}n^{r-2}d_2\frac{(r-1)!}{n^{r-1}} = 100(r-1)!d_2e^{-d_1}|I_1| \leq |I_1|/(2d).$$

Recalling (47) and (55), we have $|I_1|/d \geq ne^{-d_1}/(2d) = n^{1-o(1)}$. It follows by a Chernoff bound that with probability at least $1 - \exp(-\Omega(|I_1|/d)) = o(n^{-100})$ the actual number of annoying blue edges is at most $|I_1|/d$.

We condition on the set of annoying edges present in H_2 , assuming there are at most $|I_1|/d$ of them.

Let $I'_1 \subseteq I_1$ be the set of vertices of I_1 not in any annoying blue edges, so

$$|I'_1| \geq |I_1| - r|I_1|/d = |I_1|(1 - r/d) \geq ne^{-d_1}(1 - o(\gamma)). \quad (56)$$

A vertex $v \in I'_1$ is in P_0 if and only if it is in at most one acceptable blue edge. Noting that we have not yet tested the acceptable edges for their presence in H_2 , for a given vertex $v \in I'_1$, this event has probability

$$\pi = (1 - p_2)^{\binom{b}{r-1}} + \binom{b}{r-1}p_2(1 - p_2)^{\binom{b}{r-1}-1} \geq \binom{b}{r-1}p_2(1 - p_2)^{\binom{b}{r-1}}.$$

Recall from (52) that $p_2\binom{b}{r-1} = d_2(1 + O(e^{-d_1}))$. Since $p_2^2n^{r-1} = O(n^{-r+1}(\log n)^2)$, it easily follows that

$$\pi \geq d_2e^{-d_2}(1 - o(\gamma)). \quad (57)$$

Since the sets of potential acceptable edges meeting different vertices in I'_1 are disjoint, the events $v \in P_0$ are (conditionally) independent for different $v \in I'_1$, so the conditional distribution of $|P_0|$ is binomial $\text{Bin}(|I'_1|, \pi)$. Now by (56), (57) and (46) we have

$$|I'_1|\pi \geq ne^{-d_1}d_2e^{-d_2}(1 - o(\gamma)) \geq nd_2e^{-d}(1 - o(\gamma)).$$

Since $\gamma^2nd_2e^{-d} \sim \gamma^2nde^{-d}$ is much larger than $\log n$ by (44), a Chernoff bound shows that with (conditional) probability at least $1 - \exp(-\Omega(\gamma^2|I'_1|\pi)) = 1 - o(n^{-100})$ we have

$$|P_0| \geq (1 - \gamma/2)|I'_1|\pi \geq (1 - \gamma)nd_2e^{-d},$$

for n large enough. Thus $\mathbb{P}(\mathcal{G}^3) \geq 1 - o(n^{-100})$, as claimed. \square

Corollary 5.6. *We have $\mathbb{P}(\mathcal{G}) = 1 - o(n^{-100})$, where $\mathcal{G} = \mathcal{G}_1 \cap \mathcal{G}_2 \cap \mathcal{G}_3$.*

Proof. Immediate from (48), (49) and Lemma 5.5. \square

Before turning to the proof of Theorem 1.5 we give a simple observation about local limit theorems. Recall that a sequence $((X_n, Y_n))$ of \mathbb{Z}^2 -valued random variables satisfies a local limit theorem (an LLT) with parameters $(\mu_X(n), \mu_Y(n))$ and $(\sigma_X^2(n), \sigma_Y^2(n))$ if, suppressing the dependence on n , we have

$$\sup_{(x,y) \in \mathbb{Z}^2} |\mathbb{P}(X_n = x, Y_n = y) - f_n(x - \mu_X, y - \mu_Y)| = o(1/(\sigma_X \sigma_Y)),$$

where

$$f_n(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y} \exp\left(-\frac{x^2}{2\sigma_X^2} - \frac{y^2}{2\sigma_Y^2}\right). \quad (58)$$

We make some simple observations about such results.

Lemma 5.7. *If (X_n, Y_n) satisfies an LLT with parameters $(\mu_X(n), \mu_Y(n))$ and $(\sigma_X^2(n), \sigma_Y^2(n))$, then it also satisfies an LLT with parameters $(\tilde{\mu}_X(n), \tilde{\mu}_Y(n))$ and $(\tilde{\sigma}_X^2(n), \tilde{\sigma}_Y^2(n))$ whenever $\tilde{\mu}_X(n) = \mu_X(n) + o(\sigma_X(n))$, $\tilde{\mu}_Y(n) = \mu_Y(n) + o(\sigma_Y(n))$, $\tilde{\sigma}_X(n) \sim \sigma_X(n)$ and $\tilde{\sigma}_Y(n) \sim \sigma_Y(n)$.*

Proof. This is a standard result; we omit the proof which is just calculation. \square

Lemma 5.8. *Suppose that (X_n, Y_n) satisfies an LLT with parameters $(\mu_X(n), \mu_Y(n))$ and $(\sigma_X^2(n), \sigma_Y^2(n))$. Suppose also that (A_n, B_n) is independent of (X_n, Y_n) , and that*

$$\text{Var}[A_n] = o(\sigma_X^2) \quad \text{and} \quad \text{Var}[B_n] = o(\sigma_Y^2).$$

Then $(A_n + X_n, B_n + Y_n)$ satisfies an LLT with parameters $(\mathbb{E}[A_n] + \mu_X(n), \mathbb{E}[B_n] + \mu_Y(n))$ and $(\sigma_X^2(n), \sigma_Y^2(n))$.

Proof. This is a simple consequence of the fact that A_n and B_n are concentrated on the relevant scales, namely $\mu_X = \mu_X(n)$ and $\mu_Y = \mu_Y(n)$, together with the fact that $\exp(-x^2/2)$ is Lipschitz as a function of x (with constant $e^{-1/2}$). Indeed, we may write

$$\begin{aligned} \mathbb{P}(A_n + X_n = x, B_n + Y_n = y) &= \mathbb{E}_{A_n, B_n} [\mathbb{P}(X_n = x - A_n, Y_n = y - B_n)] \\ &= \mathbb{E}_{A_n, B_n} [f_n(x - A_n - \mu_X, y - B_n - \mu_Y)] + o(1/(\sigma_X \sigma_Y)), \end{aligned}$$

where f_n is defined as in (58) and in the second step we applied the LLT for (X_n, Y_n) . Since $\mathbb{E}[|A_n - \mathbb{E}[A_n]|] \leq \sqrt{\text{Var} A_n} = o(\sigma_X)$ and similarly $\mathbb{E}[|B_n - \mathbb{E}[B_n]|] = o(\sigma_Y)$, the result follows from the Lipschitz property of $e^{-x^2/2}$. \square

As in Theorem 1.5, set

$$\sigma_L^2 = \sigma_L^2(n) = ne^{-d} \quad \text{and} \quad \sigma_M^2 = \sigma_M^2(n) = \frac{dn}{r}.$$

Most of the time we shall suppress the dependence on n in the notation. The final ingredient in our proof is that the ‘cooked’ versions X and Y of the numbers E_{per} and E_{int} of peripheral and internal blue edges satisfy an LLT.

Lemma 5.9. *The random variables $X = X_n$ and $Y = Y_n$ defined in (54) satisfy a local limit theorem with parameters $(\mathbb{E}[X], \mathbb{E}[Y])$ and (σ_L^2, σ_M^2) .*

Proof. By Lemma 5.4, X and Y are independent of each other, and have binomial distributions $\text{Bin}(i, \pi)$ and $\text{Bin}(S, p_2)$. Recalling (51) and (53) we have

$$\text{Var}[X] = i\pi(1 - \pi) \sim i(1 - \pi) \sim i/d_2 \sim ne^{-d} = \sigma_L^2. \quad (59)$$

Also, from (50),

$$\text{Var}[Y] = Sp_2(1 - p_2) \sim Sp_2 \sim \frac{n^r}{r!} d_2 \frac{(r-1)!}{n^{r-1}} = \frac{d_2 n}{r} \sim \frac{dn}{r} = \sigma_M^2. \quad (60)$$

It is well known (and easy to verify, for example from the formula for a binomial coefficient) that a sequence of binomial random variables $\text{Bin}(N_n, p_n)$ with $\sigma_n^2 = N_n p_n(1 - p_n) \rightarrow \infty$ satisfies a univariate local limit theorem with parameters $N_n p_n$ and σ_n^2 . This statement extends in an obvious way to a pair of independent binomial random variables; this extension, and Lemma 5.7, give the result. \square

We now have all the pieces in place to prove our local limit theorem.

Proof of Theorem 1.5. We have already proved the estimates (17) and (18) for the mean and variance of $L_1 = L_1(H_{n,p}^r)$ and $M_1 = M_1(H_{n,p}^r)$ in Lemmas 4.1, 4.2 and 4.3. It remains to prove the local limit theorem. Set

$$L_1^* = L + X \quad \text{and} \quad M_1^* = M + X + Y,$$

where X and Y are defined in (54) and L and M are as in Lemma 5.4. By that lemma we have $L_1 = L_1^*$ and $M_1 = M_1^*$ whenever \mathcal{G} holds. By Corollary 5.6 we have $\mathbb{P}(\mathcal{G}^c) = o(n^{-100})$. It thus suffices to show that (L_1^*, M_1^*) satisfies a bivariate local limit theorem with parameters $(\mathbb{E}[L_1^*], \mathbb{E}[M_1^*])$ for the means and (σ_L^2, σ_M^2) for the variances. Furthermore, since $\sigma_L^2 = o(\sigma_M^2)$, setting $\tilde{M}_1 = M_1^* - L_1^*$, this is equivalent to showing that (L_1^*, \tilde{M}_1) satisfies a local limit theorem with parameters $(\mathbb{E}[L_1^*], \mathbb{E}[\tilde{M}_1])$ and (σ_L^2, σ_M^2) .

By Lemma 5.4, (X, Y) is independent of (H_1, H_2^-) . Since L and M are determined by H_1 and H_2^- , we see that (X, Y) is independent of (L, M) , and hence of $(L, M - L)$. Now

$$(L_1^*, \tilde{M}_1) = (L, M - L) + (X, Y), \quad (61)$$

with the summands independent. Since $L_1^* = L_1$ with probability $1 - o(n^{-100})$, we have

$$\text{Var}[L_1^*] \sim \text{Var}[L_1] \sim ne^{-d} = \sigma_L^2,$$

by Lemma 4.1. Similarly, $\text{Var}[M_1^*] \sim \text{Var}[M_1] \sim dn/r = \sigma_M^2$ by Lemma 4.3. Since $\sigma_L = o(\sigma_M)$ it follows that

$$\text{Var}[\tilde{M}_1] \sim \sigma_M^2.$$

From (59) and (60) we have

$$\text{Var}[X] \sim \sigma_L^2 \quad \text{and} \quad \text{Var}[Y] \sim \sigma_M^2.$$

From the independence in (61), we have $\text{Var}[L_1^*] = \text{Var}[L] + \text{Var}[X]$, which implies $\text{Var}[L] = o(\sigma_L^2)$. Similarly, $\text{Var}[M - L] = o(\sigma_M^2)$. The result now follows from Lemma 5.9, the independence in (61), and Lemma 5.8. \square

6 Proof of Theorem 1.1

Our aim in this section is to deduce our enumerative result, Theorem 1.1, from the probabilistic one, Theorem 1.5. We start with a lemma giving the asymptotic behaviour of the quantity $\xi = \xi(\bar{d})$ appearing in Theorem 1.1.

Recall that for $r \geq 3$ we define a function Φ_r on $(0, 1)$ by

$$\Phi_r(\xi) = \frac{\log(1/\xi)(1 - \xi^r)}{(1 - \xi^{r-1})(1 - \xi)},$$

and that Φ_r is a decreasing bijection between $(0, 1)$ and $(r/(r-1), \infty)$. If $r \geq 3$ then

$$\Phi_r(\xi) = \log(1/\xi)(1 + \xi + O(\xi^2)) \sim \log(1/\xi) \quad \text{as} \quad \xi \rightarrow 0. \quad (62)$$

Recall also that $F_r(\bar{d})$ is then defined by (3) with $\xi = \Phi_r^{-1}(\bar{d})$.

Lemma 6.1. *Fix $r \geq 3$. For $\bar{d} > r/(r-1)$ let $\xi(\bar{d}) = \Phi_r^{-1}(\bar{d})$. Then as $\bar{d} \rightarrow \infty$ we have*

$$\xi = e^{-\bar{d}} + \bar{d}e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}}) \quad (63)$$

and

$$F_r(\bar{d}) = e^{-\bar{d}} + \frac{\bar{d}+1}{2}e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}}). \quad (64)$$

For $r = 2$, we have $\xi = e^{-\bar{d}} + 2\bar{d}e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}})$ and $F_2(\bar{d}) = e^{-\bar{d}} + (\bar{d} + 1/2)e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}})$.

Proof. Since Φ_r is a decreasing bijection from $(0, 1)$ to $(r/(r-1), \infty)$, as $\bar{d} \rightarrow \infty$ we have $\xi = \Phi_r^{-1}(\bar{d}) \rightarrow 0$. Hence, for $r \geq 3$, from (62) we have

$$\log \xi = -\bar{d}(1 - \xi + O(\xi^2)). \quad (65)$$

It follows (multiplying by ξ) that $\bar{d}\xi \rightarrow 0$. Also, from (65),

$$\xi = e^{-\bar{d}}e^{\bar{d}\xi + O(\bar{d}\xi^2)} = e^{-\bar{d}}(1 + \bar{d}\xi + O(\bar{d}^2\xi^2)) = e^{-\bar{d}} + \bar{d}e^{-2\bar{d}} + O(\bar{d}^2e^{-3\bar{d}}).$$

From (3), for $r \geq 3$ we have

$$F_r(\bar{d}) = -\bar{d}(\xi + \xi^2/2) - (\xi + \xi^2) \log \xi + \xi + \xi^2/2 + O(\xi^3\bar{d} + \xi^3|\log \xi| + \xi^3).$$

Since $|\log \xi| \sim \bar{d}$ as $\bar{d} \rightarrow \infty$, we may write the error term as $O(\bar{d}\xi^3)$. Substituting in (65), it follows that

$$\begin{aligned} F_r(\bar{d}) &= -\bar{d}\xi - \bar{d}\xi^2/2 + \bar{d}\xi + \bar{d}\xi^2 - \bar{d}\xi^2 + \xi + \xi^2/2 + O(\bar{d}^2\xi^3) \\ &= \xi - (\bar{d} - 1)\xi^2/2 + O(\bar{d}^2\xi^3). \end{aligned}$$

Substituting in (63) gives (64). We omit the (similar) calculations for $r = 2$. \square

Recall that we write $C_r(s, m)$ for the number of connected r -uniform hypergraphs on $[s]$, and $P_r(s, m)$ for the probability that an m -edge r -uniform hypergraph on $[s]$ chosen uniformly at random is connected. Clearly, $P_r(s, m) = C_r(s, m)/\binom{N}{m}$, where $N = \binom{s}{r}$. Thus the asymptotic formulae (4) and (5) for $P_r(s, m)$ and $C_r(s, m)$ are equivalent modulo a calculation, which we now carry out.

Lemma 6.2. *Let $r \geq 2$ and let $m = m(s) = o(s^{4/3})$. If $r \geq 3$ then*

$$\binom{\binom{s}{r}}{m} \sim \frac{s^{rm}}{m!r!^m} e^{-(r-1)\bar{d}/2}, \quad (66)$$

as $s \rightarrow \infty$, where $\bar{d} = rm/s$. If $r = 2$, then

$$\binom{\binom{s}{r}}{m} \sim \frac{s^{rm}}{m!r!^m} e^{-(r-1)\bar{d}/2 - \bar{d}^2/4}. \quad (67)$$

Proof. Let

$$N = \binom{s}{r} = \frac{s(s-1) \cdots (s-r+1)}{r!} = \frac{s^r}{r!} e^{-\binom{r}{2}/s + O(s^{-2})}.$$

Since $m = o(s^{4/3}) = o(s^2)$, we have

$$N^m \sim \frac{s^{rm}}{r!^m} e^{-\binom{r}{2}m/s} = \frac{s^{rm}}{r!^m} e^{-(r-1)\bar{d}/2}.$$

Since $N = \Theta(s^r)$, if $r \geq 3$ then $m^2 = o(N)$. Thus $\binom{N}{m} \sim N^m/m!$, giving (66).

For (67), suppose that $r = 2$. Then $N = s^2/2(1 + O(1/s))$ and $m = \bar{d}s/2$. Since m/N and m^3/N^2 are $o(1)$, and $\bar{d}^2/s = O(m^2/s^3) = o(1)$, we have

$$\begin{aligned} \binom{N}{m} &= \frac{N^m}{m!} \exp(-m^2/(2N) + O(m/N + m^3/N^2)) \\ &\sim \frac{N^m}{m!} \exp\left(-\frac{\bar{d}^2 s^2}{4s^2}(1 + O(1/s))\right) \sim \frac{N^m}{m!} e^{-\bar{d}^2/4}, \end{aligned}$$

and (67) follows. \square

We are finally ready to deduce Theorem 1.1 from Theorem 1.5.

Proof of Theorem 1.1. Throughout, we consider a function $m = m(s)$ with $m/s \rightarrow \infty$; all asymptotics are as $s \rightarrow \infty$. Much of the time we suppress the dependence on s in the notation. We write

$$\bar{d} = \bar{d}(s) = \frac{rm}{s}$$

for the average degree of an r -uniform hypergraph with s vertices and m edges.

Let us first deal with a simple case: when $\log s - \bar{d}$ is bounded above. Passing to a subsequence, we may assume that either (i) $\log s - \bar{d} \rightarrow c$ for some constant

$c \in \mathbb{R}$, or (ii) $\log s - \bar{d} \rightarrow -\infty$. Let $H_{s,m}^r$ be a hypergraph on $[s]$ with m edges, chosen uniformly at random from all such hypergraphs. Let $X = X_s$ denote the number of isolated vertices in $H_{s,m}^r$. In case (i), a simple calculation shows that

$$\mathbb{E}[X] = s \binom{s-1}{r} \binom{s}{m}^{-1} \rightarrow e^c.$$

(This is to be expected, since the probability that a vertex is isolated is asymptotically $e^{-\bar{d}}$.) Similarly, for any fixed k the k th factorial moment satisfies

$$\mathbb{E}[X(X-1)\cdots(X-k+1)] = s(s-1)\cdots(s-k+1) \binom{s-k}{r} \binom{s}{m}^{-1} \rightarrow e^{kc}.$$

Now by a standard result (see, e.g., Theorem 1.22 in [7]) it follows that $X = X_s$ converges in distribution to a Poisson distribution with mean e^c as $s \rightarrow \infty$, and in particular, that $\mathbb{P}(X = 0) \rightarrow \exp(e^{-c})$. A very simple argument counting cuts shows that in this range, with high probability $H_{s,m}^r$ has no component of size between 2 and $s/2$, so the probability $P_r(s, m)$ that $H_{s,m}^r$ is connected satisfies

$$P_r(s, m) = \mathbb{P}(X = 0) + o(1) \sim \exp(-e^c) \sim \exp(-se^{-\bar{d}}).$$

Since $\bar{d} = \log s + O(1)$, we have $s\bar{d}e^{-2\bar{d}} \rightarrow 0$, so by Lemma 6.1 $sF_r(\bar{d}) = se^{-\bar{d}} + o(1)$. Thus we have $P_r(s, m) \sim \exp(-sF_r(\bar{d}))$, proving (4) in this case. For case (ii), it follows from the above and monotonicity that $P_r(s, m) \rightarrow 1$, which again agrees with (4). In both cases (i) and (ii), provided $m = o(s^{4/3})$ (which (5) assumes), relation (5) follows immediately from (4) and Lemma 6.2.

In proving Theorem 1.1, passing to a subsequence, we may assume either that $\log s - \bar{d}$ is bounded above, or that $\log s - \bar{d} \rightarrow \infty$. We have covered the first case above. From now on we thus assume that

$$\bar{d} \rightarrow \infty \quad \text{and} \quad \log s - \bar{d} \rightarrow \infty \tag{68}$$

as $s \rightarrow \infty$. Then $m = \bar{d}s/r = O(s \log s) = o(s^{4/3})$, so by Lemma 6.2 either of (4) and (5) implies the other. We shall prove (5).

When s is large enough, we have $\bar{d} > r/(r-1)$; we assume this from now on. Then there is a unique solution $\xi = \xi(s)$ to (2), i.e., to

$$\Phi_r(\xi) = \bar{d} = \frac{rm}{s}.$$

Since $\bar{d} \rightarrow \infty$ as $s \rightarrow \infty$ we have $\xi \rightarrow 0$. Set

$$d = d(s) = \frac{\log(1/\xi)}{1 - \xi^{r-1}}, \tag{69}$$

noting that $d(s) \rightarrow \infty$. Then ξ and d solve the equation (15) (which is just (69) rearranged). Also, set

$$\tilde{n} = \tilde{n}(s) = \frac{s}{1 - \xi}.$$

The reason for these choices is that then we have

$$\tilde{n}(1 - \xi) = s \quad (70)$$

and

$$\frac{d(1 - \xi^r)}{r} \tilde{n} = \frac{(1 - \xi)\Phi_r(\xi)}{r} \tilde{n} = \frac{\Phi_r(\xi)}{r} s = m. \quad (71)$$

We would like to apply Theorem 1.5 with the parameters \tilde{n} and d just defined; one trivial but annoying difficulty is that \tilde{n} is not an integer. So set

$$n = n(s) = \lceil \tilde{n} \rceil.$$

Since $n = \tilde{n} + O(1)$, from (70), (71) and the fact that $0 < \xi < 1$ we see that

$$n(1 - \xi) = s + O(1) \quad \text{and} \quad \frac{d(1 - \xi^r)}{r} n = m + O(d). \quad (72)$$

We next verify that n and d satisfy the assumptions of Theorem 1.5, i.e., that $n \rightarrow \infty$, and d and $\log n - d \rightarrow \infty$. (The theorem assumes $d = d(n)$ is defined for every n , but there is no problem considering only a subsequence.) Certainly, $n \geq \tilde{n} \geq s \rightarrow \infty$ as $s \rightarrow \infty$. We have already noted that $d = d(s) \rightarrow \infty$. For the last condition, by (69) and Lemma 6.1 we have

$$d = \log(1/\xi)(1 + O(\xi^{r-1})) = \bar{d} + O(\bar{d}e^{-\bar{d}}).$$

In particular,

$$d = \bar{d} + o(1). \quad (73)$$

Since $n = \tilde{n} + O(1) = s/(1 - \xi) + O(1) \sim s$ we thus have

$$\log n - d = \log s - \bar{d} + o(1) \rightarrow \infty. \quad (74)$$

Thus all conditions of Theorem 1.5 are satisfied.

As in the statement of Theorem 1.5, set

$$p = d \frac{(r-1)!}{n^{r-1}}. \quad (75)$$

In this section, d and n are functions of s , so this defines a function $p(s)$; as usual, we suppress the dependence on s . Let $\sigma_L^2 = ne^{-d}$ and $\sigma_M^2 = nd/r$, as in (19). Under the assumptions of Theorem 1.5 (which we have just verified) we have $\sigma_L \rightarrow \infty$ and $d = o(\sigma_M)$. Hence, by (72) and (17), the values s and m are within $o(1)$ standard deviations of the expectations of the numbers L_1 and M_1 of vertices and edges in the largest component of the (binomial) random hypergraph $H_{n,p}^r$. Thus, by Theorem 1.5,

$$\mathbb{P}(L_1 = s, M_1 = m) \sim \frac{1}{2\pi\sigma_L\sigma_M} \sim \frac{\sqrt{r}e^{d/2}}{2\pi n\sqrt{d}}. \quad (76)$$

Recalling that $n \sim s$, for s large enough we have $n < 2s$, so the hypergraph $H_{n,p}^r$ can have at most one component with s or more vertices. Writing $Z =$

$Z(H_{n,p}^r)$ for the number of components with s vertices and m edges, we thus have

$$\mathbb{P}(L_1 = s, M_1 = m) = \mathbb{P}(Z = 1) = \mathbb{E}[Z].$$

Hence, by linearity of expectation,

$$\mathbb{P}(L_1 = s, M_1 = m) = \binom{n}{s} C_r(s, m) p^m (1-p)^{M-m}, \quad (77)$$

where

$$M = \binom{n}{r} - \binom{n-s}{r}$$

is the number of possible hyperedges meeting a given set of s vertices.

From (76) and (77) we see that

$$C_r(s, m) \sim \frac{\sqrt{r} e^{d/2}}{2\pi n \sqrt{d}} \binom{n}{s}^{-1} (p^m (1-p)^{M-m})^{-1}. \quad (78)$$

The rest of the proof is ‘just’ calculation, but this calculation is not so simple. A significant hindrance is that we would eventually like to work in terms of s , m , $\bar{d} = rm/s$ and the implicitly defined $\xi = \Phi_r^{-1}(\bar{d})$. The quantity $\tilde{n} = s/(1-\xi)$ is a simple function of these variables, but $n = \lceil \tilde{n} \rceil$ is not. Morally speaking, rounding to n should make no difference, but showing this seems to require some work: because of the large exponents appearing in (78), the very small relative change of replacing n by \tilde{n} in the various factors in (78) can change these factors by a large amount even though, as we shall see, it does not significantly change (a suitably adapted form of) the whole formula. The last factor in (78) is perhaps the hardest to deal with; fortunately, we can use a trick, relating it to a binomial probability. The key point (established below) is that Mp is rather close to m .

Recall that, crudely,

$$n \sim \tilde{n} \sim s,$$

and, from Lemma 6.1 and (73), that

$$\xi \sim e^{-\bar{d}} \sim e^{-d}. \quad (79)$$

Thus,

$$\xi n \sim n e^{-\bar{d}} \sim s e^{-\bar{d}} \rightarrow \infty, \quad (80)$$

using (68) in the last step. Turning to $n - s$, recalling (70) and that $n = \lceil \tilde{n} \rceil$, we have the rather accurate bound

$$n - s = \tilde{n} - s + O(1) = \xi \tilde{n} + O(1) = \xi n + O(1). \quad (81)$$

We shall need this later, though often the simpler consequence

$$n - s \sim \xi n \sim n e^{-\bar{d}} \quad (82)$$

will suffice.

Since

$$r! \binom{x}{r} = x(x-1) \cdots (x-(r-1)) = x^r - \binom{r}{2} x^{r-1} + O(x^{r-2}),$$

we have

$$\begin{aligned} M &= \frac{n^r - \binom{r}{2} n^{r-1} - (n-s)^r}{r!} + O(n^{r-2} + (n-s)^{r-1}) \\ &= \frac{n^r}{r!} \left(1 - \binom{r}{2} \frac{1}{n} - \left(1 - \frac{s}{n} \right)^r \right) + O(n^{r-2} + n^{r-1} e^{-(r-1)\bar{d}}), \end{aligned}$$

using (82) in the last step. Since $r-1 \geq 1$ and $ne^{-\bar{d}} \rightarrow \infty$, we have $n^{r-2} + n^{r-1} e^{-(r-1)\bar{d}} = O(n^{r-1} e^{-\bar{d}})$. As the ‘cross term’ $(\binom{r}{2}/n)(1-s/n)^r$ is of order $O((n-s)^r n^{-r-1}) = O(n^{-1} e^{-\bar{d}})$, we thus have

$$M = \frac{n^r}{r!} \left(1 - \binom{r}{2} \frac{1}{n} \right) \left(1 - \left(1 - \frac{s}{n} \right)^r \right) \left(1 + O(n^{-1} e^{-\bar{d}}) \right). \quad (83)$$

We shall need this accurate estimate later; for the moment, something simpler suffices. From (81) we have

$$1 - \frac{s}{n} = \frac{n-s}{n} = \frac{\xi n + O(1)}{n} = \xi + O(1/n). \quad (84)$$

Hence (83) implies the cruder bound

$$M = \frac{n^r}{r!} (1 - \xi^r) (1 + O(n^{-1})).$$

Recalling the definition (75) of p , it follows that

$$Mp = \frac{dn}{r} (1 - \xi^r) + O(d) = m + O(d),$$

where the last step is from (72). Now $p = o(1)$, while certainly $m = \bar{d}s/r \rightarrow \infty$ and $d = o(\sqrt{m})$ (since $d \leq \log n$ and $m/n \sim m/s \rightarrow \infty$). It follows that the probability that a binomial random variable $\text{Bin}(M, p)$ takes the value m is asymptotically

$$\frac{1}{\sqrt{2\pi Mp(1-p)}} \sim \frac{1}{\sqrt{2\pi Mp}} \sim \frac{1}{\sqrt{2\pi m}}.$$

In other words,

$$\binom{M}{m} p^m (1-p)^{M-m} \sim (2\pi m)^{-1/2}.$$

Combining this with (78) we see that

$$C_r(s, m) \sim \frac{\sqrt{r} e^{d/2}}{2\pi n \sqrt{d}} \binom{n}{s}^{-1} \binom{M}{m} \sqrt{2\pi m}. \quad (85)$$

From (73) we have $d \sim \bar{d}$ and $e^{d/2} \sim e^{\bar{d}/2}$. Since $n \sim s$ and $rm/s = \bar{d}$, we may simplify (85) slightly to obtain

$$C_r(s, m) \sim \frac{e^{\bar{d}/2}}{\sqrt{2\pi s}} \binom{n}{s}^{-1} \binom{M}{m}. \quad (86)$$

This formula may appear appealingly concise, but unfortunately it still involves n , defined in a slightly unpleasant way (involving rounding), both directly and in the definition of M . So we continue with our manipulations.

Firstly, note for later that, from (84),

$$\frac{s}{n} = 1 - \xi + O(1/n). \quad (87)$$

As $s \rightarrow \infty$ we certainly have $n \rightarrow \infty$ and $n - s \rightarrow \infty$ (see (82) and (80)), so by Stirling's formula and the estimate $s \sim n$ we have

$$\binom{n}{s}^{-1} = \frac{(n-s)!s!}{n!} \sim \sqrt{2\pi(n-s)}(1-s/n)^{n-s}(s/n)^s.$$

Since $\frac{n-s}{s} \sim \frac{n-s}{n} \sim e^{-\bar{d}}$ by (82), it follows from (86) that

$$C_r(s, m) \sim (1-s/n)^{n-s}(s/n)^s \binom{M}{m}. \quad (88)$$

Now $M \sim n^r/r!$, while $m = \bar{d}s/r \sim \bar{d}n/r$. Since $\bar{d} = O(\log n)$, for $r \geq 3$ it follows that $m^2 = O(n^2 \log^2 n) = o(M)$. Hence

$$\binom{M}{m} \sim \frac{M^m}{m!}. \quad (89)$$

For $r = 2$ we have $m = \bar{d}s/2$ and $M = n^2/2(1 + O(1/n)) = s^2/2(1 + O(\xi)) = s^2/2(1 + O(e^{-\bar{d}}))$, using (87) and (79). Arguing as for (67), it follows that

$$\binom{M}{m} \sim \frac{M^m}{m!} e^{-\bar{d}^2/4}. \quad (90)$$

Since, in any case, $m = o(n^2)$, we have

$$\left(1 - \binom{r}{2} \frac{1}{n}\right)^m \sim e^{-\binom{r}{2} \frac{m}{n}} \sim e^{-\binom{r}{2} \frac{m}{s}} = e^{-(r-1)\bar{d}/2},$$

where in the second step we used that $s/n = 1 + O(e^{-\bar{d}}) = 1 + o(1/\bar{d})$ (from (87)) and $m/n = O(\bar{d})$. Also, since $m = O(\bar{d}n)$, we have

$$\left(1 + O(n^{-1}e^{-\bar{d}})\right)^m = \exp(O(\bar{d}e^{-\bar{d}})) \sim 1.$$

Combining these estimates with (89) and (83), for $r \geq 3$ we find that

$$\binom{M}{m} \sim \frac{n^{rm}}{m!r!^m} e^{-(r-1)\bar{d}/2} (1 - (1-s/n)^r)^m. \quad (91)$$

For $r = 2$ we obtain the same formula with an extra factor of $e^{-\bar{d}^2/4}$. We write the formulae in the rest of the proof for the case $r \geq 3$; the remaining estimates apply just as well when $r = 2$, with the factor $e^{-\bar{d}^2/4}$ inserted where appropriate. From (88) and (91), we see that

$$C_r(s, m) \sim e^{-(r-1)\bar{d}/2} \frac{n^{rm}}{m!r!^m} (1 - (1 - s/n)^r)^m (1 - s/n)^{n-s} (s/n)^s.$$

We may rewrite this as

$$C_r(s, m) \sim e^{-(r-1)\bar{d}/2} \frac{s^{rm}}{m!r!^m} (s/n)^{-rm} (1 - (1 - s/n)^r)^m ((1 - s/n)^{n/s-1} (s/n))^s,$$

and hence as

$$C_r(s, m) \sim e^{-(r-1)\bar{d}/2} \frac{s^{rm}}{m!r!^m} (x^{-r} (1 - (1 - x)^r))^m ((1 - x)^{1/x-1} x)^s, \quad (92)$$

where $x = s/n$. We would like to replace x by $s/\tilde{n} = 1 - \xi$. First, we substitute $y = 1 - x$, obtaining

$$C_r(s, m) \sim e^{-(r-1)\bar{d}/2} \frac{s^{rm}}{m!r!^m} g(y)^m h(y)^s,$$

where

$$g(y) = (1 - y)^{-r} (1 - y^r) \quad \text{and} \quad h(y) = y^{y/(1-y)} (1 - y).$$

Set

$$f(y) = g(y)^m h(y)^s.$$

It is straightforward to check that

$$(\log g(y))' = r \frac{1 - y^{r-1}}{(1 - y^r)(1 - y)} = r + O(y)$$

as $y \rightarrow 0$, and

$$(\log h(y))' = \frac{\log y}{(1 - y)^2} = \log y + O(y|\log y|) = \log y + o(1),$$

so

$$(\log f(y))' = rm + s \log y + O(my) + o(s).$$

We wish to compare $f(\xi)$ with $f(1 - s/n)$. From (84) we have $1 - s/n = \xi + O(1/n)$. Hence we need only consider values of y with $|y - \xi| = O(1/n)$. In this range,

$$\log y = \log \xi + O(1/(n\xi)) = \log \xi + o(1),$$

recalling that $n\xi \rightarrow \infty$. By Lemma 6.1, $\log \xi = -\bar{d} + o(1) = -rm/s + o(1)$. Hence

$$(\log f(y))' = O(my) + o(s) = o(s),$$

where in the final step we used that $my = O(\bar{d}s\xi) = O(\bar{d}se^{-\bar{d}}) = o(s)$. Since $1 - s/n = \xi + O(1/n) = \xi + O(1/s)$ it follows that

$$f(1 - s/n) \sim f(\xi).$$

This is exactly what we need to allow us to replace $x = s/n$ by $x = 1 - \xi$ in (92).

In conclusion, writing $\rho = 1 - \xi$, for $r \geq 3$ we have

$$\begin{aligned} C_r(s, m) &\sim e^{-(r-1)\bar{d}/2} \frac{s^{rm}}{m!r!^m} (\rho^{-r}(1 - (1 - \rho)^r))^m ((1 - \rho)^{1/\rho-1}\rho)^s \\ &= e^{-(r-1)\bar{d}/2} \frac{s^{rm}}{m!r!^m} \exp(-sF_r(\bar{d})), \end{aligned}$$

where the last step is from (7).

When $r = 2$ we obtain the same formula with an extra factor of $e^{-\bar{d}^2/4}$, from using (90) in place of (89). This proves (5).

As noted above, (4) follows from (5) by Lemma 6.2, so the proof is complete. \square

7 Appendix

In this appendix we briefly show that Theorem 1.4 does indeed extend the asymptotic formula given by Bender, Canfield and McKay [6]. (It does not quite imply their result, since they have an explicit bound on the $1 + o(1)$ error term.)

Writing $P_2(s, m)$ for the probability that a random m -edge graph on $[s]$ is connected, Bender, Canfield and McKay showed that whenever $m = m(s)$ satisfies $m - s \rightarrow \infty$ and $m \leq \binom{s}{2} - s$, then

$$P_2(s, t) \sim e^{a(x)} \left(\frac{2e^{-x}y^{1-x}}{\sqrt{1-y^2}} \right)^s, \quad (93)$$

where $x = m/s$, $y = y(x)$ is defined implicitly by

$$2xy = \log \left(\frac{1+y}{1-y} \right), \quad (94)$$

and

$$a(x) = x(x+1)(1-y) + \log(1-x+xy) - \frac{1}{2} \log(1-x+xy^2). \quad (95)$$

Here we have changed the notation to match ours, and have simplified the more precise error term given in [6]. Note that in our notation x is simply $\bar{d}/2$.

Let $\bar{d} = 2m/s$ and define ξ as in (2) (with $r = 2$), so

$$\bar{d} = \Phi_2(\xi) = \log(1/\xi) \frac{1+\xi}{1-\xi}. \quad (96)$$

Set

$$y = \frac{1 - \xi}{1 + \xi}. \quad (97)$$

Then from (96) we have $\bar{d}y = \log(1/\xi) = \log\left(\frac{1+y}{1-y}\right)$, so (97) defines the same $y = y(\bar{d})$ as in [6].

Now using (97), $x = \bar{d}/2$ and (96) to write everything in terms of ξ , one can check that

$$\log\left(\frac{2e^{-x}y^{1-x}}{\sqrt{1-y^2}}\right) = -F_2(\bar{d}), \quad (98)$$

where $F_2(\bar{d})$ is defined in (3). (In fact, this computation is carried out in the appendix to [8].)

Similarly, writing $G_2(\bar{d})$ (defined in (10)) and $a(x)$ and hence $\exp(a(x))$ as a function of ξ , using, for example, Maple, one can verify that

$$\exp(a(x)) = G_2(\bar{d}). \quad (99)$$

Indeed, it turns out that

$$x(x+1)(1-y) = \frac{2\bar{d}\xi + \bar{d}^2\xi}{2(1+\xi)} = g_2(\bar{d}),$$

$$1 - x + xy = (1+\xi)^{-1}(1+\xi - \bar{d}\xi) = (1+\xi)^{-1}a_2(\bar{d}),$$

and

$$1 - x + xy^2 = (1+\xi)^{-2}((1+\xi)^2 - 2\bar{d}\xi) = (1+\xi)^{-2}b_2(\bar{d}).$$

These combine to give

$$\exp(a(x)) = \frac{a_2(\bar{d})}{\sqrt{b_2(\bar{d})}} e^{g_2(\bar{d})} = G_2(\bar{d})$$

as claimed. By (98) and (99) the $r = 2$ case of the formula (12) in Theorem 1.4 does indeed match (93).

References

- [1] T. Andriamampianina and V. Ravelomanana, Enumeration of connected uniform hypergraphs, Proceedings of 17th International Conference on Formal Power Series and Algebraic Combinatorics, Taormina (FPSAC 2005), (2005) pp. 387–398.
- [2] K.B. Athreya and P.E. Ney, *Branching processes*, Springer, Berlin (1972).
- [3] M. Behrisch, A. Coja-Oghlan and M. Kang, The order of the giant component of random hypergraphs, *Random Struct. Alg.* **36** (2010), 149–184.

- [4] M. Behrisch, A. Coja-Oghlan and M. Kang, Local limit theorems for the giant component of random hypergraphs, *Combin. Probab. Comput.* **23** (2014), 331–366.
- [5] M. Behrisch, A. Coja-Oghlan and M. Kang, The asymptotic number of connected d -uniform hypergraphs, *Combin. Probab. Comput.* **23** (2014), 367–385, with a Corrigendum *Combin. Probab. Comput.* **24** (2015), 373–375.
- [6] E.A. Bender, E.R. Canfield and B.D. McKay, The asymptotic number of labeled connected graphs with a given number of vertices and edges, *Random Struct. Alg.* **1**, 127–169 (1990).
- [7] B. Bollobás, Random Graphs, *Cambridge Studies in Advanced Mathematics*, vol. 73, second edn. Cambridge University Press, Cambridge (2001).
- [8] B. Bollobás and O. Riordan, Counting connected hypergraphs via the probabilistic method, *Combin. Probab. Comput.* **25** (2016), 21–75.
- [9] A. Coja-Oghlan, C. Moore and V. Sanwalani, Counting connected graphs and hypergraphs via the probabilistic method, *Random Struct. Alg.* **31** (2007), 288–329.
- [10] M. Karoński and T. Łuczak, The number of connected sparsely edged uniform hypergraphs, *Discrete Math.* **171** (1997), 153–167.
- [11] B.I. Selivanov, Enumeration of homogeneous hypergraphs with a simple cycle structure, *Kombinatornyi Anal.* **2** (1972), 60–67.