CrossMark

J. Appl. Probab. **59,** 278–296 (2022)
*doi:*10.1017/jpr.2021.50

THE CONTAINMENT PROFILE OF HYPER-RECURSIVE TREES

JOSHUA SPARKS,* **
SRINIVASAN BALAJI,* AND
HOSAM MAHMOUD,* *The George Washington University*

Abstract

We investigate vertex levels of containment in a random hypergraph grown in the spirit of a recursive tree. We consider a local profile tracking the evolution of the containment of a particular vertex over time, and a global profile concerned with counts of the number of vertices of a particular containment level.

For the local containment profile, we obtain the exact mean, variance, and probability distribution in terms of standard combinatorial quantities such as generalized harmonic numbers and Stirling numbers of the first kind. Asymptotically, we observe phases: the early vertices have an asymptotically normal distribution, intermediate vertices have a Poisson distribution, and late vertices have a degenerate distribution.

As for the global containment profile, we establish an asymptotically normal distribution for the number of vertices at the smallest containment level as well as their covariances with the number of vertices at the second smallest containment level and the variances of these numbers. The orders in the variance–covariance matrix establish concentration laws.

Keywords: Hypergraph; limit law; phase transition; martingale

2020 Mathematics Subject Classification: Primary 05082; 90B15
Secondary 60C05; 60F05

1. Hyper-recursive trees

Hypergraphs are generalizations of graphs. In a hypergraph, we have vertices and hyperedges consisting of collections of vertices. Also, the recursive tree is a well-studied structure; see [2], [7], [8], and [15], among many other sources. In this paper we propose a generalization of recursive trees to become hypergraphs.

A hyper-recursive tree with parameter (hyperedge size) θ grows as follows. Initially, there are θ originating vertices, all labeled with 0. These θ vertices constitute the first hyperedge. At each subsequent step, a vertex is added to the structure. The incoming vertex chooses $\theta - 1$ existing vertices to co-share a hyperedge. A vertex joining at time n is labeled with n. The choice of the vertices for the new hyperedge is done uniformly at random, with all subsets of vertices of size $\theta - 1$ being equally likely. The usual recursive tree is one with the parameter $\theta = 2$. Figure 1 illustrates the growth of a hyper-recursive tree at time $n = 2$. In this example $\theta = 3$, and the hyperedge appearing at step i is labeled e_i.

Received 30 July 2020; revision received 20 June 2021.
* Postal address: Department of Statistics, The George Washington University, Washington, DC 20052, USA.
** Email address: josparks@gwu.edu

© The Author(s), 2022. Published by Cambridge University Press on behalf of Applied Probability Trust.

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

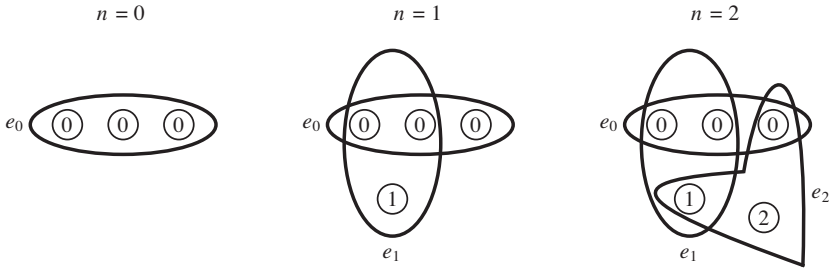


FIGURE 1. A hyper-recursive tree grown in two steps ($\theta = 3$).

Remark 1. For $\theta \geq 3$, the hyper-recursive tree is not a tree at all. We call such a hypergraph by this name only to preserve the historic origin of these structures and frame them as a generalization of genuine recursive trees ($\theta = 2$).

2. Scope

Our interest is in a local profile of the *level of containment* for a vertex, which is the number of hyperedges containing it. Let $C_{n,k}^{(\theta)}$ be the number of hyperedges containing vertex $k \geq 1$ at time $n \geq 0$. For instance, in the hyper-recursive tree in Figure 1, at $n = 2$, the vertex labeled 1 is at containment level 2, while the vertex labeled 2 is at containment level 1, and so $C_{2,1}^{(3)} = 2$ and $C_{2,2}^{(3)} = 1$. Each of the originators (all labeled with 0) may evolve differently and have different levels of containment at time n. However, the levels of these originators have the same distribution at time n, so we can choose a representative among them to include the case $k = 0$.

We later see that a vertex labeled $k = k(n)$ has an asymptotic distribution depending on the relation between k and n. Asymptotically, earlier vertices have Gaussian distributions and later vertices have shifted Poisson distributions (Theorem 3). The notions of 'early' and 'late' are made precise below, with nuances obtained from refinements into very early, early, intermediate, late, and very late vertices. We also discuss how the phases merge at the seam lines.

Moreover, we investigate a global profile of containment. Let $X_{n,i}^{(\theta)}$ be the number of vertices at containment level i, that is, the number of vertices contained in i edges after n evolutionary steps. For instance, in the example shown in Figure 1, we have $X_{2,1}^{(3)} = 1$, $X_{2,2}^{(3)} = 4$, and $X_{2,i}^{(3)} = 0$, for $i \geq 3$. To get a glimpse into the interaction between different levels of containment, we compute the variance–covariance matrix of $(X_{n,1}^{(\theta)}, X_{n,2}^{(\theta)})$ in Theorem 4. Furthermore, we develop a Gaussian distribution for $X_{n,1}^{(\theta)}$ in Theorem 5.

As an outline for the rest the paper, Section 3 will cover some preliminaries and notation for later use. In Section 4 we set up a moment generating function for $C_{n,k}^{(\theta)}$, from which we extract an exact distribution involving Stirling numbers of the first kind. The section contains a subsection where we portray the different phases in the asymptotic distribution.

Section 5 concerns the global containment profile. The derivation appears in incremental units. In Section 5.1 we set up a stochastic recurrence for $(X_{n,1}^{(\theta)}, X_{n,2}^{(\theta)})$. In Section 5.2 we use that recurrence to find means, variances, and covariances of these two components. In Section 5.3, we quantify the asymptotic error in $\mathcal{L}_1$, a necessary device for the martingale central limit theorem to go through. The variable $X_{n,1}^{(\theta)}$ is not a martingale *per se*: it is transformed into one

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

in Section 5.4. The martingale central limit theorem is worked out in Section 5.5 to give a Gaussian law.

3. Notation

Let $\tau_n^{(\theta)}$ be the cardinality of the vertex set (*size*) of the hyper-recursive tree at time n (right after the insertion of vertex n). We sometimes refer to n as the *age* of the hyper-recursive tree. Note that

$$\tau_n^{(\theta)} = n + \theta. \tag{1}$$

The exact results are represented in terms of Pochhammer's symbol for the rising factorial. The m-times rising factorial of a real number x is

$$\langle x \rangle_m = x(x+1)\cdots(x+m-1),$$

with the interpretation that $\langle x \rangle_0 = 1$. We make use of such an expression in its form as a generating function of the signless Stirling numbers of the first kind, namely

$$\langle x \rangle_m = \sum_{i=1}^{m} \begin{bmatrix} m \\ i \end{bmatrix} x^i, \tag{2}$$

where $\begin{bmatrix} m \\ i \end{bmatrix}$ is the ith (signless) Stirling number of order m of the first kind, a count of the number of permutations of $\{1, 2, \ldots, m\}$ that have i cycles. For properties of Stirling numbers, we refer the reader to [1] and [4].

Average values and variances contain generalized harmonic numbers. These are $H_n^{(s)}(x) = \sum_{k=1}^{n} 1/(k+x)^s$, for integer $n \geq 0$ and real s, $x \geq 0$. The superscript is often dropped when it is 1, and the number $H_n^{(1)}(0)$ is often written as H_n; we follow this convention. It is well known that, for any fixed x, as $n \to \infty$, we have

$$H_n(x) = H_n^{(1)}(x) = \ln(n) - \psi(x) - \frac{1}{x} + \mathrm{O}\left(\frac{1}{n}\right), \tag{3}$$

$$H_n^{(2)}(x) = \psi(1, x+1) + \mathrm{O}\left(\frac{1}{n}\right), \tag{4}$$

where $\psi(\cdot)$ is the digamma function and $\psi(1, \cdot)$ is the trigamma function. Note that $-\psi(x) - 1/x$ converges to Euler's constant γ, as $x \to 0$.

In the asymptotic analysis, we utilize the Stirling approximation of the ratio of growing Gamma functions [17]. Namely, for fixed a and b in $\mathbb{R}$, we have

$$\frac{\Gamma(x+a)}{\Gamma(x+b)} \sim x^{a-b}\left(1 + \frac{(a-b)(a+b-1)}{2x} + \mathrm{O}\left(\frac{1}{x^2}\right)\right) \quad \text{as } x \to \infty. \tag{5}$$

This approximation is applicable even if $a = a(x)$ and $b = b(x)$ grow slowly with x.

The sample taken at time n is drawn without replacement, and so the number of vertices in it at containment levels $1, \ldots, k$ have a (conditional) hypergeometric distribution. We use the notation Hypergeo$(\tau, n_1, n_2, \ldots, n_r; s)$ for the multivariate hypergeometric random vector, in which the ith component is the number of balls of color i that appear in a sample of size s drawn from an urn containing τ balls, of which n_i balls are of color i, for $i = 1, \ldots, r$. This multivariate hypergeometric distribution is standard and can be found in classical books on

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

distribution theory, such as [9]. In particular, we need the mean, variances, and covariance for a bivariate marginal distribution. Suppose Y_i is the number of balls of color i, for $i = 1, \ldots, r$, that appear in the sample. Then $(Y_i, Y_j, \tau - Y_i - Y_j)$ have a trivariate hypergeometric distribution like Hypergeo $(\tau, n_i, n_j, \tau - n_i - n_j; s)$, with Y_i distributed like Hypergeo $(\tau, n_i, \tau - n_i; s)$. Later, we utilize the formulas

$$\mathbb{E}[Y_i] = \frac{n_i}{\tau}\, s,$$
$$\mathrm{Var}\,[Y_i] = \frac{n_i(\tau - n_i)(\tau - s)}{\tau^2(\tau - 1)}\, s,$$
$$\mathrm{Cov}\,[Y_i, Y_j] = -\frac{n_i n_j(\tau - s)}{\tau^2(\tau - 1)}\, s;$$

see [9].

For two random variables X and Y, with induced measures $\mathbb{P}$ and $\mathbb{Q}$, respectively, on the real line, we define the total variation distance between the two random variables (or rather between their measures) by

$$d_{TV}(X, Y) = \sup_{A \in \mathcal{B}} |\mathbb{P}(A) - \mathbb{Q}(A)|,$$

where A is an element of the Borel sigma algebra $\mathcal{B}$ on $\mathbb{R}$. For a sequence of random variables in $\mathbb{N}_0$, convergence in total variation distance is equivalent to convergence in distribution [5].

To derive limit distributions with rates of convergence, we use the following version of Chen–Stein approximation [13].

Theorem 1. *Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = p_i$, $i = 1, 2, \ldots, n$. Let S_n be the sum $\sum_{i=1}^n X_i$ and $\lambda_n = \mathbb{E}[S_n] = \sum_{i=1}^n p_i$. We then have*

$$d_{TV}(S_n, \mathrm{Poi}\,(\lambda_n)) \le \min\left\{1, \frac{1}{\lambda_n}\right\} \sum_{i=1}^n p_i^2.$$

A useful agent for proofs based on the Poisson approximation is the following.

Lemma 1. (See [10].) *If $b > a$ and X and Y, respectively, are distributed like* Poi (b) *and* Poi (a), *then we have*

$$d_{TV}(X, Y) \le b - a.$$

To develop martingale differences, we use the backward difference operator ∇. Acting on a function h_n, this operator stands for $\nabla h_n = h_n - h_{n-1}$.

4. Local containment profile

Let $\mathbb{I}_{n,k}^{(\theta)}$ be an indicator of the event that vertex n chooses vertex k in the hyperedge appearing at time n. The indicator $\mathbb{I}_{n,k}^{(\theta)}$ is a Bernoulli random variable that assumes the value 1 with probability

$$\binom{\tau_{n-1}^{(\theta)} - 1}{\theta - 2} \Big/ \binom{\tau_{n-1}^{(\theta)}}{\theta - 1} = (\theta - 1)/\tau_{n-1}^{(\theta)}.$$

This indicator has the moment generating function

$$\psi_{n,k}^{(\theta)}(t) := \mathbb{E}\left[\mathrm{e}^{\mathbb{I}_{n,k}^{(\theta)} t}\right] = 1 - \frac{\theta - 1}{\tau_{n-1}^{(\theta)}} + \frac{(\theta - 1)\mathrm{e}^t}{\tau_{n-1}^{(\theta)}}. \tag{6}$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

Adding the indicators' contributions to $C_{n,k}$ across the various stages from when vertex k appears up to time n, we get the representation

$$C_{n,k}^{(\theta)} = 1 + \mathbb{I}_{k+1,k}^{(\theta)} + \mathbb{I}_{k+2,k}^{(\theta)} + \cdots + \mathbb{I}_{n,k}^{(\theta)} \tag{7}$$

as independent (but not identically distributed) indicator random variables. Recall that when $k = 0$, we are tracking a chosen representative among the originators (all labeled with 0).

Proposition 1. *Let $C_{n,k}^{(\theta)}$ be the containment level of the vertex labeled $k \geq 0$ in a hyper-recursive tree of edge size θ at age n. We have*

$$\mathbb{E}\left[C_{n,k}^{(\theta)}\right] = 1 + (\theta-1)(H_n(\theta-1) - H_k(\theta-1))$$

$$= \begin{cases} (\theta-1)\left(\ln(n) - \psi(\theta-1) - \dfrac{1}{\theta-1}\right) & \\ \quad +1 - (\theta-1)H_k(\theta-1) + \mathrm{O}\left(\dfrac{1}{n}\right) & \text{if } k \geq 0 \text{ is fixed}, \\ (\theta-1)\ln\left(\dfrac{n}{k}\right) + 1 + \mathrm{O}\left(\dfrac{1}{k}\right) & \text{if } n \geq k \to \infty. \end{cases}$$

and

$$\mathrm{Var}\left[C_{n,k}^{(\theta)}\right] = (\theta-1)(H_n(\theta-1) - H_k(\theta-1)) - (\theta-1)^2(H_n^{(2)}(\theta-1) - H_k^{(2)}(\theta-1))$$

$$= \begin{cases} (\theta-1)\ln(n) + \mathrm{O}(1) & \text{if } k \geq 0 \text{ is fixed}, \\ (\theta-1)\ln\left(\dfrac{n}{k}\right) + \mathrm{O}\left(\dfrac{1}{k}\right) & \text{if } n \geq k \to \infty. \end{cases}$$

Proof. Taking expectations of (7), we find

$$\begin{aligned} \mathbb{E}\left[C_{n,k}^{(\theta)}\right] &= 1 + \mathbb{E}\left[\mathbb{I}_{k+1,k}^{(\theta)}\right] + \mathbb{E}\left[\mathbb{I}_{k+2,k}^{(\theta)}\right] + \cdots + \mathbb{E}\left[\mathbb{I}_{n,k}^{(\theta)}\right] \\ &= 1 + \sum_{i=k+1}^{n} \frac{\theta-1}{\tau_{i-1}^{(\theta)}} \\ &= 1 + (\theta-1)\sum_{i=k+1}^{n} \frac{1}{i-1+\theta} \\ &= 1 + (\theta-1)(H_n(\theta-1) - H_k(\theta-1)). \end{aligned}$$

The asymptotic average follows from the approximation in (3).

By the independence of the indicators in (7), we similarly have

$$\begin{aligned} \mathrm{Var}\left[C_{n,k}^{(\theta)}\right] &= \sum_{i=k+1}^{n} \mathrm{Var}\left[\mathbb{I}_{n,k}^{(\theta)}\right] \\ &= \sum_{i=k+1}^{n} \frac{\theta-1}{i+\theta-1} - \sum_{i=k+1}^{n} \frac{(\theta-1)^2}{(i+\theta-1)^2} \\ &= (\theta-1)(H_n(\theta-1) - H_k(\theta-1)) - (\theta-1)^2(H_n^{(2)}(\theta-1) - H_k^{(2)}(\theta-1)). \end{aligned}$$

The asymptotic variance follows from the approximations in (3) and (4). □

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

Lemma 2. *The moment generating function* $\phi_{n,k}^{(\theta)}(t) = \mathbb{E}\left[\mathrm{e}^{C_{n,k}^{(\theta)} t}\right]$ *of* $C_{n,k}^{(\theta)}$ *is given by*

$$\phi_{n,k}^{(\theta)}(t) = \mathrm{e}^t \prod_{i=k}^{n-1} \frac{i+1+(\theta-1)\mathrm{e}^t}{i+\theta}.$$

Proof. The representation (7) as a sum of independent random variables gives rise to

$$\begin{aligned}\phi_{n,k}^{(\theta)}(t) &= \mathbb{E}\left[\mathrm{e}^{C_{n,k}^{(\theta)} t}\right] \\ &= \mathbb{E}\left[\mathrm{e}^{(1+\mathbb{I}_{k+1,k}^{(\theta)}+\mathbb{I}_{k+2,k}^{(\theta)}+\cdots+\mathbb{I}_{n,k}^{(\theta)})t}\right] \\ &= \mathrm{e}^t\, \psi_{k+1,k}^{(\theta)}(t)\, \psi_{k+2,k}^{(\theta)}(t) \cdots\ \psi_{n,k}^{(\theta)}(t) \quad \text{(by independence).}\end{aligned}$$

Plug in (6) and the sizes of the hyper-recursive trees in (1), and the statement follows after simplification. □

From Lemma 2, we develop an exact distribution.

Theorem 2. *For* $n \geq 1$ *and* $0 \leq k \leq n$, *let* $C_{n,k}^{(\theta)}$ *be the level of containment of the vertex* k *in a hyper-recursive tree with edge size* θ *at age* n. *For* $r \geq 1$, *we have*

$$\mathbb{P}(C_{n,k}^{(\theta)} = r) = \frac{(\theta-1)^{(r-1)}}{\langle k+\theta\rangle_{n-k}} \sum_{i=r-1}^{n-k} \begin{bmatrix} n-k \\ i \end{bmatrix} \binom{i}{r-1} (k+1)^{i-r+1}.$$

Proof. To obtain a probability generating function $\zeta_{n,k}^{(\theta)}(u) = \sum_{i=0}^{\infty} \mathbb{P}(C_{n,k}^{(\theta)} = i)u^i$, we substitute $\ln(u)$ for t in the generating function of Lemma 2. We obtain

$$\begin{aligned}\zeta_{n,k}^{(\theta)}(u) &= u\, \frac{(k+1+(\theta-1)u)\ldots(n+(\theta-1)u)}{(k+\theta)\cdots(n+\theta-1)} \\ &= u\, \frac{\langle k+1+(\theta-1)u\rangle_{n-k}}{\langle k+\theta\rangle_{n-k}}.\end{aligned}$$

Using the generating function (2), we write

$$\begin{aligned}\zeta_{n,k}^{(\theta)}(u) &= \frac{u}{\langle k+\theta\rangle_{n-k}} \sum_{i=0}^{n-k} \begin{bmatrix} n-k \\ i \end{bmatrix} (k+1+(\theta-1)u)^i \\ &= \frac{u}{\langle k+\theta\rangle_{n-k}} \sum_{i=0}^{n-k} \begin{bmatrix} n-k \\ i \end{bmatrix} \sum_{m=0}^{i} (k+1)^{i-m}(\theta-1)^m u^m \binom{i}{m} \\ &= \frac{u}{\langle k+\theta\rangle_{n-k}} \sum_{m=0}^{n-k} (\theta-1)^m u^m \sum_{i=m}^{n-k} \begin{bmatrix} n-k \\ i \end{bmatrix} \binom{i}{m} (k+1)^{i-m}.\end{aligned}$$

The exact distribution in the statement of the theorem follows upon extracting coefficients. □

4.1. Phases in the local containment profile of a vertex

Proposition 1 shows that the mean and variance of $C_{n,k}^{(\theta)}$ experience a phase change, as k increases relative to n. For instance, for fixed k, the mean is asymptotic to $(\theta-1)\ln(n)$, as

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

$n \to \infty$, and k can only alter lower-order asymptotics. However, a phase transition occurs when k grows to infinity with n but remains $o(n)$, such as the case $k = k(n) = \lceil 3n^{3/4} + 2\sqrt{n} - \pi \rceil$. In this range, k provides essential leading-term asymptotics. The vertices that appear in the entire range in which $k = o(n)$ are 'early'.

For the linear 'intermediate' range, $k \sim \alpha n$, for $0 < \alpha < 1$, such as the case $\lceil \frac{5}{7}n + 3\sqrt{n} + 6e \rceil$, the asymptotic mean is $1 + (\theta - 1)\ln(1/\alpha)$. Vertices in the range $k \sim n$ are considered 'late'. The asymptotic mean of late vertices is just 1, showing that the late arrivals, such as the case $k = \lfloor n - 5\ln(n+1) + 14 \rfloor$, have negligible probability of participating in the recruiting events.

It is possible to conceive of a bizarre relation between k and n, such as

$$k = \left\lfloor \left(\frac{1}{2} + \frac{(-1)^n}{3} \right) n \right\rfloor,$$

for which the mean containment level oscillates without settling on any asymptotic average. In these cases we have no convergence in the mean, variance, or distribution. Such an oscillating case is not likely to appear in practice.

Theorem 3. *Let $0 \le k \le n$ and let $C_{n,k}^{(\theta)}$ be the containment level of the vertex k in a hyper-recursive tree with edge size θ at age n. We have the following.*

(i) *For k fixed,*

$$\frac{C_{n,k}^{(\theta)} - (\theta - 1)\ln(n)}{\sqrt{\ln(n)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \theta - 1).$$

(ii) *For $k \to \infty$ and $k = o(n)$,*

$$\frac{C_{n,k}^{(\theta)} - (\theta - 1)\ln(n/k)}{\sqrt{\ln(n/k)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \theta - 1).$$

(iii) *For $k \sim \alpha n$ and $0 < \alpha < 1$,*

$$C_{n,k}^{(\theta)} \xrightarrow{\mathcal{D}} 1 + \mathrm{Poi}\left((\theta - 1)\ln\left(\frac{1}{\alpha} \right) \right).$$

(iv) *For $k = n - o(n)$,*

$$C_{n,k}^{(\theta)} \xrightarrow{\mathcal{P}} 1.$$

Proof. With $k_n = o(n)$ in parts (i) and (ii), $C_{n,k}^{(\theta)} - 1$ is the sum of an increasing number of indicators (see (7)), and the variance of $C_{n,k}^{(\theta)}$ increases with n without bound. So $C_{n,k}^{(\theta)}$ is amenable to Lindeberg's central limit theorem, as stated.

Part (iii) is amenable to Chen–Stein Poisson approximation. This line can also give us the rate of convergence in total variation distance, an added advantage. Consider the range $k_n \sim \alpha n$, for $\alpha \in (0, 1]$. Set $\lambda = (\theta - 1)\ln(1/\alpha)$. In our context, we have

$$d_{TV}\big(C_{n,k}^{(\theta)} - 1, \mathrm{Poi}(\lambda)\big) \le d_{TV}\big(C_{n,k}^{(\theta)} - 1, \mathrm{Poi}(\lambda_n)\big) + d_{TV}\big(\mathrm{Poi}(\lambda_n), \mathrm{Poi}(\lambda)\big). \tag{8}$$

Proposition 1 states that τ_n is of order $\ln n$. We have $\tau_n \ge 1$ for all $n \ge 0$. Theorem 1 and the asymptotic equivalent in (4) give us

$$d_{TV}\big(C_{n,k}^{(\theta)} - 1, \mathrm{Poi}(\lambda_n)\big) \le \frac{1}{\lambda_n} \sum_{i=k}^{n-1} \frac{(\theta - 1)^2}{(i + \theta)^2} = \mathrm{O}\left(\frac{1}{\ln n} \right) \times \mathrm{O}\left(\frac{1}{n} \right) = \left(\frac{1}{n \ln n} \right).$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

Lemma 1 asserts that

$$d_{TV}(\text{Poi}(\lambda_n), \text{Poi}(\lambda)) \leq |\lambda_n - \lambda| = \left| \sum_{i=k}^{n-1} \frac{\theta - 1}{i + \theta} - (\theta - 1) \ln\left(\frac{1}{\alpha}\right) \right|.$$

The asymptotic approximation (3) simplifies this upper bound to

$$d_{TV}(\text{Poi}(\lambda_n), \text{Poi}(\lambda)) \leq (\theta - 1) \left| \ln(n) + \text{O}\left(\frac{1}{n}\right) - \left(\ln(k) + \text{O}\left(\frac{1}{k}\right) \right) - \ln\left(\frac{1}{\alpha}\right) \right|,$$

which is $\text{O}(1/n)$. The sum of the two terms on the right-hand side of (8) is $\text{O}(1/n)$. This establishes part (iii), with $\text{O}(1/n)$ rate of convergence in total variation distance.

Then $C_{n,k}^{(\theta)}$ degenerates to a constant in the case $\alpha = 1$, where we get $C_{n,k}^{(\theta)} \xrightarrow{\mathcal{D}} 1$. Convergence in distribution to a constant implies convergence in probability, as stated in part (iv) of the theorem. □

Remark 2. Phase changes of the type in Theorem 3 have been observed in [3] and [11].

Remark 3. At the seam lines between the phases, the change is not abrupt. In fact, the phases 'flow into each other' in a natural way, and the role of k is negligible, so long as $\ln(k) = \text{o}(\sqrt{\ln(n)})$. Past this threshold, $\ln(k)$ cannot be neglected relative to $\sqrt{\ln(n)}$, and k must be included in the convergence. Again, the limit in part (iii) converges to that in part (iv), as $\alpha \to 1$.

We conclude this section by noting that Theorem 3 can be derived in a number of different ways. We systematically used the moment generating function to derive the exact distribution and all the parts of Theorem 3. Parts (i) and (ii) of Theorem 3 are amenable to Lindeberg's central limit theorem, since all the indicators in (7) are bounded and the variance of $C_{n,k}^{(\theta)}$ increases with n without a bound.

Part (iii) is amenable to Chen–Stein Poisson approximation. This line can also give us the rate of convergence in total variation distance, an added advantage. We denote such a distance between the measures of X and Y by $d_{TV}(X, Y)$. We outline this argument to find a $\text{O}(1/n)$ rate.

5. Global containment profile

In this section we look at a profile of the hyper-recursive tree determined by a raw count of the number of vertices at a particular containment level. Such a profile is global, as it cannot be determined without looking at all the vertices in the entire hyper-recursive tree.

We defined $X_{n,i}^{(\theta)}$ to be the number of vertices contained in exactly i hyperedges. To discern how the different containment levels interact, we investigate the mean of the row vector $(X_{n,1}^{(\theta)}, X_{n,2}^{(\theta)})$ and its covariance matrix.

We start with stochastic recurrences, from which we proceed to a calculation of the exact mean and variances, which lead us to concentration laws. Finally we establish a central limit theorem for the vertices at the smallest level of containment, i.e. the vertices contained in one hyperedge.

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

5.1. Stochastic recurrences

We discuss here recurrence equations that hold on the stochastic path. Let $Q_{n,i}^{(\theta)}$ be the number of vertices at containment level $i = 1, 2$ that appear in the sample chosen to construct the nth hyperedge. The row vector

$$\left(Q_{n,1}^{(\theta)}, Q_{n,2}^{(\theta)}, \tau_{n-1}^{(\theta)} - Q_{n,1}^{(\theta)} - Q_{n,2}^{(\theta)}\right)$$

has a (conditional) trivariate hypergeometric distribution that selects a sample of size $\theta - 1$ from among $\tau_{n-1}^{(\theta)}$ vertices, of which $X_{n,i}^{(\theta)}$ are at containment level i, for $i = 1, 2$. That is, the components of this row vector have the conditional joint distribution

$$\mathbb{P}\left(Q_{n,1}^{(\theta)} = q_1, Q_{n,2}^{(\theta)} = q_2 \mid X_{n-1,1}^{(\theta)}, X_{n-1,2}^{(\theta)}\right) = \frac{\binom{X_{n-1,1}^{(\theta)}}{q_1}\binom{X_{n-1,2}^{(\theta)}}{q_2}\binom{\tau_{n-1}^{(\theta)} - X_{n-1,1}^{(\theta)} - X_{n-1,2}^{(\theta)}}{\theta - 1 - q_1 - q_2}}{\binom{\tau_{n-1}^{(\theta)}}{\theta - 1}}.$$

The conditional means, variances, and covariance are specified in Section 3.

In the construction of the nth hyperedge, each vertex at containment level 1 in the sample upgrades to containment level 2, and the newly added vertex at step n is at containment level 1. Thus we have the stochastic recurrence

$$X_{n,1}^{(\theta)} = X_{n-1,1}^{(\theta)} - Q_{n,1}^{(\theta)} + 1. \tag{9}$$

Likewise, each vertex at containment level 2 in the sample upgrades to containment level 3. However, the $Q_{n,1}^{(\theta)}$ vertices at containment level 1 in the sample all move to containment level 2, giving rise to the stochastic recurrence

$$X_{n,2}^{(\theta)} = X_{n-1,2}^{(\theta)} - Q_{n,2}^{(\theta)} + Q_{n,1}^{(\theta)}. \tag{10}$$

5.2. The mean and covariance matrix

The pair of stochastic equations (9) and (10) is sufficient to determine the means exactly and the quadratic order moments asymptotically.

Proposition 2. *Let $X_{n,i}^{(\theta)}$ be the number of vertices contained in exactly i hyperedges, for $i = 1, 2$, of a recursive hyper-recursive tree of edge size θ. We have the mean vector*

$$\begin{pmatrix}\mathbb{E}\left[X_{n,1}^{(\theta)}\right] \\ \mathbb{E}\left[X_{n,2}^{(\theta)}\right]\end{pmatrix} = \begin{pmatrix} \frac{1}{\theta}\, n + 1 + (\theta - 1)\,\Gamma(\theta)\frac{\Gamma(n+1)}{\Gamma(n+\theta)} \\ \frac{(\theta-1)}{\theta^2}(n+\theta) + \Gamma(\theta)\frac{\Gamma(n+1)}{\Gamma(n+\theta)}\left((\theta-1)^2 H_n - \frac{\theta-1}{\theta}\right)\end{pmatrix}$$

$$= \begin{pmatrix} \frac{n}{\theta} + 1 + \mathrm{O}\left(\frac{1}{n^{\theta-1}}\right) \\ \frac{(\theta-1)}{\theta^2}(n+\theta) + \mathrm{O}\left(\frac{\ln(n)}{n^{\theta-1}}\right)\end{pmatrix}.$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

Proof. Let $\mathbb{F}_n^{(\theta)}$ be the sigma field generated by the first n hyperedge additions. Conditioning the stochastic relations (9) and (10) on $\mathbb{F}_{n-1}^{(\theta)}$, we obtain

$$\mathbb{E}\big[X_{n,1}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] = X_{n-1,1} - \mathbb{E}\big[Q_{n,1}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] + 1,$$
$$\mathbb{E}\big[X_{n,2}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] = X_{n-1,2}^{(\theta)} - \mathbb{E}\big[Q_{n,2}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] + \mathbb{E}\big[Q_{n,1}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big].$$

As discussed, the random variables $Q_{n,1}^{(\theta)}$, $Q_{n,2}^{(\theta)}$, and $\tau_{n-1}^{(\theta)} - Q_{n,1}^{(\theta)} - Q_{n,2}^{(\theta)}$ have a (conditional) trivariate hypergeometric marginal distribution; the conditional expectations are

$$\mathbb{E}\big[X_{n,1}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] = X_{n-1,1}^{(\theta)} - \frac{X_{n-1,1}^{(\theta)}}{\tau_{n-1}^{(\theta)}}(\theta-1) + 1, \tag{11}$$
$$\mathbb{E}\big[X_{n,2}^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] = X_{n-1,2}^{(\theta)} - \frac{X_{n-1,2}}{\tau_{n-1}^{(\theta)}}(\theta-1) + \frac{X_{n-1,1}^{(\theta)}}{\tau_{n-1}^{(\theta)}}(\theta-1);$$

see the formulas in Section 3. Taking an iterated average, simultaneous recurrences can be written as

$$\mathbb{E}\big[X_{n,1}^{(\theta)}\big] = \left(1 - \frac{\theta-1}{\tau_{n-1}^{(\theta)}}\right)\mathbb{E}\big[X_{n-1,1}^{(\theta)}\big] + 1 = \frac{n}{n+\theta-1}\mathbb{E}\big[X_{n-1,1}^{(\theta)}\big] + 1, \tag{12}$$
$$\begin{aligned}\mathbb{E}\big[X_{n,2}^{(\theta)}\big] &= \left(1 - \frac{\theta-1}{\tau_{n-1}^{(\theta)}}\right)\mathbb{E}\big[X_{n-1,2}^{(\theta)}\big] + \frac{\theta-1}{\tau_{n-1}^{(\theta)}}\mathbb{E}\big[X_{n-1,1}^{(\theta)}\big] \\ &= \frac{n}{n+\theta-1}\mathbb{E}\big[X_{n-1,2}^{(\theta)}\big] + \frac{\theta-1}{n+\theta-1}\mathbb{E}\big[X_{n-1,1}^{(\theta)}\big].\end{aligned} \tag{13}$$

The first of these two equations is self-contained, while the second has to wait for the solution of the first to be bootstrapped into it. The first equation has the standard form

$$y_n = g_n y_{n-1} + h_n, \tag{14}$$

with solution

$$y_n = \sum_{i=1}^{n} h_i \prod_{j=i+1}^{n} g_j + y_0 \prod_{j=1}^{n} g_j.$$

We solve (12), with $g_n = n/(n+\theta-1)$ and $h_n = 1$, and obtain

$$\begin{aligned}\mathbb{E}\big[X_{n,1}^{(\theta)}\big] &= \sum_{i=1}^{n}\prod_{j=i+1}^{n}\frac{j}{j+\theta-1} + \theta\prod_{j=1}^{n}\frac{j}{j+\theta-1} \\ &= \frac{\Gamma(n+1)}{\Gamma(n+\theta)}\sum_{i=1}^{n}\frac{\Gamma(i+\theta)}{\Gamma(i+1)} + \frac{\theta\,\Gamma(\theta)\,\Gamma(n+1)}{\Gamma(n+\theta)}.\end{aligned}$$

To simplify the sum, we use a known identity, which is namely

$$\sum_{i=1}^{n}\frac{\Gamma(i+\alpha)}{\Gamma(i+\beta)} = \frac{\Gamma(n+\alpha+1)}{(\alpha-\beta+1)\Gamma(n+\beta)} - \frac{\Gamma(\alpha+1)}{(\alpha-\beta+1)\Gamma(\beta)}, \tag{15}$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

for any given $\alpha, \beta \in \mathbb{R}^+$, such that $\beta \neq \alpha + 1$. Applying this identity with $\alpha = \theta \geq 2$ and $\beta = 1$, we obtain the stated result after some straightforward simplification.

The asymptotic formula given is a consequence of the Stirling approximation of the ratio of gamma functions (see (5)).

With $\mathbb{E}[X_{n,1}^{(\theta)}]$ in hand, we can bootstrap it into the recurrence for $\mathbb{E}[X_{n,2}^{(\theta)}]$ to put that recursion in the form (14) too. The solution follows similar steps to those used in solving the recurrence for $\mathbb{E}[X_{n,1}^{(\theta)}]$, and we only highlight the chief steps. The recurrence (13) is

$$\begin{aligned}\mathbb{E}\big[X_{n,2}^{(\theta)}\big] &= \frac{n}{n+\theta-1}\,\mathbb{E}\big[X_{n-1,2}^{(\theta)}\big] + \left(\frac{\theta-1}{n+\theta-1}\right)\left(\frac{n-1}{\theta} + 1 + \frac{(\theta-1)\,\Gamma(\theta)\,\Gamma(n)}{\Gamma(n+\theta-1)}\right)\\ &= \frac{n}{n+\theta-1}\,\mathbb{E}\big[X_{n-1,2}^{(\theta)}\big] + \left(\frac{\theta-1}{\theta} + \frac{(\theta-1)^2\,\Gamma(\theta)\,\Gamma(n)}{\Gamma(n+\theta)}\right).\end{aligned}$$

which is in the form (14). An exact solution follows after simplification. The asymptotic formula given is a consequence of the Stirling approximation of the ratio of gamma functions (see (5)). □

Theorem 4. *Let $X_{n,i}^{(\theta)}$ be the number of vertices contained in exactly i hyperedges in a hyper-recursive tree with parameter θ at age n, for $i = 1, 2$. Let $\boldsymbol{\Sigma}_n$ be the corresponding covariance matrix. Upon scaling by n, the covariance matrix converges (as $n \to \infty$) as follows:*

$$\frac{1}{n}\boldsymbol{\Sigma_n} \to \begin{pmatrix} \dfrac{(\theta-1)^2}{\theta^2(2\theta-1)} & -\dfrac{(\theta-1)^2(\theta^2+2\theta-1)}{\theta^3(2\theta-1)^2} \\ -\dfrac{(\theta-1)^2(\theta^2+2\theta-1)}{\theta^3(2\theta-1)^2} & \dfrac{(\theta-1)^2(6\theta^4-6\theta^3+8\theta^2-5\theta+1)}{\theta^4(2\theta-1)^3} \end{pmatrix}.$$

Proof. It is folklore that variance computation is very lengthy, a phenomenon called the *combinatorial explosion*. We only highlight the salient points.

The starting point for variance–covariance computation is the pair of stochastic recurrences (9) and (10), from which we can get stochastic recurrence relations for the second-order moments. We take the square of each of these equations, as well as their product:

$$\begin{aligned}\big(X_{n,1}^{(\theta)}\big)^2 &= \big(X_{n-1,1}^{(\theta)}\big)^2 + \big(Q_{n,1}^{(\theta)}\big)^2 + 1 - 2X_{n-1,1}^{(\theta)}Q_{n,1}^{(\theta)} + 2X_{n-1,1}^{(\theta)} - 2Q_{n,1}^{(\theta)},\\ X_{n,1}^{(\theta)}X_{n,2}^{(\theta)} &= X_{n-1,1}^{(\theta)}X_{n-1,2}^{(\theta)} - X_{n-1,1}^{(\theta)}Q_{n,2}^{(\theta)} + X_{n-1,1}^{(\theta)}Q_{n,1}^{(\theta)}\\ &\quad - X_{n-1,2}^{(\theta)}Q_{n,1}^{(\theta)} + Q_{n-1,1}^{(\theta)}Q_{n,2}^{(\theta)} - \big(Q_{n,1}^{(\theta)}\big)^2\\ &\quad + X_{n-1,2}^{(\theta)} - Q_{n,2}^{(\theta)} + Q_{n,1}^{(\theta)},\\ \big(X_{n,2}^{(\theta)}\big)^2 &= \big(X_{n-1,2}^{(\theta)}\big)^2 + \big(Q_{n,2}^{(\theta)}\big)^2 + \big(Q_{n,1}^{(\theta)}\big)^2 - 2X_{n-1,1}^{(\theta)}Q_{n,2}^{(\theta)}\\ &\quad + 2X_{n-1,2}^{(\theta)}Q_{n,1}^{(\theta)} - 2Q_{n,1}^{(\theta)}Q_{n,2}^{(\theta)}.\end{aligned}$$

We next take the expectation (conditioned on $\mathbb{F}_{n-1}^{(\theta)}$) and use the conditional trivariate hypergeometric distribution of $(Q_{n,1}^{(\theta)}, Q_{n,2}^{(\theta)}, \tau_{n-1}^{(\theta)} - Q_{n,1}^{(\theta)} - Q_{n,2}^{(\theta)})$, which comes in terms of $(X_{n-1,1}^{(\theta)}, X_{n-1,2}^{(\theta)})$. So an iterated expectation on each conditional recurrence gives us three unconditional recurrence equations in $\mathbb{E}\big[(X_{n,1}^{(\theta)})^2\big]$, $\mathbb{E}\big[X_{n,1}^{(\theta)}X_{n,2}^{(\theta)}\big]$, and $\mathbb{E}\big[(X_{n,2}^{(\theta)})^2\big]$.

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

It is evident that we need a bootstrapping technique. The recurrence equation for $\mathbb{E}\big[(X_{n,1}^{(\theta)})^2\big]$ is self-contained, going back only to $\mathbb{E}\big[(X_{n-1,1}^{(\theta)})^2\big]$, so we can start there. It has the form (14), with solution

$$\mathbb{E}\big[\big(X_{n,1}^{(\theta)}\big)^2\big] = \frac{n^2}{\theta^2} + \frac{5\theta^2 - 4\theta + 1}{\theta^2(2\theta - 1)}\, n + \mathrm{O}(1).$$

The recurrence equation for $\mathbb{E}\big[X_{n,1}^{(\theta)} X_{n,2}^{(\theta)}\big]$ involves $\mathbb{E}\big[\big(X_{n-1,1}^{(\theta)}\big)^2\big]$, which is now available. So, after all, the equation is of the form (14) (with some of the terms in h_n specified only asymptotically). We obtain the asymptotic solution

$$\mathbb{E}\big[X_{n,1}^{(\theta)} X_{n,2}^{(\theta)}\big] = \frac{\theta - 1}{\theta^2}\, n^2 + \frac{7\theta^4 - 16\theta^3 + 14\theta^2 - 6\theta + 1}{\theta^3(2\theta - 1)^2}\, n + \mathrm{O}(\ln(n)).$$

Similarly, the recurrence equation for $\mathbb{E}\big[(X_{n,2}^{(\theta)})^2\big]$ involves $\mathbb{E}\big[(X_{n-1,1}^{(\theta)})^2\big]$ and $\mathbb{E}\big[X_{n-1,1}^{(\theta)} X_{n-1,2}^{(\theta)}\big]$, which are both available now. Again, the equation is of the form (14) (with h_n specified only asymptotically). This gives the solution

$$\mathbb{E}\big[(X_{n,2}^{(\theta)})^2\big] = \frac{(\theta - 1)^2}{\theta^4}\, n^2 + \frac{(22\theta^4 - 30\theta^3 + 20\theta^2 - 7\theta + 1)(\theta - 1)^2}{\theta^4(2\theta - 1)^3}\, n + \mathrm{O}(\ln(n)).$$

For the variance, from the second moments we subtract the expectations of the squares of the first moments, and for the covariance, we subtract $\mathbb{E}[X_{n,1}^{(\theta)}]\,\mathbb{E}[X_{n,2}^{(\theta)}]$ from the mixed moment. Huge cancelations take place, removing the n^2-term from the expression and leaving the covariance matrix convergence as stated. A reader interested in checking out all the details can find them in the extended arXiv version [16]. □

5.3. Concentration laws

We determine approximations of $X_{n,1}^{(\theta)}$ and $X_{n,2}^{(\theta)}$ by the leading asymptotic equivalents of their means. A sequence of random variables Y_n is $\mathrm{O}_{\mathcal{L}_1}(g(n))$ when there exist a positive constant A and a positive integer n_0, such that $\mathbb{E}[|Y_n|] \le A|g(n)|$ for all $n \ge n_0$. Errors in the $\mathrm{O}_{\mathcal{L}_1}$ sense are of lower order.

Lemma 3. *As $n \to \infty$, we have the asymptotic approximation*

$$\begin{pmatrix} X_{n,1}^{(\theta)} \\ X_{n,2}^{(\theta)} \end{pmatrix} = \begin{pmatrix} \dfrac{n}{\theta} + \mathrm{O}_{\mathcal{L}_1}(\sqrt{n}) \\ \dfrac{(\theta - 1)n}{\theta^2} + \mathrm{O}_{\mathcal{L}_1}(\sqrt{n}) \end{pmatrix}.$$

Proof. From the asymptotics of the mean and variance, as given in Proposition 2 and Theorem 4, we have

$$\begin{aligned} \mathbb{E}\left[\left(X_{n,1}^{(\theta)} - \frac{n}{\theta}\right)^2\right] &= \mathbb{E}\left[\left(\left(X_{n,1}^{(\theta)} - \mathbb{E}\big[X_{n,1}^{(\theta)}\big]\right) + \left(\mathbb{E}\big[X_{n,1}^{(\theta)}\big] - \frac{n}{\theta}\right)\right)^2\right] \\ &= \mathrm{Var}\big[X_{n,1}^{(\theta)}\big] + \left(\mathbb{E}\big[X_{n,1}^{(\theta)}\big] - \frac{n}{\theta}\right)^2 \\ &= \mathrm{O}(n). \end{aligned}$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

So, by Jensen's inequality

$$\mathbb{E}\left[\left|X_{n,1}^{(\theta)} - \frac{n}{\theta}\right|\right] \leq \sqrt{\mathbb{E}\left[\left(X_{n,1}^{(\theta)} - \frac{n}{\theta}\right)^2\right]} = \mathrm{O}(\sqrt{n}).$$

It follows that

$$X_{n,1}^{(\theta)} = \frac{n}{\theta} + \mathrm{O}_{\mathcal{L}_1}(\sqrt{n}).$$

The proof of the asymptotic approximation for $X_{n,2}^{(\theta)}$ is quite similar. □

5.4. Martingalization

We perform a martingale transform on $X_{n,1}^{(\theta)}$. Let $M_n^{(\theta)} = r_n^{(\theta)} X_{n,1}^{(\theta)} + s_n^{(\theta)}$, for deterministic – but yet-to-be specified – factors $r_n^{(\theta)}$ and $s_n^{(\theta)}$ that render $M_n^{(\theta)}$ a martingale. For such martingalization, using (11), we write

$$\begin{aligned}
\mathbb{E}\big[M_n^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] &= \mathbb{E}\big[r_n^{(\theta)} X_{n,1}^{(\theta)} + s_n^{(\theta)} \mid \mathbb{F}_{n-1}^{(\theta)}\big] \\
&= \left(1 - \frac{\theta - 1}{\tau_{n-1}^{(\theta)}}\right) r_n^{(\theta)} X_{n-1,1}^{(\theta)} + r_n^{(\theta)} + s_n^{(\theta)} \\
&= M_{n-1}^{(\theta)} \\
&= r_{n-1}^{(\theta)} X_{n-1,1}^{(\theta)} + s_{n-1}^{(\theta)}.
\end{aligned}$$

This is possible if

$$\begin{aligned}
r_{n-1}^{(\theta)} &= \left(1 - \frac{\theta - 1}{\tau_{n-1}^{(\theta)}}\right) r_n^{(\theta)}, \\
s_{n-1}^{(\theta)} &= r_n^{(\theta)} + s_n^{(\theta)}.
\end{aligned}$$

The factor $r_n^{(\theta)}$ should satisfy the recurrence

$$r_n^{(\theta)} = \left(\frac{\tau_{n-1}^{(\theta)}}{\tau_{n-1}^{(\theta)} - \theta + 1}\right) r_{n-1} = \left(\frac{n + \theta - 1}{n}\right) r_{n-1}^{(\theta)},$$

which unwinds into

$$r_n^{(\theta)} = \frac{(n + \theta - 1)(n + \theta - 2) \cdots \theta}{n(n-1) \times \cdots \times 1} \, r_0^{(\theta)} = \frac{\Gamma(n + \theta)}{\Gamma(\theta)\,\Gamma(n+1)} \, r_0^{(\theta)}$$

for any arbitrary $r_0^{(\theta)} \in \mathbb{R}$; for simplicity, we take $r_0^{(\theta)} = 1$.

The factor $s_n^{(\theta)}$ should satisfy the recurrence

$$s_n^{(\theta)} = s_{n-1}^{(\theta)} - r_n^{(\theta)},$$

which unwinds into

$$s_n^{(\theta)} = s_0^{(\theta)} - \sum_{i=1}^{n} r_i^{(\theta)}$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

for any arbitrary $s_0^{(\theta)} \in \mathbb{R}$; we take $s_0^{(\theta)} = 0$. Using the identity (15) once again, we simplify $s_n^{(\theta)}$ to

$$s_n^{(\theta)} = -\frac{\Gamma(n+\theta+1)}{\theta\,\Gamma(\theta)\,\Gamma(n+1)} + 1.$$

Thus we deduce that

$$M_n^{(\theta)} = \frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1)}\,X_{n,1}^{(\theta)} - \frac{\Gamma(n+\theta+1)}{\theta\,\Gamma(\theta)\,\Gamma(n+1)} + 1$$

is a martingale.

Asymptotics of $r_n^{(\theta)}$, $s_n^{(\theta)}$ and their backward differences are useful in the ensuing analysis.

Lemma 4. *As $n \to \infty$, we have the asymptotics*

$$r_n^{(\theta)} \sim \frac{n^{\theta-1}}{\Gamma(\theta)}, \quad r_n^{(\theta)} - r_{n-1}^{(\theta)} = \mathrm{O}(n^{\theta-2}),$$
$$s_n^{(\theta)} \sim -\frac{n^{\theta}}{\theta\,\Gamma(\theta)}, \quad s_n^{(\theta)} - s_{n-1}^{(\theta)} = \mathrm{O}(n^{\theta-1}).$$

Proof. Examine the forms of $r_n^{(\theta)}$ and $s_n^{(\theta)}$. Their asymptotic equivalents follow from the Stirling approximation in (5).

Further, we have

$$\begin{aligned} r_n^{(\theta)} - r_{n-1}^{(\theta)} &= \frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1)} - \frac{\Gamma(n-1+\theta)}{\Gamma(\theta)\,\Gamma(n)} \\ &= \frac{\Gamma(n-1+\theta)}{\Gamma(\theta)\,\Gamma(n)}\left(\frac{n-1+\theta}{n} - 1\right) \\ &\sim \frac{\theta-1}{\Gamma(\theta)}\,n^{\theta-2} \end{aligned}$$

and

$$\begin{aligned} s_n^{(\theta)} - s_{n-1}^{(\theta)} &= -\frac{\Gamma(n+\theta+1)}{\theta\,\Gamma(\theta)\,\Gamma(n+1)} + \frac{\Gamma(n+\theta)}{\theta\,\Gamma(\theta)\,\Gamma(n)} \\ &= -\frac{\Gamma(n+\theta)}{\theta\,\Gamma(\theta)\,\Gamma(n)}\left(\frac{n+\theta}{n} - 1\right) \\ &= -\frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1)} \\ &\sim \frac{n^{\theta-1}}{\Gamma(\theta)}. \end{aligned}$$

□

Corollary 1. *For large enough positive constants K_1, K_2, and K_3, we have*

$$r_n^{(\theta)} \le K_1 n^{\theta-1}, \quad r_n^{(\theta)} - r_{n-1}^{(\theta)} \le K_2 n^{\theta-2}, \quad s_n^{(\theta)} - s_{n-1}^{(\theta)} = K_3 n^{\theta-1}$$

for all $n \ge 1$.

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

5.5. Gaussian limit law

In this subsection we obtain an asymptotic Gaussian law for $X_{n,1}^{(\theta)}$ by verifying the conditions of the martingale central limit theorem for $M_n^{(\theta)}$. We use the conditional Lindeberg condition and the conditional variance condition in [6, pages 57–59].

The conditional Lindeberg condition requires that for some positive sequence $\xi_n^{(\theta)}$ and for any $\varepsilon > 0$ we have

$$U_n^{(\theta)} := \sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{\nabla M_j^{(\theta)}}{\xi_n^{(\theta)}}\right)^2 \mathbb{I}_{\{|\nabla M_j^{(\theta)}/\xi_n^{(\theta)}|>\varepsilon\}} \,\middle|\, \mathbb{F}_{j-1}^{(\theta)}\right] \xrightarrow{\mathcal{P}} 0,$$

and the conditional variance condition requires that, for some random variable $G_\theta \not\equiv 0$, we have

$$V_n^{(\theta)} := \sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{\nabla M_j^{(\theta)}}{\xi_n^{(\theta)}}\right)^2 \,\middle|\, \mathbb{F}_{j-1}^{(\theta)}\right] \xrightarrow{\mathcal{P}} G_\theta.$$

When these conditions are satisfied, we get

$$\frac{M_n^{(\theta)}}{\xi_n^{(\theta)}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, G_\theta),$$

where the right-hand side is a mixture of normally distributed random variates, with mixing variance G_θ. In our case we find that G_θ is a constant, so the mixture has only one normal random variate in it, which has a deterministic variance. In the case of the number of vertices at containment level 1, it turns out that $\xi_n^{(\theta)}$ is $n^{\theta-1/2}$.

The following uniform bound paves the way to the verification of the two conditions of the martingale central limit theorem.

Lemma 5. *The absolute differences* $|\nabla M_j^{(\theta)}|/n^{\theta-1}$ *are uniformly bounded in* $j = 1, \ldots, n$.

Proof. By the construction of the martingale, for each $1 \le j \le n$, we have

$$\begin{aligned}
\left|\nabla M_j^{(\theta)}\right| &= \left|M_j^{(\theta)} - M_{j-1}^{(\theta)}\right| \\
&= \left|\left(r_j^{(\theta)} X_{j,1}^{(\theta)} + s_j^{(\theta)}\right) - \left(r_{j-1}^{(\theta)} X_{j-1,1}^{(\theta)} + s_{j-1}^{(\theta)}\right)\right| \\
&\le \left|r_j^{(\theta)}\left(X_{j-1,1}^{(\theta)} - Q_{j,1}^{(\theta)} + 1\right) - r_{j-1} X_{j-1,1}^{(\theta)}\right| + \left|s_j^{(\theta)} - s_{j-1}^{(\theta)}\right| \\
&\le \left|r_j^{(\theta)} - r_{j-1}^{(\theta)}\right| \tau_{j-1}^{(\theta)} + r_j^{(\theta)} Q_{j,1}^{(\theta)} + r_j^{(\theta)} + \left|s_j^{(\theta)} - s_{j-1}^{(\theta)}\right|.
\end{aligned}$$

Recall that $Q_{n,1}^{(\theta)}$ is a hypergeometric random variable representing the number of vertices at containment level 1 in a sample of size $\theta - 1$. Its maximum value is $\theta - 1$.

From the bounds in Corollary 1, we obtain

$$\begin{aligned}
\left|\frac{\nabla M_j^{(\theta)}}{n^{\theta-1}}\right| &\le \frac{1}{n^{\theta-1}}\left(K_2 j^{\theta-2} \tau_{n-1}^{(\theta)} + (\theta-1) K_1 j^{\theta-1} + K_1 j^{\theta-1} + K_3 n^{\theta-1}\right) \\
&\le \frac{1}{n^{\theta-1}}\left(K_2 n^{\theta-2}(n+\theta-1) + \theta K_1 n^{\theta-1}\right) + K_3 \\
&= K_2 + \frac{(\theta-1)K_2}{n} + \theta K_1 + K_3 \\
&\le K_2 + (\theta-1)K_2 + \theta K_1 + K_3.
\end{aligned}$$

□

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

Lemma 6. *For any $\varepsilon > 0$, we have*

$$U_n^{(\theta)} = \sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{\nabla M_j^{(\theta)}}{n^{\theta-1/2}} \right)^2 \mathbb{I}_{\{|\nabla M_j^{(\theta)}/n^{\theta-1/2}| > \varepsilon\}} \,\middle|\, \mathbb{F}_{j-1}^{(\theta)} \right] \xrightarrow{\mathcal{P}} 0.$$

Proof. By the uniform bound established in Lemma 5, for every $\varepsilon > 0$, there exists a natural number $n_0(\varepsilon)$, such that for all $n \geq n_0(\varepsilon)$, the sets $\{|\nabla M_j^{(\theta)}| > \varepsilon n^{\theta-1/2}\}$ are empty, which implies that the sequence U_n converges almost surely to 0. This almost-sure convergence is stronger than the required in-probability convergence. □

Lemma 7.

$$V_n^{(\theta)} = \sum_{j=1}^{n} \mathbb{E}\left[\left(\frac{\nabla M_j^{(\theta)}}{n^{\theta-1/2}} \right)^2 \,\middle|\, \mathbb{F}_{j-1}^{(\theta)} \right] \xrightarrow{\mathcal{P}} \frac{\theta - 1}{\theta^2 (2\theta - 1)\, \Gamma^2(\theta)}.$$

Proof. This is a rather lengthy calculation, but for the large part it goes in the same vein as the computations we encountered in the proof of Theorem 4. So we only outline the salient features in the long chain of calculations.

Write

$$\begin{aligned} V_n^{(\theta)} &= \frac{1}{n^{2\theta-1}} \sum_{j=1}^{n} \mathbb{E}\big[\big(\nabla\big(r_j^{(\theta)} X_{j,1}^{(\theta)} \big) + \nabla s_j^{(\theta)} \big)^2 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] \\ &= \frac{1}{n^{2\theta-1}} \sum_{j=1}^{n} \mathbb{E}\big[\big(\nabla\big(r_j^{(\theta)} X_{j,1}^{(\theta)} \big) \big)^2 + 2 \big(\nabla\big(r_j^{(\theta)} X_{j,1}^{(\theta)} \big) \big) \nabla s_j^{(\theta)} + \big(\nabla s_j^{(\theta)} \big)^2 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] \\ &:= \frac{1}{n^{2\theta-1}} \sum_{j=1}^{n} \big(A_j^{(\theta)} + B_j^{(\theta)} + D_j^{(\theta)} \big). \end{aligned}$$

We take up each part separately, starting with

$$\begin{aligned} A_j^{(\theta)} &:= \mathbb{E}\big[\big(\nabla\big(r_j^{(\theta)} X_{j,1}^{(\theta)} \big) \big)^2 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] \\ &= \mathbb{E}\big[\big(r_j^{(\theta)} X_{j,1}^{(\theta)} - r_{j-1}^{(\theta)} X_{j-1,1}^{(\theta)} \big)^2 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] \\ &= \big(r_j^{(\theta)} \big)^2 \mathbb{E}\big[\big(X_{j,1}^{(\theta)} \big)^2 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] + \big(r_{j-1}^{(\theta)} \big)^2 \big(X_{j-1,1}^{(\theta)} \big)^2 \\ &\quad - 2 r_j^{(\theta)} r_{j-1}^{(\theta)} X_{j-1,1}^{(\theta)} \, \mathbb{E}\big[X_{j,1}^{(\theta)} \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] \\ &= \big(r_j^{(\theta)} \big)^2 \mathbb{E}\big[\big(X_{j-1,1}^{(\theta)} - Q_{j,1}^{(\theta)} + 1 \big)^2 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big] + \big(r_{j-1}^{(\theta)} \big)^2 \big(X_{j-1,1}^{(\theta)} \big)^2 \\ &\quad - 2 r_j^{(\theta)} r_{j-1}^{(\theta)} X_{j-1,1}^{(\theta)} \, \mathbb{E}\big[X_{j-1,1}^{(\theta)} - Q_{j,1}^{(\theta)} + 1 \,|\, \mathbb{F}_{j-1}^{(\theta)} \big], \end{aligned}$$

where we substituted the right-hand side of (9) for $X_{j,1}^{(\theta)}$. Upon expanding, we get conditional expectations (given $\mathbb{F}_{j-1}^{(\theta)}$) of both $(Q_{j,1}^{(\theta)})^2$ and $Q_{j,1}^{(\theta)}$. The variable $Q_{j,1}^{(\theta)}$ is

$$\text{Hypergeo}\big(\tau_{j-1}^{(\theta)}, X_{j-1,1}^{(\theta)}, \theta - 1 \big).$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

The required conditional expectations are obtained from the hypergeometric distribution; see Section 3. The second part is

$$\begin{aligned} B_j^{(\theta)} &= \mathbb{E}\big[2\big(\nabla\big(r_j^{(\theta)}X_{j,1}^{(\theta)}\big)\big)\nabla s_j^{(\theta)} \mid \mathbb{F}_{j-1}^{(\theta)}\big] \\ &= 2\,\mathbb{E}\big[\big(r_j^{(\theta)}X_{j,1}^{(\theta)} - r_{j-1}^{(\theta)}X_{j-1,1}^{(\theta)}\big)\big(s_j^{(\theta)} - s_{j-1}^{(\theta)}\big) \mid \mathbb{F}_{j-1}^{(\theta)}\big] \\ &= 2\,r_j^{(\theta)}\big(s_j^{(\theta)} - s_{j-1}^{(\theta)}\big)\,\mathbb{E}\big[X_{j,1}^{(\theta)} \mid \mathbb{F}_{j-1}^{(\theta)}\big] - 2r_{j-1}^{(\theta)}\big(s_j^{(\theta)} - s_{j-1}^{(\theta)}\big)X_{j-1,1}^{(\theta)}. \end{aligned}$$

The third part is

$$D_j^{(\theta)} = \mathbb{E}\big[\big(\nabla s_j^{(\theta)}\big)^2 \mid \mathbb{F}_{j-1}^{(\theta)}\big] = \big(s_j^{(\theta)} - s_{j-1}^{(\theta)}\big)^2.$$

We now put the three parts together and get an expression for $V_n^{(\theta)}$ as a sum, in which the summand is in terms of $r_j^{(\theta)}$ and $s_j^{(\theta)}$, and their backward differences, as well as $X_{j-1,1}^{(\theta)}$. For simplified asymptotics, we use the asymptotic equivalents in Lemma 3 for $X_{j-1,1}^{(\theta)}$, and for $r_j^{(\theta)}$, $s_j^{(\theta)}$, $\nabla r_j^{(\theta)}$, $\nabla s_j^{(\theta)}$, we use the asymptotics in Lemma 4.

Huge cancelations take place, leaving

$$\begin{aligned} V_n^{(\theta)} &= \frac{1}{n^{2\theta-1}}\left(\sum_{j=1}^{n}\frac{(\theta-1)^2}{\theta^2\,\Gamma^2(\theta)}j^{2\theta-2} + \mathrm{O}_{\mathcal{L}_1}(j^{2\theta-3})\right) \\ &= \frac{(\theta-1)^2}{\theta^2(2\theta-1)\,\Gamma^2(\theta)} + \mathrm{O}_{\mathcal{L}_1}\left(\frac{1}{n}\right) \\ &\xrightarrow{\mathcal{L}_1} \frac{(\theta-1)^2}{\theta^2(2\theta-1)\,\Gamma^2(\theta)}. \end{aligned}$$

This $\mathcal{L}_1$ convergence is stronger than the required in-probability convergence. □

Theorem 5. *Let $X_{n,1}^{(\theta)}$ be the number of vertices in hyper-recursive tree with hyperedges of size θ at age n. Then, as $n \to \infty$, we have*

$$\frac{X_{n,1}^{(\theta)} - n/\theta}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{(\theta-1)^2}{\theta^2(2\theta-1)}\right).$$

Proof. Having checked the conditions for the martingale central limit theorem, we can ascertain that

$$\begin{aligned} \frac{M_n^{\theta}}{n^{\theta-1/2}} &= \frac{\frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1)}\,X_{n,1}^{(\theta)} - \frac{\Gamma(n+\theta+1)}{\theta\,\Gamma(\theta)\,\Gamma(n+1)} + 1}{n^{\theta-1/2}} \\ &= \frac{\frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1)}\,X_{n,1}^{(\theta)} - \frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1)}\left(\frac{n}{\theta}\right) - \frac{\Gamma(n+\theta)}{\Gamma(\theta)\,\Gamma(n+1))} + 1}{n^{\theta-1/2}} \\ &\xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{(\theta-1)^2}{\theta^2(2\theta-1)\,\Gamma^2(\theta)}\right). \end{aligned}$$

We have

$$\left(-\frac{\Gamma(n+\theta)}{\Gamma(n+1)\Gamma(\theta)} + 1\right)(n^{1/2-\theta}) \to 0,$$

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

and so an application of Slutsky's theorem [14] allows us to remove this term. By the Stirling approximation in (5), we have

$$\frac{n^{\theta-1}\Gamma(n+1)}{\Gamma(n+\theta)} \to 1.$$

An application of Slutsky's theorem [14] yields

$$\frac{\frac{1}{\Gamma(\theta)}\left(X_{n,1}^{(\theta)} - \frac{n}{\theta}\right)}{\sqrt{n}} \xrightarrow{\mathcal{D}} N\left(0, \frac{(\theta-1)^2}{\theta^2(2\theta-1)\,\Gamma^2(\theta)}\right),$$

which is an equivalent statement to the one given in the theorem. □

Remark 4. In the very special case $\theta = 2$, the hyper-recursive tree is the standard uniform recursive tree. In this case $X_{n,1}^{(2)}$ is just a count of the leaves in the tree. Theorem 5 recovers the result in [12] and generalizes it.

Acknowledgements

We wish to thank the referees for suggesting a redirection of the proof that led to a finer presentation.

Funding information

There are no funding bodies to thank relating to this creation of this article.

Competing interests

There were no competing interests to declare which arose during the preparation or publication process of this article.

References

[1] David, F. and Barton, E. (1962). *Combinatorial Chance*. Charles Griffin, London.

[2] Drmota, M. (2009). *Random Trees: An Interplay Between Combinatorics and Probability*. Springer, New York.

[3] Feng, Q., Mahmoud, H. and Panholzer, A. (2008). Phase changes in subtree varieties in random recursive trees and binary search trees. *SIAM J. Discrete Math.* **22**, 160–184.

[4] Graham, R., Knuth, D. and Patashnik, O. (1994). *Concrete Mathematics*. Addison-Wesley, Reading, MA.

[5] Gut, A. (2013). *Probability: A Graduate Course*, 2nd edn. Springer.

[6] Hall, P. and Heyde, C. (1980). *Martingale Limit Theory and its Application*. Academic Press, New York.

[7] Hofri, M. and Mahmoud, H. (2018). *Algorithmics of Nonuniformity: Tools and Paradigms*. CRC Press, Boca Raton, FL.

[8] Frieze, A. M. and Karoński, M. (2015). *Introduction to Random Graphs*. Cambridge University Press.

[9] Kendall, M., Stuart, A. and Ord, K. (1987). *Advanced Theory of Statistics*, Vol. I, *Distribution Theory*. Oxford University Press.

[10] Lyon, M. and Mahmoud, H. (2020). Trees grown under young-age preferential attachment. *J. Appl. Prob.* **57**, 911–927.

[11] Mahmoud, H. (2019). Local and global degree profiles of randomly grown self-similar hooking networks under uniform and preferential attachment. *Adv. Appl. Math.* **111**, 101930.

[12] Mahmoud, H. and Smythe, R. (1992). Asymptotic joint normality of outdegrees of nodes in random recursive trees. *Random Structures Algorithms* **3**, 255–266.

[13] Ross, N. (2011). Fundamentals of Stein's method. *Prob. Surveys* **8**, 210–293.

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press

[14] SLUTSKY, E. (1925). Über stochastische Asymptoten und Grenzwerte. *Metron* **5**, 3–89.
[15] SMYTHE, R. AND MAHMOUD, H. (1996). A survey of recursive trees. *Theory Prob. Math. Statist.* **51**, 1–29 (appeared in Ukrainian in 1994).
[16] SPARKS, J., BALAJI, S. AND MAHMOUD, H. (2021). The containment profile of hyperrecursive trees. Available at arXiv:2101.06359.
[17] TRICOMI, F. AND ERDÉLYI, A. (1951). The asymptotic expansion of a ratio of gamma functions. *Pacific J. Math.* **1**, 133–142.
[18] WILLIAMS, D. (1991). *Probability with Martingales*. Cambridge University Press, New York.

https://doi.org/10.1017/jpr.2021.50 Published online by Cambridge University Press