HIGH DEGREES OF RANDOM RECURSIVE TREES

LOUIGI ADDARIO-BERRY AND LAURA ESLAVA

Abstract. For $n \geq 1$, let T_n be a random recursive tree (RRT) on the vertex set $[n] = \{1, \ldots, n\}$. Let $\deg_{T_n}(v)$ be the degree of vertex v in T_n, that is, the number of children of v in T_n. Devroye and Lu [6] showed that the maximum degree Δ_n of T_n satisfies $\Delta_n / \lfloor \log_2 n \rfloor \to 1$ almost surely; Goh and Schmutz [7] showed distributional convergence of $\Delta_n - \lfloor \log_2 n \rfloor$ along suitable subsequences. In this work we show how a version of Kingman's coalescent can be used to access much finer properties of the degree distribution in T_n.

For any $i \in \mathbb{Z}$, let $X_i^{(n)} = |\{v \in [n] : \deg_{T_n}(v) = \lfloor \log n \rfloor + i\}|$. Also, let $\mathcal{P}$ be a Poisson point process on $\mathbb{R}$ with rate function $\lambda(x) = 2^{-x} \cdot \ln 2$. We show that, up to lattice effects, the vectors $(X_i^{(n)}, i \in \mathbb{Z})$ converge weakly in distribution to $(\mathcal{P}[i, i+1), i \in \mathbb{Z})$. We also prove asymptotic normality of $X_i^{(n)}$ when $i = i(n) \to -\infty$ slowly, and obtain precise asymptotics for $\mathbf{P}(\Delta_n - \log_2 n > i)$ when $i(n) \to \infty$ and $i(n)/\log n$ is not too large. Our results recover and extends the previous results on maximal and near-maximal degrees in random recursive trees.

1. Statement of results

The process of random recursive trees $(T_n, n \geq 1)$ is defined as follows. T_1 has a single node with label 1, which its root. The tree T_{n+1} is obtained from T_n by directing an edge from a new vertex $n+1$ to $v \in [n]$; the choice of v is uniformly random and independent for each $n \in \mathbb{N}$. We call T_n a random recursive tree (RRT) of size n.

As a consequence of the construction, vertex-labels in T_n increase along root-to-leaf paths. Rooted labelled trees with such property are called *increasing trees*. It is not difficult to see that, in fact, T_n is uniformly chosen among the set $\mathcal{I}_n$ of increasing trees with vertex set $[n]$.

We write $\deg_{T_n}(v)$ to denote the number of children of v in T_n. The degree distribution of T_n is encoded by the variables $Z_i^{(n)} = |\{v \in [n] : \deg_{T_n}(v) = i\}|$, for $i \geq 0$. In fact, the study of RRT's started with a paper by Na and Rapoport [12] in which they obtained, for any *fixed* $i \geq 0$, the convergence $\mathbb{E}(Z_i^{(n)})/n \to 2^{-i-1}$ as $n \to \infty$. Mahmoud and Smythe [11] derived the asymptotic joint normality of $Z_i^{(n)}$ for $i \in \{0, 1, 2\}$; Janson [8] extended the joint normality to $Z_i^{(n)}$ for $i \geq 0$ and gave explicit formulae for the covariance matrix (this is not an exhaustive account of the results concerning the random variables $Z_i^{(n)}$).

The above results concern typical degrees; the focus in this work is large degree vertices, and in particular the maximum degree in T_n, which we denote $\Delta_n = \max_{v \in [n]} \deg_{T_n}(v)$. For the rest of the paper we write log to denote logarithms with base 2, and ln to denote natural logarithms. For $n \in \mathbb{N}$ let $\varepsilon_n = \log n - \lfloor \log n \rfloor$.

A heuristic to find the order of Δ_n is that, if $\mathbb{E}(Z_i^{(n)}) \approx n 2^{-i-1}$ were to hold for all i, as it does when i is fixed, then we would have $\mathbb{E}(Z_{\lfloor \log n \rfloor}^{(n)}) \approx n 2^{-\lfloor \log n \rfloor - 1} = 2^{-1+\varepsilon_n}$.

Date: July 21, 2015.

2010 *Mathematics Subject Classification.* 60C05, 05C80.

This heuristic suggests that Δ_n is of order $\log n$. This is indeed the case: Szymanski [14] proved that $\mathbf{E}[\Delta_n]/\log n \to 1$ as $n \to \infty$, and Devroye and Lu [6] later established that $\Delta_n/\log n \to 1$ a.s.. Finally, Goh and Schmutz [7] showed that $\Delta_n - \lfloor \log n \rfloor$ converges in distribution along suitable subsequences, and identified the possible limiting laws.

Since we focus on maximal degrees, it is useful to let

$$X_i^{(n)} = Z_{i+\lfloor \log n \rfloor}^{(n)} = |\{v \in [n] : \deg_{T_n}(v) = \lfloor \log n \rfloor + i\}|,$$

for $n \in \mathbb{N}$ and $i \geq -\lfloor \log n \rfloor$. The following is a simplified version of one of our main results.

Theorem 1.1. *Fix $\varepsilon \in [0,1]$. Let $(n_l)_{l \geq 1}$ be an increasing sequence of integers satisfying $\varepsilon_{n_l} \to \varepsilon$ as $l \to \infty$. Then, as $l \to \infty$*

$$(X_i^{(n_l)},\, i \in \mathbb{Z}) \xrightarrow{\mathrm{d}} (P_i^{\varepsilon},\, i \in \mathbb{Z})$$

jointly for all $i \in \mathbb{Z}$ where the P_i^{ε} are independent Poisson r.v.'s with mean $2^{-i-1+\varepsilon}$.

The random variables $X_i^{(n)}$ do not converge in distribution as $n \to \infty$ without taking subsequences; this is essentially a lattice effect caused by the floor $\lfloor \log n \rfloor$ in the definition of $X_i^{(n)}$.

Theorem 1.1 can be stated in terms of weak convergence of point processes (which is equivalent to convergence of finite dimensional distributions (FDD's); see Theorem 11.1.VII in [4]). In fact, we will also prove convergence (along subsequences) of

$$X_{\geq i}^{(n)} = \sum_{k \geq i} X_k^{(n)} = |\{v \in [n] : \deg_{T_n}(v) \geq \lfloor \log n \rfloor + i\}|.$$

This is useful as it yields information about Δ_n which cannot be derived from Theorem 1.1. We formulate this result as a statement about convergence of point processes, and now provide the relevant definitions. Let $\mathbb{Z}^* = \mathbb{Z} \cup \{\infty\}$. Endow $\mathbb{Z}^*$ with the metric defined by $\mathrm{d}(i,j) = |2^{-j} - 2^{-i}|$ and $\mathrm{d}(i,\infty) = 2^{-i}$ for $i, j \in \mathbb{Z}$. Let $\mathcal{M}_{\mathbb{Z}^*}^{\#}$ be the space of boundedly finite measures of $\mathbb{Z}^*$.

Let $\mathcal{P}$ be a Poisson point process on $\mathbb{R}$ with rate function $\lambda(x) = 2^{-x} \cdot \ln 2$. For each $\varepsilon \in [0,1]$ let $\mathcal{P}^{\varepsilon}$ be the point process on $\mathbb{Z}^*$ given by

$$\mathcal{P}^{\varepsilon} = \sum_{x \in \mathcal{P}} \delta_{\lfloor x + \varepsilon \rfloor}.$$

Similarly, for all $n \in \mathbb{N}$ let

$$\mathcal{P}^{(n)} = \sum_{v \in [n]} \delta_{\deg_{T_n}(v) - \lfloor \log n \rfloor}.$$

Then, for each $i \in \mathbb{Z}$ we have that

$$\mathcal{P}^{\varepsilon}(\{i\}) := |\{x \in \mathcal{P} : \lfloor x + \varepsilon \rfloor = i\}| = |\{x \in \mathcal{P} : x \in [i - \varepsilon, i + 1 - \varepsilon)\}|$$

has distribution $\mathrm{Poi}(2^{-i-1+\varepsilon})$; also $\mathcal{P}^{(n)}(\{i\}) = X_i^{(n)}$. We abuse notation by writing, e.g., $\mathcal{P}^{(n)}(i) = \mathcal{P}^{(n)}(\{i\})$.

It is clear that $\mathcal{P}^{(n)}$ and $\mathcal{P}^{\varepsilon}$ are elements of $\mathcal{M}_{\mathbb{Z}^*}^{\#}$. The advantage of working on the state space to $\mathbb{Z}^*$ is that intervals $[k, \infty]$ are compact. In particular, the convergence of FDD's of $\mathcal{P}^{(n_l)}$ implies the convergence in distribution of $X_{\geq i}^{(n_l)} = \mathcal{P}^{(n_l)}[i, \infty)$.

Theorem 1.2. *Fix $\varepsilon \in [0,1]$. Let $(n_l)_{l\geq 1}$ be an increasing sequence of integers satisfying $\varepsilon_{n_l} \to \varepsilon$ as $l \to \infty$. Then in $\mathcal{M}^{\#}_{\mathbb{Z}^*}$, $\mathcal{P}^{(n_l)}$ converges weakly to $\mathcal{P}^{\varepsilon}$ as $l \to \infty$. Equivalently, for any $i < i' \in \mathbb{Z}$, jointly as $l \to \infty$*

$$(X_i^{(n_l)}, \dots, X_{i'-1}^{(n_l)}, X_{\geq i}^{(n_l)}) \xrightarrow{\mathrm{d}} (\mathcal{P}^{\varepsilon}(i), \dots, \mathcal{P}^{\varepsilon}(i'-1), \mathcal{P}^{\varepsilon}[i', \infty)).$$

Note that Theorem 1.1 follows from Theorem 1.2. We finish this section stating two additional results. The first is an extension of the main theorem from [7], that result being essentially the case $i = O(1)$.

Theorem 1.3. *For any $i = i(n)$ with $i + \log n < 2 \ln n$ and $\liminf_{n\to\infty} i(n) > -\infty$,*

$$\mathbf{P}(\Delta_n \geq \lfloor \log n \rfloor + i) = (1 - \exp\{-2^{-i+\varepsilon_n}\})(1 + o(1)).$$

When $i = O(1)$, the assertion of Theorem 1.3 is a straight-forward consequence of Theorem 1.2. For the case that $i(n) \to \infty$ we use estimates for the first and second moments of $X_{\geq i}^{(n)}$; note that $\{\Delta_n < \lfloor \log n \rfloor + i\} = \{X_{\geq i}^{(n)} = 0\}$.

Finally, we also obtain the asymptotic normality for $X_i^{(n)}$ when i tends to $-\infty$ slowly enough.

Theorem 1.4. *If $i = i(n) \to -\infty$ and $i = o(\ln n)$, then as $n \to \infty$*

$$\frac{X_i^{(n)} - 2^{-i-1+\varepsilon_n}}{\sqrt{2^{-i-1+\varepsilon_n}}} \xrightarrow{\mathrm{d}} N(0,1).$$

Remark 1.5. *Up to lattice effects, Theorems 1.2 and 1.4 extend the range of $i = i(n)$ for which the heuristic that $Z_i^{(n)} \approx n2^{-i-1}$ holds.*

A key novelty of our approach is that for each n we use *Kingman's coalescent* to generate a tree $T^{(n)}$ whose vertex degrees $\{\deg_{T^{(n)}}(v)\}_{v\in[n]}$ are exchangeable but otherwise have the same law as degrees in T_n. (See [2], Chapter 2 for a description of Kingman's coalescent, and [1], Section 2.2 for a description of the connection with random recursive trees which we exploit in this paper.) By this we mean that if $\sigma : [n] \to [n]$ is a uniformly random permutation then the following distributional identiy holds:

$$(\deg_{T^{(n)}}(v),\ v \in [n]) \stackrel{\mathrm{d}}{=} (\deg_{T_n}(\sigma(v)),\ v \in [n]). \tag{1}$$

We describe the trees $T^{(n)}$, $n \in \mathbb{N}$ in Section 3.

An essentially equivalent construction was used by Devroye [5] to study union-find trees. In [13], Pittel related the results of [5] on union-find trees to the height of RRT's. It is worth mentioning that both Kingman's coalescent and the union-find trees can be equivalently represented as binary trees or, as we will see in Section 3, as RRT's. Aside from the works [5] and [13], it seems that the use of Kingman's coalescent or of union-find trees to study RRT's is rare. However, it turns out to provide just the right perspective for studying high degree vertices.

2. Outline

In this section we sketch the approach used in the paper. The proofs of the theorems relay on the computation of the moments of the FDD's of $\mathcal{P}^{(n)}$; these estimates are given in Proposition 2.1. In particular, the proofs of Theorems 1.2 and 1.4 use the method of moments (e.g., see [9] Section 6.1, and [3] Section 1.5).

Any FDD of $\mathcal{P}^{(n)}$ can be recovered from suitable marginals of the joint distribution of $(X_i^{(n_l)}, \dots, X_{i'-1}^{(n_l)}, X_{\geq i'}^{(n_l)})$ for some $i < i' \in \mathbb{Z}$. For simplicity, we focus for the moment

on collections of variables $X_i^{(n)}, \dots, X_{i'}^{(n)}$ for $i \le i'$. For $r \in \mathbb{R}$ and $a \in \mathbb{N}$ write $(r)_a = r(r-1)\cdots(r-a+1)$, also let $(r)_0 = 1$. We will prove that for any non-negative integers $a_i, \dots, a_{i'}$, as $n \to \infty$, we have

$$\mathbf{E}\left[\prod_{i\le k\le i'} (X_k^{(n)})_{a_k}\right] - \prod_{i\le k\le i'} \left(2^{-(k+1)+\varepsilon_n}\right)^{a_k} \to 0. \tag{2}$$

This immediately yields Theorem 1.1.

By the linearity of expectation, proving (2) reduces to understanding the probabilities

$$\mathbf{P}\left(\deg_{T_n}(v_k) = \lfloor \log n \rfloor + i_k,\, k \in [K]\right) \tag{3}$$

for all $i_1, \dots i_K \in \mathbb{N}$ and $v_1, \dots v_K \in [n]$, $K \in \mathbb{N}$; see Section 6 for more details.

In the standard model for RRT's described at the beginning, $\deg_{T_n}(v)$ is a sum of Bernoulli variables:

$$\deg_{T_n}(v) = \sum_{v<u\le n} \mathbf{1}_{\{u\to v\}}.$$

The lack of symmetry of the degrees $\{\deg_{T_n}(v)\}_{v\in[n]}$ complicates the analysis of (3). In proving that $\Delta_n/\log n \overset{\text{a.s.}}{\to} 1$, Devroye and Lu [6] used that $\{\deg_{T_n}(v)\}_{v\in[n]}$ are negatively orthant dependent (see [10] for a definition), which in particular means that for all $S \subset [n]$ and $m_1, \dots, m_n \in \mathbb{N}$

$$\mathbf{P}\left(\deg_{T_n}(v) \ge m_v,\, v \in S\right) \le \prod_{v\in S} \mathbf{P}\left(\deg_{T_n}(v) \ge m_v\right) \tag{4}$$

and then obtained upper bounds for $\mathbf{P}\left(deg_{T_n}(v) \ge c \ln n\right)$ for each $v \in [n]$.

One approach to studying high degrees in T_n would be to obtain matching lower bounds for $\mathbf{P}\left(\deg_{T_n}(v) \ge m_v,\, v \in S\right)$, with uniform error terms even when m_v is large. Instead, we study trees $T^{(n)}$, mentioned in (1), above, for which we can obtain precise asymptotics for the analogous probabilities

$$\mathbf{P}\left(\deg_{T^{(n)}}(v) \ge m_v,\, v \in [K]\right). \tag{5}$$

The core of the paper lies in Proposition 4.2, which gives precise estimates of (5) for $m_1, \dots, m_K$ in a suitable range. Broadly speaking, $\deg_{T^{(n)}}(v)$ depends on a set of random *selection times* $\mathcal{S}_v$ and the first streak of heads in a sequence of $|\mathcal{S}_v|$ fair coin flips. As mentioned in the previous section, the degrees of $T^{(n)}$ have the same distribution as the degrees in T_n. Consequently, our estimation of (5) allows us to obtain the following moments estimate.

Proposition 2.1. *For all $c \in (0,2)$ and $K \in \mathbb{N}$ there is $\alpha = \alpha(c,K) > 0$ such that the following holds. Fix any integers i, i' with $0 < i + \log_n < i' + \log_n < c \ln n$. Then for any non-negative integers $a_i, \dots, a_{i'}$ with $a_i + \ldots + a_{i'} = K$, we have*

$$\mathbf{E}\left[(X_{\ge i'}^{(n)})_{a_{i'}} \prod_{i\le k<i'} (X_k^{(n)})_{a_k}\right] = \left(2^{-i'+\varepsilon_n}\right)^{a_{i'}} \prod_{i\le k<i'} \left(2^{-(k+1)+\varepsilon_n}\right)^{a_k} (1 + o(n^{-\alpha})).$$

Equipped with Proposition 2.1, the proofs of the theorems are straightforward. The rest of the paper is organized as follows. In Section 3, we explain how to define the trees $T^{(n)}$ using Kingman's coalescent and establish the distributional relation between $T^{(n)}$ and the RRT; see Corollary 3.4. In Section 4, we define the random sets $(\mathcal{S}_v, v \in T^{(n)})$ and explain their relation with degrees in $T^{(n)}$. The proof of Proposition 4.2, which is our estimate of

(5), is completed with the study of $|\mathcal{S}_v|$, which can be found in Section 5. Finally, the proof of Proposition 2.1 is given in Section 6 and the proof of Theorems 1.2-1.4 are in Section 7.

3. Random Recursive Trees and Kingman's coalescent

In this section we give a representation of Kingman's coalescent in terms of labelled forests, and relate it to RRT's. All trees in the remainder of the paper are rooted, and we write $r(t)$ for the root of tree t. By convention, edges of a tree are directed towards the root of the tree and we write uv to denote an edge directed from u to v. A forest f is a set of trees whose vertex sets are pairwise disjoint. The vertex set of a forest, denoted $V(f)$, is the union of the vertex sets of its trees. Similarly, $E(f)$ denotes the set of edges in the trees of f. For $n \geq 1$, let

$$\mathcal{F}_n = \{f : V(f) = [n]\}$$

be the set of forests with vertex set $[n]$.

A sequence $C = (f_1, \ldots, f_n)$ of elements of $\mathcal{F}_n$ is an n-chain if f_1 is the forest in $\mathcal{F}_n$ with n one-vertex trees and, for $1 \leq i < n$, f_{i+1} is obtained from f_i by adding a directed edge between the roots of some pair of trees in f_i. If $(f_1, \ldots, f_n)$ is an n-chain then for $1 \leq i \leq n$, the forest f_i consists of $n + 1 - i$ trees, and in this case we list its elements in increasing order of their smallest-labelled vertex as $t_1^{(i)}, \ldots, t_{n+1-i}^{(i)}$.

Definition 3.1. *Kingman's n-coalescent is the random n-chain $\mathbf{C} = (F_1, \ldots, F_n)$ built as follows. Independently for each $1 \leq i \leq n-1$ let $\{a_i, b_i\}$ be a random pair uniformly chosen from $\{\{a, b\} : 1 \leq a < b \leq n + 1 - i\}$ and let ξ_i be independent with* Bernoulli(1/2) *distribution.*

For $1 \leq i < n$, construct F_{i+1} from F_i as follows. If $\xi_i = 1$ then add an edge from $r(T_{b_i}^{(i)})$ to $r(T_{a_i}^{(i)})$ and if $\xi_i = 0$ then add an edge from $r(T_{a_i}^{(i)})$ to $r(T_{b_i}^{(i)})$. The forest F_{i+1} consists of the new tree and the remaining $n - 1 - i$ unaltered trees from F_i.

For an example of the process see Figure 1.

Lemma 3.2. *Let $\mathcal{CF}_n$ be the set of n-chains of elements in $\mathcal{F}_n$. Then $|\mathcal{CF}_n| = n!(n-1)!$ and Kingman's n-coalescent is a uniformly random element of $\mathcal{CF}_n$.*

Proof. Fix an n-chain $(f_1, \ldots, f_n) \in \mathcal{CF}_n$. Then

$$\mathbf{P}\left((F_1, \ldots, F_n) = (f_1, \ldots, f_n)\right) = \prod_{k=1}^{n-1} \mathbf{P}\left(F_{k+1} = f_{k+1} | F_j = f_j,\, 1 \leq j \leq k\right).$$

Among the $(n + 1 - k)(n - k)$ possible oriented edges between roots of f_k, there is exactly one whose addition yields f_{k+1}. It follows that the k-th term in the above product is $((n+1-k)(n-k))^{-1}$, so $\mathbf{P}\left((F_1, \ldots, F_n) = (f_1, \ldots, f_n)\right) = [n!(n-1)!]^{-1}$. The result follows since this expression does not depend on $(f_1, \ldots, f_n) \in \mathcal{CF}_n$. □

Recall that $\mathcal{I}_n$ is the set of increasing trees with vertex set $[n]$. It is not difficult to see that $|\mathcal{I}_n| = (n-1)!$ and that a RRT is a uniformly random element of $\mathcal{I}_n$.

There is a natural mapping ϕ between n-chains and increasing trees. Given an n-chain $C = (f_1, \ldots, f_n)$, write $t^{(n)} := t_1^{(n)}$ for the unique tree in f_n. Let $L_C^- : E(t^{(n)}) \to [n-1]$ be defined as follows. For each $e \in E(t^{(n)})$, let

$$L_C^-(e) = \max\{i \in [n-1] :\, e \notin E(t^{(i)})\}.$$

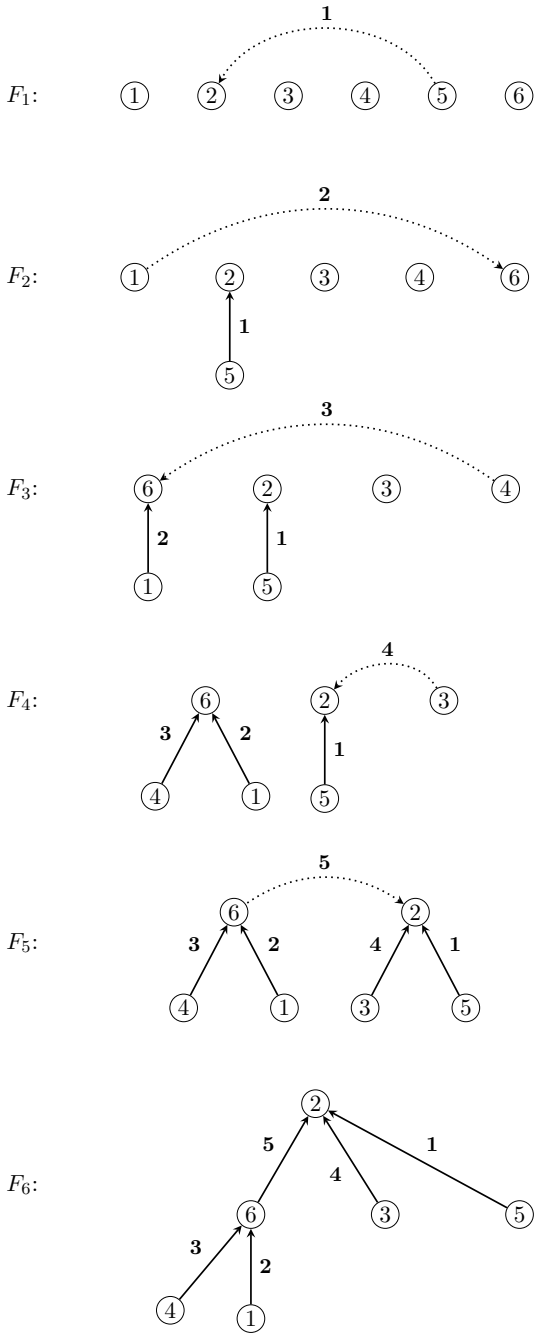

FIGURE 1. An example of Kingman's n-coalescent $\mathbf{C} = (F_1, \dots, F_n)$ for $n = 6$. For $1 \leq i < n$, F_i has, in dotted line, the edge in $E(F_{i+1}) \setminus E(F_i)$. Edges are marked with their time of addition; this is the function $L^-_{\mathbf{C}}$ defined after Lemma 3.2. In this instance, $\xi_1 = \xi_3 = \xi_4 = 1$, $\xi_2 = \xi_5 = 0$ and $\{a_1, b_1\} = \{2, 5\}$, $\{a_2, b_2\} = \{1, 5\}$, $\{a_3, b_3\} = \{1, 4\}$, $\{a_4, b_4\} = \{2, 3\}$, $\{a_5, b_5\} = \{1, 2\}$.

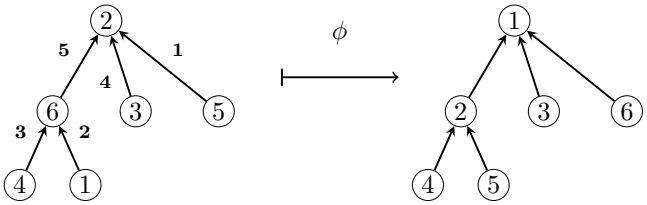

FIGURE 2. On the left a tree $t^{(n)}$; edges are marked with L_C^-, from which the n-chain $C = (f_1, \ldots, f_n)$ can be recovered. On the right, the increasing tree $\phi(f_1, \ldots, f_n)$; it has the shape of $t^{(n)}$ and the vertex labels $\{L_C(v),\, v \in V(t^{(n)})\}$.

We think of L_C^- as a function that keeps track of the *time of addition* of the edges along the n-chain C. Now, we define a vertex labelling $L_C : V(t^{(n)}) \to [n]$ as follows. Let $L_C(r(t^{(n)})) = 1$ and for each $uv \in E(t^{(n)})$, let

$$L_C(u) = n + 1 - L_C^-(uv);$$

then $L_C(u)$ is the number of trees in the forest just before uv is added.

Note that for each $i \in [n-1]$, the new edge in f_{i+1} joins the roots of two trees in f_i and is directed towards the root of the resulting tree. Thus, the labels $\{L_C^-(e),\, e \in E(t^{(n)})\}$ increase along all paths in $t^{(n)}$ towards the root $r(t^{(n)})$ and consequently, the labels $\{L_C(v),\, v \in V(t^{(n)})\}$ increase along root-to-leaf paths in $t^{(n)}$. This shows that relabelling the vertices of $t^{(n)}$ with L_C yields an increasing tree (specifically, an element of $\mathcal{I}_n$). See Figure 2 for an example.

Proposition 3.3. *Let $\phi : \mathcal{CF} \to \mathcal{I}_n$ be defined as follows. For an n-chain $C = (f_1, \ldots, f_n)$ let $\phi(C)$ be the tree obtained from $t^{(n)}$ by relabelling its vertices with L_C. Then $\phi(\mathbf{C})$, the push-forward of Kingman's n-coalescent by ϕ, has the law of a RRT of size n.*

Proof. First, we prove that ϕ is onto. Fix an increasing tree $t \in \mathcal{I}_n$. For each $j \in V(t) \setminus \{1\}$, let $v_j \in V(t)$ be such that $jv_j \in E(t)$, recall that edges are directed toward the root of t, thus v_j is uniquely defined. For each $1 < j \le n$, let $e_{n-j+1} = ju_j$.

Now construct an n-chain C as follows. Let f_1 be the forest with n one-vertex trees. For each $1 < i \le n$ construct f_i from f_{i-1} by adding the edge e_{i-1}. In other words, for each $1 \le i < n$, $L_C^-(e_i) = i$ and so $L_C(n+1-i) = n+1-L_C^-(e_i) = n+1-i$; also since $r(t) = 1$, we have $L_C(1) = 1$. Consequently, $\phi(C) = t$.

We claim that $|\phi^{-1}(t)| \ge n!$ for any $t \in \mathcal{I}_n$. To see this, consider an n-chain C and a permutation $\sigma : [n] \to [n]$. Let C_σ be the n-chain obtained from C by permuting the vertices in each forest of C by σ. Since $L_C(v)$ depends only on the time of addition of its outgoing edge (if any), it follows that $\phi(C) = \phi(C_\sigma)$ for all permutations σ. By Lemma 3.2, this shows that ϕ is $n!$-to-1 and that $\phi(\mathbf{C})$ is a uniform element in $\mathcal{I}_n$. □

Since $\phi(\mathbf{C})$ preserves the shape of $T^{(n)}$ and only relabels its vertices, the degrees in $T^{(n)}$ and $\phi(\mathbf{C})$ are equal as multisets: $\{deg_{T^{(n)}}(v)\}_{v\in[n]} = \{deg_{\phi(\mathbf{C})}(v)\}_{v\in[n]}$. This immediately gives the following key corollary of Proposition 3.3, on which the rest of the paper relies.

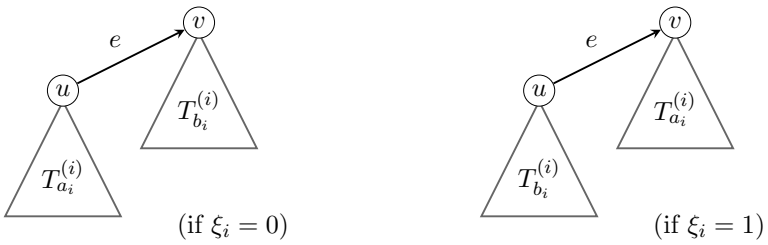

FIGURE 3. If v is a root in $T^{(i)}_{a_i} \cup T^{(i)}_{b_i}$ and ξ_i favours v, then v increases its degree and remains a root in F_{i+1}.

Corollary 3.4. *For all $n \in \mathbb{N}$, we have the following equality in distribution holds jointly for all $i \in \mathbb{Z}$,*

$$X_i^{(n)} \stackrel{\mathrm{d}}{=} |\{v \in [n] : \deg_{T^{(n)}}(v) = \lfloor \log n \rfloor + i\}|.$$

We now proceed to the study of the joint distribution of the vertex degrees in $T^{(n)}$.

4. Degree distribution: Selection sets and coin flips

By construction, the vertex degrees $\{\deg_{T^{(n)}}(v)\}_{v\in[n]}$ are exchangeable. Our next goal is to explain how to approximate (5); that is, for any fixed $k \in \mathbb{N}$ and integers $m_1, \dots, m_k < 2 \ln n$, to obtain estimates for $\mathbf{P}\left(\deg_{T^{(n)}}(v) \geq m_v,\ v \in [k]\right)$.

The key to analyse the degrees in $T^{(n)}$ is to understand how the degrees of a vertex $v \in [n]$ change in Kingman's coalescent $\mathbf{C} = (F_1, \dots, F_n)$. For any vertex v and $1 \leq i \leq n$, denote $\deg_{F_i}(v)$ the number of children of v in F_i. Also, we will simply write $\deg(v) = \deg_{F_n}(v) = \deg_{T^{(n)}}(v)$. For each $1 \leq i < n$, if $\xi_i = 1$ we say that ξ_i *favours* the vertices of $T^{(i)}_{a_i}$, and otherwise that it favours the vertices of $T^{(i)}_{b_i}$. For $v \in [n]$, let

$$\mathcal{S}_v = \{i \in [n-1] : v \in T^{(i)}_{a_i} \cup T^{(i)}_{b_i}\}.$$

For any vertex v, and $1 \leq i < n$, $\deg_{F_{i+1}}(v)$ increases by one only if v is a root in F_i, $i \in \mathcal{S}_v$ and ξ_i favours v; see Figure 3. Conversely, let $p_v = \min\{i \in \mathcal{S}_v,\ \xi_i \text{ does not favour } v\}$, then the first F_{i+1} in which v is not a root is exactly $i = p_v$. In this case, in F_{p_v+1} there is an outgoing edge from v, and v is not a root of any subsequent forests. As a consequence, $\deg_{F_j}(v) = \deg_{F_{p_v}}(v)$ for $p_v < j \leq n$.

Fact 4.1. *For $v \in [n]$, $\deg(v) = \deg_{F_{p_v}}(v) = |\mathcal{S}_v \cap [p_v - 1]|$.*

In other words, $\deg(v)$ depends only on its first streak of favourable random variables ξ_i with $i \in \mathcal{S}_v$. More precisely, given $|\mathcal{S}_v|$, the degree $\deg(v)$ is distributed as $\min\{|\mathcal{S}_v|, G\}$, where G is a Geo(1/2) r.v. independent of $\mathcal{S}_v$. We will see that, asymptotically, $\deg(v)$ has a geometric distribution. More strongly, for any fixed k, the following proposition shows that the variables $\{\deg_{T^{(n)}}(v)\}_{v\in[k]}$ are asymptotically independent, even when conditioned to be quite large.

Proposition 4.2. *Fix $c \in (0, 2)$ and $k \in \mathbb{N}$. There exists $\alpha = \alpha(c, k) > 0$ such that uniformly over positive integers $m_1, \dots, m_k < c \ln n$,*

$$\mathbf{P}\left(\deg_{T^{(n)}}(v) \geq m_v,\ v \in [k]\right) = 2^{-\sum_v m_v}(1 + o(n^{-\alpha})).$$

In this section we obtain the upper bound in Proposition 4.2 and explain what are the estimates needed to obtain the lower bound. In particular, we will require tail bounds for the random variables $|\mathcal{S}_v|$; this will be developed in Section 5.

Lemma 4.3. *For any $k \in \mathbb{N}$ and positive integers $m_1, \dots, m_k < n$,*

$$\mathbf{P}\left(\deg(v) \geq m_v,\, v \in [k]\right) \leq 2^{-\sum_v m_v}\, \mathbf{P}\left(|\mathcal{S}_v| \geq m_v,\, v \in [k]\right).$$

Equality holds for $k = 1$.

Proof. For each $v \in [k]$ list $\mathcal{S}_v$ in increasing order as $(i_{v,j},\, 1 \leq j \leq |\mathcal{S}_v|)$. Let $\mathcal{A}$ be the set of sequences $A = (A_1, \dots, A_k)$ satisfying $A_v \subset [n-1]$ and $|A_v| = m_v$ for all $v \in [k]$. For every $A \in \mathcal{A}$, let D_A be the event that $|\mathcal{S}_v| \geq m_v$ and $\{i_{v,1}, \dots, i_{v,m_v}\} = A_v$, for all $v \in [k]$. By Fact 4.1, if $\deg(v) \geq m_v$ then necessarily $|\mathcal{S}_v| \geq m_v$ so

$$\{\deg(v) \geq m_v,\, v \in [k]\} \cap D_A = \{\xi_{i_{v,j}} \text{ favours } v \text{ for all } j \in [m_v],\, v \in [k]\} \cap D_A.$$

Now, ξ_i are i.i.d Bernoulli(1/2) r.v.'s. Thus, if D_A has positive probability then

$$\mathbf{P}\left(\xi_{i_{v,j}} \text{ favours } v \text{ for all } j \in [m_v],\, v \in [k] | D_A\right) = \begin{cases} 2^{-\sum_v m_v} & \text{if } |A_u \cap A_v| = 0,\, \forall u \neq v \in [k] \\ 0 & \text{o.w.} \end{cases}$$

The second case follows from the fact that if $i \in \mathcal{S}_u \cap \mathcal{S}_v$ for some $u \neq v$, then ξ_i cannot favour both u and v. The events $(D_A,\, A \in \mathcal{A})$ are pairwise disjoint, and if $\deg(v) \geq m_v$ for all $v \in [k]$ then one of the events D_A must occur. It follows that

$$\begin{aligned} \mathbf{P}\left(\deg(v) \geq m_v,\, v \in [k]\right) &= \sum_{A \in \mathcal{A}} \mathbf{P}\left(D_A,\, \deg(v) \geq m_v,\, v \in [k]\right) \\ &\leq \sum_{A \in \mathcal{A}} 2^{-\sum_v m_v}\, \mathbf{P}\left(D_A\right) \\ &= 2^{-\sum_v m_v}\, \mathbf{P}\left(|S_v| \geq m_v,\, v \in [k]\right). \end{aligned}$$

The second line holds with equality when $k = 1$; we get the desired upper bounds. □

For the lower bounds we restrict to events D_A where the sets A_v are already disjoint. To do so, we consider instead the vertex degrees in F_I for some $I < n$. For $k \geq 2$ let

$$\tau_k = \min\{i \in [n-1] : \{a_i, b_i\} \subset [k]\}.$$

Since $F_i \subset F_j$ for all $i \leq j \in [n]$ we have that for any $I < n$

$$\begin{aligned} \mathbf{P}\left(\deg(v) \geq m_v,\, v \in [k]\right) &\geq \mathbf{P}\left(\deg_{F_{I+1}}(v) \geq m_v,\, v \in [k]\right) \\ &\geq \mathbf{P}\left(I < \tau_k,\, \deg_{F_{I+1}}(v) \geq m_v,\, v \in [k]\right). \end{aligned} \tag{6}$$

Recall that trees in F_i are listed in increasing order of their least elements; this implies that indices of the trees of vertices $1, \dots, k$ do not change until two trees indexed by $a, b \leq k$ are merged. Therefore, for all $v \in [k]$, $v \in T_v^{(i)}$ for $i \leq \tau_k$. This implies the sets $\{\mathcal{S}_v \cap [\tau_k - 1],\, v \in [k]\}$ are pairwise disjoint. These observations allow us to obtain a lower bound analogous to Lemma 4.3.

Lemma 4.4. *For any positive integers $k \geq 2$ and $m_1, \dots, m_k, I < n$,*

$$\mathbf{P}\left(\deg(v) \geq m_v,\, v \in [k]\right) \geq 2^{-\sum_v m_v} \mathbf{P}\left(I < \tau_k, |S_v \cap [I]| \geq m_v, v \in [k]\right).$$

Proof. By (6), it suffices to bound $\mathbf{P}\left(I < \tau_k,\, \deg_{F_{I+1}}(v) \geq m_v,\, v \in [k]\right)$.

Let $\mathcal{A}^*$ be the set of sequences $A = (A_1, \ldots, A_k)$ of pairwise disjoint subsets of $[I]$ satisfying $|A_v| = m_v$ for all $v \in [k]$. For each $A \in \mathcal{A}^*$, let D_A be the event that for all $v \in [k]$, $\{i_{v,j}, \ldots, i_{v,m_v}\} = A_v$ (and so $|\mathcal{S}_v \cap [I]| \geq m_v$).

As in the proof of Lemma 4.3, we have that

$$\{\deg_{F_{I+1}}(v) \geq m_v,\, v \in [k]\} \cap D_A = \{\xi_{i_{v,j}} \text{ favours } v \text{ for all } j \in [m_v],\, v \in [k]\} \cap D_A.$$

In this case, the sets A_v are pairwise disjoint. If $\mathbf{P}(D_A) > 0$ then

$$\mathbf{P}\left(\xi_{i_{v,j}} \text{ favours } v \text{ for all } j \in [m_v],\, v \in [k] | D_A\right) = 2^{-\sum_v m_v}.$$

Recall that $I < \tau_k$ if and only if the sets $\{S_v \cap [I],\, v \in [k]\}$ are pairwise disjoint; that is, if one of the events D_A occur. We then have

$$\begin{aligned}
\mathbf{P}\left(I < \tau_k,\, \deg_{F_{I+1}}(v) \geq m_v,\, v \in [k]\right) &= \sum_{A \in \mathcal{A}^*} \mathbf{P}\left(D_A,\, \deg_{F_{I+1}}(v) \geq m_v,\, v \in [k]\right) \\
&= \sum_{A \in \mathcal{A}^*} 2^{-\sum_v m_v}\, \mathbf{P}(D_A) \\
&= 2^{-\sum_v m_v}\, \mathbf{P}\left(I < \tau_k,\, |S_v \cap [I]| \geq m_v,\, v \in [k]\right). \qquad \square
\end{aligned}$$

To use Lemma 4.4 we need tail bounds for $|\mathcal{S}_v \cap [I]|$ for some suitable $I < n$; these are provided by the following proposition.

Proposition 4.5. *Let $c \in (0,2)$ and $\varepsilon \in (0,1)$ satisfy $c < 2(1-\varepsilon)$. Then there exists $\beta = \beta(c,\varepsilon) > 0$ such that for any vertex v,*

$$\mathbf{P}\left(|\mathcal{S}_v \cap [n - \lceil n^\varepsilon \rceil]| < c \ln n\right) = o(n^{-\beta}).$$

The proof of this proposition is given in the next section. The following lemma is the last ingredient for Proposition 4.2.

Lemma 4.6. *Fix an integer $k \geq 2$ and let $\varepsilon \in (0,1)$. Then, for n large enough,*

$$\mathbf{P}\left(\tau_k \leq n - \lceil n^\varepsilon \rceil\right) \leq \frac{2k^2}{\lceil n^\varepsilon \rceil - 1}.$$

Proof. By the definition of τ_k, if $\tau_k > n - \lceil n^\varepsilon \rceil$ then $\{a_i, b_i\} \not\subset [k]$ for all $1 \leq i \leq n - \lceil n^\varepsilon \rceil$. The events that $\{a_i, b_i\} \not\subset [k]$ are independent for distinct i and $\mathbf{P}\left(\{a_i, b_i\} \subset [k]\right) = \frac{k(k-1)}{(n+1-i)(n-i)}$, so we have that

$$\mathbf{P}\left(\tau_k > n - \lceil n^\varepsilon \rceil\right) = \prod_{i=1}^{n - \lceil n^\varepsilon \rceil} \left(1 - \frac{k(k-1)}{(n+1-i)(n-i)}\right) \geq 1 - \sum_{i=1}^{n - \lceil n^\varepsilon \rceil} \frac{2k^2}{(n-i)^2}$$

The last inequality holds for n large enough. Since $\sum_{j=m}^{\infty} j^{-2} \leq \int_{m-1}^{\infty} x^{-2} dx = (m-1)^{-1}$, we get

$$\mathbf{P}\left(\tau_k \leq n - \lceil n^\varepsilon \rceil\right) \leq \sum_{i=1}^{n - \lceil n^\varepsilon \rceil} \frac{2k^2}{(n-i)^2} \leq \sum_{j = \lceil n^\varepsilon \rceil}^{\infty} \frac{2k^2}{j^2} = \frac{2k^2}{\lceil n^\varepsilon \rceil - 1}. \qquad \square$$

We finish this section with the proof of Proposition 4.2 assuming Proposition 4.5.

Proof of Proposition 4.2. Fix $c \in (0,2)$, $k \in \mathbb{N}$ and let $m_1, \dots, m_k < c \ln n$ be positive integers. Let $\varepsilon = (2-c)/4$ so that Proposition 4.5 holds for some $\beta(c) = \beta(c,\varepsilon) > 0$. For $k = 1$, the result follows from the equality in Lemma 4.3 and Proposition 4.5 since

$$\mathbf{P}\left(|\mathcal{S}_1| < m_1\right) \le \mathbf{P}\left(|\mathcal{S}_1 \cap [n - \lceil n^{\varepsilon} \rceil]| < c \ln n\right) = o(n^{-\beta}).$$

For $k \ge 2$, the upper bound is likewise established immediately by Lemma 4.3. For the lower bound, letting $I = n - \lceil n^{\varepsilon} \rceil$, by Lemma 4.6 and Proposition 4.5 we have

$$\mathbf{P}\left(I < \tau_k, |\mathcal{S}_v \cap [I]| \ge m_v, v \in [k]\right) \ge 1 - \mathbf{P}\left(I \ge \tau_k\right) - \sum_{v \in [k]} \mathbf{P}\left(|\mathcal{S}_v \cap [I]| < m_v\right) \ge 1 - o(n^{-\alpha}),$$

where $\alpha = \min\{\beta, \varepsilon\}$. By Lemma 4.4, it follows that

$$\mathbf{P}\left(\deg(v) \ge m_v,\ v \in [k]\right) = 2^{-\sum_v m_v}(1 + o(n^{-\alpha})),$$

as required. □

We now proceed to the proof of Proposition 4.5.

5. Sizes of selection sets: A stick-breaking process

In this section we study the sizes of the selection sets $|\mathcal{S}_v|$, $v \in [n]$. More precisely we obtain tail bounds for $|\mathcal{S}_v \cap [n - \lceil n^{\varepsilon} \rceil]|$ for $\varepsilon \in [0,1)$. Recall that vertices in $T^{(n)}$ are exchangeable and so it suffices to consider $\mathcal{S}_1$. Also, trees in F_i are ordered in increasing order of their least elements. In particular, $1 \in V(T_1^{(i)})$ for all $1 \le i \le n$. Thus

$$\mathcal{S}_1 = \{i \in [n-1] : 1 \in \{a_i, b_i\}\};$$

recall that $\{a_i, b_i\}$ is a uniformly chosen pair of distinct elements of $[n+1-i]$.

Write $N = |\mathcal{S}_1|$ and list $\mathcal{S}_1 = \{S_1, S_2, \dots, S_N\}$ so that S_k are in increasing order. For $1 \le k \le N$, let $R_k = n - S_k$ be the number of remaining trees in F_{S_k+1}; for convenience we write $R_0 = n$, and let $S_k = n$ and $R_k = 0$ for $k > N$. For all $m, I \in [n]$ we have the equality of events

$$\{|\mathcal{S}_1 \cap [I]| < m\} = \{S_m > I\} = \{R_m < n - I\}. \tag{7}$$

Thus, to prove Proposition 4.5 it suffices to prove, that for $c \in (0,2)$ and $\varepsilon \in (0,1)$ satisfying $c < 2(1-\varepsilon)$, there exists $\beta = \beta(c,\varepsilon) > 0$ such that

$$\mathbf{P}\left(R_{\lfloor c \ln n \rfloor} < \lceil n^{\varepsilon} \rceil\right) = o(n^{-\beta}). \tag{8}$$

The proof of (8) occupies the remainder of the section.

For any $i \in [n-1]$, we have $\mathbf{P}\left(1 \in \{a_i, b_i\}\right) = \frac{2}{n+1-i}$. Consequently,

$$\mathbf{P}\left(S_1 = i\right) = \mathbf{P}\left(1 \in \{a_i, b_i\}\right) \prod_{j=1}^{i-1} \mathbf{P}\left(1 \notin \{a_j, b_j\}\right) = \frac{2}{n+1-i} \prod_{j=1}^{i-1} \frac{n-j-1}{n-j+1} = \frac{2(n-i)}{n(n-1)}.$$

The probability above can be expressed as

$$\mathbf{P}\left(R_1 = j\right) = \frac{2j}{n(n-1)} = \mathbf{P}\left(\max\{\lfloor nU \rfloor, \lfloor nU' \rfloor\} = j | \lfloor nU \rfloor \ne \lfloor nU' \rfloor\right)$$

where U, U' are independent uniform r.v.'s on $[0,1]$. Since $\mathbf{P}\left(\lfloor nU \rfloor = \lfloor nU' \rfloor\right) = n^{-1}$, for n large, $n^{-1}R_1$ can be approximated with

$$\max\{\lfloor nU \rfloor, \lfloor nU' \rfloor\} = \lfloor n \max\{U, U'\} \rfloor \sim n \max\{U, U'\};$$

recall that $\max\{U, U'\}$ is Beta$(2,1)$ distributed. Next, for all $1 \leq m \leq n$, the coalescence dynamics of $(F_{n-m+1}, \ldots, F_n)$ are precisely those of Kingman's m-coalescent. It follows that for all $1 \leq k \leq n-1$ and $1 \leq j < m \leq n-k$,

$$\mathbf{P}\left(R_{k+1} = j | R_k = m\right) = \frac{2j}{m(m-1)}\,. \tag{9}$$

Thus, provided R_k is large enough, $\frac{R_{k+1}}{R_k}$ is approximately Beta$(2,1)$-distributed.

As an aside, this argument shows that $\mathcal{S}_1$ is the result of a stick-breaking process that can be approximated with a rescaled Poisson-Dirichlet PD$(0,2)$ distribution (for the definition, see [2], Section 1.3).

We now formalize this heuristic. Let $(B_k, k \geq 1)$ be i.i.d. random variables with distribution Beta$(2,1)$. Write $Q_0 = n$ and for $k \in \mathbb{N}$, let $Q_{k+1} = \lfloor Q_k B_{k+1} \rfloor$.

Lemma 5.1. *Let $\varepsilon \in (0,1)$ and $1 \leq m < n$. Then*

$$\mathbf{P}\left(R_m < n^\varepsilon\right) \leq \mathbf{P}\left(Q_m < n^\varepsilon\right) + \frac{m}{\lceil n^\varepsilon \rceil}.$$

Proof. We prove the lemma by constructing an explicit couplling of $(Q_k,\, k \geq 0)$ and $(R_k,\, k \geq 0)$ as follows. Independently for each $0 \leq i \leq n$ let U_i, W_i and Y_i be independent random variables, with U_i and W_i taken to be Unif$[0,1]$-distributed, and with the distribution of Y_i given as follows. First, $\mathbf{P}\left(Y_i = 0\right) = 1$ for $i \in \{0,1\}$. Next for $i > 1$, Y_i is such that $\mathbf{P}\left(Y_i = j\right) = 2j/(i(i-1))$ for $i \in [j-1]$. Note that, conditional on $\lfloor iU_i \rfloor \neq \lfloor iW_i \rfloor$, $\max\{\lfloor iU_i \rfloor, \lfloor iW_i \rfloor\}$ is distributed as Y_i.

Let $Q_0 = n$, and for $k \geq 0$ let $Q_{k+1} = \max\{\lfloor Q_k U_{Q_k} \rfloor, \lfloor Q_k W_{Q_k} \rfloor\}$. Next, let $R_0 = n$, and for $k \geq 0$, given R_k,

$$R_{k+1} = \begin{cases} \max\{\lfloor R_k U_{R_k} \rfloor, \lfloor R_k W_{R_k} \rfloor\} & \text{if } \lfloor R_k U_{R_k} \rfloor \neq \lfloor R_k W_{R_k} \rfloor \\ Y_{R_k} & \text{otherwise} \end{cases}$$

It is straight-forward that $(Q_k, k \geq 0)$ has the desired distribution; for $(R_k, k \geq 0)$ this follows by (9).

Now,

$$\begin{aligned} \mathbf{P}\left(R_m < n^\varepsilon\right) &\leq \mathbf{P}\left(R_m \leq n^\varepsilon,\, Q_m < n^\varepsilon\right) + \mathbf{P}\left(R_m \leq n^\varepsilon,\, Q_m \geq n^\varepsilon\right) \\ &\leq \mathbf{P}\left(Q_m < n^\varepsilon\right) + \mathbf{P}\left(Q_m \geq n^\varepsilon,\, R_m \neq Q_m\right). \end{aligned}$$

If $R_m \neq Q_m$, then there is exactly one $k \leq m$ for which $R_k \neq Q_k$ and $R_l = Q_l$ for all $1 \leq l < k$. For such k we then have $\lfloor Q_{k-1} U_{Q_{k-1}} \rfloor = \lfloor Q_{k-1} W_{Q_{k-1}} \rfloor$. Furthermore, if $Q_m \geq n^\varepsilon$, then $Q_{k-1} \geq n^\varepsilon$. Thus,

$$\begin{aligned} \mathbf{P}\left(Q_m \geq n^\varepsilon,\, R_m \neq Q_m\right) &\leq \sum_{k=1}^{m} \mathbf{P}\left(Q_{k-1} \geq n^\varepsilon,\, R_k \neq Q_k,\, R_l = Q_l,\, 1 \leq l < k\right) \\ &\leq \sum_{k=1}^{m} \mathbf{P}\left(Q_{k-1} \geq n^\varepsilon,\, \lfloor Q_{k-1} U_{Q_{k-1}} \rfloor = \lfloor Q_{k-1} W_{Q_{k-1}} \rfloor\right) \\ &\leq \sum_{k=1}^{m} \mathbf{P}\left(\lfloor Q_{k-1} U_{Q_{k-1}} \rfloor = \lfloor Q_{k-1} W_{Q_{k-1}} \rfloor | Q_{k-1} \geq n^\varepsilon\right) \\ &\leq \frac{m}{\lceil n^\varepsilon \rceil}. \end{aligned}$$

The result follows. □

Now, it remains to obtain tail bounds for Q_m; these are obtained by a standard Chernoff bound on the variables $X_k = -\ln B_k$.

Proposition 5.2. *Let $c \in (0,2)$ and $\varepsilon \in (0,1)$ satisfy $c < 2(1-\varepsilon)$. Then there exists $\gamma = \gamma(c,\varepsilon) > 0$ such that $\mathbf{P}\left(Q_{\lfloor c\ln n\rfloor} < n^{\varepsilon}\right) \le n^{-\gamma}$.*

Proof. Write $m = \lfloor c\ln n\rfloor$ and let $\delta > 0$ satisfy $c < 2\delta < 2(1-\varepsilon)$. Note that $Q_1 = \lfloor nB_1\rfloor \ge nB_1 - 1$; by induction it follows easily that $Q_m \ge nB_1\cdots B_m - m$. Thus, for n large enough, $Q_m < n^{\varepsilon}$ implies

$$nB_1B_2\cdots B_m \le n^{\varepsilon} + m < n^{1-\delta}.$$

Consequently,

$$\mathbf{P}\left(Q_m < n^{\varepsilon}\right) \le \mathbf{P}\left(B_1B_2\cdots B_m < n^{-\delta}\right).$$

Let $X_k = -\ln B_k$ for $k \in \mathbb{N}$. Then

$$\mathbf{P}\left(B_1B_2\cdots B_m < n^{-\delta}\right) = \mathbf{P}\left(\sum_{k\le m} X_k > \delta\ln n\right).$$

Since $\mathbf{E}[X_1] = \frac{1}{2}$ and $\delta > c/2$, we have $\delta\ln n > \mathbf{E}\left[\sum_{k\le m} X_k\right]$. Also, for $\lambda < 2$

$$\mathbf{E}\left[e^{\lambda X_1}\right] = \mathbf{E}\left[B_1^{-\lambda}\right] = \frac{2}{2-\lambda};$$

therefore,

$$\mathbf{P}\left(\sum_{k\le m} X_k > \delta\ln n\right) \le \min_{\lambda\in(0,2)} \frac{\mathbf{E}\left[B_1^{-\lambda}\right]^m}{n^{\lambda\delta}} = \min_{\lambda\in(0,2)} n^{-\lambda\delta + c\ln(\frac{2}{2-\lambda})}.$$

The function $f(\lambda) = -\lambda\delta + c\ln(\frac{2}{2-\lambda})$ is convex in $(0,2)$ and the minimum is achieved at $\lambda^* = (2 - \frac{c}{\delta})$. Consequently, for n large enough,

$$\mathbf{P}\left(Q_m < n^{\varepsilon}\right) \le \mathbf{P}\left(\sum_{k\le m} X_k > \delta\ln n\right) \le n^{-(2\delta - c\ln(2\delta e/c))}.$$

It is straightforward that for $c \in (0,2)$ fixed, $\gamma(c,\delta) = 2\delta - c\ln(2\delta e/c)$ is, indeed, positive for $c/2 < \delta < 1$. Since δ can be chosen as a function of c and ε, the result follows. □

Proof of Proposition 4.5. Let $c \in (0,2)$ and $\varepsilon \in (0,1)$ satisfy $c < 2(1-\varepsilon)$. Writing $m = \lfloor c\ln n\rfloor$, (7) and Lemma 5.1 gives

$$\mathbf{P}\left(|\mathcal{S}_1 \cap [n - \lceil n^{\varepsilon}\rceil]| < c\ln n\right) \le \mathbf{P}\left(R_m \le \lceil n^{\varepsilon}\rceil\right) \le 2\mathbf{P}\left(Q_m \le \lceil n^{\varepsilon}\rceil\right) + \frac{m}{n^{\varepsilon}},$$

the left-hand side is $O(n^{-\beta})$ by Proposition 5.2 where we take $\beta = \min\{\varepsilon,\gamma\}$. The result follows. □

6. Proof of Proposition 2.1

By Corollary 3.4 we can study vertex degrees in $T^{(n)}$ and derive conclusions about the variables $X_i^{(n)}, X_{\ge i}^{(n)}$, $i \in \mathbb{Z}$. Recall that we write $\deg(v) = \deg_{T^{(n)}}(v)$, for $v \in [n]$.

Lemma 6.1. *For any $k \in \mathbb{N}$ and integers $m_1, \dots, m_k$,*

$$\mathbf{P}\left(\deg(u) = m_u,\, u \in [k]\right) = \sum_{j=0}^{k} \sum_{\substack{S \subset [k] \\ |S|=j}} (-1)^j \mathbf{P}\left(\deg(u) \geq m_u + \mathbf{1}_{[u \in S]},\, u \in [k]\right).$$

Furthermore, for $k' \in \mathbb{N}$ and integers $m_{k+1}, \dots m_{k+k'}$,

$$\begin{aligned}&\mathbf{P}\left(\deg(u) = m_u,\, \deg(v) \geq m_v,\, 1 \leq u \leq k < v \leq k+k'\right)\\ =&\sum_{j=0}^{k} \sum_{\substack{S \subset [k] \\ |S|=j}} (-1)^j \mathbf{P}\left(\deg(v) \geq m_v + \mathbf{1}_{[v \in S]},\, v \in [k+k']\right).\end{aligned}$$

Proof. The second equation follows by intersecting the event $\{\deg(v) \geq m_v,\, k < v \leq k+k'\}$ along all probabilities in the first equation. The first is straightforwardly proved using the inclusion-exclusion principle. □

We are now ready to prove Proposition 2.1.

Proof of Proposition 2.1. Let $c \in (0,2)$ and $K \in \mathbb{N}$. Let $i < i'$ be integers such that $0 < i + \log_n < i' + \log_n < c \ln n$ and let a_j, $i \leq j \leq i'$ be non-negative integers with $a_i + \cdots a_{i'} = K$. We are interested in the factorial moments $\mathbf{E}\left[(X^{(n)}_{\geq i'})_{a_{i'}} \prod_{i \leq k < i'} (X^{(n)}_k)_{a_k}\right]$.

For $i \leq k \leq i'$, for each v with $\sum_{l=i}^{k-1} a_l < v \leq \sum_{l=i}^{k} a_l$ let $m_v = \lfloor \log n \rfloor + k$. Let $K' = K - a_{i'}$, by Corollary 3.4 and the exchangeability of the vertex degrees of $T^{(n)}$,

$$\begin{aligned}\mathbf{E}\left[(X^{(n)}_{\geq i'})_{a_{i'}} \prod_{i \leq k < i'} (X^{(n)}_k)_{a_k}\right] &= (n)_K \mathbf{P}\left(\deg(u) = m_u,\, \deg(v) \geq m_v,\, 1 \leq u \leq K' < v \leq K\right)\\ &= (n)_K \sum_{l=0}^{K'} \sum_{\substack{S \subset [K'] \\ |S|=l}} (-1)^l \mathbf{P}\left(\deg(v) \geq m_v + \mathbf{1}_{[v \in S]},\, v \in [K]\right),\end{aligned}$$

the last equality by Lemma 6.1. At this point we can apply Proposition 4.2 to each of the terms. Since $m_v \leq c \ln n$ for $v \in [K]$, there is $\alpha' = \alpha'(c, K) > 0$ such that

$$\begin{aligned}&\sum_{l=0}^{K'} \sum_{\substack{S \subset [K'] \\ |S|=l}} (-1)^l \mathbf{P}\left(\deg(v) \geq m_v + \mathbf{1}_{[v \in S]},\, v \in [K]\right)\\ =&\sum_{l=0}^{K'} \sum_{\substack{S \subset [K'] \\ |S|=l}} (-1)^l 2^{-l-\sum_v m_v}(1 + o(n^{-\alpha'}))\\ =&2^{-\sum_v m_v}(1 + o(n^{-\alpha'})) \sum_{l=0}^{K'} \sum_{\substack{S \subset [K'] \\ |S|=l}} (-1)^l 2^{-l}\\ =&2^{-K'-\sum_v m_v}(1 + o(n^{-\alpha'})).\end{aligned}$$

Using that $(n)_K = n^K(1+o(n^{-1}))$, we get

$$\mathbf{E}\left[(X^{(n)}_{\geq i'})_{a_{i'}} \prod_{i\leq k<i'} (X^{(n)}_k)_{a_k}\right] = 2^{K\log n - K' - \sum_{v=1}^K m_v}(1+o(n^{-\alpha}));$$

where $\alpha = \min\{\alpha', 1\}$. Finally, to complete the proof, note that

$$\begin{aligned} K\log n - K' - \sum_{v=1}^K m_v &= \sum_{v=K'+1}^K (\log n - m_v) + \sum_{v=1}^{K'} (\log n - 1 - m_v) \\ &= (-i' + \varepsilon_n)a_{i'} + \sum_{k=i}^{i'-1}(-k-1+\varepsilon_n)a_k. \end{aligned}$$

□

7. Proofs of the main theorems

Proof of Theorem 1.2. By Theorem 11.1.VII of [4], weak convergence in $\mathcal{M}^{\#}_{\mathbb{Z}^*}$ is equivalent to convergence of FDD's, that is, convergence of every finite family of bounded continuity sets; see Definition 11.1.IV of [4]. For any point process ξ on $\mathbb{Z}$ and any $i \in \mathbb{Z}$, we have that $\mathbb{Z} \cap [i, \infty)$ is a bounded stochastic continuity set for the underlying measure of ξ in $\mathcal{M}^{\#}_{\mathbb{Z}^*}$. Thus, any FDD of ξ can be recovered from suitable marginals of the joint distribution of $(\xi(i), \dots, \xi(i-1'), \xi[i,\infty))$ for some $i < i' \in \mathbb{Z}$.

Let $\varepsilon \in [0,1]$ and $(n_l)_{l\geq 1}$ be an increasing sequence with $\varepsilon_{n_l} \to \varepsilon$. The goal then is to prove that, for any integers $i < i'$, the joint distribution of

$$X^{(n_l)}_i, \dots, X^{(n_l)}_{i'-1}, X^{(n_l)}_{\geq i'}$$

converges to the joint distribution of

$$\mathcal{P}^{\varepsilon}(i), \dots, \mathcal{P}^{\varepsilon}(i'-1), \mathcal{P}^{\varepsilon}[i', \infty),$$

that is, to the law of independent Poisson r.v.'s with parameters $2^{-i-1+\varepsilon}, \dots, 2^{-i'-2+\varepsilon}, 2^{-i'+\varepsilon}$.

We compute the limit of the factorial moments of $X^{(n_l)}_i, \dots, X^{(n_l)}_{i'-1}, X^{(n_l)}_{\geq i'}$. For any non-negative integers $a_i, \dots, a_{i'}$, by Proposition 2.1,

$$\begin{aligned} \mathbf{E}\left[(X^{(n)}_{\geq i'})_{a_{i'}} \prod_{i\leq k<i'} (X^{(n)}_k)_{a_k}\right] &= \left(2^{-i'+\varepsilon_n}\right)^{a_{i'}} \prod_{i\leq k<i'} \left(2^{-(k+1)+\varepsilon_n}\right)^{a_k} (1+o(n^{-\alpha})) \\ &\to \left(2^{-i'+\varepsilon}\right)^{a_{i'}} \prod_{i\leq k<i'} \left(2^{-(k+1)+\varepsilon}\right)^{a_k}, \end{aligned}$$

as $n_l \to \infty$. The limit correspond to the factorial moment

$$\mathbf{E}\left[(\mathcal{P}^{\varepsilon}[i',\infty))_{a'_i} \prod_{i\leq k<i'} (\mathcal{P}^{\varepsilon}(k))_{a_k}\right].$$

The result follows (by, e.g. Theorem 6.10 of [9]). □

Proof of Theorem 1.3. Since $\{\Delta_n \geq \lfloor \log n \rfloor + i\} = \{X^{(n)}_{\geq i} > 0\}$, we need only to estimate $\mathbf{P}\left(X^{(n)}_{\geq i} > 0\right)$. If $i = O(1)$, then $\exp\{-2^{-i+\varepsilon_n}\} = O(1)$ and so it suffices to prove that

$$\mathbf{P}\left(X^{(n)}_{\geq i} = 0\right) - \exp\{-2^{-i+\varepsilon_n}\}) \to 0,$$

as $n \to \infty$. This follows from Theorem 1.2 and the subsubsequence principle. Suppose that there exists $\delta > 0$ and a subsequence n_k for which $|\mathbf{P}\left(X_{\geq i}^{(n_k)} = 0\right) - \exp\{-2^{-i+\varepsilon_{n_k}}\}| > \delta$. Since $\{\varepsilon_{n_k}\}_{k\geq 1}$ is a bounded set there is a subsubsequence n_{k_l} such that $\varepsilon_{n_{k_l}} \to \varepsilon$ for some $\varepsilon \in [0,1]$. By Theorem 1.2, $\mathbf{P}\left(X_{\geq i}^{(n_{k_l})} = 0\right) \to \exp\{-2^{-i+\varepsilon}\}$; this contradicts our assumption on the subsequence n_k.

Now consider the case $i \to \infty$ with $i + \log_n < 2\ln n$. By a standard inclusion-exclusion argument (see, e.g., [3] Corollary 1.11),

$$\mathbf{P}\left(X_{\geq i}^{(n)} = 0\right) = \sum_{r=0}^{n} (-1)^r \frac{\mathbf{E}\left[(X_{\geq i}^{(n)})_r\right]}{r!}, \tag{10}$$

and this sum has the so called *alternating inequalities* property; this means that partial sums alternatively serve as upper and lower bounds for $\mathbf{P}\left(X_{\geq i}^{(n)} = 0\right)$. Consequently [1],

$$\mathbf{E}\left[X_{\geq i}^{(n)}\right] - \frac{1}{2}\mathbf{E}\left[(X_{\geq i}^{(n)})_2\right] \leq \mathbf{P}\left(X_{\geq i}^{(n)} > 0\right) \leq \mathbf{E}\left[X_{\geq i}^{(n)}\right]. \tag{11}$$

Using Proposition 2.1 and the fact that $i \to \infty$, we have that $\mathbf{E}\left[X_{\geq i}^{(n)}\right] = 2^{-i+\varepsilon_n}(1+o(1))$ and

$$\mathbf{E}\left[X_{\geq i}^{(n)}\right] - \frac{1}{2}\mathbf{E}\left[(X_{\geq i}^{(n)})_2\right] = 2^{-i+\varepsilon_n}(1+o(1)) = (1-\exp\{-2^{-i+\varepsilon_n}\})(1+o(1)).$$

The result follows. □

Proof of Theorem 1.4. We again use the method of moments. By Theorem 1.24 of [3], it suffices to prove that, as $n \to \infty$

$$\mathbf{E}\left[(X_i^{(n)})_a\right] - (2^{-i-1+\varepsilon_n})^a = o(2^{-(i+1-\varepsilon_n)b}), \tag{12}$$

for all fixed $1 \leq a \leq b$. Since $i = o(\ln n)$, we have that $2^{-i-1+\varepsilon_n} = n^{o(1)}$. On the other hand, by Proposition 2.1 there is $\alpha > 0$ such that

$$\mathbf{E}\left[(X_i^{(n)})_a\right] - (2^{-i-1+\varepsilon_n})^a = o(n^{-\alpha}2^{-(i+\varepsilon_n)a}) = n^{-\alpha+o(1)} = o(n^{o(1)}).$$

Therefore, condition (12) is satisfied and the proof is complete. □

References

[1] Louigi Addario-Berry. Partition functions of discrete coalescents: from Cayley's formula to Frieze's $\zeta(3)$ limit theorem. In *XI Symposium on Probability and Stochastic Processes*, volume 68 of *Progress in Probability*, Basel, 2015. Birkhauser.

[2] Nathanaël Berestycki. *Recent progress in coalescent theory*, volume 16 of *Ensaios Matemáticos [Mathematical Surveys]*. Sociedade Brasileira de Matemática, Rio de Janeiro, 2009. ISBN 978-85-85818-40-1.

[3] Béla Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, second edition, 2001. ISBN 0-521-80920-7; 0-521-79722-5. doi: 10.1017/CBO9780511814068. URL http://dx.doi.org/10.1017/CBO9780511814068.

[1] A similar lower bound for $\mathbf{P}\left(X_{\geq i}^{(n)} > 0\right)$ could be obtained from Paley-Zigmund's inequality.

[4] D. J. Daley and D. Vere-Jones. *An introduction to the theory of point processes. Vol. II.* Probability and its Applications (New York). Springer, New York, second edition, 2008. ISBN 978-0-387-21337-8. doi: 10.1007/978-0-387-49835-5. URL http://dx.doi.org/10.1007/978-0-387-49835-5. General theory and structure.

[5] L. Devroye. Branching processes in the analysis of the heights of trees. *Acta Inform.*, 24(3):277–298, 1987. ISSN 0001-5903. doi: 10.1007/BF00265991. URL http://dx.doi.org/10.1007/BF00265991.

[6] Luc Devroye and Jiang Lu. The strong convergence of maximal degrees in uniform random recursive trees and dags. *Random Structures Algorithms*, 7(1):1–14, 1995. ISSN 1042-9832. doi: 10.1002/rsa.3240070102. URL http://dx.doi.org/10.1002/rsa.3240070102.

[7] William Goh and Eric Schmutz. Limit distribution for the maximum degree of a random recursive tree. *J. Comput. Appl. Math.*, 142(1):61–82, 2002. ISSN 0377-0427. doi: 10.1016/S0377-0427(01)00460-5. URL http://dx.doi.org/10.1016/S0377-0427(01)00460-5. Probabilistic methods in combinatorics and combinatorial optimization.

[8] Svante Janson. Asymptotic degree distribution in random recursive trees. *Random Structures Algorithms*, 26(1-2):69–83, 2005. ISSN 1042-9832. doi: 10.1002/rsa.20046. URL http://dx.doi.org/10.1002/rsa.20046.

[9] Svante Janson, Tomasz Łuczak, and Andrzej Rucinski. *Random graphs*. Wiley-Interscience Series in Discrete Mathematics and Optimization. Wiley-Interscience, New York, 2000. ISBN 0-471-17541-2. doi: 10.1002/9781118032718. URL http://dx.doi.org/10.1002/9781118032718.

[10] Kumar Joag-Dev and Frank Proschan. Negative association of random variables, with applications. *Ann. Statist.*, 11(1):286–295, 1983. ISSN 0090-5364. doi: 10.1214/aos/1176346079. URL http://dx.doi.org/10.1214/aos/1176346079.

[11] Hosam M. Mahmoud and R. T. Smythe. Asymptotic joint normality of outdegrees of nodes in random recursive trees. *Random Structures Algorithms*, 3(3):255–266, 1992. ISSN 1042-9832. doi: 10.1002/rsa.3240030305. URL http://dx.doi.org/10.1002/rsa.3240030305.

[12] Hwa Sung Na and Anatol Rapoport. Distribution of nodes of a tree by degree. *Math. Biosci.*, 6:313–329, 1970. ISSN 0025-5564.

[13] Boris Pittel. Note on the heights of random recursive trees and random m-ary search trees. *Random Structures Algorithms*, 5(2):337–347, 1994. ISSN 1042-9832. doi: 10.1002/rsa.3240050207. URL http://dx.doi.org/10.1002/rsa.3240050207.

[14] Jerzy Szymański. On the maximum degree and the height of a random recursive tree. In *Random graphs '87 (Poznań, 1987)*, pages 313–324. Wiley, Chichester, 1990.

E-mail address: louigi@problab.ca

E-mail address: laura.eslavafernandez@mail.mcgill.ca

Department of Mathematics and Statistics, McGill University, Montreal, Canada