Canad. Math. Bull. Vol. **64** (3), 2021, pp. 605–623
http://dx.doi.org/10.4153/S0008439520000685

Small trees in supercritical random forests

Tao Lei

Abstract. We study the scaling limit of a random forest with prescribed degree sequence in the regime that the largest tree consists of all but a vanishing fraction of nodes. We give a description of the limit of the forest consisting of the small trees, by relating a plane forest to a marked cyclic forest and its corresponding skip-free walk.

1 Introduction

A plane tree is a finite rooted tree in which the children of each node are ordered. A plane forest is a finite sequence of plane trees.

Fix a plane tree T and a plane forest $F = (T_1, \dots, T_c)$. The node set of F is $v(F) = \bigsqcup_{i \le c} v(T_i)$, where $v(T_i)$ is the node set of T_i and where $\bigsqcup$ denotes disjoint union. For a node $v \in v(T)$, by the *degree* of v, we mean the number of children of v in T. We denote this quantity $k_T(v)$. The degree of $v \in v(F)$, denoted $k_F(v)$, is its degree in its tree, so if $v \in v(T_i)$, then $k_F(v) = k_{T_i}(v)$. For F, we let $F^{\downarrow}$ be the sequence of reordering $\{T_1, \dots, T_c\}$ in decreasing order of number of nodes, breaking ties by the original order of appearance in F.

For $i \ge 0$, let $s^i(T) = \#\{v \in v(T) : k_T(v) = i\}$ and define $s^i(F)$ accordingly, so $s^i(F) = \sum_{j \le c} s^i(T_j)$. The *degree sequences* of T and of F are $s(T) = (s^i(T), i \ge 0)$ and $s(F) = (s^i(F), i \ge 0)$, respectively. Any sequence $s = (s^i, i \ge 0)$ of non-negative integers with $\sum_{i \ge 0} s^i < \infty$ and with $\sum_{i \ge 0} is^i < \sum_{i \ge 0} s^i$ is the degree sequence of some tree or forest. More precisely, writing $c(s) := \sum_{i \ge 0}(1 - i)s^i > 0$, then any forest with degree sequence s consists of exactly $c(s)$ trees.

The goal of this paper is to study the asymptotic structure of large random forests with a given degree sequence, in the *supercritical* finite variance regime. The supercritical regime is defined by $c_n = o(\sqrt{n})$, where c_n and n are the the number of tree components and vertices of the n-th forest, respectively. In this setting, the limiting forest typically consists of a single large tree containing all but a vanishing fraction of the nodes. The existence of this giant tree is the reason we call this regime supercritical. The scaling limit of this tree is $\mathcal{T}$, the Brownian Continuum Random Tree (CRT) introduced by Aldous in [2–4]. The remaining nodes form another random forest, which can be expected to have its own scaling limit (with an appropriate rescaling, which should be different from that of the large tree). The contributions of this paper confirm that the above picture is correct and yield a

Received by the editors January 23, 2018; revised March 30, 2020.
Published online on Cambridge Core September 29, 2020.
AMS subject classification: **60C05**, 05C05, 05C80.
Keywords: Random forests, continuum random trees, degree sequence, GHP convergence.

pleasingly straightforward description, which we now provide, for the joint scaling limit of the large tree and the small trees.

Let $B = (B(t), t \geq 0)$ be a linear Brownian motion. For $t \geq 0$, let $R(t) = B(t) - \inf(B(s), s \leq t)$; the process $R = (R(t), t \geq 0)$ is Brownian motion reflected at its running minimum. Let $Z = \{t \geq 0 : R(t) = 0\}$ be the zero set of R. By definition, this is also the set of times at which B is equal to its running minimum.

Now let $\tau(x) = \inf(t : B(t) \leq -x)$ for $x \geq 0$, and let $Z(x) = Z \cap [0, \tau(x)]$. For $\sigma > 0$, the relative complement $[0, \tau(\frac{1}{\sigma})] \setminus Z(\frac{1}{\sigma})$ is almost surely a countable collection of intervals with distinct lengths, and with total length $\tau(\frac{1}{\sigma})$. List these intervals in decreasing order of length as $((g_i, d_i), i \geq 1)$.

For $i \geq 1$, let $\mathcal{T}_i$ be the continuum random tree coded by $2B_i$, where

$$B_i = \big(B(g_i + t) - B(g_i), 0 \leq t \leq d_i - g_i\big) = \big(R(g_i + t) - R(g_i), 0 \leq t \leq d_i - g_i\big);$$

this construction is described in more detail and generality in Section 1.1. Then the scaling limit of the small trees has the law of the sequence $\mathcal{F} = (\mathcal{T}_i^{\downarrow}, i \geq 1)$, which is a decreasing reordering of $(\mathcal{T}_i, i \geq 1)$ according to $(d_i - g_i, i \geq 1)$.

For any probability distribution $q = (q^{(i)}, i \geq 0)$ on $\mathbb{N}$, we let $\sigma^2(q) = \sum_{i \geq 0} i^2 q^{(i)}$.

Theorem 1.1 *Fix a sequence $p = (p^i, i \geq 0)$ with $\sum_{i\geq 0} p^i = 1 = \sum_{i \geq 0} i p^i$ and with $\sigma^2 := \sigma^2(p) \in (0, \infty)$. For each $n \geq 1$, let $s_n = (s_n^i, i \geq 0)$ be a degree sequence with $\sum_{i \geq 0} s_n^i = n$ and write $p_n = (p_n^i, i \geq 0) = (s_n^i/n, i \geq 0)$ and $c_n = c(s_n)$.*

Let F_n be a uniformly random plane forest with degree sequence s_n. Let $\hat{F}_n = (T_{n,i}^{\downarrow}, 2 \leq i \leq c_n)$ be the decreasing reordering of F_n, excluding the largest tree $T_{n,1}^{\downarrow}$. Suppose that $p_n \to p$ in L^2 and $c_n = o(n^{1/2})$; then

$$\left(\frac{\sigma(p_n) T_{n,1}^{\downarrow}}{n^{1/2}}, \frac{\sigma(p_n)\hat{F}_n}{c_n}, \frac{n - |T_{n,1}^{\downarrow}|}{c_n^2} \right) \xrightarrow{d} \left(\mathcal{T}, \mathcal{F}, \tau\Big(\frac{1}{\sigma}\Big)\right),$$

where the first coordinate of the joint convergence is in the GHP sense, the second coordinate is in the sense of coordinatewise GHP convergence, and $\mathcal{T}$ and $\mathcal{F}$ are independent.

Remarks

- The condition that $\sum_{i\geq 0} s_n^i = n$ in Theorem 1.1 is for notational convenience; all proofs carry through with only cosmetic changes provided that $|s_n| = \sum_{i \geq 0} s_n^i \to \infty$ as $n \to \infty$, that $|s_n|^{-1} \cdot s_n \to p$ in L^2 and that $c_n = o(|s_n|^{1/2})$.
- Fix a critical, finite variance offspring distribution ν, and let $\mathcal{F}_n$ be a forest of c_n independent Galton–Watson (ν) trees with offspring distribution ν, conditioned to have total progeny n. It is not hard to check, as in [6], that with high probability the degree sequence of $\mathcal{F}_n$ satisfies the conditions of Theorem 1.1, so the distributional convergence of the theorem also applies to $\mathcal{F}_n$. The convergence of the third coordinate, in the Galton–Watson setting, appears as [15, Theorem 2.1.5], and provides a new proof and different perspective on that result; the convergence of the second coordinate strengthens and generalizes and removes a moment assumption from [7, Theorem 1.7].

The field of scaling limits of large random structures is motivated by the seminal papers [2–4] by Aldous, where he introduced the concept of the Brownian Continuum Random Tree and showed that a critical Galton–Watson tree conditioned on its size has the CRT as its limiting object. To be more specific, our work here is a natural generalization of [6] where it is shown that under natural hypotheses on the degree sequences, after suitable normalization, uniformly random trees with given degree sequence converge to CRT as sizes of trees tend to infinity. Let n be the number of nodes of the forest. In this paper, we work on uniformly random forests where the number of trees is $o(n^{1/2})$; a previous paper [12] addressed the asymptotic behaviour of such forests in the regime where the number of trees is of order $n^{1/2}$.

Outline of the Section In the remainder of this section, we first briefly introduce the concepts required to understand the statement of Theorem 1.1 rigorously. In Section 1.2, we describe the key ingredients of the proof of our main theorem. In Section 1.3, we explain how to deduce Theorem 1.1 from the results of Section 1.2 and outline the remaining sections of the article.

1.1 Concepts

Real Trees We briefly recall the concepts of real trees and real trees coded by continuous functions, which are necessary for understanding the construction of $\mathcal{F}$. A more lengthy presentation about the probabilistic aspects of real trees can be found in [10, 11].

Definition 1.2 A compact metric space (T, d) is a *real tree* if the following hold for every $a, b \in T$:

(i) There is a unique isometric map $f_{a,b} : [0, d(a, b)] \to T$ such that $f_{a,b}(0) = a$ and $f_{a,b}(d(a, b)) = b$.
(ii) If q is a continuous injective map from $[0, 1]$ into T, such that $q(0) = a$ and $q(1) = b$, we have $q([0, 1]) = f_{a,b}([0, d(a, b)])$.

Now we show a way of constructing real trees from continuous functions. Let $g : [0, \infty) \to [0, \infty)$ be a continuous function with compact support and such that $g(0) = 0$. For every $s, t \geq 0$, let

$$d_g^{\circ}(s, t) = g(s) + g(t) - 2m_g(s, t),$$

where

$$m_g(s, t) = \min_{s \wedge t \leq r \leq s \vee t} g(r).$$

The function d_g° is a pseudometric on $[0, \infty)$. Define an equivalence relation $\sim$ on $[0, \infty)$ by setting $s \sim t$ if and only if $d_g^{\circ}(s, t) = 0$. Then let $T_g = [0, \infty)/\sim$ and let d_g be the induced distance on T_g. Then (T_g, d_g) is a real tree (see, *e.g.*, [11, Theorem 2.2]).

GHP convergence Let (X,d) and (X',d') be compact metric spaces. Then the *Gromov-Hausdorff distance* between (X,d) and (X',d') is given by

$$d_{GH}((X,d),(X',d')) = \inf_{\phi,\phi',Z} d_H^Z\big(\phi(X),\phi'(X')\big),$$

where the infimum is taken over all isometric embeddings $\phi : X \hookrightarrow Z$ and $\phi' : X' \hookrightarrow Z$ into some common Polish metric space (Z,d^Z) and d_H^Z denotes the Hausdorff distance between compact subsets of Z, that is,

$$d_H^Z(A,B) = \inf\{\varepsilon > 0 : A \subset B^\varepsilon, B \subset A^\varepsilon\},$$

where A^ε is the ε-*enlargement* of A:

$$A^\varepsilon = \{z \in Z : \inf_{y\in A} d^Z(y,z) < \varepsilon\}.$$

Note that strictly speaking d_{GH} is not a distance since different compact metric spaces can have GH distance zero.

A *measured metric space* $\mathcal{X} = (X,d,\mu)$ is a metric space (X,d) with a finite Borel measure μ. Let $\mathcal{X} = (X,d,\mu)$ and $\mathcal{X}' = (X',d',\mu')$ be two compact measured metric spaces. They are *GHP-isometric* if there exists an isometric one-to-one map $\Phi : X \to X'$ such that $\Phi_*\mu = \mu'$ where $\Phi_*\mu$ is the *push forward* of measure μ to (X',d'), that is, $\Phi_*\mu(A) = \mu(\Phi^{-1}(A))$ for $A \in \mathcal{B}(X')$. In this case, call Φ a *GHP-isometry*. Suppose both $\mathcal{X}$ and $\mathcal{X}'$ are compact; then define the Gromov–Hausdorff–Prokhorov distance as:

$$d_{GHP}(\mathcal{X},\mathcal{X}') = \inf_{\Phi,\Phi',Z}\big(d_H^Z(\Phi(X),\Phi'(X')) + d_P^Z(\Phi_*\mu,\Phi'_*\mu')\big),$$

where the infimum is taken over all GHP-isometric embeddings $\Phi : X \hookrightarrow Z$ and $\Phi' : X' \hookrightarrow Z$ into some common Polish metric space (Z,d^Z), and d_P^Z denotes the Prokhorov distance between finite Borel measures on Z, that is,

$$d_P^Z(\mu,\nu) = \inf\{\varepsilon > 0 : \mu(A) \le \nu(A^\varepsilon) + \varepsilon, \nu(A) \le \mu(A^\varepsilon) + \varepsilon \text{ for any closed set } A\}.$$

Let $\mathbb{K}$ denote the set of GHP-isometry classes of compact measured metric spaces and we identify $\mathcal{X}$ with its GHP-isometry class.

Theorem 1.3 ([1, Theorem 2.5]) *The function d_{GHP} defines a metric on $\mathbb{K}$ and the space $(\mathbb{K}, d_{GHP})$ is a Polish metric space.*

We next define coordinatewise GHP convergence of sequences of measured metric spaces. For $\mathbf{X}_n = (\mathcal{X}_{n,j}, j \ge 1), \mathbf{X} = (\mathcal{X}_j, j \ge 1)$ in $\mathbb{K}^{\mathbb{N}}$, we say that $\mathbf{X}_n$ converges to $\mathbf{X}$ in *coordinatewise GHP sense* if for every $j \in \mathbb{N}$,

$$\sup_{1\le l\le j} d_{GHP}(\mathcal{X}_{n,l},\mathcal{X}_l) \to 0 \text{ as } n \to \infty.$$

Now to understand the statement of Theorem 1.1 in the rigorous way, we are viewing $T_{n,1}^{\downarrow}$ and each tree component of $\hat{F}_n$ as measured metric space where the distance is rescaled graph distance and the measure is the uniform measure putting mass $1/n$ on each node of $T_{n,1}^{\downarrow}$ and $\hat{F}_n$.

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

1.2 Functional Convergence and Proof of Theorem 1.1

Given a degree sequence $s = (s^i, i \geq 0)$ with $|s| = n$, we let $d(s) \in \mathbb{Z}^n_{\geq 0}$ be the vector whose entries are weakly increasing and with s^i entries equal to i, for each $i \geq 0$. For example, if $s = (3, 2, 0, 1, 0, 0, \dots)$ with $s^i = 0$ for $i \geq 4$, then $d(s) = (0, 0, 0, 1, 1, 3)$. Suppose we have a sequence of degree sequences $(s_n)_{n\in\mathbb{N}}$, with $s_n = (s^i_n, i \geq 0)$, $|s_n| = n$, $c_n := c(s_n) = o(n^{1/2})$, and $n^{-1} \cdot s_n \to p$ in L^2 for some distribution $p = (p^i, i \geq 0)$ with mean 1 and finite variance σ^2 on $\mathbb{N}$. Let $C_{n,1}, \dots, C_{n,n}$ be a uniformly random permutation of $d(s_n)$. For $1 \leq k \leq n$, let $X_{n,k} = C_{n,k} - 1$, and set $S_{n,k} = \sum_{j=1}^{k} X_{n,j}$. The sequence $(S_{n,k})_k$ is called *Lukasiewicz path* in the literature (*e.g.*, [11]). Our proof makes use of the following functional convergence theorem.

Theorem 1.4 *We have the following convergence of processes:*

$$\left(\frac{1}{c_n} S_{n,\lfloor tc_n^2 \rfloor}, t \geq 0\right) \xrightarrow{d} \left(\sigma B(t), t \geq 0\right) \tag{1.1}$$

where $(B(t), t \geq 0)$ is standard Brownian Motion.

Remark Technically $S_{n,\lfloor tc_n^2 \rfloor}$ is only defined for $t \leq n/c_n^2$. One can think of the process $(S_{n,k}, k \in \mathbb{N})$ in the theorem as the cumulative sum of

$$X^{(0)}_{n,1}, \dots, X^{(0)}_{n,n}, X^{(1)}_{n,1}, \dots, X^{(1)}_{n,n}, X^{(2)}_{n,1}, \dots, X^{(2)}_{n,n} \dots$$

where for each $j \in \mathbb{N}$, $X^{(j)}_{n,1}, \dots, X^{(j)}_{n,n}$ is an independent copy of $X_{n,1}, \dots, X_{n,n}$. Here, for the purpose of readability, we slightly abuse the notations. But this will not affect the proof later as $c_n^2 = o(n)$.

The proof of Theorem 1.4 will be given in Section 3. Theorem 1.4 will yield a description of the asymptotic behaviour of the sizes of all but the largest tree of F_n.

Corollary 1.5 *We have*

$$\left(\frac{|T^{\downarrow}_{n,i+1}|}{c_n^2}, i \geq 1\right) \xrightarrow{d} (g_i - d_i, i \geq 1) \text{ in } L^1,$$

where $((g_i, d_i), i \geq 1)$ are the excursion intervals of $(R(t), t \leq \tau(\frac{1}{\sigma}))$ in decreasing order of length.

Corollary 1.5 is equivalent to the assertions that

$$\frac{1}{c_n^2} \sum_{i\geq 2} |T^{\downarrow}_{n,i}| \xrightarrow{d} \tau\left(\frac{1}{\sigma}\right) = \sum_{i\geq 1} (g_i - d_i), \tag{1.2}$$

and that for any fixed $j \in \mathbb{N}$,

$$\left(\frac{|T^{\downarrow}_{n,2}|}{c_n^2}, \frac{|T^{\downarrow}_{n,3}|}{c_n^2}, \dots, \frac{|T^{\downarrow}_{n,j}|}{c_n^2}\right) \xrightarrow{d} (g_1 - d_1, g_2 - d_2, \dots, g_{j-1} - d_{j-1}). \tag{1.3}$$

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

We will prove this corollary in Section 3. To describe the limit structure of each tree, we appeal to the following theorem in [6].

Theorem 1.6 *Let $\{s_n, n \geq 1\}$ be a sequence of degree sequences such that $c(s_n) = 1, |s_n| = n$, and $\Delta_n := \max\{i : s_n^i \neq 0\} = o(n^{1/2})$ as $n \to \infty$. Suppose that there exists a distribution $p = (p^i, i \geq 0)$ on $\mathbb{N}$ with mean 1 such that $p_n = (s_n^i/n, i \geq 0)$ converges to p coordinatewise and such that $\sigma(p_n) \to \sigma(p) \in (0, \infty)$. Let $\mathbb{T}_n$ be the random plane tree under $\mathbb{P}_{s_n}$, the uniform measure on the set of plane trees with degree sequence s_n and let $d_{\mathbb{T}_n}$ be the graph distance in $\mathbb{T}_n$. Then when $n \to \infty$,*

$$\left(\mathbb{T}_n, \frac{\sigma(p_n)}{\sqrt{n}} d_{\mathbb{T}_n}\right) \xrightarrow{d} \mathcal{T}$$

in the GHP sense.

To apply Theorem 1.6 to each $T^{\downarrow}_{n,i}$, we also need to verify that the hypotheses of Theorem 1.1 imply those of Theorem 1.6. For fixed integers $i \geq 0$ and $l \geq 1$, let

$$p^i_{n,l} := \frac{|\{v \in T^{\downarrow}_{n,l} : k(v) = i\}|}{|T^{\downarrow}_{n,l}|} \text{ and } p_{n,l} = (p^i_{n,l}, i \geq 0).$$

In Section 4, we prove the following assertions:

$$\text{for any fixed } i \geq 0 \text{ and } l \geq 1,\ p^i_{n,l} - p^i_n \xrightarrow{p} 0, \text{ as } n \longrightarrow \infty, \tag{1.4}$$

and

$$\text{for any } l \geq 1,\ \sigma^2(p_{n,l}) - \sigma^2(p_n) \xrightarrow{p} 0, \text{ as } n \longrightarrow \infty. \tag{1.5}$$

Note that once these two conditions are verified, it follows that for any fixed $l \geq 1$,

$$\max\{i : p^i_{n,l} \neq 0\} = o_p(|T^{\downarrow}_{n,l}|^{1/2}) \text{ as } n \to \infty;$$

see, *e.g.*, [12, Lemma A.1].

1.3 Proof of Theorem 1.1

Now we are ready to give the proof of Theorem 1.1, assuming the results of Section 1.2.

Proof It suffices to prove that for any fixed $j \in \mathbb{N}$,

$$\left(\frac{\sigma(p_n)T^{\downarrow}_{n,1}}{n^{1/2}}, \left(\frac{\sigma(p_n)T^{\downarrow}_{n,l}}{c_n}, 2 \leq l \leq j\right), \frac{n - |T^{\downarrow}_{n,1}|}{c_n^2}\right) \xrightarrow{d} \left(\mathcal{T}, (\mathcal{T}^{\downarrow}_1, \dots, \mathcal{T}^{\downarrow}_{j-1}), \tau\left(\frac{1}{\sigma}\right)\right).$$

The convergence of the third coordinate is simply (1.2). This, in particular, implies that $\frac{|T^{\downarrow}_{n,1}|}{n} \xrightarrow{p} 1$. Since $p_n \to p$ in L^2, it straightforwardly follows that with probability $1 - o(1)$, $(T^{\downarrow}_{n,1}, n \geq 1)$ satisfies the conditions of Theorem 1.6; this yields the convergence of the first coordinate. With (1.4) and (1.5), we can also apply Theorem 1.6 to

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

each $T^{\downarrow}_{n,l}$ with $l \geq 2$ and conclude that

$$\frac{\sigma(p_{n,l})}{|T^{\downarrow}_{n,l}|^{1/2}} T^{\downarrow}_{n,l} \xrightarrow{d} \mathcal{T}.$$

Since the trees $(T^{\downarrow}_{n,l}, l \geq 1)$ are conditionally independent given their degree sequences, it follows that

$$\left(\frac{\sigma(p_{n,l})}{|T^{\downarrow}_{n,l}|^{1/2}} T^{\downarrow}_{n,l}, 2 \leq l \leq j \right) \xrightarrow{d} (\widetilde{\mathcal{T}}_{l-1}, 2 \leq l \leq j),$$

where $(\widetilde{\mathcal{T}}_l)_{l \in \mathbb{N}}$ are independent copies of $\mathcal{T}$. Using (1.5) again, together with (1.3) and Brownian scaling, the convergence of the second coordinate then follows. ■

Outline of the Rest of the Paper In Section 2, we describe a combinatorial construction that associates a *marked cyclic forest* with the concatenation of a sequence of first passage bridges, followed by one skip-free bridge. This construction is what links Theorem 1.4 with random forests. In Section 3, we give the proof of Theorem 1.4 and Corollary 1.5. Finally in Section 4, we prove (1.4) and (1.5) using martingale concentration inequalities.

2 Coding Marked Forests by Skip-free Walks

We call a sequence of integers $\mathbf{b} = (b(0), b(1), \dots, b(n))$ a *skip-free bridge* if

$$b(0) = 0, b(n) = -1 \quad \text{and} \quad \forall 0 \leq i \leq n-1,\ b(i+1) - b(i) \geq -1.$$

If $\mathbf{b}$ is a skip-free bridge and $\min_i \{i : b(i) = -1\} = n$, then we call $\mathbf{b}$ a *first passage bridge*. Given a skip-free bridge $\mathbf{b}$ and a positive integer $k \leq n$, we define a skip-free bridge $\mathbf{b}^{(k)}$ as follows. First, for $1 \leq i \leq n$, let $b(n+i) = b(n) + b(i) = -1 + b(i)$. Then for $0 \leq i \leq n$, let $\mathbf{b}^{(k)}(i) = b(k+i) - b(k)$. Let $[n] = \{1, \dots, n\}$. We have the following elementary lemma as a variant of the classical ballot theorem.

Lemma 2.1 ([16, Lemma 6.1]) *Fix a skip-free bridge $\mathbf{b} = (b(i), 0 \leq i \leq n)$, and let $r = r(\mathbf{b}) \in [n]$ be minimal so that $b(r) = \min(b(i), i \leq n)$. Then $\mathbf{b}^{(r)}$ is a first passage bridge, and r is the only such value in $[n]$.*

Lemma 2.1 is illustrated by Figure 1(a) and (b). In Figure 1(a), we have a skip-free bridge $\mathbf{b} = (0, -1, -1, -2, -1, 1, 0, -1)$. The vertical dashed line shows the position of $\mathbf{b}$ attaining its minimum for the first time, hence the unique position for the cyclic shift to transform $\mathbf{b}$ to a first passage bridge, as claimed by Lemma 2.1. The resulting first skip-free bridge, with steps $\mathbf{b}^{(3)}$, is shown in Figure 1(b).

A plane tree is a rooted tree T in which the children of each node have a left-to-right order. Recall that for a plane tree T and a node $v \in v(T)$, we write $k_T(v)$ to denote the degree of v in T. We also write $\operatorname{lex}(T) = (k_T(u_1), \dots, k_T(u_{|T|}))$ where $(u_i, 1 \leq i \leq |T|) = (u_i(T), 1 \leq i \leq |T|)$ are nodes of T listed in lexicographic order, *i.e.*, depth-first search order on $v(T)$. This order is "lexicographic" in the sense that the plane tree can

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

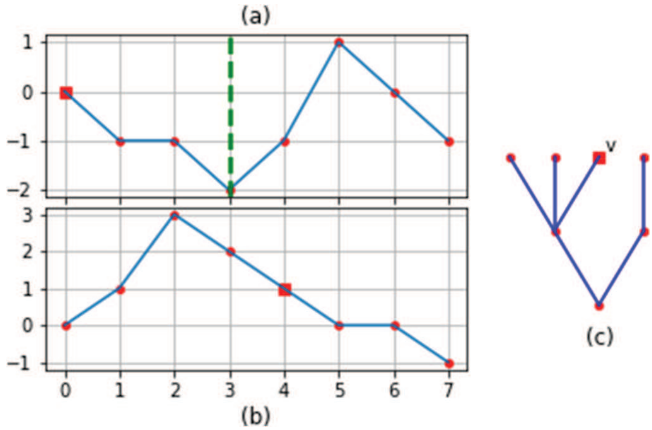

Figure 1: (a) A skip-free bridge; (b) the corresponding marked first-passage bridge; (c) the corresponding marked tree.

be represented as a subset of $\bigcup_{n=0}^{\infty} \mathbb{N}^n$, where $\mathbb{N} = \{1,2,\dots\}$ and $\mathbb{N}^0 = \{\varnothing\}$, as used in [11].

For any sequence $\mathbf{c} = (c_1,\dots,c_n) \in \mathbb{R}^n$, we write $W_{\mathbf{c}}(j) = \sum_{i=1}^{j}(c_i - 1)$ for $j \in [n]$, let $W_{\mathbf{c}}(0) = 0$ and make $W_{\mathbf{c}}$ a continuous function on $[0,n]$ by linear interpolation. A classical bijection between plane trees and first passage bridges associates to a tree T its *depth-first walk* $(W_{\mathrm{lex}(T)}(i), 0 \le i \le n)$; see, *e.g.*, [16, Chapter 6]. We build on this bijection below.

For a plane tree T and $v \in v(T)$, we call the pair (T,v) a *marked tree* and call v the *mark*. The bijection between first passage bridges and plane trees also leads to a bijection between skip-free bridges and marked trees. This bijection, depicted in Figure 1, is specified as follows. For a skip-free bridge $\mathbf{b}$, let $r = r(\mathbf{b})$ as in Lemma 2.1, let $\mathbf{b}' = \mathbf{b}^{(r)}$ be the first passage bridge corresponding to $\mathbf{b}$, and let T be the plane tree with depth-first walk $\mathbf{b}'$. Then the marked node is $v = u_{|T|-r+1}(T)$, the $(|T| - r + 1)$st node of T in lexicographic order. The mark v is denoted by a red square in Figure 1.

A *marked forest* is a pair (F,v) where F is a plane forest and $v \in v(F)$. We refer to v as the *mark* of (F,v). A *marked cyclic forest* is a marked forest with its mark in its last tree; it is so named because we can equivalently view such a forest as having its trees arranged around a cycle.

Fix an integer sequence $W = (W_i : 0 \le i \le n)$ with $W_0 = 0$, $W_n = -k$, and $W_i - W_{i-1} \ge -1$ for all $1 \le i \le n$. The bijections described above allow us to view W as a marked cyclic forest $(F,v) = (F(W), v(W))$ consisting of $k-1$ trees and one marked tree, as follows. For integer $b < 0$, let $\tau(b) = \inf\{t \in \mathbb{N} : W_t \le b\}$. For $1 \le j \le k-1$, let T_j be the tree whose depth-first walk is $(W_i - W_{\tau(-(j-1))} : \tau(-(j-1)) \le i \le \tau(-j))$. Let (T_k, v) be the marked tree corresponding to skip-free bridge $(W_i - W_{\tau(-(k-1))} : \tau(-(k-1)) \le i \le n)$. Then $(F(W), v(W)) = ((T_1,\dots,T_k), v)$. We call W the *coding walk* of the forest, and note that the coding is bijective: W can be recovered from $(F(W), v(W))$ as the concatenation of the first-passage bridges which code $T_1,\dots,T_{k-1}$ and the skip-free bridge that codes (T_k, v). This bijection is illustrated in Figures 2 and 3. In Figure 2, the whole sequence is decomposed into three segments

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

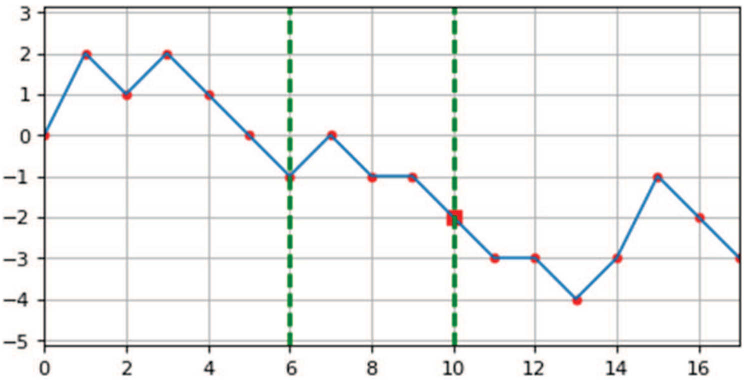

Figure 2: A skip-free walk $W = (W_i : 0 \le i \le 17)$.

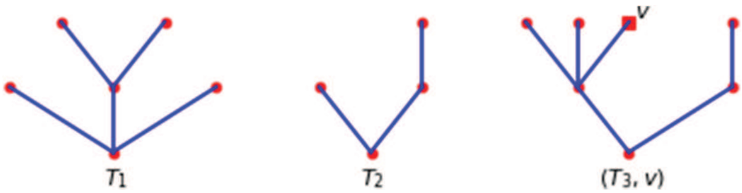

Figure 3: The marked forest $(F(W), v(W)) = ((T_1, T_2, T_3), v)$.

(divided by vertical dashed lines). The first two segments are first passage bridges, hence correspond to plane trees T_1, T_2. The last part is a skip-free bridge, and hence corresponds to a marked tree (T_3, v) and the node v is again depicted by a square mark. These trees are shown in Figure 3.

Given a degree sequence $s = (s^{(i)}, i \ge 0)$ with $\sum_{i\ge 0} s^i = n$, recall from Section 1.2 that $d(s) \in \mathbb{Z}^n_{\ge 0}$ is the vector whose entries are weakly increasing and with $s^{(i)}$ entries equal to i, for each $i \ge 0$. Let $\mathrm{D}(s)$ be the set of sequences $d \in \mathbb{Z}^n_{\ge 0}$ that are permutations of $d(s)$ (there are $n!/(\prod_i s^i!)$ of them). Let $\mathrm{MCF}(s)$ be the set of all marked cyclic forests with degree sequence s. By the correspondence we developed previously, the following lemma is immediate.

Lemma 2.2 *Fix a degree sequence $s = (s^{(i)}, i \ge 0)$ with $\sum_{i\ge 0} s^i = n$. Let π be a uniformly random permutation of $[n]$, and let $W = W_{\pi(d(s))}$. Then the marked cyclic forest $(F(W), v(W))$ coded by W is uniformly distributed on MCF(s).*

In particular, we have the following corollary.

Corollary 2.3 *Let $s = (s^{(i)}, i \ge 0)$ with $\sum_{i\ge 0} s^i = n$. Let (F, v) be a uniformly random element of MCF(s), and let M be the total number of nodes in the non-marked trees of (F, v). Let π be a uniformly random permutation of $[n]$ and let $W : [0, n] \to \mathbb{R}$, $W(t) = W_{\pi(d(s))}(t)$. Then*

$$M \overset{d}{=} \inf\{t : W(t) = -c(s) + 1\}. \tag{2.1}$$

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

We will also need the following easy fact connecting linear forests with marked cyclic forests.

Lemma 2.4 *Fix a degree sequence $s = (s^{(i)}, i \geq 0)$, and let F be a uniformly random linear forest with degree sequence s, and let (F^*, v) be the marked cyclic forest obtained from F by marking a uniformly random node and applying the requisite cyclic shift of the trees of F. Then (F^*, v) is a uniformly random element of $MCF(s)$.*

Proof Let $\mathrm{F}(s)$ be the set of all plane forests with degree sequence s. The operation of marking a node induces an n-to- $c(s)$ map from $\mathrm{F}(s)$ to $\mathrm{MCF}(s)$, from which the lemma is immediate. ∎

The preceding lemma allows us to relate the random forest F_n from Theorem 1.1 with the skip-free walk $S_n = (S_{n,k}, 0 \leq k \leq n)$ from Theorem 1.4. Let $(F_n^*, v_n) = ((T_{n,k}, 1 \leq k \leq c_n), v_n)$ be obtained from F_n by marking a uniformly random node and applying the requisite cyclic shift of the trees of F_n. Then we can couple F_n and S_n so that $S_n = (S_{n,j}, 0 \leq j \leq n)$ is the coding walk of (F_n^*, v_n). We work with such a coupling for the remainder of the paper.

3 Convergence of the Coding Processes

The goal of this section is to prove Theorem 1.4 and Corollary 1.5. To achieve that, we decompose the walk process into two random processes. To be precise, let $d_n := \frac{n^{1/2}}{c_n}$ and fix a sequence $(t_n)_{n \in \mathbb{N}}$, such that $t_n = o(d_n)$ and $t_n = \omega(1)$. This is possible, since $d_n \to \infty$ as $n \to \infty$ by our assumption that $c_n = o(n^{1/2})$. We consider the following two processes. Let $(M_{n,k}, k \leq n)$ be as follows: $M_{n,0} = 0$, and for $k \geq 1$,

$$M_{n,k} - M_{n,k-1} = X_{n,k} \mathbb{1}_{|X_{n,k}| < t_n}.$$

Similarly, let $(R_{n,k}, k \leq n)$ be given by $R_{n,0} = 0$, and for $k \geq 1$,

$$R_{n,k} - R_{n,k-1} = X_{n,k} \mathbb{1}_{|X_{n,k}| \geq t_n}.$$

Then clearly we have $S_{n,k} = M_{n,k} + R_{n,k}$ for all $k \leq n$. Define the following quantity:

$$\mu_n^+ := \sum_{j \geq t_n + 1} (j-1) \frac{s_n^j}{n}.$$

Theorem 1.4 is an immediate consequence of the following two results:

$$\left(\frac{1}{c_n} (M_{n,\lfloor tc_n^2 \rfloor} + \mu_n^+ \lfloor tc_n^2 \rfloor), t \geq 0 \right) \xrightarrow{d} (\sigma B(t), t \geq 0), \tag{3.1}$$

$$\left(\frac{1}{c_n} (R_{n,\lfloor tc_n^2 \rfloor} - \mu_n^+ \lfloor tc_n^2 \rfloor), t \geq 0 \right) \xrightarrow{d} 0, \tag{3.2}$$

where 0 denotes a process Z such that $\mathbf{P}\{Z(t) = 0, \ \forall t \geq 0\} = 1$. For (3.1), we are going to use the following theorem from [8].

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

Theorem 3.1 ([8, Theorem 4]) *Suppose an urn U contains n balls, each marked by an element of a set S of cardinality $c < n/2$. Let H_{Uk} be the distribution of k draws made at random without replacement from U, and let M_{Uk} be the distribution of k draws made at random with replacement. Then the two probabilities on S^k satisfy*

$$\|H_{Uk} - M_{Uk}\| \le 2ck/n,$$

where $\|\cdot\|$ denotes the total variation distance.

Proof of (3.1) Let $(\tilde{X}_{n,k}, k \le n)$ be i.i.d. with the law of $X_{n,1}\mathbb{1}_{|X_{n,1}|<t_n}$, set $\tilde{M}_{n,0} = 0$ and for $k \ge 1$, let

$$\tilde{M}_{n,k} = \sum_{j=1}^{k} \tilde{X}_{n,j}.$$

Now apply Theorem 3.1 with urn U containing n balls, with s_n^j balls marked by $j-1$ for $0 \le j \le t_n$, $j \ne 1$, and $s_n^1 + \sum_{j>t_n} s_n^j$ balls marked by 0, with $S = \{-1, 0, 1, \dots, t_n - 1\}$, and with $k = k(n) = \lfloor n/d_n \rfloor$. This yields that

$$\left\|(X_{n,j}\mathbb{1}_{|X_{n,j}|<t_n}, j \le k) - (\tilde{X}_{n,j}, j \le k)\right\| \le \frac{2(t_n+1)\lfloor n/d_n)\rfloor}{n} = 4\frac{t_n}{d_n},$$

so for all Borel $B \subset \mathbb{R}^k$,

$$\left|\mathbf{P}\left\{(M_{n,j}, j \le k(n)) \in B\right\} - \mathbf{P}\left\{(\tilde{M}_{n,j}, j \le k(n)) \in B\right\}\right| \le 4\frac{t_n}{d_n}.$$

Since $t_n = o(d_n)$ and $k(n) = \lfloor n/d_n \rfloor > d_n \cdot c_n^2 - 1 = \omega(c_n^2)$, this implies that to establish (3.1) it suffices to prove that

$$\left(\frac{1}{c_n}\left(\tilde{M}_{n,\lfloor tc_n^2 \rfloor} + \mu_n^+ \lfloor tc_n^2 \rfloor\right), t \ge 0\right) \xrightarrow{d} \left(\sigma B(t), t \ge 0\right). \tag{3.3}$$

Note that

$$\mathbf{E}\tilde{X}_{n,1} = \sum_{j \le t_n} (j-1)\frac{s_n^j}{n} = \frac{1}{n}\sum_j (j-1)s_n^j - \sum_{j \ge t_n+1} (j-1)\frac{s_n^j}{n} = -\frac{c_n}{n} - \mu_n^+. \tag{3.4}$$

Define σ_n^- by setting

$$\begin{aligned}(\sigma_n^-)^2 &:= Var(\tilde{X}_{n,1}) = \mathbf{E}\left[\tilde{X}_{n,1}^2\right] - \mathbf{E}\left[\tilde{X}_{n,1}\right]^2 \\ &= \sum_{j \le t_n} (j-1)^2 \frac{s_n^j}{n} - \left(-\mu_n^+ - \frac{c_n}{n}\right)^2.\end{aligned} \tag{3.5}$$

Applying Donsker's theorem to the process $(\tilde{M}_{n,k} + k(\mu_n^+ + \frac{c_n}{n}), k \ge 0)$, we obtain that

$$\left(\frac{1}{a}(\tilde{M}_{n,\lfloor ta^2 \rfloor} + \mu_n^+ \lfloor ta^2 \rfloor) + \frac{c_n \lfloor ta^2 \rfloor}{na}, t \ge 0\right) \xrightarrow{d} \left(\sigma_n^- B(t), t \ge 0\right), \tag{3.6}$$

as $a \to \infty$.

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

Since $\sum_{j\geq 0}(j-1)p^j$ is convergent, for any prescribed $\delta > 0$, we can find L large such that $\sum_{j>L}(j-1)p^j < \delta/2$. By our assumption that $n^{-1}\cdot s_n \to p$ in L^2, for this fixed L, we can find n large such that $\sum_{j>L}(j-1)\frac{s_n^j}{n} < \delta$. Since $t_n \to \infty$, we must have $\mu_n^+ \leq \sum_{j>L}(j-1)\frac{s_n^j}{n} < \delta$ for n large enough, *i.e.*,

$$\mu_n^+ \longrightarrow 0 \text{ as } n \longrightarrow \infty. \tag{3.7}$$

Similarly, the assumption that $n^{-1}\cdot s_n \to p$ in L^2 implies that

$$\sigma_n^2 := \sum_j j(j-1)\frac{s_n^j}{n} \longrightarrow \sigma^2 < \infty,$$

so

$$(\sigma_n^+)^2 \longrightarrow 0 \text{ as } n \longrightarrow \infty, \tag{3.8}$$

where we let $(\sigma_n^+)^2 := \sum_{j\geq t_n+1} j(j-1)\frac{s_n^j}{n}$. Using (3.4) and (3.5), we have

$$\sigma_n^2 - (\sigma_n^-)^2 = (\sigma_n^+)^2 - \Big(\mu_n^+ + \frac{c_n}{n}\Big)\Big(1 - \mu_n^+ - \frac{c_n}{n}\Big).$$

Using (3.7), (3.8), and the fact that $c_n = o(\sqrt{n})$, the last quantity

$$(\sigma_n^+)^2 - \Big(\mu_n^+ + \frac{c_n}{n}\Big)\Big(1 - \mu_n^+ - \frac{c_n}{n}\Big) \longrightarrow 0 \text{ as } n \longrightarrow \infty,$$

so $\sigma_n^- \to \sigma$ as $n \to \infty$. Taking $a = c_n$ in (3.6), then letting $n \to \infty$, now yields that

$$\left(\frac{1}{c_n}\big(\tilde{M}_{n,\lfloor tc_n^2\rfloor} + \mu_n^+\lfloor tc_n^2\rfloor\big) + \frac{\lfloor tc_n^2\rfloor}{n}, t \geq 0\right) \xrightarrow{d} \big(\sigma B(t), t \geq 0\big).$$

Since $c_n^2 = o(n)$, (3.3) follows. ∎

To prove (3.2), we need the following result concerning *dilation*. Recall (or see, *e.g.*, [5]) that given real random variables U, V, we say U is a *dilation* of V if there exist random variables $\hat{U}, \hat{V}$ such that

$$\hat{U} \stackrel{d}{=} U, \quad \hat{V} \stackrel{d}{=} V, \quad \text{and} \quad \mathbf{E}\big[\hat{U}|\hat{V}\big] = \hat{V}.$$

Proposition 3.2 ([5, Proposition 20.6]) *Suppose $X_1,\dots,X_k$ and $X_1^*,\dots,X_k^*$ are samples from the same finite population $x_1,\dots,x_n$, without replacement and with replacement, respectively. Let $S_k = \sum_{i=1}^k X_i$, $S_k^* = \sum_{i=1}^k X_i^*$. Then S_k^* is a dilation of S_k. In particular, $\mathbf{E}\big[\phi(S_k^*)\big] \geq \mathbf{E}\big[\phi(S_k)\big]$ for all continuous convex function $\phi : \mathbb{R} \to \mathbb{R}$.*

Proof of (3.2) We prove that for all $\varepsilon > 0$, we have

$$\limsup_{n\to\infty} \mathbf{P}\left\{\max_{i\leq c_n^2/\varepsilon}\left|\frac{R_{n,i} - i\mu_n^+}{c_n}\right| > \varepsilon\right\} \leq \varepsilon,$$

this immediately implies (3.2). Fix n and let $\mathrm{c}_1,\dots,\mathrm{c}_n$ be such that $|\{1 \leq k \leq n : \mathrm{c}_k = j\}| = s_n^j$. Let $C_1,\dots,C_n$ be a uniformly random permutation of $\mathrm{c}_1,\dots,\mathrm{c}_n$. Fix $t_n \in \mathbb{N}$.

Define $(R_i, 0 \le i \le n)$ as follows: let $R_0 = 0$, and for $i \ge 0$, let

$$R_{i+1} = \begin{cases} R_i + C_i - 1, & \text{if } C_i \ge t_n + 1; \\ R_i, & \text{if } C_i \le t_n. \end{cases}$$

For $0 \le i \le n$, let $\mathcal{F}_i = \sigma(C_1, \dots, C_i)$. Since $R_n = n\mu_n^+$ and the process $(R_i, 0 \le i \le n)$ has exchangeable increments,

$$\mathbf{E}\left[R_{i+1} \mid \mathcal{F}_i\right] = R_i + \frac{n\mu_n^+ - R_i}{n - i}. \tag{3.9}$$

Now let $K_i = \mathbf{E}\left[R_{i+1} - R_i \mid \mathcal{F}_i\right] = \frac{n\mu_n^+ - R_i}{n-i}$. Then using (3.9), we have

$$\mathbf{E}\left[K_{i+1} \mid \mathcal{F}_i\right] = \frac{n\mu_n^+ - R_i}{n - (i+1)} - \frac{n\mu_n^+ - R_i}{(n-i)(n-(i+1))} = K_i.$$

Hence, K_i is an $\mathcal{F}_i$−martingale.

Since for any $0 \le i \le s$,

$$\frac{n\mu_n^+ - R_i}{n - i} = \mu_n^+ + \frac{i\mu_n^+ - R_i}{n - i},$$

and μ_n^+ is a constant, if we define $\tilde{K}_i = \frac{i\mu_n^+ - R_i}{n-i}$, then $\tilde{K}_i$ is also an $\mathcal{F}_i$−martingale. It follows that for any $\varepsilon > 0$,

$$\begin{aligned} \mathbf{P}\left\{\frac{1}{c_n} \max_{i \le s} \left|i\mu_n^+ - R_i\right| > \varepsilon\right\} &\le \frac{n^2}{\varepsilon^2 c_n^2} \mathbf{E}\left[\left(\max_{i \le s} \frac{\left|i\mu_n^+ - R_i\right|}{n - i}\right)^2\right] \\ &\le \frac{4n^2 \mathbf{E}\left[(s\mu_n^+ - R_s)^2\right]}{\varepsilon^2 c_n^2 (n - s)^2}, \end{aligned} \tag{3.10}$$

where in the first line we use Markov's inequality, and in the last line we use the L^2 maximal inequality for martingales (see, *e.g.*, [9, Theorem 5.4.3]).

Since the process $(R_s, 0 \le s \le n)$ has exchangeable increments, we have $\mathbf{E}R_s = s\mu_n^+$. Let $R_s^* = \sum_{i \le s} J_i$ where $J_1, \dots, J_s$ are i.i.d. random variables with $J_1 \overset{d}{=} R_1$. Then Proposition 3.2 gives

$$\begin{aligned} \mathbf{E}\left[R_s^2\right] \le \mathbf{E}\left[R_s^{*2}\right] &= \mathbf{E}\left[(J_1 + \dots + J_s)^2\right] = s\mathbf{E}\left[J_1^2\right] + s(s-1)(\mathbf{E}J_1)^2 \\ &= s(\sigma_n^{+2} - \mu_n^+) + s(s-1)\mu_n^{+2}. \end{aligned}$$

Therefore,

$$\mathbf{E}\left[(s\mu_n^+ - R_s)^2\right] = \mathbf{E}\left[R_s^2\right] - s^2\mu_n^{+2} \le s(\sigma_n^{+2} - \mu_n^+) - s\mu_n^{+2} \le s\sigma_n^{+2}.$$

Now take $s = s(n) = c_n^2/\varepsilon$ in (3.10). For n large, this is less than $n/2$, so $(n-s)^2 > n^2/4$, and we obtain

$$\mathbf{P}\left\{\frac{1}{c_n} \max_{i \le c_n^2/\varepsilon} \left|i\mu_n^+ - R_i\right| > \varepsilon\right\} \le \frac{16s\sigma_n^{+2}}{\varepsilon^2 c_n^2} = \frac{16\sigma_n^{+2}}{\varepsilon^3} \le \varepsilon,$$

the last inequality holding for n large, since $\sigma_n^+ \to 0$ as $n \to \infty$. This completes the proof. ∎

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

Recall that in Section 1, we let $\tau(x) = \inf(t : B(t) \le -x)$ for $x \ge 0$. If we let $\tau_n = \sum_{1\le i<c_n} |T_{n,i}| = n - |T_{n,c_n}|$ be the total size of non-marked trees of (F_n^*, v_n), then since S_n is the coding process of (F_n^*, v_n) as argued in the last paragraph of Section 2, using (2.1), we have

$$\tau_n = \inf\left\{k : S_{n,k} = -(c_n - 1)\right\}.$$

From this, we immediately get the following corollary of Theorem 1.4.

Corollary 3.3 *Given the assumptions in Theorem 1.1, we have*

$$\frac{\tau_n}{c_n^2} \xrightarrow{d} \tau\left(\frac{1}{\sigma}\right), \tag{3.11}$$

where $\tau(x) = \inf(t : B(t) \le -x)$ for $x \ge 0$ and $(B(t), t \ge 0)$ is standard Brownian Motion.

Remark Note that the right-hand side of (3.11) has density $\frac{1}{\sigma\sqrt{2\pi t^3}} \exp\left(-\frac{1}{2t\sigma^2}\right) dt$; see, *e.g.*, [17, Theorem 6.9].

The corollary above in fact tells us something about the size of the largest tree $T_{n,1}^{\downarrow}$.

Corollary 3.4 *For a marked cyclic forest (F, v), let $MT(F, v)$ denote the marked tree, i.e., the tree of F containing v. Then*

$$\mathbf{P}\left\{MT(F_n^*, v_n) = T_{n,1}^{\downarrow}\right\} \longrightarrow 1$$

as $n \to \infty$.

Proof It is clear that

$$\begin{aligned}\mathbf{P}\left\{MT(F_n^*, v_n) \ne T_{n,1}^{\downarrow}\right\} &\le \mathbf{P}\left\{|MT(F_n^*, v_n)| < n/2\right\}\\ &= \mathbf{P}\left\{\tau_n > n/2\right\} = \mathbf{P}\left\{\frac{\tau_n}{c_n^2} > \frac{n}{2c_n^2}\right\} \longrightarrow 0,\end{aligned}$$

where in the last line, the first equation is by Lemma 2.4, and the final convergence is by Corollary 3.3 and the assumption $c_n^2 = o(n)$. ∎

Now we are ready to prove Corollary 1.5.

Proof of Corollary 1.5 As noted, it suffices to prove (1.2) and (1.3). Corollary 3.3 and Corollary 3.4 together imply (1.2).

For (1.3), first note that by Lemma 2.2, the process $S_n = (S_{n,k}, 0 \le k \le n)$ has the same law as the coding walk $W(F_n)$ of F_n. Applying Corollary 3.4 then yields that the law of $(|T_{n,2}^{\downarrow}|, \dots, |T_{n,j}^{\downarrow}|)$ is asymptotically equivalent to the law of $(g_1^n - d_1^n, \dots, g_{j-1}^n - d_{j-1}^n)$, the first $j-1$ ranked excursion lengths of S_n above its running minimum before time τ_n. Using this equivalence, (1.3) now follows from Theorem 1.4 by the Portmanteau Theorem ([14, Theorem 12.6]), since the vector $(g_1 - d_1, \dots, g_{j-1} - d_{j-1})$ has a density. ∎

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

4 Empirical Degree Sequences of Trees

In this section, we aim to prove (1.4) and (1.5).

For $i \geq 0$ and $x \leq n$, let

$$Q_n^i(x) := |\{1 \leq j \leq x : C_{n,j} = i\}|,$$

where $(C_{n,1}, \ldots, C_{n,n})$ is a uniformly random permutation of $d(s_n)$. Let $S_{n,k} = \sum_{j=1}^{k}(C_{n,j} - 1)$. Let $\mathcal{F}_j = \sigma(C_{n,1}, \ldots, C_{n,j})$. Since $Q_n^i(n) = s_n^i = np_n^i$ and the process $(Q_n^i(j), 0 \leq j \leq n)$ has exchangeable increments,

$$\mathbf{E}\left[Q_n^i(j+1) \mid \mathcal{F}_j\right] = Q_n^i(j) + \frac{np_n^i - Q_n^i(j)}{n-j}.$$

Setting $K_j = \frac{np_n^i - Q_n^i(j)}{n-j} = \mathbf{E}\left[Q_n^i(j+1) - Q_n^i(j) \mid \mathcal{F}_j\right]$ for $0 \leq j \leq n-1$, then

$$\mathbf{E}\left[K_{j+1} \mid \mathcal{F}_j\right] = \frac{np_n^i - Q_n^i(j)}{n-(j+1)} - \frac{np_n^i - Q_n^i(j)}{(n-(j+1))(n-j)} = K_j,$$

so K_j is an $\mathcal{F}_j$–martingale. If we let $\tilde{K}_j = \frac{jp_n^i - Q_n^i(j)}{n-j}$, then $\tilde{K}_j = K_j - p_n^i$, so $\tilde{K}_j$ is also an $\mathcal{F}_j$–martingale.

We now use the following martingale bound from [13]. Let $\{X_j\}_{j=0}^n$ be a bounded martingale adapted to a filtration $\{\mathcal{F}_j\}_{j=0}^n$. Let $V = \sum_{j=0}^{n-1} var\{X_{j+1} \mid \mathcal{F}_j\}$, where

$$var\{X_{j+1} \mid \mathcal{F}_j\} := \mathbf{E}\left[(X_{j+1} - X_j)^2 \mid \mathcal{F}_j\right] = \mathbf{E}\left[X_{j+1}^2 \mid \mathcal{F}_j\right] - X_j^2.$$

Let

$$v = \text{ess sup } V \quad \text{and} \quad b = \max_{0 \leq j \leq n-1} \text{ess sup}(X_{j+1} - X_j \mid \mathcal{F}_j).$$

Theorem 4.1 ([13, Theorem 3.15]) *For any $t \geq 0$,*

$$\mathbf{P}\left\{\max_{0 \leq j \leq n} X_j \geq t\right\} \leq \exp\left(-\frac{t^2}{2v(1 + bt/(3v))}\right).$$

We will apply this theorem to bound the fluctuations of $Q_n^i(s)$.

Proposition 4.2 *For any $0 < t < 1$, we have*

(4.1)
$$\mathbf{P}\left\{\exists s > c_n : \left|p_n^i - \frac{Q_n^i(s)}{s}\right| \geq t\right\} \leq 2\exp\left(-\frac{3t^2 c_n}{5}\right).$$

Proof It is not hard to show that for any $0 \leq j \leq n-2$,

$$var\{\tilde{K}_{j+1} \mid \mathcal{F}_j\} \leq \frac{1}{4} \cdot \frac{1}{(n-(j+1))^2};$$

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

see, *e.g.*, [12, Lemma 3.2]. Thus, for $1 \le x \le n-2$,

$$V = \sum_{j=0}^{x-1} var\{\tilde{K}_{j+1} \mid \mathcal{F}_j\} \le \frac{1}{4}\sum_{j=0}^{x-1}\frac{1}{(n-(j+1))^2}$$
$$\le \frac{1}{4}\int_{n-x-1}^{n-1}\frac{1}{m^2}\mathrm{d}m = \frac{x}{4(n-1)(n-x-1)}.$$

On the other hand, for $0 \le j \le x-1$, if $Q_n^i(j+1) = Q_n^i(j)$, then

$$|\tilde{K}_{j+1} - \tilde{K}_j| = \Big|\frac{np_n^i - Q_n^i(j)}{(n-(j+1))(n-j)}\Big|$$
$$= \Big|\frac{np_n^i - Q_n^i(j)}{n-j}\Big| \cdot \Big|\frac{1}{n-(j+1)}\Big| \le \frac{1}{n-x},$$

where in the last inequality we use $j \le x-1$ and the fact that $\frac{np_n^i - Q_n^i(j)}{n-j}$ is the proportion of $C_{n,j+1},\dots,C_{n,n}$ equalling to i (hence a value between 0 and 1). While if $Q_n^i(j+1) = Q_n^i(j)+1$, then

$$|\tilde{K}_{j+1} - \tilde{K}_j| = \left|\frac{np_n^i - Q_n^i(j)}{(n-(j+1))(n-j)} - \frac{1}{n-(j+1)}\right| \le \frac{1}{n-x}.$$

Applying Theorem 4.1 to both $\{\tilde{K}_j\}_{j=0}^x$ and $\{-\tilde{K}_j\}_{j=0}^x$ with $x = n - c_n$, we have

$$v \le \frac{n-c_n}{4(n-1)(c_n-1)} \le \frac{1}{4(c_n-1)} \le \frac{1}{2c_n},\ b \le \frac{1}{c_n}.$$

Hence, for $t \le 1$,

$$\mathbf{P}\left\{\max_{0\le j\le n-c_n}\left|p_n^i - \frac{np_n^i - Q_n^i(j)}{n-j}\right| \ge t\right\} \le 2\exp\left(-\frac{t^2}{\frac{1}{c_n}+\frac{2t}{3c_n}}\right) \le 2\exp\left(-\frac{3t^2c_n}{5}\right).$$

Using the exchangeability of $C_{n,1},\dots,C_{n,n}$, it follows that

$$\mathbf{P}\left\{\exists s > c_n : |p_n^i - \frac{Q_n^i(s)}{s}| \ge t\right\} = \mathbf{P}\left\{\max_{0\le j\le n-c_n}\Big|p_n^i - \frac{np_n^i - Q_n^i(j)}{n-j}\Big| \ge t\right\}$$
$$\le 2\exp\Big(-\frac{3t^2c_n}{5}\Big). \qquad \blacksquare$$

We next give the proofs of (1.4) and (1.5). In both proofs we use the coupling between F_n, (F_n^*, v_n) and S_n explained at the end of Section 2.

Proof of (1.4) Fix $i \ge 0$ and $l \ge 2$. By Corollary 3.4, with high probability $T_{n,1}^{\downarrow} = T_{n,c_n}$, *i.e.*, $T_{n,1}^{\downarrow}$ is the last tree of (F_n^*, v_n), in which case $T_{n,l}^{\downarrow} = T_{n,j}$ for some $j < c_n$. Recall that $\tau_n = \sum_{1\le k<c_n}|T_{n,k}|$.

Let $1 \le j < c_n$, and suppose $|\{v \in T_{n,j} : k(v) = i\}|/|T_{n,j}| \notin [p_n^i - \delta, p_n^i + \delta]$. Suppose that $|T_{n,j}| > \delta c_n^2 > c_n$ and $\tau_n < c_n^3$. Then there must exist $m > c_n$ and $1 \le u \le \tau_n - m$ such that

$$\left|\frac{|\{t \in [m] : C_{n,u+t} = i\}|}{m} - p_n^i\right| > \delta.$$

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

By union bound and the exchangeability of $(C_{n,1},\dots,C_{n,n})$, and by our assumption $\tau_n < c_n^3$, the probability of this is bounded above by $c_n^3\mathbf{P}\left\{\exists m > c_n : \left|\frac{Q_n^i(m)}{m} - p_n^i\right| > \delta\right\}$. Thus, for $l \geq 2$, for n large enough that $\delta c_n^2 > c_n$, we have

$$\mathbf{P}\left\{t|p_{n,l}^i - p_n^i| > \delta\right\} \leq \mathbf{P}\left\{\tau_n > c_n^3\right\} + \mathbf{P}\left\{|T_{n,l}^{\downarrow}| < \delta c_n^2\right\} + \mathbf{P}\left\{T_{n,1}^{\downarrow} \neq T_{n,c_n}\right\} + c_n^3\mathbf{P}\left\{\max_{s>c_n}\left|p_n^i - \frac{Q_n^i(s)}{s}\right| > \delta\right\}.$$

For any $\varepsilon > 0, \mathbf{P}\left\{\tau_n > c_n^3\right\} < \varepsilon/3$ by Corollary 3.3 for n large enough, and $\mathbf{P}\left\{|T_{n,l}^{\downarrow}| < \delta c_n^2\right\} < \varepsilon/3$ by Corollary 1.5. The second last probability tends to zero by Corollary 3.4. And for the last probability, for n large enough, $\sqrt{\frac{5}{3}}c_n^{-1/3} < \delta$; hence, Proposition 4.2 gives upper bound $2c_n^3\exp(-c_n^{1/3})$, which tends to zero. Thus, $\mathbf{P}\left\{|p_{n,l}^i - p_n^i| > \delta\right\} < \varepsilon$ for n large; this proves (1.4) for $i \geq 0$ and $l > 1$.

Finally, since $|T_{n,1}^{\downarrow}|/n \to 1$, the fact that $|p_{n,1}^i - p_n^i| \to 0$ in probability for each $i \geq 0$ is immediate. ■

Proof of (1.5) Fix $\varepsilon > 0$. By Corollary 3.3, we can pick $M > 0$ large enough such that for n large enough,

$$\mathbf{P}\left\{\tau_n > Mc_n^2\right\} < \varepsilon. \tag{4.2}$$

By Corollary 3.4, we have that for n large enough,

$$\mathbf{P}\left\{T_{n,c_n} \neq T_{n,1}^{\downarrow}\right\} < \varepsilon\,. \tag{4.3}$$

For fixed $l \geq 2$ and for this $\varepsilon > 0$, there exists $\delta > 0$ such that $\mathbf{P}\left\{g_{l-1} - d_{l-1} \leq \delta\right\} < \varepsilon/2$, so by Corollary 1.5, for n large,

$$\mathbf{P}\left\{\frac{|T_{n,l}^{\downarrow}|}{c_n^2} \leq \delta\right\} < \varepsilon. \tag{4.4}$$

Next we fix $t > 0$ large enough such that

$$\mathbf{E}\left[C_{n,1}^2\mathbb{1}_{C_{n,1}\geq t}\right] < \frac{\varepsilon^2\delta}{M} \quad \text{and} \quad \sum_{i>t} i^2 p_n^i < \varepsilon\,; \tag{4.5}$$

this is possible, since $p_n = (p_n^i, i \geq 0) \to p = (p^i, i \geq 0)$ in L^2. For fixed $l \geq 2$, we have

$$\begin{aligned}|\sigma^2(p_{n,l}) - \sigma^2(p_n)| &\leq \left|\sum_{i\leq t} i^2(p_{n,l}^i - p_n^i)\right| + \sum_{i>t} i^2 p_n^i + \sum_{i>t} i^2 p_{n,l}^i \\ &\leq \left|\sum_{i\leq t} i^2(p_{n,l}^i - p_n^i)\right| + \varepsilon + \sum_{i>t} i^2 p_{n,l}^i,\end{aligned} \tag{4.6}$$

where we use (4.5) in the second line.

https://doi.org/10.4153/S0008439520000685 Published online by Cambridge University Press

Let $L_n = \sum_{j \le Mc_n^2} C_{n,j}^2 \mathbb{1}_{C_{n,j} \ge t}$. If $T_{n,c_n} = T_{n,1}^{\downarrow}$ and $\tau_n \le Mc_n^2$, then $\sum_{i>t} i^2 p_{n,l}^i \le L_n / |T_{n,l}^{\downarrow}|$. Hence,

$$
\begin{aligned}
&\mathbf{P}\left\{|\sigma^2(p_{n,l}) - \sigma^2(p_n)| \ge 3\varepsilon\right\} \\
&\quad \le \mathbf{P}\left\{|\sigma^2(p_{n,l}) - \sigma^2(p_n)| \ge 3\varepsilon, \tau_n \le Mc_n^2, T_{n,c_n} = T_{n,1}^{\downarrow}, \frac{|T_{n,l}^{\downarrow}|}{c_n^2} > \delta\right\} \\
&\qquad + \mathbf{P}\left\{\tau_n > Mc_n^2\right\} + \mathbf{P}\left\{T_{n,c_n} \ne T_{n,1}^{\downarrow}\right\} + \mathbf{P}\left\{\frac{|T_{n,l}^{\downarrow}|}{c_n^2} \le \delta\right\} \\
&\quad \le \mathbf{P}\left\{\left|\sum_{i \le t} i^2 (p_{n,l}^i - p_n^i)\right| \ge \varepsilon\right\} + \mathbf{P}\left\{\frac{L_n}{|T_{n,l}^{\downarrow}|} > \varepsilon, \frac{|T_{n,l}^{\downarrow}|}{c_n^2} > \delta\right\} + 3\varepsilon \qquad (4.7)
\end{aligned}
$$

where we use (4.2), (4.3), (4.4), (4.6) and the aforementioned stochastic dominance in the last line.

Since t is fixed, we can use (1.4) to conclude that the first summand of (4.7) can be made arbitrarily small by taking n large enough. For the second summand, note that by exchangeability and (4.5),

$$
\mathbf{E} L_n = Mc_n^2 \mathbf{E}\left[C_{n,1}^2 \mathbb{1}_{C_{n,1} \ge t}\right] < c_n^2 \varepsilon^2 \delta ,
$$

so

$$
\mathbf{P}\left\{\frac{L_n}{|T_{n,l}^{\downarrow}|} > \varepsilon, \frac{|T_{n,l}^{\downarrow}|}{c_n^2} > \delta\right\} \le \mathbf{P}\left\{\frac{L_n}{c_n^2} > \varepsilon\delta\right\} \le \frac{\mathbf{E}\left[\frac{L_n}{c_n^2}\right]}{\varepsilon\delta} < \varepsilon.
$$

This completes the proof of (1.5) for $l \ge 2$. Again since $|T_{n,1}^{\downarrow}|/n \to 1$, (1.5) is immediate for the $l = 1$ case. ■

Acknowledgement I am grateful for my supervisor Prof. Louigi Addario-Berry, who suggested the project and gave numerous helpful insights and suggestions during our discussions.

References

[1] R. Abraham, J.-F. Delmas, and P. Hoscheit, *A note on the Gromov–Hausdorff–Prokhorov distance between (locally) compact metric measure spaces*. Electron. J. Probab. **18**(2013), no. 14, 21. http://dx.doi.org/10.1214/EJP.v18-2116

[2] D. Aldous, *The continuum random tree. I*. Ann. Probab. **19**(1991), 1–28.

[3] D. Aldous, *The continuum random tree. II. an overview*. In: Stochastic analysis (Durham, 1990), London Mathematical Society Lecture Note Series, 167, Cambridge University Press, Cambridge, UK, 1991, pp. 23–70. http://dx.doi.org/10.1017/CB9780511662980.003

[4] D. Aldous, *The continuum random tree. III*. Ann. Probab. **21**(1993), 248–289.

[5] D. J. Aldous, *Exchangeability and related topics*. In: École d'été de probabilités de Saint-Flour, XIII—1983, Lecture Notes in Mathematics, 1117, Springer, Berlin, Germany, 1985, pp. 1–198. http://dx.doi.org/10.1007/BFb0099421

[6] N. Broutin and J.-F. Marckert, *Asymptotics of trees with a prescribed degree sequence and applications*. Random Struct. Algor. **44**(2014), 290–316. http://dx.doi.org/10.1002/rsa.20463

[7] I. A. Cheplyukova, *The emergence of a giant tree in a random forest*. Discrete Math. Appl. **8**(1998), 17–33. http://dx.doi.org/10.1515/dma.1998.8.1.17

[8] P. Diaconis and D. Freedman, *Finite exchangeable sequences.* Ann. Probab. **8**(1980), 745–764.

[9] R. Durrett, *Probability: theory and examples*. 4th ed., Cambridge Series in Statistical and Probabilistic Mathematics, 31, Cambridge University Press, Cambridge, UK, 2010. http://dx.doi.org/10.1017/CB09780511779398

[10] S. N. Evans, *Probability and real trees*. Lecture Notes in Mathematics, 1920, Springer, Berlin, Germany, 2008. http://dx.doi.org/10.1007/978-3-540-74798-7

[11] J.-F. Le Gall, *Random trees and applications*. Probab. Surv. **2**(2005), 245–311. http://dx.doi.org/10.1214/154957805100000140

[12] T. Lei, *Scaling limit of random forests with prescribed degree sequences.* Bernouilli **25**(2019), 2409–2438.

[13] C. McDiarmid, *Concentration*. In: Probabilistic methods for algorithmic discrete mathematics, Algorithms Combinations, 16, Springer, Berlin, Germany, 1998, pp. 195–248. http://dx.doi.org/10.1007/978-3-662-12788-9-6

[14] P. Mörters and Y. Peres, *Brownian motion*. Cambridge Series in Statistical and Probabilistic Mathematics, 30, Cambridge University Press, Cambridge, UK, 2010. http://dx.doi.org/10.1017/CBO9780511750489

[15] Y. L. Pavlov, *Random forests*. VSP, Utrecht, 2000.

[16] J. Pitman, *Combinatorial stochastic processes*. Lecture Notes in Mathematics, 1875, Springer-Verlag, Berlin, Germany, 2006.

[17] R. L. Schilling and L. Partzsch, *Brownian motion*. De Gruyter, Berlin, Germany, 2012. http://dx.doi.org/10.1515/9783110278989

Department of Mathematics and Statistics, McGill University, 805 Sherbrooke Street West, Montréal, QC H3A 0B9, Canada
e-mail: tao.lei@mail.mcgill.ca
URL: http://www.math.mcgill.ca/~tlei/