Combinatorics, Probability and Computing (2019) 28, 81–99. © Cambridge University Press 2018
doi:10.1017/S0963548318000226

Random Recursive Trees and Preferential Attachment Trees are Random Split Trees

SVANTE JANSON[†]

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden
(e-mail: svante.janson@math.uu.se,
http://www.math.uu.se/svante-janson)

Received 17 June 2017; first published online 21 May 2018

We consider linear preferential attachment trees, and show that they can be regarded as random split trees in the sense of Devroye (1999), although with infinite potential branching. In particular, this applies to the random recursive tree and the standard preferential attachment tree. An application is given to the sum over all pairs of nodes of the common number of ancestors.

2010 *Mathematics subject classification*: Primary 60C05
Secondary 05C05

1. Introduction

The purpose of this paper is to show that the linear preferential attachment trees, a class of random trees that includes and generalizes both the random recursive tree and the standard preferential attachment tree, can be regarded as random split trees in the sense of Devroye [11], although with infinite (potential) branching.

Recall that the random recursive tree is an unordered rooted tree that is constructed by adding nodes one by one, with each node attached as the child of an existing node chosen uniformly at random; see *e.g.* [12, Section 1.3.1]. The general preferential attachment tree is constructed in a similar way, but for each new node, its parent is chosen among the existing nodes with the probability of choosing a node v proportional to $w_{d(v)}$, where $d(v)$ is the outdegree (number of existing children) of v, and $w_0, w_1, \dots$ is a given sequence of weights. The constant choice $w_k = 1$ thus gives the random recursive tree. The preferential attachment tree made popular by Barabási and Albert [3] (as a special case of more general preferential attachment graphs) is given by the choice $w_k = k + 1$; this coincides with the plane oriented recursive tree earlier introduced by Szymański [31]. We shall here

[†] Partly supported by the Knut and Alice Wallenberg Foundation.

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

consider the more general linear case

$$w_k = \chi k + \rho \tag{1.1}$$

for some real parameters χ and $\rho > 0$, which was introduced (at least for $\chi \geqslant 0$) by Pittel [29]. Thus the random recursive tree is obtained for $\chi = 0$ and $\rho = 1$, while the standard preferential attachment tree is the case $\chi = \rho = 1$. We allow $\chi < 0$, but in that case we have to assume that $\rho/|\chi|$ is an integer, say m, in order to avoid negative weights. (We then have $w_m = 0$ so a node never gets more than m children, and w_k for $k > m$ are irrelevant; see further Section 6.) See also [16, Section 6] and the further references given there. We denote the random linear preferential attachment tree with n nodes and weights (1.1) by $T_n^{\chi,\rho}$.

Remark 1. Note that multiplying all w_k by the same positive constant will not change the trees, so only the ratio χ/ρ is important. Hence we may normalize the parameters in some way when convenient; however, different normalizations are convenient in different situations, and therefore we keep the general and more flexible assumptions above unless we say otherwise.

Note also that our assumptions imply $w_1 = \chi + \rho > 0$ except in the case $\chi = -\rho$, when $w_1 = 0$ and $T_n^{\chi,\rho}$ deterministically is a path. We will usually ignore that trivial case hereafter, and assume $\chi + \rho > 0$.

Remark 2. The three cases $\chi > 0$, $\chi = 0$ and $\chi < 0$ give the three classes of very simple increasing trees defined and characterized by Panholzer and Prodinger [27]; see also [4] and [12, Section 1.3.3]. In fact, it suffices to consider $\chi = 1$, $\chi = 0$ and $\chi = -1$ (see Remark 1). Then, $\chi = 0$ yields the random recursive tree, as mentioned above, $\chi = 1$ yields the generalized plane oriented recursive tree, and $\chi = -1$ (and $\rho = m \in \mathbb{N}$) yields the m-ary increasing tree; see further Section 6.

Random split trees were defined by Devroye [11] as rooted trees generated by a certain recursive procedure using a stream of balls added to the root. We only need a simple but important special case (the case $s = 1$, $s_0 = 1$, $s_1 = 0$ in the notation of [11]), in which case the general definition simplifies to the following (we use $\mathcal{P}$ and P_i instead of $\mathcal{V}$ and V_i in [11]):

Let $b \geqslant 2$ be fixed and let $\mathcal{P} = (P_i)_1^b$ be a random vector of probabilities: in other words, $P_i \geqslant 0$ and $\sum_{i=1}^b P_i = 1$. Let $\mathcal{T}_b$ be the infinite rooted tree where each node has b children, labelled $1, \dots, b$, and give each node $v \in \mathcal{T}_b$ an independent copy $\mathcal{P}^{(v)} = (P_i^{(v)})_1^b$ of $\mathcal{P}$. (These vectors are thus random, but chosen only once and fixed during the construction.) Each node in $\mathcal{T}_b$ may hold one ball; if it does, we say that the node is *full*. Initially all nodes are empty. Balls arrive, one by one, to the root of $\mathcal{T}_b$, and move (instantaneously) according to the following rules.

(i) A ball arriving at an empty node stays there, making the node full.

(ii) A ball arriving at a node v that is already full continues to a child of v; the child is chosen at random, with child i chosen with probability $P_i^{(v)}$. Given the vectors $\mathcal{P}^{(v)}$, all these choices are made independently of each other.

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

The random split tree $T_n = T_n^{\mathcal{P}}$ is the subtree of $\mathcal{T}_b$ consisting of the nodes that contain the first n balls. Note that the parameters apart from n in (this version of) the construction are b and the random b-dimensional probability vector $\mathcal{P}$ (or rather its distribution); $\mathcal{P}$ is called the *split vector*.

Devroye [11] gives several examples of this construction (and also of other instances of his general definition). One of them is the random binary search tree, which is obtained with $b = 2$ and $\mathcal{P} = (U, 1-U)$, with $U \sim U(0,1)$, the uniform distribution on $[0,1]$. The main purpose of the definition of random split trees is that they encompass many different examples of random trees that have been studied separately; the introduction of split trees made it possible to treat them together. Some general results were proved in [11], and further results and examples have been added by other authors; see for example [10, 15].

Devroye [11] considers only finite b, yielding trees T_n where each node has at most b children, but the definition above of random split trees extends to $b = \infty$, when each node can have an unlimited number of children. This is the case that we shall use. (Note that random recursive trees and linear preferential attachment trees with $\chi > 0$ do not have bounded degrees; see Section 6 for the case $\chi < 0$.) Our purpose is to show that with this extension, linear preferential attachment trees are also random split trees.

Remark 3. The general preferential attachment tree is usually considered as an unordered tree. However, it is often convenient to label the children of each node by $1, 2, 3, \dots$ in the order that they appear; hence we can also regard the tree as a ordered tree. Thus both the preferential attachment trees and the split trees considered in the present paper can be regarded as subtrees of the infinite Ulam–Harris(–Neveu) tree $\mathcal{T}_\infty$, which is the infinite rooted ordered tree where every node has a countably infinite set of children, labelled $1, 2, 3, \dots$. (The nodes of $\mathcal{T}_\infty$ are all finite strings $\imath_1 \dots \imath_m \in \mathbb{N}^* := \bigcup_0^\infty \mathbb{N}^m$, with the empty string as the root.)

One advantage of this is that it makes it possible to talk unambiguously about inclusions among the trees. We note that both constructions above yield random sequences of trees $(T_n^{\chi,\rho})_{n=1}^\infty$ and $(T_n^{\mathcal{P}})_{n=1}^\infty$ that are increasing: $T_n^{\chi,\rho} \subset T_{n+1}^{\chi,\rho}$ and $T_n^{\mathcal{P}} \subset T_{n+1}^{\mathcal{P}}$.

Remark 4. The random split tree, on the other hand, is defined as an ordered tree, with the potential children of a node labelled $1, 2, \dots$. Note that these do not have to appear in order: child 2 may appear before child 1, for example.

We can always consider the random split tree as unordered by ignoring the labels. If we do so, any (possibly random) permutation of the random probabilities P_i yields the same unordered split tree. (In particular, if b is finite, then it is natural to permute $(P_i)_1^b$ uniformly at random, thus making all P_i have the same (marginal) distribution [11]. However, we cannot do that when $b = \infty$.)

Using the GEM and Poisson–Dirichlet distributions defined in Section 2, we can state our main result as follows. The proof is given in Section 3, using Kingman's paintbox representation of exchangeable partitions. (Appendix A.2 gives an alternative, but related, argument using exchangeable sequences instead.) In fact, the result can be said to be implicit in [28] and [5]; see *e.g.* [5, Corollary 2.6].

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

Theorem 1.1. *Let (χ,ρ) be as above, and assume $\chi+\rho>0$. Then, provided the trees are regarded as unordered trees, the linear preferential attachment tree $T_n^{\chi,\rho}$ has, for every n, the same distribution as the random split tree $T_n^{\mathcal{P}}$ with $b=\infty$ and*

$$\mathcal{P}\sim \mathrm{GEM}(\chi/(\chi+\rho),\rho/(\chi+\rho)).$$

Further, (re)labelling the children of each node in order of appearance, the sequences $(T_n^{\chi,\rho})_1^\infty$ and $(T_n^{\mathcal{P}})_1^\infty$ of random trees have the same distribution.

The same results also hold if we let $\mathcal{P}$ have the Poisson–Dirichlet distribution $\mathrm{PD}(\chi/(\chi+\rho),\rho/(\chi+\rho))$. □

The result extends to the trivial case $\chi+\rho=0$, with $\mathcal{P}\sim\mathrm{GEM}(0,0)=\mathrm{PD}(0,0)$, that is, $P_1=1$; in this case T_n is a path.

Corollary 1.2. *The sequence of random recursive trees $(T_n)_1^\infty=(T_n^{0,1})_1^\infty$ has the same distribution as the sequence of random split trees $(T_n^{\mathcal{P}})_1^\infty$ with $\mathcal{P}\sim\mathrm{GEM}(0,1)$ or $\mathcal{P}\sim\mathrm{PD}(0,1)$ (as unordered trees).* □

Recall that the split vector $\mathrm{PD}(0,1)$ appearing here also appears as, for example, the asymptotic distribution of the (scaled) sizes of the cycles in a random permutation; see *e.g.* [28, Section 3.1].

Corollary 1.3. *The sequence of standard preferential attachment trees $(T_n)_1^\infty=(T_n^{1,1})_1^\infty$ has the same distribution as the sequence of random split trees $(T_n^{\mathcal{P}})_1^\infty$ with $\mathcal{P}\sim\mathrm{GEM}(1/2,1/2)$ or $\mathcal{P}\sim\mathrm{PD}(1/2,1/2)$ (as unordered trees).* □

Note that in Theorem 1.1 and its corollaries above, it is important that we ignore the original labels, and either regard the trees as unordered, or (re)label the children of each node in order of appearance (see Remark 3); random split trees with the original labelling are different (see Remark 4). In the case $\chi<0$, there is also a version for labelled trees (see Theorem 6.1).

We give an application of Theorem 1.1 in Section 5.

2. Notation

If T is a rooted tree, and v is a node in T, then T^v denotes the subtree of T consisting of v and all its descendants. (Thus T^v is rooted at v.)

A *principal subtree* (also called branch) of T is a subtree T^v where v is a child of the root o of T. Thus the node set $V(T)$ of T is partitioned into $\{o\}$ and the node sets $V(T^{v_i})$ of the principal subtrees.

For a (general) preferential attachment tree, with a given weight sequence $(w_k)_k$, the weight of a node v is $w_{d(v)}$, where $d(v)$ is the outdegree of v. The (total) weight $w(S)$ of a set S of nodes is the sum of the weights of the nodes in S; if T' is a tree, we write $w(T')$ for $w(V(T'))$.

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

The Beta distribution $B(\alpha,\beta)$ is for $\alpha,\beta>0$, as usual, the distribution on $[0,1]$ with density function $cx^{\alpha-1}(1-x)^{\beta-1}$, with the normalization factor $c=\Gamma(\alpha+\beta)/(\Gamma(\alpha)\Gamma(\beta))$. We allow also the limiting cases $B(0,\beta):=\delta_0$ $(\beta>0)$ and $B(\alpha,0):=\delta_1$ $(\alpha>0)$, that is, the distributions of the deterministic variables 0 and 1, respectively.

The GEM distribution $\mathrm{GEM}(\alpha,\theta)$ is the distribution of a random infinite vector of probabilities $(P_i)_1^\infty$ that can be represented as

$$P_i = Z_i\prod_{j=1}^{i-1}(1-Z_j), \quad j\geqslant 1, \tag{2.1}$$

where the Z_j are independent random variables with Beta distributions

$$Z_j \sim B(1-\alpha,\theta+j\alpha). \tag{2.2}$$

Note that (2.1) has the interpretation that $P_1=Z_1$, P_2 is a fraction Z_2 of the remaining probability $1-P_1$, P_3 is a fraction Z_3 of the remainder $1-P_1-P_2=(1-Z_1)(1-Z_2)$, and so on. Here the parameters α and θ are assumed to satisfy $-\infty<\alpha<1$ and $\theta+\alpha\geqslant 0$; furthermore, if $\alpha<0$, then $\theta/|\alpha|$ has to be an integer. (If $\alpha<0$ and $\theta=m|\alpha|$, then $Z_m=1$, and thus (2.1) yields $P_i=0$ for all $i>m$; hence it does not matter that Z_j really is defined only for $j\leqslant m$ in this case.) See further *e.g.* [28, Section 3.2].

The Poisson–Dirichlet distribution $\mathrm{PD}(\alpha,\theta)$ is the distribution of the random infinite vector $(\hat P_i)_1^\infty$ obtained by reordering $(P_i)_1^\infty\sim\mathrm{GEM}(\alpha,\theta)$ in decreasing order.

3. Proof of Theorem 1.1

Lemma 3.1. *With the linear weights* (1.1)*, a tree T with m nodes has total weight $w(T)=(m-1)\chi+m\rho=m(\chi+\rho)-\chi$.*

Proof. Let the nodes have outdegrees $d_1,\dots,d_m$. Then $\sum_{i=1}^m d_i=m-1$, and the weight of the tree is thus

$$w(T)=\sum_{i=1}^m(\chi d_i+\rho)=\chi\sum_{i=1}^m d_i+m\rho=(m-1)\chi+m\rho.$$

□

Lemma 3.2. *Consider the sequence of linear preferential attachment trees $(T_n)_1^\infty=(T_n^{\chi,\rho})_1^\infty$, with the children of the root labelled in order of appearance. Let $N_j(n):=|T_n^j|$, the size of the jth principal subtree of T_n. Then $N_j(n)/n\to P_j$ a.s. as $n\to\infty$, for every $j\geqslant 1$ and some random variables P_j with the distribution* $\mathrm{GEM}(\chi/(\chi+\rho),\rho/(\chi+\rho))$. *(In the trivial case $\chi+\rho=0$, interpret this as* $\mathrm{GEM}(0,0)$.)

Proof. The case $\chi+\rho=0$ is trivial, with $N_1(n)=n-1$ and $P_1=1$. Hence we may assume that $\chi+\rho>0$. Furthermore (see Remark 1), we may and shall, for convenience, assume that

$$\chi+\rho=1. \tag{3.1}$$

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

The lemma now follows from Pitman [28, Theorem 3.2], which is stated for 'the Chinese restaurant with the (α, θ) seating plan', since we may regard the principal subtrees as tables in a Chinese restaurant (ignoring the root), and then the preferential attachment model with (1.1) translates into the (χ, ρ) seating plan as defined in [28]. (See the bijection between recursive trees and permutations in [12, Section 6.1.1], which yields this correspondence; the uniform case treated there is the case $(\chi, \rho) = (0, 1)$, which yields the usual Chinese restaurant process.)

For completeness, we give a direct proof using Pólya urns in Appendix A.1. □

Proof of Theorem 1.1. Recall that in the (general) preferential attachment tree, the parent u of a new node is chosen to be a node v with probability proportional to the current weight $w_{d(v)}$ of the node. We can make this random choice in several steps, by first deciding randomly whether u is the root or not, and if not, which principal subtree it belongs to, making this choice with probabilities proportional to the total weights of these sets of nodes. If u is chosen to be in a subtree T^w, we then continue recursively inside this tree, by deciding randomly whether u is the root of T^w or not, and if not, which principal subtree of T^w it belongs to, again with probabilities proportional to the total weights, and so on.

Consequently, the general preferential attachment tree can be constructed recursively using a stream of new nodes (or balls) similarly to the random split tree, with the following rules.

(i′) A ball arriving at an empty node stays there, making the node full.

(ii′) A ball arriving at a node v that is already full continues to a child of v. The child is chosen at random. If v has d children $v_1, \dots, v_d$, then the ball is passed to child i with probability $cw(T^{v_i})$ for each $i = 1, \dots, m$, and to the new child $m+1$ with probability $cw(v) = c(\chi d + \rho)$, where $c = 1/w(T^v)$ is a positive normalization factor.

Thus both the random split trees and the linear preferential attachment trees can be constructed recursively, and in order to show Theorem 1.1, it suffices to show that the two constructions yield the same result at the root, that is, balls after the first are passed on to the children of the root in the same way in both random trees (provided we ignore the order of the children, or (re)label the children in order of appearance).

Consider the linear preferential attachment tree with the construction above. As in the proof of Lemma 3.2, we may assume that (3.1) holds.

Label the children of the root in order of appearance (see Remark 3). The first ball stays at the root, while all others are passed on; we label each ball after the first by the label of the child of the root that it is passed to. This gives a random sequence $(X_i)_{i=1}^{\infty}$ of labels in $\mathbb{N}$ (where X_i is the label of ball $i+1$, the ith ball that is passed on). By construction, the random sequence $(X_i)_i$ is such that the first 1 appears before the first 2, which comes before the first 3, and so on; we call a finite or infinite sequence $(x_i)_i$ of labels in $\mathbb{N}$ *acceptable* if it has this property.

Let $(x_i)_1^n$ be a finite acceptable sequence of length $n \geqslant 0$, and let n_k be the number of times k appears in the sequence; further, let d_n be the largest label in the sequence, so $n_k \geqslant 1$ if $1 \leqslant k \leqslant d_n$, but $n_k = 0$ if $k > d$. If $(X_i)_1^n = (x_i)_1^n$, then the subtree T^k with label k has

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

n_k nodes, and thus by Lemma 3.1 and our assumption (3.1) weight $n_k(\chi+\rho)-\chi = n_k-\chi$, provided $k \leqslant d_n$, while the root has weight $\chi d_n + \rho$. Hence, by the construction above, noting that the tree has $n+1$ nodes and thus by Lemma 3.1 weight $(n+1)-\chi = n+\rho$,

$$\mathbb{P}(X_{n+1}=k \mid (X_i)_1^n = (x_i)_1^n) = \begin{cases} (n_k-\chi)/(n+\rho) & 1 \leqslant k \leqslant d_n, \\ (d_n\chi+\rho)/(n+\rho) & k = d_n+1. \end{cases} \tag{3.2}$$

It follows by multiplying these probabilities for $n=0$ to $N-1$ and rearranging factors in the numerator (or by induction) that, letting $d := d_N$ and $N_k := n_k$ for $n=N$,

$$\mathbb{P}((X_i)_1^N = (x_i)_1^N) = \frac{\prod_{j=0}^{d-1}(j\chi+\rho)\prod_{k=1}^{d}\prod_{n_k=1}^{N_k-1}(n_k-\chi)}{\prod_{n=0}^{N-1}(n+\rho)}. \tag{3.3}$$

In particular, note that this probability depends on the sequence $(x_i)_1^N$ only through the numbers N_k. Consequently, if $(x_i')_1^N$ is another acceptable sequence that is a permutation of $(x_i)_1^N$, then

$$\mathbb{P}((X_i)_1^N = (x_i)_1^N) = \mathbb{P}((X_i)_1^N = (x_i')_1^N). \tag{3.4}$$

Return to the infinite sequence $(X_i)_1^\infty$. This sequence encodes a partition of $\mathbb{N}$ into the sets $A_j := \{k \in \mathbb{N} : X_k = j\}$, and interpreted in this way, (3.4) says that the random partition $\{A_j\}_j$ of $\mathbb{N}$ is an exchangeable random partition; see *e.g.* [5, Section 2.3.2] or [28, Chapter 2]. (See Appendix A.2 for a version of the argument without using the theory of exchangeable partitions.) By Kingman's paintbox representation theorem [5, 22, 23, 28], any exchangeable random partition of $\mathbb{N}$ can be constructed as follows from some random subprobability vector $(P_i)_1^\infty$, that is, a random vector with $P_i \geqslant 0$ and $\sum_i P_i \leqslant 1$. Let $P_\infty := 1 - \sum_{i<\infty} P_i \geqslant 0$. Let $Y_i \in \mathbb{N} \cup \{\infty\}$ be i.i.d. random variables with the distribution $(P_i)_1^\infty$. Then the equivalence classes are $\{i : Y_i = k\}$ for each $k < \infty$, and the singletons $\{i\}$ for each i with $Y_i = \infty$.

In the present case, Lemma 3.2 shows that every principal subtree T^j satisfies either $|T^j(n)| \to \infty$ as $n \to \infty$, or $T^j(n)$ is empty for all n (when $\chi < 0$ and $\rho = m|\chi|$ with $m < j$). Hence, the equivalence classes defined by $(X_i)_1^\infty$ are either empty or infinite, so there are no singletons. Thus $P_\infty = 0$, and $(P_i)_1^\infty$ is a random probability vector. Moreover, the paintbox construction is precisely what the split tree construction (i)–(ii) does at the root, provided we ignore the labels on the children.

Consequently, the sequence of random split trees $T_n^{\mathcal{P}}$ with this random split vector $\mathcal{P} = (P_i)_1^\infty$ has the same distribution as the sequence $(T_n^{\chi,\rho})_1^\infty$, provided that we ignore the labels of the children, or (equivalently) relabel the children of a node in the split trees by their order of appearance. It remains to identify the split vector $\mathcal{P}$.

Let T_n^j be the principal subtree of the split tree $T_n^{\mathcal{P}}$ whose root is labelled j, and let $N_j(n) := |T_n^j|$. Then, by the law of large numbers, as $n \to \infty$,

$$N_j(n)/n \xrightarrow{\text{a.s.}} P_j, \quad j \geqslant 1. \tag{3.5}$$

Recall that we may permute the probabilities P_i arbitrarily (see Remark 4). Let us relabel the children of the root in their order of appearance, and permute the P_i correspondingly; thus (3.5) still holds. Moreover, we have shown that the tree also can be regarded as a

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

linear preferential attachment tree, and with this labelling of the children, Lemma 3.2 applies. Consequently, (3.5) and Lemma 3.2 yield $(P_i)_1^\infty \sim \mathrm{GEM}(\chi,\rho)$.

Finally, $\mathrm{PD}(\chi,\rho)$ is by definition a permutation of $\mathrm{GEM}(\chi,\rho)$, and thus these two split vectors define random split trees with the same distribution (as unordered trees). □

4. An auxiliary result

In the theory of random split trees, an important role is played by the random variable W defined as a size-biased sample from the split vector $\mathcal{P}$; in other words, we first sample $\mathcal{P}=(P_i)_1^\infty$, then sample $I\in\mathbb{N}$ with the distribution $\mathbb{P}(I=i)=P_i$, and finally let $W:=P_I$. Consequently, for any $r\geqslant 0$,

$$\mathbb{E}\,W^r=\mathbb{E}\sum_i P_iP_i^t=\sum_i \mathbb{E}\,P_i^{t+1}. \tag{4.1}$$

We have a simple result for the distribution of W in our case.

Lemma 4.1. *For the random split tree in Theorem 1.1, $W\sim B(\rho/(\chi+\rho),1)$. Thus W has density function $\gamma x^{\gamma-1}$ on $(0,1)$, where $\gamma=\rho/(\chi+\rho)$.*

Proof. Let X_n be the number of nodes in T_n that are descendants of the first node added after the root. In the split tree $T_n^{\mathcal{P}}$, let I be the label of the subtree containing the first node added after the root. Conditioned on the split vector $\mathcal{P}$ at the root, by definition $\mathbb{P}(I=i\mid\mathcal{P})=P_i$. Furthermore, still conditioned on $\mathcal{P}$, the law of large numbers yields that if $I=i$, then $X_n/n\xrightarrow{\text{a.s.}}P_i$. Hence, $X_n/n\xrightarrow{\text{a.s.}}P_I=W$.

On the other hand, in the preferential attachment tree $T_n^{\chi,\rho}$ with children labelled in order of appearance, the first node after the root always gets label 1 and thus in the notation of Lemma 3.2, $X_n=N_1(n)$. Consequently, Lemma 3.2 implies $X_n/n\xrightarrow{\text{a.s.}}P_1$. Since Theorem 1.1 implies that X_n has the same distribution in the two cases, $W\overset{\text{d}}{=}P_1$. Furthermore, by (2.1)–(2.2), assuming again for simplicity (3.1), $P_1=Z_1\sim B(1-\chi,\chi+\rho)=B(\rho,1)$. □

Thus $W\overset{\text{d}}{=}P_1$ for our GEM distribution. This is only a special case of the general result that rearranging the P_i in size-biased order preserves $\mathrm{GEM}(\alpha,\theta)$ for any pair of parameters; see [28, Section 3.2].

Example 1. By Lemma 4.1 we have $\mathbb{E}\,W=\gamma/(\gamma+1)$, and thus by (4.1)

$$\sum_{i=1}^\infty \mathbb{E}\,P_i^2=\mathbb{E}\,W=\frac{\rho}{\chi+2\rho}. \tag{4.2}$$

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

It is possible to calculate the sum in (4.2) directly, using the definitions (2.1)–(2.2), but the calculation is rather complicated:

$$\begin{aligned}\sum_{i=1}^{\infty}\mathbb{E}\,P_i^2 &= \sum_{i=1}^{\infty}\mathbb{E}\,Z_i^2\prod_{j<i}\mathbb{E}(1-Z_j)^2\\ &= \sum_{i=1}^{\infty}\frac{(1-\alpha)(2-\alpha)\prod_1^{i-1}(\theta+j\alpha)(\theta+1+j\alpha)}{\prod_1^{i}(\theta+1+(j-1)\alpha)(\theta+2+(j-1)\alpha)}\\ &= \frac{(1-\alpha)(2-\alpha)}{(\theta+1)(\theta+2)}\sum_{i=1}^{\infty}\prod_{1}^{i-1}\frac{\theta+j\alpha}{\theta+2+j\alpha}.\end{aligned}\tag{4.3}$$

The last sum can be summed, for example by writing it as a hypergeometric function $F(\theta/\alpha+1,1;(\theta+2)/\alpha+1;1)$ and using Gauss's formula [26, (15.4.20)], leading to (4.2). The proof above seems simpler.

5. An application

Devroye [11] showed general results on the height and insertion depth for split trees, and used them to give results for various examples. The theorems in [11] assume that the split vectors are finite, so the trees have bounded degrees, but they may be extended to the present case, using *e.g.* (for the height) results on branching random walks [6, 7] and methods of [8, 9]. However, for the linear preferential attachment trees, the height and insertion depth are well known by other methods (see *e.g.* [16, 29]); hence we give another application.

For a rooted tree T, let $h(v)$ denote the depth of a node v, that is, its distance to the root. Furthermore, for two nodes v and w, let $v\wedge w$ denote their last common ancestor. We define

$$Y = Y(T) := \sum_{v\neq w} h(v\wedge w),\tag{5.1}$$

summing over all pairs of distinct nodes. For definiteness, we sum over ordered pairs; summing over unordered pairs is the same except for a factor $1/2$. We may modify the definition by including the case $v=w$; this adds the total pathlength which a.s. is of order $O(n\log n)$ (see (5.17) below) and thus does not affect our asymptotic result.

The parameter $Y(T)$ occurs in various contexts. For example, if $\hat{W}(T)$ denotes the Wiener index and $\hat{P}(T)$ the total pathlength of T, then $Y(T)=\hat{W}(T)-(n-1)\hat{P}(T)$; see [17]. Hence, for the random recursive tree and binary search tree considered in [25], the theorems there imply convergence of Y_n/n^2 in distribution. We extend this to convergence a.s., and to all linear preferential attachment trees, with characterizations of the limit distribution Q that are different from the one given in [25].

Theorem 5.1. *Consider random split trees $T_n^{\mathcal{P}}$ of the type defined in the Introduction for some random split vector $\mathcal{P}=(P_i)_1^\infty$, and let $Y_n := Y(T_n^{\mathcal{P}})$ be given by* (5.1). *Assume that with positive probability, $0<P_i<1$ for some i. Then there exists a random variable Q*

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

such that $Y_n/n^2 \xrightarrow{\text{a.s.}} Q$ *as* $n \to \infty$. *Furthermore,* Q *has the representation in* (5.8) *below and satisfies*

$$\mathbb{E}\, Q = \frac{1}{1 - \mathbb{E}\sum_i P_i^2} - 1 < \infty, \tag{5.2}$$

and the distributional fixed point equation

$$Q \overset{\text{d}}{=} \sum_{i=1}^{\infty} P_i^2 (1 + Q^{(i)}), \tag{5.3}$$

with all $Q^{(i)}$ *independent of each other and of* $(P_i)_1^\infty$, *and with* $Q^{(i)} \overset{\text{d}}{=} Q$.

If W *is the size-biased splitting variable defined in Section 4, then also*

$$\mathbb{E}\, Q = \frac{\mathbb{E}\, W}{1 - \mathbb{E}\, W}. \tag{5.4}$$

Higher moments may be calculated from (5.3) or (5.8), with some effort.

Proof. We modify the definition of split trees by never placing a ball in a node; we use rule (ii) for all nodes, and thus each ball travels along an infinite path, chosen randomly with probabilities determined by the split vectors at the visited nodes. Let $X_{k,i}$ be the number of the child chosen by ball k at the ith node it visits, and let $\mathbf{X}_k := (X_{k,i})_{i=1}^\infty$. Label the nodes of $\mathcal{T}_\infty$ by strings in $\mathbb{N}^*$ as in Remark 3. Then the path of ball k is $\emptyset$, $X_{k,1}$, $X_{k,1}X_{k,2}$, ..., visiting the nodes labelled by initial segments of $\mathbf{X}_k$. Note that conditioned on the split vectors $\mathcal{V}^{(v)}$ for all $v \in \mathcal{T}_\infty$, the sequences $\mathbf{X}_k$ are i.i.d. random infinite sequences with the distribution

$$\mathbb{P}(X_{k,j} = i_j,\ 1 \leqslant j \leqslant m) = \prod_{j=1}^{m} P_{i_j}^{(i_1 \cdots i_{j-1})}. \tag{5.5}$$

For two sequences $\mathbf{X}, \mathbf{X}' \in \mathbb{N}^\infty$, let

$$f(\mathbf{X}, \mathbf{X}') := \min\{i : X_i \neq X_i'\} - 1, \tag{5.6}$$

that is, the length of the longest common initial segment. Let v_k be the node in T_n that contains ball k, and note that if neither v_k nor v_ℓ is an ancestor of the other, then $h(v_k \wedge v_\ell) = f(\mathbf{X}_k, \mathbf{X}_\ell)$.

We define, as an approximation of Y_n,

$$\widehat{Y}_n := \sum_{k,\ell \leqslant n,\, k \neq \ell} f(\mathbf{X}_k, \mathbf{X}_\ell) = 2 \sum_{\ell < k \leqslant n} f(\mathbf{X}_k, \mathbf{X}_\ell). \tag{5.7}$$

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

Condition on all split vectors $\mathcal{P}^{(v)}$. Then, using (5.5),

$$\begin{aligned}\mathbb{E}(f(\mathbf{X}_1,\mathbf{X}_2) \mid \{\mathcal{P}^{(v)}, v\in\mathcal{T}_\infty\}) &= \mathbb{E}\sum_{m=1}^{\infty}\sum_{i_1,\dots,i_m\in\mathbb{N}} \mathbf{1}\{X_{1,j}=X_{2,j}=i_j \text{ for } j=1,\dots,m\}\\ &= \sum_{m=1}^{\infty}\sum_{i_1,\dots,i_m\in\mathbb{N}}\Bigl(\prod_{j=1}^{m} P_{i_j}^{(i_1\cdots i_{j-1})}\Bigr)^2 =: Q. \end{aligned} \tag{5.8}$$

Hence, since the split vectors are i.i.d.,

$$\begin{aligned}\mathbb{E}\, f(\mathbf{X}_1,\mathbf{X}_2) = \mathbb{E}\, Q &= \sum_{m=1}^{\infty}\sum_{i_1,\dots,i_m\in\mathbb{N}}\prod_{j=1}^{m}\mathbb{E}\, P_{i_j}^2 = \sum_{m=1}^{\infty}\Bigl(\sum_i \mathbb{E}\, P_i^2\Bigr)^m\\ &= \frac{1}{1-\sum_i \mathbb{E}\, P_i^2} - 1. \end{aligned} \tag{5.9}$$

Since $\sum_i P_i^2 \leqslant \sum_i P_i = 1$, with strict inequality with positive probability, $\mathbb{E}\sum_i P_i^2 < 1$, and thus (5.9) shows that $\mathbb{E}\, f(\mathbf{X}_1,\mathbf{X}_2) < \infty$. Consequently, a.s.,

$$Q = \mathbb{E}(f(\mathbf{X}_1,\mathbf{X}_2) \mid \{\mathcal{P}^{(v)}, v\in\mathcal{T}_\infty\}) < \infty. \tag{5.10}$$

Condition again on all split vectors $\mathcal{P}^{(v)}$. Then the random sequences $\mathbf{X}_k$ are i.i.d., and thus (5.7) is a U-statistic. Hence, we can apply the strong law of large numbers for U-statistics by Hoeffding [14], which shows that a.s.

$$\frac{\widehat{Y}_n}{n(n-1)} \to \mathbb{E}(f(\mathbf{X}_1,\mathbf{X}_2) \mid \{\mathcal{P}^{(v)}, v\in\mathcal{T}_\infty\}) = Q. \tag{5.11}$$

Consequently, also unconditionally,

$$\frac{\widehat{Y}_n}{n(n-1)} \xrightarrow{\text{a.s.}} Q. \tag{5.12}$$

It remains only to prove that $(\widehat{Y}_n - Y_n)/n^2 \xrightarrow{\text{a.s.}} 0$, since we already have shown (5.2), which implies (5.4) by (4.1), and (5.3) follows from the representation (5.8).

As noted above, if $\ell < k$, then $h(v_k \wedge v_\ell) = f(\mathbf{X}_k,\mathbf{X}_\ell)$ except possibly when v_ℓ is an ancestor of v_k; furthermore, in the latter case

$$0 \leqslant h(v_k \wedge v_\ell) \leqslant f(\mathbf{X}_k,\mathbf{X}_\ell). \tag{5.13}$$

Let $H_n := \max\{h(v) : v \in T_n\}$ be the height of $T_n = T_n^{\mathcal{P}}$, and let

$$H_n^* := \max\{f(\mathbf{X}_k,\mathbf{X}_\ell) : \ell < k \leqslant n\}.$$

Since a node v_k has at most H_n ancestors, it follows from (5.13) that, writing $v \prec w$ when v is ancestor of w,

$$0 \leqslant \widehat{Y}_n - Y_n = 2\sum_{k=1}^{n}\sum_{v_l \prec v_k}(f(\mathbf{X}_k,\mathbf{X}_\ell) - h(v_k\wedge v_\ell)) \leqslant 2nH_nH_n^*. \tag{5.14}$$

Furthermore, there is some node v_k with $h(v_k) = H_n$, and if v_ℓ is its parent, then $f(\mathbf{X}_k,\mathbf{X}_\ell) \geqslant H_n - 1$; hence, $H_n \leqslant H_n^* + 1$.

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

Let $m = m_n := \lceil c \log n \rceil$, where $c > 0$ is a constant chosen later. Then, arguing similarly to (5.8)–(5.9),

$$\begin{aligned}\mathbb{P}(f(\mathbf{X}_1, \mathbf{X}_2) \geqslant m \mid \{\mathcal{P}^{(v)}, v \in \mathcal{T}_\infty\}) &= \mathbb{E} \sum_{i_1,\dots,i_m \in \mathbb{N}} \mathbf{1}\{X_{1,j} = X_{2,j} = i_j \text{ for } j = 1,\dots,m\} \\ &= \sum_{i_1,\dots,i_m \in \mathbb{N}} \left(\prod_{j=1}^{m} P_{i_j}^{(i_1 \cdots i_{j-1})} \right)^2 \end{aligned} \tag{5.15}$$

and thus, letting $a := \sum_i \mathbb{E}\, P_i^2 < 1$,

$$\mathbb{P}(f(\mathbf{X}_1, \mathbf{X}_2) \geqslant m) = \sum_{i_1,\dots,i_m \in \mathbb{N}} \prod_{j=1}^{m} \mathbb{E}\, P_{i_j}^2 = a^m. \tag{5.16}$$

By symmetry, we thus have

$$\mathbb{P}(H_n^* \geqslant m) \leqslant \sum_{\ell < k \leqslant n} \mathbb{P}(f(\mathbf{X}_k, \mathbf{X}_\ell) \geqslant m) \leqslant n^2 a^m \leqslant n^2 a^{c \log n} \leqslant n^{-2},$$

provided we choose $c \geqslant 4/|\log a|$. Consequently, by the Borel–Cantelli lemma, a.s. $H_n^* \leqslant m - 1 \leqslant c \log n$ for all large n. Hence, a.s. for all large n,

$$H_n \leqslant H^* + 1 \leqslant c \log n + 1, \tag{5.17}$$

and (5.14) shows that a.s. $\widehat{Y}_n - Y_n = O(n \log^2 n)$. In particular, $(\widehat{Y}_n - Y_n)/n^2 \xrightarrow{\text{a.s.}} 0$, which, as mentioned above, together with (5.12) completes the proof. □

Corollary 5.2. *Let $Y_n := Y(T_n^{\chi,\rho})$ be given by* (5.1) *for the linear preferential attachment tree $T_n^{\chi,\rho}$, and assume $\chi + \rho > 0$. Then $Y_n/n^2 \xrightarrow{\text{a.s.}} Q$ for some random variable Q with*

$$\mathbb{E}\, Q = \frac{\rho}{\chi + \rho}. \tag{5.18}$$

Proof. This is immediate by Theorems 1.1 and 5.1, using (5.4) and (4.2) to obtain (5.18). □

6. The case $\chi < 0$: m-ary increasing trees

In this section we consider the case $\chi < 0$ of linear preferential attachment trees further; as noted above, this case has some special features. By Remark 1, we may assume $\chi = -1$, and then by our assumptions, $\rho > 0$ is necessarily an integer, say $\rho = m \in \mathbb{N}$. As mentioned in Remark 1, the case $m = 1$ is trivial, with $T_n^{-1,1}$ a path, so we are mainly interested in $m \in \{2, 3, \dots\}$.

By (1.1), $w_m = 0$, and thus no node in $T_n^{-1,m}$ will get more than m children. In other words, the trees will all have outdegrees bounded by m. It follows from Lemma 3.2, or directly from (2.1)–(2.2), that if, as in Theorem 1.1,

$$(P_i)_1^\infty \sim \mathrm{GEM}\left(-\frac{1}{m-1}, \frac{m}{m-1}\right),$$

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

then $P_j = 0$ for $j > m$. Consequently, in this case, the split tree can be defined using a finite split vector $(P_j)_1^b$ as in Devroye's original definition (with $b = m$).

Recall that an m-ary tree is a rooted tree where each node has at most m children, and the children are labelled by distinct numbers in $\{1,\dots,m\}$; in other words, a node has m potential children, labelled $1,\dots,m$, although not all of these have to be present. (Potential children that are not nodes are known as external nodes.) The m-ary trees can also be defined as the subtrees of the infinite m-ary tree $\mathcal{T}_m$ that contain the root. Note that m-ary trees are ordered, but that the labelling includes more information than just the order of children (for vertices of degree less than m).

It is natural to regard the trees $T_n^{-1,m}$ as m-ary trees by labelling the children of a node by $1,\dots,m$ in (uniformly) random order. It is then easy to see that the construction above, with $w_k = m - k$ by (1.1), is equivalent to adding each new node at random uniformly over all positions where it may be placed in the infinite tree $\mathcal{T}_m$, that is, by converting a uniformly chosen random external node to a node; see [12, Section 1.3.3]. Regarded in this way, the trees $T_n^{-1,m}$ are called m-ary increasing trees (or m-ary recursive trees). See also [4, Example 1].

Example 2. The case $\chi = -1$, $m = 2$ gives, using the construction above with m-ary (binary) trees and external nodes, the random binary search tree. As mentioned in the Introduction, the binary search tree was one of the original examples of random split trees in [11], with the split vector $(U, 1-U)$ where $U \sim U(0,1)$.

Our Theorem 1.1 also exhibits the binary search tree as a random split tree, but with split vector $(P_1, 1-P_1) \sim \mathrm{GEM}(-1,2)$ and thus, by (2.2), $P_1 = Z_1 \sim B(2,1)$. There is no contradiction, since we consider the trees as unordered in Theorem 1.1, and thus any (possibly random) permutation of the split vector yields the same trees; in this case, it is easily seen that reordering (P_1, P_2) uniformly at random yields $(U, 1-U)$. ($P_1 \sim B(2,1)$ has density $2x$, and $P_2 = 1 - P_1$ thus density $2(1-x)$, leading to a density 1 for a uniformly random choice of one of them.)

There are many other split vectors yielding the same unordered trees. For example, Theorem 1.1 gives $\mathrm{PD}(-1,2)$ as one of them. By definition, $\mathrm{PD}(-1,2)$ is obtained by ordering $\mathrm{GEM}(-1,2)$ in decreasing order; by the discussion above, this is equivalent to ordering $(U, 1-U)$ in decreasing order, and it follows that the split vector $(\hat{P}_1, \hat{P}_2) \sim \mathrm{PD}(-1,2)$ has $\hat{P}_1 \sim U(1/2,1)$ and $\hat{P}_2 = 1 - \hat{P}_1$.

For the binary search tree, Devroye's original symmetric choice $(U, 1-U)$ for the split vector has the advantage that, by symmetry, the random split tree then coincides with the binary search tree also as binary trees.

Remark 5. For $m > 2$, the m-ary increasing tree considered here is not the same as the m-ary search tree; the latter is also a random split tree [11], but not of the simple type studied here.

Example 2 shows that when $m = 2$, we may see the m-ary increasing tree as a random split tree also when regarded as an m-ary tree, and not only as an unordered tree as in Theorem 1.1. We show next that this extends to $m > 2$. Recall that the Dirichlet

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

distribution $\mathrm{Dir}(\alpha_1,\dots,\alpha_m)$ is a distribution of probability vectors $(X_1,\dots,X_m)$, that is, random vectors with $X_i \geqslant 0$ and $\sum_1^m X_i = 1$; the distribution has the density function $cx_1^{\alpha_1-1}\cdots x_m^{\alpha_m-1}\,\mathrm{d}x_1\cdots\mathrm{d}x_{m-1}$ with the normalization factor

$$c = \Gamma(\alpha_1+\cdots+\alpha_m)/\prod_1^m \Gamma(\alpha_i).$$

Theorem 6.1. *Let $m \geqslant 2$. The sequence of m-ary increasing trees $(T_n)_1^\infty = (T_n^{-1,m})_1^\infty$, considered as m-ary trees, has the same distribution as the sequence of random split trees $(T_n^{\mathcal{P}})_1^\infty$ with the split vector*

$$\mathcal{P} = (P_i)_1^m \sim \mathrm{Dir}\left(\frac{1}{m-1},\dots,\frac{1}{m-1}\right).$$

Proof. By Theorem 1.1, the sequence of m-ary increasing trees $(T_n^{-1,m})_n$ has, as unordered trees, the same distribution as the random split trees $(T_n^{\mathcal{P}'})_n$, where

$$\mathcal{P}' = (P_i')_1^\infty \sim \mathrm{GEM}\left(-\frac{1}{m-1},\frac{m}{m-1}\right).$$

As noted above, $P_j' = 0$ for $j > m$, so we may as well use the finite split vector $(P_i')_1^m$. Let $\mathcal{P} = (P_i)_1^m$ be a uniformly random permutation of $(P_i')_1^m$. Then, as sequences of unordered trees, $(T_n^{\mathcal{P}})_n \overset{\mathrm{d}}{=} (T_n^{\mathcal{P}'})_n \overset{\mathrm{d}}{=} (T_n^{-1,m})_n$. Moreover, regarded as m-ary trees, both $(T_n^{\mathcal{P}})_n$ and $(T_n^{-1,m})_n$ are, by symmetry, invariant under random relabellings of the children of each node. Consequently, $(T_n^{\mathcal{P}})_n \overset{\mathrm{d}}{=} (T_n^{-1,m})_n$ also as m-ary trees, as claimed.

It remains to identify the split vector $\mathcal{P}$. The definition as a random permutation of $(P_i')_1^m$ does not seem very convenient; instead we use a variation of the argument in Appendix A.1 for Lemma 3.2. We may assume that $T_n = T_n^{-1,m} = T_n^{\mathcal{P}}$, as m-ary trees, for all $n \geqslant 1$. Let $N_j(n)$ be the number of nodes and $N_j^e(n)$ the number of external nodes in the principal subtree T_n^j (now using the given labelling of the children of the root). It is easy to see that $N_j^e(n) = (m-1)N_j(n)+1$.

Consider first T_n as the random split tree $T_n^{\mathcal{P}}$; then the law of large numbers yields, by conditioning on the split vector $\mathcal{P}$ at the root,

$$N_j(n)/n \xrightarrow{\text{a.s.}} P_j, \quad j = 1,\dots,m. \tag{6.1}$$

Next, consider T_n as the m-ary increasing tree $T_n^{-1,m}$, and regard the external nodes in T_n^j as balls with colour j. Then the external nodes evolve as a Pólya urn with m colours, starting with one ball of each colour and at each round adding $m-1$ balls of the same colour as the drawn one. Then (see *e.g.* [2] or [20, Section 4.7.1]) the vector of proportions $(N_j^e(n)/((m-1)n+1))_{j=1}^m$ of the different colours converges a.s. to a random vector with a symmetric Dirichlet distribution $\mathrm{Dir}(1/(m-1),\dots,1/(m-1))$. Hence the vector $(N_j(n)/n)_j$ converges to the same limit. This combined with (6.1) shows that

$$\mathcal{P} \sim \mathrm{Dir}\left(\frac{1}{m-1},\dots,\frac{1}{m-1}\right). \qquad \square$$

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

Remark 6. If we modify the proof above by considering one N_j at a time, using a sequence of two-colour Pólya urns as in Appendix A.1, we obtain a representation (2.1) of the Dirichlet distributed split vector with

$$Z_j \sim B\left(\frac{1}{m-1}, \frac{m-j}{m-1}\right), \quad j = 1, \dots, m;$$

cf. the similar but different (2.2). (This representation can also be seen directly.)

Remark 7. Broutin, Devroye, McLeish and de la Salle [9] study a general model of random trees that generalizes split trees (with bounded outdegrees) by allowing more general mechanisms to split the nodes (or balls) than the ones considered in the present paper. (The main difference is that the splits are only asymptotically given by a single split vector $\mathcal{V}$.) Their examples include the m-ary increasing tree, and also increasing trees as defined by Bergeron, Flajolet and Salvy [4] with much more general weights, assuming only a finite maximum outdegree m; they show that some properties of such trees asymptotically depend only on m, and in particular that the distribution of subtree sizes $(N_j(n)/n)_1^d$ converges to the Dirichlet distribution $\mathrm{Dir}(1/(m-1), \dots, 1/(m-1))$ seen also in Theorem 6.1 above. (Recall that Theorem 6.1, while for a special case only, is an exact representation for all n and not only an asymptotic result.)

There is no analogue of Theorem 6.1 for $\chi \geqslant 0$, since then the split vector is infinite, and symmetrization is not possible.

Acknowledgement

I thank Cecilia Holmgren for helpful discussions.

Appendix: Two alternative proofs

We give here two alternative arguments, a direct proof of Lemma 3.2 and an alternative version of part of the proof of Theorem 1.1 without using Kingman's theory of exchangeable partitions. We do this both for completeness and because we find the alternative and more direct arguments interesting. (For the proof of Theorem 1.1, it should be noted that the two arguments, although stated using different concepts, are closely related; see the proof of Kingman's paintbox theorem by Aldous [1, §11].)

A.1. A direct proof of Lemma 3.2

We often write N_k for $N_k(n)$.

Consider first the evolution of the first principal subtree T_n^1. Let us colour all nodes in T_n^1 red and all other nodes white. If at some stage there are $r = N_1 \geqslant 1$ red nodes and w white nodes, and thus $n = r + w$ nodes in total, then the total weight R of the red nodes is, using Lemma 3.1,

$$R = w(T_n^1) = r - \chi = N_1 - \chi, \tag{A.1}$$

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

while the total weight of all nodes is $w(T_n) = n - \chi$, and thus the total weight W of the white nodes is

$$W = w(T_n) - w(T_n^1) = (n - \chi) - (r - \chi) = n - r = w. \tag{A.2}$$

By (A.1)–(A.2), adding a new red node increases R by 1, but does not change W, while adding a new white node increases W by 1 but does not change R. Moreover, by definition, the probabilities that the next new node is red or white are proportional to R and W. In other words, the total red and white weights R and W evolve as a Pólya urn with balls of two colours, where a ball is draw at random and replaced together with a new ball of the same colour. (See *e.g.* [13, 30] and, even earlier, [24].) Note that while the classical description of Pólya urns considers the numbers of balls of different colours, and thus implicitly assumes that these are integers, the weights considered here may be arbitrary positive real numbers; however, it has been noted many times that this extension of the original definition does not change the results; see *e.g.* [18, Remark 4.2] and *cf.* [19] for the related case of branching processes.

In our case, the first node is the root, which is white, and the second node is its first child, which is the root of the principal subtree T^1 and thus is red. Hence, the Pólya urn just described starts (at $n = 2$) with $r = w = 1$, and thus by (A.1)–(A.2) $R = 1 - \chi$ and $W = 1$.

It is well known that for a Pólya urn of the type just described (adding one new ball each time, of the same colour as the drawn one), with initial (non-random) values R_0 and W_0 of the weights, the red proportion in the urn, *i.e.* $R/(R + W)$, converges a.s. to a random variable $Z \sim B(R_0, W_0)$. (Convergence in distribution follows easily from the simple exact formula for the distribution of the sequence of the first N draws [24]; convergence a.s. follows by the martingale convergence theorem, or by exchangeability and de Finetti's theorem. See also [20, Sections 4.2 and 6.3.3].) Consequently, in our case, $R/(R + W) \xrightarrow{\text{a.s.}} Z_1 \sim B(1 - \chi, 1)$, and thus by (A.1)–(A.2) $N_1(n)/n \xrightarrow{\text{a.s.}} Z_1 \sim B(1 - \chi, 1)$. Note that this is consistent with (2.2), with $(\alpha, \theta) = (\chi, \rho)$, since we assume (3.1). Furthermore, by definition (2.1), we have $P_1 = Z_1$, and thus $N_1(n)/n \xrightarrow{\text{a.s.}} P_1$.

We next consider N_2, then N_3, and so on. In general, for the kth principal subtree, we suppose by induction that $N_i(n)/n \xrightarrow{\text{a.s.}} P_i$ for $1 \leqslant i < k$, with P_i given by (2.1) for some independent random variables Z_i satisfying (2.2), $i < k$. We now colour all nodes in the principal subtree T_n^k red, all nodes in $T_n^1, \dots, T_n^{k-1}$ black, and the remaining ones white. We then ignore all black nodes, and consider only the (random) times that a new node is added and becomes red or white. Arguing as above, we see that if there are $r = N_k \geqslant 1$ red and w white nodes, then the red and white total weights R and W are given by

$$R = w(T_n^k) = r - \chi = N_k - \chi, \tag{A.3}$$

$$W = w(T_n) - \sum_{i=1}^{k} w(T_n^i) = (n - \chi) - \sum_{i=1}^{k} (N_i - \chi) = w + (k - 1)\chi. \tag{A.4}$$

Moreover, (R, W) evolve as a Pólya urn as soon as there is a red node. When the first red node appears, there is only one white node (the root), since then T^j is empty for $j > k$. Consequently, then $r = w = 1$, and (A.3)–(A.4) show that the Pólya urn now starts with

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

$R = 1 - \chi$ and $W = 1 + (k-1)\chi = k\chi + \rho$. Since the total number of non-black nodes is $n - \sum_{i<k} N_i$, it follows that, as $n \to \infty$,

$$\frac{N_k(n)}{n - \sum_{i<k} N_i(n)} \xrightarrow{\text{a.s.}} Z_k, \tag{A.5}$$

for some random variable $Z_k \sim B(1-\chi, k\chi + \rho)$, again consistent with (2.2). Moreover, this Pólya urn is independent of what happens inside the black subtrees, and thus Z_k is independent of $Z_1, \dots, Z_{k-1}$. We have, by (A.5), the inductive hypothesis and (2.1),

$$\begin{aligned}\frac{N_k(n)}{n} &= \frac{N_k(n)}{n - \sum_{i<k} N_i(n)} \cdot \frac{n - \sum_{i<k} N_i(n)}{n} \\ &\xrightarrow{\text{a.s.}} Z_k \left(1 - \sum_{i<k} P_i\right) = Z_k \prod_{i<k} (1 - Z_i) = P_k.\end{aligned} \tag{A.6}$$

This completes the proof. □

A.2. An alternative argument in the proof of Theorem 1.1

Equation (3.4) shows a kind of limited exchangeability for the infinite sequence $(X_i)_1^\infty$, limited because we only consider acceptable sequences, that is, the first appearance of each label is in the natural order. We eliminate this restriction by a random relabelling of the principal subtrees. Let $(U_i)_1^\infty$ be an i.i.d. sequence of $U(0,1)$ random variables, independent of everything else, and relabel the balls passed to subtree i by U_i. Then the sequence of new labels is $(U_{X_i})_1^\infty$, and it follows from (3.4) and symmetry that this sequence is exchangeable, that is, its distribution is invariant under arbitrary permutations. Hence, by de Finetti's theorem [21, Theorem 11.10], there exists a random probability measure $\mathbf{P}$ on $[0,1]$ such that the conditional distribution of $(U_{X_i})_1^\infty$ given $\mathbf{P}$ a.s. equals the distribution of an i.i.d. sequence of random variables with the distribution $\mathbf{P}$.

As in the proof in Section 3, every principal subtree T^j satisfies by Lemma 3.2 either $|T^j(n)| \to \infty$ as $n \to \infty$, or $T^j(n) = \emptyset$ for all n. Hence, a.s. there exists some (random) index ℓ such that $X_\ell = X_1$, and thus $U_{X_\ell} = U_{X_1}$. It follows that the random measure $\mathbf{P}$ a.s. has no continuous part, so $\mathbf{P} = \sum_{i=1}^\infty P_i \delta_{\xi_i}$, for some random variables $P_i \geqslant 0$ and (distinct) random points $\xi_i \in [0,1]$, with $\sum_i P_i = 1$. (We allow $P_i = 0$, and can thus write $\mathbf{P}$ as an infinite sum even if its support happens to be finite.)

The labels ξ_i serve only to distinguish the subtrees, and we may now relabel again, replacing ξ_i by i. After this relabelling, the sequence (X_i) has become a sequence which conditioned on $\mathcal{P} := (P_i)_1^\infty$ is an i.i.d. sequence with each variable having the distribution $\mathcal{P}$. In other words, up to a (random) permutation of the children, the rules (i′)–(ii′) yield the same result as the split tree rules (i)–(ii) given in the Introduction, using the split vector $\mathcal{P} = (P_i)_1^\infty$.

It remains to identify this split vector, which is done as in Section 3, using (3.5) and Lemma 3.2. □

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

References

[1] Aldous, D. J. (1985) Exchangeability and related topics. In *École d'Été de Probabilités de Saint-Flour XIII, 1983*, Vol. 1117 of Lecture Notes in Mathematics, Springer, pp. 1–198.

[2] Athreya, K. B. (1969) On a characteristic property of Polya's urn. *Studia Sci. Math. Hungar.* **4** 31–35.

[3] Barabási, A.-L. and Albert, R. (1999) Emergence of scaling in random networks. *Science* **286** (5439), 509–512.

[4] Bergeron, F., Flajolet, P. and Salvy, B. (1992) Varieties of increasing trees. In *CAAP '92: 17th Colloquium on Trees in Algebra and Programming*, Vol. 581 of Lecture Notes in Computer Science, Springer, pp. 24–48.

[5] Bertoin, J. (2006) *Random Fragmentation and Coagulation Processes*, Cambridge University Press.

[6] Biggins, J. D. (1976) The first- and last-birth problems for a multitype age-dependent branching process. *Adv. Appl. Probab.* **8** 446–459.

[7] Biggins, J. D. (1977) Chernoff's theorem in the branching random walk. *J. Appl. Probab.* **14** 630–636.

[8] Broutin, N. and Devroye, L. (2006) Large deviations for the weighted height of an extended class of trees. *Algorithmica* **46** 271–297.

[9] Broutin, N., Devroye, L., McLeish, E. and de la Salle, M. (2008) The height of increasing trees. *Random Struct. Alg.* **32** 494–518.

[10] Broutin, N. and Holmgren, C. (2012) The total path length of split trees. *Ann. Appl. Probab.* **22** 1745–1777.

[11] Devroye, L. (1999) Universal limit laws for depths in random trees. *SIAM J. Comput.* **28** 409–432.

[12] Drmota, M. (2009) *Random Trees*, Springer.

[13] Eggenberger, F. and Pólya, G. (1923) Über die Statistik verketteter Vorgänge. *Zeitschrift Angew. Math. Mech.* **3** 279–289.

[14] Hoeffding, W. (1961) The strong law of large numbers for U-statistics. Institute of Statistics, University of North Carolina, Mimeograph series 302. https://repository.lib.ncsu.edu/handle/1840.4/2128

[15] Holmgren, C. (2012) Novel characteristic of split trees by use of renewal theory. *Electron. J. Probab.* **17** # 5.

[16] Holmgren, C. and Janson, S. (2017) Fringe trees, Crump–Mode–Jagers branching processes and m-ary search trees. *Probab. Surveys* **14** 53–154.

[17] Janson, S. (2003) The Wiener index of simply generated random trees. *Random Struct. Alg.* **22** 337–358.

[18] Janson, S. (2004) Functional limit theorems for multitype branching processes and generalized Pólya urns. *Stoch. Process. Appl.* **110** 177–245.

[19] Jiřina, M. (1958) Stochastic branching processes with continuous state space. *Czechoslovak Math. J.* **8** 292–313.

[20] Johnson, N. L. and Kotz, S. (1977) *Urn Models and Their Application: An Approach to Modern Discrete Probability Theory*, Wiley.

[21] Kallenberg, O. (2002) *Foundations of Modern Probability*, second edition, Springer.

[22] Kingman, J. F. C. (1978) The representation of partition structures. *J. London Math. Soc.* (2) **18** 374–380.

[23] Kingman, J. F. C.(1982) The coalescent. *Stochastic Process. Appl.* **13** 235–248.

[24] Markov, A. A. (1917) Sur quelques formules limites du calcul des probabilités (Russian). *Bulletin de l'Académie Impériale des Sciences, Petrograd* **11** 177–186.

[25] Neininger, R. (2002) The Wiener index of random trees. *Combin. Probab. Comput.* **11** 587–597.

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press

[26] Olver, F. W. J., Lozier, D. W., Boisvert, R. F. and Clark, C. W., eds (2010) *NIST Handbook of Mathematical Functions*, Cambridge University Press. Also available as *NIST Digital Library of Mathematical Functions*, http://dlmf.nist.gov/
[27] Panholzer, A. and Prodinger, H. (2007) Level of nodes in increasing trees revisited. *Random Struct. Alg.* **31** 203–226.
[28] Pitman, J. (2006) *Combinatorial Stochastic Processes*, *École d'Été de Probabilités de Saint-Flour XXXII, 2002*, Vol. 1875 of Lecture Notes in Mathematics, Springer.
[29] Pittel, B. (1994) Note on the heights of random recursive trees and random m-ary search trees. *Random Struct. Alg.* **5** 337–347.
[30] Pólya, G. (1930) Sur quelques points de la théorie des probabilités. *Ann. Inst. Poincaré* **1** 117–161.
[31] Szymański, J. (1987) On a nonuniform random recursive tree. *Ann. Discrete Math.* **33** 297–306.

https://doi.org/10.1017/S0963548318000226 Published online by Cambridge University Press