THE CRITICAL BETA-SPLITTING RANDOM TREE III: THE EXCHANGEABLE PARTITION REPRESENTATION AND THE FRINGE TREE

DAVID J. ALDOUS AND SVANTE JANSON

ABSTRACT. In the critical beta-splitting model of a random n-leaf rooted tree, clades are recursively split into sub-clades, and a clade of m leaves is split into sub-clades containing i and $m-i$ leaves with probabilities $\propto 1/(i(m-i))$. Study of structure theory and explicit quantitative aspects of the model is an active research topic. It turns out that many results have several different proofs, and detailed studies of analytic proofs are given in [9] (via analysis of recursions) and [7] (via Mellin transforms). This article describes two core probabilistic methods for studying $n \to \infty$ asymptotics of the basic finite-n-leaf models.

(i) There is a canonical embedding into a continuous-time model, that is a random tree CTCS(n) on n leaves with real-valued edge lengths, and this model turns out to be more convenient to study. The family $(\mathrm{CTCS}(n), n \geq 2)$ is consistent under a "delete random leaf and prune" operation. That leads to an explicit inductive construction (the *growth algorithm*) of $(\mathrm{CTCS}(n), n \geq 2)$ as n increases, and then to a limit structure CTCS(∞) which can be formalized via exchangeable partitions, in some ways analogous to the Brownian continuum random tree.

(ii) There is an explicit description of the limit *fringe distribution* relative to a random leaf, whose graphical representation is essentially the format of the cladogram representation of biological phylogenies.

Keywords: Exchangeable partition, fringe tree, Markov chain, phylogenetic tree, random tree, subordinator.

MSC 60C05; 05C05, 60G09, 92B10

CONTENTS

Date: December 11, 2024.

SJ supported by the Knut and Alice Wallenberg Foundation and the Swedish Research Council.

1. Introduction

This article is part of a broad project [6, 7, 8, 9] studying a certain random tree model. The model is defined by recursively splitting a given set of leaves such that a set of m leaves is split into subsets containing i and $m-i$ leaves with probabilities proportional to $1/i(m-i)$; see Section 2 for details. The model arose [5] as a toy model for phylogenetic trees, designed to mimic the uneven splits observed in real world examples (see Section 5.2). The model turns out to have a rich mathematical structure. There are many questions one can ask and many different proof techniques can be exploited. Indeed several of the key results each have several quite different proofs, a fact which may be of pedagogical interest.

This article provides an introduction to the model, emphasizing connections with previous work, and describes core probabilistic methods for studying $n \to \infty$ asymptotics of the basic finite-n-leaf models. A detailed technical study of some aspects via the analysis of recursions is given in [9], and another detailed technical study of other aspects via Mellin transforms will be given in [7]. In parallel, a document

[6] will be maintained, to summarize known results and provide more heuristics and open problems.

We do find it convenient to adopt the biological term *clade* for the set of leaves in a subtree, that is the elements in a subset somewhere in the splitting process. There is an obvious correspondence between the clades and the total $2n-1$ nodes of the binary tree, where leaves correspond to the n clades of size 1 and internal nodes to the $n-1$ larger clades.

1.1. **Outline of results.**

The most studied aspects of the model have been centered on the CLT for leaf heights, proved by different methods in [9, 27, 30], with further related "height" results in [7, 9]. Basic such results are mentioned in Section 3.1, but this article is essentially independent of those results.

- In Section 2.3 we describe the consistency property (Theorem 2.3) for n-leaf trees and the resulting representation of a limit tree CTCS(∞) via an exchangeable random partition of $\mathbb{N}$.
- For finite n the "number of subclades along a path to a uniform random leaf" is a certain continuous-time Markov chain that we call the *harmonic descent* chain (Section 3.2). The probability of visiting a given state (subclade size) i has an explicit formula $a(i)$ in the $n \to \infty$ limit. This *occupation measure* Theorem 3.1 (Section 3.3) has been proved originally in [8] and then [26].
- This leads in Section 4 to an exact description of the "number of subclades along a path to a uniform random leaf on the infinite boundary" process within CTCS(∞), in terms of a certain subordinator (Theorem 4.5).
- In Section 5 we observe that the "occupation measure" Theorem 3.1 leads to an explicit description (Theorem 5.1) of the asymptotic *fringe tree*, many of whose properties have yet to be investigated. The fringe tree is essentially the way that real-world phylogenies are drawn as *cladograms*, and we illustrate a real example alongside a realization of our model.
- In Section 6 we give a novel third proof of the occupation measure theorem, as a first indication of the power of Mellin transform methods.
- In Section 7.2 we discuss analogies with the Brownian continuum random tree.

2. The critical beta-splitting model of random trees

In this section we give the basic construction of the critical beta-splitting random tree. In fact we will give several different versions: we define a discrete-time version DTCS(n) and a continuous-time version CTCS(n); furthermore, in both cases we define ordered and unordered versions. Moreover, for the unordered version of CTCS(n), we define an explicit growth process (CTCS(n)$, n = 1, 2, 3, \dots$) that constructs CTCS(n) for all n jointly in a natural way by adding leaves one by one.

2.1. **The ordered versions.**

For $m \geq 2$, consider the probability distribution $(q(m,i),\ 1 \leq i \leq m-1)$ constructed to be proportional to $\frac{1}{i(m-i)}$. Explicitly

$$q(m,i) = \frac{m}{2h_{m-1}} \cdot \frac{1}{i(m-i)} = \frac{1}{2h_{m-1}}\Big(\frac{1}{i} + \frac{1}{m-i}\Big), \qquad 1 \leq i \leq m-1, \tag{2.1}$$

where h_{m-1} is the harmonic sum

$$h_{m-1} := \sum_{i=1}^{m-1} \frac{1}{i}. \tag{2.2}$$

Now fix $n \geq 2$. Consider the process of constructing a random binary tree with n leaves, labelled $1, \dots, n$, by recursively splitting the integer interval $[n] := \{1, 2, \dots, n\}$ of leaves as follows. First specify that there is a left edge and a right edge at the root, leading to a left subtree which will have the $\mathcal{L}_n$ leaves $\{1, \dots, \mathcal{L}_n\}$ and a right subtree which will have the $\mathcal{R}_n = n - \mathcal{L}_n$ leaves $\{\mathcal{L}_n + 1, \dots, n\}$, where $\mathcal{L}_n$ (and also $\mathcal{R}_n$, by symmetry) has distribution $q(n, \cdot)$. Recursively, a subinterval with $m \geq 2$ leaves is split into two subintervals of random size from the distribution $q(m, \cdot)$. Continue until reaching intervals of size 1, which are the leaves. This yields a binary tree with the given n leaves; each of the $n-1$ splits corresponds to an internal node (including the root). For completeness, we also include the case $n = 1$, in which there are no splits and the tree just consists of the root. Figure 1 (left) illustrates schematically the construction as interval-splitting.

For our purposes, it will be convenient to draw the binary trees in a non-standard way, shown in Figure 1 (center and right). Instead of drawing two edges from an internal node to its children as usual, we draw one vertical line to a "branchpoint" (representing the split but *not* regarded as a node in the tree) followed by two horizontal lines to the children. We regard the horizontal lines as having length 0; we may (for obvious practical reasons) draw them with arbitrary lengths in the figures, but these lengths have no significance. In Figure 1 (right), the horizontal lines to leaves are drawn with their true length 0.

We regard the splitting process as evolving in time, and consider two versions, one in discrete time and one in continuous time, which we call DTCS(n) and CTCS(n), respectively.[1] In DTCS(n), the root clade splits at time 1, its children at time 2, and so on. In CTCS(n), a clade with $m \geq 2$ leaves is split at rate h_{m-1}, that is after an Exp(h_{m-1}) random time (independent of everything else).[2] In both versions, we start at time 0. Each node is regarded as born at the time the corresponding clade appears; we also regard this time as the height of the node in the tree. Hence, in the DTCS(n), all edges have length 1 and the height equals the usual graph distance to the root; in CTCS(n) the edges have different, random, lengths.

In other words, in our graphical representation of the binary tree, each clade with size > 1 is represented by one vertical line (and conversely); this line thus has length 1 in DTCS(n) and has length Exp(h_{m-1}) for a clade of size m in CTCS(n).

Recall that apart from edge lengths, DTCS(n) and CTCS(n) define the same binary tree; in particular, we can always recover DTCS(n) from CTCS(n) by ignoring the edge lengths. Figure 2 shows a schematic realization of CTCS(20) as a "continuization" of the realization of DTCS(20) in Figure 1. Figure 3 shows an actual

[1] DTCS and CTCS are abbreviations for Discrete Time Critical Splitting and Continuous Time Critical Splitting, for reasons explained in Section 7.1.

[2] Exp(r) denotes a random variable with an exponential distribution with rate r, and thus expectation $1/r$.

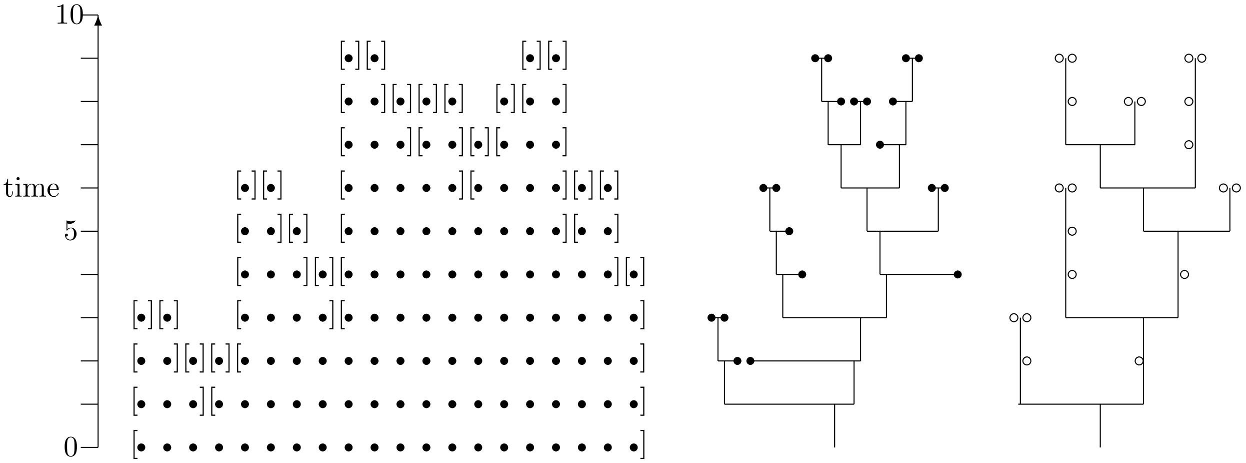


FIGURE 1. Equivalent representations of a realization of DTCS(20).

realization of CTCS(400). Figure 2 shows also, as an example, a distinguished leaf (11); the path from the root to the distinguished leaf 11 passes through successive clades

$$[[1,20]],\ [[4,20]],\ [[5,20]],\ [[9,20]],\ [[9,19]],\ [[9,17]],[[9,13]],\ [[9,11]],\ [[11]]$$

which have successive sizes (number of leaves) $20, 17, 16, 12, 11, 9, 5, 3, 1$.

The choice of rate h_{m-1} in the definition of CTCS(n) may seem arbitrary, but it is justified by the consistency result below (Theorem 2.3, see also (2.4) in its proof), which suggests that h_{m-1} is the canonical choice of splitting rates for the continuization. Note that we, equivalently, can say that a clade of size $m \geq 2$ splits into two clades of sizes i and $m - i$ (taking, as always below, the left subclade first for definiteness) with rate

$$\widehat{q}(m,i) := h_{m-1}q(m,i) = \frac{m}{2i(m-i)} = \frac{1}{2i} + \frac{1}{2(m-i)}, \tag{2.3}$$

for every $1 \leq i \leq m - 1$.

Regarding terminology, remember that "time" and "height"[3] are the same. Within the mathematical analysis of random processes we generally follow the usual "time" convention, while in stating results we generally use the tree-related terminology of "height".

Remark 2.1. In our representations of the trees, we stop at each leaf. It is sometimes advantageous to consider an *extended representation* where we add a vertical line to infinity from each leaf; then every clade is represented by a vertical line, extending from the time the clade is created until it splits (if ever). This is particularly attractive for CTCS(n), since leaves split with rate $0 = h_0$ (i.e., never), and thus the extended representation has for each clade of size m (including leaves) a vertical line of length Exp(h_{m-1}) (interpreted as ∞ when $m = 1$ so the rate $h_{m-1} = 0$) showing the interval of time that the clade lives. In this representation, the leaves

[3] Or depth, if one draws trees upside-down.

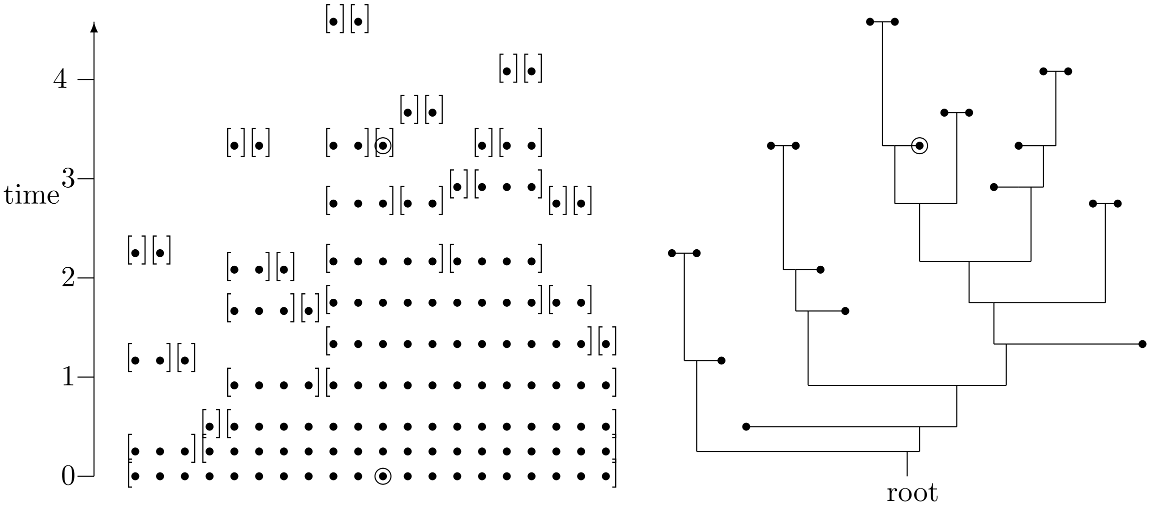


FIGURE 2. Equivalent representations of a realization of CTCS(20). One distinguished leaf is marked.

are located at the bottoms of the infinite lines (i.e., where the lines branch off from other lines, going from there to infinity without further branching).

Note that in the construction above, we have labelled the leaves $1, \dots, n$ from left to right. This is sometimes convenient, but it is often artificial. In particular, the leaves are not equivalent; for example 1 and n are the only leaves that can have height 1, and quantities such as the expectation of the height of leaf i (in either CTCS(n) or DTCS(n)) will depend on i. We call this version of CTCS(n) and DTCS(n) *ordered*; in the next subsection we consider an alternative.

2.2. **The unordered versions.** In the versions of the construction in Section 2.1, the leaves are ordered before we start. Conceptually we are instead usually thinking of recursively splitting a set of objects which have labels (so they are distinguishable) but without any prior structure on the label-set. Without loss of generality, we can still assume that the set of labels is $[n]$, but now, each time a clade of size m is to be split into a left subclade of size i and a right subclade of size $m - i$, we choose the left subclade uniformly at random among all $\binom{m}{i}$ subsets of size i. Otherwise, the construction is exactly as in Section 2.1. This yields the *unordered* versions of DTCS(n) and CTCS(n).

Note that we may obtain the unordered versions from the ordered ones by applying a uniform random permutation to the labels of the leaves. Conversely, we may obtain the ordered versions from the unordered ones by relabelling the vertices in order from left to right. Consequently, any properties of the tree that do not depend on the labels of the leaves are the same for the ordered and unordered versions; two examples are the height of the tree (i.e., the maximum of the heights of the leaves), and, for CTCS, the sum of all edge lengths. Moreover, any properties of the path to a uniform random leaf will have the same distribution for the ordered and unordered version,

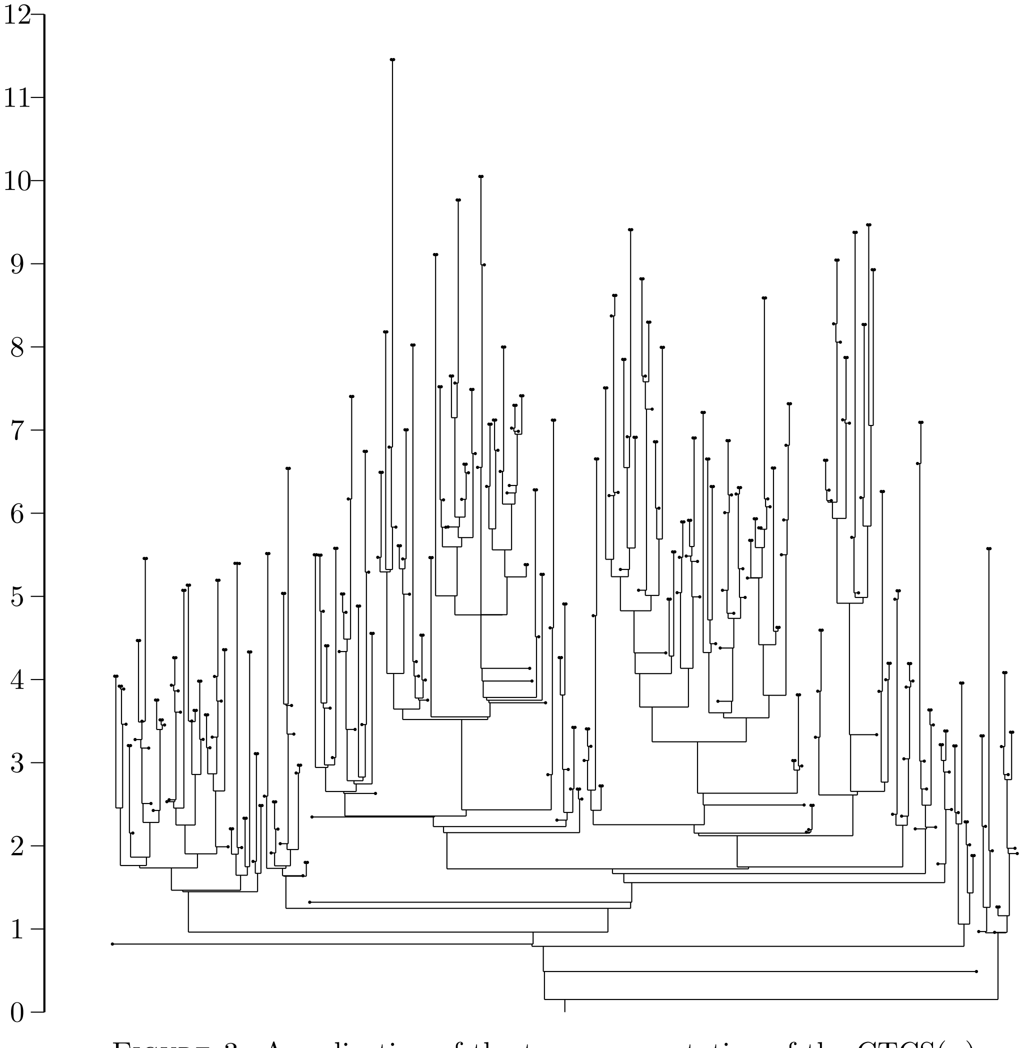
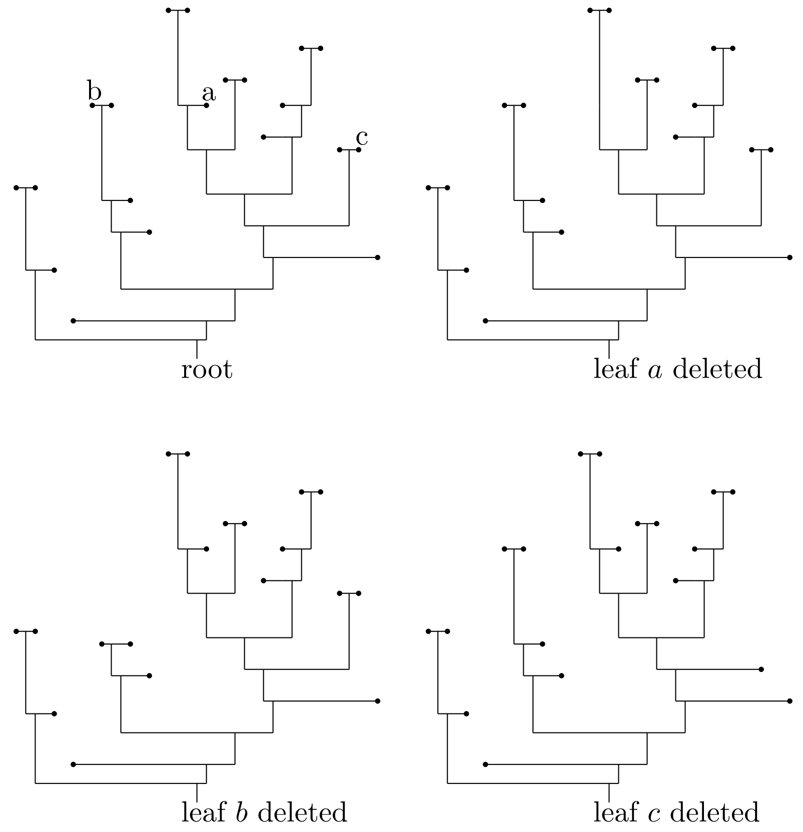


Figure 3. A realization of the tree-representation of the CTCS(n) model with $n = 400$. Drawn as in the previous Figure, so the width of subtrees above a given time level are the sizes of clades at that time.



root

leaf a deleted

leaf b deleted

leaf c deleted

Figure 4. The delete and prune operation: effect of deleting leaf a or b or c from the top left tree.

The unordered versions of DTCS(n) and CTCS(n) have for our purposes the advantage that they (by definition) are invariant under permutation of the leaf-labels. (This of course is a certain type of finite exchangeability.) Hence, for example, the "path to a uniform random leaf" is equivalent (in distribution) to "path to leaf 1". And "delete a uniform random leaf" is equivalent to "delete leaf n". (Recall that this does not hold for the ordered versions.)

Another important advantage of the unordered versions is the consistency property in the next subsection. For this reason, *in the sequel we will always use the unordered versions unless we explicitly say otherwise.*

Remark 2.2. We may also consider *unlabelled versions* where leaves are not labelled. The left/right distinction still matters, and thus the leaves can still be identified by their positions. Hence, the unlabelled versions are equivalent to the unordered versions; we may obtain the unordered versions by randomly labelling the leaves in the unlabelled versions.

2.3. **The consistency property and the growth algorithm.** Observe that all versions of the construction above are for a given n; there is no direct connection between the model (discrete or continuous) for n and the model for $n+1$. Nevertheless, for the unordered versions we have the following *consistency property.*

Note that if we delete a leaf k, then we also have to delete the internal node that is the mother of k (merging the other two edges at that node into one), and if the mother has only one other child, we also have to reduce the height of that child to its grandmother's; see Figure 4. We call this operation "delete and prune leaf k".

Theorem 2.3 (Consistency property)**.** *The operation "delete and prune leaf* $n+1$ *from* $\mathrm{CTCS}(n+1)$*" gives a tree distributed as* $\mathrm{CTCS}(n)$*, and similarly for* $\mathrm{DTCS}(n+1)$ *and* $\mathrm{DTCS}(n)$*.*

This consistency property is in fact a special case of [23, Theorem 1 and Proposition 3], where *all* consistent splitting rules are characterized using the theory of homogeneous fragmentation processes; the connection to such processes will be discussed in Section 4 below. We will give a simple direct proof of Theorem 2.3 in Section 2.4, which furthermore leads to a proof of the growth algorithm in Theorem 2.4 below. An alternative, elementary but longer, proof is given in Appendix A.

Theorem 2.3 implies that if we start at some large N and repeatedly delete and prune the last leaf, we obtain a realization of the sequence $(\mathrm{CTCS}(n))_{n=1}^{N}$ with the correct marginal distributions. By Kolmogorov's extension theorem, there exists an infinite *consistent growth process* $(\mathrm{CTCS}(n), n = 1, 2, 3, \ldots)$ such that, for each n, "delete and prune leaf n from the realization of $\mathrm{CTCS}(n+1)$" gives exactly the realization of $\mathrm{CTCS}(n)$. Conversely, the realization of $\mathrm{CTCS}(n+1)$ is obtained by adding a new leaf $n+1$ to $\mathrm{CTCS}(n)$ at the appropriate place (i.e., at a random place with a specified distribution). It turns out that this addition can be described by the following explicit *growth algorithm.*

In the context of *growth* of trees, it is more evocative to use the word *buds* instead of *leaves*, which we use in the following. In Figure 4 we see *side-buds* such as a, and *bud-pairs* such as b, c.

Theorem 2.4 (The growth algorithm)**.** *Given a realization of* $\mathrm{CTCS}(n)$ *for some* $n \geq 1$*:*

(1) *Pick a uniform random bud; move up the path from the root toward that bud. A* stop *event occurs at rate* $= 1/$*(size of clade from current position).*
(2) *If* stop *before reaching the target bud, make a side-bud at that point, random on left or right.*
(3) *Otherwise, extend the target bud into a branch of* $\mathrm{Exp}(1)$ *length to make a bud-pair.*

Then the result is a realization of $\mathrm{CTCS}(n+1)$*. Consequently, we obtain a realization of the growth process* $\big(\mathrm{CTCS}(n), n = 1, 2, \ldots\big)$ *by starting with* $\mathrm{CTCS}(1)$*, which has a single bud at the root, and then repeating this algorithm ad infinitum.*

The proof is given in Section 2.4; an alternative proof, which also gives explicit formulas for probability densities, is given in Appendix A.

To visualize the growth step, the addition of a new bud can happen in one of three qualitative ways, illustrated in Figure 5, as the reverse of the "cut" in Figure 4. The addition is either what we will call a *side-bud addition* (case **a** in Figure 5) in which a side-bud is attached at the interior of some existing edge, or a *branch extension* (case **b**) in which one bud of a terminal pair grows into a new branch to a terminal pair of buds, or a *side-bud extension* (case **c**) in which a side-bud grows into a new branch with two terminal buds.

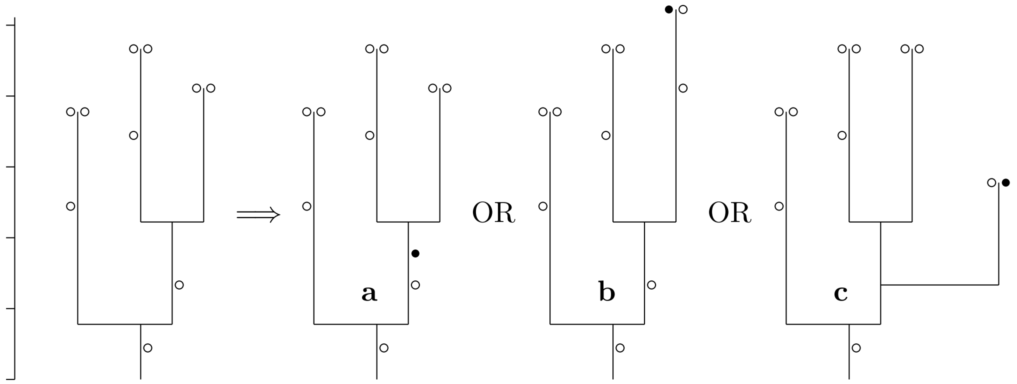


FIGURE 5. Possible transitions from CTCS(10) to CTCS(11): the added bud is •.

Remark 2.5. Using the extended representation in Remark 2.1, the growth algorithm in Theorem 2.4 has an even simpler description:

(1) Pick a uniform random path to infinity (corresponding to a uniform random bud); move up this path from the root toward infinity. A stop event occurs at rate = 1/(size of clade from current position).
(2) At stop, make a side-bud at that point, random on left or right. Add a vertical line from the new bud to infinity. If the current clade at stop had size 1, so stop occurred on the line from some bud to infinity, move also that bud up along the line to the same height as the new bud.

Corollary 2.6. *Let B_n denote the height of the branchpoint between the paths to two uniform random distinct leaves of* CTCS(n). *Then, for each $n \geq 2$, B_n has exactly* Exp(1) *distribution.*

Proof. By exchangeability, B_n has the same distribution as the height of the branchpoint between the paths to leaves 1 and 2 in CTCS(n). If we consider the growth process given by the growth algorithm in Theorem 2.4, then this branchpoint remains the same in CTCS(n) for all $n \geq 2$. (Leaves may move to higher positions, but branchpoints will not move.) Hence, $B_n \overset{\mathrm{d}}{=} B_2$, which by definition has the distribution $\mathrm{Exp}(h_1) = \mathrm{Exp}(1)$. □

Remark 2.7. Note that the growth algorithm is for the continuous-time CTCS(n) only, since it depends on the lengths of the edges. Kolmogorov's extension theorem applies to DTCS(n) too and yields a consistent sequence (DTCS(n)$, n = 1, 2, 3, \ldots$), but this is of less interest since both leaves and internal nodes move towards infinity as new leaves are added, while for CTCS(n), all internal vertices (branchpoints) remain at the same height when new leaves are added, as seen in Corollary 2.6. This happens because in the continuous-time model there is an offsetting feature, that the initial splitting rate h_{n-1} is increasing with n, which remarkably compensates exactly in Corollary 2.6

Remark 2.8. Using the growth process (CTCS(n)$, n = 1, 2, \ldots$), we can define a limiting object CTCS(∞) as the union $\bigcup_{n=1}^{\infty}$ CTCS(n), suitably interpreted. This is perhaps best done with the version in Remark 2.5 with lines to ∞. In that version, by ignoring the buds (which may move when adding a new bud) and considering the lines only, the lines form a type of tree structure that grows with n by adding new lines to ∞ at random branchpoints.

We can regard CTCS(∞) as a (non-compact) real tree – see e.g. [15]. Note that this is not the usual kind of "locally finite" infinite tree[4], because a realization has a countable infinite dense set of branchpoints. We will not study this limit object further; instead we consider a different (though conceptually equivalent) formalization in Section 4.

Here is another viewpoint on the existence of the limit object. For any given buds i and j with $i < j$, the branchpoint B_{ij} between the paths from the root to i and j in CTCS(n) is the same for all $n \geq j$ (its height is Exp(1) by Corollary 2.6). In the growth process above, if we consider only the instances where we happen to move past B_{ij}, and record whether we turn towards i or towards j, then this process can be modelled by a Pólya urn; consequently, the proportion of leaves in each branch converges a.s. to a random non-zero limit.

2.4. Proof of the consistency property and growth algorithm.

Proof of Theorem 2.3. It suffices to consider CTCS(n), since the result for DTCS(n) then follows by ignoring edge-lengths.

Recall from (2.3) that the rate at which a clade of size m splits into two clades of sizes i and $m - i$ is $\widehat{q}(m, i)$ given by (2.3).

Consider CTCS($n + 1$), but kill leaf $n + 1$ and replace it with an invisible ghost. Consider a clade in the tree with m visible elements plus the ghost. This clade really has $m + 1$ elements and thus splits with rate h_m in CTCS($n + 1$), but the two cases when only the ghost is split off from the rest are invisible. A visible split into subsets with j and $m - j$ visible elements may have the ghost in either of the two, and so, taking into account the probability that the ghost appears in the proper subclade, the rate is by (2.3)

$$\begin{aligned}(2.4) \qquad & \frac{j+1}{m+1}\widehat{q}(m+1, j+1) + \frac{m+1-j}{m+1}\widehat{q}(m+1, j) \\ & \qquad = \frac{j+1}{2(j+1)(m-j)} + \frac{m+1-j}{2j(m+1-j)} = \frac{1}{2(m-j)} + \frac{1}{2j} = \widehat{q}(m, i).\end{aligned}$$

[4] Such as a supercritical Galton-Watson tree.

In other words, the ghost does not affect the visible splitting rates. Hence, if we delete and prune leaf $n+1$ from CTCS$(n+1)$, we obtain CTCS(n), which proves Theorem 2.3. □

Proof of Theorem 2.4. We continue to consider CTCS$(n+1)$ with leaf $n+1$ replaced by an invisible ghost. In the argument above, we see from the calculation in (2.4) that if a clade containing the ghost and m visible leaves splits into new clades of visible sizes j and $m-j$, then the ghost will be in the left clade, of size j, with probability

$$\frac{1/(2(m-j))}{\widehat{q}(m,j)} = \frac{j}{m}. \tag{2.5}$$

In other words, the ghost moves as if it accompanies a uniformly chosen visible leaf in the clade. Note also that when the ghost belongs to a clade with m visible elements, it splits off on its own at the rate

$$\frac{2}{m+1}\widehat{q}(m+1,1) = \frac{1}{m}. \tag{2.6}$$

(This follows also because the splitting rate in CTCS$(n+1)$ is h_m, of which the visible splits have rate h_{m-1}; hence the rate of an invisible split is $h_m - h_{m-1}$.) This means that given CTCS(n), the life of the ghost can (up to identity in distribution) be described by: Choose a leaf in CTCS(n) uniformly at random, and follow the branch towards it. With rate 1/(current size of the clade (excluding the ghost)), branch off alone; if the ghost reaches the chosen leaf, continue together with it and branch off from it with the same rate (now 1).

We may thus construct CTCS$(n+1)$ from CTCS(n) by the procedure just described, but giving life to the ghost as leaf $n+1$. This gives precisely the growth algorithm in Theorem 2.4 (or Remark 2.5). □

Remark 2.9 (Alternative proofs)**.** As said above, Theorem 2.3 follows also from general results in [23], but we do not know any analog of Theorem 2.4 in the generality studied there. Before finding the rather “conceptual” proofs above, we found a more pedestrian argument based on explicitly describing the joint distribution of (CTCS$(n+1)$, CTCS(n)). That argument is given in Appendix A. There is also a direct (not using the consistency theorem) proof of the branchpoint result (Corollary 2.6) via stochastic calculus – see Appendix B. Another discussion of exchangeability and consistency of random tree models can be found in [24] but we do not see any direct application to our model.

3. Leaf height and the harmonic descent chain

3.1. **Leaf height.** Before continuing to study a formalization of the limit process CTCS(∞) and its quantitative properties (Section 4.1), let us describe some relevant quantitative work on another aspect of the model, which is *leaf height.* We let D_n be the height of a uniform random leaf ℓ in CTCS(n), and let L_n be the height of a uniform random leaf ℓ in DTCS(n). Equivalently, recalling the relation between DTCS(n) and CTCS(n), D_n is the total length of the path from the root to ℓ in

CTCS(n), while L_n is the *hop-height*, i.e., the number of edges on the path, in either of DTCS(n) or CTCS(n).

Recall that in the unordered versions (which we normally use), D_n and L_n can just as well be defined by taking the path to a fixed leaf $\ell \in [n]$, for example $\ell = 1$.

The limit behavior of both D_n and L_n is studied in great detail in [9], though here we consider only D_n. It is easy to see that $t_n := \mathbb{E}[D_n]$ satisfies the recurrence

$$t_n = \tfrac{1}{h_{n-1}}\Big(1 + \sum_{i=1}^{n-1} \tfrac{t_i}{n-i}\Big);\ n \geq 2 \tag{3.1}$$

with $t_1 = 0$. One can see the first order result $\mathbb{E}[D_n] \sim \frac{6}{\pi^2}\log n$ heuristically by plugging $c \log n$ into the recursion and taking the natural first-order approximation to the right side; the constant c emerges as the inverse of the constant

$$\int_0^1 \tfrac{\log(1/x)}{1-x}\, \mathrm{d}x = \zeta(2) = \tfrac{\pi^2}{6} \tag{3.2}$$

and this heuristic goes back to [5]. It has recently been proved [9, Theorems 1.1 and 1.7]

$$\mathbb{E}[D_n] = \tfrac{1}{\zeta(2)} \log n + O(1), \tag{3.3}$$

$$\mathrm{var}(D_n) = (1 + o(1))\tfrac{2\zeta(3)}{\zeta^3(2)} \log n \tag{3.4}$$

and the corresponding CLT holds for D_n. These and related results (and analogs for L_n) are proved in [9] by detailed analysis of recursions analogous to (3.1); further results will be given in [7]. After the preprint version of [9] was posted, alternative proofs of the CLT have been announced: see [27, 30]. Related weaker results about tree height, that is maximum leaf height, are given in [6, 9].

3.2. **The harmonic descent chain.** We can characterize D_n in an alternate way, as follows. In the discrete construction, the sequence of clade sizes along the path from the root to ℓ is the discrete-time Markov chain, starting in state n, whose transition ($m \to i$) probabilities $q^*(m, i)$ are obtained by size-biasing the $q(m, \cdot)$ distribution; so

$$q^*(m,i) := \tfrac{2i}{m} q(m,i) = \tfrac{1}{h_{m-1}} \cdot \tfrac{1}{m-i}, \qquad 1 \leq i \leq m-1,\ m \geq 2 \tag{3.5}$$

from (2.1). Because the continuous-time CTCS process exits m at rate h_{m-1}, the continuous-time process of clade sizes as one moves at speed 1 along the path is the continuous-time Markov process on states $\{1, 2, 3, \ldots\}$ with transition rates

$$\lambda_{m,i} := \tfrac{1}{m-i}, \qquad 1 \leq i \leq m-1,\ m \geq 2 \tag{3.6}$$

with state 1 absorbing. So D_n is the absorption time for this chain, started at state n. Let us call this the (continuous-time) *harmonic descent* (HD) chain.[5]

The HD chain is relevant to the current article in two ways. First, there is a simple probabilistic heuristic for the behavior of the harmonic descent chain, leading to the

[5] *Descent* is a reminder that the chain is decreasing. Despite its simple form, the HD chain has apparently never been studied before.

approximation (3.7) below. Write $\mathbf{X} = (X_t, t \geq 0)$ for the HD chain with rates (3.6), or $\mathbf{X}^{(n)} = (X_t^{(n)}, t \geq 0)$ for this chain starting with $X_0^{(n)} = n$. The key idea is to study the process $\log \mathbf{X} = (\log X_t,\ t \geq 0)$. By considering its transitions, one quickly sees that, for large n, there should be a good approximation

$$\log X_t^{(n)} \approx \log n - Y_t \text{ while } Y_t < \log n \tag{3.7}$$

where $(Y_t, 0 \leq t < \infty)$ is the subordinator with *Lévy measure* ψ_∞ and corresponding σ-finite density f_∞ on $(0,\infty)$ defined as

$$\psi_\infty[a,\infty) := -\log(1-e^{-a}); \quad f_\infty(a) := \tfrac{e^{-a}}{1-e^{-a}}, \qquad 0 < a < \infty. \tag{3.8}$$

Recall that a *subordinator* [13] is the continuous-time analog of the discrete-time process of partial sums of i.i.d. positive summands: informally

$$\mathbb{P}(Y_{t+dt} - Y_t \in \mathrm{d}a) = f_\infty(a)\,\mathrm{d}a\,\mathrm{d}t. \tag{3.9}$$

Such a subordinator satisfies the law of large numbers

$$t^{-1} Y_t \to \rho \qquad \text{a.s. as } t \to \infty \tag{3.10}$$

where the limit is the mean

$$\rho = \int_0^\infty \psi_\infty[a,\infty)\,\mathrm{d}a = \int_0^\infty -\log(1-e^{-a})\,\mathrm{d}a = \pi^2/6. \tag{3.11}$$

So the approximation (3.7) provides a heuristic explanation of why $\frac{D_n}{\log n} \to 6/\pi^2$, and by the CLT for subordinators one can derive a heuristic for the explicit form (3.4) of the variance. This method can, with some effort, be made into a proof of the CLT – see [6]. But instead of asymptotics of CTCS(n), we shall show in Section 4 that the subordinator arises *exactly* within the limit structure CTCS(∞).

3.3. **The occupation measure.** Here is the second way in which the HD chain is relevant to this article. The chain describes the number of descendant leaves of a node, as one moves at speed 1 along the path from the root to a uniform random leaf. We study the “occupation measure”, that is

$$a(n,i) := \text{probability that the chain started at state } n \text{ is ever in state } i. \tag{3.12}$$

So $a(n,n) = a(n,1) = 1$. To see the relevance of $a(n,i)$ to the tree model, we let $N_n(j)$ be the number of subtrees of CTCS(n) that have j leaves; thus, for $j \geq 2$, $N_n(j)$ is the number of internal nodes of CTCS(n) that have exactly j leaves as descendants. Then, conditioned on CTCS(n), the number of leaves that are in some subtree with i leaves is $iN_n(i)$, and thus the (conditional) probability that a random leaf is in such a subtree is $iN_n(i)/n$. Taking the expectation we find

$$a(n,i) = \frac{i\,\mathbb{E}\,[N_n(i)]}{n} \tag{3.13}$$

and, conversely,

$$\mathbb{E}\,[N_n(i)] = na(n,i)/i. \tag{3.14}$$

It seems very intuitive (but not obvious at a rigorous level) that the limits $a(i) = \lim_{n\to\infty} a(n,i)$ exist. Note that $\sum_{i=2}^n a(n,i)/h_{i-1}$ is just the mean absorption time $\mathbb{E}[D_n]$, so (from (3.3)) we anticipate that, assuming the limits exist,

$$\sum_{i=2}^{n} \tfrac{a(i)}{\log i} \sim \mathbb{E}[D_n] \sim (6/\pi^2)\log n \text{ as } n \to \infty. \tag{3.15}$$

This in turn suggests

$$a(i) \sim \tfrac{6}{\pi^2}\tfrac{\log i}{i} \text{ as } i \to \infty. \tag{3.16}$$

However, there seems no intuitive reason to think there should be some simple formula for the limits $a(i)$. So the following result was surprising to us.

Theorem 3.1 (Occupation measure)**.** *For each $i = 2, 3, \ldots$,*

$$a(i) := \lim_{n\to\infty} a(n,i) = \frac{6h_{i-1}}{\pi^2(i-1)}. \tag{3.17}$$

And $a(1) = 1$.

This is the starting point for our analysis of the *fringe distribution* in Section 5. We currently know 3 quite different proofs of Theorem 3.1.

1. One method [8] (straightforward in outline, though somewhat tedious in detail)[6] is to first prove by coupling that the limits $a(i)$ exist. The limits must satisfy a certain infinite set of equations; the one solution $\frac{6h_{i-1}}{\pi^2(i-1)}$ was found by inspired guesswork. Then check that the solution is unique.

2. Iksanov [26] repeats his method for proving the CLT [27] by exploiting the exact relationship with regenerative composition structures, enabling a shorter derivation of Theorem 3.1 from known results in that theory. This methodology is clearly worth further consideration.

3. In Section 6 we give a third proof, illustrating how to exploit the exchangeable representation of CTCS(∞).

4. The exchangeable partitions representation

In Remark 2.8 we discussed briefly a limiting object CTCS(∞), which formally is a real tree. In this section we will define and study another, related, limiting object, which formally is a nested family $(\Pi(t))_{t\geq 0}$ of partitions of $\mathbb{N}$. This uses an existing formalism via Kingman's theory of exchangeable partitions; a standard reference is [13, Section 2.3] – see also [12] and [36, Chapter 2]. The key feature of this approach is Kingman's *paintbox theorem*, which is stated in our setting in Theorem 4.2 below.

The relation between trees and nested families of partitions has been used at least since [21]. For completeness, we develop it below in detail for our case; we refer also to [23] where this relation is studied in a more general situation. (See also Section 7.2 for further discussion.) The idea is simple: Given a finite tree with edge-lengths and leaves labelled $1, \ldots, n$ we define a partition $\Pi(t)$ of $[n]$ for each $t \geq 0$ by cutting

[6] A simplification of that proof has been found by Luca Pratelli and Pietro Rigo (personal communication).

the tree at time (=height) t; conversely, it is easy to see that, provided there are no vertices with outdegree 1, the tree is determined by this family of partitions. This extends to infinite trees, and for exchangeable infinite trees, such as CTCS(∞), we obtain a family of exchangeable partitions and can employ Kingman's theory.

4.1. **Exchangeable partitions.** Fix a level (time) $t \geq 0$. For each n, the clades of CTCS(n) at time t define a partition $\Pi^{[n]}(t)$ of $[n] := \{1, \dots, n\}$. If we represent the tree CTCS(n) as in Remark 2.1, with lines extending to infinity from each node, then $\Pi^{[n]}(t)$ is the partition obtained by cutting the tree CTCS(n) at level t; that is, i and j are in the same part if and only if the branchpoint separating the paths to leaves i and j has height $> t$.

We use the consistent growth process to define CTCS(n) for all $n \geq 1$, and then these partitions $\Pi^{[n]}(t)$ are consistent and define a partition $\Pi(t)$ of $\mathbb{N} := \{1, 2, \dots\}$ into clades at time t. Explicitly, i and j (with $i, j \in \mathbb{N}$) are in the same part if and only if the branchpoint separating the paths to leaves i and j has height $> t$, in CTCS(n) for any $n \geq \max(i, j)$. In other words, $\Pi(t)$ is the partition of $\mathbb{N}$ into the clades defined by the infinite tree CTCS(∞). Obviously, $\Pi(0)$ is the trivial partition into a single class.

Because each CTCS(n) is exchangeable, $\Pi(t)$ is an exchangeable random partition of $\mathbb{N}$, so we can exploit the theory of exchangeable partitions. Denote the clades at time t, that is the parts of $\Pi(t)$, by $\Pi(t)_1, \Pi(t)_2, \dots$, enumerated in order of the least elements. In particular, the clade of leaf 1 is $\Pi(t)_1$. The clades $\Pi(t)_\ell$ are thus subsets of $\mathbb{N}$, and the clades of CTCS(n) are the sets $\Pi(t)_\ell \cap [n]$ that are non-empty.

Writing $| \cdot |$ for cardinality, it is easy to show the following (proofs of the results stated here are given in Section 4.6).

Lemma 4.1. *Let $t > 0$. Then, a.s., all clades $\Pi(t)_\ell$ are infinite, that is $|\Pi(t)_\ell| = \infty$ for every $\ell \geq 1$.*

Write, for $\ell, n \geq 1$,

$$K_{t,\ell}^{(n)} := \bigl|\Pi(t)_\ell \cap [n]\bigr|; \tag{4.1}$$

the sequence $K_{t,1}^{(n)}, K_{t,2}^{(n)}, \dots$ is thus the sequence of *sizes* of the clades in CTCS(n), extended by 0's to an infinite sequence. Lemma 4.1 shows that for every $t > 0$, $K_{t,\ell}^{(n)} \to \infty$ as $n \to \infty$ for every ℓ. By Kingman's fundamental result [13, Theorem 2.1], the asymptotic proportionate clade sizes, that is the limits

$$P_{t,\ell} := \lim_{n\to\infty} \frac{K_{t,\ell}^{(n)}}{n}, \tag{4.2}$$

exist a.s. for every $\ell \geq 1$, and the random partition $\Pi(t)$ may be reconstructed (in distribution) from the limits $(P_{t,\ell})_\ell$ by Kingman's paintbox construction, which we state as the following theorem. Obviously $P_{0,\ell} = \delta_{1\ell}$.

Theorem 4.2. *Let $t \geq 0$.*

(i) *If $t > 0$, then a.s. each $P_{t,\ell} \in (0, 1)$, and $\sum_\ell P_{t,\ell} = 1$.*

(ii) *Given a realization of* $(P_{t,\ell})_\ell$*, give each integer* $i \in \mathbb{N}$ *a random color* ℓ*, with probability distribution* $(P_{t,\ell})_\ell$*, independently for different* i*. These colors define a random partition of* $\mathbb{N}$*, which has the same distribution as* $\Pi(t)$*.*

Note that the paintbox construction in Theorem 4.2 starts with the limits $P_{t,\ell}$, but gives as the result (in distribution) $\Pi(t)$ and thus also the partition $\Pi^{[n]}(t) = \Pi(t) \cap [n]$ for every finite n.

Regarding CTCS(∞) as a real tree, the process $(P_{t,1}, t \geq 0)$ is the relative size of the subclade at time t, as one moves at speed 1 down the path to a uniform random leaf on the infinite boundary.

4.2. **The homogeneous fragmentation process.** We have in Section 4.1 studied a fixed t; now consider the family of nested partitions $(\Pi(t))_{t\geq 0}$. It is easy to see that this is a *homogeneous fragmentation process* as defined in [13, Definition 3.2]. To verify this, it suffices by [13, Lemma 3.4] to show that $(\Pi^{[n]}(t))_{t\geq 0}$ is a homogeneous fragmentation process for each $n \geq 1$, which follows directly from the definition of CTCS(n).

For $n \geq 1$, the family of nested partitions $(\Pi^{[n]}(t))_{t\geq 0}$ determines when a clade splits in CTCS(n), and how the clade splits, except for which subclade is left and which is right. Hence, the process $(\Pi^{[n]}(t))_{t\geq 0}$ determines CTCS(n) up to the order of the children at each vertex; conversely, CTCS(n) determines the partitions $\Pi^{[n]}(t)$ by definition. Consequently, if we ignore the ordering of children in CTCS(n) (which in any case is uniformly random), the process $(\Pi(t))_{t\geq 0}$ determines the entire growth process $(\text{CTCS}(n))_{n\geq 1}$ and conversely.

The conclusion is that we may regard the homogeneous fragmentation process $(\Pi(t))_{t\geq 0}$ of partitions of $\mathbb{N}$ as another representation of the limit object CTCS(∞). We continue to develop some properties of $(\Pi(t))_{t\geq 0}$; some of them will later be used to study CTCS(n) and DTCS(n).

Remark 4.3. In this paper we start with the concrete definition of DTCS(n) and CTCS(n) in Section 2, and then find explicitly in the present section the corresponding homogeneous fragmentation process $(\Pi(t))$. An alternative approach, suggested to us by Bénédicte Haas and using [23], is to start with a general homogeneous fragmentation process $(\Pi(t))_{t\geq 0}$, which can be defined as in [13] by an erosion coefficient $\mathbf{c}$ and a dislocation measure $\boldsymbol{\nu}$ (see Section 4.5); then the restrictions of the partitions $\Pi(t)$ to $[n]$ correspond to a random tree T_n in continuous time and it is easy to see that the family $(T_n)_n$ of random trees is consistent.[7] Moreover, if we choose the erosion coefficient 0 and the dislocation measure defined by (4.12) in Section 4.5 below, then [23, Theorem 1, (2)] shows (cf. the calculation in (4.13)) that this family of random trees has the correct splitting probabilities $q(m, i)$; furthermore, (4.8) below then shows that the splitting rate is h_{m-1}. Consequently, this constructs DTCS(n) and CTCS(n) starting from the correct homogeneous fragmentation process.

See also Remark 4.6.

[7] One of the results in [23] is the converse: every consistent family of random trees obtained by some splitting rule can be obtained in this way.

4.3. **Self-similarity.** As in Section 4.1, consider the version of the tree CTCS(n) with all branches extended up to ∞ (see Remark 2.1) and cut it at a fixed height t, but now consider also the continuation to higher levels; that is, we consider the tree CTCS(n) restricted to times $u \geq t$, which defines a forest $F_t^{(n)}$. The trees in the forest $F_t^{(n)}$ then correspond to the clades at height t in CTCS(n).

The roots are all at height t, but we may make an obvious time translation so that all roots have height 0.

As n grows, we have the following self-similar behaviour as a consequence of the growth algorithm; this can be seen as a consequence of the fact that the fragmentation process $(\Pi(t))_{t\geq 0}$ is homogeneous (see Section 4.2 and [13, p. 119]), but we give also an elementary direct proof in Section 4.6.

Theorem 4.4. *Let $t \geq 0$ be fixed and let n grow from 1 to ∞. At each increase of n, either one of the trees in $F_t^{(n)}$ gets a new leaf, or a new tree consisting only of a root is added to $F_t^{(n)}$; in either case all other trees in $F_t^{(n)}$ remain unchanged. Moreover, each tree in $F_t^{(n)}$, considered only when it is born or increases in size, grows as a copy of the process* CTCS(n)*, and different trees grow as independent copies.*

4.4. **The subordinator within** CTCS(∞)**.** Let us consider the clade containing a given (or random) node and see how it develops as time increases; by exchangeability, we may consider the clade containing 1.

For given n the process $(K_{t,1}^{(n)}, t \geq 0)$ at (4.1) of the clade size is the harmonic descent (Section 3.2) chain $(X_t^{[n]}, t \geq 0)$ started at state n. We have described informally the approximation (3.7) of this $(K_{t,1}^{(n)}, t \geq 0)$ by the subordinator $(Y_t, 0 \leq t < \infty)$ with Lévy measure ψ_∞ and corresponding σ-finite density f_∞ on $(0, \infty)$ defined in (3.8), which we for convenience repeat:

$$\psi_\infty[a, \infty) := -\log(1 - e^{-a}); \quad f_\infty(a) := \tfrac{e^{-a}}{1-e^{-a}}, \quad 0 < a < \infty. \tag{4.3}$$

The next theorem says that this becomes exact in the $n \to \infty$ limit given by (4.2). We note that by [13, Theorem 3.2], a.s. the limit $P_{t,1}$ in (4.2) exists for all $t \geq 0$ simultaneously.

Theorem 4.5. *Define $Y_t := -\log P_{t,1}$. Then $(Y_t, 0 \leq t < \infty)$ is the subordinator given by (4.3). Moreover, for $t \geq 0$ and complex s with $\Re s > -1$,*

$$\mathbb{E}\,[P_{t,1}^s] = \mathbb{E}\,[e^{-sY_t}] = e^{-t(\psi(s+1)-\psi(1))} \tag{4.4}$$

where $\psi(z) := \Gamma'(z)/\Gamma(z)$ is the digamma function.

We prove Theorem 4.5 in Section 4.6 by calculating moments.

As noted after Theorem 4.2, for finite n the partition of CTCS(n) into clades at a fixed level t can be also described by the limits $P_{t,\ell}$. Similarly, considering only $\ell = 1$ but all $t \geq 0$ simultaneously, the harmonic descent chain describing the size of the first clade can be reconstructed (in distribution) for any finite n from the process $P_{t,1}$, or equivalently from the subordinator Y_t, as shown by Iksanov [26, 27].

4.5. **Jump rates and dislocation measure.** The general theory of homogeneous fragmentation processes in [13, Sections 3.1–3.2] includes several further objects associated with such processes that can be used to study and characterize them. In this subsection we calculate the objects below for the process $(\Pi(t))_{t\geq 0}$. The results of this subsection will not be used in the present paper, but the results are included both for possible future use and to illustrate more aspects of the general theory that apply to our setting.

For $n \in \mathbb{N} \cup \{\infty\}$, let $\mathcal{P}_n$ denote the set of partitions of $[n]$, where $[\infty] := \mathbb{N}$. The trivial partition into a single class is denoted $\mathbf{1}_{[n]}$. Let $\mathcal{P}'_n := \mathcal{P}_n \setminus \{\mathbf{1}_{[n]}\}$. We denote the parts of the partition π by $\pi_1, \pi_2, \dots$, in order of their least elements; the number of parts is $|\pi| \geq 1$. Thus $\mathcal{P}'_n := \{\pi \in \mathcal{P}_n : |\pi| \geq 2\}$.

4.5.1. The *jump rates* q_π [13, p. 121] are defined for (finite) partitions $\pi \in \bigcup_{1\leq n<\infty} \mathcal{P}'_n$ as the jump rates from $\mathbf{1}_{[n]}$ in the (Markov) process $\Pi^{[n]}(t)$. They are thus equal to the rate that the initial clade $[n]$ splits in CTCS(n) according to the partition π. Hence,

$$q_\pi = 0 \qquad \text{if } |\pi| \geq 3, \tag{4.5}$$

while if $|\pi| = 2$, then by (2.3), noting that we now specify the parts as subsets of $[n]$ (and not just their sizes as in (2.3)), and, on the other hand, that we ignore the left/right distinction which gives a factor 2,

$$q_\pi = \frac{2}{\binom{n}{|\pi_1|}} \widehat{q}(n, |\pi_1|) = 2\frac{|\pi_1|!\,|\pi_2|!}{n!} \frac{n}{2|\pi_1||\pi_2|} = \frac{(|\pi_1|-1)!\,(|\pi_2|-1)!}{(n-1)!}. \tag{4.6}$$

4.5.2. The *splitting rate* [13, p. 122] is a (possibly infinite) measure $\boldsymbol{\mu}$ on $\mathcal{P}_\infty$ with $\boldsymbol{\mu}\{\mathbf{1}_{[\infty]}\} = 0$ characterized by

$$\boldsymbol{\mu}\bigl\{\pi' \in \mathcal{P}_\infty : \pi'|_{[n]} = \pi\bigr\} = q_\pi \tag{4.7}$$

for every finite n and every $\pi \in \mathcal{P}'_n$. It follows from (4.5) that $\boldsymbol{\mu}$ is supported on the set $\{\pi \in \mathcal{P}_\infty : |\pi| = 2\}$. We note that (4.5)–(4.7) yield

$$\boldsymbol{\mu}\bigl\{\pi' \in \mathcal{P}_\infty : \pi'|_{[n]} \neq \mathbf{1}_{[n]}\bigr\} = \sum_{\pi\in\mathcal{P}'_n} q_\pi = \sum_{i=1}^{n-1} \widehat{q}(n,i) = h_{n-1}. \tag{4.8}$$

It follows that the total mass $\boldsymbol{\mu}(\mathcal{P}'_\infty) = \infty$.

4.5.3. In general, the splitting rate can be decomposed as a sum of two measures which are determined by the *erosion coefficient* $\mathbf{c}$ and the *dislocation measure* $\boldsymbol{\nu}$, respectively [13, Theorem 3.1 and p. 128]. The erosion coefficient equals the mass $\boldsymbol{\mu}(\boldsymbol{\epsilon}^{(1)})$ of the partition $\boldsymbol{\epsilon}^{(1)} \in \mathcal{P}_\infty$ with two blocks: $\{1\}$ and $\mathbb{N} \setminus \{1\}$. For every $n \geq 1$, we have by (4.7) and (4.6), with $\boldsymbol{\epsilon}^{(1)}_n := \boldsymbol{\epsilon}^{(1)}|_{[n]}$, the partition of $[n]$ into $\{1\}$ and $\{2, \dots, n\}$,

$$\mathbf{c} = \boldsymbol{\mu}(\boldsymbol{\epsilon}^{(1)}) \leq \boldsymbol{\mu}\bigl\{\pi' \in \mathcal{P}_\infty : \pi'|_{[n]} = \boldsymbol{\epsilon}^{(1)}_n\bigr\} = q_{\boldsymbol{\epsilon}^{(1)}_n} = \frac{1}{n-1}. \tag{4.9}$$

Thus, the erosion coefficient $\mathbf{c} = 0$.

4.5.4. The dislocation measure $\boldsymbol{\nu}$ is a (possibly infinite) measure on the space $\mathcal{P}_{\mathfrak{m}}$ of *mass partitions*, where a mass partition $\mathbf{s}$ is an infinite sequence $s_1 \geq s_2 \geq \cdots \geq 0$ such that $\sum_{i=1}^{\infty} s_i \leq 1$ [13, Definition 2.1]. The space $\mathcal{P}_{\mathfrak{m}}$ is a compact metric space, see [13]. Each mass partition $\mathbf{s}$ defines a random partition of $\mathbb{N}$ by the paintbox construction in Theorem 4.2(ii) (with obvious change of notation, and allowing for missing mass if $\sum_i s_i < 1$, see [13, Lemma 2.7]); the distribution of this random partition is denoted $\boldsymbol{\rho}_{\mathbf{s}}$. The dislocation measure $\boldsymbol{\nu}$ is characterized by, assuming for simplicity that $\mathbf{c} = 0$ as in our case, see [13, pp. 126–128],

$$\boldsymbol{\mu} = \int_{\mathcal{P}_{\mathfrak{m}}} \boldsymbol{\rho}_{\mathbf{s}} \,\mathrm{d}\boldsymbol{\nu}(\mathbf{s}). \tag{4.10}$$

Also, or as a consequence of (4.10) and $\boldsymbol{\mu}(\mathbf{1}_{[\infty]}) = 0$, $\boldsymbol{\nu}$ has no mass at the point $(1, 0, 0, \dots) \in \mathcal{P}_{\mathfrak{m}}$. (Note that in Theorem 4.2, we used the paintbox construction for a fixed t; here we use it for the splitting rate $\boldsymbol{\mu}$, which can be seen as a version for infinitesimally small t.)

It is easy to see that if $\mathbf{s}$ has at least 3 non-zero terms, or if $\sum_i s_i < 1$, then $\boldsymbol{\rho}_{\mathbf{s}}$ gives a positive probability to the set of partitions with more than two parts; since $\boldsymbol{\mu}$ gives mass 0 to such partitions, $\boldsymbol{\nu}$ is concentrated on the set of mass partitions $\mathbf{s}_x := (x, 1-x, 0, \dots)$ for $x \in [\frac{1}{2}, 1)$. (We need $x \geq \frac{1}{2}$ since $s_1 \geq s_2$ is assumed.) Given a partition $\pi \in \mathcal{P}_n$ with two parts of sizes i and $n-i$ (with $1 \leq i \leq n-1$), the paintbox construction using $\mathbf{s}_x$ yields a probability, using $\boldsymbol{\rho}_{\mathbf{s}_x}$ to denote also the induced probability distribution on $\mathcal{P}_n$,

$$\boldsymbol{\rho}_{\mathbf{s}_x}(\pi) = x^i(1-x)^{n-i} + x^{n-i}(1-x)^i. \tag{4.11}$$

We claim that $\boldsymbol{\nu}$ is the (infinite) measure on $\mathcal{P}_{\mathfrak{m}}$ obtained as the push-forward by the map $x \mapsto \mathbf{s}_x$ of the measure

$$\mathrm{d}\widetilde{\boldsymbol{\nu}} := \frac{\mathrm{d}x}{x(1-x)} \qquad \text{on } [\tfrac{1}{2}, 1). \tag{4.12}$$

To verify this, it suffices to calculate for a partition $\pi \in \mathcal{P}_n$ as above, using (4.11), (4.12), and (4.6),

$$\begin{aligned} \int_{1/2}^{1} \boldsymbol{\rho}_{\mathbf{s}_x}(\pi)\,\mathrm{d}\widetilde{\boldsymbol{\nu}}(x) &= \int_{1/2}^{1} \Bigl(x^i(1-x)^{n-i} + x^{n-i}(1-x)^i\Bigr)\frac{\mathrm{d}x}{x(1-x)} \\ &= \int_0^1 x^i(1-x)^{n-i}\frac{\mathrm{d}x}{x(1-x)} = \frac{\Gamma(i)\,\Gamma(n-i)}{\Gamma(n)} = q_\pi, \end{aligned} \tag{4.13}$$

which by (4.7) verifies (4.10).

Remark 4.6. The dislocation measure[8] $\widetilde{\boldsymbol{\nu}}$ in (4.12) appears alternatively in the definition of $\mathrm{DTCS}(n)$ in [5]. In fact, [5, Section 4] considers first a general construction of random binary splits. To split a clade with n leaves, the leaves are represented by i.i.d. uniformly distributed random points in $(0,1)$, and then the unit interval is split at a random point X with a given density $f(x)$ in $(0,1)$; we condition on this giving a proper split. The beta-splitting model is defined in [5] for $-1 < \beta < \infty$

[8] In the symmetric version with $x \in (0,1)$.

using this construction with the beta density $f(x) = c_\beta x^\beta (1-x)^\beta$; for $-2 < \beta \leq -1$. Here $f(x)$ is not a probability density, but the calculation of splitting probabilities still makes sense, and defines the model. For $\beta = -1$, this calculation is just (4.13).

As mentioned before, the framework of exchangeable partitions have been used by Haas et al [23, 22] in somewhat similar contexts – see Section 7.2 for further discussion.

4.6. **Proofs.**

Proof of Lemma 4.1. It is easily seen from the growth algorithm that a.s., as n grows to ∞:

1. Infinitely many buds of height $< t$ are added, and thus $\Pi(t)_\ell \neq \emptyset$ for every $\ell \geq 1$, and
2. Once a clade $\Pi^{[n]}(t)_\ell$ is non-empty, new leaves will be added to it an infinite number of times.

The result follows. □

Proof of Theorem 4.2. First, obviously $P_{t,\ell} \in [0,1]$, and $\sum_\ell P_{t,\ell} \leq 1$ by Fatou's lemma. Part (ii) is Kingman's paintbox construction [13, Theorem 12.1], stated for the special case when $\sum_\ell P_{t,\ell} = 1$. This holds a.s. since otherwise the general version of the paintbox construction would imply that $|\Pi(t)_\ell| = 1$ for some ℓ [13, Proposition 2.8(iii)], which is ruled out by Lemma 4.1. □

Proof of Theorem 4.4. Consider the effect on the forest $F_t^{(n)}$ of adding a new leaf by the growth algorithm. We have the following cases:

(i) If the algorithm stops at height $u < t$, then CTCS(n) gets a new leaf (bud) there, which means that $F_t^{(n)}$ gets a new tree consisting of a root only.

(ii) If the target leaf has height $\geq t$ and the algorithm does not stop before reaching height t, then the algorithm will continue in the tree containing the target exactly as it would if acting on this tree separately. All other trees in $F_t^{(n)}$ remain unchanged. Note also that the probability of reaching height t is the same for all target leaves in a given tree in $F_t^{(n)}$; hence the conditional distribution of the target leaf, given the tree in $F_t^{(n)}$ that it belongs to, is uniform.

(iii) If the target leaf has height $< t$ and the algorithm does not stop until reaching the target, then the target leaf is extended into a branch of Exp(1) length L ending with a bud-pair.

If $u + L < t$, then the two buds in the pair define separate singleton trees in $F_t^{(n+1)}$, and thus the net effect is to add a new tree consisting of a root only to $F_t^{(n)}$.

On the other hand, if $u + L \geq t$, then the tree in $F_t^{(n)}$ consisting of the target leaf (only) becomes a tree with two leaves at the end of a branch of length $L + u - t$. Since the exponential distribution has no memory, also this branch length has Exp(1) (conditional) distribution, and thus this tree has the distribution of CTCS(2).

All cases conform to the description in the statement. □

Proof of Theorem 4.5. This is, apart from the explicit formula (4.4), an instance of [13, Theorem 3.2]. Nevertheless, we find it instructive to give an explicit proof, partly using the same arguments as [13]. We prove the theorem in 3 steps.

Step 1. Y_t is a subordinator. Recall that $K^{(n)}_{t,1}$ is the size of the first clade of CTCS(n) at time t. Consider two fixed times t and $t+h$, where $h>0$. Then Theorem 4.4 and (4.2) imply that a.s., as $n\to\infty$,

$$\frac{K^{(n)}_{t+h,1}}{K^{(n)}_{t,1}} \to P'_{h,1}, \tag{4.14}$$

where $P'_{h,1}$ is a copy of $P_{h,1}$ that is independent of $P_{t,1}$. Consequently, a.s.

$$\frac{K^{(n)}_{t+h,1}}{n} = \frac{K^{(n)}_{t+h,1}}{K^{(n)}_{t,1}} \cdot \frac{K^{(n)}_{t,1}}{n} \to P'_{h,1}P_{t,1} \tag{4.15}$$

and thus

$$P_{t+h,1} = P_{t,1}P'_{h,1}. \tag{4.16}$$

Hence, $Y_t := -\log P_{t,1}$ is an increasing stochastic process with stationary independent increments, i.e., a subordinator. Note that $Y_t<\infty$ a.s. since $P_{t,1}>0$ by Theorem 4.2.

Step 2. The Lévy measure is given by (4.3)*.* In order to verify this, we calculate moments. Let $k\geq 0$. By the paintbox construction in Theorem 4.2,

$$\mathbb{P}\bigl(\Pi(t)_1\cap[k+1]=[k+1] \mid (P_{t,\ell})_{\ell=1}^{\infty}\bigr) = \sum_{\ell=1}^{\infty} P_{t,\ell}^{k+1}. \tag{4.17}$$

Furthermore, also as a consequence of the paintbox construction, $P_{t,1}$ has the same distribution as a size-biased sample of $(P_{t,\ell})_{\ell=1}^{\infty}$ [13, Proposition 2.8], and thus [13, Corollary 2.4]

$$\mathbb{E}\,[P_{t,1}^{k}] = \mathbb{E}\Bigl[\sum_{\ell=1}^{\infty} P_{t,\ell}^{k+1}\Bigr]. \tag{4.18}$$

Consequently, (4.18) and (4.17) together with the definition of $\Pi(t)_1$ in Section 4.1 yield

$$\begin{aligned}\mathbb{E}\,[P_{t,1}^{k}] &= \mathbb{P}\bigl(\Pi(t)_1\cap[k+1]=[k+1]\bigr)\\ &= \mathbb{P}\bigl(2,3,\dots,k+1\in\Pi(t)_1\bigr)\\ &= \mathbb{P}\bigl(\text{CTCS}(k+1)\text{ has no branchpoint with height }\leq t\bigr).\end{aligned} \tag{4.19}$$

The latter event occurs if and only if for each $j\leq k$, in the inductive construction by the growth algorithm of CTCS($j+1$) from CTCS(j), there is no stop at height $\leq t$. Since the subclade at the current position then has size j for all times $\leq t$, the probability of this happening at step j, given that it has happened so far, is $\exp(-t/j)$. Consequently,

$$\mathbb{E}\,[P_{t,1}^k] = \prod_{j=1}^k e^{-t/j} = \exp\Bigl(-t\sum_{j=1}^k \frac{1}{j}\Bigr) = e^{-h_k t}. \tag{4.20}$$

On the other hand, let $\mathcal{Y}_t$ be the subordinator with Lévy measure given by (4.3). Then, by definition, for any real $s \geq 0$,

$$\mathbb{E}\,[e^{-s\mathcal{Y}_t}] = \exp\Bigl(-t\int_0^\infty (1-e^{-sx})f_\infty(x)\,\mathrm{d}x\Bigr) = \exp\Bigl(-t\int_0^\infty \frac{1-e^{-sx}}{1-e^{-x}}e^{-x}\,\mathrm{d}x\Bigr). \tag{4.21}$$

In particular, if $s = k$ is an integer,

$$\begin{aligned}\mathbb{E}\,[\bigl(e^{-\mathcal{Y}_t}\bigr)^k] = \mathbb{E}\,[e^{-k\mathcal{Y}_t}] &= \exp\Bigl(-t\int_0^\infty \frac{e^{-x}-e^{-(k+1)x}}{1-e^{-x}}\,\mathrm{d}x\Bigr) = \exp\Bigl(-t\int_0^\infty \sum_{j=1}^k e^{-jx}\,\mathrm{d}x\Bigr) \\ &= \exp\Bigl(-t\sum_{j=1}^k \frac{1}{j}\Bigr) = e^{-h_k t}.\end{aligned} \tag{4.22}$$

Consequently, for any $t \geq 0$, (4.20) and (4.22) show that $\mathbb{E}\,[P_{t,1}^k] = \mathbb{E}\,[\bigl(e^{-\mathcal{Y}_t}\bigr)^k]$ for all $k \geq 1$, and thus, by the method of moments [9] $P_{t,1} \overset{\mathrm{d}}{=} e^{-\mathcal{Y}_t}$. Thus $Y_t = -\log P_{t,1} \overset{\mathrm{d}}{=} \mathcal{Y}_t$.

This calculation is for a fixed $t \geq 0$, but we know that the process (Y_t) is a subordinator, and thus the distribution of the entire process is determined by, say, the distribution of Y_1.

Step 3. The moment formula. For any complex s with $\Re s > -1$, we have [35, 5.9.16] (as is easily verified by standard arguments)

$$\int_0^\infty \frac{1-e^{-sx}}{1-e^{-x}}e^{-x}\,\mathrm{d}x = \int_0^\infty \frac{e^{-x}-e^{-(s+1)x}}{1-e^{-x}}\,\mathrm{d}x = \psi(s+1)-\psi(1), \tag{4.23}$$

generalizing the formula for integer s in (4.22). Hence, (4.21) yields, for $t \geq 0$ and $\Re s > -1$,

$$\mathbb{E}\,[P_{t,1}^s] = \mathbb{E}\,[e^{-sY_t}] = \mathbb{E}\,[e^{-s\mathcal{Y}_t}] = e^{-t(\psi(s+1)-\psi(1))}, \tag{4.24}$$

which shows (4.4) and completes the proof of the theorem. □

Remark 4.7. [13, Theorem 3.2] gives a general formula relating the moment and Laplace transform in (4.4) to the dislocation measure in (4.12). This can be used to show (4.4), although we preferred above a calculation using the growth algorithm; conversely, using [13, footnote on p. 135], this formula can be used to show (4.12) from (4.4).

[9] The random variables on both sides are bounded, with values in $[0,1]$.

5. The occupation measure and the fringe process

5.1. **The (limit) fringe tree.** To be consistent with the *cladogram* representation described below, we work here in the discrete time DTCS(n) setting: the definition (3.12) of $a(n,i)$ is of course unchanged in discrete time.

The motivation for Theorem 3.1 involves the (asymptotic) *fringe tree* for the random tree model DTCS(n), that is the $n \to \infty$ local weak limit of the tree relative to a typical leaf. See [2, 25, 29] for general accounts of fringe trees, which for us[10] are random locally finite trees with a distinguished leaf. It will be straightforward to verify that the fringe tree can be described in terms of the limits $(a(i), i \geq 1)$ as follows.

Theorem 5.1. (a) *The sequence of clade sizes as one moves away from the distinguished leaf is the discrete time "reverse HD" Markov chain started at state* 1*, whose "upward" transition probabilities $q^{\uparrow}(i,j)$ are given by*

$$q^{\uparrow}(i,j) = \frac{a(j)}{a(i)} q^*(j,i), \tag{5.1}$$

which, from the explicit formula (3.17) *for $a(i)$, becomes*

$$q^{\uparrow}(1,j) = 6\pi^{-2} \tfrac{1}{(j-1)(j-1)}, \qquad j \geq 2 \tag{5.2}$$

$$q^{\uparrow}(i,j) = \tfrac{i-1}{(j-1)(j-i)h_{i-1}}, \qquad 2 \leq i < j. \tag{5.3}$$

(b) *At each such upward step $i \to j$, there is the sibling clade of size $j-i$, and this clade is distributed as* DTCS$(j-i)$*, independently for each step. This sibling clade is randomly on the left or right side.*

One can check that (5.3) is a *probability* distribution by observing

$$\sum_{j>i} \tfrac{1}{(j-1)(j-i)} = \sum_{j>i} \tfrac{1}{i-1}\left(\tfrac{1}{j-i} - \tfrac{1}{j-1}\right) = \tfrac{h_{i-1}}{i-1}. \tag{5.4}$$

Proof. (a): This is a simple exercise in reversing a Markov chain. Let $1 = k_1 < k_2 < \cdots < k_\ell$ be a finite sequence of integers. If $n \geq k_\ell$, then the probability that the HD chain started at n ends with $k_\ell, k_{\ell-1}, \ldots, k_1 = 1$ is, by the definition of $a(n,i)$ and recalling the transition probabilities $q^*(m,i)$ of the HD chain in (3.5),

$$a(n,k_\ell) \prod_{i=1}^{\ell-1} q^*(k_{\ell+1-i}, k_{\ell-i}) = \prod_{j=1}^{\ell-1} \frac{a(n,k_{j+1})}{a(n,k_j)} q^*(k_{j+1}, k_j). \tag{5.5}$$

So the reverse chain is a Markov chain with transition probabilities

$$q_n^{\uparrow}(i,j) = \frac{a(n,j)}{a(n,i)} q^*(j,i), \qquad i < j \leq n. \tag{5.6}$$

Taking the limit as $n \to \infty$, we obtain by (3.17) and (3.5) the transition probabilities (5.1)–(5.3).

(b): Obvious. □

[10] The general accounts take limits relative to a random *node*, but for our leaf-labelled trees it is more natural to use leaves. In the terminology of [2, 25] these are *extended* fringe trees.

Remark 5.2. Using (3.5), we can also write (5.6) as

$$i^{-1}a(n,i)q^{\uparrow}(i,j) = 2j^{-1}a(n,j)q(j,i) = j^{-1}a(n,j)(q(j,i)+q(j,j-i)) \tag{5.7}$$

where the two sides both are

$$n^{-1}\,\mathbb{E}\,[\text{number of splits } j \to (i,j-i) \text{ or } (j-i,i) \text{ in DTCS}(n)], \tag{5.8}$$

calculated in the two different directions.

5.2. **Fringe trees and cladograms.** As mentioned before, the model arose as a toy model for phylogenetic trees, designed to mimic the uneven splits observed in real world examples. The small-scale study [10] suggests that in splits $m \to (i, m-i)$ in real-world phylogenetic trees, the median size of the smaller subtree scales roughly as $m^{1/2}$. That data is not consistent with more classical random tree models, where the median size would be $O(\log m)$ or $\Theta(m)$, but this $m^{1/2}$ median property does hold for our particular model. Figure 6 compares a simulation of DTCS(77) with a real cladogram on 77 species; these appear visually quite similar. As shown in that figure, a cladogram is typically drawn upwards from the leaves, and we draw the fringe tree in the same way. That is, one should visualize a fringe tree[11] as in Figure 6 (top), but with leaves labelled as $\ldots, -2, -1, -0, 1, 2, \ldots$.

5.3. **Properties of the fringe tree.** For large n, a realization of DTCS(n) will contain many copies of small clades in its fringe. In the asymptotic fringe tree, the probability that a given leaf is in *some* clade χ of size i is just $a(i)$. Because a clade of size i has the DTCS(i) distribution, we can then calculate (numerically via recursion) the probability $p(\chi)$ that a leaf is in a *specific clade* χ of size i. Some numerical results are shown in Figure 7. In that figure we have grouped clades with the same *shape*, meaning that (as in the biology use) we do not distinguish left and right branches. Figure 7 compares these model predictions with the data from a small set of real cladograms[12] – 10 cladograms with a total of 995 species. Further data will be given in [6], and it is clear that the current model gives a better fit than other models such as those in [29, Appendix A] and [28]. Note that the models treated in [29] are precisely the cases $\beta = \infty, 0, -3/2$ of the beta-splitting tree [5].

But also one can use the fringe tree to study asymptotics of statistics of DTCS(n) or CTCS(n), for statistics which depend only on the structure of the tree near the leaves. In particular, the number $N_n(\chi)$ of copies of a size-i clade χ in DTCS(n) will satisfy $n^{-1}\,\mathbb{E}\,[N_n(\chi)] \to p(\chi)/i$. By analogy with CLT results for other random tree models [25, section 14], and because occurrences of a given χ are only locally dependent, we expect a CLT for $N_n(\chi)$, but have not attempted a proof.

[11] There is no biological significance to the positioning of left/right branches, though a common convention is to position the larger subclade to the right. In our model, branches are randomly positioned left/right, but in drawing Figure 6 (top) we followed the biological convention for visual comparison.

[12] Dragonflies [33], eagles [32], elms [40], gamebirds [14], ladybirds [34], parrots[41], primates [20], sharks [39], snakes [16], swallows [37]

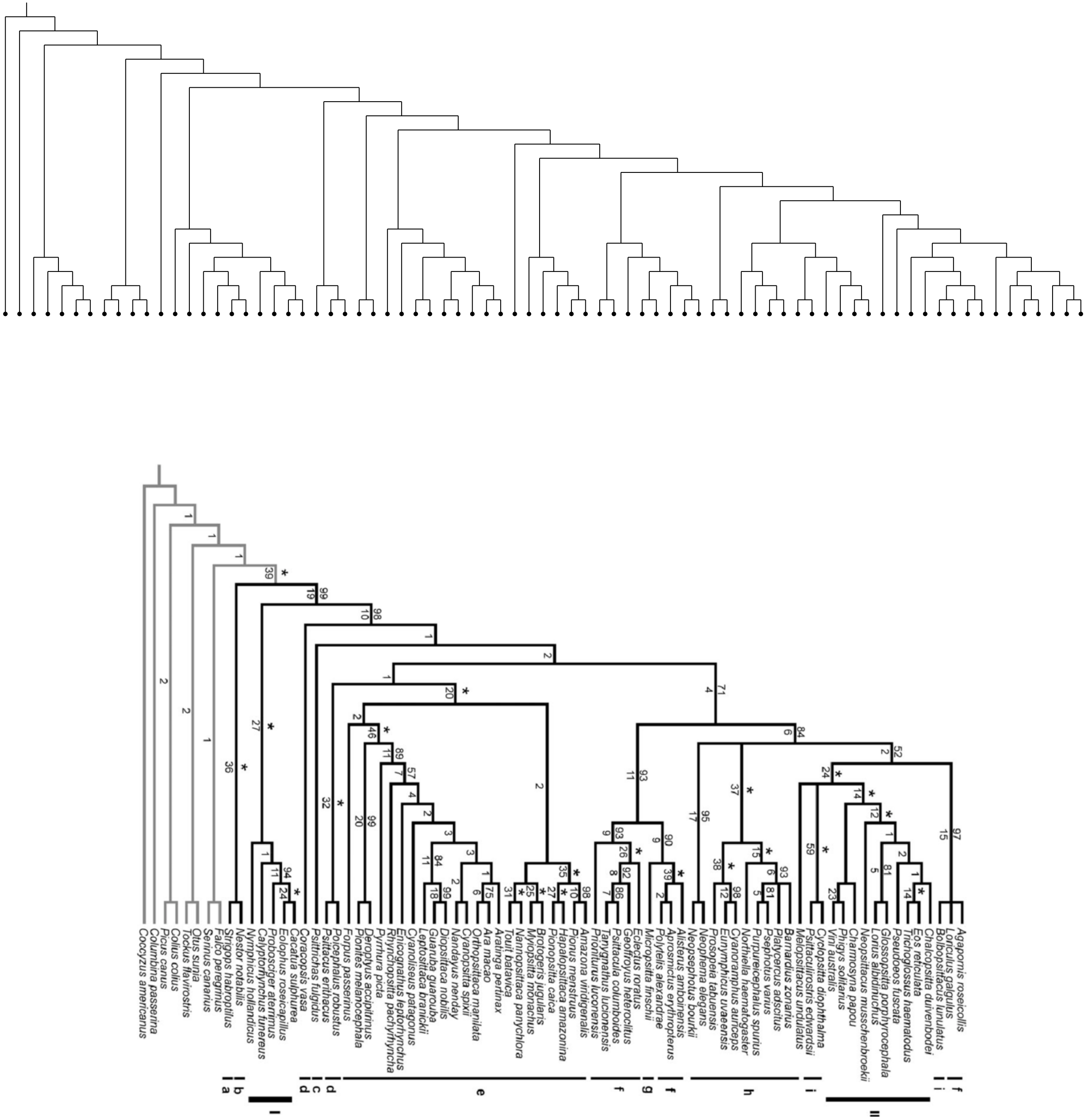


Figure 6. Bottom: cladogram showing phylogenetics of 77 parrot species, from [41]. Top: simulation of DTCS(77), drawn as fringe distribution in the style of biological cladograms.

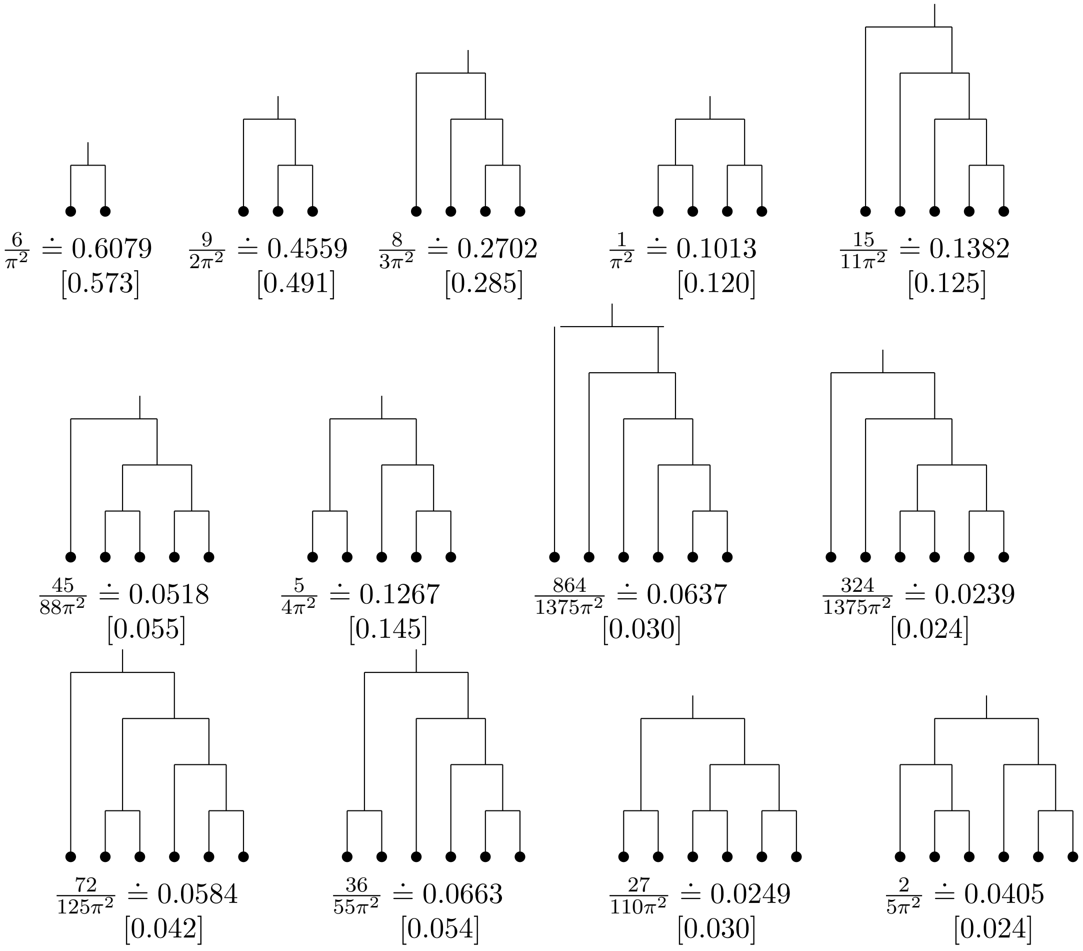

$\frac{6}{\pi^2} \doteq 0.6079$
[0.573]
$\frac{9}{2\pi^2} \doteq 0.4559$
[0.491]
$\frac{8}{3\pi^2} \doteq 0.2702$
[0.285]
$\frac{1}{\pi^2} \doteq 0.1013$
[0.120]
$\frac{15}{11\pi^2} \doteq 0.1382$
[0.125]
$\frac{45}{88\pi^2} \doteq 0.0518$
[0.055]
$\frac{5}{4\pi^2} \doteq 0.1267$
[0.145]
$\frac{864}{1375\pi^2} \doteq 0.0637$
[0.030]
$\frac{324}{1375\pi^2} \doteq 0.0239$
[0.024]
$\frac{72}{125\pi^2} \doteq 0.0584$
[0.042]
$\frac{36}{55\pi^2} \doteq 0.0663$
[0.054]
$\frac{27}{110\pi^2} \doteq 0.0249$
[0.030]
$\frac{2}{5\pi^2} \doteq 0.0405$
[0.024]

FIGURE 7. Proportions of leaves in clades of a given shape, for each shape with $2-6$ leaves in the fringe tree. The top number is from our model, the bottom number $[\cdots]$ from our small data set.

5.4. **Combinatorial questions.** Regarding the number $N_n(\chi)$ of copies of a clade χ in DTCS(n), there are aspects which have not been studied (even within the usual random tree models). For example one could study distributions of the following:

- The number $K_n := \sum_\chi 1_{(N_n(\chi)\geq 1)}$ of different-shape clades within (a realization of) DTCS(n).
- The largest clade that appears more than once within DTCS(n).
- The smallest clade that does not appear within DTCS(n).

The difficulty is that, although one can calculate each $p(\chi)$ numerically, we do not have a useful explicit description of the set of probabilities $(p(\chi) : \ |\chi| = m)$ of size-m clades.

6. Proving Theorem 3.1 via study of CTCS(∞) and Mellin transforms

Having the exchangeable formalization of CTCS(∞) leads to an alternate proof of the second foundational result (Theorem 3.1). This is rather surprising, because convergence of CTCS(n) to CTCS(∞) seems a kind of "global" convergence, whereas the asymptotic fringe is a "local" limit. It turns out that the Mellin transform method can be used to study many other aspects of the model, including leaf heights. Work in progress will appear in [7], and the following is included here as an illustration of the methodology,

Recall notation from Section 4. The central idea of the proof is to define an infinite measure Υ on $(0,1)$ by

$$\Upsilon := \int_0^\infty \mathcal{L}(P_{t,1})\,\mathrm{d}t. \tag{6.1}$$

Formula (4.4) immediately tells us the moments of the measure Υ:

$$\int_0^1 x^{s-1}\,\mathrm{d}\Upsilon(x) = \int_0^\infty \mathbb{E}\,[P_{t,1}^{s-1}]\,\mathrm{d}t = \frac{1}{\psi(s)-\psi(1)}, \qquad \Re s > 1. \tag{6.2}$$

So this is the Mellin transform of Υ. We do not know how to invert the transform to obtain an explicit formula for Υ, but what is relevant to the current proof is the behavior of Υ near 0, as follows.

Lemma 6.1. *Let Υ be the infinite measure on $(0,1)$ having the Mellin transform (6.2). Then Υ is absolutely continuous, with a continuous density $\upsilon(x)$ on $(0,1)$ that satisfies*

$$\upsilon(x) = \frac{6}{\pi^2 x} + O\bigl(x^{-s_1} + x^{-s_1}|\log x|^{-1}\bigr), \tag{6.3}$$

uniformly for $x \in (0,1)$, where $s_1 \doteq -0.567$ is the largest negative root of $\psi(s) = \psi(1)$. In particular, for $x \in (0,\frac{1}{2})$ say,

$$\upsilon(x) = \frac{6}{\pi^2 x} + O\bigl(x^{-s_1}\bigr) \text{ as } x \downarrow 0. \tag{6.4}$$

The (quite technical) proof of this "inversion" lemma is given in Appendix C.

6.1. Deriving Theorem 3.1.

Here we show how to derive Theorem 3.1 via the exchangeable representation and Theorem 4.5 and Lemma 6.1.

For $j \geq 2$ let, as in Section 3.3, $N_n(j)$ be the number of internal nodes of CTCS(n) that have exactly j leaves as descendants. Similarly, let $N_n(j;t)$ be the number of clades of CTCS(n) at time t that have size exactly j. Let

$$e_n(j) := \mathbb{E}\,[N_n(j)], \qquad e_n(j;t) := \mathbb{E}\,[N_n(j;t)]. \tag{6.5}$$

The integral $\int_0^\infty N_n(j;t)\,\mathrm{d}t$ equals the sum of the lifetimes of all clades of size j that ever appear in CTCS(n). Because these lifetimes have expectation $1/h_{j-1}$ (and are independent of the structure), we have

$$\int_0^\infty e_n(j;t)\,\mathrm{d}t = \mathbb{E}\Bigl[\int_0^\infty N_n(j;t)\,\mathrm{d}t\Bigr] = \frac{1}{h_{j-1}}\,\mathbb{E}\,[N_n(j)] = \frac{e_n(j)}{h_{j-1}}. \tag{6.6}$$

As noted previously in (3.13),

$$a(n,j) = \tfrac{j}{n} e_n(j) \tag{6.7}$$

so to prove Theorem 3.1 it will suffice to study the behavior of $e_n(j)$.

To start to calculate $e_n(j)$, use the paintbox construction in Theorem 4.2 to see that

> conditioned on $(P_{t,\ell})_{\ell=1}^{\infty}$, the probability that a given set of j leaves form a clade at time t equals $\sum_\ell P_{t,\ell}^j (1-P_{t,\ell})^{n-j}$.

Thus, by recalling that $P_{t,1}$ can be regarded as a size-biased sample of $(P_{t,\ell})_{\ell=1}^{\infty}$ [13, Corollary 2.4], we see

$$e_n(j;t) = \binom{n}{j} \mathbb{E}\Big[\sum_\ell P_{t,\ell}^j (1-P_{t,\ell})^{n-j}\Big] = \binom{n}{j} \mathbb{E}\big[P_{t,1}^{j-1}(1-P_{t,1})^{n-j}\big]. \tag{6.8}$$

Recall that Υ is the infinite measure on $(0,1)$ given by

$$\Upsilon := \int_0^\infty \mathcal{L}(P_{t,1})\,\mathrm{d}t. \tag{6.9}$$

Then (6.8) yields

$$\int_0^\infty e_n(j;t)\,\mathrm{d}t = \binom{n}{j} \int_0^1 x^{j-1}(1-x)^{n-j}\,\mathrm{d}\Upsilon(x) \tag{6.10}$$

and thus

$$\frac{1}{n}\int_0^\infty e_n(j;t)\,\mathrm{d}t \sim \frac{n^{j-1}}{j!} \int_0^1 x^{j-1}(1-x)^{n-j}\,\mathrm{d}\Upsilon(x) \qquad \text{as } n \to \infty. \tag{6.11}$$

Lemma 6.1 gives us the relevant information about the density $\upsilon(x)$ of Υ, and then we complete a proof of Theorem 3.1 as follows. By (6.6), (6.10) and Lemma 6.1, we have

$$\frac{1}{h_{j-1}}\,\frac{e_n(j)}{n} = \frac{1}{n}\int_0^\infty e_n(j;t)\,\mathrm{d}t = \frac{1}{n}\binom{n}{j}\int_0^1 x^{j-1}(1-x)^{n-j}\upsilon(x)\,\mathrm{d}x. \tag{6.12}$$

Substitute $\upsilon(x)$ from (6.3). The main term becomes, using a standard beta integral,

$$\begin{aligned} \frac{1}{n}\binom{n}{j}\int_0^1 x^{j-1}(1-x)^{n-j}\frac{6}{\pi^2}x^{-1}\,\mathrm{d}x &= \frac{6}{\pi^2}\frac{1}{n}\binom{n}{j}\int_0^1 x^{j-2}(1-x)^{n-j}\,\mathrm{d}x \\ &= \frac{6}{\pi^2}\frac{1}{n}\binom{n}{j}\frac{\Gamma(j-1)\Gamma(n-j+1)}{\Gamma(n)} \\ &= \frac{6}{\pi^2}\frac{1}{n}\binom{n}{j}\frac{(j-2)!\,(n-j)!}{(n-1)!} \\ &= \frac{6}{\pi^2}\frac{1}{j(j-1)}. \end{aligned} \tag{6.13}$$

The contribution from the error term in (6.3) has absolute value at most, letting C denote unimportant constants (not necessarily the same), and using $-\log x > 1-x$,

$$\begin{aligned} C\frac{1}{n}\binom{n}{j}\int_0^1 x^{j-1}(1-x)^{n-j}x^{-s_1}\bigl(1+|\log x|^{-1}\bigr)\,\mathrm{d}x \\ &\le Cn^{j-1}\int_0^1 x^{j-s_1-1}(1-x)^{n-j-1}\,\mathrm{d}x \\ &\le Cn^{j-1}\int_0^\infty x^{j-s_1-1}e^{-(n-j-1)x}\,\mathrm{d}x \\ &= Cn^{j-1}(n-j-1)^{-j+s_1} \\ &= O\bigl(n^{s_1-1}\bigr). \end{aligned} \tag{6.14}$$

This is $o(1)$ as $n\to\infty$, and thus from (6.12) and (6.13), for every fixed $j\ge 2$,

$$\frac{e_n(j)}{n}\to\frac{6}{\pi^2}\frac{h_{j-1}}{j(j-1)}. \tag{6.15}$$

Then by (6.7) we get the assertion of Theorem 3.1: for $j\ge 2$,

$$a(n,j)\to\frac{6}{\pi^2}\frac{h_{j-1}}{j-1}=:a(j) \tag{6.16}$$

with the bound

$$|a(n,j)-a(j)|=O\bigl(n^{s_1-1}\bigr)\text{ as }n\to\infty. \tag{6.17}$$

□

7. Final remarks

7.1. The general beta-splitting model. The mathematical theme of [5] was to introduce the *beta-splitting model* with split probabilities

$$q(n,i)=\frac{1}{a_n(\beta)}\;\frac{\Gamma(\beta+i+1)\Gamma(\beta+n-i+1)}{\Gamma(i+1)\Gamma(n-i+1)},\;1\le i\le n-1 \tag{7.1}$$

with a parameter $-2\le\beta\le\infty$ and normalizing constant $a_n(\beta)$. The qualitative behavior of the model is different for $\beta>-1$ than for $\beta<-1$; in the former case the height (number of edges to the root) of a typical leaf grows as order $\log n$, and in the latter case as order $n^{-\beta-1}$. In this article we are studying the *critical*[13] case $\beta=-1$. The general beta-splitting model is often mentioned in the mathematical biology literature on phylogenetics as one of several simple stochastic models. See [31, 38] for recent overviews of that literature.

[13] Hence our terminology CS for *critical splitting*. But note that *critical* in our context is quite different from the usual *critical* in the context of branching processes or percolation.

7.2. **Analogies with and differences from the Brownian CRT.** As noted in the introduction, this article is part of a broad project investigating the random tree model. The document [6], which will be periodically updated, is intended to provide an overview: statements of results, proofs not given elsewhere, heuristics and open problems, data from phylogenetic trees, and general discussion. Let us discuss below only one aspect of the project.

The best known continuous limit of finite random tree models is the Brownian continuum random tree (CRT) [3, 4, 15, 18], which is a scaling limit of conditioned Galton-Watson trees and other "uniform random tree" models. Our CTCS(∞) model can also be regarded as a scaling limit[14] of CTCS(n). How do these compare?

(**a**) The most convenient formalization of the Brownian CRT is as a random *measured metric space*, with the Gromov-Hausdorff-Prokhorov topology [1] on the set of all such spaces. So one automatically has a notion of convergence in distribution. Our formalization of CTCS(∞) via exchangeable partitions is less amenable to rephrasing as a random element of some metric space. For instance it is easy to visualize Brownian motion [11] on a realization of the CRT, but it seems harder to visualize a stochastic process on a realization of CTCS(∞).

(**b**) Our consistency result, that CTCS(n) is consistent as n increases, and exchangeable over the random leaves, constitutes one general approach to the construction of continuum random trees (CRTs) [4, 15].

(**c**) Our explicit inductive construction (growth algorithm) is analogous to the line-breaking constructions of the Brownian CRT [3] and stable trees [19].

(**d**) Haas et al [23] and subsequent work such as [22] have given a detailed general treatment of self-similar fragmentations via exchangeable partitions, though the focus there is on characterizations and on models like the $-2 < \beta < -1$ case of the beta-splitting model. On the range $-2 < \beta < -1$, such models have limits which are qualitatively analogous to the Brownian continuum random tree, which is the case $\beta = -2$.

(**e**) It is implausible that CTCS(∞) is as "universal" a limit as the Brownian CRT has proved to be, but nevertheless one can ask *Are there superficially different discrete models whose limit is the same* CTCS(∞)*?* The key feature of our model seems to be the subordinator approximation (3.7): can this arise in some other model?

Acknowledgments

Thanks to Boris Pittel for extensive interactions regarding this project. Thanks to Serte Donderwinkel for pointing out a gap in an early version, and to Jim Pitman and David Clancy and Prabhanka Deka for helpful comments on early versions. Thanks especially to Bénédicte Haas for his careful explanation of how our setting fits into the general theory of exchangeable random partitions, which is the basis of our Section 4, as noted in Remark 4.3.

[14] The "scaling" in CTCS(n) arises from the initial splitting rate being h_{n-1}; in the finite uniform random tree model we scale edge lengths to be order $n^{-1/2}$.

For recent alternative proofs mentioned in the text we thank Brett Kolesnik, Luca Pratelli and Pietro Rigo, and in particular Alexander Iksanov, whose observation of the connection with regenerative composition structures may lead to interesting further results.

Appendix A. More on the consistency property and growth algorithm

We give in this section our original proof of the consistency property Theorem 2.3 and the growth algorithm Theorem 2.4. This proof uses only straightforward (although rather long) calculations of (conditional) probabilities.

Recall the statement of Theorem 2.3: *The operation "delete and prune leaf* $n+1$ *from* CTCS$(n+1)$*" gives a tree distributed as* CTCS(n).

Consider a pair of trees $(\mathbf{t}_n, \mathbf{t}_{n+1})$ with n and $n+1$ leaves, in which $\mathbf{t}_n$ can be obtained from $\mathbf{t}_{n+1}$ via the "delete a leaf and prune" operation in Figure 4. So $\mathbf{t}_{n+1}$ arises by adding a new bud to $\mathbf{t}_n$, which can happen in one of three qualitative ways illustrated in Figure 5.

We will do explicit calculations for (CTCS(3), CTCS(4)) in the following section. This enables one to guess the growth algorithm in Theorem 2.4, which we verify for general n in Section A.2. In the argument below, it will be convenient to use unlabelled leaves, cf. Remark 2.2.

A.1. **A starting step.** The distribution of CTCS(n) is specified by the shape of the tree and the probability density of the edge-lengths. For $n = 3$ there are only two possible shapes, as $\mathbf{t}$ in Figure 8 and as its "reflection" with the side-bud on the left instead of the right. There are two edge-lengths (a, b). Clearly the density of CTCS(3) is

$$f(\mathbf{t}; a, b) = \tfrac{1}{2} h_2 e^{-h_2 a} \cdot e^{-b}; \quad 0 < a, b < \infty \tag{A.1}$$

and the probability of $\mathbf{t}$ is 1/2. There are 7 shapes of CTCS(4) that are consistent with this $\mathbf{t}$, shown as $\mathbf{t}_1, \mathbf{t}_2, \mathbf{t}_3, \mathbf{t}_4$ in Figure 8, together with the "reflected" forms of $\mathbf{t}_1$ and $\mathbf{t}_3$ and $\mathbf{t}_4$ (the added side-bud involves the other side; drawn as $\hat{\mathbf{t}}_1, \hat{\mathbf{t}}_3, \hat{\mathbf{t}}_4$) which will be accounted for as $q(\cdot,\cdot) + q(\cdot,\cdot)$ terms in the calculation below.[15] The densities of these shapes involve 3 edge-lengths (a, b, c), calculated below as $f_i^+(a, b, c)$. We also calculate the marginals $f_i(a, b) = \int f_i^+(a, b, c)\, \mathrm{d}c$.

The consistency assertion that we wish to verify is the assertion, for $f = f(\mathbf{t}; \cdot, \cdot)$ as at (A.1),

$$f \overset{?}{=} \tfrac{1}{4} f_1 + \tfrac{2}{4} f_2 + \tfrac{1}{4} f_3 + \tfrac{2}{4} f_4 \tag{A.2}$$

[15] That is, f_1^+ is the density of $\mathbf{t}_1$ plus the density of $\hat{\mathbf{t}}_1$.

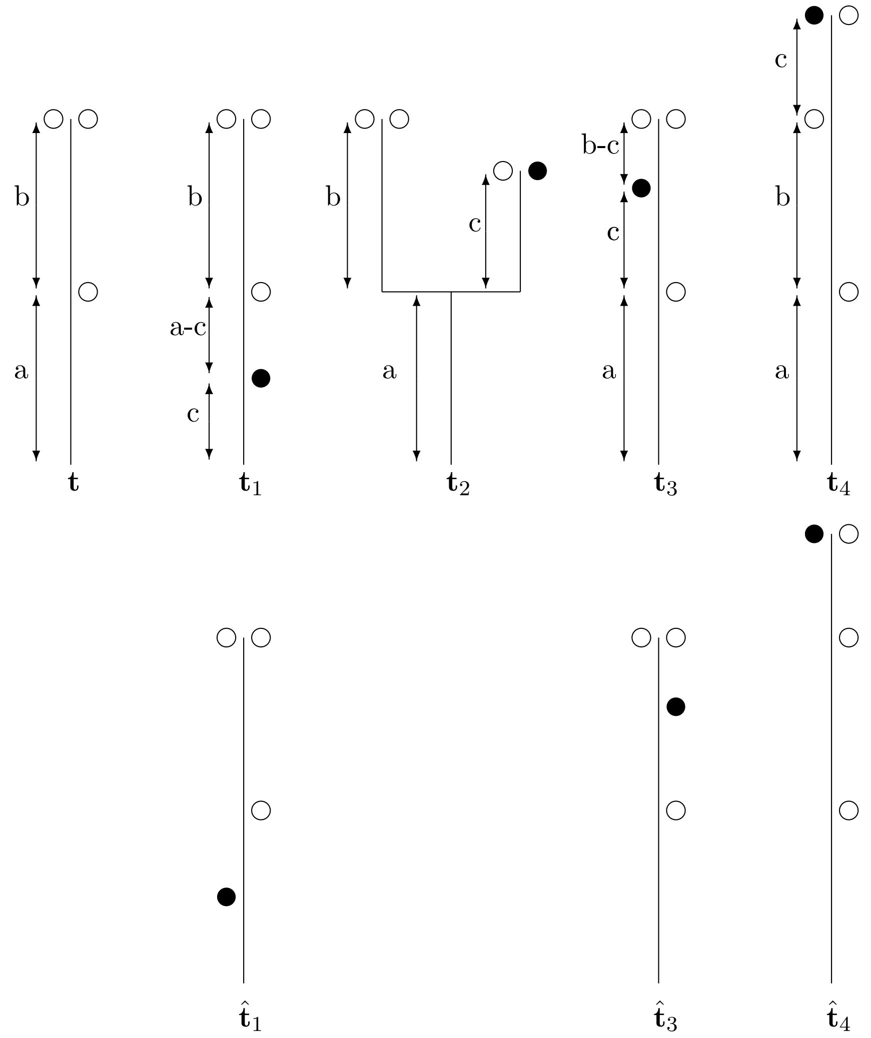


FIGURE 8. The possible transitions from $\mathbf{t}$: the added bud is $\bullet$.

where the fractions denote the probability that deleting a random bud gives $\mathbf{t}$ with the given edge-lengths (a, b). From the definition of CTCS(4) we can calculate

$$f_1^+(a,b,c) = (q(4,1)+q(4,3))\cdot h_3 e^{-h_3 c}\cdot q(3,2)\cdot h_2 e^{-h_2(a-c)}\cdot e^{-b} \tag{A.3}$$

$$\begin{aligned} f_1(a,b) &= (q(4,1)+q(4,3))\cdot h_3\cdot 3(1-e^{-a/3})\cdot q(3,2)\cdot h_2 e^{-h_2 a}\cdot e^{-b} \\ &= 3(e^{-h_2 a}-e^{-h_3 a})\cdot e^{-b} \end{aligned} \tag{A.4}$$

$$f_2^+(a,b,c) = q(4,2)\cdot h_3 e^{-h_3 a}\cdot e^{-b}e^{-c} \tag{A.5}$$

$$\begin{aligned} f_2(a,b) &= q(4,2)\cdot h_3 e^{-h_3 a}\cdot e^{-b} \\ &= \tfrac{1}{2}e^{-h_3 a}\cdot e^{-b} \end{aligned} \tag{A.6}$$

$$f_3^+(a,b,c) = q(4,3)\cdot h_3 e^{-h_3 a}\cdot (q(3,2)+q(3,1))\cdot h_2 e^{-h_2 c}\cdot e^{-(b-c)} \tag{A.7}$$

$$\begin{aligned} f_3(a,b) &= q(4,3)\cdot h_3 e^{-h_3 a}\cdot (q(3,2)+q(3,1))\cdot h_2\cdot 2(1-e^{-b/2})\cdot e^{-b} \\ &= 2e^{-h_3 a}(1-e^{-b/2})\cdot e^{-b} \end{aligned} \tag{A.8}$$

$$f_4^+(a,b,c) = q(4,3)\cdot h_3 e^{-h_3 a}\cdot (q(3,2)+q(3,1))\cdot h_2 e^{-h_2 b}\cdot e^{-c} \tag{A.9}$$

$$\begin{aligned} f_4(a,b) &= q(4,3)\cdot h_3 e^{-h_3 a}\cdot (q(3,2)+q(3,1))\cdot h_2 e^{-h_2 b} \\ &= e^{-h_3 a}\cdot e^{-h_2 b}. \end{aligned} \tag{A.10}$$

From this we can verify (A.2).

This argument is not so illuminating, but we can immediately derive the conditional distribution of CTCS(4) given that CTCS(3) is $(\mathbf{t}, a, b)$. Writing $g_i(c|a,b)$ for the conditional density of shape $\mathbf{t}_i$ or $\hat{\mathbf{t}}_i$ and additional edge length c, and $p(\mathbf{t}_i|a,b) = \int g_i(c|a,b)\ dc$ for the conditional probability of shape $\mathbf{t}_i$ or $\hat{\mathbf{t}}_i$, we have

$$g_1(c|a,b) = \frac{\frac{1}{4}f_1^+(a,b,c)}{f(a,b)} = \tfrac{1}{3}e^{-c/3}; \quad p(\mathbf{t}_1|a,b) = 1 - e^{-a/3} \tag{A.11}$$

$$g_2(c|a,b) = \frac{\frac{1}{2}f_2^+(a,b,c)}{f(a,b)} = \tfrac{1}{3}e^{-a/3} \cdot e^{-c}; \quad p(\mathbf{t}_2|a,b) = \tfrac{1}{3}e^{-a/3} \tag{A.12}$$

$$g_3(c|a,b) = \frac{\frac{1}{4}f_3^+(a,b,c)}{f(a,b)} = \tfrac{1}{3}e^{-a/3} \cdot e^{-c/2}; \quad p(\mathbf{t}_3|a,b) = \tfrac{2}{3}e^{-a/3}(1 - e^{-b/2}) \tag{A.13}$$

$$g_4(c|a,b) = \frac{\frac{1}{2}f_4^+(a,b,c)}{f(a,b)} = \tfrac{2}{3}e^{-a/3}e^{-b/2} \cdot e^{-c}; \quad p(\mathbf{t}_4|a,b) = \tfrac{2}{3}e^{-a/3} \cdot e^{-b/2}. \tag{A.14}$$

One can now see that these are the conditional probabilities that arise from the growth algorithm in Theorem 2.4 which we for convenience repeat:

Given a realization of CTCS(n) *for some* $n \geq 1$ *(above,* $n = 3$*):*

(1) *Pick a uniform random bud; move up the path from the root toward that bud. A* stop *event occurs at rate* $= 1/$*(size of clade from current position).*
(2) *If* stop *before reaching the target bud, make a side-bud at that point, random on left or right. (As in* $\mathbf{t}_1$ *or* $\mathbf{t}_3$ *above.)*
(3) *Otherwise, extend the target bud into a branch of* Exp(1) *length to make a bud-pair. (As in* $\mathbf{t}_2$ *or* $\mathbf{t}_4$ *above.)*

Figure 8 indicated three of these possibilities $(\mathbf{t}_1, \mathbf{t}_3, \mathbf{t}_4)$ when the chosen target bud was at the top right. The "rate" is $1/3$ until the side-bud, and then $1/2$. Note that case $\mathbf{t}_2$ arises as an "extend the target bud" for a different target bud.

A.2. **The general step.** To set up a calculation, we consider the side-bud addition case first, illustrated by the example in Figure 9, where the left diagram shows the relevant part of $\mathbf{t}_n$ and the right diagram shows the side-bud addition making $\mathbf{t}_{n+1}$. The ℓ_i are edge-lengths and the (n_i) are clade sizes. The side-bud is attached to some edge, in Figure 9 an edge at edge-height 4 with length ℓ_4 and defining a clade of size $n_4 \geq 2$. The new bud splits that edge into edges of length α and $\ell_4 - \alpha$. The probability density function on a given tree is a product of terms for each edge. Table 1 shows the terms for the edges where the terms differ between the two trees – these are only the edges on the path from the root to the added bud. The first three lines in Table 1 refer to the edges below the old edge into which the new bud is inserted, and the bottom line refers to that old edge.

Because $h_{n-1}q(n,k) = \frac{n}{2k(n-k)}$ the ratios right/left of each of the first 3 lines in Table 1 equal

$$\frac{n_i + 1}{n_i} \cdot \frac{n_{i+1}}{n_{i+1} + 1} \cdot \exp(-\ell_i/n_i), \qquad i = 1, 2, 3. \tag{A.15}$$

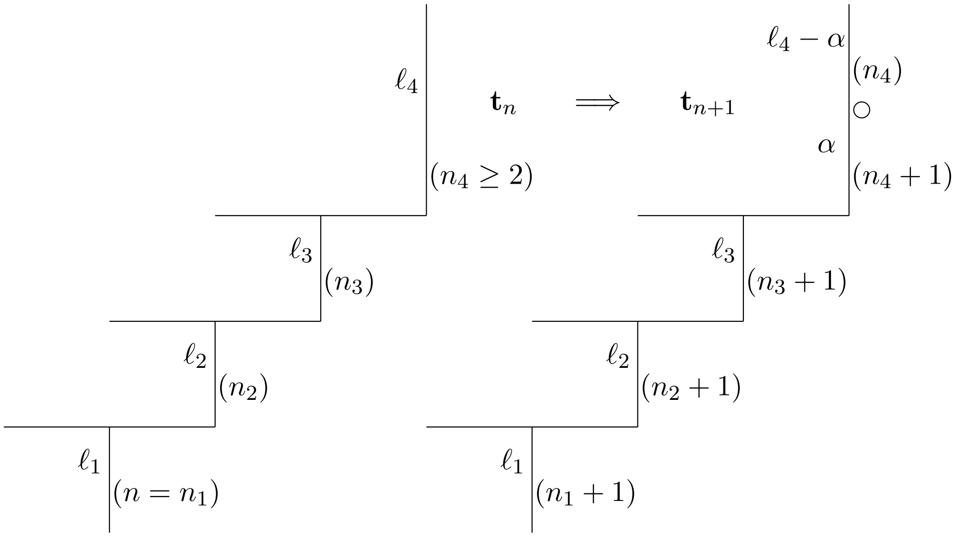


Figure 9. Growing via a side-bud addition

left tree	right tree
$h_{n_1-1}\exp(-h_{n_1-1}\ell_1)\mathrm{d}\ell_1 \cdot q(n_1,n_2)$	$h_{n_1}\exp(-h_{n_1}\ell_1)\mathrm{d}\ell_1 \cdot q(n_1+1,n_2+1)$
$h_{n_2-1}\exp(-h_{n_2-1}\ell_2)\mathrm{d}\ell_2 \cdot q(n_2,n_3)$	$h_{n_2}\exp(-h_{n_2}\ell_2)\mathrm{d}\ell_2 \cdot q(n_2+1,n_3+1)$
$h_{n_3-1}\exp(-h_{n_3-1}\ell_3)\mathrm{d}\ell_3 \cdot q(n_3,n_4)$	$h_{n_3}\exp(-h_{n_3}\ell_3)\mathrm{d}\ell_3 \cdot q(n_3+1,n_4+1)$
$h_{n_4-1}\exp(-h_{n_4-1}\ell_4)\mathrm{d}\ell_4$	$h_{n_4}\exp(-h_{n_4}\alpha)\mathrm{d}\alpha \cdot q(n_4+1,1)$ $\cdot h_{n_4-1}\exp(-h_{n_4-1}(\ell_4-\alpha))\mathrm{d}\ell_4$

Table 1. Differing terms in density product (side-bud case)

$h_{n_4-1}\exp(-h_{n_4-1}\ell_4)\mathrm{d}\ell_4$	$h_{n_4}\exp(-h_{n_4}\ell_4)\mathrm{d}\ell_4 \cdot q(n_4+1,1)$ $\cdot h_1\exp(-h_1\beta)\mathrm{d}\beta$

Table 2. Differing terms in density product (branch extension case)

The corresponding ratio for the final term equals

$$\frac{n_4+1}{2n_4} \cdot \exp(-\alpha/n_4)\,\mathrm{d}\alpha. \tag{A.16}$$

Combining terms, the ratio of densities equals

$$\frac{n+1}{2n} \cdot \exp(-\ell_1/n_1 - \ell_2/n_2 - \ell_3/n_3 - \alpha/n_4)\,\mathrm{d}\alpha. \tag{A.17}$$

In obtaining $\mathbf{t}_n$ from $\mathbf{t}_{n+1}$ we chose one of $n+1$ buds to delete, so finally the conditional density of CTCS(n+1) given CTCS(n) at $(\mathbf{t}_{n+1}|\mathbf{t}_n)$ equals

$$\frac{1}{2n} \cdot \exp(-\ell_1/n_1 - \ell_2/n_2 - \ell_3/n_3 - \alpha/n_4)\,\mathrm{d}\alpha. \tag{A.18}$$

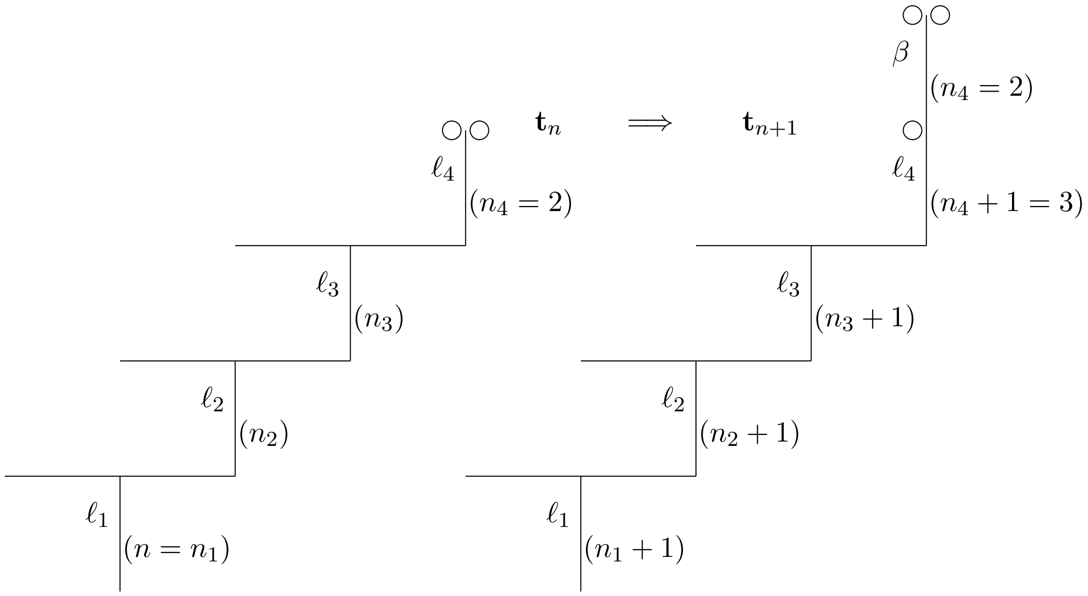


FIGURE 10. Growing via a branch extension

We need to check that this agrees with the growth algorithm. According to the algorithm, the conditional density is a product of terms

- n_4/n: the chance that the target bud is in the relevant clade;
- $\exp(-\ell_1/n_1 - \ell_2/n_2 - \ell_3/n_3)$: the chance of not stopping before reaching the edge of length ℓ_4;
- $\frac{1}{n_4}\exp(-\alpha/n_4)\,\mathrm{d}\alpha$: the chance of stopping in $\mathrm{d}\alpha$;
- $1/2$: chance of placing side-bud on right side.

And this agrees with (A.18).

That was the side-bud addition case. Now consider the branch extension case, illustrated in Figure 10. In this case, $\mathbf{t}_n$ has an edge terminating in two buds. Then $\mathbf{t}_{n+1}$ is obtained by extending the branch by an extra edge of some length β to two terminal buds, leaving one bud as a side-bud. Comparing the densities of $\mathbf{t}_n$ and $\mathbf{t}_{n+1}$ in this case, the first 3 lines are the same as in Table 1, and the 4th is shown in Table 2. Following the previous argument we derive the conditional density in a format similar to (A.18):

$$\frac{1}{2n} \cdot \exp(-\ell_1/n_1 - \ell_2/n_2 - \ell_3/n_3 - \ell_4/n_4 - \beta)\,\mathrm{d}\beta. \tag{A.19}$$

Again this agrees with the growth algorithm. The third case, the side-bud extension, is similar. □

Appendix B. Another proof of Proposition 2.6

In CTCS(n) write $(X_n(i,t), i \geq 1)$ for the clade sizes at time t and consider

$$Q_n(t) = \sum_i X_n^2(i,t). \tag{B.1}$$

Note that, when a size-m clade is split, the effect on sum-of-squares of clade sizes has expectation

$$\sum_{i=1}^{m-1} (m^2 - i^2 - (m-i)^2)\ q(m,i) = \frac{m}{2h_{m-1}} \sum_{i=1}^{m-1} 2 = \frac{m(m-1)}{h_{m-1}}. \tag{B.2}$$

If we chose some arbitrary rates $r(m,n)$ for splitting a size-m clade, then

$$\mathbb{E}\left[Q_n(t) - Q_n(t+\mathrm{d}t)|\mathcal{F}_t\right] = \sum_i r(X_n(i,t),n)\ \frac{X_n(i,t)(X_n(i,t)-1)}{h_{X_n(i,t)-1}}\,\mathrm{d}t. \tag{B.3}$$

So by choosing $r(m,n) = h_{m-1}$ we obtain

$$\mathbb{E}\left[Q_n(t) - Q_n(t+\mathrm{d}t)|\mathcal{F}_t\right] = (Q_n(t) - n)\,\mathrm{d}t. \tag{B.4}$$

Because $Q_n(0) = n^2$ we obtain the exact formula

$$\mathbb{E}\left[Q_n(t)\right] = n + (n^2-n)e^{-t},\ 0 \leq t < \infty. \tag{B.5}$$

Now we are studying the height B_n of the branchpoint between the paths to two uniform random distinct leaves of CTCS(n). The conditional probability that both sampled leaves are in clade i at time t equals $\frac{1}{n(n-1)} X_n(i,t)(X_n(i,t)-1)$. So

$$\begin{aligned}\mathbb{P}(B_n > t) &= \tfrac{1}{n(n-1)}\,\mathbb{E}\,[\sum_i X_n(i,t)(X_n(i,t)-1)] \\ &= \tfrac{1}{n(n-1)}\,\mathbb{E}\left[Q_n(t) - n\right] \\ &= e^{-t} \text{ by (B.5)}.\end{aligned} \tag{B.6}$$

Appendix C. Proof of Lemma 6.1 by Mellin inversion

Proof of Lemma 6.1. We begin by noting that the Mellin transform $1/\bigl(\psi(s)-\psi(1)\bigr)$ in (6.2) extends to a meromorphic function in the entire complex plane. The poles are the roots of

$$\psi(s) = \psi(1). \tag{C.1}$$

Obviously, $s_0 := 1$ is a pole. Its residue is

$$\operatorname{Res}_{s=1} \frac{1}{\psi(s)-\psi(1)} = \frac{1}{\psi'(1)} = \frac{6}{\pi^2}, \tag{C.2}$$

using the well known formula $\psi'(1) = \pi^2/6$ [35, 5.4.12] (see also (C.15) below). As shown in Lemma C.1 below the other poles are real and negative, and thus can be ordered $0 > s_1 > s_2 > \dots$. In particular, there are no other poles in the half-plane $\Re s > s_1$, with $s_1 \doteq -0.567$.

We cannot immediately use standard results on Mellin inversion[16] (as in [17, Theorem 2(i)]) because the Mellin transform in (6.2) decreases too slowly as $\Im s \to \pm\infty$ to be integrable on a vertical line $\Re s = c$. In fact, Stirling's formula implies (see e.g. [35, 5.11.2]) that

$$\psi(s) = \log s + o(1) = \log|s| + O(1) = \log|\Im s| + O(1) \tag{C.3}$$

as $\Im s \to \infty$ with s in, for example, any half-plane $\Re s \geq c$.

We overcome this problem by differentiating the Mellin transform, but we first subtract the leading term corresponding to the pole at 1. Since Υ is an infinite measure, we first replace it by ν defined by $\mathrm{d}\nu(x) = x\,\mathrm{d}\Upsilon(x)$; note that ν is also a measure on $(0,1)$, and taking $s = 2$ in (6.2) shows that ν is a finite measure.

Next, define ν_0 as the measure $(6/\pi^2)\,\mathrm{d}x$ on $(0,1)$, and let ν_Δ be the (finite) signed measure $\nu - \nu_0$. Then ν_Δ has the Mellin transform, by (6.2),

$$\begin{aligned}\widetilde{\nu_\Delta}(s) &:= \int_0^1 x^{s-1}\,\mathrm{d}\nu_\Delta(x) = \int_0^1 x^s\,\mathrm{d}\Upsilon(x) - \frac{6}{\pi^2}\int_0^1 x^{s-1}\,\mathrm{d}x \\ &= \frac{1}{\psi(s+1)-\psi(1)} - \frac{6}{\pi^2 s}, \qquad \Re s > 0.\end{aligned} \tag{C.4}$$

We may here differentiate under the integral sign, which gives

$$\widetilde{\nu_\Delta}'(s) := \int_0^1 (\log x)x^{s-1}\,\mathrm{d}\nu_\Delta(x) \tag{C.5}$$

$$= -\frac{\psi'(s+1)}{(\psi(s+1)-\psi(1))^2} + \frac{6}{\pi^2 s^2}, \qquad \Re s > 0. \tag{C.6}$$

The Mellin transform $\widetilde{\nu_\Delta}(s)$ extends to a meromorphic function in $\mathbb{C}$ with (simple) poles $(s_i - 1)_1^\infty$; note that there is no pole at $s_0 - 1 = 0$, since the residues there of the two terms in (C.4) cancel by (C.2). Furthermore, the formula (C.6) for $\widetilde{\nu_\Delta}'(s)$ holds for all s (although the integral in (C.5) diverges unless $\Re s > 0$).

For any real c we have, on the vertical line $\Re s = c$, as $\Im s \to \pm\infty$, that $\psi(s) \sim \log|s|$ by (C.3), and also, by careful differentiation of (C.3) or by [35, 5.15.8], that $\psi'(s) \sim s^{-1}$. It follows from (C.6) that $\widetilde{\nu_\Delta}'(s) = O(|s|^{-1}\log^{-2}|s|)$ on the line $\Re s = c$, for $|\Im s| \geq 2$ say, and thus $\widetilde{\nu_\Delta}'$ is integrable on this line unless c is one of the poles $s_i - 1$. In particular, taking $c = 1$ and thus $s = 1 + u\mathrm{i}$ ($u \in \mathbb{R}$), we see that the function

$$\widetilde{\nu_\Delta}'(1+\mathrm{i}u) = \int_0^1 x^{\mathrm{i}u}\log(x)\,\mathrm{d}\nu_\Delta(x) \tag{C.7}$$

is integrable. The change of variables $x = e^{-y}$ shows that the function (C.7) is the Fourier transform of the signed measure on $\mathbb{R}_+$ that corresponds to $\log(x)\,\mathrm{d}\nu_\Delta(x)$. This measure on $\mathbb{R}_+$ is thus a finite signed measure with integrable Fourier transform, which implies that it is absolutely continuous with a continuous density. Reversing the change of variables, we thus see that the signed measure $\log(x)\,\mathrm{d}\nu_\Delta(x)$ is absolutely continuous with a continuous density on $(0,1)$. Moreover, denoting this

[16] And we cannot use [17, Theorem 2(ii)] since we do not know that Υ has a density that is locally of bounded variation.

density by $h(x)$, we obtain the standard inversion formula for the Mellin transform [17, Theorem 2(i)], [35, 1.14.35]:

$$h(x) = \frac{1}{2\pi \mathrm{i}} \int_{c-\infty \mathrm{i}}^{c+\infty \mathrm{i}} x^{-s} \widetilde{\nu_\Delta}'(s)\,\mathrm{d}s, \qquad x > 0, \tag{C.8}$$

with $c = 1$. Furthermore, the integrand in (C.8) is analytic in the half-plane $\Re s > s_1 - 1$, and the estimates above of $\psi(s)$ and $\psi'(s)$ are uniform for $\Re s$ in any compact interval and, say, $|\Im s| \geq 2$. Consequently, we may shift the line of integration in (C.8) to any $c > s_1 - 1$. Taking absolute values in (C.8), and recalling that $\widetilde{\nu_\Delta}'(s)$ is integrable on the line, then yields

$$h(x) = O\bigl(x^{-c}\bigr) \tag{C.9}$$

for any $c > s_1 - 1$.

Reversing the transformations above, we see that ν_Δ has the density $(\log x)^{-1} h(x)$, and thus ν has the density $(\log x)^{-1} h(x) + 6/\pi^2$, and, finally, that Υ has the density

$$\upsilon(x) := \frac{\mathrm{d}\Upsilon}{\mathrm{d}x} = \frac{1}{x}\frac{\mathrm{d}\nu}{\mathrm{d}x} = \frac{6}{\pi^2 x} + \frac{1}{x \log x} h(x), \qquad 0 < x < 1. \tag{C.10}$$

Furthermore, (C.10) and (C.9) have the form of the claimed estimate (6.3), although with the weaker error term $O(x^{-s_1-\varepsilon} |\log x|^{-1})$ for any $\varepsilon > 0$.

To obtain the (stronger) claimed error term, we note that the residue of $\nu_\Delta(s)$ at $s_1 - 1$ is $r_1 := 1/\psi'(s_1)$. Let ν_1 be the measure $r_1 x^{1-s_1}\,\mathrm{d}x$ on $(0,1)$; then ν_1 has Mellin transform

$$\widetilde{\nu_1}(s) = r_1 \int_0^1 x^{s-1} x^{1-s_1}\,\mathrm{d}x = \frac{r_1}{s+1-s_1}, \qquad \Re s > s_1 - 1. \tag{C.11}$$

It follows from (C.4) and (C.11) that the signed measure $\nu - \nu_0 - \nu_1 = \nu_\Delta - \nu_1$ has the Mellin transform

$$\frac{1}{\psi(s+1) - \psi(1)} - \frac{6}{\pi^2 s} - \frac{r_1}{s+1-s_1}, \tag{C.12}$$

which is an analytic function in the half plane $\Re s > s_2 - 1$. Hence, the same argument as above yields the estimate

$$\upsilon(x) = \frac{6}{\pi^2 x} + \frac{1}{\psi'(s_1)} x^{-s_1} + O\bigl(x^{-s_2-\varepsilon} |\log x|^{-1}\bigr), \qquad x \downarrow 0, \tag{C.13}$$

for any $\varepsilon > 0$, which in particular yields (6.3). □

C.1. **A lemma on the digamma function.**

Lemma C.1. *The roots of the equation $\psi(s) = \psi(1)$ are all real and can be enumerated in decreasing order as $s_0 = 1 > s_1 > s_2 > \dots$, with $s_i \in (-i, -(i-1))$ for $i \geq 1$. Numerically, $s_1 \doteq -0.567$ and $s_2 \doteq -1.628$.*

Proof. Recall that $\psi(s)$ is a meromorphic function of s, with poles at $0, -1, -2, \dots$. For any other complex s we have the standard formulas [35, 5.7.6 and 5.15.1]

$$\psi(s) = -\gamma + \sum_{k=0}^{\infty}\Bigl(\frac{1}{k+1} - \frac{1}{k+s}\Bigr), \tag{C.14}$$

$$\psi'(s) = \sum_{k=0}^{\infty} \frac{1}{(k+s)^2}. \tag{C.15}$$

If $\Im s > 0$, then $\Im(1/(k+s)) < 0$ for all k and thus (C.14) implies $\Im\psi(s) > 0$. Similarly, if $\Im s < 0$, then $\Im\psi(s) < 0$. Consequently, all roots of $\psi(s) = \psi(1)$ are real.

For real s, (C.15) shows that $\psi'(s) > 0$. We can write $\mathbb{R} \setminus \{\text{the poles}\} = \bigcup_{i=0}^{\infty} I_i$ with $I_0 := (0, \infty)$ and $I_i := (-i, -(i-1))$ for $i \geq 1$; it then follows that $\psi(s)$ is strictly increasing in each interval I_i. Moreover, by (C.14) (or general principles), at the poles we have the limits $\psi(-i-0) = +\infty$ and $\psi(-i+0) = -\infty$ $(i \geq 0)$. Consequently, $\psi(s) = \psi(1)$ has exactly one root s_i in each I_i, and obviously the positive root is $s_0 = 1$. (See also the graph of $\psi(s)$ in [35, Figure 5.3.3].) □

References

[1] Romain Abraham, Jean-François Delmas, and Patrick Hoscheit. A note on the Gromov-Hausdorff-Prokhorov distance between (locally) compact metric measure spaces. *Electron. J. Probab.*, 18:no. 14, 21 pp., 2013.

[2] David Aldous. Asymptotic fringe distributions for general families of random trees. *Ann. Appl. Probab.*, 1(2):228–266, 1991.

[3] David Aldous. The continuum random tree. II. An overview. In *Stochastic analysis (Durham, 1990)*, volume 167 of *London Math. Soc. Lecture Note Ser.*, pages 23–70. Cambridge Univ. Press, Cambridge, 1991.

[4] David Aldous. The continuum random tree. III. *Ann. Probab.*, 21(1):248–289, 1993.

[5] David Aldous. Probability distributions on cladograms. In *Random discrete structures (Minneapolis, MN, 1993)*, volume 76 of *IMA Vol. Math. Appl.*, pages 1–18. Springer, New York, 1996.

[6] David Aldous and Svante Janson. The critical beta-splitting random tree II: Overview and open problems. arXiv 2303.02529v2, 2024.

[7] David Aldous and Svante Janson. The critical beta-splitting random tree IV: Mellin analysis of leaf height. In preparation, 2024.

[8] David Aldous, Svante Janson, and Xiaodan Li. The harmonic descent chain. *Electron. Commun. Probab.* 29 (2024), paper no. 77, 1–10.

[9] David Aldous and Boris Pittel. The critical beta-splitting random tree I: Heights and related results. arXiv:2302.05066, 2023. To appear in *Ann. Appl. Probab.*

[10] David Aldous. Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.*, 16(1):23–34, 2001.

[11] George Andriopoulos, David A. Croydon, Vlad Margarint and Laurent Menard. On the cover time of Brownian motion on the Brownian continuum random tree. arXiv:2410.03922, 2024.

[12] Jean Bertoin. Homogeneous fragmentation processes. *Probab. Theory Related Fields*, 121(3):301–318, 2001.

[13] Jean Bertoin. *Random fragmentation and coagulation processes*, volume 102 of *Cambridge Studies in Advanced Mathematics.* Cambridge Univ. Press, Cambridge, 2006.

[14] Timothy M. Crowe, Rauri C.K. Bowie, Paulette Bloomer, Tshifhiwa G. Mandiwana, Terry A.J. Hedderson, Ettore Randi, Sergio L. Pereira, and Julia Wakeling. Phylogenetics, biogeography

and classification of, and character evolution in, gamebirds (Aves: Galliformes): effects of character exclusion, data partitioning and missing data. *Cladistics*, 22(6):495–532, 2006.

[15] Steven N. Evans. *Probability and real trees*, volume 1920 of *Lecture Notes in Mathematics*. Lectures from the 35th Summer School on Probability Theory held in Saint-Flour, July 6–23, 2005. Springer, Berlin, 2008.

[16] Alex Figueroa, Alexander D. McKelvy, L. Lee Grismer, Charles D. Bell, and Simon P. Lailvaux. A species-level phylogeny of extant snakes with description of a new colubrid subfamily and genus. *PLOS ONE*, 11(9):e0161070, 2016.

[17] Philippe Flajolet, Xavier Gourdon, and Philippe Dumas. Mellin transforms and asymptotics: harmonic sums. *Theoret. Comput. Sci.*, 144(1-2):3–58, 1995.

[18] Christina Goldschmidt. Scaling limits of random trees and random graphs. In *Random graphs, phase transitions, and the Gaussian free field*, volume 304 of *Springer Proc. Math. Stat.*, pages 1–33. Springer, Cham, 2020.

[19] Christina Goldschmidt, Bénédicte Haas, and Delphin Sénizergues. Stable graphs: distributions and line-breaking construction. *Ann. H. Lebesgue*, 5:841–904, 2022.

[20] Morris Goodman, Lawrence I. Grossman, and Derek E. Wildman. Moving primate genomics beyond the chimpanzee genome. *TRENDS in Genetics*, 21(9):511–517, 2005.

[21] Bénédicte Haas and Grégory Miermont. The genealogy of self-similar fragmentations with negative index as a continuum random tree. *Electron. J. Probab.* 9(4):57–97, 2004.

[22] Bénédicte Haas and Grégory Miermont. Scaling limits of Markov branching trees with applications to Galton-Watson and random unordered trees. *Ann. Probab.*, 40(6):2589–2666, 2012.

[23] Bénédicte Haas, Grégory Miermont, Jim Pitman, and Matthias Winkel. Continuum tree asymptotics of discrete fragmentations and applications to phylogenetic models. *Ann. Probab.*, 36(5):1790–1837, 2008.

[24] Benjamin Hollering and Seth Sullivant. Exchangeable and sampling-consistent distributions on rooted binary trees. *J. Appl. Probab.*, 59(1):60–80, 2022.

[25] Cecilia Holmgren and Svante Janson. Fringe trees, Crump-Mode-Jagers branching processes and m-ary search trees. *Probab. Surv.*, 14:53–154, 2017.

[26] Alexander Iksanov. A comment on the article 'The harmonic descent chain' by D. J. Aldous, S. Janson and X. Li. arxiv 2412.06826, 2024.

[27] Alexander Iksanov. Another proof of CLT for critical beta-splitting tree. Unpublished, 2024.

[28] Jasper Ischebeck. Central limit theorems for fringe trees in patricia tries. arXiv 2305.14900, 2023.

[29] Svante Janson. Fringe trees of Patricia tries and compressed binary search trees. arXiv 2405.01239, 2024.

[30] Brett Kolesnik. Critical beta-splitting, via contraction. arXiv 2404.16021, 2024.

[31] Amaury Lambert. Probabilistic models for the (sub)tree(s) of life. *Braz. J. Probab. Stat.*, 31(3):415–475, 2017.

[32] Heather R.L. Lerner and David P. Mindell. Phylogeny of eagles, Old World vultures, and other Accipitridae based on nuclear and mitochondrial DNA. *Molecular Phylogenetics and Evolution*, 37(2):327–346, 2005.

[33] Harald Letsch. *Phylogeny of Anisoptera (Insecta: Odonata): promises and limitations of a new alignment approach*. PhD thesis, Rheinische Friedrich-Wilhelms-Universität in Bonn, 2007.

[34] Alexandra Magro, E. Lecompte, F. Magne, J.-L. Hemptinne, and B. Crouau-Roy. Phylogeny of ladybirds (Coleoptera: Coccinellidae): are the subfamilies monophyletic? *Molecular Phylogenetics and Evolution*, 54(3):833–848, 2010.

[35] Frank W. J. Olver, Daniel W. Lozier, Ronald F. Boisvert, and Charles W. Clark, editors. *NIST handbook of mathematical functions*. U.S. Department of Commerce, National Institute of Standards and Technology, Washington, DC; Cambridge Univ. Press, Cambridge, 2010. Also available as *NIST Digital Library of Mathematical Functions*, `http://dlmf.nist.gov/`

[36] J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics*. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002. Springer-Verlag, Berlin, 2006.

[37] Frederick H. Sheldon, Linda A. Whittingham, Robert G. Moyle, Beth Slikas, and David W. Winkler. Phylogeny of swallows (Aves: Hirundinidae) estimated from nuclear and mitochondrial DNA sequences. *Molecular Phylogenetics and Evolution*, 35(1):254–270, 2005.

[38] Mike Steel. *Phylogeny—discrete and random processes in evolution*, volume 89 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2016.

[39] Ximena Vélez-Zuazo and Ingi Agnarsson. Shark tales: a molecular species-level phylogeny of sharks (Selachimorpha, Chondrichthyes). *Molecular Phylogenetics and Evolution*, 58(2):207–217, 2011.

[40] Alan T. Whittemore, Ryan S. Fuller, Bethany H. Brown, Marlene Hahn, Linus Gog, Jaime A. Weber, and Andrew L. Hipp. Phylogeny, biogeography, and classification of the elms (*Ulmus*). *Systematic Botany*, 46(3):711–727, 2021.

[41] Timothy F. Wright, Erin E. Schirtzinger, Tania Matsumoto, Jessica R. Eberhard, Gary R. Graves, Juan J. Sanchez, Sara Capelli, Heinrich Müller, Julia Scharpegge, Geoffrey K. Chambers, Robert C. Fleischer. A multilocus molecular phylogeny of the parrots (Psittaciformes): support for a Gondwanan origin during the Cretaceous. *Molecular Biology and Evolution*, 25:2141–2156, 2008.

Department of Statistics, 367 Evans Hall # 3860, U.C. Berkeley CA 94720
Email address: aldousdj@berkeley.edu
URL: www.stat.berkeley.edu/users/aldous.

Department of Mathematics, Uppsala University, PO Box 480, SE-751 06 Uppsala, Sweden
Email address: svante.janson@math.uu.se
URL: www2.math.uu.se/∼svante