Discrete Mathematics and Theoretical Computer Science DMTCS vol. **21**:4, 2019, #8

On the centroid of increasing trees

Kevin Durant Stephan Wagner*

Department of Mathematical Sciences, Stellenbosch University, South Africa

received 27th Feb. 2018, revised 29th Nov. 2018, accepted 3rd Aug. 2019.

A centroid node in a tree is a node for which the sum of the distances to all other nodes attains its minimum, or equivalently a node with the property that none of its branches contains more than half of the other nodes. We generalise some known results regarding the behaviour of centroid nodes in random recursive trees (due to Moon) to the class of very simple increasing trees, which also includes the families of plane-oriented and d-ary increasing trees. In particular, we derive limits of distributions and moments for the depth and label of the centroid node nearest to the root, as well as for the size of the subtree rooted at this node.

Keywords: centroid, increasing tree, limit distribution

1 Introduction

Increasing trees are rooted, labelled trees in which paths away from the root are labelled in increasing order. Apart from their practical relevance to problems in computer science (see Bergeron et al. (1992) for a useful summary), they possess a large structural variety that is especially interesting from a combinatorial point of view. This variety stems from a weighting scheme in which a tree's nodes are assigned weights according to their out-degrees (that is, their number of children) and the weight of the tree is defined to be the product of those of its nodes. Each set of weights gives rise to a different family of trees, $\mathcal{T}$, which may emphasise or suppress certain structural features within the class; say by restricting the possible out-degrees of nodes, or by assigning significance to the order of their children.

Given such a family of trees, one can pose combinatorial questions in terms of the subset $\mathcal{T}_n \subset \mathcal{T}$ of trees of size n (that is, the set of trees made up of n nodes), from which trees are drawn at random according to their relative weights. For example, one might be interested in the sum of the weights of the trees in this set (Bergeron et al., 1992), the height and profile of a typical tree (Drmota, 2009, Chapter 6), or the asymptotic distribution of the number of subtrees of a fixed size over all elements of $\mathcal{T}_n$ (Fuchs, 2012).

Our interest here is in centroid nodes of large, random increasing trees. We say that a node v in a graph is a *centroid* if it minimises the average distance to any other node in the graph (as opposed to a *central point* or *centre*, which minimises the maximum). An equivalent definition for trees can be given in terms of branches. The *branches* of a tree T at node v are the maximal subtrees of T that do not contain v.

*Supported by the National Research Foundation of South Africa, grant 96236.

ISSN 1365–8050

We will also use the shorthand "branches of T" for the branches of T at its root node. The members of the branch of v containing the root are *ancestral* nodes, while members of the remaining branches are *descendants*. It is well known that v is a centroid if each of the tree's branches at v contains at most half of the tree's nodes (Zelinka, 1968).

Furthermore, every tree has either one or two centroids—however the latter case only occurs when the size n of the tree is even, the centroid nodes are adjacent, and the largest branch of each has exactly $n/2$ nodes (essentially, when there is a centroid edge), see Jordan (1869).

Specifically, the probability of a random increasing tree having two centroids vanishes at a rate of $\Theta(1/n)$ as n tends to infinity (see Lemma 15), so it seems reasonable to focus solely on parameters that involve the first centroid—i.e., the one closest to the tree's root. All of the parameters we consider here—depth, label, and number of descendants—lead to random variables over the set of trees of size n, and as $n \to \infty$, these random variables converge to limits that can be described in full.

Before we can summarise these or any other known results properly, we must highlight the subclass of *very simple increasing trees*, which are particularly interesting (from a combinatorial point of view, at least) because they can be described and characterised using a straightforward growth process: begin with the root node 1, and repeatedly attach nodes to the tree according to certain probabilistic rules, determined by the family's out-degree weights. The simplest such family is that of recursive trees, in which trees are formed by attaching new nodes uniformly at random to existing ones.

1.1 New and existing results

For the family of recursive trees, Moon (2002) has presented a number of noteworthy results that detail both the depth and label of the centroid lying nearest to the root node. Similar results on simply generated trees can be found in two papers by Meir and Moon dedicated to the topic of centroid nodes (Meir and Moon, 2002; Moon, 1985). Letting $M = \lfloor (n-1)/2 \rfloor$, it turns out that the average depth of the centroid in a recursive tree of size n is $M/(n-M)$, and the expected label of the centroid is:

$$\frac{1}{2} + \frac{n(n+1)}{2(n-M)(n+1-M)}. \tag{1}$$

It follows that as n tends to infinity, the mean depth and label of the nearest centroid approach 1 and $5/2$ respectively—an early indication of the strong correlation between the root and centroid nodes of an increasing tree.

Indeed, Moon also showed that the probability that the centroid and root coincide tends to $1 - \ln 2 \approx 0.31$ as $n \to \infty$ (along with the limiting probability for an arbitrary node; see Corollary 12 below), and that the probability of the centroid having depth at least h tends to $(\ln 2)^h/h!$. Furthermore, it was shown that the probability that the ancestral branch of the centroid has B nodes is:

$$\frac{n}{(n-B)(n-B+1)} \left(1 - \sum_{b=\lceil (n+1)/2 \rceil}^{\lfloor n-B-1 \rfloor} 1/h \right), \tag{2}$$

and that the mean of the proportion of the tree accounted for by this branch approaches $(\ln 2)^2/2 \approx 0.24$.

Although exact finite expressions such as (1) and (2) do not seem to be within reach for the broader class of very simple increasing trees, we will show here that the behaviour of the centroid can be described both precisely and in some generality in the limit $n \to \infty$. In particular, our theorems will be formulated

in terms of limiting distributions and limits of moments, and all of the above (asymptotic) results will follow from them as special cases. This focus on asymptotic behaviour and full limiting distributions is one of two notable differences between our work and that of Moon, the other being methodological: whereas the above results were obtained mostly using elementary counting arguments, our exposition relies on generating functions, singularity analysis, and a variety of other tools from the field of analytic combinatorics.

Briefly, our results are as follows: the depth of the nearest centroid in a large very simple increasing tree converges to a discrete limiting distribution that is concentrated around the root, i.e., one that decreases exponentially for larger integer values. The associated moments converge as well, and we find, for example, that the expected depth of the centroid in a random plane-oriented or binary increasing tree tends to $1/2$ and 2 respectively.

Similarly, the label of the centroid converges to a discrete limiting distribution that favours values close to 1. The mean label of the centroid in plane-oriented and binary trees tends to $7/4$ and 4, respectively, and it will follow from the limiting distribution of the centroid's label that the probability of the root and centroid coinciding tends to 0.59 and 0 in these two families. In all of these results it is noticeable that the root is further from the centroid in a binary increasing family than in any other type of increasing tree.

Lastly, we show that the limiting distribution for the proportion of the tree accounted for by the centroid's ancestral branch is a combination of a point measure at 1 and a decreasing density on $[1/2, 1)$. Its expected value approaches (roughly) 0.13 and 0.38 in plane-oriented and binary increasing trees respectively. Several of these results are partially recognisable in Figure 1.

The remainder of this paper is organised as follows: we describe families of very simple increasing trees in more detail in Section 2, and then address the depth, label, and subtree size of the centroid in Sections 3, 4, and 5, respectively.

As a final introductory note, we should mention that there are several interesting results involving centroids of another, similar class of random trees—the so-called *simply generated* families. Like increasing trees these are rooted, weighted trees, but without the additional restriction on paths leading away from the root. By design, both plane (Catalan) and labelled (Cayley) trees can be seen as simply generated families.

In particular, it is known that the centroid of a large simply generated tree typically has exactly three branches of size $\Theta(n)$ (Aldous, 1994, Theorem 4), and that in the limit $n \to \infty$ each of these branches itself behaves like a scaled, random simply generated tree. Furthermore, the proportion of the tree accounted for by the centroid's ancestral branch approches 0.41, while the proportions for the remaining two branches tend to 0.44 and 0.15 (Meir and Moon, 2002). Remarkably, these results apply to the class of simply generated trees as a whole, because all simply generated families share the same limiting object: the continuum random tree (Aldous, 1991).

On the other hand, there are also a number of results that describe the behaviour of a complementary centrality measure, *betweenness centrality*, within these two classes of trees. The betweenness of a node in a tree is the number of paths passing through that node, and one can view it as an alternative to the average distance metric that underlies the definition of a centroid. In fact the average distance from a node to any other in a graph also gives rise to a centrality measure, known simply as (inverse) *closeness centrality*. From this perspective, a centroid is simply a node with maximal closeness. Both of these measures have proved useful in practical applications (Girvan and Newman, 2002; Goh et al., 2002; Shah and Zaman, 2011).

One finds, for example, that betweenness centrality must be rescaled linearly to obtain limiting distri-

butions over families of both simply generated and increasing trees, implying that betweenness is typically $\Theta(n)$ in both classes. (A notable exception is that nodes close to the root of an increasing tree tend to have betweenness that is $\Theta(n^2)$.) Moreover, the probability that the centroid also has maximal betweenness approaches 0.62 and 0.87 in labelled (simply generated) and recursive (increasing) trees respectively (Durant, 2017; Durant and Wagner, 2017).

2 Very simple increasing trees

As mentioned above, increasing trees are rooted, labelled trees that satisfy a certain increasing property. One can state this property recursively: each of an increasing tree's branches is itself an increasing tree whose labels are all larger than that of the root.

In addition, specific families of increasing trees are defined concretely by coupling a non-negative weight ϕ_i to each node according to its out-degree i, and then letting the weight $\omega(T)$ of the tree be the product of the per-node weights. It is typical to assume $\phi_0 \neq 0$, and $\phi_i > 0$ for some $i \geq 2$ (Bergeron et al., 1992). Trees are understood to be plane in this context, i.e., the order of the children of a node matters. However, for some choices of weights (notably, the choice $\phi_i = \frac{1}{i!}$, which yields recursive trees, see the discussion later), it is more natural to interpret the resulting family of weighted trees as non-plane.

With these concepts in mind, one can construct a generating function for a family $\mathcal{T}$ of increasing trees. We let y_n denote the total weight of increasing trees in $\mathcal{T}$ with n nodes. The exponential generating function will be denoted by $y(x)$:

$$y(x) = \sum_{n \geq 1} \frac{y_n}{n!} x^n.$$

Recall that the symbolic act of removing the node with the lowest label from every object in a class is represented analytically by the differential operator $y(x) \to y'(x)$ (Flajolet and Sedgewick, 2009, Section VII.9.2). We have:

$$y'(x) = \sum_{T \in \mathcal{T}} \omega(T) \frac{x^{|T|-1}}{(|T|-1)!} = \phi(y(x)), \tag{3}$$

in which $\phi(u) = \sum_i \phi_i u^i$ is the family's characteristic weight function. Note that increasing trees are labelled structures, so all of the generating functions we make use of here will be *exponential* generating functions.

The so-called very simple increasing trees are a subclass of the increasing trees defined in this way, and are important for two reasons: one, because the most familiar examples of increasing trees fall into this category; and two, because every family of very simple increasing trees can be characterised by a few simple, useful properties.

Lemma 1 (Panholzer and Prodinger (2007, Lemma 5))**.** *Let $\mathcal{T}$ be a family of increasing trees; then $\mathcal{T}$ is a family of* very simple *increasing trees if the following (equivalent) properties hold:*

- *the total weight $|\mathcal{T}_n|$ of trees of size n, also denoted by y_n, satisfies $y_{n+1}/y_n = c_1 n + c_2$ for certain $c_1, c_2 \in \mathbb{R}$;*
- *repeatedly pruning the node with the largest label from a random tree in $\mathcal{T}$ yields another, smaller, random tree; and*

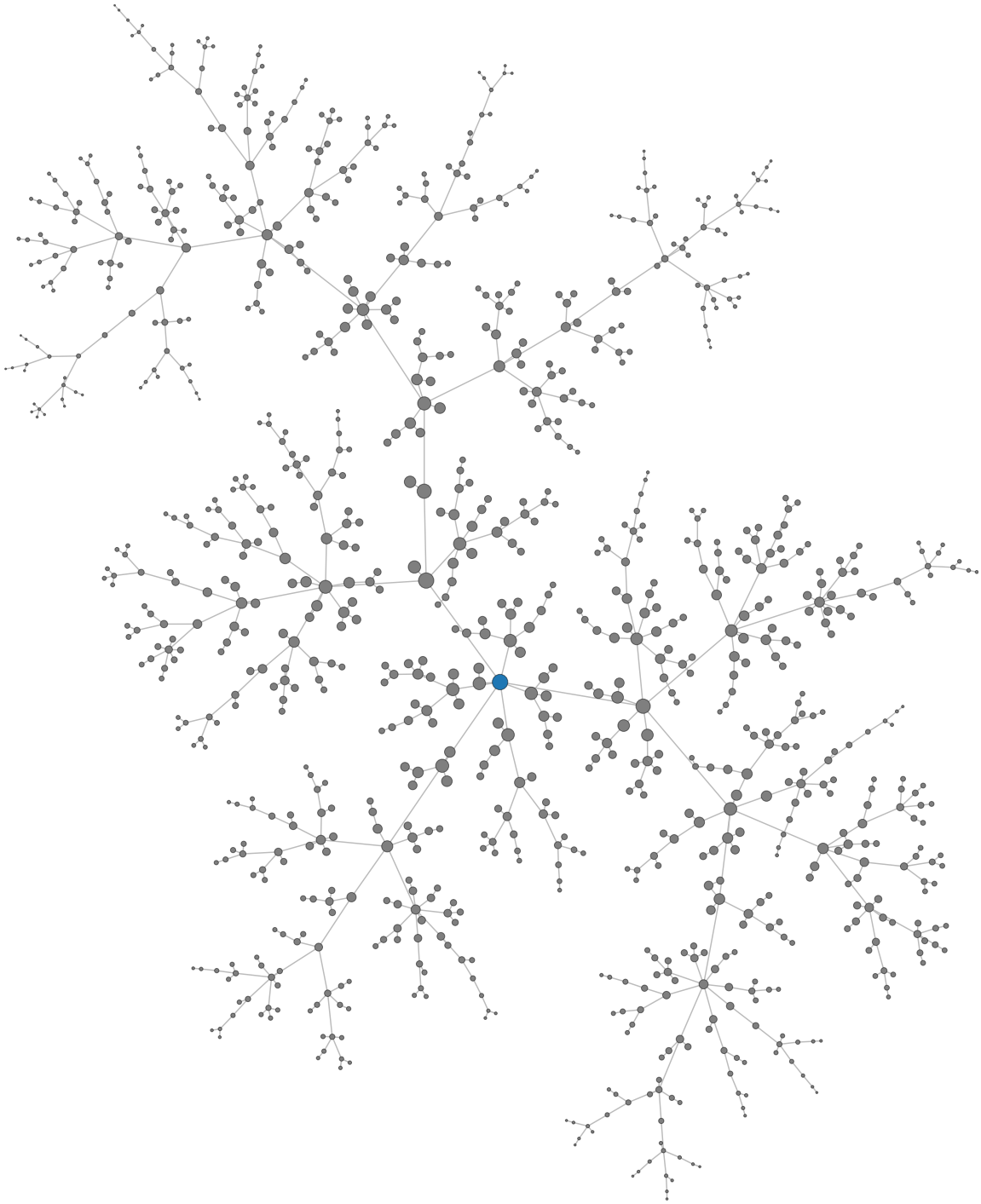

Fig. 1: A random recursive tree with $n = 1000$ nodes. The larger a node, the smaller its average distance to other nodes, and the centroid is depicted in blue. In this particular example, the label of the centroid is 2, and the root is the node of degree 4 directly to its right. (This figure can also be compared to Figures 1 and 2 of (Durant, 2017).)

- *trees can be constructed by way of a probabilistic growth process, where the n-th node is attached to one of the previous $n-1$ nodes in the n-th step, and the probability of being attached to a given node is proportional to a linear function of its outdegree.*

More directly, families of very simple increasing trees can be identified by their characteristic functions, which always correspond to one of the following:

- *general recursive* trees:
$$\phi(u) = \exp(c_1 u), \text{ with } c_1 > 0;$$
- *general plane-oriented* trees:
$$\phi(u) = (1 + c_2 u)^{1+c_1/c_2}, \text{ with } c_2 < 0 \text{ and } c_1/c_2 < -1;$$
- *general d-ary increasing* trees:
$$\phi(u) = (1 + c_2 u)^{1+c_1/c_2}, \text{ with } c_2 > 0 \text{ and } c_1/c_2 \in \mathbb{Z}_{>0}.$$

Note that in the above characteristic functions we have implicitly assumed that the weight assigned to a leaf node is 1, since the coefficient of u^0 in $\phi(u)$—which we write as $[u^0]\phi(u)$—equals 1 in all three of them. We have called these families 'general' because each one has a certain standard form, corresponding to certain fixed values of c_1 and c_2. These are:

- *recursive* trees ($c_1 = 1$):
$$\phi(u) = \exp(u) \implies y(x) = -\log(1-x);$$
- *plane-oriented* trees ($c_2 = -1$ and $c_1 = 2$):
$$\phi(u) = (1-u)^{-1} \implies y(x) = 1 - \sqrt{1-2x};$$
- *d-ary increasing* trees ($c_2 = 1$ and $c_1 = d - 1 \in \mathbb{Z}_{>0}$):
$$\phi(u) = (1+u)^d \implies y(x) = -1 + (1-(d-1)x)^{-1/(d-1)}.$$

Throughout the next few sections we will work with the more general forms, only referring to specific cases for the sake of intuition or examples.

The probabilistic growth process mentioned in Lemma 1 was described briefly for the case of recursive trees in Section 1.1, but is worth repeating in more detail here. For recursive trees, the process starts with a root node—always labelled 1—and at each step, node n is attached to one of the $n-1$ previous nodes, uniformly at random. The tree obtained after the nth step is random (relative to its weight) in $\mathcal{T}_n$. Clearly, the number of recursive trees of size n satisfies $y_n = (n-1)y_{n-1} = (n-1)!$.

The processes for plane-oriented and d-ary increasing trees are similar, but with attachment probabilities that depend on the out-degrees of the existing nodes. Firstly, in the plane-oriented case, a node with m children is viewed as having $m+1$ distinct attachment points. Since there are a total of $2n-1$ such points in a tree of size n, the number of plane-oriented trees is $y_n = (2n-3)y_{n-1} = (2n-3)!!$. Secondly,

the defining characteristic of a d-ary tree is that each of its nodes starts with d attachment points, so that $y_n = ((d-1)n+1)y_{n-1}$. Note that these counts all abide by the general rule $y_n = (c_1 n + c_2)y_{n-1}$ of Lemma 1.

Generally, the probability of attaching node n to one of the previous $n-1$ nodes whose outdegree is k is equal to $(c_1 + c_2 - c_2 k)/((n-1)c_1 + c_2)$. Note that this quotient only depends on c_2/c_1 rather than the precise values of c_1 and c_2, and that the proportionality factor $c_1 + c_2 - c_2 k$ is independent of n.

For recursive trees, the probability simplifies to $1/(n-1)$ for all nodes; for plane-oriented trees, we have $c_1 + c_2 - c_2 k = k + 1$, so that the probability of attaching a new node to an existing node increases with the number of children that node already has (so-called preferential attachment). For d-ary trees, we have $c_1 + c_2 - c_2 k = d - k$, which is decreasing in k and becomes 0 for $k = d$, so that no more nodes can be attached once a node reaches full capacity.

Remark 1. Since the probability that the new node n is attached to a specific earlier node only depends on n and the current outdegree, but not the shape of the rest of the tree, it follows that the size and shape of the subtree rooted at k are independent of the history, i.e., the shape of the tree induced by the first k nodes. This fact will be used a few times in our arguments.

2.1 A common generating function

The standard expressions for the generating function $y(x)$ that we have mentioned above are all quite familiar and amenable to analysis, and this is true of the generating functions for the more general families of very simple increasing trees as well. However, for our purposes, it will be somewhat more concise to work with the derived generating function $y'(x)$, which has a common, manageable form for all families:

$$y'(x) = (1 - c_1 x)^{-(1+(c_2/c_1))}, \text{ with } c_1 \neq 0. \tag{4}$$

This expression is not only a necessary property of very simple families, but a sufficient one as well. To see this, note that it is usual to assume that $y_n \geq 0$ for $n > 0$, along with $y_0 = 0$. Since $y_n = \prod_{j=1}^{n-1}(c_1 j + c_2)$, this is the case only if $c_1 > 0$ and $1 + c_2/c_1 > 0$, which corresponds to general recursive and plane-oriented trees when $c_2 = 0$ and $c_2 < 0$, respectively.

Another typical assumption is that $\phi_i = [u^i]\phi(u) \geq 0$ for $i > 0$, with $\phi_0 > 0$ and $\phi_i > 0$ for some $i \geq 2$. This can also be applied here, since by integrating $y'(x)$ when $c_2 > 0$ (the constant is determined using $y_0 = 0$) it follows that $y'(x) = \phi(y(x)) = (1 + c_2 y(x))^{1+(c_1/c_2)}$. This characteristic function satisfies the assumption only if $c_1/c_2 \in \mathbb{Z}_{\geq 0}$, although the case $c_1/c_2 = 0$ is generally excluded to avoid the family of 'path' trees.

Thus it is possible to derive results that are specific to very simple increasing trees by working with the derived form (4) and assuming the non-negativity of the y_n and ϕ_i. That being said, our results will remain unaffected even if we weaken the constraints on y_n and ϕ_i slightly, so for the sections to come we define α and assume that:

$$y'(x) = (1 - c_1 x)^{-\alpha}, \text{ with } \alpha = 1 + \frac{c_2}{c_1} > 0.$$

In particular, we can write $y_n = c_1^{n-1}\alpha^{\overline{n-1}}$, in which the latter term is a rising factorial. Recursive, plane-oriented, and d-ary increasing trees now correspond to $\alpha = 1$, $\alpha = 1/2$, and $\alpha = d/(d-1)$ respectively (for binary increasing trees, $\alpha = 2$).

3 The depth of the centroid

The remaining sections deal with the depth, label, and subtree size of the centroid, respectively. Intuitively, one can expect the root node to play a large part in this analysis, because the branches of a random increasing tree are typically very well balanced. To be more specific: the expected height of a random very simple increasing tree is $\Theta(\log n)$ (Drmota, 2009, Chapter 6), and the depths of its nodes follow a normal distribution with both mean and variance of order $\log n$ (Bergeron et al., 1992). (Recall, e.g., that a strict binary tree of size n has a height of *at least* $\Theta(\log n)$.) This implies that the average distance from the root to other nodes in the tree remains relatively small, leading one to surmise that the centroid will lie at least somewhat close to the root.

The starting point for all of our results is the observation that in a tree of size n, the node with label k is on the path between the root and the centroid (always considering the nearer if there are two) if and only if k has at least $\lfloor n/2 \rfloor$ descendent nodes (Moon, 2002). We let $\Lambda_k(1/2)$ mark the occurrence of this event (later, we will also introduce a more general version $\Lambda_k(\sigma)$). Our first goal will be to derive a closed form for the probability $P_n(\Lambda_k(1/2))$, which, due to its reliance on the label k, will be obtained via the exponential generating function $y^{(k)}(x)$. By combining this closed form with a known probability generating function for the depth of node k, we can describe the event that k is both at depth h and on the path from the root to the centroid, and by marginalising over k, we arrive at a generating function that yields, in the limit $n \to \infty$, the behaviour of the depth of the (nearest) centroid node.

3.1 The probability of a node appearing on the centroid path

There is, of course, a natural extension of $\Lambda_k(1/2)$ to the event $\Lambda_k(\sigma)$ that node k has at least $\lfloor \sigma n \rfloor$ descendants, where $1/2 \leq \sigma < 1$, and in fact the closed form we desire for $P_n(\Lambda_k(1/2))$ is simply a special case of a similar expression for the more general probability $P_n(\Lambda_k(\sigma))$. Although we have no need for it in determining the depth and label of the centroid, this more general version will in fact be required in Section 5, when considering the size of the centroid branch containing the root. Since the two derivations are essentially identical, we deal with the variable case immediately.

Consider the exponential generating function $y^{(k)}(x)$, which counts trees (of size at least k) as if the nodes 1 through k were 'size-less'. We have:

$$\begin{aligned} y^{(k)}(x) &= c_1^{k-1}\alpha^{\overline{k-1}}(1-c_1x)^{-(\alpha+k-1)} \\ &= y_k \cdot y'(x) \cdot (1-c_1x)^{-(k-1)}, \end{aligned}$$

where the three terms can be interpreted as accounting for the configurations of the first k nodes, the subtree rooted at node k, and the remaining subtrees, respectively. A simple combinatorial interpretation for the latter can for example be given for d-ary trees, where $c_1 = d-1$, $\alpha = d/(d-1)$ and $y(x) = -1 + (1-(d-1)x)^{-1/(d-1)}$. When the structure of the first k nodes is fixed, there are $(d-1)(k-1)$ places where a (possibly empty) subtree can be attached to one of the first $k-1$ nodes. This is represented by the exponential generating function $(1+y(x))^{(d-1)(k-1)} = (1-c_1x)^{-(k-1)}$.

With this in mind, the number (more accurately: the total weight) of trees of size n in which k has m descendent nodes is:

$$y_k \binom{n-k}{m} \big(m!\,[x^m]y'(x)\big)\Big((n-k-m)!\,[x^{n-k-m}](1-c_1x)^{-(k-1)}\Big),$$

and the proportion of trees in which k's descendants number at least $\lfloor \sigma n \rfloor$ is, for $k > 1$:

$$P_n(\Lambda_k(\sigma)) = \frac{y_k(n-k)!}{y_n} \sum_{m=\lfloor \sigma n \rfloor}^{n-k} \big([x^m]y'(x)\big)\Big([x^{n-k-m}](1-c_1x)^{-(k-1)}\Big)$$
$$= \sum_{m=\lfloor \sigma n \rfloor}^{n-k} \binom{\alpha+m-1}{m}\binom{n-m-2}{k-2}\Big/\binom{\alpha+n-2}{n-k}, \tag{5}$$

where we have used the fact that $y_n = (n-k)!\,[x^{n-k}]y^{(k)}(x)$.

As long as the label k is small relative to the tree's size n, we have an asymptotic formula for $P_n(\Lambda_k(\sigma))$ that is given in the following theorem.

Theorem 2. *For $1 < k < \lceil (1-\sigma)n \rceil$ such that $k = o(n^{1/4})$, the probability that the node with label k has at least $\lfloor \sigma n \rfloor$ descendants satisfies:*

$$P_n(\Lambda_k(\sigma)) = I_{1-\sigma}(k-1,\alpha)\left(1 + O\left(\frac{k^2}{\sqrt{n}}\right)\right), \tag{6}$$

where the error term is uniform in σ over subsets of the form $[1/2, 1-\delta)$, for $0 < \delta < 1/2$, and

$$I_x(a,b) = \frac{B(x;a,b)}{B(a,b)} = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\,dt}{\int_0^1 t^{a-1}(1-t)^{b-1}\,dt}$$

is the regularised incomplete beta function.

Proof: The main step in going from (5) to (6) is an application of the Euler-Maclaurin formula; but first, note that as n grows:

$$P_n(\Lambda_k(\sigma)) = \frac{\displaystyle\sum_{m=\lfloor \sigma n \rfloor}^{n-k} \frac{m^{\alpha-1}}{\Gamma(\alpha)}\left(1+O\left(\frac{1}{m}\right)\right)\binom{n-m-2}{k-2}}{\dfrac{(n-k)^{\alpha+k-2}}{\Gamma(\alpha+k-1)}\left(1+\dfrac{(\alpha+k-1)^{\underline{2}}}{2(n-k)}+O\left(\dfrac{k^4}{(n-k)^2}\right)\right)}$$
$$= \frac{n^{-(\alpha+k-2)}}{B(k-1,\alpha)} \sum_{m=k}^{\lceil (1-\sigma)n \rceil} (n-m)^{\alpha-1}(m-2)^{\underline{k-2}}\left(1+O\left(\frac{k^2}{n}\right)\right).$$

Splitting the sum at an intermediate value $n^{1-\varepsilon}$, where $0 < \varepsilon < 1$, reveals that the contribution from smaller values of m is minimal:

$$n^{-(\alpha+k-2)} \sum_{m=k}^{\lceil n^{1-\varepsilon} \rceil - 1} (n-m)^{\alpha-1}(m-2)^{\underline{k-2}} = O\Big(n^{-(k-1)\varepsilon}\Big). \tag{7}$$

It suffices now to apply the Euler-Maclaurin formula to the dominant portion of the sum:

$$n^{-(\alpha+k-2)} \sum_{m=\lceil n^{1-\varepsilon} \rceil}^{\lceil (1-\sigma)n \rceil} (n-m)^{\alpha-1} m^{k-2}\left(1+O\left(\frac{k^2}{m}\right)\right)$$

$$= n^{-(\alpha+k-2)} \int_{\lceil n^{1-\varepsilon} \rceil}^{\lceil (1-\sigma)n \rceil} (n-u)^{\alpha-1} u^{k-2}\, du \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right) + O\left(\frac{1}{n}\right)$$
$$= \int_{n^{-\varepsilon}+O(1/n)}^{1-\sigma+O(1/n)} t^{k-2}(1-t)^{\alpha-1}\, dt \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right). \tag{8}$$

Note that in absorbing the error term $O(1/n)$ we have implicitly treated σ as a constant with respect to n, and as such, the uniformity of the error term only holds as long as σ is not allowed to tend arbitrarily close to 1. Also, as long as $\varepsilon \geq 1/k$, the term $k^2/n^{1-\varepsilon}$ is not smaller in order than that of equation (7), and the contribution of the first portion of the sum can be ignored. This is the case for all $k > 1$ when $1/2 \leq \varepsilon < 1$.

As n grows, the bounds of the integral in (8) approach 0 and $1-\sigma$ at the following rates:

$$\int_0^{n^{-\varepsilon}+O(1/n)} t^{k-2}(1-t)^{\alpha-1}\, dt = O\left(n^{-(k-1)\varepsilon}\right),$$

and

$$\int_{1-\sigma}^{1-\sigma+O(1/n)} t^{k-2}(1-t)^{\alpha-1}\, dt = O\left(\frac{1}{n}\right).$$

Noting that these terms are also of orders lower than $k^2/n^{1-\varepsilon}$ when $1/2 \leq \varepsilon < 1$, we see that the probability of node k having at least $\lfloor \sigma n \rfloor$ descendants can be written, for $1 < k = o(n^{(1-\varepsilon)/2})$, as:

$$P_n(\Lambda_k(\sigma)) = \frac{1}{B(k-1,\alpha)} \int_0^{1-\sigma} t^{k-2}(1-t)^{\alpha-1}\, dt \left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right)$$
$$= I_{1-\sigma}(k-1,\alpha)\left(1 + O\left(\frac{k^2}{n^{1-\varepsilon}}\right)\right).$$

□

We mention also that an alternative representation of $P_n(\Lambda_k(\sigma))$ can be obtained using the binomial theorem, since

$$\int_0^{1-\sigma} t^{k-2}(1-t)^{\alpha-1}\, dt = B(k-1,\alpha) - \int_{1-\sigma}^1 t^{k-2}(1-t)^{\alpha-1}\, dt$$
$$= B(k-1,\alpha) - \sum_{l=0}^{k-2} \binom{k-2}{l} \frac{(-1)^l}{l+\alpha} \sigma^{l+\alpha}.$$

Corollary 3. *For $1 < k = o(n^{1/4})$, the probability that node k is on the path from the root to the (nearest) centroid node satisfies:*

$$P_n(\Lambda_k(1/2)) = I_{1/2}(k-1,\alpha)\left(1 + O\left(\frac{k^2}{\sqrt{n}}\right)\right).$$

Finally, we give limiting probabilities for the event that node k is on the path to the centroid in the two simplest families of very simple increasing trees, for which the incomplete beta function can be easily simplified:

Corollary 4. *For recursive trees:*

$$\lim_{n\to\infty} P_n(\Lambda_k(1/2)) = I_{1/2}(k-1,1) = 2^{-(k-1)}.$$

For binary increasing trees:

$$\lim_{n\to\infty} P_n(\Lambda_k(1/2)) = I_{1/2}(k-1,2) = (k+1)2^{-k}.$$

3.2 A uniform bound on the path probability

In addition to the asymptotic expression for $P_n(\Lambda_k(\sigma))$ given above, we can show that the probability of a specific node k appearing on the path to the centroid not only vanishes for large k, but does so exponentially in k and uniformly over n. In particular:

$$P_n(\Lambda_k(1/2)) \le \frac{\alpha^{\overline{k-1}}}{(k-1)!} 2^{-(k-2)}. \tag{9}$$

It is this fact, in combination with Corollary 3, that will allow us to derive limiting distributions for events that depend on $P_n(\Lambda_k(1/2))$.

Once again a more general form of this result will be required later, in Section 5, so to avoid a repeated derivation we give the version for variable σ here.

Lemma 5. *For $k \ge 1$ and $1/2 \le \sigma < 1$, the probability that node k has at least $\lfloor \sigma n \rfloor$ descendants in a tree of size $n \ge 3$ is subject to an upper bound that decreases exponentially with k:*

$$P_n(\Lambda_k(\sigma)) \le \frac{3}{\sigma} \frac{\alpha^{\overline{k-1}}}{(k-1)!} (1-\sigma)^{k-1}. \tag{10}$$

Proof: Firstly, take note of the following inequality involving binomial coefficients: if $\alpha \in \mathbb{R}_{\ge 0}$ and $m \le n \in \mathbb{Z}_{\ge 0}$, then:

$$\begin{aligned}
\binom{m-1+\alpha}{m} &= \frac{(\alpha+m-1)^{\underline{m}}}{m!} \\
&\le \frac{(\alpha+n)\cdots(\alpha+m)}{n\cdots m} \frac{(\alpha+m-1)^{\underline{m}}}{m!} \\
&= \frac{n+\alpha}{m} \binom{n-1+\alpha}{n}.
\end{aligned} \tag{11}$$

Since $P_n(\Lambda_1(\sigma)) = 1$, and $P_n(\Lambda_k(\sigma)) = 0$ whenever $k > \lceil (1-\sigma)n \rceil$, we need only consider $1 < k \le \lceil (1-\sigma)n \rceil$. In this case:

$$\begin{aligned}
P_n(\Lambda_k(\sigma)) &= \sum_{m=\lfloor \sigma n \rfloor}^{n-k} \binom{m-1+\alpha}{m} \binom{n-m-2}{k-2} \Big/ \binom{n-2+\alpha}{n-k} \\
&\le \sum_{m=\lfloor \sigma n \rfloor}^{n-k} \frac{n-k+\alpha}{m} \binom{n-k-1+\alpha}{n-k} \binom{n-m-2}{k-2} \Big/ \binom{n-2+\alpha}{n-k}
\end{aligned}$$

$$\begin{aligned}
&\leq \sum_{m=\lfloor \sigma n \rfloor}^{n-k} \frac{n-k+\alpha}{\lfloor \sigma n \rfloor} \binom{n-k-1+\alpha}{n-k} \binom{n-m-2}{k-2} \Bigg/ \binom{n-2+\alpha}{n-k} \\
&= \frac{n-k+\alpha}{\lfloor \sigma n \rfloor} \binom{n-k-1+\alpha}{n-k} \binom{\lceil (1-\sigma)n \rceil - 1}{k-1} \Bigg/ \binom{n-2+\alpha}{n-k} \\
&= \frac{n-k+\alpha}{\lfloor \sigma n \rfloor} \frac{(\lceil (1-\sigma)n \rceil - 1)^{\underline{k-1}}}{(k-1)!} \frac{\Gamma(n-k+\alpha)}{\Gamma(n-1+\alpha)} \frac{\Gamma(k-1+\alpha)}{\Gamma(\alpha)} \\
&= \frac{\alpha^{\overline{k-1}}}{(k-1)!} \frac{n-k+\alpha}{\lfloor \sigma n \rfloor} \frac{(\lceil (1-\sigma)n \rceil - 1)^{\underline{k-1}}}{(n-2+\alpha)^{\underline{k-1}}} \\
&\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!} \frac{n-k+\alpha}{n-2} \frac{n(n-2)\cdots(n-2k+4)}{(n-2+\alpha)\cdots(n-k+\alpha)} \frac{(1-\sigma)^{k-1}}{\sigma} \\
&\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!} \frac{n}{n-2} \frac{(1-\sigma)^{k-1}}{\sigma}.
\end{aligned}$$

This yields the stated bound whenever $n \geq 3$. In the specific case $\sigma = 1/2$, the final two lines can be improved slightly, resulting in equation (9):

$$\begin{aligned}
\frac{\alpha^{\overline{k-1}}}{(k-1)!} \frac{n-k+\alpha}{\lfloor n/2 \rfloor} \frac{(\lceil n/2 \rceil - 1)^{\underline{k-1}}}{(n-2+\alpha)^{\underline{k-1}}} &\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!} \frac{(\lceil n/2 \rceil - 2)^{\underline{k-2}}}{(n-2+\alpha)^{\underline{k-2}}} \\
&= \frac{\alpha^{\overline{k-1}}}{(k-1)!} \prod_{j=1}^{k-2} \frac{\lceil n/2 \rceil - 1 - j}{n-1+\alpha-j} \\
&\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!} \prod_{j=1}^{k-2} \frac{(n+1)/2 - 1 - j}{n-1-j} \\
&= \frac{\alpha^{\overline{k-1}}}{(k-1)!} 2^{-(k-2)} \prod_{j=1}^{k-2} \frac{n-1-2j}{n-1-j} \\
&\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!} 2^{-(k-2)}.
\end{aligned}$$

□

3.3 A limiting distribution for the depth of the centroid

Let $D = D(T)$ denote the depth of the centroid—that is, the number of edges on the path from the root to the centroid node (the nearer if there are two)—in a random tree T. As mentioned earlier, $D(T) \geq h$ if and only if there is a vertex at depth h that is on this path. Breaking this event down per label, we may say that the depth of the centroid is at least h if and only if for some label k, node k has both depth h and is present on the path to the centroid. Since these per-label events are disjoint, and the depth of node k is

independent of the size of the subtree rooted at k (see Remark 1), we have

$$P_n(D \geq h) = \sum_{k \geq 1} P_n(D_k = h) P_n(\Lambda_k(1/2)),$$

in which D_k is a random variable over the possible depths $h \in \mathbb{Z}_{\geq 0}$ of node k. This random variable has a known probability generating function (Panholzer and Prodinger, 2007):

$$\sum_{h \geq 0} P_n(D_k = h) v^h = \prod_{j=0}^{k-2} \frac{\alpha v + j}{\alpha + j} = \frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}, \tag{12}$$

which is independent of n, implying that $P_n(D_k = h) = P(D_k = h)$ is as well. Combining these two expressions yields a (complementary cumulative) probability generating function for the depth of the centroid:

$$\begin{aligned} C_n(v) = \sum_{h \geq 0} P_n(D \geq h) v^h &= \sum_{k \geq 1} P_n(\Lambda_k(1/2)) \sum_{h \geq 0} P(D_k = h) v^h \\ &= \sum_{k \geq 1} P_n(\Lambda_k(1/2)) \frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}. \end{aligned} \tag{13}$$

Our goal is to make use of the asymptotic form of $P_n(\Lambda_k(1/2))$ given in Corollary 3 to find the limit of this generating function, and then simply extract the desired probabilities as coefficients.

Theorem 6. *The depth $D(\mathcal{T}_n)$ of the centroid node in a random tree of size n converges in probability to the discrete random variable $\mathcal{D}$ supported by $\mathbb{Z}_{\geq 0}$ and with cumulative distribution function:*

$$P(\mathcal{D} \geq h) = \left(\frac{\alpha}{\alpha - 1}\right)^h \left(1 - 2^{1-\alpha} \sum_{j=0}^{h-1} \frac{((\alpha - 1)\ln 2)^j}{j!}\right)$$

and mass function:

$$P(\mathcal{D} = h) = \frac{\alpha^h}{(\alpha - 1)^{h+1}} \left[2^{1-\alpha} \left(\sum_{j=0}^{h-1} \frac{\big((\alpha - 1)\ln 2\big)^j}{j!} + \frac{\alpha\big((\alpha - 1)\ln 2\big)^h}{h!}\right) - 1\right].$$

In the case $\alpha = 1$ (corresponding to recursive trees), the expressions are undefined. Singular cases such as this—which will appear throughout this paper—can be derived by direct calculations or as limits of the more general cases.

Corollary 7 (Moon (2002))**.** *For recursive trees:*

$$\lim_{n \to \infty} P_n(D \geq h) = \frac{(\ln 2)^h}{h!}$$

and

$$\lim_{n \to \infty} P_n(D = h) = \frac{(\ln 2)^h}{h!} - \frac{(\ln 2)^{h+1}}{(h+1)!}.$$

Proof of Theorem 6: Letting n be large, and fixing $K = \lfloor n^{1/4-\varepsilon} \rfloor$ for an arbitrarily small $\varepsilon > 0$, Corollary 3 and equation (13) above imply:

$$C_n(v) = 1 + \sum_{k=2}^{K} I_{1/2}(k-1,\alpha)\left(1 + O\left(\frac{k^2}{\sqrt{n}}\right)\right)\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}} + \sum_{k=K+1}^{\lceil n/2 \rceil} P_n(\Lambda_k(1/2))\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}.$$

As $n \to \infty$, assuming $|v| < 1$, the second sum tends to 0: recalling the inequalities (9) and (11), we get

$$\begin{aligned}
\sum_{k=K+1}^{\lceil n/2 \rceil}\left|P_n(\Lambda_k(1/2))\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}\right| &\leq \sum_{k=K+1}^{\lceil n/2 \rceil}\frac{(\alpha v)^{\overline{k-1}}}{(k-1)!}2^{-(k-2)} \\
&\leq \sum_{k=K+1}^{\lceil n/2 \rceil}\frac{\alpha^{\overline{k-1}}}{(k-1)!}2^{-(k-2)} \\
&= 2\sum_{k=K}^{\lceil n/2 \rceil - 1}\binom{k-1+\alpha}{k}2^{-k} \\
&\leq n\frac{\frac{n}{2}+\alpha}{\lfloor n^{1/4-\varepsilon} \rfloor}\binom{\frac{n}{2}-1+\alpha}{\frac{n}{2}}2^{-\lfloor n^{1/4-\varepsilon} \rfloor} \\
&= O\left(\frac{n^{\alpha+3/4+\varepsilon}}{2^{\lfloor n^{1/4-\varepsilon} \rfloor}}\right) \xrightarrow[n\to\infty]{} 0,
\end{aligned}$$

where we have made use of the upper bound on $P_n(\Lambda_k(1/2))$ given in equation (9), as well as the binomial inequality (11). Similarly, the tail which replaces this sum is negligible:

$$\begin{aligned}
\sum_{k>K}\left|I_{1/2}(k-1,\alpha)\frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}}\right| &\leq \sum_{k>K} B(k-1,\alpha)^{-1}\int_0^{1/2} t^{k-2}(1-t)^{\alpha-1}\,dt \\
&\leq \alpha\sum_{k>K}\binom{\alpha+k-2}{k-2}2^{-(k-1)} \\
&= O\big(K^\alpha 2^{-K}\big) \xrightarrow[n\to\infty]{} 0.
\end{aligned}$$

Thus, for values of v within the unit circle, the pointwise limit of $C_n(v)$ is:

$$\begin{aligned}
C(v) = \lim_{n\to\infty} C_n(v) &= 1 + \sum_{k\geq 2} I_{1/2}(k-1,\alpha)\frac{\Gamma(\alpha v+k-1)\Gamma(\alpha)}{\Gamma(\alpha v)\Gamma(\alpha+k-1)} \\
&= 1 + \sum_{k\geq 2}\frac{\Gamma(\alpha v+k-1)}{\Gamma(\alpha v)\Gamma(k-1)}\int_0^{1/2} t^{k-2}(1-t)^{\alpha-1}\,dt
\end{aligned}$$

$$= 1 + \alpha v \int_0^{1/2} (1-t)^{\alpha-1} \sum_{k\geq 2} \binom{\alpha v + k - 2}{k-2} t^{k-2}\, dt$$
$$= 1 + \alpha v \int_0^{1/2} (1-t)^{-\alpha(v-1)-2}\, dt$$
$$= 1 + \frac{\alpha v}{\alpha(v-1)+1}\Big(2^{\alpha(v-1)+1} - 1\Big). \tag{14}$$

Considering now the probability generating function involving the mass $P_n(D = h)$—say $A_n(v)$—we see that this too converges:

$$A_n(v) = \sum_{h\geq 0} P_n(D=h)v^h = C_n(v) - \frac{C_n(v)-1}{v}$$
$$\to C(v) - \frac{C(v)-1}{v}$$
$$= 1 + \frac{\alpha(v-1)}{\alpha(v-1)+1}\Big(2^{\alpha(v-1)+1} - 1\Big) = A(v). \tag{15}$$

Since pointwise convergence of probability generating functions (in this case $A_n(v) \to A(v)$) implies convergence in probability of their distributions (Flajolet and Sedgewick, 2009, Theorem IX.1), we have the stated theorem via $P_n(D \geq h) \to [v^h]C(v)$. The limiting mass function follows naturally.

Let us finally remark that for the special case $\alpha = 1$, equation (14) directly yields

$$\lim_{n\to\infty} P_n(D \geq h) = [v^h]2^v = \frac{(\ln 2)^h}{h!}.$$

□

3.4 Moments of the depth distribution

To close our discussion of the centroid's depth, we consider the moments of $D(T)$ as the size of the tree tends to infinity. More specifically, with $C_n(v)$ and $A_n(v)$ as they were in the proof of Theorem 6 (along with their respective limits), we are interested in:

$$\lim_{n\to\infty} E_n(D^{\underline{m}}) = \lim_{n\to\infty} \sum_{h\geq m} h^{\underline{m}} P_n(D=h) = \lim_{n\to\infty} A_n^{(m)}(1).$$

We show firstly that the moments of $P_n(D = h)$ converge to those of its limiting distribution $P(D = h)$, and then, instead of dealing with $A_n(v)$ directly, derive the moments' asymptotic behaviour using the limiting generating function $A(v)$.

Lemma 8. *The moments of the distribution of the centroid's depth $D(\mathcal{T}_n)$ converge to those of $\mathcal{D}$, i.e.:*

$$\lim_{n\to\infty} E_n(D^m) = E(\mathcal{D}^m).$$

Proof: This follows from Lemma 5 and Lebesgue's dominated convergence theorem, which states that if (f_n) is a sequence of real-valued functions, and g a function such that $|f_n| \leq g$ for all n, then if $\int g < \infty$, one has $\lim_{n\to\infty} \int f_n = \int \lim_{n\to\infty} f_n$.

For our purposes, let $f_n(h) = h^{\underline{m}} P_n(D = h)$. The bound on $P_n(\Lambda_k(1/2))$ given in equation (9) then leads to a similar one (also uniform over n) on $P_n(D = h)$, as follows (recall that $[v^h]C_n(v) = P_n(D \geq h)$):

$$\begin{aligned} P_n(D = h) \leq [v^h]C_n(v) &= [v^h] \sum_{k\geq 1} P_n(\Lambda_k(1/2)) \frac{(\alpha v)^{\overline{k-1}}}{\alpha^{\overline{k-1}}} \\ &\leq [v^h] \sum_{k\geq 1} \binom{\alpha v + k - 2}{k-1} 2^{-(k-2)} \\ &= [v^h] 2^{1+\alpha v} = 2\frac{(\alpha \ln 2)^h}{h!}. \end{aligned}$$

Since the range of the random variable D is countable, the integrals in Lebesgue's dominated convergence theorem become sums, and we are left with

$$2 \sum_{h\geq 0} h^{\underline{m}} \frac{(\alpha \ln 2)^h}{h!} = (\alpha \ln 2)^m \, 2^{\alpha+1} < \infty.$$

Thus the factorial moments of the distributions $P_n(D = h)$ converge to those of $P(\mathcal{D} = h)$, and since the usual higher-order moments $E(\mathcal{D}^m)$ are (finite) linear combinations of the factorial moments, convergence holds for them as well. □

With convergence established, all that remains is to compute the moments of $\mathcal{D}$ by making use of its probability generating function $A(v) = \sum_{h\geq 0} P(\mathcal{D} = h)v^h$, since $E(\mathcal{D}^{\underline{m}}) = A^{(m)}(1)$.

Theorem 9. *The limit of the mth factorial moment of the centroid's depth $D(\mathcal{T}_n)$ is given by:*

$$E(\mathcal{D}^{\underline{m}}) = m\alpha^m \left(2 \sum_{j=0}^{m-2} \binom{m-1}{j} (-1)^j j! \, (\ln 2)^{m-1-j} + (-1)^{m-1}(m-1)! \right).$$

In particular, the limits of its mean and variance are:

$$\begin{aligned} E(\mathcal{D}) &= \alpha, \\ V(\mathcal{D}) &= \alpha^2(4 \ln 2 - 3) + \alpha. \end{aligned}$$

Proof: The calculation can be simplified slightly by writing the derivative of the expression in equation (15) as:

$$\begin{aligned} A^{(m)}(v) &= \frac{d^m}{dv^m} \left[\alpha(v-1) \cdot \big(1 + \alpha(v-1)\big)^{-1} \cdot \Big(2^{\alpha(v-1)+1} - 1\Big) \right] \\ &= \frac{d^m}{dv^m} [a(v) \cdot b(v) \cdot c(v)] \end{aligned}$$

$$= \sum_{i+j+k=m} \binom{m}{i,j,k} a^{(i)}(v)\, b^{(j)}(v)\, c^{(k)}(v).$$

Since we are interested in $v = 1$, note that $a'(1) = \alpha$, but $a^{(i)}(1) = 0$ when $i \neq 1$. Furthermore, $b^{(j)}(1) = (-\alpha)^j j!$ for $j \geq 0$, and $c^{(k)}(1) = 2(\alpha \ln 2)^k$ for $k > 0$, whereas $c(1) = 1$. This leads to:

$$\begin{aligned} E(\mathcal{D}^{\underline{m}}) &= m\alpha \sum_{j+k=m-1} \binom{m-1}{j,k} b^{(j)}(1)\, c^{(k)}(1) \\ &= m\alpha^m \left(2 \sum_{j=0}^{m-2} \binom{m-1}{j} (-1)^j j!\, (\ln 2)^{m-1-j} + (-1)^{m-1}(m-1)! \right). \end{aligned}$$

The mean is computed more simply as $C(1) - 1 = \alpha$, and the second factorial moment is $E(\mathcal{D}^{\underline{2}}) = 2\alpha^2(2 \ln 2 - 1)$. □

Finally, we make two small remarks again: that the limits of the mean and variance of the depth of the centroid in a random recursive tree (1 and $4 \ln 2 - 2$ respectively) were previously given by Moon (2002); and that Theorem 9 implies that the mean and variance of the centroid's depth are greatest (in the limit) in the case of binary increasing trees ($\alpha = 2$).

4 The label of the centroid

Our second task regarding the centroid of an increasing tree is to describe its label, which we will denote by $L = L(T)$.

Since we will have no need for the general event $\Lambda_k(\sigma)$ throughout this section, let us adopt the shorthand $\Lambda_k = \Lambda_k(1/2)$ to denote the presence of label k on the path between the root and centroid nodes. A key observation is that the event $L(T) = k$ can be expressed in terms of the presence (or lack thereof) of nodes $k, k+1, \ldots$ on this path, namely:

$$\begin{aligned} P_n(L = k) = P_n(\Lambda_k) &- P_n(\Lambda_k \cap \Lambda_{k+1}) \\ &- P_n(\Lambda_k \cap \overline{\Lambda}_{k+1} \cap \Lambda_{k+2}) \\ &- P_n(\Lambda_k \cap \overline{\Lambda}_{k+1} \cap \overline{\Lambda}_{k+2} \cap \Lambda_{k+3}) \\ &- \cdots . \end{aligned} \tag{16}$$

Here $\overline{\Lambda}_l$ is the complement of Λ_l, i.e., it is the event that node l is *not* on the path to the centroid. Equation (16) simply states that the centroid has label k if and only if k is on the path to the centroid, but none of the nodes $k+1, k+2, \ldots$ are. One can write a similar expression for the probability that the centroid's label is at least k:

$$\begin{aligned} P_n(L \geq k) = P_n(\Lambda_k) &+ P_n(\overline{\Lambda}_k \cap \Lambda_{k+1}) \\ &+ P_n(\overline{\Lambda}_k \cap \overline{\Lambda}_{k+1} \cap \Lambda_{k+2}) \\ &+ P_n(\overline{\Lambda}_k \cap \overline{\Lambda}_{k+1} \cap \overline{\Lambda}_{k+2} \cap \Lambda_{k+3}) \\ &+ \cdots . \end{aligned} \tag{17}$$

The composite event $\Lambda_k \cap \overline{\Lambda}_{k+1} \cap \cdots \cap \overline{\Lambda}_{k+j-1} \cap \Lambda_{k+j}$ in equation (16) requires that k and $k+j$ appear on the path to the centroid, but none of $k+1, \ldots, k+j-1$ do. This occurs if and only if $k+j$ is on the path and has node k as its parent. (This is a simpler condition than the one required by equation (17), which would be that node $k+j$ is on the path and its parent is any one of the nodes $1, \ldots, k-1$.)

Let $A_l(T)$ be the random variable that yields node l's parent, which, if we consider the increasing tree's probabilistic growth process, is the node l was 'attached' to. Then we are interested in $P_n(\Lambda_{k+j} \cap (A_{k+j} = k))$, for fixed k and j, as $n \to \infty$. Because the size of the subtree consisting of node $k+j$ and its descendants is independent of the node to which $k+j$ was attached (see Remark 1), we have:

$$P_n\big(\Lambda_{k+j} \cap (A_{k+j} = k)\big) = P_n(\Lambda_{k+j}) P(A_{k+j} = k), \tag{18}$$

where the second probability in the product is independent of the size n of the tree. Since by Corollary 3 we already know the asymptotic behaviour of $P_n(\Lambda_{k+j})$, we need an expression for the probability that node $k+j$ attaches to node k. Such an expression can be found in the literature:

Lemma 10 (see Dobrow and Smythe (1996); Kuba and Wagner (2010))**.** *For $k, j \in \mathbb{Z}_{>0}$, the probability that the parent of node $k+j$ has label k is given by:*

$$P(A_{k+j} = k) = \frac{\alpha k^{\overline{j-1}}}{(\alpha + k - 1)^{\overline{j}}},$$

which does not depend on the size of the tree.

Note that in the case of recursive trees ($\alpha = 1$), the probability of a particular attachment is $P(A_{k+j} = k) = 1/(k+j-1)$, which agrees with the family's growth process.

4.1 A limiting distribution for the label of the centroid

Following on from equations (16) and (18), we can write the probability of the centroid assuming a certain label k in terms of the events Λ_k and $A_{k+j} = k$ (for which we have closed-form expressions):

$$P_n(L = k) = P_n(\Lambda_k) - \sum_{j \geq 1} P_n(\Lambda_{k+j}) P_n(A_{k+j} = k).$$

As a consequence, we arrive at the desired result of this section:

Theorem 11. *The label $L(\mathcal{T}_n)$ of the centroid node in a random tree of size n converges in probability to a discrete random variable $\mathcal{L}$ supported by $\mathbb{Z}_{\geq 0}$ and with mass function:*

$$P(\mathcal{L} = k) = \begin{cases} 1 - \frac{\alpha}{\alpha - 1}\big(1 - 2^{-(\alpha-1)}\big) & \text{if } k = 1, \\ I_{1/2}(k-1, \alpha) - \frac{\alpha}{\alpha-1} I_{1/2}(k, \alpha - 1) & \text{otherwise.} \end{cases}$$

An alternative form, which holds for $k \geq 1$, is given by:

$$P(\mathcal{L} = k) = -\frac{1}{\alpha - 1} I_{1/2}(k, \alpha) + \left(1 + \frac{\alpha}{\alpha - 1}\right) \binom{\alpha + k - 2}{k - 1} 2^{-(\alpha + k - 1)}. \tag{19}$$

Proof: Recalling the asymptotic expression for $P_n(\Lambda_k)$ (Corollary 3), assume now that n is large, and fix J so that $k + J = \lfloor n^{1/4-\varepsilon} \rfloor$, for some arbitrarily small, positive ε. Then:

$$\begin{aligned} P_n(L = k) &= P_n(\Lambda_k) - \sum_{j=1}^{\lceil n/2 \rceil - k} P_n(\Lambda_{k+j}) P_n(A_{k+j} = k) \\ &= P_n(\Lambda_k) - \sum_{j=1}^{J} I_{1/2}(k+j-1, \alpha) \frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}} \left(1 + O\left(\frac{(k+j)^2}{\sqrt{n}}\right)\right) \\ &\quad - \sum_{j=J+1}^{\lceil n/2 \rceil - k} P_n(\Lambda_{k+j}) \frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}}, \end{aligned}$$

As in the proof of Theorem 6, which dealt with the depth of the centroid, the upper bound for $P_n(\Lambda_k)$ given in equation (9) implies that the sum over larger labels vanishes as n grows:

$$\begin{aligned} \sum_{j=J+1}^{\lceil n/2 \rceil - k} P_n(\Lambda_{k+j}) \frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}} &\leq \sum_{j=J+1}^{\lceil n/2 \rceil - k} \frac{\alpha^{\overline{k+j-1}}}{(k+j-1)!} \frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}} 2^{-(k+j-2)} \\ &= \frac{\alpha^{\overline{k-1}}}{(k-1)!} \sum_{j=J+1}^{\lceil n/2 \rceil - k} \frac{\alpha}{k+j-1} 2^{-(k+j-2)} \\ &\leq \frac{\alpha^{\overline{k-1}}}{(k-1)!} \frac{\alpha n}{\lfloor n^{1/4-\varepsilon} \rfloor} 2^{-\lfloor n^{1/4-\varepsilon} \rfloor} \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

Also, the extension of the first sum to an infinite one is permissible, since:

$$\begin{aligned} \sum_{j>J} I_{1/2}(k+j-1, \alpha) \frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}} &= \sum_{j>J} \frac{\alpha \Gamma(\alpha+k-1)}{\Gamma(\alpha)\Gamma(k)} \int_0^{1/2} t^{k+j-2} (1-t)^{\alpha-1} \, dt \\ &\leq \alpha \binom{\alpha+k-2}{k-1} \sum_{j>J} 2^{-(k+j-1)} \\ &= O\left(2^{-(k+J-1)}\right) \xrightarrow[n\to\infty]{} 0. \end{aligned}$$

Letting $n \to \infty$, and assuming $k > 1$, we obtain the limiting probability $P(\mathcal{L} = k)$:

$$\begin{aligned} P(\mathcal{L} = k) &= \lim_{n\to\infty} P_n(L = k) \\ &= I_{1/2}(k-1,\alpha) - \sum_{j\geq 1} I_{1/2}(k+j-1,\alpha)\frac{\alpha k^{\overline{j-1}}}{(\alpha+k-1)^{\overline{j}}} \\ &= I_{1/2}(k-1,\alpha) - \alpha\frac{\Gamma(\alpha+k-1)}{\Gamma(k)\Gamma(\alpha)}\int_0^{1/2}(1-t)^{\alpha-1}\sum_{j\geq 1} t^{k+j-2}\,dt \qquad (20) \\ &= I_{1/2}(k-1,\alpha) - \frac{\alpha}{\alpha-1}B(k,\alpha-1)^{-1}\int_0^{1/2} t^{k-1}(1-t)^{\alpha-2}\,dt \\ &= I_{1/2}(k-1,\alpha) - \frac{\alpha}{\alpha-1}I_{1/2}(k,\alpha-1). \end{aligned}$$

When $k = 1$, the first incomplete beta function is replaced by 1. The consolidated form given in equation (19) is due to the following property of the incomplete beta function:

$$I_x(a-1,b) - \frac{x^{a-1}(1-x)^b}{(a-1)B(a-1,b)} = I_x(a,b) = I_x(a,b-1) + \frac{x^a(1-x)^{b-1}}{(b-1)B(a,b-1)},$$

□

Repeating (20) for $\alpha = 1$ yields:

Corollary 12 (Moon (2002))**.** *For recursive trees:*

$$\lim_{n\to\infty} P_n(L = k) = 2^{-(k-1)} - \sum_{j\geq k}\frac{2^{-j}}{j}.$$

In particular (for recursive trees), $\lim_n P_n(L = 1) = 1 - \ln 2$.

4.2 Moments of the label distribution

Just as we did when dealing with the depth of the centroid, we can apply Lebesgue's dominated convergence theorem to prove convergence of the moments of $P_n(L = k)$ to those of $P(\mathcal{L} = k)$, and then derive their limits using the more convenient form of the limiting distribution. In the present case of the centroid's label, however, the proof of convergence is almost immediate.

Lemma 13. *The moments of the distribution of the centroid's label $L(\mathcal{T}_n)$ converge to those of $\mathcal{L}$, i.e.:*

$$\lim_{n\to\infty} E_n(L^m) = E(\mathcal{L}^m).$$

Proof: The line of argument is the same as that which was used for Lemma 8: we must find a uniform bound $g(k)$ for $k^{\underline{m}}P_n(L = k)$ such that $\sum_k k^{\underline{m}} g(k)$ converges. Once again by equation (9):

$$\sum_{k\geq 1} k^{\underline{m}} P_n(L = k) \leq \sum_{k\geq m} k^{\underline{m}} P_n(\Lambda_k)$$

$$\begin{aligned}
&\leq \sum_{k\geq m} \frac{k^{\underline{m}}\,\alpha^{\overline{k-1}}}{(k-1)!} 2^{-(k-2)} \\
&= 2^{-(m-2)} \sum_{k\geq m-1} \binom{\alpha+k-1}{k} (k+1)^{\underline{m}}\, 2^{-(k-m+1)} \\
&= 2^{-(m-2)} \frac{d^m}{du^m} \left[u(1-u)^{-\alpha}\right]_{u=1/2} < \infty.
\end{aligned}$$

□

Theorem 14. *The limit of the mth factorial moment of the centroid's label $L(\mathcal{T}_n)$ is given by:*

$$E(\mathcal{L}^{\underline{m}}) = \frac{4m^2+2\alpha m+\alpha-2}{m+1}\alpha^{\overline{m-1}}.$$

In particular, the limits of its mean and variance are:

$$\begin{aligned}
E(\mathcal{L}) &= 1+\frac{3}{2}\alpha, \\
V(\mathcal{L}) &= -\frac{7}{12}\alpha^2+\frac{19}{6}\alpha.
\end{aligned}$$

Proof: The factorial moments of the limiting distribution can be computed directly using equation (19):

$$\begin{aligned}
\sum_{k\geq 1} k^{\underline{m}} P(\mathcal{L}=k) &= -\frac{1}{\alpha-1}\sum_{k\geq 1} k^{\underline{m}} I_{1/2}(k,\alpha) \\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right)\sum_{k\geq 1}\binom{\alpha+k-2}{k-1} k^{\underline{m}}\, 2^{-(\alpha+k-1)} \\
&= -\frac{1}{\alpha-1}\int_0^{1/2}(1-t)^{\alpha-1}\sum_{k\geq 1}\binom{\alpha+k-1}{k-1}\alpha k^{\underline{m}} t^{k-1}\,dt \\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right)\sum_{k\geq 1}\binom{\alpha+k-2}{k-1} k^{\underline{m}}\, 2^{-(\alpha+k-1)} \\
&= -\frac{\alpha}{\alpha-1}\int_0^{1/2}(1-t)^{\alpha-1} t^{m-1}\frac{d^m}{dt^m}\left[t(1-t)^{-(\alpha+1)}\right] dt \\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right) 2^{-(\alpha+m-1)}\frac{d^m}{dt^m}\left[t(1-t)^{-\alpha}\right]_{t=1/2} \\
&= -\frac{\alpha}{\alpha-1}\int_0^{1/2}\sum_{i=0}^{1}\binom{m}{i}(\alpha+1)^{\overline{m-i}}\, t^{m-i}(1-t)^{-(m+2-i)}\,dt \\
&\quad + \left(1+\frac{\alpha}{\alpha-1}\right)\left(\alpha^{\overline{m}}+m\alpha^{\overline{m-1}}\right).
\end{aligned}$$

Noting that the integrals within the sum are all of a common, solvable form:

$$\int_0^{1/2} t^m (1-t)^{-(m+2)}\, dt = \frac{1}{m+1}\left[\left(\frac{t}{1-t}\right)^{m+1}\right]_{t=0}^{1/2} = \frac{1}{m+1},$$

the mth factorial moment reduces to:

$$\begin{aligned}\sum_{k\geq 1} k^{\underline{m}} P(\mathcal{L}=k) &= -\frac{\alpha}{\alpha-1}\left(\frac{(\alpha+1)^{\overline{m}}}{m+1} + \frac{m(\alpha+1)^{\overline{m-1}}}{m}\right) \\ &\quad + \left(1+\frac{\alpha}{\alpha-1}\right)\left(\alpha^{\overline{m}} + m\alpha^{\overline{m-1}}\right) \\ &= \left(1+\frac{\alpha}{\alpha-1}\right) m\alpha^{\overline{m-1}} + 2\alpha^{\overline{m}} - \frac{1}{\alpha-1}\frac{\alpha^{\overline{m+1}}}{m+1} \\ &= \frac{4m^2+2\alpha m+\alpha-2}{m+1}\alpha^{\overline{m-1}}.\end{aligned}$$

□

Once again, the fact that the expected label of the centroid in a random recursive tree tends to $5/2$ was first proved by Moon (2002). And lastly, but not unexpectedly, it follows from Theorem 14 that binary increasing trees ($\alpha = 2$) lead to the greatest eventual mean and variance.

5 The size of the centroid's root branch

Our third and final set of results involving the centroid of an increasing tree involve its ancestral branch. In a way this is the most interesting of the centroid's branches, because its descendent branches behave mostly (in particular, their number and sizes do) like those of the root branches of a random increasing tree—albeit under the extra condition that no one branch contains more than $\lfloor n/2 \rfloor$ nodes.

It is also worth noting that these results are interesting for another reason: they can be contrasted with the case of simply generated trees. We have already mentioned that almost all simply generated trees of size n have three large centroid branches that together contain most of the tree's nodes, and in fact Meir and Moon (2002) have proved (among other things) that the size of the centroid's ancestral branch, divided by n, tends to $\sqrt{2}-1 \approx 0.414$ as $n \to \infty$ (independently of the specific family of simply generated trees). Our main goal in this section is an analogue of this result for increasing trees, however we will phrase it (analogously) in terms of the size of the subtree rooted at the centroid. This sort of comparison was not possible for the results of the previous two sections, simply due to the fact that the depth and label (where applicable) of the centroid in a simply generated tree are relatively uninformative, because roots or specifically labelled nodes in simply generated trees are, for the most part, no different from randomly selected nodes.

Let $S(T)$ denote the size of the subtree containing the centroid and all of its descendent branches, and $P_n(S = m = \lfloor \theta n \rfloor)$ the relevant probability distribution. Since the ancestral branch contains at most $\lfloor n/2 \rfloor$ nodes, the ranges of m and θ are $\{\lceil n/2 \rceil, \ldots, n\}$ and $[1/2, 1]$ respectively, with $m = n$ characterising the case in which the root and centroid coincide (for which Theorem 11 already provides a limiting probability).

5.1 A preliminary equation

The event that the centroid's subtree is made up of m nodes (where $n/2 \leq m < n$) can be decomposed into a pair of simpler events: firstly, that the tree contains a subtree of size m (there can be at most one); and secondly, given the presence of such a subtree, that its root is the centroid. This second event can be stated more explicitly as the case, in a tree of size m, that the root is the only node with at least $\lfloor n/2 \rfloor$ descendants.

It is here that we will make use of $P_n(\Lambda_k(\sigma))$, which was introduced in Section 3.1 as a generalisation of the 'path' probability $P_n(\Lambda_k(1/2))$. Let $X_m(T)$ mark the existence of a subtree of size m in a tree T, and let F_j be the label of node j's parent, so that $F_j = 1$ characterises the root's children. The probability that the centroid's subtree contains exactly m nodes can be expressed as:

$$\begin{aligned} P_n(S = m) &= P_n(X_m)\Big(1 - P_m\Big(\bigcup\nolimits_{j\geq 2} \Lambda_j\big(\tfrac{n}{2m}\big)\Big)\Big) \\ &= P_n(X_m)\left(1 - \sum_{j=2}^{m} P_m\big(F_j = 1 \cap \Lambda_j\big(\tfrac{n}{2m}\big)\big)\right) \\ &= P_n(X_m) \cdot \big(1 - A_m\big(\tfrac{n}{2m}\big)\big), \end{aligned} \tag{21}$$

where $A_m\big(\frac{n}{2m}\big)$ is used as an abbreviation for $\sum_{j=2}^{m} P_m\big(F_j = 1 \cap \Lambda_j\big(\frac{n}{2m}\big)\big)$. The probabilities that appear within the sum refer to disjoint events, since at most one of the subtree's root branches can contain $\lfloor n/2 \rfloor + 1$ nodes. We note, as we did for equation (18), that the size of the subtree rooted at j is independent of the node j was attached to, so that:

$$P_m\big(F_j = 1 \cap \Lambda_j\big(\tfrac{n}{2m}\big)\big) = P(F_j = 1) P_m\big(\Lambda_j\big(\tfrac{n}{2m}\big)\big).$$

Consider the first probability $P_n(X_m)$ in equation (21), of the event that a tree of n nodes contains a subtree of size m. Since there can be at most one, this probability can be rephrased as the expected number of such subtrees. This problem in turn is known as the *subtree size profile* of a tree, and has recently been studied for various families of increasing trees. In fact, the expected proportion of nodes with $m-1$ descendants (each forming a rooted subtree of size m) has already been given explicitly for the most interesting families, see (Fuchs, 2012, Section 3 and Theorem 4.1). Letting $U_m = U_m(T)$ denote the number of subtrees of size m in a random tree, we perform the derivation in a more general way here.

Lemma 15. *For $1 \leq m < n$, the expected number of subtrees of size m in a random very simple increasing tree of size n is given by:*

$$E_n(U_m) = \frac{\alpha(\alpha + n - 1)}{(\alpha + m)(\alpha + m - 1)}.$$

Proof: Firstly, note that $E_n(U_m) = \sum_l P_n(S_l = m)$, where S_l is the size of the subtree rooted at l. Now $P_n(S_l = m)$ is the probabilty that node l has $m - 1$ descendants, which was derived in Section 3.1 (see in particular (5)) to be

$$P_n(S_l = m) = \binom{\alpha + m - 2}{m - 1}\binom{n - m - 1}{l - 2} \Big/ \binom{\alpha + n - 2}{n - l},$$

which can be rewitten as

$$P_n(S_l = m) = \alpha B(n-m, \alpha+m-1)\binom{\alpha+l-2}{l-2}\binom{n-l}{m-1}.$$

Summing over possible labels (the root is omitted since $m < n$) yields:

$$\begin{aligned} E_n(U_m) &= \sum_{l=2}^{n-m+1} P_n(S_l = m) \\ &= \alpha B(n-m, \alpha+m-1) \sum_{l=2}^{n-m+1} \binom{\alpha+l-2}{l-2}\binom{n-l}{m-1} \\ &= \alpha B(n-m, \alpha+m-1)\binom{\alpha+n-1}{n-m-1}. \end{aligned}$$

in which the final step is due to the Chu-Vandermonde identity, once the numerators of the binomial coefficients have been converted to constants:

$$\sum_{l=0}^{n-m-1} \binom{\alpha+l}{l}\binom{n-l-2}{m-1} = (-1)^{n-m-1} \sum_{l=0}^{n-m-1} \binom{-\alpha-1}{l}\binom{-m}{n-m-1-l}.$$

The stated result is obtained after simplifying. □

Corollary 16. *For $n/2 \leq m < n$, the probability that a tree of size n contains a subtree of size m is:*

$$P_n(X_m) = \frac{\alpha(\alpha+n-1)}{(\alpha+m)(\alpha+m-1)}.$$

5.2 The probability that the root of a subtree is the centroid

The second probability in equation (21), denoted by $1 - A_m(n/(2m))$, accounts for the cases in which the root of a subtree of size m is the centroid of the entire tree, where $m/n = \theta$ for some fixed θ. By the aforementioned independence argument, we have:

$$A_m\left(\tfrac{n}{2m}\right) = \sum_{j=2}^{m} P(F_j = 1) P_m\left(\Lambda_j\left(\tfrac{n}{2m}\right)\right), \tag{22}$$

in which both of the terms contained within the sum are manageable—the first by Lemma 10:

$$P(F_j = 1) = \frac{\alpha(j-2)!}{\alpha^{\overline{j-1}}} = \alpha B(j-1, \alpha),$$

and the second due to Theorem 2 and Lemma 5, which provide an asymptotic form and an upper bound respectively.

Lemma 17. *For $n/2 \leq m < n$ in a tree of size n, the probability that the root of a subtree of size m is* not *the centroid of the tree satisfies, for all $0 < \varepsilon < 1/2$:*

$$A_m\left(\tfrac{n}{2m}\right) = \frac{\alpha}{\alpha-1}\left(1 - \left(\tfrac{n}{2m}\right)^{\alpha-1}\right) + O\left(n^{-\varepsilon}\right).$$

Proof: The sum given in equation (22) can be split at a value small enough for Theorem 2 to be applied, say $J = \lfloor m^{1/4-\varepsilon/2} \rfloor$ so that $J^2/\sqrt{m} = O(m^{-\varepsilon})$:

$$A_m\big(\tfrac{n}{2m}\big) = \sum_{j=2}^{J} P(F_j = 1) I_{1-\frac{n}{2m}}(j-1, \alpha)\left(1 + O\left(\frac{j^2}{\sqrt{m}}\right)\right) + \sum_{j=J+1}^{m} P(F_j = 1) P_m\big(\Lambda_j\big(\tfrac{n}{2m}\big)\big).$$

Applying the bound of Lemma 5 affirms that the second sum is small for large values of J:

$$\begin{aligned}\sum_{j=J+1}^{m} P(F_j = 1) P_m\big(\Lambda_j\big(\tfrac{n}{2m}\big)\big) &\le 6\alpha \frac{m}{n} \sum_{j=J+1}^{m} B(j-1, \alpha) \frac{\alpha^{\overline{j-1}}}{(j-1)!} \Big(1 - \frac{n}{2m}\Big)^{j-1} \\ &\le 6\alpha \frac{m}{n} \sum_{j=J+1}^{m} \Big(1 - \frac{n}{2m}\Big)^{j-1} \\ &< 12\alpha \Big(\frac{m}{n}\Big)^2 \Big(1 - \frac{n}{2m}\Big)^{J} \xrightarrow[m\to\infty]{} 0,\end{aligned}$$

Similarly, extending the first sum to an infinite one has an effect that vanishes as m and n grow:

$$\begin{aligned}\sum_{j>J} P(F_j = 1) I_{1-\frac{n}{2m}}(j-1, \alpha) &= \alpha \sum_{j>J} \int_0^{1-\frac{n}{2m}} t^{j-2}(1-t)^{\alpha-1}\, dt \\ &\le \alpha \sum_{j>J} \Big(1 - \frac{n}{2m}\Big)^{j-1} \xrightarrow[m\to\infty]{} 0.\end{aligned}$$

Combined, these two substitutions give the asymptotic behaviour of $A_m(\frac{n}{2m})$:

$$\begin{aligned}A_m\big(\tfrac{n}{2m}\big) &= \alpha \sum_{j=2}^{J} \int_0^{1-\frac{n}{2m}} t^{j-2}(1-t)^{\alpha-1}\, dt \left(1 + O\left(\frac{J^2}{\sqrt{m}}\right)\right) + O\left(\Big(1 - \frac{n}{2m}\Big)^{J}\right) \\ &= \alpha \int_0^{1-\frac{n}{2m}} (1-t)^{\alpha-2}\, dt \big(1 + O\big(m^{-\varepsilon}\big)\big) \\ &= \frac{\alpha}{\alpha - 1}\left(1 - \Big(\frac{n}{2m}\Big)^{\alpha-1}\right) + O\big(n^{-\varepsilon}\big),\end{aligned}$$

which can be compared to the case $m = n$ as given in Theorem 11. □

When dealing with recursive trees, the final step is slightly different, resulting in:

Corollary 18. *For recursive trees:*

$$A_m\big(\tfrac{n}{2m}\big) \sim \ln\big(\tfrac{2m}{n}\big).$$

5.3 The distribution of the size of the centroid's subtree

Now that $P_n(X_m)$ is known explicitly (Corollary (16)), and $A_m(\frac{n}{2m})$ asymptotically, we are ready to derive an expression for the distribution of $S(T)$. In the light of equation (21), which states that:

$$P_n(S=m) = P_n(X_m) \cdot \big(1 - A_m\big(\tfrac{n}{2m}\big)\big),$$

we have the main theorem of this section:

Lemma 19. *For $n/2 \leq m < n$ and any $0 < \varepsilon < 1/2$, the probability that the centroid has $m-1$ descendent nodes is given by:*

$$P_n(S=m) = \frac{4}{n}\frac{\alpha}{\alpha-1}\left(\alpha\Big(\frac{n}{2m}\Big)^{\alpha+1} - \Big(\frac{n}{2m}\Big)^2\right) + O\big(n^{-1-\varepsilon}\big).$$

Proof: The result is simply an application of Corollary (16) and Lemma 17 to the above expression:

$$\begin{aligned} P_n(S=m) &= \left(1 - \frac{\alpha}{\alpha-1}\left(1 - \Big(\frac{n}{2m}\Big)^{\alpha-1}\right) + O\big(n^{-\varepsilon}\big)\right)\frac{\alpha(\alpha+n-1)}{(\alpha+m)(\alpha+m-1)} \\ &= \left(1 - \frac{\alpha}{\alpha-1}\left(1 - \Big(\frac{n}{2m}\Big)^{\alpha-1}\right)\right)\frac{\alpha}{n}\Big(\frac{n}{m}\Big)^2 + O\big(n^{-1-\varepsilon}\big) \\ &= \frac{4}{n}\frac{\alpha}{\alpha-1}\left(\alpha\Big(\frac{n}{2m}\Big)^{\alpha+1} - \Big(\frac{n}{2m}\Big)^2\right) + O\big(n^{-1-\varepsilon}\big). \end{aligned}$$

□

Combined with the special case $m = n$ (Theorem 11), this asymptotically describes the size of the centroid's subtree, and thus the size $n - m$ of its root branch as well.

As with the distributions of the centroid's depth and label, we can also show convergence to a limiting distribution; however in this case, although the finite probability distributions are discrete, the limiting distribution is a mixture of a continuous distribution with support $[1/2, 1)$ and a point measure at 1. The notion of convergence is also slightly different, in that it is weaker than those of the previous two sections.

Theorem 20. *The proportion $S(\mathcal{T}_n)/n$ of nodes accounted for by the subtree consisting of the centroid and all of its descendants in a random tree of size n converges in distribution to the random variable $\mathcal{S}$, defined on $[1/2, 1)$ by the density:*

$$f(\theta) = 4\frac{\alpha}{\alpha-1}\Big(\alpha(2\theta)^{-(\alpha+1)} - (2\theta)^{-2}\Big),$$

and at the boundary $\theta = 1$ by the point measure:

$$P(\mathcal{S}=1) = 1 - \frac{\alpha}{\alpha-1}\Big(1 - 2^{-(\alpha-1)}\Big).$$

Proof: Consider the cumulative distribution function arising from Lemma 19:

$$\begin{aligned} P_n(S \leq \sigma n) &= \frac{4}{n}\frac{\alpha}{\alpha-1}\sum_{m=\lceil n/2\rceil}^{\lfloor \sigma n\rfloor}\left(\alpha\Big(\frac{n}{2m}\Big)^{\alpha+1} - \Big(\frac{n}{2m}\Big)^2\right) + O\big(n^{-\varepsilon}\big) \\ &= 4\frac{\alpha}{\alpha-1}\int_{1/2}^{\sigma}\Big(\alpha 2\theta)^{-(\alpha+1)} - (2\theta)^{-2}\Big)\,d\theta + O\big(n^{-\varepsilon}\big). \end{aligned}$$

Note that the error term—which traces back to Theorem 2 via Lemma 17—is uniform in σ over subsets of the form $[1/2, 1-\delta)$. And since each point of continuity (there is discontinuity at 1) is contained in such a subset, this makes explicit (for $\sigma < 1$) the convergence of $P_n(S \leq \sigma n)$ to the continuous distribution with the stated density. The point measure simply corresponds to $P(\mathcal{L} = 1)$. □

Once again, the result for recursive trees differs slightly:

Corollary 21. *For recursive trees:*

$$f(\theta) = \frac{1 - \ln(2\theta)}{\theta^2} \text{ on } [\tfrac{1}{2}, 1) \quad \text{and} \quad P(\mathcal{S} = 1) = 1 - \ln 2.$$

5.4 Moments of the subtree's size distribution

Finally, we can detail the limiting behaviour of the moments of $S(T)$, and in particular, the expected size of the centroid's subtree. This is simply mechanical, since our random variables have bounded support, so that convergence in distribution implies convergence of moments.

Theorem 22. *The moments of the distribution of the proportion $S(\mathcal{T}_n)/n$ of the tree accounted for by the centroid and its descendants converge to those of $\mathcal{S}$, i.e.:*

$$\lim_{n\to\infty} E_n((S/n)^r) = E(\mathcal{S}^r).$$

The limit of the rth moment satisfies:

$$E(\mathcal{S}^r) = P(\mathcal{S} = 1) + \frac{\alpha}{\alpha - 1}\left(\frac{\alpha}{\alpha - r}\left(2^{-(r-1)} - 2^{-(\alpha-1)}\right) - \frac{1}{r-1}\left(1 - 2^{-(r-1)}\right)\right).$$

In particular, the limit of its mean is:

$$E(\mathcal{S}) = 1 + \frac{\alpha}{(\alpha-1)^2}\left(1 - 2^{-(\alpha-1)} - (\alpha - 1)\ln 2\right).$$

Proof: For $\alpha \notin \{1, r\}$ and $r \in \mathbb{Z}_{>0}$, and with $f(\theta)$ as in Theorem 20, we have:

$$\begin{aligned}
E(\mathcal{S}^r) &= P(\mathcal{S} = 1) + \int_{1/2}^{1} \theta^r f(\theta)\, d\theta \\
&= P(\mathcal{S} = 1) + 2^{-(r-1)} \frac{\alpha}{\alpha - 1} \int_1^2 \mu^r \left(\alpha\mu^{-(\alpha+1)} - \mu^{-2}\right) d\mu \\
&= P(\mathcal{S} = 1) + 2^{-(r-1)} \frac{\alpha}{\alpha - 1} \left[-\frac{\alpha}{\alpha - r}\mu^{r-\alpha} - \frac{1}{r-1}\mu^{r-1}\right]_1^2 \\
&= P(\mathcal{S} = 1) + \frac{\alpha}{\alpha - 1}\left(\frac{\alpha}{\alpha - r}\left(2^{-(r-1)} - 2^{-(\alpha-1)}\right) - \frac{1}{r-1}\left(1 - 2^{-(r-1)}\right)\right).
\end{aligned}$$

□

For plane-oriented and binary increasing trees ($\alpha = 1/2$ and $\alpha = 2$), this yields means of $3 - 2\sqrt{2} + \ln 2 \approx 0.86$ and $2 - 2\ln 2 \approx 0.61$ respectively. It is worth noting that the proof's requirement that $\alpha \neq r$

is weak for two reasons: because the standard definition of very simple increasing trees deals with the range $0 < \alpha \leq 2$, and because singular cases such as these can be seen as limits of the above function, or, in the case of recursive trees, be derived directly from Corollaries 12 and 21. For instance:

$$E(\mathcal{S})|_{\alpha=1} = 1 - \frac{1}{2}(\ln 2)^2 \approx 0.76,$$
$$E(\mathcal{S}^r)|_{\alpha=1} = 1 + \frac{r}{(r-1)^2}\left(1 - 2^{-(r-1)}\right) - \frac{r}{r-1}\ln 2,$$
$$E(\mathcal{S}^2)|_{\alpha=2} = 2\ln 2 - 1.$$

The limit of the mean in the case of recursive trees was given by Moon (2002). The asymptotic variances for plane-oriented, recursive, and binary increasing trees are approximately 0.03, 0.04, and 0.01 respectively.

6 Concluding remarks

The behaviour of the (nearest) centroid in a large, random very simple increasing tree is now reasonably clear, and in fact quite consistent across the entire subclass: one expects the centroid to lie, on average, within two edges of the root (for the usual case $0 < \alpha \leq 2$), and its root branch to account for a significant portion of the entire tree.

That being said, there are still a number of related questions that could be raised and investigated: for example, we might consider other parameters of the centroid—in the simplest case, its degree—or ask to what extent the above behaviour generalises to the entire class of increasing trees, which do not necessarily satisfy the properties of Lemma 1.

Perhaps more interestingly, one could attempt to characterise the distribution of the average distance from a node to the other nodes in a tree—the closeness centrality—in a way similar to that which has been done for betweenness centrality (Durant and Wagner, 2017); or, as an alternative definition of a tree's most 'central' node, consider its central points (i.e., nodes with minimal eccentricity). Finally, there are classes of random trees other than simply generated and increasing trees that we have not mentioned at all here—the most notable being the various families of search trees (Drmota, 2009, Section 1.4). One would expect (or at least hope) that several of these problems will also be amenable to the tools and methods of analytic combinatorics.

References

D. Aldous. The continuum random tree. II. An overview. In M. T. Barlow and N. H. Bingham, editors, *Stochastic Analysis*, pages 23–70. Cambridge University Press, 1991.

D. Aldous. Recursive self-similarity for random trees, random triangulations and Brownian excursion. *Ann. Probab.*, 22(2):527–545, 1994.

F. Bergeron, P. Flajolet, and B. Salvy. Varieties of increasing trees. In J.-C. Raoult, editor, *Lecture Notes in Computer Science*, volume 581, pages 24–48. Springer, 1992.

R. P. Dobrow and R. T. Smythe. Poisson approximations for functionals of random trees. *Random Struct. Algorithms*, 9(1–2):79–92, 1996.

M. Drmota. *Random Trees: An Interplay Between Combinatorics and Probability*. Springer, Vienna, 1st edition, 2009. ISBN 9783211753552.

K. Durant. *Centrality in Random Trees*. PhD thesis, Stellenbosch University, December 2017. `http://scholar.sun.ac.za/handle/10019.1/102861`.

K. Durant and S. Wagner. On the distribution of betweenness centrality in random trees. *Theor. Comput. Sci.*, 699:33–52, 2017.

P. Flajolet and R. Sedgewick. *Analytic Combinatorics*. Cambridge University Press, New York, 1st edition, 2009. ISBN 9780521898065.

M. Fuchs. Limit theorems for subtree size profiles of increasing trees. *Comb. Probab. Comput.*, 21(3): 412–441, 2012.

M. Girvan and M. E. J. Newman. Community structure in social and biological networks. *Proc. Natl. Acad. Sci. USA*, 99(12):7821–7826, 2002.

K.-I. Goh, E. Oh, H. Jeong, B. Kahng, and D. Kim. Classification of scale-free networks. *Proc. Natl. Acad. Sci. USA*, 99(20):12583–12588, 2002.

C. Jordan. Sur les assemblages des lignes. *J. Reine Angew. Math.*, 70:185–190, 1869.

M. Kuba and S. Wagner. On the distribution of depths in increasing trees. *Electron. J. Comb.*, 17(R137), 2010.

A. Meir and J. W. Moon. On centroid branches of trees from certain families. *Discrete Math.*, 250(1–3): 153–170, 2002.

J. W. Moon. On the expected distance from the centroid of a tree. *Ars Comb.*, 20:263–276, 1985.

J. W. Moon. On the centroid of recursive trees. *Australas. J. Comb.*, 25:211–219, 2002.

A. Panholzer and H. Prodinger. Level of nodes in increasing trees revisited. *Random Struct. Algorithms*, 31(2):203–226, 2007.

D. Shah and T. Zaman. Rumors in a network: who's the culprit? *IEEE Trans. Inf. Theory*, 57(8):5163–5181, 2011.

B. Zelinka. Medians and peripherians of trees. *Arch. Math.*, 4(2):87–95, 1968.