The Annals of Applied Probability
2025, Vol. 35, No. 1, 158–195
https://doi.org/10.1214/24-AAP2112

THE CRITICAL BETA-SPLITTING RANDOM TREE I: HEIGHTS AND RELATED RESULTS

BY DAVID ALDOUS[1,a] AND BORIS PITTEL[2,b]

[1]*Department of Statistics, University of California, Berkeley,* [a]*aldous@stat.berkeley.edu*
[2]*Department of Mathematics, The Ohio State University,* [b]*pittel.1@osu.edu*

In the critical beta-splitting model of a random n-leaf binary tree, leaf-sets are recursively split into subsets, and a set of m leaves is split into subsets containing i and $m-i$ leaves with probabilities proportional to $1/i(m-i)$. We study the continuous-time model in which the holding time before that split is exponential with rate h_{m-1}, the harmonic number. We (sharply) evaluate the first two moments of the time-height D_n and of the edge-height L_n of a uniform random leaf (i.e., the length of the path from the root to the leaf), and prove the corresponding CLTs. We study the correlation between the heights of two random leaves of the same tree realization, and analyze the expected number of splits necessary for a set of t leaves to partially or completely break away from each other. We give tail bounds for the time-height and the edge-height of the *tree*, that is, the maximal leaf heights. We show that there is a limit distribution for the size of a uniform random subtree, and derive the asymptotics of the mean size. Our proofs are based on asymptotic analysis of the attendant (sum-type) recurrences. The essential idea is to replace such a recursive equality by a pair of recursive inequalities for which matching asymptotic solutions can be found, allowing one to bound, both ways, the elusive explicit solution of the recursive equality. This reliance on recursive inequalities necessitates usage of Laplace transforms rather than Fourier characteristic functions.

1. Introduction. The topic of this paper is a certain random tree model, described below, introduced and discussed briefly in 1996 in [3] but not subsequently studied. There is a slight "applied" motivation as a toy model of phylogenetic trees (see Section 1.3), but our purpose here is to commence a theoretical study of the model. A key point is that it has qualitatively different properties from those of the well-studied random tree models that can be found in (for instance) [7, 12].

One fundamental question about a random tree concerns the height of leaves. It turns out that this question, and many extensions, can be answered extremely sharply via an analysis of recursions, and this paper gives a thorough and detailed account of the range of questions that can be answered by this methodology.

In addressing the Applied Probability community, let us observe that there are also other questions about the model that can be studied via a wide range of general modern probability techniques. Some such work in progress is briefly described in a final Section 3, and a current overview can be found in the preprint [4].

1.1. *The model.* For $m \geq 2$, consider the distribution $(q(m,i), 1 \leq i \leq m-1)$ constructed to be proportional to $\frac{1}{i(m-i)}$. Explicitly (by writing $\frac{1}{i(m-i)} = (\frac{1}{i} + \frac{1}{m-i})/m$)

$$q(m,i) = \frac{m}{2h_{m-1}} \cdot \frac{1}{i(m-i)}, \quad 1 \leq i \leq m-1, \tag{1}$$

Received February 2023; revised April 2024.
MSC2020 subject classifications. Primary 60C05; secondary 05C05, 92B10.
Key words and phrases. Markov chain, phylogenetic tree, random tree, recurrence.

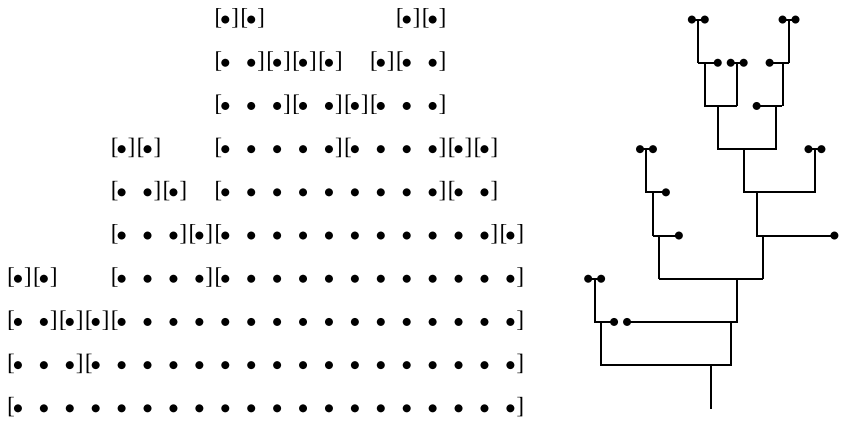

FIG. 1. *The discrete time construction for* $n = 20$. *In the tree, by* edges *we mean the* $n - 1$ *vertical edges. The leaves have edge-heights from* 2 *to* 9.

where h_{m-1} is the harmonic sum $\sum_{i=1}^{m-1} 1/i$. Now fix $n \geq 2$. Consider the process of constructing a random tree by recursively splitting the integer interval $[n] = \{1, 2, \ldots, n\}$ 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 L_n leaves $\{1, \ldots, L_n\}$ and a right subtree which will have the $R_n = n - L_n$ leaves $\{L_n + 1, \ldots, n\}$, where L_n (and also 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 process has a natural tree structure, illustrated schematically[1] in Figure 1. In this discrete-time construction we regard the edges of the tree as having length 1. It turns out (see Section 1.3) to be convenient to consider the continuous-time construction in which a size-m interval is split at rate h_{m-1}, that is, after an exponential(h_{m-1}) *holding time*. Once constructed, it is natural to identify "time" with "distance": a leaf that appears at time t has *time-height* t. Of course the discrete-time model is implicit within the continuous-time model, and a leaf which appears after ℓ splits has *edge-height* ℓ.

We call the continuous-time model the *critical beta-splitting random tree*, but must emphasize that the word *critical* does not have its usual meaning within branching processes. Instead, among the one-parameter family of splitting probabilities with

$$q(m, i) \propto i^{\beta}(m - i)^{\beta}, \quad -2 < \beta < \infty \tag{2}$$

our parameter value $\beta = -1$ is *critical* in the sense that leaf-heights change from order $n^{-\beta-1}$ to order $\log n$ at that value, as noted many years ago when this family was introduced [3].

Finally, our results do not use the leaf-labels $\{1, 2, \ldots, n\}$ in the interval-splitting construction. Instead they involve a uniform *random* leaf. Equivalently, one could take a uniform random permutation of labels and then talk about the leaf with some arbitrary label.

1.2. *Outline of results*. Our main focus is on two related random variables associated with the continuous-time random tree on n leaves:

- D_n = time-height of a uniform random leaf;
- L_n = edge-height of a uniform random leaf.

We start with sharp asymptotic formulas for the moments of D_n and L_n. They are of considerable interest in their own right, and also because the techniques are then extended for

[1] Actual simulations appear in [4].

analysis of the limiting distributions, with the moments estimates enabling us to guess what those distributions should be.

Write $\zeta(\cdot)$ for the Riemann zeta-function, $\zeta(r) := \sum_{j=1}^{\infty} \frac{1}{j^r}$, $(r > 1)$. Note that $\zeta(2) = \pi^2/6$ and that $\zeta^{-1}(2)$ below means $1/\zeta(2)$, not the inverse function. Write γ for the Euler–Masceroni constant, which will appear frequently in our analysis: $\sum_{j=1}^{n} \frac{1}{j} = \log n + \gamma + O(n^{-1})$. Asymptotics are as $n \to \infty$.

THEOREM 1.1.

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

$$\operatorname{var}(D_n) = \big(1 + o(1)\big)\frac{2\zeta(3)}{\zeta^3(2)}\log n,$$

and, contingent on a numerically supported "h-ansatz" (see Section 2.2),

$$\mathbb{E}[D_n] = \zeta^{-1}(2)\log n + c_0 - \frac{1}{2\zeta(2)}n^{-1} + O\big(n^{-2}\big)$$

for a constant c_0 estimated numerically, and

$$\operatorname{var}(D_n) = \frac{2\zeta(3)}{\zeta^3(2)}\log n + O(1).$$

THEOREM 1.2.

$$\mathbb{E}[L_n] = \frac{1}{2\zeta(2)}\log^2 n + \frac{\gamma\zeta(2) + \zeta(3)}{\zeta^2(2)}\log n + O(1),$$

$$\operatorname{var}(L_n) = \frac{2\zeta(3)}{3\zeta^3(2)}\log^3 n + O\big(\log^2 n\big).$$

The various parts of Theorem 1.1 are proved in Sections 2.1–2.4 and 2.7, and Theorem 1.2 is proved in Section 2.8. These theorems immediately yield the WLLNs (weak laws of large numbers) for D_n and L_n, with rates, as follows.

COROLLARY 1.3. *In probability*

$$\mathbb{P}\left(\left|\frac{D_n}{\mathbb{E}[D_n]} - 1\right| \geq \varepsilon\right), \mathbb{P}\left(\left|\frac{L_n}{\mathbb{E}[L_n]} - 1\right| \geq \varepsilon\right) = O\big(\varepsilon^{-2}\log^{-1} n\big).$$

Consider next the time-height $\mathcal{D}_n$ and the edge-height $\mathcal{L}_n$ of the random tree itself, that is, the largest time length and the largest edge length of a path from the root to a leaf. By upper-bounding the Laplace transforms of D_n and L_n, we prove in Sections 2.6 and 2.9

THEOREM 1.4. *There exists $\rho > 0$ such that for all $\varepsilon \in (0, 1)$ we have*

$$\mathbb{P}\big(\mathcal{D}_n \geq (2 + \varepsilon)\log n\big) \leq \frac{1}{n^{\rho\varepsilon}},$$

THEOREM 1.5. *Let $\beta = \min_{\alpha > 1/\log 2}[\alpha + \frac{4\alpha^2\zeta(3)}{\alpha\log 2 - 1}] \approx 42.9$. For $\varepsilon \in (0, 1)$,*

$$\mathbb{P}\big(\mathcal{L}_n \geq (1 + \varepsilon)\beta\log^2 n\big) \leq \exp\big(-\Theta(\varepsilon\log n)\big).$$

We conjecture that both $\frac{\mathcal{D}_n}{\log n}$ and $\frac{\mathcal{L}_n}{\log^2 n}$ converge, in probability, to constants.

The definitions of D_n and L_n involve two levels of randomness, the random tree and the random leaf within the tree. To study the interaction between levels, it is natural to consider the correlation between the heights of two leaves within the same realization of the random tree. Write $D_n^{(1)}$ and $D_n^{(2)}$ for the time-heights of two distinct leaves chosen uniformly from all pairs of leaves. We study the correlation coefficient defined by

$$r_n = \frac{\mathbb{E}[D_n^{(1)} D_n^{(2)}] - \mathbb{E}^2[D_n]}{\mathrm{Var}(D_n)},$$

and prove in Section 2.5

THEOREM 1.6. *Contingent on the h-ansatz, $r_n = O(\log^{-1} n)$, that is, asymptotically $D_n^{(1)}$ and $D_n^{(2)}$ are uncorrelated.*

We conjecture that a similar result holds for the correlation coefficient of $L_n^{(1)}$ and $L_n^{(2)}$, the edge-heights of two distinct, uniformly random leaves, *independently* of the h-ansatz.

Returning to properties of D_n and L_n, in Sections 2.7 and 2.10 we will prove the CLTs corresponding to the means and variances in Theorems 1.1 and 1.2.

THEOREM 1.7. *In distribution, and with all their moments,*

$$\frac{D_n - \zeta^{-1}(2)\log n}{\sqrt{\frac{2\zeta(3)}{\zeta^3(2)}\log n}}, \frac{L_n - (2\zeta(2))^{-1}\log^2 n}{\sqrt{\frac{2\zeta(3)}{3\zeta^3(2)}\log^3 n}} \Longrightarrow \mathrm{Normal}(0,1).$$

The sharp asymptotic estimates of the moments of D_n and L_n, and the ample numeric evidence in the case of D_n, provided a compelling evidence that both D_n and L_n must be asymptotically normal. However, the proof of Theorem 1.7 does not use these estimates, providing instead an alternative verification of the *leading* terms in those estimates, without relying on the h-ansatz.

After posting the original preprint version of this article, alternative proofs of these CLTs have appeared in preprints. Via a martingale CLT [4] (continuous model); via the contraction method [13] (discrete model); and via the theory of regenerative composition structures [11] (discrete model). Presumably these methods can also be applied to the alternate model. Of course, in Theorem 1.7 there is presumably *joint convergence* to a bivariate Gaussian limit. It would be interesting to see which method would be best for proving such joint convergence.

Like Theorems 1.4 and 1.5, the proof of Theorem 1.7 is based on showing convergence of the Laplace transform for the (properly centered and scaled) leaf height to that of Normal(0, 1). Why Laplace, but not Fourier? Because, even though there is enough independence to optimistically expect asymptotic normality, our variables are too far from being the sums of essentially independent terms. So, the best we could do is to use recurrences to bound the (real-valued) Laplace transforms recursively both ways, by those of the Normals, whose parameters we choose to satisfy, asymptotically, the respective recursive inequalities. The added feature here is that we get convergence of the moments as well.

Leaving Laplace versus Fourier issue aside, there are many cases when a limited moment information and the recursive nature of the process can be used to establish asymptotic normality, but the standard techniques hardly apply; see [8, 15, 17–20]. The concrete details vary substantially, of course. For instance, in [19] it was shown that the total number of linear extensions of the random, tree-induced, partial order is lognormal, by showing convergence

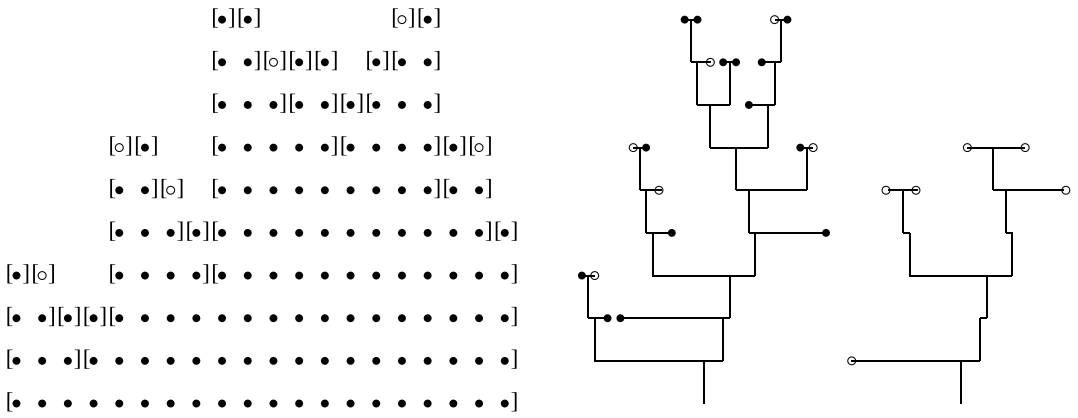

FIG. 2. *The left and center diagrams show $t = 6$ leaves $\circ$ in the $n = 20$-leaf tree in Figure 1. The right diagram is the pruned spanning tree on those leaves, with 8 edges.*

of all semi-invariants, rather than of the Laplace transforms. In [20], for the proof of a two-dimensional CLT for the number of vertices and arcs in the giant strong component of the random digraph, boundedness of the Fourier transform made it indispensable. The unifying feature of these diverse arguments is the recurrence equation for the chosen transform.

The structure theory studied in [4] involves the notion of *pruned spanning tree*, illustrated in Figure 2, and here we study its edge-length. Given a set T of $t := |T| < n$ leaves of the tree on n leaves, there is spanning tree on those leaves and the root; the edges of the spanning tree are the union of the edges on the paths to these leaves. Now we can "prune" this spanning tree by cutting the end segment of each path back to the internal vertex v where it branches from the other paths; the spanning tree on those branchpoints v forms the pruned spanning tree. Equivalently, the edges of the pruned spanning tree are the edges in the paths from the root to vertices v such that *each* of the two subtrees rooted at v's children has at least one leaf from T. Write $S^*_{n,t}$ for the number of such edges, when T is a uniform random choice of $t < n$ leaves. In Section 2.11 we prove the following theorem.

THEOREM 1.8. *With $B(t_1, t_2) = \frac{\Gamma(t_1)\Gamma(t_2)}{\Gamma(t)}$, we have*

$$\mathbb{E}[S^*_{n,t}] = \alpha(t)\log n + O(1), \quad \alpha(t) = \Big(h_{t-1} - \sum_{t_1+t_2=t} B(t_1, t_2)\Big)^{-1},$$

along with a related result (Proposition 2.14) for the edge-height of the first branch-point in the pruned tree. At long last, the Riemann zeta-function has suddenly loosened its grip, and appropriately the beta-function has taken the stage.

Finally in Section 2.12 we prove

THEOREM 1.9. *Let $\mathbf{u}_n := \{u_n(t)\}_{t\le n}$ be the distribution of the number of leaves in a subtree rooted at a uniform random vertex, that is, one of the $2n-1$ leaves or branchpoints. The sequence $\{\mathbf{u}_n\}_{n\ge1}$ converges to a proper distribution $\mathbf{u}$. However $\sum_{t\ge1} t u_n(t) \sim \frac{3}{2\pi^2}\log^2 n$.*

1.3. *Motivation and background.* The one-parameter family at (2) was introduced in [3] in 1996 as a toy model for phylogenetic trees. It was observed in [5] that, in splits $m \to (i, m-i)$ in real-world phylogenetic trees, the median size of the smaller subtree scaled roughly as $m^{1/2}$. This data is not consistent with more classical random tree models, where

the median size would be $O(\log m)$ or $\Theta(m)$. However, in the "critical" case $\beta = -1$ of the model studied in this paper, that median size is indeed order $m^{1/2}$, because

$$2 \sum_{i<m^{1/2}} q(m,i) \to \frac{1}{2} \quad \text{as } m \to \infty.$$

Of course the model is not a biologically meaningful "forwards in time" model, but is a mathematically basic model that could be used for comparison with more realistic models [14] that are deliberately constructed in the mathematical biology literature to reproduce features of real phylogenetic trees. The distributional results of this model might therefore be useful for some future "applied" project of that kind.

2. The proofs. Let τ_ν be the holding time before a split of a subset of size ν. So τ_ν has exponential distribution with rate $h_{\nu-1}$. By the definition of the splitting process, for $\nu \geq 2$ we have, with $q(\nu,i) = \frac{\nu}{2h_{\nu-1}} \frac{1}{i(\nu-i)}$ as at (1),

$$D_\nu = \begin{cases} \tau_\nu + D_i, & \text{with probability } q(\nu,i)\dfrac{i}{\nu}, i = 1,\dots,\nu-1, \\ \tau_\nu + D_{\nu-i}, & \text{with probability } q(\nu,i)\dfrac{\nu-i}{\nu}, i = 1,\dots,\nu-1. \end{cases}$$

Introduce $\phi_\nu(u) = \mathbb{E}[e^{uD_\nu}]$, the Laplace transform of the distribution of D_ν; so, $\phi_1(u) = 1$. The equation above implies that for $\nu \geq 2$,

$$\begin{aligned} \phi_\nu(u) &= \sum_{k=1}^{\nu-1} q(\nu,k)\left(\frac{k}{\nu}\,\mathbb{E}\big[\exp\big(u(\tau_\nu + D_k)\big)\big] + \frac{\nu-k}{\nu}\,\mathbb{E}\big[\exp\big(u(\tau_\nu + D_{\nu-k})\big)\big]\right) \\ &= 2\,\mathbb{E}\big[\exp(u\tau_\nu)\big] \sum_{k=1}^{\nu-1} \frac{k}{\nu}\phi_k(u) q_{\nu,k} = \frac{1}{h_{\nu-1}-u} \sum_{k=1}^{\nu-1} \frac{\phi_k(u)}{\nu-k}. \end{aligned} \tag{3}$$

Furthermore, introduce $f_\nu(u) = \mathbb{E}[e^{uL_\nu}]$, the Laplace transform of the distribution of L_ν; so $f_1(u) = 1$. In this case we have, for $\nu \geq 2$,

$$L_\nu = \begin{cases} 1 + L_i, & \text{with probability } q(\nu,i)\dfrac{i}{\nu}, i = 1,\dots,\nu-1, \\ 1 + L_{\nu-i}, & \text{with probability } q(\nu,i)\dfrac{\nu-i}{\nu}, i = 1,\dots,\nu-1. \end{cases}$$

Therefore,

$$\begin{aligned} f_\nu(u) &= \sum_{k=1}^{\nu-1} q(\nu,k)\left(\frac{k}{\nu}\,\mathbb{E}\big[\exp\big(u(1 + L_k)\big)\big] + \frac{\nu-k}{\nu}\,\mathbb{E}\big[\exp\big(u(1 + L_{\nu-k})\big)\big]\right) \\ &= 2e^u \sum_{k=1}^{\nu-1} \frac{k}{\nu} f_k(u) q_{\nu,k} = \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{f_k(u)}{\nu-k}. \end{aligned} \tag{4}$$

In particular, we make extensive use of the following fundamental recurrence for $\mathbb{E}[D_\nu]$:

$$\mathbb{E}[D_\nu] = \frac{1}{h_{\nu-1}}\left(1 + \sum_{k=1}^{\nu-1} \frac{\mathbb{E}[D_k]}{\nu-k}\right). \tag{5}$$

This follows directly from the hold-jump construction of the random tree, or by differentiating both sides of (3) at $u = 0$.

2.1. *The moments of D_n.* Our first result includes one part of Theorem 1.1.

PROPOSITION 2.1.

$$\zeta^{-1}(2)\log n \le \mathbb{E}[D_n] \le \max\{0, 1+\log(n-1)\}, \quad n \ge 2,$$
$$\mathbb{E}[D_n] = \zeta^{-1}(2)\log n + O(1).$$

PROOF. The proof has three steps.

(i) Let us prove that $\mathbb{E}[D_n] \ge \frac{6}{\pi^2}\log n$. Introduce $\theta_n = A\log n$. Then $\mathbb{E}[D_1] = 0 = \theta_1$. If we find A such that

$$\theta_n \le \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1}\frac{\theta_k}{n-k}\right), \quad n \ge 2, \tag{6}$$

then, by induction on n, $\mathbb{E}[D_n] \ge \theta_n$ for all $n \ge 1$. We compute

$$\begin{aligned}\frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1}\frac{\theta_k}{n-k}\right) &= \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1}\frac{A\log k}{n-k}\right)\\ &= \frac{1}{h_{n-1}}\left(1 + A(\log n)h_{n-1} + A\sum_{k=1}^{n-1}\frac{\log(k/n)}{n(1-k/n)}\right)\\ &= \theta_n + \frac{1}{h_{n-1}}\left(1 + A\sum_{k=1}^{n-1}\frac{\log(k/n)}{n(1-k/n)}\right)\\ &\ge \theta_n + \frac{1}{h_{n-1}}\left(1 - A\int_0^1 \frac{\log(1/x)}{1-x}\,dx\right).\end{aligned}$$

The inequality holds since the integrand is positive and decreasing. Since

$$\int_0^1 \frac{\log(1/x)}{1-x}\,dx = \sum_{j\ge 0}\int_0^1 x^j \log(1/x)\,dx = \sum_{j\ge 0}\frac{1}{(j+1)^2} = \zeta(2) = \frac{\pi^2}{6},$$

we deduce that (6) holds if we select $A = \frac{6}{\pi^2} = \zeta^{-1}(2)$.

NOTE. The proof above is the harbinger of things to come, including the next part. The seemingly naive idea is to replace a recurrence equality by a recurrence *inequality* for which an exact solution can be found and then to use it to upper bound the otherwise-unattainable solution of the recurrence *equality*. Needless to say, it is critically important to have a good guess as to how that "hidden" solution behaves asymptotically.

(ii) Let us prove that $\mathbb{E}[D_n] \le f(n) := \max\{0, 1+\log(n-1)\}$ for $n \ge 2$. This is true for $n = 1, 2$ since $\mathbb{E}[D_1] = 0$, $\mathbb{E}[D_2] = 1$. Notice that $1 + \log(x-1) \le x - 1$ for $x \in (1, 2]$. So $f(x) \le g(x)$, $\forall x > 1$, where $g(x) = x - 1$ for $x \in [1, 2]$, $g(x) = 1 + \log(x-1)$ for $x \ge 2$, and $g(x)$ is concave for $x \ge 1$. So, similar to (6), it is enough to show that $g(n)$ satisfies

$$g(n) \ge \frac{1}{h_{n-1}}\left(1 + \sum_{i=1}^{n-1}\frac{g(i)}{n-i}\right), \quad n \ge 2. \tag{7}$$

By concavity of $g(x)$ for $x \ge 1$, we have

$$\begin{aligned}\frac{1}{h_{n-1}}\left(1 + \sum_{i=1}^{n-1}\frac{g(i)}{n-i}\right) &\le \frac{1}{h_{n-1}} + g\left(\sum_{i=1}^{n-1}\frac{i}{n-i}\right)\\ &= \frac{1}{h_{n-1}} + g\left(n - \frac{n-1}{h_{n-1}}\right) \le \frac{1}{h_{n-1}} + g(n) - g'(n)\left(\frac{n-1}{h_{n-1}}\right),\end{aligned}$$

which is exactly $g(n)$, since $g'(n) = \frac{1}{n-1}$ for $n > 1$.

(iii) Write $\mathbb{E}[D_n] = \frac{6}{\pi^2}\log n + u_n$, so that $u_n \geq 0$ and $u_1 = 0$. Let us prove that $u_n = O(1)$. Using (5) we have

$$\begin{aligned} u_n &= \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1} \frac{u_k}{n-k}\right) + \frac{6}{\pi^2}\left(\frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\log k}{n-k} - \log n\right) \\ &= \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1} \frac{u_k}{n-k}\right) + \frac{6}{\pi^2 h_{n-1}}\sum_{k=1}^{n-1}\frac{\log(k/n)}{n-k}. \end{aligned} \tag{8}$$

The proof of (iii) depends on the following rather sharp asymptotic formula for the last sum, which we believe to be new. We defer the proof of the lemma.

LEMMA 2.2.

$$\sum_{k=1}^{n-1}\frac{\log(k/n)}{n-k} = -\zeta(2) + \frac{\log(2\pi e)}{2n} + \frac{\log n}{12n^2} + O\big(n^{-2}\big).$$

Granted this estimate, the recurrence (8) becomes

$$\begin{aligned} u_n = \frac{\zeta^{-1}(2)}{h_{n-1}}&\left(\frac{\log(2\pi e n)}{2n} + \frac{\log n}{12n^2} + O\big(n^{-2}\big)\right) \\ &+ \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{u_k}{n-k}, \quad n \geq 2, u_1 = 0. \end{aligned} \tag{9}$$

It is easy to check that the sequence $x_n := \frac{n-1}{n}$ satisfies the recurrence

$$x_n = \frac{1}{n} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{x_k}{n-k}, \quad n \geq 2, x_1 = 0.$$

As the explicit term on the RHS of (9) is asymptotic to $\frac{\zeta^{-1}(2)}{2n}$, we can deduce that $u_n = O(1)$, establishing (iii). Indeed, by the triangle inequality, the equation (9) implies that

$$|u_n| \leq \frac{c}{n} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{|u_k|}{n-k}.$$

By induction on n, this inequality coupled with the recurrence for x_n imply that $|u_n| \leq 2cx_n \leq 2c$. □

Proof of Lemma 2.2. First, we have, for $n \geq 2$

$$\begin{aligned} \sum_{k=1}^{n-1}\frac{\log(k/n)}{n-k} &= \sum_{k=1}^{n-1}\left(\frac{\log(k/n)}{n} + \frac{k\log(k/n)}{n(n-k)}\right) \\ &= \frac{1}{n}\log\frac{(n-1)!}{n^{n-1}} + \sum_{k=1}^{n-1}\frac{(k/n)\log(k/n)}{n-k}. \end{aligned} \tag{10}$$

By Euler's summation formula (Graham, Knuth, and Patashnik [10], (9.78)), if $f(x)$ is a smooth differentiable function for $x \in [a, b]$ such that the even derivatives are all of the same

sign, then for every $m \geq 1$,

$$\sum_{a\leq k<b} f(k) = \int_a^b f(x)\,dx - \frac{1}{2}f(x)\Big|_a^b + \sum_{\ell=1}^{m} \frac{B_{2\ell}}{(2\ell)!} f^{(2\ell-1)}(x)\Big|_a^b + \theta_m \frac{B_{2m+2}}{(2m+2)!} f^{(2m+1)}(x)\Big|_a^b. \tag{11}$$

Here $\theta_m \in (0,1)$ and $\{B_{2\ell}\}$ are even Bernoulli numbers, defined by $\frac{z}{e^z-1} = \sum_{\mu\geq 0} B_\mu \frac{z^\mu}{\mu!}$. The equation (11) was used in [10] to show that

$$\sum_{1\leq k<n} \log k = n\log n - n + \frac{1}{2}\log\frac{2\pi}{n} + \sum_{\ell=1}^{m} \frac{B_{2\ell}}{2\ell(2\ell-1)n^{2\ell-1}} + \theta_{m,n}\frac{B_{2m+2}}{(2m+2)(2m+1)n^{2m+1}},$$

$\theta_{m,n} \in (0,1)$. Here $f(x) = \log x$, so that $f^{(2\ell)}(x) < 0$ for $x \geq 1$ and $\ell \geq 1$. Using this estimate for $m = 1$, we obtain a sharp version of Stirling's formula:

$$\frac{1}{n}\log\frac{(n-1)!}{n^{n-1}} = -1 + \frac{\log(2\pi n)}{2n} + O\big(n^{-2}\big). \tag{12}$$

Consider the sum in the bottom RHS of (10). This time, take $f(x) = \frac{(x/n)\log(x/n)}{n-x}$, $x \in [1, n-1]$, and $f(n) := -\frac{1}{n}$. Let us show that $f^{(2\ell)}(x) > 0$ for $x \in (0,n)$, or equivalently that $g^{(2\ell)}(y) > 0$ for $y \in (0,1)$, where $g(y) := \frac{y\log y}{1-y}$. We have

$$g(y) = -\log y + \frac{\log y}{1-y} = -\log y - \sum_{j\geq 1}\frac{(1-y)^{j-1}}{j}$$

$$= -\log(1-z) - \sum_{j\geq 1}\frac{z^{j-1}}{j}, \quad z := 1-y.$$

So, we need to show that

$$\big(-\log(1-z)\big)^{(2\ell)} \geq \Big(\sum_{j\geq 1}\frac{z^{j-1}}{j}\Big)^{(2\ell)},$$

or equivalently that

$$\frac{(2\ell-1)!}{(1-z)^{2\ell}} \geq \sum_{j>2\ell}\frac{(j-1)_{2\ell}z^{j-1-2\ell}}{j}.$$

This inequality will follow if we prove a stronger inequality,[2] namely that, for every $\nu \geq 0$,

$$[z^\nu]\frac{(2\ell-1)!}{(1-z)^{2\ell}} \geq [z^\nu]\sum_{j>2\ell}\frac{(j-1)_{2\ell}z^{j-1-2\ell}}{j}.$$

But this is equivalent to

$$(2\ell+\nu-1)! \geq \frac{(2\ell+\nu)!}{2\ell+\nu+1},$$

[2] $[z^\nu]$ denotes the coefficient of z^ν.

which is obviously true. Therefore, applying (11), we have, with $\theta'_{m,n} \in (0,1)$,

$$\begin{aligned}\sum_{k=1}^{n-1} \frac{(k/n)\log(k/n)}{n-k} &= \int_{1/n}^{1} g(y)\,dy - \frac{1}{2n} g(y)\Big|_{1/n}^{1} \\ &\quad + \sum_{\ell=1}^{m} \frac{B_{2\ell}}{n^{2\ell}(2\ell)!} g^{(2\ell-1)}(y)\Big|_{1/n}^{1} \\ &\quad + \theta'_{m,n} \frac{B_{2m+2}}{n^{2m+2}(2m+2)!} g^{(2m+1)}(y)\Big|_{1/n}^{1}.\end{aligned} \tag{13}$$

For the first terms in (13)

$$\begin{aligned}\int_{1/n}^{1} g(y)\,dy &= \int_0^1 \frac{y\log y}{1-y}\,dy - \sum_{j\ge 1}\int_0^{1/n} y^j \log y\,dy \\ &= -\zeta(2) + 1 + (\log n)\sum_{j\ge 2} n^{-j} j^{-1} + \sum_{j\ge 2} n^{-j} j^{-2}; \\ g(y)|_{1/n}^{1} &= -1 + \frac{\log n}{n-1}.\end{aligned}$$

The integrals were evaluated using the more general identities (23) and (24) later.

For the next term in (13) we need $g^{(2\ell-1)}(y)|_{1/n}^{1}$. We use the Newton–Leibniz formula and evaluate $g^{(2\ell-1)}(1/n)$ and $g^{(2\ell-1)}(1)$ using respectively

$$\begin{aligned}g^{(2\ell-1)}(y) &= \sum_{j=0}^{2\ell-1} \binom{2\ell-1}{j} (\log y)^{(j)} \Big(\frac{y}{1-y}\Big)^{(2\ell-1-j)}, \\ g^{(2\ell-1)}(y) &= \sum_{j=0}^{2\ell-1} \binom{2\ell-1}{j} y^{(j)} \Big(\frac{\log y}{1-y}\Big)^{(2\ell-1-j)}.\end{aligned}$$

In the second sum there are only two nonzero terms, for $j=0$ and $j=1$, and using $\frac{\log y}{1-y} = -\sum_{j\ge 1} \frac{(1-y)^{j-1}}{j}$ we obtain, with some work, that

$$g^{(2\ell-1)}(1) = -\frac{(2\ell-2)!}{2\ell}.$$

For $g^{(2\ell-1)}(1/n)$, we use $(\frac{y}{1-y})^{(\mu)} = (\frac{1}{1-y})^{(\mu)}$ for $\mu > 0$, and after some more protracted work we obtain

$$g^{(2\ell-1)}(1/n) = -(\log n)\frac{(2\ell-1)!}{(1-n^{-1})^{2\ell}} + \sum_{j=1}^{2\ell-2} n^j \cdot \frac{(2\ell-1)_j}{j(1-n^{-1})^{2\ell-j}} + n^{2\ell-2}\cdot\frac{(2\ell-2)!}{1-1/n}.$$

Therefore,

$$\begin{aligned}g^{(2\ell-1)}(y)|_{1/n}^{1} &= -\frac{(2\ell-2)!}{2\ell} + (\log n)\frac{(2\ell-1)!}{(1-n^{-1})^{2\ell}} \\ &\quad - \sum_{j=1}^{2\ell-2} n^j \cdot \frac{(2\ell-1)_j}{j(1-n^{-1})^{2\ell-j}} - n^{2\ell-2}\cdot\frac{(2\ell-2)!}{1-1/n}.\end{aligned}$$

This term enters the RHS of (13) with the factor $n^{-2\ell}$, making the product of order n^{-2} regardless of $m \geq 1$. And the remainder term in (13) is of order n^{-2}, again independently of $m \geq 1$. So we choose the simplest $m = 1$. Collecting all the pieces we transform (13) into

$$\sum_{k=1}^{n-1} \frac{(k/n)\log(k/n)}{n-k} = -\zeta(2) + 1 + \frac{1}{2n} + \frac{\log n}{12n^2} + O\big(n^{-2}\big). \tag{14}$$

So, combining (10), (12), and (14), we have

$$\sum_{k=1}^{n-1} \frac{\log(k/n)}{n-k} = -\zeta(2) + \frac{\log(2\pi e)}{2n} + \frac{\log n}{12n^2} + O\big(n^{-2}\big)$$

which is the assertion of Lemma 2.2.

This completes the proof of Proposition 2.1.

2.2. *An ansatz for sharper results.* Knowing that $\mathbb{E}[D_n] = \zeta^{-1}(2)\log n + O(1)$, it seems natural to seek more refined estimates by imagining that

$$\mathbb{E}[D_n] = \zeta^{-1}(2)\log n + \sum_{j\geq 0} c_j n^{-j}$$

almost satisfies the recurrence, and then calculating c_j. Let us call this the *h-ansatz*, being analogous to a known expansion for h_n. So to use this ansatz we write

$$w_n := \sum_{j\geq 0} c_j n^{-j}$$

and seek to identify the c_j from the recurrence (9), which we rewrite as follows:

$$\begin{aligned} w_n &= \frac{d_1 \log n}{n h_{n-1}} + \frac{d_2}{n h_{n-1}} + \frac{1}{h_{n-1}} \sum_{k=2}^{n-1} \frac{w_k}{n-k}, \quad n \geq 2, \\ d_1 &= \frac{\zeta^{-1}(2)}{2}, \quad d_2 = \frac{\zeta^{-1}(2)}{2}\log(2\pi e). \end{aligned} \tag{15}$$

Here

$$\frac{\log n}{h_{n-1}} = 1 - \frac{\gamma}{\log n} + O\big(\log^{-2} n\big),$$

where

$$\gamma := 1 - \sum_{j=2}^{\infty} \frac{\zeta(j)-1}{j} \approx 0.5772156649, \tag{16}$$

is the Euler–Masceroni constant coming from $h_\nu = \log \nu + \gamma + O(\nu^{-1})$, [10]. For $n \geq 3$, using $\frac{1}{k(n-k)} = n^{-1}(\frac{1}{k} + \frac{1}{n-k})$, we have

$$\begin{aligned} \sum_{k=2}^{n-1} \frac{w_k}{n-k} &= \sum_{j\geq 0} c_j \sum_{k=2}^{n-1} \frac{1}{k^j (n-k)} \\ &= c_0 \left(h_{n-1} - \frac{1}{n-1} \right) + c_1 n^{-1} \left(2h_{n-1} - \frac{n}{n-1} \right) \\ &\quad + n^{-1} \sum_{j\geq 2} c_j \sum_{k=2}^{n-1} \left(\frac{1}{k^j} + \frac{1}{k^{j-1}(n-k)} \right) \end{aligned}$$

$$= c_0\left(h_{n-1} - \frac{1}{n-1}\right) + c_1 n^{-1}\left(2h_{n-1} - \frac{n}{n-1}\right) + n^{-1}\sum_{j\geq 2} c_j(\zeta(j)-1) + O(n^{-2}\log n).$$

Therefore,

$$\frac{d_1 \log n}{nh_{n-1}} + \frac{d_2}{nh_{n-1}} + \frac{1}{h_{n-1}}\sum_{k=2}^{n-1}\frac{w_k}{n-k} - w_n = \frac{d_1+c_1}{n} + \frac{1}{nh_{n-1}}\left(-d_1\gamma + d_2 - c_0 - c_1 + \sum_{j\geq 2} c_j(\zeta(j)-1)\right) + O(n^{-2}).$$

So, selecting

$$c_1 = -d_1 = -\frac{3}{\pi^2}, \qquad c_0 = d_2 + \sum_{j\geq 2}\left(c_j + \frac{d_1}{j}\right)(\zeta(j)-1), \tag{17}$$

(as suggested by (16)) we have

$$\frac{d_1 \log n}{nh_{n-1}} + \frac{d_2}{nh_{n-1}} + \frac{1}{h_{n-1}}\sum_{k=2}^{n-1}\frac{w_k}{n-k} - w_n = O(n^{-2}).$$

Therefore, $w_n = \sum_{j\geq 0} c_j n^{-j}$ satisfies (9) within the additive error $O(n^{-2})$, provided that $\{c_j\}_{j\geq 0}$ satisfies (17). It is worth noticing that c_0 is well defined for every $\{c_j\}_{j\geq 2}$ provided that the series in (17) converges. The constant c_0 can be viewed as a counterpart of the Euler–Masceroni constant γ. Strikingly, c_0 depends on *all* c_j, $j \geq 2$, while c_1 is determined uniquely from the requirement that w_n satisfies (15) within $O(n^{-2})$ error.

So the conclusion is as follows.

PROPOSITION 2.3. *Assuming the h-ansatz, there exists a constant c_0 such that*

$$\mathbb{E}[D_n] = \frac{6}{\pi^2}\log n + c_0 - \frac{3}{\pi^2}n^{-1} + O(n^{-2}). \tag{18}$$

This is another part of Theorem 1.1. One can calculate $\mathbb{E}[D_n]$ numerically via the basic recurrence, and doing so up to $n = 400,000$ gives a good fit[3] to (18) with $c_0 = 0.7951556604.....$ We do not have a conjecture for the explicit value of c_0.

In what follows, we will use only a weak corollary of (18), namely

$$\mathbb{E}[D_n] = \frac{6}{\pi^2}\log n + c_0 + O(n^{-1}). \tag{19}$$

Paradoxically, the actual value of c_0 will be immaterial as well.

2.3. *The recursion for variance.* Parallel to the recursion (5) for expectations, here is the recursion for variance.

LEMMA 2.4. *Setting $v_n := \mathrm{var}(D_n)$, we have*

$$v_n = \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{v_k + (\mathbb{E}[D_n] - \mathbb{E}[D_k])^2}{n-k}. \tag{20}$$

[3]Taking the coefficient of n^{-1} as unknown, the fit to this data is 0.30408, compared to $\frac{3}{\pi^2} = 0.30396$.

PROOF. Differentiating twice both sides of (3) at $u=0$, we get

$$\begin{aligned}\mathbb{E}[D_n^2] &= \frac{2}{h_{n-1}^3}\cdot h_{n-1} + \frac{2}{h_{n-1}^2}\sum_{k=1}^{n-1}\frac{\mathbb{E}[D_k]}{n-k} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\mathbb{E}[D_k^2]}{n-i}\\ &= \frac{2}{h_{n-1}^2}\left(1+\sum_{k=1}^{n-1}\frac{\mathbb{E}[D_k]}{n-k}\right) + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\mathbb{E}[D_k^2]}{n-k}\\ &= \frac{2\,\mathbb{E}[D_n]}{h_{n-1}} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\mathbb{E}[D_k^2]}{n-k}.\end{aligned}$$

Since $v_n = \mathbb{E}[D_n^2] - \mathbb{E}^2[D_n]$, the equation above becomes

$$v_n = \frac{2\,\mathbb{E}[D_n]}{h_{n-1}} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{v_k + \mathbb{E}^2[D_k]}{n-k} - \mathbb{E}^2[D_n].$$

The identity (20) holds because, by (5),

$$\frac{2\,\mathbb{E}[D_n]}{h_{n-1}} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\mathbb{E}^2[D_k]}{n-k} - \mathbb{E}^2[D_n] = \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{(\mathbb{E}[D_n]-\mathbb{E}[D_k])^2}{n-k}.$$ □

NOTE. The equation (46) could be obtained by using the "law of total variance". We preferred the above derivation as more direct in the present context, inconceivable without Laplace transform. Besides, the similar argument will be used later to derive a recurrence for variance of the edge length of the random path. It will be almost the "same" as (20), but with an unexpected, if not shocking, additive term -1 on the RHS.

2.4. *Sharp estimates of* $\operatorname{var}(D_n)$. Assuming the h-ansatz, and using (20), we are able to obtain the following sharp estimate, asserted as part of Theorem 1.1.

PROPOSITION 2.5. *Contingent on the h-ansatz,*

$$v_n = \frac{2\zeta(3)}{\zeta^3(2)}\log n + O(1). \quad n \ge 2.$$

NOTE. It is the term $(\mathbb{E}[D_n]-\mathbb{E}[D_k])^2$ in (20) that necessitates our reliance on the h-ansatz. Comfortingly, the first-order result $\operatorname{var}(D_n) \sim \frac{2\zeta(3)}{\zeta^3(2)}\log n$ follows from the CLT proof in Section 2.7, independently of the h-ansatz.

PROOF. By (19), we have

$$\begin{aligned}\big(\mathbb{E}[D_n]-\mathbb{E}[D_k]\big)^2 &= \zeta^{-2}(2)\big(\log(n/k)+O\big(k^{-1}\big)\big)^2\\ &= \zeta^{-2}(2)\big(\log^2(n/k)+O\big(k^{-1}\log(n/k)\big)+O\big(k^{-2}\big)\big).\end{aligned} \tag{21}$$

We need the estimates

$$\begin{aligned}\sum_{k=1}^{n-1}\frac{\log(n/k)}{k(n-k)} &= n^{-1}\sum_{k=1}^{n-1}\big(k^{-1}+(n-k)^{-1}\big)\log(n/k) = O\big(n^{-1}\log^2 n\big),\\ \sum_{k=1}^{n-1}\frac{1}{k^2(n-k)} &= n^{-1}\sum_{k=1}^{n-1}\big(k^{-2}+n^{-1}\big(k^{-1}+(n-k)^{-1}\big)\big) = O\big(n^{-1}\big).\end{aligned}$$

Consider the dominant term in (21). Observe that the function $\frac{\log^2(n/x)}{n-x}$ is convex. So, using (11) for $m=0$, we obtain

$$\begin{aligned}\sum_{k=1}^{n-1}\frac{\log^2(n/k)}{n-k} &= \int_1^n \frac{\log^2(n/x)}{n-x}\,dx + O\big(n^{-1}\log^2 n\big)\\ &= \int_0^1 \frac{\log^2(1/x)}{1-x}\,dx + O\big(n^{-1}\log^2 n\big)\\ &= 2\zeta(3) + O\big(n^{-1}\log^2 n\big).\end{aligned} \tag{22}$$

To explain the final equality, by induction on r and integrating by parts, we obtain

$$\int_0^1 z^j \log^r z\,dz = (-1)^r \frac{r!}{(j+1)^{r+1}}. \tag{23}$$

Consequently

$$\int_0^1 \frac{\log^r z}{1-z}\,dz = \int_0^1 (\log^r z)\sum_{j\ge 0} z^j\,dz = (-1)^r r!\zeta(r+1), \quad r\ge 1 \tag{24}$$

used for $r=2$ at (22). Now the recursion in Lemma 2.4 becomes

$$v_n = \frac{1}{h_{n-1}}\left(\frac{2\zeta(3)}{\zeta^2(2)} + O\big(n^{-1}\log^2 n\big) + \sum_{k=1}^{n-1}\frac{v_k}{n-k}\right).$$

Recalling that

$$\mathbb{E}[D_n] = \frac{1}{h_{n-1}}\left(1 + \sum_{k=1}^{n-1}\frac{\mathbb{E}[D_k]}{n-k}\right),$$

it follows that $w_n := |v_n - \frac{2\zeta(3)}{\zeta^2(2)}\,\mathbb{E}[D_n]|$ satisfies

$$w_n \le \frac{1}{h_{n-1}}\left(cn^{-1}\log^2 n + \sum_{k=1}^{n-1}\frac{w_k}{n-k}\right), \quad n\ge 2, w_1 = 0, \tag{25}$$

for some constant $c>0$. Let us prove that the sequence

$$z_n := c\left(\log^2(14) - \frac{\log^2(14n)}{n}\right)$$

satisfies

$$z_n \ge \frac{1}{h_{n-1}}\left(cn^{-1}\log^2 n + \sum_{k=1}^{n-1}\frac{z_k}{n-k}\right), \quad n\ge 2. \tag{26}$$

Because $z_1 = 0 = w_1$, we will get then, predictably by induction using (25), that $w_n \le z_n$. Let us prove (26). For $g(x) := -\frac{\log^2(14x)}{x}$, we have

$$\begin{aligned} g'(x) &= x^{-2}\big(\log^2(14x) - 2\log(14x)\big),\\ g''(x) &= -\frac{2}{x^3}\big[\log^2(14x) - 3\log(14x) + 1\big] < 0, \quad x\ge 1,\end{aligned}$$

because $\log(14) > 2.63 > \frac{3+\sqrt{5}}{2}$, the larger of two roots of $x^2 - 3x + 1$. Therefore, $g(x)$ is *concave* on $[1, \infty)$. So,

$$\begin{aligned}\frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{g(k)}{n-k} &\le g\left(\frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{k}{n-k}\right) = g\left(n - \frac{n-1}{h_{n-1}}\right)\\ &\le g(n) - g'(n)\frac{n-1}{h_{n-1}}\\ &= g(n) - n^{-2}\bigl(\log^2(14n) - 2\log(14n)\bigr)\frac{n-1}{h_{n-1}}.\end{aligned}$$

Since $z_k = c(\log^2(14) + g(k))$, we obtain then

$$\begin{aligned}&\frac{1}{h_{n-1}}\left(\frac{c\log^2 n}{n} + \sum_{k=1}^{n-1}\frac{z_k}{n-k}\right)\\ &\le z_n + \frac{c}{h_{n-1}}\left[\frac{\log^2 n}{n} - n^{-2}(n-1)\bigl(\log^2(14n) - 2\log(14n)\bigr)\right] < z_n,\end{aligned}$$

because the expression within square brackets is easily shown to be negative for $n \ge 2$. This establishes (26). □

2.5. *How correlated are leaf-heights?* Recall the statement of Theorem 1.6, copied below as Theorem 2.6. To study the interaction between the two levels of randomness, it is natural to consider the correlation between leaf heights. Write $D_n^{(1)}$ and $D_n^{(2)}$ for the time-heights, within the same realization of the random tree, of two distinct leaves chosen uniformly over pairs of leaves. Both time-heights individually are distributed as D_n, the time height of the uniformly random leaf. We study the correlation coefficient defined by

$$r_n = \frac{\mathbb{E}[D_n^{(1)} D_n^{(2)}] - \mathbb{E}^2[D_n]}{\mathrm{Var}(D_n)}.$$

THEOREM 2.6. *Contingent on the h-ansatz,* $r_n = O(\log^{-1} n)$.

PROOF. Recall the splitting distribution $n \to (L_n, R_n)$ at (1):

$$\mathbb{P}(L_n = i) = q(n,i) = \frac{n}{2h_{n-1}}\frac{1}{i(n-i)} = q(n, n-i), \quad 1 \le i \le n-1. \tag{27}$$

There is a natural recursion for $Z_\nu := D_\nu^{(1)} \cdot D_\nu^{(2)}$, as follows:

$$Z_\nu \stackrel{d}{=} \begin{cases} (\tau_\nu + D_i^{(1)})(\tau_\nu + D_i^{(2)}), & \text{with probability } q(\nu,i)\cdot\dfrac{(i)_2}{(\nu)_2},\\ (\tau_\nu + D_{\nu-i}^{(1)})(\tau_\nu + D_{\nu-i}^{(2)}), & \text{with probability } q(\nu,i)\cdot\dfrac{(\nu-i)_2}{(\nu)_2},\\ (\tau_\nu + D_i^{(1)})(\tau_\nu + D_{\nu-i}^{(2)}), & \text{with probability } q(\nu,i)\cdot\dfrac{i(\nu-i)}{(\nu)_2},\\ (\tau_\nu + D_i^{(2)})(\tau_\nu + D_{\nu-i}^{(1)}), & \text{with probability } q(\nu,i)\cdot\dfrac{i(\nu-i)}{(\nu)_2}. \end{cases} \tag{28}$$

Here τ_ν is the exponential$(h_{\nu-1})$ hold time. The first two cases correspond to the two leaves being in the same subtree, so their heights are dependent, whereas the last two cases correspond to the two leaves being in the different subtrees, so their heights are (conditionally) independent.

Consequently

$$\begin{aligned}\mathbb{E}[Z_\nu|L_\nu=i]&=\Big(\frac{2}{h_{\nu-1}^2}+\frac{2}{h_{\nu-1}}\,\mathbb{E}[D_i]+\mathbb{E}[Z_i]\Big)\frac{(i)_2}{(\nu)_2}\\&\quad+\Big(\frac{2}{h_{\nu-1}^2}+\frac{2}{h_{\nu-1}}\,\mathbb{E}[D_{\nu-i}]+\mathbb{E}[Z_{\nu-i}]\Big)\frac{(\nu-i)_2}{(\nu)_2}\\&\quad+2\Big(\frac{2}{h_{\nu-1}^2}+\frac{1}{h_{\nu-1}}\big(\mathbb{E}[D_i]+\mathbb{E}[D_{\nu-i}]\big)+\mathbb{E}[D_i]\cdot\mathbb{E}[D_{\nu-i}]\Big)\frac{i(\nu-i)}{(\nu)_2},\end{aligned}$$

or, with a bit of algebra,

$$\begin{aligned}\mathbb{E}[Z_\nu|L_\nu=i]&=\frac{2}{h_{\nu-1}^2}+\frac{2i\,\mathbb{E}[D_i]}{\nu h_{\nu-1}}+\frac{2(\nu-i)\,\mathbb{E}[D_{\nu-i}]}{\nu h_{\nu-1}}\\&\quad+\frac{1}{(\nu)_2}\big((i)_2\,\mathbb{E}[Z_i]+(\nu-i)_2\,\mathbb{E}[Z_{\nu-i}]+2i(\nu-i)\,\mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}]\big).\end{aligned}$$

Using (27) we obtain then

$$\begin{aligned}\mathbb{E}[Z_\nu]&=\sum_{i=1}^{\nu-1}q(\nu,i)\,\mathbb{E}[Z_\nu|L_\nu=i]\\&=\frac{2}{h_{\nu-1}^2}+\frac{2}{h_{\nu-1}^2}\sum_{i=1}^{\nu-1}\frac{\mathbb{E}[D_i]}{\nu-i}\\&\quad+\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}]+\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\,\mathbb{E}[Z_i]}{\nu-i}.\end{aligned}$$

So, using $\mathbb{E}[D_\nu]=\frac{1}{h_{\nu-1}}(1+\sum_{i=1}^{\nu-1}\frac{\mathbb{E}[D_i]}{\nu-i})$, we arrive at

$$\begin{aligned}\mathbb{E}[Z_\nu]&=\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\,\mathbb{E}[Z_i]}{\nu-i}\\&\quad+\frac{2\,\mathbb{E}[D_\nu]}{h_{\nu-1}}+\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}].\end{aligned}\tag{29}$$

We use (29) to sharply estimate $\mathbb{E}[Z_\nu]$ and then estimate $r_n=\frac{\mathbb{E}[Z_\nu]-\mathbb{E}^2[D_n]}{\operatorname{Var}(D_n)}$. To start,

$$\frac{2\,\mathbb{E}[D_\nu]}{h_{\nu-1}}=2\zeta^{-1}(2)+O\big(\log^{-1}\nu\big).$$

Second,

$$\begin{aligned}\mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}]&=\big[\zeta^{-1}(2)\log i+c_0+O\big(i^{-1}\big)\big]\\&\quad\times\big[\zeta^{-1}(2)\log(\nu-i)+c_0+O\big((\nu-i)^{-1}\big)\big].\end{aligned}$$

The leading contribution to $\sum_i\mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}]$ comes from

$$\begin{aligned}&\zeta^{-2}(2)\sum_{i=1}^{\nu-1}\log i\cdot\log(\nu-i)\\&\quad=\zeta^{-2}(2)(\nu-1)\log^2\nu+2\zeta^{-2}(2)\log\nu\sum_{i=1}^{\nu-1}\log(i/\nu)\end{aligned}$$

$$+\zeta^{-2}(2)\sum_{i=1}^{\nu-1}\log(i/\nu)\log\bigl((\nu-i)/\nu\bigr)$$

$$=\zeta^{-2}(2)\nu\log^2\nu+2\zeta^{-2}(2)\nu\log\nu\int_0^1\log x\,dx+O(\nu)$$

$$=\zeta^{-2}(2)\bigl(\nu\log^2\nu-2\nu\log\nu\bigr)+O(\nu).$$

The secondary contribution to $\sum_i \mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}]$ comes from $c_0\zeta^{-1}(2)(\log i+\log(\nu-i))$, and it equals $2c_0\zeta^{-1}(2)\nu\log\nu+O(\nu)$. The terms c_0, $O(i^{-1})$, $O((\nu-i)^{-1})$ contribute jointly $O(\nu)$. Altogether,

$$\sum_{i=1}^{\nu-1}\mathbb{E}[D_i]\,\mathbb{E}[D_{\nu-i}]=\zeta^{-2}(2)\bigl(\nu\log^2\nu-2\nu\log\nu\bigr)+2c_0\zeta^{-1}(2)\nu\log\nu+O(\nu).$$

Therefore, the equation (29) becomes

$$\mathbb{E}[Z_\nu]=\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\,\mathbb{E}[Z_i]}{\nu-i}+2\zeta^{-1}(2)$$
$$+\zeta^{-2}(2)(\log\nu-2-\gamma)+2c_0\zeta^{-1}(2)+O\bigl(\log^{-1}\nu\bigr). \tag{30}$$

Let us look at an approximate solution $\widetilde{E}(\nu):=A\log^2\nu+B\log\nu$. The RHS of the above equation is

$$\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)(A\log^2 i+B\log i)}{\nu-i}+2\zeta^{-1}(2)$$
$$+\zeta^{-2}(2)(\log\nu-2-\gamma)+2c_0\zeta^{-1}(2)+O\bigl(\log^{-1}\nu\bigr).$$

Here, since $\sum_i \frac{i-1}{\nu-i}=(\nu-1)(h_{\nu-1}-1)$, we have

$$\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\log^2 i}{\nu-i}=\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)(\log(i/\nu)+\log\nu)^2}{\nu-i}$$

$$=\frac{h_{\nu-1}-1}{h_{\nu-1}}\log^2\nu+\frac{2\log\nu}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\log(i/\nu)}{\nu-i}$$

$$+\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\log^2(i/\nu)}{\nu-i}$$

$$=\log^2\nu-\log\nu+\gamma+2\int_0^1\frac{x\log x}{1-x}\,dx+O\bigl(\log^{-1}\nu\bigr)$$

$$=\log^2\nu-\log\nu+\gamma+2\bigl(1-\zeta(2)\bigr)+O\bigl(\log^{-1}\nu\bigr),$$

and

$$\frac{1}{(\nu-1)h_{\nu-1}}\sum_{i=1}^{\nu-1}\frac{(i-1)\log i}{\nu-i}=\log\nu-1+O\bigl(\log^{-1}\nu\bigr).$$

Therefore, with $\widetilde{E}(\cdot)$ instead of $\mathbb{E}[Z_\cdot]$, the RHS of the equation (30) becomes

$$A\bigl(\log^2\nu-\log\nu+\gamma+2\bigl(1-\zeta(2)\bigr)\bigr)+B(\log\nu-1)$$
$$+\zeta^{-2}(2)(\log\nu-2-\gamma)+2(c_0+1)\zeta^{-1}(2)+O\bigl(\log^{-1}\nu\bigr).$$

And we need this to be equal to $\widetilde{E}(\nu) := A\log^2\nu + B\log\nu$ within an additive error $O(\log^{-1}\nu)$, meaning that

$$-A + B + \zeta^{-2}(2) = B,$$

$$A\big[\gamma + 2\big(1-\zeta(2)\big)\big] - B - (2+\gamma)\zeta^{-2}(2) + 2(c_0+1)\zeta^{-1}(2) = 0,$$

or explicitly

$$A = \zeta^{-2}(2), \qquad B = 2c_0\zeta^{-1}(2). \tag{31}$$

With these A and B, our approximation $\widetilde{E}(\nu)$ satisfies the same equation (30) as $\mathbb{E}[Z_\nu]$, excluding an exact value of the remainder term $O(\log^{-1}\nu)$, of course. Consequently, $\Delta(\nu) := |\mathbb{E}[Z_\nu] - \widetilde{E}(\nu)|$ satisfies

$$\Delta(\nu) \le \frac{1}{(\nu-1)h_{\nu-1}} \sum_{i=1}^{\nu-1} \frac{(i-1)\Delta(i)}{\nu-i} + O\big(\log^{-1}\nu\big), \quad \Delta(1) = 0. \tag{32}$$

With $\mathcal{U}_\nu := (\nu-1)\Delta(\nu)$, the resulting equation is a special case of the later equation (65) with the remainder term $O(\nu^{t-1}\log^{-1}\nu)$, when $t=2$. Applying the bound for the solution proved there, we obtain that $\mathcal{U}_\nu = O(\nu)$, or that $\Delta(\nu) = O(1)$. Thus

$$\mathbb{E}[Z_\nu] = A\log^2\nu + B\log\nu + O(1).$$

Combining this formula with (31), $r_n = \frac{\mathbb{E}[Z_\nu] - \mathbb{E}^2[D_n]}{\operatorname{Var}(D_n)}$ and $\mathbb{E}[D_n] = \zeta^{-1}(2)\log n + c_0 + O(n^{-1})$, we compute

$$r_n = \frac{\zeta^{-2}(2)\log^2 n + 2c_0\zeta^{-1}(2)\log n - (\zeta^{-1}(2)\log n + c_0)^2 + O(1)}{\frac{2\zeta(3)}{\zeta^3(2)}\log n + O(1)} = O\big(\log^{-1} n\big). \qquad \square$$

NOTE. We do not need the h-ansatz in the rest of the paper.

2.6. *Bounding the time-height of the random tree.* Consider now the time-height $\mathcal{D}_n$ of the random tree itself, that is, the maximum leaf time-height. We re-state Theorem 1.4, together with a tail bound on D_n.

PROPOSITION 2.7. (i) *For some $\rho > 0$ and all $\varepsilon \in (0, \pi^2/6 - 1)$,*

$$\mathbb{P}\left(D_n \ge \frac{6}{\pi^2}(1+\varepsilon)\log n\right) = O\big(n^{-\rho\varepsilon}\big).$$

(ii) *For some ρ' and all $\varepsilon \in (0,1)$,*

$$\mathbb{P}\big(\mathcal{D}_n \ge 2(1+\varepsilon)\log n\big) = O\big(n^{-\rho'\varepsilon}\big).$$

PROOF. (i) Since the tree with ν leaves has $\nu-1$ nonleaf vertices, rather crudely D_ν is stochastically dominated by the sum of $\nu-1$ independent exponentials with rate 1. Therefore, for $u < 1$, the Laplace transform $\phi_\nu(u) := \mathbb{E}[e^{uD_\nu}]$ is bounded above by $(1-u)^{-\nu}$. Recall (3):

$$\phi_\nu(u) = \frac{1}{h_{\nu-1} - u} \sum_{k=1}^{\nu-1} \frac{\phi_k(u)}{\nu-k}, \quad \nu \ge 2.$$

Pick $\varepsilon' < \varepsilon$ and introduce $\alpha = \frac{6}{\pi^2}(1+\varepsilon')$ (so $\alpha < 1$) and $\psi_\nu(u) = \exp(u\alpha\log\nu)$. Let us prove that

$$\psi_\nu(u) \ge \frac{1}{h_{\nu-1} - u} \sum_{k=1}^{\nu-1} \frac{\psi_k(u)}{\nu-k}, \tag{33}$$

if $u \in (0,1)$ is sufficiently small, and $\nu > 1$ sufficiently large.

First note that

$$\psi_k(u) = \psi_\nu(u)\exp\bigl(u\alpha\log(k/\nu)\bigr), \quad k \le \nu.$$

Therefore,

$$\begin{aligned}\frac{1}{\psi_\nu(u)(h_{\nu-1}-u)}\sum_{k=1}^{\nu-1}\frac{\psi_k(u)}{\nu-k} &= \frac{1}{h_{\nu-1}-u}\sum_{k=1}^{\nu-1}\frac{\exp(u\alpha\log(k/\nu))}{\nu-k}\\ &= \left(1-\frac{u}{h_{\nu-1}}\right)^{-1}\cdot\left(1+\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\exp(u\alpha\log(k/\nu))-1}{\nu-k}\right)\\ &= \left(1+\frac{u}{h_{\nu-1}}+O\left(\frac{u^2}{h_{\nu-1}^2}\right)\right)\\ &\quad\times\left[1+\frac{u}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\alpha\log(k/\nu)}{\nu-k}+O\left(\frac{u^2}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\log^2(k/\nu)}{\nu-k}\right)\right];\end{aligned}$$

(where we used $|e^x-1-x|\le x^2/2$, for $x\le 0$). So, since $\alpha=\zeta^{-1}(2)(1+\varepsilon')$,

$$\begin{aligned}\frac{1}{\psi_\nu(u)(h_{\nu-1}-u)}\sum_{k=1}^{\nu-1}\frac{\psi_k(u)}{\nu-k} &= 1+\frac{u}{h_{\nu-1}}\left(1+\alpha\sum_{k=1}^{\nu-1}\frac{\log(k/\nu)}{\nu-k}\right)+O\left(\frac{u^2}{h_{\nu-1}}\right)\\ &\le 1+\frac{u}{h_{\nu-1}}\left(1+\alpha\left(-\zeta(2)+\frac{\log(\nu e)}{\nu-1}\right)+O\left(\frac{u^2}{h_{\nu-1}}\right)\right)\\ &= 1-\frac{u}{h_{\nu-1}}\left(\varepsilon'-\zeta^{-1}(2)(1+\varepsilon)\frac{\log(\nu e)}{\nu-1}\right)+O\left(\frac{u^2}{h_{\nu-1}}\right).\end{aligned}\tag{34}$$

To justify the inequality above: $\frac{\log x}{1-x}$ increases for $x\le 1$, so that

$$\begin{aligned}\sum_{k=1}^{\nu-1}\frac{\log(k/\nu)}{\nu-k} &\le \int_0^1\frac{\log x}{1-x}\,dx-\int_0^{1/\nu}\frac{\log x}{1-x}\,dx\\ &\le -\zeta(2)+\frac{\nu}{\nu-1}\int_0^{1/\nu}\log(1/x)\,dx=-\zeta(2)+\frac{\log(\nu e)}{\nu-1}.\end{aligned}$$

The big-O term is uniform over all $u\in(0,1)$ and $\nu>1$. It follows then from (34) that there exist $u(\varepsilon')\in(0,1)$ and $\nu(\varepsilon')>1$ such that (33) holds for $u\in(0,u(\varepsilon'))$ and $\nu\ge\nu(\varepsilon')$. Furthermore, for $u\in(0,u(\varepsilon'))$ and $\nu\le\nu(\varepsilon')$,

$$\frac{\phi_\nu(u)}{\psi_\nu(u)}\le A(\varepsilon'):=\frac{(1-u(\varepsilon'))^{-\nu(\varepsilon')}}{\exp(u(\varepsilon')\alpha\log(\nu(\varepsilon'))},$$

since, for $\alpha\le 1$, $\frac{(1-u)^{-\nu}}{\exp(u\alpha\log\nu)}$ attains its maximum on $[0,\nu(\varepsilon')]$ at $\nu(\varepsilon')$. Combining this inequality with (33), by induction on ν we obtain that $\phi_\nu(u)\le A(\varepsilon')\psi_\nu(u)$ for all $\nu>1$ and $u\le u':=u(\varepsilon')$. The rest is easy:

$$\begin{aligned}\mathbb{P}\left(D_n\ge\frac{6}{\pi^2}(1+\varepsilon)\log n\right) &\le \frac{\mathbb{E}[\exp(u'D_n)]}{\exp(u'\frac{6}{\pi^2}(1+\varepsilon)\log n)}\le\frac{A(\varepsilon')\psi_\nu(u')}{\exp(u'\frac{6}{\pi^2}(1+\varepsilon)\log n)}\\ &\le A(\varepsilon')\exp\left[u'\left(\alpha-\frac{6}{\pi^2}(1+\varepsilon)\right)\log n\right]=\frac{A(\varepsilon')}{n^{\frac{6u'}{\pi^2}(\varepsilon-\varepsilon')}}.\end{aligned}$$

(ii) Predictably, we will use the union bound, which makes it necessary to upper-bound $\mathbb{P}(D_n \geq 2(1+\varepsilon)\log n)$. To this end, we use a cruder version of the argument in the part (i). Set $\alpha = 1 + \varepsilon/2$ and choose $u = \frac{1}{\alpha}$. Denoting $z_\nu = u/h_{\nu-1}$ we bound

$$\begin{aligned}\frac{1}{\psi_\nu(u)(h_{\nu-1}-u)}\sum_{k=1}^{\nu-1}\frac{\psi_k(u)}{\nu-k} &= \frac{1}{h_{\nu-1}-u}\sum_{k=1}^{\nu-1}\frac{\exp(u\alpha\log(k/\nu))}{\nu-k}\\ &= \frac{h_{\nu-1}}{h_{\nu-1}-u}\cdot\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{k/\nu}{\nu-k} = \frac{h_{\nu-1}}{h_{\nu-1}-u}\cdot\Big(1-\frac{\nu-1}{\nu h_{\nu-1}}\Big)^{u\alpha}\\ &\leq \exp\Big(-\log(1-z_\nu) - z_\nu\frac{\alpha(\nu-1)}{\nu}\Big).\end{aligned}$$

Since $z_\nu \to 0$, the last expression is below 1 for $\nu \in [\nu(\alpha), n]$. Therefore, arguing closely to the part (i), we see that $\phi_n(u) = O(\psi_n(u))$. Consequently

$$\mathbb{P}\big(D_n \geq 2(1+\varepsilon)\log n\big) = O\bigg(\frac{\psi_n(u)}{\exp(2u(1+\varepsilon)\log n)}\bigg) = O\big(n^{-\frac{2(1+\varepsilon)}{1+\varepsilon/2}+1}\big),$$

implying, by the union bound, that

$$\mathbb{P}\big(\mathcal{D}_n \geq 2(1+\varepsilon)\log n\big) \leq n\,\mathbb{P}\big(D_n \geq 2(1+\varepsilon)\log n\big) = O\big(n^{-\frac{2(1+\varepsilon)}{1+\varepsilon/2}+2}\big) = O\big(n^{-\frac{\varepsilon}{1+\varepsilon/2}}\big). \quad \square$$

2.7. *Asymptotic normality of D_n.* Here is one part of Theorem 1.7.

PROPOSITION 2.8. *In distribution, and with all of its moments,*

$$\frac{D_n - \zeta^{-1}(2)\log n}{\sqrt{\frac{2\zeta(3)}{\zeta^3(2)}\log n}} \Longrightarrow \text{Normal}(0,1).$$

In particular, this provides a proof of the first-order result

$$\text{var}(D_n) \sim \frac{2\zeta(3)}{\zeta^3(2)}\log n,$$

without having to rely on the h-ansatz, as stated in Theorem 1.1.

PROOF. By a general theorem due to Curtis [6], it suffices to show that for $|u| = \Theta(\log^{-1/2} n)$ and properly chosen $\alpha_1, \alpha_2 > 0$, the Laplace transform $\phi_n(u) = \mathbb{E}[e^{uD_n}]$ satisfies

$$\phi_n(u) = \big(1+o(1)\big)\exp\big[\big(u\alpha_1 + u^2\alpha_2\big)\log n\big]. \tag{35}$$

Recall from (3) that

$$\phi_\nu(u) = \frac{1}{h_{\nu-1}-u}\sum_{k=1}^{\nu-1}\frac{\phi_k(u)}{\nu-k}, \quad \nu \geq 2. \tag{36}$$

Define a function

$$\Psi_\nu(u) = \exp\big[\big(u\alpha_1 + u^2\alpha_2\big)\log\nu\big], \quad \nu \in [1, n];$$

obviously $\Psi_1(u) = 1 = \phi_1(u)$. We will use induction on ν to prove a stronger result, namely that there exist α_1 and α_2 such that for $|u| = \Theta(\log^{-1/2} n)$, the ratio $\frac{\phi_\nu(u)}{\Psi_\nu(u)}$ converges to 1,

uniformly over $n \geq \nu \to \infty$, sufficiently fast. Pick $\delta \in (0, 1/6)$, and set $\nu_n = \lceil \exp(\log^\delta n) \rceil$, so in particular $u \log \nu_n \to 0$. Introduce $\Psi^*_\nu(u) := 1 + u\alpha \log \nu$. For $u > 0$, we have

(37)
$$\begin{aligned}
&\frac{1}{(h_{\nu-1} - u)\Psi^*_\nu(u)} \sum_{k=1}^{\nu-1} \frac{\Psi^*_k(u)}{\nu - k} \\
&\quad = \frac{1}{h_{\nu-1}} \sum_{j_1, j_2 \geq 0} u^{j_1+j_2} \biggl(\frac{1}{h_{\nu-1}}\biggr)^{j_1} (-\alpha \log \nu)^{j_2} \\
&\qquad \times \biggl((1 + u\alpha \log \nu) h_{\nu-1} + u\alpha \sum_{k=1}^{\nu-1} \frac{\log(k/\nu)}{\nu - k}\biggr) \\
&\quad = \biggl(1 + u\biggl(\frac{1}{h_{\nu-1}} - \alpha \log \nu\biggr) + O\bigl(\alpha^2 u^2 \log^2 \nu\bigr)\biggr) \cdot \biggl(1 + u\alpha \log \nu - \frac{u}{h_{\nu-1}} \Theta(\alpha)\biggr) \\
&\quad = 1 + \frac{u}{h_{\nu-1}} \bigl(1 - \Theta(\alpha)\bigr) + O\bigl((\alpha^2 + 1) u^2 \log^2 \nu\bigr) \\
&\quad \begin{cases} > 1, & \text{if } \nu \leq \nu_n, \alpha > 0 \text{ and small,} \\ < 1, & \text{if } \nu \leq \nu_n, \alpha > 0 \text{ and large.} \end{cases}
\end{aligned}$$

(For the bottom part we used $\delta < \frac{1}{6}$.) And the inequalities are interchanged if $u < 0$. Combining this with (36), we conclude that $\phi_\nu(u) = 1 + O(|u| \log \nu) = \exp(O(|u| \log \nu))$, uniformly for $\nu \leq \nu_n$. So, for bounded α_1, α_2,

$$\lim_{n \to \infty} \max_{\nu \leq \nu_n} \biggl| \frac{\phi_\nu(u)}{\Psi_\nu(u)} - 1 \biggr| = 0. \tag{38}$$

Thus, we need to prove existence of α_1, α_2 such that the property above holds for $\nu \geq \nu_n$, as well. To this end, let us determine α_1 and α_2 from the condition that $\Psi_\nu(u)$, $(\nu \in [\nu_n, n])$, satisfies the recursive inequality

$$\Psi_\nu(u) \geq (\leq) \frac{1}{h_{\nu-1} - u} \biggl(\sum_{k=1}^{\nu-1} \frac{\Psi_k(u)}{\nu - k}\biggr), \quad \nu \in [\nu_n, n]. \tag{39}$$

First of all, we have

$$\Psi_k(u) = \Psi_\nu(u) \exp\bigl[\bigl(u\alpha_1 + u^2 \alpha_2\bigr) \log(k/\nu)\bigr], \quad k \leq \nu.$$

Therefore,

$$\begin{aligned}
&\frac{1}{\Psi_\nu(u)(h_{\nu-1} - u)} \sum_{k=1}^{\nu-1} \frac{\Psi_k(u)}{\nu - k} \\
&\quad = \frac{1}{h_{\nu-1} - u} \sum_{k=1}^{\nu-1} \frac{\exp[(u\alpha_1 + u^2\alpha_2) \log(k/\nu)]}{\nu - k} \\
&\quad = \biggl(1 - \frac{u}{h_{\nu-1}}\biggr)^{-1} \cdot \biggl(1 + \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\exp[(u\alpha_1 + u^2\alpha_2) \log(k/\nu)] - 1}{\nu - k}\biggr) \qquad (40) \\
&\quad = \biggl(1 - \frac{u}{h_{\nu-1}}\biggr)^{-1} \\
&\qquad \cdot \biggl(1 + \frac{1}{h_{\nu-1}} \int_0^1 \frac{\exp[(u\alpha_1 + u^2\alpha_2) \log x] - 1}{1 - x}\, dx + O\biggl(\frac{|u| \log \nu_n}{\nu_n}\biggr)\biggr).
\end{aligned}$$

In the final line, the bottom integral does not depend on ν. Let us first justify the remainder term. Define $f(k/\nu)$ as the kth term in the previous sum, $(k < \nu)$, and, for continuity, set $f(\nu/\nu) = -\nu^{-1}(u\alpha_1 + u^2\alpha_2)$. It can be checked that $f_k''(k/\nu)$ does not change its sign on $[1, \nu]$. So, replacing the sum with the integral for k varying continuously from 1 to ν, we introduce the error on the order of the sum of absolute values of

$$f(k/\nu)|_1^\nu \quad \text{and} \quad f_k'(k/\nu)|_1^\nu.$$

The dominant contribution to each of these terms comes from $k = 1$. Since for $\sigma \in (0, 1)$ the function $\frac{z^\sigma - 1}{z}$ decreases for $z \geq \sigma^{-1} \log \frac{1}{1-\sigma}$, we bound

$$\bigl|f(1/\nu)\bigr| \leq \frac{\exp(|u\alpha_1 + u^2\alpha_2| \log \nu_n) - 1}{\nu_n} = O\bigl(\nu_n^{-1}|u| \log \nu_n\bigr).$$

And the bound for $|f_k'(1/\nu)|$ is even better. So the sum in question is of order $O(\frac{|u| \log \nu_n}{\nu_n})$ uniformly for $\nu \geq \nu_n$. Extending the resulting integral to the full $[0, \nu]$, we introduce the second error on the order of

$$\int_0^1 \frac{\exp[(u\alpha_1 + u^2\alpha_2) \log(k/\nu)] - 1}{\nu - k}\, dk = O\biggl(\frac{|u| \log \nu_n}{\nu_n}\biggr). \tag{41}$$

The sum of the two error terms is $O(\nu^{-1}|u| \log \nu)$, and dividing it by $h_{\nu-1}$ we get $O(\frac{|u|}{\nu_n})$.

Let us sharply estimate the bottom integral in (40). By (41), the contribution to this integral coming from $x \in (0, 1/\nu_n]$ is $O(\nu_n^{-1}|u| \log \nu_n)$. And for $x \in [1/\nu_n, 1]$, we have $|u| \log(1/x) \leq |u| \log \nu_n \to 0$, that is, we can use the Taylor expansion

$$\begin{aligned}\frac{\exp[(u\alpha_1 + u^2\alpha_2) \log x] - 1}{1 - x} &= \frac{(u\alpha_1 + u^2\alpha_2) \log x}{1 - x} + \frac{(u\alpha_1 + u^2\alpha_2)^2 \log^2 x}{2(1 - x)} \\ &\quad + O\biggl(\frac{|u|^3 \log^3(1/x)}{1 - x}\biggr).\end{aligned}$$

This means that, at the price of the error term of the order $\nu_n^{-1}|u| \log \nu_n + |u|^3 \int_0^1 \frac{\log^3(1/x)}{1-x}\, dx$, we can use the expansion above for all $x \in (0, 1]$.

So, using (24), we obtain

$$\begin{aligned}&\frac{1}{h_{\nu-1}} \int_0^1 \frac{\exp[(u\alpha_1 + u^2\alpha_2) \log x] - 1}{1 - x}\, dx \\ &\quad = -\frac{\alpha_1\zeta(2)u}{h_{\nu-1}} + \frac{u^2}{h_{\nu-1}}\bigl[\alpha_1^2\zeta(3) - \alpha_2\zeta(2)\bigr] + O\biggl(\frac{|u|^3}{h_{\nu-1}} + \nu_n^{-1}|u|\biggr).\end{aligned}$$

Consequently, for $\nu \geq \nu_n (= \lceil \exp(\log^\delta n) \rceil)$,

$$\begin{aligned}\frac{1}{\Psi_\nu(u)(h_{\nu-1} - u)} \sum_{k=1}^{\nu-1} \frac{\Psi_k(u)}{\nu - k} &= 1 + \frac{u}{h_{\nu-1}}\bigl(1 - \alpha_1\zeta(2)\bigr) \\ &\quad + \frac{u^2}{h_{\nu-1}}\bigl[\alpha_1^2\zeta(3) - \alpha_2\zeta(2)\bigr] + O\biggl(\frac{|u|^3}{h_{\nu-1}} + \frac{|u|}{\nu_n}\biggr) \\ &= 1 + \frac{u}{h_{\nu-1}}\bigl(1 - \alpha_1\zeta(2)\bigr) + O\biggl(\frac{|u|^3}{h_{\nu-1}}\biggr),\end{aligned} \tag{42}$$

if we select $\alpha_2 = \frac{\alpha_1^2\zeta(3)}{\zeta(2)}$, which we certainly do. Suppose $u > 0$; set $\alpha_1 = \zeta^{-1}(2) + u^b$, $b \in (1, 2)$. Then, uniformly for $\nu \in [\nu_n, n]$, we have

$$1 + \frac{u}{h_{\nu-1}}\bigl(1 - \alpha_1\zeta(2)\bigr) + O\biggl(\frac{|u|^3}{h_{\nu-1}}\biggr) = 1 - \frac{\zeta^{-1}(2)u^{b+1}}{h_{\nu-1}}\bigl(1 + O(u^{2-b})\bigr) < 1.$$

So, (42) becomes

$$\frac{1}{h_{\nu-1}-u}\sum_{k=1}^{\nu-1}\frac{\Psi_k(u)}{\nu-k}\le\Psi_\nu(u).$$

This equation and the equation (36) together imply, by induction on $\nu\in[\nu_n,n]$, that $\limsup_{n\to\infty}\max_{\nu\in[\nu_n,n]}\frac{\phi_\nu(u)}{\Psi_\nu(u)}\le 1$. Now,

$$\begin{aligned}\Psi_\nu(u)&=\exp\bigl[(u\alpha_1+u^2\alpha_2)\log\nu\bigr]\\&=\exp\biggl[\Bigl(u\zeta^{-1}(2)+u^2\frac{\zeta(3)}{\zeta^3(2)}\Bigr)\log\nu+O\bigl(u^{b+1}\log\nu\bigr)\biggr]\\&\sim\exp\biggl[\Bigl(u\zeta^{-1}(2)+u^2\frac{\zeta(3)}{\zeta^3(2)}\Bigr)\log\nu\biggr],\end{aligned}$$

since $u^{b+1}\log n=O(\log^{-\frac{b-1}{2}}n)$ and $b>1$. Therefore,

$$\limsup_{n\to\infty}\max_{\nu\in[\nu_n,n]}\frac{\phi_\nu(u)}{\Psi_\nu(u)}\le 1.$$

Analogously, setting $\alpha_1=\zeta^{-1}(2)-u^b$, we have

$$\liminf_{n\to\infty}\min_{\nu\in[\nu_n,n]}\frac{\phi_\nu(u)}{\Psi_\nu(u)}\ge 1.$$

So, for $u=\Theta(\log^{-1/2}n)>0$ we have

$$\lim_{n\to\infty}\frac{\phi_n(u)}{\exp[(u\zeta^{-1}(2)+u^2\frac{\zeta(3)}{\zeta^3(2)})\log n]}=1.$$

The case $u<0$ is completely similar, so that the last equation holds for $u=-\Theta(\log^{-1/2}n)<0$ as well. □

2.8. *The moments of edge-heights of the leaves.* Recall that L_n denotes the edge-height of a uniform random leaf. In this section we prove Theorem 1.2 via the two propositions below.

PROPOSITION 2.9.

$$\mathbb{E}[L_n]=\frac{1}{2\zeta(2)}\log^2 n+\frac{\gamma\zeta(2)+\zeta(3)}{\zeta^2(2)}\log n+O(1). \tag{43}$$

PROOF. The straightforward recurrence for $\mathbb{E}[L_\nu]$ is

$$\mathbb{E}[L_\nu]=1+\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\mathbb{E}[L_k]}{\nu-k}. \tag{44}$$

Write $\mathbb{E}[L_\nu]=A\log^2\nu+B\log\nu+u_\nu$, so that $u_1=0$. We need to show that $u_\nu=O(1)$, if we select A and B appropriately. (Sure enough, these will be the constants in the claim.) Using (44), we have

$$\begin{aligned}u_\nu=1&+\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{u_k}{\nu-k}+A\biggl(\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\log^2 k}{\nu-k}-\log^2\nu\biggr)\\&+B\biggl(\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\log k}{\nu-k}-\log\nu\biggr).\end{aligned} \tag{45}$$

Here, by (14),

$$\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\log k}{\nu-k}-\log\nu=\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\log(k/\nu)}{\nu-k}=-\frac{\zeta(2)}{h_{\nu-1}}+\frac{\log(2\pi e)}{\nu h_{\nu-1}}+O\bigl(\nu^{-2}\bigr),$$

and, combining the equation above with (22), we also have

$$\begin{aligned}&\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\log^2 k}{\nu-k}-\log^2\nu\\&\quad=\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\log(k/\nu)\cdot(\log(k/\nu)+2\log\nu)}{\nu-k}\\&\quad=\frac{2\zeta(3)}{h_{\nu-1}}+O\bigl(\nu^{-1}\log\nu\bigr)+2\biggl(-\frac{\zeta(2)\log\nu}{h_{\nu-1}}+\frac{\log(2\pi e)\log\nu}{\nu h_{\nu-1}}+O\bigl(\nu^{-2}\log\nu\bigr)\biggr).\end{aligned}$$

Plugging the estimates above into (45) and using $\log\nu=h_{\nu-1}-\gamma+O(\nu^{-1})$, we get

$$\begin{aligned}u_\nu=\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{u_k}{\nu-k}+\bigl(1-2A\zeta(2)\bigr)+\frac{1}{h_{\nu-1}}\bigl[2A\bigl(\gamma\zeta(2)+\zeta(3)\bigr)-B\zeta(2)\bigr]\\+O\bigl(\nu^{-1}\log\nu\bigr).\end{aligned}$$

So, selecting A and B such that the (A,B)-dependent coefficients are both zeros, that is, $A=\frac{1}{2\zeta(2)}$, $B=\frac{\gamma\zeta(2)+\zeta(3)}{\zeta(2)}$, we arrive at

$$u_\nu=\frac{1}{h_{\nu-1}}\biggl(\sum_{k\in[\nu-1]}\frac{u_k}{\nu-k}+O\bigl(\nu^{-1}\log^2\nu\bigr)\biggr).$$

From the proof of Proposition 2.5 (starting with (25)), it follows that $u_\nu=O(1)$. □

PROPOSITION 2.10. $\operatorname{var}(L_n)=\frac{2\zeta(3)}{3\zeta^3(2)}\log^3 n+O(\log^2 n)$.

PROOF. (i) The key is

LEMMA 2.11. *Setting* $\bar v_n:=\operatorname{var}(L_n)$, *we have*

$$\bar v_n=-1+\frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\bar v_k+(\mathbb{E}[L_n]-\mathbb{E}[L_k])^2}{n-k}. \tag{46}$$

NOTE. In particular, $\bar v_2=0$ as it should be, since $L_2\equiv 1$, unlike D_2 which is distributed exponentially with rate 1.

PROOF. Differentiating twice both sides of (4) at $u=0$, we get

$$\begin{aligned}\mathbb{E}\bigl[L_\nu^2\bigr]&=1+\frac{1}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\mathbb{E}[L_k^2]}{\nu-k}+\frac{2}{h_{\nu-1}}\sum_{k\in[\nu-1]}\frac{\mathbb{E}[L_k]}{\nu-k}\\&=2\,\mathbb{E}[L_\nu]-1+\frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\mathbb{E}[L_k^2]}{n-k}\\&=2\,\mathbb{E}[L_\nu]-1+\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\mathbb{E}^2[L_k]}{\nu-k}+\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\bar v_k}{\nu-k}.\end{aligned}$$

Since $\bar{v}_\nu = \mathbb{E}[L_\nu^2] - \mathbb{E}^2[L_\nu]$, the equation above becomes

$$\bar{v}_\nu = 2E[L_\nu] - 1 + \frac{1}{h_{n-1}} \sum_{k=1}^{\nu-1} \frac{\bar{v}_k + \mathbb{E}^2[L_k]}{\nu - k} - \mathbb{E}^2[L_\nu],$$

and it is easy to check that this equation is equivalent to the claim. □

(ii) Using Proposition 2.9, we compute, for $A = \frac{1}{2\zeta(2)}$, $B = \frac{\gamma\zeta(2)+\zeta(3)}{\zeta(2)}$,

$$\begin{aligned}\big(\mathbb{E}[L_\nu] - \mathbb{E}[L_k]\big)^2 &= \big(A\big(\log^2 \nu - \log^2 k\big) + B(\log \nu - \log k) + O(1)\big)^2 \\ &= \big[2A\big(\log(k/\nu)\big)\log \nu\big]^2 + O\big[\mathcal{P}\big(\log(\nu/k)\big)\log \nu\big],\end{aligned}$$

where $\mathcal{P}(\eta)$ is a fourth-degree polynomial. Therefore, invoking (22), we have

$$\begin{aligned}\frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{(\mathbb{E}[L_\nu] - \mathbb{E}[L_k])^2}{\nu - k} &= \frac{4A^2 \log^2 \nu}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\log^2(k/\nu)}{\nu - k} + O(1) \\ &= \frac{8A^2\zeta(3)\log^2 \nu}{h_{\nu-1}} + O(1) = 8A^2\zeta(3)\log \nu + O(1).\end{aligned}$$

So, since $A = \frac{1}{2\zeta(2)}$, the equation (46) becomes

$$\bar{v}_\nu = \frac{2\zeta(3)}{\zeta(2)^2}\log \nu + O(1) + \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\bar{v}_k}{\nu - k}. \tag{47}$$

Let us use this recurrence to show that, for appropriately chosen A^*,

$$\bar{v}_\nu = \mathcal{V}_\nu + O\big(\log^2 \nu\big), \quad \mathcal{V}_\nu := A^* \log^3 \nu.$$

Here $O(\log^2 \nu)$ is uniform over all $\nu \geq 2$. We compute

$$\begin{aligned}\frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\log^3 k}{\nu - k} &= \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{(\log \nu + \log(k/\nu))^3}{\nu - k} \\ &= \frac{1}{h_{\nu-1}} \Bigg(\log^3 \nu h_{\nu-1} + 3\log^2 \nu \sum_{k=1}^{\nu-1} \frac{\log(k/\nu)}{\nu - k} \\ &\quad + 3\log \nu \sum_{k=1}^{\nu-1} \frac{\log^2(k/\nu)}{\nu - k} + \sum_{k=1}^{\nu-1} \frac{\log^3(k/\nu)}{\nu - k} \Bigg) \\ &= \log^3 \nu + \frac{3\log^2 \nu}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\log(k/\nu)}{\nu - k} + O(1) \\ &= \log^3 \nu - 3\zeta(2)\log \nu + O(1).\end{aligned}$$

It follows that

$$\begin{aligned}&\frac{2\zeta(3)}{\zeta(2)^2}\log \nu + \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\mathcal{V}_k}{\nu - k} \\ &\quad = \mathcal{V}_\nu + \left(\frac{2\zeta(3)}{\zeta(2)^2} - 3A^*\zeta(2)\right)\log \nu + O(1) = \mathcal{V}_\nu + O(1),\end{aligned}$$

if we select $A^* = \frac{2\zeta(3)}{3\zeta^3(2)}$. Combining this equation with (47), we obtain that $W_\nu := \bar{v}_\nu - \mathcal{V}_\nu$ satisfies

$$W_\nu = O(1) + \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{W_k}{\nu - k}.$$

Comparing with (44), we see that $W_\nu = O(E[L_\nu]) = O(\log^2 \nu)$. □

2.9. *Bounding the edge-height of the random tree.* As with the time-height in Section 2.6, we use a tail bound on the edge-height of a random leaf to obtain a tail bound for the edge-height of the tree itself.

PROPOSITION 2.12. *Let L_n denote the edge-height of the uniformly random leaf, and let $\mathcal{L}_n$ denote the largest edge-height of a leaf.*

(1) *For $\varepsilon > 0$,*

$$\mathbb{P}\left(L_n \geq \frac{3}{\pi^2}(1+\varepsilon)\log^2 n\right) = O\big(n^{-\Theta(\varepsilon)}\big).$$

(2) *Let $\beta = \min_{\alpha > 1/\log 2}[\alpha + \frac{4\alpha^2\zeta(3)}{\alpha \log 2 - 1}] \approx 42.9$. For $\varepsilon \in (0,1)$,*

$$\mathbb{P}\big(\mathcal{L}_n \geq (1+\varepsilon)\beta \log^2 n)\big) \leq \exp\big(-\Theta(\varepsilon \log n)\big).$$

PROOF. (1) First of all, $f_\nu(u) := \mathbb{E}[e^{uL_\nu}] \leq e^{u(\nu-1)}$ for $u \geq 0$. By (4), we have

$$f_\nu(u) := \mathbb{E}\big[e^{uL_\nu}\big] = \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{f_k(u)}{\nu - k}, \quad \nu \in [2, n].$$

Consider $u = \frac{v}{\log n}$, and introduce $g_k(u) = \exp(u\alpha \log^2 k)$, $\alpha > 0$ yet to be determined. Let us prove that there exists $v = v(\alpha)$ sufficiently small, and $\nu(\alpha)$ sufficiently large such that

$$g_\nu(u) \geq \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{g_k(u)}{\nu - k}, \quad \forall \nu \in \big[\nu(a), n\big]. \tag{48}$$

We compute

$$\begin{aligned} \frac{1}{g_\nu(u) h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{g_k(u)}{\nu - k} &= \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\exp[u\alpha(\log^2 k - \log^2 \nu)]}{\nu - k} \\ &= 1 + \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\exp[u\alpha(\log^2 k - \log^2 \nu)] - 1}{\nu - k} \\ &\leq 1 + \frac{u\alpha}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\log^2 k - \log^2 \nu}{\nu - k} + O\left(u^2 \log \nu \sum_{k=1}^{\nu-1} \frac{\log^2(k/\nu)}{\nu - k}\right). \end{aligned} \tag{49}$$

The big-Oh term is $O(u^2 \log \nu)$, and

$$\sum_{k=1}^{\nu-1} \frac{\log^2 k - \log^2 \nu}{\nu - k} = \sum_{k=1}^{\nu-1} \frac{\log^2(k/\nu)}{\nu - k} + 2\log \nu \sum_{k=1}^{\nu-1} \frac{\log(k/\nu)}{\nu - k} = -\frac{\pi^2}{3} \log \nu + O(1).$$

Therefore,

$$\frac{1}{g_\nu(u) h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{g_k(u)}{\nu - k} \leq 1 - \frac{\alpha u \pi^2}{3} + O\big(u \log^{-1} \nu + u^2 \log \nu\big),$$

implying that

$$\frac{e^u}{g_\nu(u)h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{g_k(u)}{\nu-k}\le 1-u\Big(\frac{\alpha\pi^2}{3}-1\Big)+O\big(u\log^{-1}\nu+u^2\log\nu\big). \tag{50}$$

Recalling that $u=\frac{v}{\log n}$, we obtain that for $\alpha>\frac{3}{\pi^2}$ there exists a sufficiently small $v(\alpha)>0$, and a sufficiently large $\nu(\alpha)$, such that for $v\le v(\alpha)$ and $n\ge\nu\ge\nu(\alpha)$, we have

$$\frac{e^u}{g_\nu(u)h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{g_k(u)}{\nu-k}\le 1. \tag{51}$$

Furthermore, for $u\le\frac{v(\alpha)}{\log n}$ and $\nu\le\nu(\alpha)$,

$$\frac{f_\nu(u)}{g_\nu(u)}\le\frac{\exp(u\nu)}{\exp(u\alpha\log^2\nu)}\le\exp\Big[\frac{v(\alpha)\nu(\alpha)}{\log n}\Big]. \tag{52}$$

By induction on $\nu\in[\nu(a),n]$, it follows that for those ν's

$$f_\nu(u)\le\exp\Big[\frac{v(\alpha)\nu(\alpha)}{\log n}\Big]g_\nu(u)\quad\Longrightarrow\quad f_n(u)\le\exp\Big[\frac{v(\alpha)\nu(\alpha)}{\log n}\Big]g_n(u),$$

provided that $\alpha>\frac{3}{\pi^2}$, and $u\le\frac{v(\alpha)}{\log n}$. So, given $\varepsilon>0$, we set $\alpha=\frac{3}{\pi^2}(1+\varepsilon/2)$, $\alpha'=\frac{3}{\pi^2}(1+\varepsilon)$, and bound

$$\mathbb{P}\Big(L_n\ge\frac{3}{\pi^2}(1+\varepsilon)\log^2 n\Big)=O\Big(\frac{\exp(u\alpha\log^2 n)}{\exp(u\alpha'\log^2 n)}\Big)\Big|_{u=\frac{v(\alpha)}{\log n}}=O\big(n^{-\Theta(\varepsilon)}\big).$$

This is the assertion of (1).

(2) We need a more explicit, but cruder, version of (50), again for $u=O(\log^{-1}n)$. Instead of (49), we bound

$$\frac{1}{g_\nu(u)h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{g_k(u)}{\nu-k}\le 1+\frac{u\alpha}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\log^2 k-\log^2\nu}{\nu-k}+\frac{2\alpha^2u^2\log^2\nu}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\log^2(k/\nu)}{\nu-k}.$$

Here

$$\begin{aligned}\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\log^2 k-\log^2\nu}{\nu-k}&=\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\log^2(k/\nu)}{\nu-k}+\frac{2\log\nu}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{\log(k/\nu)}{\nu-k}\\&\le\frac{2\zeta(3)}{h_{\nu-1}}+2\log\nu\cdot\log\Big(1-\frac{\nu-1}{\nu h_{\nu-1}}\Big)\\&\le\frac{\zeta(3)}{\log\nu}-2\frac{(\nu-1)\log\nu}{h_{\nu-1}\nu}\le\frac{2\zeta(3)}{\log\nu}-\log 2,\end{aligned}$$

since $\frac{\log^2 x}{1-x}$ is increasing, $\log x$ is concave, and $\log(1+z)\le z$, $(z>-1)$. So, we replace (50) with

$$\frac{e^u}{g_\nu(u)h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{g_k(u)}{\nu-k}\le 1-u\Big(\alpha\Big(\log 2-\frac{2\zeta(3)}{\log\nu}\Big)-1\Big)+u^2\big(1+4\alpha^2\zeta(3)\log\nu\big). \tag{53}$$

Given $\alpha>0$, the coefficient by u can be made arbitrarily close to $\alpha\log 2-1$ for ν sufficiently large, thus positive if $\alpha>\frac{1}{\log 2}$. Assuming the latter, for those large ν's, still below n, the RHS

expression in (53) is below 1 if $0 < u \le \frac{\alpha \log 2 - 1}{(4+\delta)\alpha^2 \zeta(3) \log n}$, $(\delta > 0)$, and $n \ge n(\delta)$. It follows that, as $n \to \infty$, we have $f_n(u) = O(g_n(u))$. Consequently, given $\alpha' > \alpha > \frac{1}{\log 2}$,

$$\begin{aligned} \mathbb{P}(\mathcal{L}_n \ge \alpha' \log^2 n) &\le n\,\mathbb{P}(L_n \ge \alpha' \log^2 n) = O\biggl(n \frac{e^{\alpha u \log^2 n}}{e^{\alpha' u \log^2 n}}\biggr) \\ &= O\bigl(\exp\bigl[\log n - u(\alpha' - \alpha)\log^2 n\bigr]\bigr) \to 0, \end{aligned} \tag{54}$$

if

$$u = \frac{\alpha \log 2 - 1}{(4+\delta)\alpha^2 \zeta(3) \log n}, \quad \alpha' > \alpha + \frac{(4+\delta)\alpha^2 \zeta(3)}{\alpha \log 2 - 1}.$$

Set

$$\beta = \min_{\alpha > 1/\log 2} \biggl[\alpha + \frac{4\alpha^2 \zeta(3)}{\alpha \log 2 - 1}\biggr],$$

and let $\hat{\alpha}$ stand for the point where the minimum is attained. Given $\varepsilon > 0$, there exists $\delta = \delta(\varepsilon) = \Theta(\varepsilon)$ such that

$$\alpha' := (1+\varepsilon)\beta > \hat{\alpha} + \frac{(4+\delta)\hat{\alpha}^2 \zeta(3)}{\hat{\alpha} \log 2 - 1},$$

implying, by (54) with $\alpha = \hat{\alpha}$, that $\mathbb{P}(\mathcal{L}_n \ge \alpha' \log^2 n) = O(\exp(-\Theta(\varepsilon) \log n))$. □

2.10. *Asymptotic normality of L_n.* Here is the second part of Theorem 1.7.

PROPOSITION 2.13. *In distribution, and with all of its moments,*

$$\frac{L_n - (2\zeta(2))^{-1} \log^2 n}{\sqrt{\frac{2\zeta(3)}{3\zeta^3(2)} \log^3 n}} \Longrightarrow \text{Normal}(0,1).$$

PROOF. Analogously to the proof of Proposition 2.8, it would seem natural to show that for $|u| = \Theta(\log^{-3/2} n)$ and properly chosen $\alpha_1 > 0$, $\alpha_2 > 0$, the Laplace transform $f_\nu(u) = \mathbb{E}[e^{uL_\nu}]$ satisfies

$$f_\nu(u) = \bigl(1 + o(1)\bigr) g_\nu(u), \qquad g_\nu(u) := \exp\bigl(u\alpha_1 \log^2 \nu + u^2 \alpha_2 \log^3 \nu\bigr), \tag{55}$$

uniformly for $\nu \le n$. But we could only prove (55) for $0 < u \le v \log^{-3/2} n$ and a fixed $v > 0$. Fortunately, that is all we need, thanks to a relatively recent extension of the Curtis lemma: it suffices to prove convergence of the sequence of Laplace transforms for the parameter v confined to a fixed interval $(0, \sigma]$; see [16] or [21].

Recall

$$f_\nu(u) = \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{f_k(u)}{\nu - k}, \quad \nu \ge 2. \tag{56}$$

Pick $\delta \in (0, 3/4)$ and set $\nu_n = \lceil \exp(\log^\delta n) \rceil$. For a constant α, introduce $g^*_\nu(u) := \exp(u\alpha \log^2 \nu)$. For $u > 0$, we have

$$\begin{aligned} &\frac{e^u}{h_{\nu-1} g^*_\nu(u)} \sum_{k=1}^{\nu-1} \frac{g^*_k(u)}{\nu - k} \\ &\quad = \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\exp(u\alpha \log(k\nu) \log(k/\nu))}{\nu - k} \end{aligned}$$

$$
\begin{aligned}
&= \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \bigl[1 + u\alpha \log(k\nu) \log(k/\nu) + O\bigl(u^2\alpha^2 \log^{\nu} \log^2(k/\nu)\bigr)\bigr] \\
&= e^u \left[1 + \frac{u\alpha\Theta(\log \nu)}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\log(k/\nu)}{\nu - k} + O\left(u^2\alpha^2 \log \nu \sum_{k=1}^{\nu-1} \frac{\log^2(k/\nu)}{\nu - k}\right)\right] \\
&= e^u \bigl[1 - u\Theta(\alpha) + O\bigl(u^2\alpha^2 \log \nu\bigr)\bigr] = 1 + u\bigl(1 - \Theta(\alpha)\bigr) + O\bigl(u^2\bigl(1 + \alpha^2 \log \nu\bigr)\bigr) \\
&\begin{cases} > 1, & \text{if } \nu \le \nu_n, \alpha > 0 \text{ and small,} \\ < 1, & \text{if } \nu \le \nu_n, \alpha > 0 \text{ and large.} \end{cases}
\end{aligned} \tag{57}
$$

(For the second line we used $e^z = 1 + z + O(z^2/2)$, uniformly for $z < 0$. For the bottom line we used $\delta < \frac{3}{4}$.) Combining this with (56), we conclude that $f_\nu(u) = \exp(O(u \log^2 \nu))$, uniformly for $\nu \le \nu_n$. So, for bounded α_1, α_2,

$$
\lim_{n\to\infty} \max_{\nu \le \nu_n} \left| \frac{f_\nu(u)}{g_\nu(u)} - 1 \right| = 0. \tag{58}
$$

Thus, we need to prove existence of α_1, α_2 such that the analogous relation holds uniformly for all $\nu \in [\nu_n, n]$. Predictably, we select α_1 and α_2, requiring that $g_\nu(u)$ is the asymptotically best fit for the recurrence (56) for all $\nu \in [\nu_n, n]$. To begin,

$$
\begin{aligned}
g_k(u) &= g_\nu(u) \exp\bigl[u\alpha_1 G_1(k/\nu, \nu) + u^2\alpha_2 G_2(k/\nu, \nu)\bigr], \\
G_1(k/\nu, \nu) &:= \log\left(\frac{k}{\nu}\right) \log(k\nu) \le 0, \\
G_2(k/\nu, \nu) &:= \log\left(\frac{k}{\nu}\right)\left[3 \log k \log \nu + \log^2\left(\frac{k}{\nu}\right)\right] \le 0.
\end{aligned} \tag{59}
$$

And, since $G_j(k/\nu, \nu)$ are *nonpositive*, the Taylor expansion of the exponential function holds for $u > 0$, even though $|uG_1(1/n, n)| = O(\sqrt{\log n})$. (Notice that $u^2|G_2(1/\nu, \nu)| = O(1)$.) So, proceeding analogously to (40),

$$
\begin{aligned}
&\frac{e^u}{g_\nu(u) h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{g_k(u)}{\nu - k} \\
&\quad = \frac{e^u}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\exp[u\alpha_1 G_1(k/\nu, \nu) + u^2\alpha_2 G_2(k/\nu, \nu)]}{\nu - k} \\
&\quad = e^u \cdot \left(1 + \frac{1}{h_{\nu-1}} \sum_{k=1}^{\nu-1} \frac{\exp[u\alpha_1 G_1(k/\nu, \nu) + u^2\alpha_2 G_2(k/\nu, \nu)] - 1}{\nu - k}\right) \\
&\quad = e^u \cdot \Bigl(1 + \frac{1}{h_{\nu-1}} \int_0^1 \frac{\exp[u\alpha_1 G_1(x, \nu) + u^2\alpha_2 G_2(x, \nu)] - 1}{1 - x}\, dx \\
&\qquad + O\left(\frac{\exp(-u \log^2 \nu)}{\nu \log \nu}\right),
\end{aligned} \tag{60}
$$

where *uniformly* for $x \in (0, 1]$

$$
\begin{aligned}
\frac{\exp[u\alpha_1 G_1(x, \nu) + u^2\alpha_2 G_2(x, \nu)] - 1}{1 - x} &= \frac{u\alpha_1 G_1(x, \nu) + u^2\alpha_2 G_2(x, \nu)}{1 - x} \\
&\quad + \frac{(u\alpha_1 G_1(x, \nu) + u^2\alpha_2 G_2(x, \nu))^2}{2(1 - x)}
\end{aligned}
$$

$$+ O\Big(\frac{u^3 \log^3(1/x) \log^3 \nu}{1-x}\Big).$$

Using (59), and (24), we have then

$$\begin{aligned}
&\int_0^1 \frac{\exp[u\alpha_1 G_1(x,\nu) + u^2\alpha_2 G_2(x,\nu)] - 1}{1-x}\,dx \\
&\quad = \alpha_1 u\big(-2\zeta(2)\log\nu + 2\zeta(3)\big) + u^2\Big(\frac{\alpha_1^2}{2}\big(8\zeta(3)\log^2\nu - 24\zeta(4)\log\nu\big) \\
&\qquad + \alpha_2\big(-3\zeta(2)\log^2\nu + 6\zeta(3)\log\nu - 6\zeta(4)\big)\Big) + O\big(u^3\log^3\nu\big).
\end{aligned}$$

Upon expansion $e^u = 1 + u + u^2/2 + O(u^3)$, the bottom RHS in (60) then becomes

$$\begin{aligned}
&1 + u\Big(1 + \alpha_1\frac{-2\zeta(2)\log\nu + 2\zeta(3)}{h_{\nu-1}}\Big) + u^2\Big(\frac{\frac{\alpha_1^2}{2}(8\zeta(3)\log^2\nu - 24\zeta(4)\log\nu)}{h_{\nu-1}} \\
&\qquad + \frac{a_2(-3\zeta(2)\log^2\nu + 6\zeta(3)\log\nu - 6\zeta(4))}{h_{\nu-1}} + \alpha_1\frac{-2\zeta(2)\log\nu + 2\zeta(3)}{h_{\nu-1}}\Big) \\
&\qquad + O\big(u^3\log^2\nu\big) \\
&\quad = 1 + u\Big(1 + \alpha_1\frac{-2\zeta(2)\log\nu + 2\zeta(3)}{h_{\nu-1}}\Big) + O\big(u^3\log^2\nu\big),
\end{aligned} \tag{61}$$

if, leaving $\alpha_1 = \alpha_1(\nu) > 0$ to be determined shortly, we select $\alpha_2 = \alpha_2(\nu)$ to make the coefficient by u^2 equal to zero. Looking closer at the coefficient by u^2, we see that

$$\alpha_2 = \frac{4\alpha_1^2\zeta(3)}{3\zeta(2)} + O\big(\log^{-1}\nu\big).$$

The rest is short. Pick $\alpha_1 = (2\zeta(2))^{-1}(1+u^b)$, $b < 2\delta/3$. Since $u = \Theta(\log^{-3/2} n)$, the bottom expression in (61) becomes

$$\begin{aligned}
&1 + u\big(-u^b + O\big(\log^{-1}\nu\big)\big) + O\big(u^3\log^2\nu\big) \\
&\quad = 1 - u^{b+1}\big(1 + O\big(u^{-b}\log^{-1}\nu\big) + O\big(u^{2-b}\log^2 n\big)\big) \\
&\quad = 1 - u^{b+1}\big[1 + O\big((\log n)^{-\delta+3b/2}\big) + O\big((\log n)^{-1+3b/2}\big)\big] < 1.
\end{aligned}$$

So, it follows from (60) that

$$\frac{e^u}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{g_k(u)}{\nu-k} < g_\nu(u), \qquad \nu \in [\nu_n, n].$$

Combining this recursive inequality with (58), we conclude that

$$\limsup_{n\to\infty}\max_{\nu\in[\nu_n,n]}\frac{f_\nu(u)}{g_\nu(u)} \le 1.$$

Now,

$$\begin{aligned}
g_\nu(u) &= \exp\big(u\alpha_1\log^2\nu + u^2\alpha_2\log^3\nu\big) \\
&= \exp\Big[u\big((2\zeta(2))^{-1}(1+u^b)\big)\log^2\nu + u^2\Big(\frac{\zeta(3)}{3\zeta^3(2)} + o(1)\Big)\log^3\nu\Big] \\
&= \exp\Big[u(2\zeta(2))^{-1}\log^2\nu + u^2\frac{\zeta(3)}{3\zeta^3(2)}\log^3\nu + o(1) + O\big(u^{b+1}\log^2\nu\big)\Big]
\end{aligned}$$

$$= \bigl(1+o(1)\bigr)\exp\biggl[u\bigl(2\zeta(2)\bigr)^{-1}\log^2\nu + u^2\frac{\zeta(3)}{3\zeta^3(2)}\log^3\nu\biggr],$$

if we select $b > 1/3$. Since $b < 2\delta/3$, a desired b exists provided that $\delta > 1/2$, the constraint compatible with the initial restriction $\delta < 3/4$. We conclude that for $\delta \in (1/2, 3/4)$

$$\limsup_{n\to\infty}\max_{\nu\in[\nu_n,n]} f_\nu(u)\exp\biggl[-u\bigl(2\zeta(2)\bigr)^{-1}\log^2\nu - u^2\frac{\zeta(3)}{3\zeta^3(2)}\log^3\nu\biggr] \le 1.$$

Likewise, picking $\alpha_1 = (2\zeta(2))^{-1}(1-u^b)$, $b < 2\delta/3$, we obtain

$$\liminf_{n\to\infty}\min_{\nu\in[\nu_n,n]} f_\nu(u)\exp\biggl[-u\bigl(2\zeta(2)\bigr)^{-1}\log^2\nu - u^2\frac{\zeta(3)}{3\zeta^3(2)}\log^3\nu\biggr] \ge 1.$$

This verifies (55), as required. □

2.11. *How soon do the species part their ways?* Recall from Section 1.2 the notion of *pruned spanning tree* on t random leaves within the tree model on n leaves. Write $S_{n,t}$ for the edge height of the first branchpoint in the pruned tree. In other words, the number of edges from the root to the vertex after which the t sampled leaves are first split into some $(k, t-k)$ leaf subsets. Conditioned on the size k of the left subtree at the root of the tree with n leaves, the probability that the t sampled leaves are all in this left subtree is $\frac{(k)_t}{(n)_t}$. Therefore, since $q(n,k) = \frac{n}{2h_{n-1}k(n-k)}$, we obtain the recursion

$$\mathbb{E}[S_{n,t}] = 1 + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{(n/k)\,\mathbb{E}[S_{k,t}]}{n-k}\frac{(k)_t}{(n)_t}, \quad n \ge t \ge 2, \tag{62}$$

($\mathbb{E}[S_{k,1}] = 0$), or, introducing $\Phi_{n,t} = (n-1)_{t-1}\,\mathbb{E}[S_{n,t}]$,

$$\Phi_{n,t} = (n-1)_{t-1} + \frac{1}{h_{n-1}}\sum_{k=1}^{n-1}\frac{\Phi_{k,t}}{n-k}. \tag{63}$$

PROPOSITION 2.14.

$$\mathbb{E}[S_{n,t}] = \frac{\log n}{h_{t-1}} + O(1) \quad \text{as } n\to\infty.$$

PROOF. Given $\alpha > 0$, define

$$U_{\nu,t} = \Phi_{\nu,t} - \alpha\nu^{t-1}\log\nu.$$

Then, by (63), we have

$$U_{\nu,t} = (\nu-1)_{t-1} + \frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{U_{k,t}}{\nu-k} + \alpha\biggl(\frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{k^{t-1}\log k}{\nu-k} - \nu^{t-1}\log\nu\biggr), \tag{64}$$

and the coefficient by α equals

$$\begin{aligned}&\frac{\nu^{t-1}}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{(k/\nu)^{t-1}[\log\nu + \log(k/\nu)]}{\nu-k} - \nu^{t-1}\log\nu\\ &\quad= \frac{\nu^{t-1}}{h_{\nu-1}}\biggl(\log\nu\sum_{k=1}^{\nu-1}\frac{(k/\nu)^{t-1}-1}{\nu-k} + \log\nu\sum_{k=1}^{\nu-1}\frac{1}{\nu-k} + \sum_{k=1}^{\nu-1}\frac{(k/\nu)^{t-1}\log(k/\nu)}{\nu-k}\biggr)\\ &\qquad- \nu^{t-1}\log\nu\end{aligned}$$

$$= \frac{\nu^{t-1}}{h_{\nu-1}}\left(\log\nu\int_0^1 \frac{x^{t-1}-1}{1-x}\,dx + h_{\nu-1}\log\nu + O(1)\right) - \nu^{t-1}\log\nu$$
$$= -\frac{\nu^{t-1}\log\nu}{h_{\nu-1}}h_{t-1} + O\bigl(\nu^{t-1}\log^{-1}\nu\bigr).$$

So, the equation (64) becomes

$$U_{\nu,t} = (\nu-1)_{t-1} + \alpha\left(-\frac{\nu^{t-1}\log\nu}{h_{\nu-1}}h_{t-1} + O\bigl(\nu^{t-1}\log^{-1}\nu\bigr)\right) + \frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{U_{k,t}}{\nu-k}$$
$$= O\bigl(\nu^{t-1}\log^{-1}\nu\bigr) + \frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{U_{k,t}}{\nu-k} \tag{65}$$

if we choose $\alpha = \frac{1}{h_{t-1}}$. Consequently, for some constant β,

$$|U_{\nu,t}| \le \beta\nu^{t-1}\log^{-1}\nu + \frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{|U_{k,t}|}{\nu-k}.$$

For a constant B, to be chosen shortly, we have

$$\beta\nu^{t-1}\log^{-1}\nu + \frac{1}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{Bk^{t-1}}{\nu-k}$$
$$= \beta\nu^{t-1}\log^{-1}\nu + \frac{B\nu^{t-1}}{h_{\nu-1}}\sum_{k=1}^{\nu-1}\frac{(k/\nu)^{t-1}}{\nu-k}$$
$$= \beta\nu^{t-1}\log^{-1}\nu + \frac{B\nu^{t-1}}{h_{\nu-1}}\left(h_{\nu-1} + \int_0^1\frac{x^{t-1}-1}{1-x}\,dx + O\bigl(\nu^{-1}\bigr)\right)$$
$$= \beta\nu^{t-1}\log^{-1}\nu + \frac{B\nu^{t-1}}{h_{\nu-1}}\bigl(h_{\nu-1} - h_{t-1} + O\bigl(\nu^{-1}\bigr)\bigr) < B\nu^{t-1},$$

provided that

$$\beta\log^{-1}\nu - B\left(\frac{h_{t-1}}{h_{\nu-1}} + O\bigl(\nu^{-1}\bigr)\right) < 0.$$

And this inequality holds for all $\nu \ge 2$, if we choose B sufficiently large. It follows, by induction on ν, that $|U_{\nu,t}| \le B\nu^{t-1}$. Consequently

$$\Phi_{\nu,t} = \alpha\nu^{t-1}\log\nu + O\bigl(\nu^{t-1}\bigr),$$

so that

$$\mathbb{E}[S_{\nu,t}] = \frac{\Phi_{\nu,t}}{(\nu-1)_{t-1}} = \alpha\log\nu + O(1), \quad \alpha = \frac{1}{h_{t-1}}.$$ □

Within the same notion of *pruned spanning tree* on t random leaves within the tree model on n leaves, a more complicated statistic is the edge-length of the pruned tree, which we denote as $S^*_{n,t}$. To derive the counterpart of (62), notice that the total number of ways to partition the set $[n]\setminus[t]$ into two trees, the left one of cardinality k, with $t_1 \le t$ vertices from $[t]$ and the right one of cardinality $n-k$, with $t_2 = t - t_1$ remaining vertices from $[t]$, equals

$\binom{n-t}{k-t_1}$. Defining $S^*_{n,0}=0$, $S^*_{n,1}=0$, $\forall n\ge 0$, we have the recursion: for $n\ge t\ge 2$,

$$\begin{aligned}\mathbb{E}[S^*_{n,t}] &= 1+\sum_{k=1}^{n-1}\frac{n}{2h_{n-1}k(n-k)}\cdot\binom{n}{k}^{-1}\\ &\quad\times\sum_{t_1\le t}\binom{n-t}{k-t_1}\left(\mathbb{E}[S^*_{k,t_1}]+\mathbb{E}[S^*_{n-k,t_2}]\right)\\ &= 1+\sum_{k=1}^{n-1}\frac{n}{2h_{n-1}k(n-k)}\sum_{t_1\le t}\frac{(k)_{t_1}(n-k)_{t_2}}{(n)_t}\left(\mathbb{E}[S^*_{k,t_1}]+\mathbb{E}[S^*_{n-k,t_2}]\right)\\ &= 1+\frac{1}{h_{n-1}}\sum_{k=2}^{n-1}\sum_{t_1=2}^{t}\frac{(k-1)_{t_1-1}(n-k)_{t_2}}{(n-1)_{t-1}(n-k)}\mathbb{E}[S^*_{k,t_1}].\end{aligned}$$

Therefore, with $\Psi_{n,t}:=(n-1)_{t-1}\,\mathbb{E}[S^*_{n,t}]$, so that $\Psi_{n,0}=\Psi_{n,1}=0$, $\Psi_{n,t}=0$ for $n<t$, we obtain

$$\Psi_{n,t}=(n-1)_{t-1}+\frac{1}{h_{n-1}}\sum_{t_1=2}^{t}\sum_{k=2}^{n-1}\frac{(n-k)_{t_2}}{n-k}\Psi_{k,t_1},\quad n\ge t\ge 2. \tag{66}$$

This equation is similar to (63). Because of the new factor $(n-k)_{t_2}$, we will use

$$(a)_b=\sum_{j=1}^{b}s(b,j)a^j, \tag{67}$$

where $s(b,j)$ is the signed Stirling number of the first kind, so that $|s(b,j)|$ is the total number of permutations of $[b]$ with j cycles.

We now repeat the statement of Theorem 1.8.

PROPOSITION 2.15.

$$\mathbb{E}[\mathcal{S}^*_{n,t}]=\alpha(t)\log n+O(1),\quad \alpha(t)=\left(h_{t-1}-\sum_{t_1+t_2=t}\frac{(t_1-1)!(t_2-1)!}{(t-1)!}\right)^{-1}.$$

PROOF. The argument is guided by the proof of Proposition 2.14. Given $\alpha>0$, define

$$V_{\nu,t}=\Psi_{\nu,t}-\alpha\nu^{t-1}\log\nu,\quad \nu\ge t\ge 2.$$

By (66), we have

$$\begin{aligned}V_{\nu,t} &= (\nu-1)_{t-1}+\frac{1}{h_{n-1}}\sum_{t_1=2}^{t}\sum_{k=2}^{\nu-1}\frac{(\nu-k)_{t_2}}{\nu-k}V_{k,t_1}\\ &\quad+\alpha\left(\frac{1}{h_{\nu-1}}\sum_{t_1=2}^{t}\sum_{k=2}^{\nu-1}\frac{(\nu-k)_{t_2}}{\nu-k}k^{t_1-1}\log k-\nu^{t-1}\log\nu\right).\end{aligned} \tag{68}$$

Consider the factor by α. By (67),

$$\sum_{k=2}^{\nu-1}\frac{(\nu-k)_{t_2}}{\nu-k}k^{t_1-1}\log k=\sum_{j=0}^{t_2}s(t_2,j)\Sigma(\nu,t_1,j),$$

$$\Sigma(\nu,t_1,j):=\sum_{k=2}^{\nu-1}(\nu-k)^{j-1}k^{t_1-1}\log k.$$

Recalling that $t_1 > 1$, we write

$$\Sigma(\nu, t_1, 0) = \sum_{k=2}^{\nu-1} \frac{k^{t_1-1}\log k}{\nu - k} = \sum_{k=2}^{\nu-1} \frac{k^{t_1-1}(\log \nu + \log(k/\nu))}{\nu - k}$$
$$= (\log \nu)\left(\nu^{t_1-1} h_{\nu-1} + \sum_{k=2}^{\nu-1} \frac{k^{t_1-1} - \nu^{t_1-1}}{\nu - k}\right) + \sum_{k=2}^{\nu-1} \frac{k^{t_1-1}\log(k/\nu)}{\nu - k},$$

and

$$\sum_{k=2}^{\nu-1} \frac{k^{t_1-1} - \nu^{t_1-1}}{\nu - k} = \nu^{t_1-1}\left(\int_0^1 \frac{x^{t_1-1} - 1}{1-x}\,dx + O(\nu^{-1})\right)$$
$$= \nu^{t_1-1}\left(-\int_0^1 \sum_{s=0}^{t_1-2} x^s\,dx + O(\nu^{-1})\right)$$
$$= -\nu^{t_1-1} h_{t_1-1} + O(\nu^{t_1-2}),$$

while it is easy to see that $\sum_{k=2}^{\nu-1} \frac{k^{t_1-1}\log(k/\nu)}{\nu-k}$ is of order $\nu^{t_1-1}\int_0^1 \frac{x^{t_1-1}\log x}{1-x}\,dx = O(\nu^{t_1-1})$. Therefore,

$$\Sigma(\nu, t_1, 0) = (h_{\nu-1} - h_{t_1-1})\nu^{t_1-1}\log \nu + O(\nu^{t_1-1}). \tag{69}$$

Suppose that $j > 0$. Then

$$\Sigma(\nu, t_1, j) = \nu^{t_1+j-1}\left(\nu^{-1}\sum_{k=1}^{\nu-1}(1-k/\nu)^{j-1}(k/\nu)^{t_1-1}\big[\log \nu + \log(k/\nu)\big]\right)$$
$$= \nu^{t_1+j-1}\Bigg[(\log \nu)\int_0^1 (1-x)^{j-1}x^{t_1-1}\,dx$$
$$+ \int_0^1 (1-x)^{j-1}x^{t_1-1}(\log x)\,dx + O(\nu^{-1}\log \nu)\Bigg] \tag{70}$$
$$= \frac{(j-1)!(t_1-1)!}{(t_1+j-1)!}\cdot \nu^{t_1+j-1}\log \nu + O(\nu^{t_1+j-2}\log \nu),$$

and $t_1 + j - 1 \le t_1 + t_2 - 1 = t - 1$. Combining (69) and (70), and using $s(b, b) = 1$, $s(b, 0) = 0$ for $b > 0$, we have

$$\sum_{k=2}^{\nu-1} \frac{(\nu-k)_{t_2}}{\nu-k} k^{t_1-1}\log k$$
$$= (h_{\nu-1} - h_{t_1-1})\nu^{t_1-1}\log \nu + \frac{(t_2-1)!(t_1-1)!}{(t-1)!}\nu^{t-1}\log \nu + O(\nu^{t-2}\log \nu).$$

So, the factor by α in (68) is

$$\frac{\nu^{t-1}\log \nu}{h_{\nu-1}}\left(h_{\nu-1} - h_{t-1} + \sum_{t_1=1}^{t} \frac{(t_2-1)!(t_1-1)!}{(t-1)!} + O(\nu^{-1})\right) - \nu^{t-1}\log \nu$$
$$= \frac{\nu^{t-1}\log \nu}{h_{\nu-1}}\left(-h_{t-1} + \sum_{t_1=1}^{t} \frac{(t_2-1)!(t_1-1)!}{(t-1)!} + O(\nu^{-1})\right).$$

Consequently the equation (68) becomes

$$V_{\nu,t} = (\nu-1)_{t-1} + \alpha\frac{\nu^{t-1}\log\nu}{h_{\nu-1}}\left(-h_{t-1} + \sum_{t_1=1}^{t}\frac{(t_2-1)!(t_1-1)!}{(t-1)!} + O\big(\nu^{-1}\big)\right)$$
$$+ \frac{1}{h_{n-1}}\sum_{t_1=2}^{t}\sum_{k=2}^{\nu-1}\frac{(\nu-k)_{t_2}}{\nu-k}V_{k,t_1}$$
$$= O\big(\nu^{t-1}\log^{-1}\nu\big) + \frac{1}{h_{n-1}}\sum_{t_1=2}^{t}\sum_{k=2}^{\nu-1}\frac{(\nu-k)_{t_2}}{\nu-k}V_{k,t_1},$$

if we select

$$\alpha = \left(h_{t-1} - \sum_{t_1+t_2=t}\frac{(t_1-1)!(t_2-1)!}{(t-1)!}\right)^{-1}.$$

We omit the rest of the proof since it runs just like the final part of the proof of Proposition 2.14. □

2.12. *Counting the subtrees by the number of their leaves.* Since the tree with n leaves has $2n-1$ vertices, there are exactly $2n-1$ subtrees, with the number of leaves ranging, with possible gaps, from 1 to n. Let $X_n(t)$ be the number of subtrees with t leaves; so $X_n(1)=n$, $X_n(n)=1$, and $X_n(t)=0$ for $t>n$. Now, $\sum_{t\ge1}X_n(t)=2n-1$, so $\{u_n(t)\}_{t\ge1} := \{\frac{E[X_n(t)]}{2n-1}\}_{t\ge1}$ is the probability distribution of the number of leaves in the uniformly random subtree, that is, the subtree rooted at the uniformly random vertex of the whole tree. Furthermore

$$E\big[X_n(t)\big] = \frac{n}{2h_{n-1}}\sum_{j=1}^{n-1}\frac{E[X_j(t)]+E[X_{n-j}(t)]}{j(n-j)} = \frac{n}{h_{n-1}}\sum_{j=1}^{n-1}\frac{E[X_j(t)]}{j(n-j)}. \tag{71}$$

So, with $\xi_n(t) := \frac{E[X_n(t)]}{n}$, and $h_k := \sum_{j=1}^{k}\frac{1}{j}$, we have

$$\xi_n(t) = \frac{1}{h_{n-1}}\sum_{j=t}^{n-1}\frac{\xi_j(t)}{n-j}, \quad n\ge t+1, \left(\xi_t(t)=\frac{1}{t}\right), \tag{72}$$

and clearly $u_n(t)=\frac{\xi_n(t)}{2-n^{-1}}$.

THEOREM 2.16. (i) $\xi_n(t)\in[\frac{1}{t^2},\frac{1}{th_t}]$, $\frac{1}{t}\le\sum_{\tau\ge t}\xi_n(\tau)\le\frac{2}{t}$, *the last bound implying that the sequence of distributions* $\{u_n(t)\}_{t\ge1}$ *is tight.*

(ii) *For* $q\in(0,1)$, $F_n(q) := \sum_{t\ge1}q^t\xi_n(t)$ *decreases with* n. *Consequently, the sequence of distributions* $\{u_n(t)\}_{t\ge1}$ *converges to a proper distribution* $\{u(t)\}_{t\ge1}$.

(iii) *However, the expected size of the uniformly random subtree is asymptotic to* $\frac{3}{2\pi^2}\log^2 n$.

We conjecture that (ii) can be improved to the stronger assertion that $\xi_n(t)$ is decreasing with n, for each t. We are grateful to Huseyin Acan [1] for numerically verifying this for n and t below 1000.

PROOF. (i) Let us show that $\xi_n(t)\ge\frac{1}{t^2}$ for $n\ge t>1$. By (71), we have $\xi_t(t)=\frac{1}{t}$ and $\xi_{t+1}(t)=\frac{1}{th_t}$, both above $\frac{1}{t^2}$. Suppose that $n\ge t+1$ is such that $\xi_j(t)\ge\frac{1}{t^2}$ for all $j\in[t,n]$.

This is true for $n = t + 1$. For $n > t + 1$,

$$\xi_n(t) \geq \frac{\xi_t(t)}{h_{n-1}(n-t)} + \frac{1}{t^2 h_{n-1}} \sum_{j=t+1}^{n-1} \frac{a}{n-j} = \frac{1}{h_{n-1}(n-t)t} + \frac{h_{n-1-t}}{t^2 h_{n-1}}$$

$$= \frac{1}{t^2} + \frac{1}{h_{n-1}(n-t)t} + \frac{h_{n-1-t} - h_{n-1}}{t^2 h_{n-1}}$$

$$\geq \frac{1}{t^2} + \frac{1}{h_{n-1}(n-t)t} - \frac{1}{t^2 h_{n-1}} \cdot \frac{t}{n-t} = \frac{1}{t^2},$$

which completes the induction step. The proof of $\xi_n(t) \leq \frac{1}{th_t}$ is similarly reduced to showing that $\frac{(n-1)h_t}{(n-t)th_{n-1}} \leq 1$ for $n > t + 1$. This is so, as the fraction is at most $\frac{h_t}{h_{t+1}} \cdot \frac{t+1}{2t}$.

Let us prove that $\frac{1}{t} \leq \sum_{\tau \geq t} \xi_n(\tau) \leq \frac{2}{t}$. Introduce $Y_n(t) = \sum_{\tau \geq t} X_n(\tau)$, the total number of subtrees with at least t leaves, and $\eta_n(t) := \frac{\mathbb{E}[Y_n(t)]}{n} = \sum_{\tau \geq t} \xi_n(\tau)$; so $\eta_n(1) = \frac{2n-1}{n}$, and $\eta_n(n) = \frac{1}{n}$. Analogously to (71), we have

$$\eta_n(t) = \frac{1}{h_{n-1}} \sum_{j=t}^{n-1} \frac{\eta_j(t)}{n-j}, \quad n \geq t+1.$$

We need to show that $\eta_n(t) \leq \frac{2}{t}$ for all $n \geq t$. It suffices to consider $n > t > 1$. Suppose that for some $n \geq t$ and all $j \in [t, n]$ we have $\eta_j(t) \leq \frac{2}{t}$. This is definitely true for $n = t$. Then

$$\eta_{n+1}(t) = \frac{1}{h_n} \sum_{j=t}^{n} \frac{\eta_j(t)}{n+1-j} \leq \frac{2}{th_n} \sum_{j=t}^{n} \frac{1}{n+1-j} = \frac{2h_{n+1-t}}{th_n} \leq \frac{2}{t},$$

which competes the inductive proof of $\eta_n(t) \leq \frac{2}{t}$.

(ii) For $n \geq 2$, we have

$$F_n(q) = \sum_{t \geq 1} q^t \xi_n(t) = \sum_{t \geq 1} \frac{q^t}{h_{n-1}} \sum_{j=t}^{n-1} \frac{\xi_j(t)}{n-j}$$

$$= \frac{1}{h_{n-1}} \sum_{j=1}^{n-1} \frac{1}{n-j} \sum_{t=1}^{j} q^t \xi_j(t) = \frac{1}{h_{n-1}} \sum_{j=1}^{n-1} \frac{F_j(q)}{n-j}.$$

Therefore,

$$F_{n+1}(q) = \frac{1}{h_n} \sum_{j=1}^{n} \frac{F_j(q)}{n+1-j} = \frac{1}{h_n} \left(\frac{F_1(q)}{n} + \sum_{j=2}^{n} \frac{F_j(q)}{n+1-j} \right)$$

$$\leq \frac{1}{h_n} \left(\frac{F_1(q)}{n} + \sum_{j=1}^{n-1} \frac{F_j(q)}{n-j} \right) \leq \frac{1}{h_n} \left(\frac{q}{n} + h_{n-1} F_n(q) \right) \leq F_n(q),$$

since $F_n(q) \geq q\xi_n(t) = q$, and $h_n - h_{n-1} = \frac{1}{n}$. Therefore, for each $q \in (0, 1)$ there exists $F(q) = \lim_{n \to \infty} \sum_{t \geq 1} q^t \xi_n(t)$, implying existence of $\lim_{n \to \infty} \sum_{t \geq 1} q^t u_n(t) = 2F(q)$. For any weakly convergent subsequence of the distributions $\{u_{n_i}(t)\}_{t \geq 1}$, for each x in the unit disc there exists $\lim_{n_i \to \infty} \sum_{t \geq 1} x^t u_{n_i}(t)$, dependent on the subsequence, which is analytic within the disc. All these limits coincide for $x \in (0, 1)$, whence they coincide for all x within the disc, whence on the whole disc. Since the characteristic function determines the distribution uniquely, we see that the whole sequence of the distributions converges to a proper distribution.

(iii) $Z_n := \sum_{t\ge 1} t X_n(t)$ is the total number of the leaves, each leaf counted as many times as the number of the subtrees rooted at the vertices along the path from the root to the leaf, which is distributed as 1 plus L_n, the edge-length of the path to the random leaf. Therefore, $\frac{\mathbb{E}[Z_n]}{2n-1} = \frac{n}{2n-1}(1+\mathbb{E}[L_n])$, and it remains to use Proposition 2.9. □

3. Other methods. Our results here demonstrate that the analysis of recursions method is very effective at deriving sharp asymptotics for the questions addressed here. However, there are many other aspects of the model that could be studied, and a wide variety of familiar general modern probabilistic techniques that could be applied. We indicate such possibilities briefly below—see the preprint [4] for a more comprehensive account.

It is intuitively clear that for our tree model (call it $\mathcal{T}_n$, say) there should be two $n \to \infty$ limit structures:

(a) A scaling limit process, which is a fragmentation of the unit interval via some sigma-finite splitting measure.

(b) A fringe process, which is the local weak limit relative to a random leaf, describable as some marked branching process. This starts with an explicit description of the limit distribution $\{u(t)\}_{t\ge 1}$ in Theorem 2.16.

Less intuitive is a piece of structure theory. In our discrete-time model there is no simple connection between $\mathcal{T}_n$ and $\mathcal{T}_{n+1}$. In the continuous-time model with our chosen rates h_{n-1} (and this is the reason for that particular choice), [4] shows there is a nonobvious consistency property under a "delete a random leaf and prune" operation. This enables an inductive construction of a process $(\mathcal{T}_n, n = 2, 3, 4, \ldots)$, which in turn suggests the possibility of a.s. limit theorems.

Readers may notice that the issues above are somewhat analogous to those arising in the theory surrounding the Brownian continuum random tree (CRT) as a limit of certain other random tree models [2, 9]. However, in contrast to the CRT setting, the two limit processes above would not capture the asymptotics of the quantities studied in this article.

Acknowledgments We thank Svante Janson for catching several errors in earlier versions. We are grateful to Huseyin Acan for the numerics below Theorem 2.16. We thank the referees for the time and effort to repeatedly read the paper and to provide expert critical advice. We thank the Editors for the efficient reviewing process.

REFERENCES

[1] ACAN, H. Personal communication, 05/10/2023.

[2] ALDOUS, D. (1991). The continuum random tree. II. An overview. In *Stochastic Analysis* (*Durham*, 1990). *London Mathematical Society Lecture Note Series* **167** 23–70. Cambridge Univ. Press, Cambridge. MR1166406 https://doi.org/10.1017/CBO9780511662980.003

[3] ALDOUS, D. (1996). Probability distributions on cladograms. In *Random Discrete Structures* (*Minneapolis*, *MN*, 1993). *IMA Vol. Math. Appl.* **76** 1–18. Springer, New York. MR1395604 https://doi.org/10.1007/978-1-4612-0719-1_1

[4] ALDOUS, D. and JANSON, S. (2023). The critical beta-splitting random tree model II: Overview and open problems. Available at arXiv:2303.02529.

[5] ALDOUS, D. J. (2001). Stochastic models and descriptive statistics for phylogenetic trees, from Yule to today. *Statist. Sci.* **16** 23–34. MR1838600 https://doi.org/10.1214/ss/998929474

[6] CURTISS, J. H. (1942). A note on the theory of moment generating functions. *Ann. Math. Stat.* **13** 430–433. MR0007577 https://doi.org/10.1214/aoms/1177731541

[7] DRMOTA, M. (2009). *Random Trees*: *An Interplay Between Combinatorics and Probability*. Springer, Vienna. MR2484382 https://doi.org/10.1007/978-3-211-75357-6

[8] DYER, M., FRIEZE, A. and PITTEL, B. (1993). The average performance of the greedy matching algorithm. *Ann. Appl. Probab.* **3** 526–552. MR1221164

[9] EVANS, S. N. (2008). *Probability and Real Trees*. *Lecture Notes in Math*. **1920**. Springer, Berlin. Lectures from the 35th Summer School on Probability Theory held in Saint-Flour, July 6–23, 2005. MR2351587 https://doi.org/10.1007/978-3-540-74798-7

[10] GRAHAM, R. L., KNUTH, D. E. and PATASHNIK, O. (1994). *Concrete Mathematics*: *A Foundation for Computer Science*, 2nd ed. Addison-Wesley Company, Reading, MA. MR1397498

[11] IKSANOV, A. (2024). Another proof of CLT for critical beta-splitting tree. Unpublished.

[12] JANSON, S. (2012). Simply generated trees, conditioned Galton-Watson trees, random allocations and condensation: Extended abstract. In 23*rd Intern*. *Meeting on Probabilistic*, *Combinatorial*, *and Asymptotic Methods for the Analysis of Algorithms* (*AofA*'12). *Discrete Math*. *Theor*. *Comput*. *Sci*. *Proc*., *AQ* 479–490. Assoc. Discrete Math. Theor. Comput. Sci., Nancy. MR2957350

[13] KOLESNIK, B. (2024). Critical beta-splitting, via contraction. Available at arXiv:2404.16021.

[14] LAMBERT, A. (2017). Probabilistic models for the (sub)tree(s) of life. *Braz*. *J*. *Probab*. *Stat*. **31** 415–475. MR3693976 https://doi.org/10.1214/16-BJPS320

[15] MAHMOUD, H. M. and PITTEL, B. (1989). Analysis of the space of search trees under the random insertion algorithm. *J*. *Algorithms* **10** 52–75. MR0987097 https://doi.org/10.1016/0196-6774(89)90023-0

[16] MUKHERJEA, A., RAO, M. and SUEN, S. (2006). A note on moment generating functions. *Statist*. *Probab*. *Lett*. **76** 1185–1189. MR2270543 https://doi.org/10.1016/j.spl.2005.12.026

[17] PITTEL, B. (1987). An urn model for cannibal behavior. *J*. *Appl*. *Probab*. **24** 522–526. MR0889816 https://doi.org/10.1017/s0021900200031156

[18] PITTEL, B. (1990). On tree census and the giant component in sparse random graphs. *Random Structures Algorithms* **1** 311–342. MR1099795 https://doi.org/10.1002/rsa.3240010306

[19] PITTEL, B. (1999). Normal convergence problem? Two moments and a recurrence may be the clues. *Ann*. *Appl*. *Probab*. **9** 1260–1302. MR1728562 https://doi.org/10.1214/aoap/1029962872

[20] PITTEL, B. and POOLE, D. (2016). Asymptotic distribution of the numbers of vertices and arcs of the giant strong component in sparse random digraphs. *Random Structures Algorithms* **49** 3–64. MR3521273 https://doi.org/10.1002/rsa.20622

[21] YAKYMIV, A. L. (2011). A generalization of the Curtiss theorem for moment generating functions. *Mat*. *Zametki* **90** 947–952. MR2962966 https://doi.org/10.1134/S0001434611110290