The Annals of Applied Probability
2012, Vol. 22, No. 5, 1745–1777
DOI: 10.1214/11-AAP812

THE TOTAL PATH LENGTH OF SPLIT TREES

BY NICOLAS BROUTIN AND CECILIA HOLMGREN

Inria, and Inria and Cambridge University

We consider the model of random trees introduced by Devroye [*SIAM J. Comput.* **28** (1999) 409–432]. The model encompasses many important randomized algorithms and data structures. The pieces of data (items) are stored in a randomized fashion in the nodes of a tree. The total path length (sum of depths of the items) is a natural measure of the efficiency of the algorithm/data structure. Using renewal theory, we prove convergence in distribution of the total path length toward a distribution characterized uniquely by a fixed point equation. Our result covers, using a unified approach, many data structures such as binary search trees, m-ary search trees, quad trees, median-of-$(2k+1)$ trees, and simplex trees.

1. Introduction. In this paper we investigate the total path length, that is, sum of all depths, of random split trees defined by Devroye [13] (we will be more precise shortly). Split trees model a large class of efficient data structures or sorting algorithms. Some important examples of split trees are binary search trees (which are also the representation of Quicksort) [24], m-ary search trees [47], quad trees [19], median-of-$(2k+1)$ trees [4], simplex trees; all these are covered by the results in this document. The case of tries [21] and digital search trees [12] is also important in practice [54]; however, their treatment necessitates different tools, and we leave this case for later studies.

The magnitude of the depths in tree data structures naturally influences their efficiency; in the case where the tree represents the branching choices made by an algorithm, the depths are related to the running time of the algorithm. In this sense, the sum of the depths is a natural and important measure of the efficiency of tree data structures or sorting algorithms.

The path length of tree data structures has been studied by many authors, but in most cases the analyses and proofs are very much tied to a specific case. The main result of this study is to prove that for a large class of split trees, the total path length converges in distribution to a random variable characterized by some fixed point equation. In that sense our result extends the earlier studies of Rösler [50, 52] and Neininger and Rüschendorf [45] who used the so-called contraction method to show convergence in distribution of the total path length for the specific examples of the binary search trees, the median-of-$(2k+1)$ trees and quad trees. Our method actually relies on previous work of Neininger and Rüschendorf [45] who gave a

Received February 2011; revised September 2011.
MSC2010 subject classifications. Primary 05C05, 60C05; secondary 68P05.
Key words and phrases. Random tree, path length, data structure, limit distribution.

limit theorem for the path length of general split trees, under the assumption that the mean satisfies some precise asymptotic form, which we prove.

PLAN OF THE PAPER. In Section 2, we introduce the model of split trees of Devroye [13]. We also discuss previous work on the path length and similar topics. This is also the place where we state our main result, Theorem 2.1.

In Section 3, we explain our general approach, which relies heavily on previous work by Neininger and Rüschendorf [45]. These authors stated a general condition for convergence in distribution of the path length, and our contribution is to prove that it indeed holds for a large class of split trees. So Section 3 is included so that the reader has a general view of the argument.

Once we have stated the precise condition in Section 3, we will move on to explaining our approach to proving it in Section 4. Finally, in Section 5 we discuss extensions of our results.

2. Split trees and path length: Notation and background. We introduce the split tree model of Devroye [13]. Consider an infinite rooted b-ary tree (every node has b children). The nodes are identified with the set of finite words on an alphabet with b letters, $\mathcal{U} = \bigcup_{n\ge 0}\{1,\dots,b\}^n$. The root is represented by the empty word $\varnothing$. We write $u \preceq v$ to denote that u is an ancestor of v (as words, u is a prefix of v). In particular, for the empty word $\varnothing$, we have $\varnothing \preceq v$ for any $v \in \mathcal{U}$.

A split tree T^n of *cardinality* n is constructed by distributing n items (pieces of data) to the nodes $u \in \mathcal{U}$. To describe the tree, it suffices to define the number of items n_u in the subtree rooted at any node $u \in \mathcal{U}$. The tree T^n is then defined as the smallest relevant tree, that is, the subset of nodes u such that $n_u > 0$ (which is indeed a tree).

In the model, internal nodes all contain $s_0 \ge 0$ items, and external nodes can contain up to s items. The construction then resembles a divide-and-conquer procedure, where the partitioning pattern depends on a random vector of proportions. Let $\mathcal{V} = (V_1,\dots,V_b)$ satisfy $V_i \ge 0$ and $\sum_i V_i = 1$; each node $u \in \mathcal{U}$ receives an independent copy $\mathcal{V}_u$ of the random vector $\mathcal{V}$. In the following, we always assume that $\mathbf{P}(\exists i : V_i = 1) < 1$. We can now describe $(n_u, u \in \mathcal{U})$. The tree contains n items, and we naturally have $n_\varnothing$. The split procedure is then carried on from parent to children as long as $n_v > s$. Given the cardinality n_v and the split vector $\mathcal{V}_v = (V_1, V_2,\dots,V_b)$ of v, the cardinalities $(n_{v_1}, n_{v_2},\dots,n_{v_b})$ of the b subtrees rooted at $v_1, v_2,\dots,v_b$ are distributed as

$$\mathrm{Mult}(n_v - s_0 - bs_1, V_1, V_2,\dots,V_b) + (s_1, s_1,\dots,s_1), \tag{1}$$

where $0 \le s$ and $0 \le bs_1 \le s + 1 - s_0$.

Depending on the choice of parameters s_0, s_1, s and the distribution of $\mathcal{V} = (V_1,\dots,V_b)$, many important data structures may be modeled, such as binary search trees, m-ary search trees, median-of-$(2k+1)$ trees, quad trees, simplex trees (see [13]). To make sure that the model is clear and to give a hint of the wide applicability of the model, we illustrate the construction with two canonical examples.

EXAMPLE 1 (Binary search tree). The binary search tree is one of the most common data structures for sorted data. Here we assume that the data set is $\{1, \dots, n\}$. A first (uniformly) random key is drawn σ_1, and stored at the root of a binary tree. The remaining keys are then divided into two subgroups, depending on whether they are smaller or larger than σ_1. The left and right subtrees are then binary search trees built from the two subgroups $\{i : i < \sigma_1\}$ and $\{i : i > \sigma_1\}$, respectively. The sizes of the two subtrees of the root are $\sigma_1 - 1$ and $n - \sigma_1$. One easily verifies that, since σ_1 is uniform in $\{1, 2, \dots, n\}$, one has

$$(\sigma_1 - 1, n - \sigma_1) \stackrel{d}{=} \mathrm{Mult}(n - 1; U, 1 - U),$$

where U is a uniform $U(0, 1)$ random variable. Thus, a binary search tree can be described as a split tree with parameters $b = 2$, $s_0 = 1$, $s = 1$, $s_1 = 0$ and $\mathcal{V}$ is distributed as $(U, 1 - U)$ for U a random variable uniform on $[0, 1]$.

EXAMPLE 2 (Digital trees or tries). We are given n (infinite) strings $X_1, \dots, X_n$ on the alphabet $\{1, \dots, b\}$. The strings are drawn independently, and the symbols of every string are also independent with distribution on $\{1, \dots, b\}$ given by $p_1, \dots, p_b$. Each string naturally corresponds to an infinite path in the infinite complete b-ary tree, where the sequence of symbols indicates the sequence of directions to take as one walks away from the root. The trie is then defined as the smallest tree so that all the paths corresponding to the infinite strings are eventually distinguished; that is, for every string X_i, there exists a node u in the tree such that X_i is the only string with $u \preceq X_i$. The internal nodes store no data; each leaf stores a unique string. In this case, n_v is the number of strings that have prefix v, and one clearly has for the children of the root

$$(n_1, \dots, n_b) \stackrel{d}{=} \mathrm{Mult}(n; p_1, \dots, p_b).$$

The trie is thus a random split tree with parameters $s = 1, s_0 = s_1 = 0$ and $\mathcal{V} = (p_1, p_2, \dots, p_b)$ almost surely.

AN ALGORITHMIC POINT OF VIEW. Rather than using the divide-and-conquer description above, the random trees may be equivalently defined using incremental insertion of data items into an initially empty data structure. The items are labeled using $\{1, 2, \dots, n\}$ in the order of insertion. Initially, $n_u = 0$ for every $u \in U$. We first sample the i.i.d. copies of $\mathcal{V}$ that are assigned to the nodes $u \in \mathcal{U}$.

- Upon insertion, an item first trickles down along a random path from the root until it finds a leaf (i.e., a node u such that all its children $u_1, \dots, u_b$ satisfy $n_{u_i} = 0$). If the path currently corresponds to a word $v \in \mathcal{U}$, and v is not a leaf, then it is extended to v_i, the ith child of v with probability V_i, where $(V_1, \dots, V_b)$ is the copy of $\mathcal{V}$ associated with v.

- When the first phase is finished, the item is stored in a leaf, say v. The leaves can contain up to s items. So if $n_v < s$ (before the insertion), then the item is stored at v, and all the n_u for $u \preceq v$ are updated.
- If $n_v = s$, there is no space for the new item at v. With the new item, we formally have $n_v = s + 1$. In this case, s_0 of these $s + 1$ items are randomly chosen to remain at v while the other $s + 1 - s_0$ are distributed among the children $v_1, \dots, v_b$ of v. Each child receives s_1 items chosen at random. The remaining $s + 1 - s_0 - bs_1$ each choose (independently) a child v_i at random with probability V_i, where $(V_1, \dots, V_b)$ is the copy of $\mathcal{V}$ at node v. If $s_1 = s_0 = 0$, it may happen that all $s + 1$ items now lie at one child v_i, in which case the scheme is repeated until a stable position is found. [This happens with probability 1, since $\mathbf{P}(\exists i : V_i = 1) < 1$.] This last step is the reason why an item may move down when a further item is inserted.

The properties of the multinomial distribution ensure that the tree T^n obtained in this way has the correct distribution (see [13] for details).

In the present case we can assume without loss of generality that the components of $\mathcal{V}$ are identically distributed; applying a random permutation to the components would leave the path length unchanged. We now let V denote a uniformly random component of $\mathcal{V}$. So, for instance, $\mathbf{E}[V] = 1/b$ and $\mathbf{P}(V = 1) < 1/b$ by our assumption that $\mathbf{P}(\exists i : V_i = 1) < 1$.

BACKGROUND AND PREVIOUS WORK. The labeling of the items induced by the algorithm above is interesting for the analysis. Let D_i be the depth of the item labeled i when all n items have been inserted. Then, the total path length is

$$\Psi(T^n) = \sum_{i=1}^{n} D_i.$$

The analysis of the depth D_n of the last item n is thus tightly related to the analysis of $\Psi(T_n)$, and yet is much simpler since it avoids the intricate dependence between the D_i. Devroye [13] proved a weak law of large numbers and a central limit theorem for D_n in general split trees. Let Δ be a component of $(V_1, \dots, V_b)$ picked with probability proportional to its size; that is, given $(V_1, \dots, V_b)$, let $\Delta = V_j$ with probability V_j. We write

$$\begin{aligned} \mu &:= \mathbf{E}[-\ln \Delta] = b\mathbf{E}[-V \ln V] \quad \text{and} \\ \sigma^2 &:= \mathbf{Var}(\ln \Delta) = b\mathbf{E}[V \ln^2 V] - \mu^2. \end{aligned} \tag{2}$$

Note that $\mu \in (0, \infty)$ and $\sigma < \infty$. Then $D_n / \ln n$ converges in probability to μ^{-1}, and $\mathbf{E}[D_n]/\ln n \to \mu^{-1}$ (Devroye assumed that $\mathbf{P}(V = 1) = 0$, but this assumption can be relaxed as long as V satisfies $\mathbf{P}(V = 1) < 1/b$; this is done using trees in which edges are weighted by geometric random variables (see, e.g., [6, 7])). If we also have $\sigma > 0$, then

$$\frac{D_n - \mu^{-1}\ln n}{\sqrt{\sigma^2\mu^{-3}\ln n}} \to \mathcal{N}(0,1)$$

in distribution where $\mathcal{N}(0,1)$ denotes the standard Normal distribution. Note that $\sigma > 0$ precisely when V is not monoatomic, that is, if $bV \neq 1$ with positive probability.

The total path length $\Psi(T^n)$ itself has been extensively studied for specific cases of split trees. The first moment follows from that of D_n since

$$\mathbf{E}[\Psi(T^n)] = \sum_{i=1}^{n}\mathbf{E}[D_i].$$

For instance, in the binary search tree, we have [23]

$$\mathbf{E}[\Psi^{\mathrm{BST}}(T^n)] = 2n\ln n + n(2\gamma - 4) + 2\ln n + 2\gamma + 1 + \mathcal{O}(n^{-1}), \tag{3}$$

where γ is Euler's constant. For higher moments and the distribution of $\Psi(T^n)$, one needs to carefully take the dependence in the terms of the sum into account. Most studies of this type concern the model of binary search tree, or equivalently the cost of quicksort (e.g., [18, 49, 50, 55]). Let

$$Y_n := \frac{\Psi^{\mathrm{BST}}(T^n) - \mathbf{E}[\Psi^{\mathrm{BST}}(T^n)]}{n}. \tag{4}$$

Using martingale arguments, Régnier [49] showed that Y_n converges in distribution to a random variable Y. Rösler [50] showed that Y is satisfying the following distributional equality:

$$Y \stackrel{d}{=} UY + (1-U)Y^* + C(U), \tag{5}$$

where $C(u) := 2u\ln u + 2(1-u)\ln(1-u) + 1$, U is uniform on $[0,1]$, Y and $Y^* \stackrel{d}{=} Y$ are independent. He also proved that the stochastic equality in (5) actually characterizes the distribution of Y: there exists a unique solution Y of (5) such that $\mathbf{E}[Y] = 0$ and $\mathbf{Var}(Y) < \infty$. The distribution of Y is usually called the quicksort distribution. Properties of Y and the rate of convergence of Y_n to Y are studied in [17, 18, 50, 55].

The aim of the present study is to prove that the path length exhibits a similar asymptotic behavior regardless of the precise model of split tree:

THEOREM 2.1. *Let $\Psi(T^n)$ be the total path length in a general split tree with split vector $\mathcal{V} = (V_1, \dots, V_b)$. Suppose that $\mathbf{P}(\exists i : V_i = 1) < 1$. Let*

$$X_n := \frac{\Psi(T^n) - \mathbf{E}[\Psi(T^n)]}{n} \quad \textit{and} \quad C(\mathcal{V}) = 1 + \frac{1}{\mu}\sum_{i=1}^{b} V_i \ln V_i.$$

If $C(\mathcal{V}) \neq 0$ with positive probability, then $X_n \to X$ in distribution, where X is the unique solution of the fixed point equation

$$X \stackrel{d}{=} \sum_{k=1}^{b} V_k X^{(k)} + C(\mathcal{V}),$$

satisfying $\mathbf{E}[X] = 0$ and $\mathbf{Var}(X) < \infty$. Furthermore, exponential moments of X_n exist and converge $\mathbf{E}[e^{\lambda X_n}] \to \mathbf{E}[e^{\lambda X}]$ for any $\lambda \in \mathbb{R}$.

As mentioned in the Introduction, Neininger and Rüschendorf [45] proved a version of Theorem 2.1 conditional on the type of asymptotic expansion for $\mathbf{E}\Psi(T^n)$; our contribution is to prove that this expansion indeed holds (Theorem 3.1), which implies the unconditional version stated in Theorem 2.1.

We have recently been informed that, based on a Markov chain representation of Bruhn [9] and coupling arguments, Munsonius [44] has shown a result similar to our Theorem 2.1 in the special case when the distribution of V has a density with respect to Lebesgue measure.

DISCUSSION AND REMARKS ABOUT THE ASSUMPTIONS. (i) When the split vector $\mathcal{V}$ is deterministic, that is, $\mathcal{V}$ is a permutation of some fixed vector $(p_1, \dots, p_b)$, the cost function $C(\mathcal{V}) = 0$. Such a split tree is a digital tree [54]. In some sense, part of Theorem 2.1 still holds, but the limit X is trivial since $X = 0$ almost surely. The renormalization is actually too strong, since the variance in this case should be of order $n \log n$, rather than n^2 [and order n in the special case when $b\mathcal{V} = (1, \dots, 1)$]. The total path length for binary tries has been treated by Jacquet and Régnier [30]. They showed that the variance of $\Psi(T_n)$ is of order $\mathcal{O}(n)$ if $p = q$ and of order $\mathcal{O}(n \log n)$ if $p \neq q$ and that the path length is asymptotically normal. Schachinger [53] showed that, for tries with a general branch factor, the variance of the total path length for general tries is $\mathcal{O}(n \log^2 n)$. See also [34, 35].

(ii) In general, in the case of digital trees [when $C(\mathcal{V}) = 0$], it is expected that under the correct rescaling the limit distribution should be normal. Neininger and Rüschendorf [46] gave a general conditions under which limit distributions are Gaussian. The case of the binary tries is one example when this theorem can be applied as an alternative proof to the method in [30]. In general, to apply the result in [46] one needs to have approximations for the first two moments of the path length. This is the reason why we report the analysis of this case: a lot more work is required to estimate the variance to the correct order.

(iii) It might seem at first that one should have $C(\mathcal{V}) = 0$ when $\ln V$ is lattice (trie case). However, one can easily construct examples with $C(\mathcal{V}) \neq 0$ and $\ln V$ lattice: for instance, take $b = 5$ and $\mathcal{V}$ a random permutation of either $(1/2, 1/8, 1/8, 1/8, 1/8)$ or $(1/2, 1/4, 1/4, 0, 0)$, each with probability $1/2$.

(iv) Note, although it might come as a surprise since our main tool is renewal theory, Theorem 2.1 does not require any condition on arithmetic properties related to the vector $(V_1, \dots, V_b)$. In particular, it holds whether $-\ln V$ is lattice or not. However, the behavior of the average path length does depend on arithmetic properties of $\ln V$; see Theorem 3.1 later for details.

(v) Note that the limit fixed equation only depends on $\mathcal{V}$, so in particular, the limit distribution X does not depend on the parameters s, s_0 or s_1. However, the average $\mathbf{E}[\Psi(T^n)]$ should clearly depend on these parameters, although we do not prove it formally.

(vi) For the sake of simplicity, we cover only trees with bounded degree, which is usually the case for trees representing data structures. The path length of recursive trees, which do not have bounded degree, has been studied by [14, 38].

3. The contraction method for path length. The condition stated by Neininger and Rüschendorf [45] to ensure weak convergence of the path length concerns the asymptotics of the average path length. More precisely, if one has, for some constant ς,

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + \varsigma n + o(n) \tag{6}$$

and $\mathbf{P}(C(\mathcal{V}) \neq 0) > 0$, then Theorem 5.1 of [45] ensures that $X_n \to X$ in distribution. The purpose of this section is to explain why these conditions are sufficient to prove Theorem 2.1. In particular, we give the necessary background about the contraction method, and we explain the general approach that has been devised in [45]. This section is included only to put our result in context, and no new result is proved with respect to the contraction method.

Note first that (6) holds in the case of binary search trees (3). Recall that D_i is the depth of the ith item in the construction where items are inserted one after another. It is not difficult to deduce from the results on D_i by Devroye [13] that

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + nq(n)$$

with $q(n) = o(\ln n)$ (see Theorem 2.3 of [26] for a formal proof). So proving (6) reduces to proving that $q(n) \to \varsigma$ as $n \to \infty$. Our contribution is to prove that this is indeed the case as soon as the random variable V is such that $-\ln V$ is not lattice, that is, there is no $a \in \mathbb{R}$ such that $-\ln V \in a\mathbb{Z}$ almost surely. In the following, we let

$$d := \sup\{a \geq 0 : \mathbf{P}(\ln V \in a\mathbb{Z}) = 1\},$$

so that d is the span of the lattice when $d > 0$ and $\ln V$ is nonlattice when $d = 0$. More precisely, we prove:

THEOREM 3.1. *The expected value of the total path length $\Psi(T^n)$ exhibits the following asymptotics, as $n \to \infty$*:

$$\mathbf{E}[\Psi(T^n)] = \mu^{-1} n \ln n + n\varpi(\ln n) + o(n), \tag{7}$$

where μ is the constant in (2) and ϖ is a continuous periodic function of period d. In particular, if $\ln V$ is not lattice, then $d = 0$ and ϖ is constant.

If $\ln V$ is nonlattice, then Theorem 3.1 and Theorem 5.1 of [45] together prove Theorem 2.1. If the random variable $\ln V$ is lattice with span d, then Theorem 3.1 implies that $q(n) = \varpi(\ln n) + o(1)$ as $n \to \infty$, where ϖ is d-periodic. So it seems that Theorem 3.1 does not permit to conclude along the arguments by Neininger and Rüschendorf [45]. However, the techniques in [45] only require convergence of the coefficients of a certain recursive equation; this fact was used in [46] to deal with certain cases involving oscillations.

We now move on to the approach developed by Neininger and Rüschendorf [45, 46]. Let $\overline{n} = (n_1, \dots, n_b)$ denote the vector of cardinalities of the children of the root. Then we have, for $n > s$,

$$\Psi(T^n) \stackrel{d}{=} \sum_{i=1}^{b} \Psi_i(T^{n_i}) + n - s_0,$$

where $\Psi_i(T^{n_i})$ are copies of $\Psi(T^{n_i})$ that are independent conditional on $(n_1, \dots, n_b)$. Introducing the normalized total path length

$$X_n := \frac{\Psi(T^n) - \mathbf{E}[\Psi(T^n)]}{n}, \tag{8}$$

we can rewrite the distributional identity above as

$$X_n := \sum_{i=1}^{b} \frac{n_i}{n} X_{n_i} + C_n(\overline{n}),$$

where

$$C_n(\overline{n}) := 1 - \frac{s_0}{n} - \frac{\mathbf{E}[\Psi(T^n)]}{n} + \sum_{i=1}^{b} \frac{\mathbf{E}[\Psi(T^{n_i})]}{n}$$

and X_{n_i}, $i \in \{1, \dots, b\}$, are independent conditional on $(n_1, \dots, n_b)$. By definition, the vector of cardinalities $\overline{n}$ is $\mathrm{Mult}(n - s_0 - bs_1, V_1, V_2, \dots, V_b) + (s_1, s_1, \dots, s_1)$ so that

$$\left(\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_b}{n}\right) \to \mathcal{V}_\sigma = (V_1, V_2, \dots, V_b), \tag{9}$$

almost surely as $n \to \infty$. This is where (6) comes into play: it ensures that the cost $C_n(\overline{n})$ (the "toll function") in the recursive distributional equation does converge (in distribution) as $n \to \infty$. Indeed

$$\begin{aligned} C_n(\overline{n}) &= 1 + \frac{1}{n} \sum_{i=1}^{b} \mathbf{E}[\Psi(T^{n_i})] - \frac{\mathbf{E}[\Psi(T^n)]}{n} - \frac{s_0}{n} \\ &= 1 + \frac{1}{\mu} \sum_{i=1}^{b} \frac{n_i}{n} \ln \frac{n_i}{n} + \frac{1}{\mu} \left(\sum_{i=1}^{b} \frac{n_i}{n} \varpi(\ln n_i) - \varpi(\ln n) \right) + o(1). \end{aligned}$$

Now, by (9) and the continuity of ϖ, it follows that

$$\begin{aligned}C_n(\overline{n}) &= 1 + \frac{1}{\mu}\sum_{i=1}^{b}\frac{n_i}{n}\ln\frac{n_i}{n} + \frac{1}{\mu}\left(\sum_{i=1}^{b}\frac{n_i}{n}\varpi(\ln n + \ln V_i) - \varpi(\ln n)\right) + o(1)\\ &= 1 + \frac{1}{\mu}\sum_{i=1}^{b} V_i \ln V_i + o(1),\end{aligned} \tag{10}$$

since ϖ is d-periodic and $\ln V_i \in d\mathbb{Z}$ by assumption (if $d = 0$, φ is constant and the claim also holds). Note that, apart from (9), only asymptotics for the first moments are required for (10) to hold. Together (9) and (10) suggest that if X_n converges in distribution to some limit X, then X should satisfy the following fixed point equation:

$$X \stackrel{d}{=} \sum_{k=1}^{b} V_k X^{(k)} + C(\mathcal{V}) \qquad \text{where } C(\mathcal{V}) = 1 + \frac{1}{\mu}\sum_{i=1}^{b} V_i \ln V_i, \tag{11}$$

and $X^{(k)}$ are independent and identically distributed copies of X.

The point of the contraction method is to make the previous arguments rigorous, that is, to show that if the coefficients $C_n(\overline{n})$ do converge, then (11) has a unique solution X and that $X_n \to X$ in distribution; this is precisely what was done in [45, 46]. This is done by proving that the recursive map defined by (11) is a contraction in a suitable space of probability measures [48, 50, 51]. We now expose the lines of the arguments to show the extent of the results that follow from the mere convergence of the coefficients $C_n(\overline{n})$. (We claim no novelty.)

Let $\mathscr{M}_2$ be the set of probability measures with a finite second moment. For a random variable X, we write $\mathcal{D}(X)$ for its law. For $\phi \in \mathscr{M}_2$ and X a random variable with law $\mathcal{D}(X) = \phi$, define the L^2-norm by $\|X\|_2 = \mathbf{E}[X^2]^{1/2}$. We can then define a metric d_2 on $\mathscr{M}_2$ (the Mallow metric): for $\phi, \varphi \in \mathscr{M}_2$, let

$$d_2(\phi, \varphi) := \inf\|X - Y\|_2, \tag{12}$$

where the range of the infimum is the set of couples (X, Y) with marginal distributions $\mathcal{D}(X) = \phi$ and $\mathcal{D}(Y) = \varphi$. For simplicity we write $d_2(X, Y) = d_2(\phi, \varphi)$ for random variables X and Y, but note that this only depends on the marginal distributions ϕ and φ. Convergence of ϕ_n to ϕ in $(\mathscr{M}_2, d_2)$ is equivalent to weak convergence with convergence of the second moment [48]:

$$\phi_n \xrightarrow{w} \phi \quad \text{and} \quad \int x^2\, d\phi_n(x) \to \int x^2\, d\phi(x). \tag{13}$$

Let $\mathscr{M}_2^0$ be the subset of $\mathscr{M}_2$ containing distributions ϕ such that $\int x\, d\phi(x) = 0$. Define the operator $T : \mathscr{M}_2^0 \to \mathscr{M}_2^0$. For a distribution $\phi \in \mathscr{M}_2^0$, let $T(\phi)$ be the distribution of the random variable given by

$$\sum_{1\le k\le b} V_k Z^{(k)} + C(\mathcal{V}),$$

where $Z^{(i)}$ are i.i.d. random variables with distribution ϕ. Then, calculations similar to that in the proof of Lemma 3.2 in [45] yield

$$\begin{aligned} d_2(T(X), T(Y)) &\le \sum_{1\le i\le b} \mathbf{E}[V_i^2]\cdot d_2(X,Y) \\ &= b\mathbf{E}[V^2]\cdot d_2(X,Y). \end{aligned}$$

Since $b\mathbf{E}[V^2] < 1$ the operator T is a contraction in $(\mathscr{M}_2^0, d_2)$. Thus the Banach fixed point theorem implies that T has a unique fixed point. The random variable X has this fixed point as distribution. The same line of thought actually implies that $d_2(X_n, X) \to 0$. A formal proof can be found in [45]. As stated in (13), the convergence in $(\mathscr{M}_2^0, d_2)$ is strong enough to imply convergence of second moments. In particular,

$$\mathbf{Var}(\Psi(T^n)) \sim \zeta n^2,$$

where $\zeta = \mathbf{Var}(X)$. Computing $\mathbf{E}[X^2]$ using the fixed point equation, one easily obtains the following expression for ζ:

$$\zeta = \mathbf{Var}(X) = \frac{\mu^{-2}\mathbf{E}[(\sum_{i=1}^b V_i \log V_i)^2] - 1}{1 - \sum_{i=1}^b \mathbf{E}[V_i^2]}. \tag{14}$$

This expression may also be obtained using estimates based on renewal theory in the spirit of our proof of Theorem 3.1.

4. Precise asymptotics for the average path length.

4.1. *Plan of the proof of Theorem* 3.1. In the previous section, we have explained why precise asymptotics for $\mathbf{E}[\Psi(T^n)]$ imply convergence in distribution of $\Psi(T^n)$ (suitably rescaled). We now move on to the proof of Theorem 3.1.

Recall that D_i denotes the depth of the ith inserted item. Write $i \in T_u$ if the item i is stored in the subtree rooted at u. Then rearranging the sum in the definition of $\Psi(T^n)$, we see that

$$\Psi(T^n) = \sum_{i=1}^n D_i = \sum_{i=1}^n \sum_{u\ne\sigma} \mathbf{1}_{\{i\in T_u\}} = \sum_{u\ne\sigma} n_u. \tag{15}$$

Recall the following fact, which we used already in Section 3:

$$\frac{1}{n}\mathrm{Mult}(n; V_1, \dots, V_b) \to (V_1, \dots, V_b),$$

almost surely, as $n \to \infty$. We actually have a similar behavior for any random variable n_v, when v is a fixed node (so in particular, its depth does not depend on n). For a node u, the components $V_1, V_2, \dots, V_b$ of $\mathcal{V}_u$ are naturally associated to the children $u_1, u_2, \ldots, u_b$ of u, and we can define $V_{u_i} = V_i$. For the root node $\varnothing$, define $V_\varnothing = 1$. Then let

$$L_u = \prod_{v \preceq u} V_v, \tag{16}$$

where $v \preceq u$ if v is an ancestor of u. The random variables $(L_u, u \in \mathcal{U})$ define a recursive partition of $[0, 1]$, where L_u is the *length* of the interval associated with u. In general, for any fixed node u, we have

$$\frac{n_v}{n} \to L_v,$$

almost surely as $n \to \infty$. So, as long as n_v is large it should be well approximated by nL_v. This suggests that the sum in (15) be decomposed into the contributions of the top and of the fringe of the tree. We define the separation in terms of a parameter B measuring the size of the trees pending in the fringe. The lengths L_v are decreasing on any path from the root. So let R be the collection of nodes such that $r \in R$ if r has $nL_r < B$ but for all its strict ancestors v we have $nL_v \geq B$. We write $T_r, r \in R$, for the subtrees rooted at the nodes that belong to R.

Then

$$\mathbf{E}[\Psi(T^n)] = \mathbf{E}\biggl[\sum_{v \neq \varnothing} n_v \mathbf{1}_{\{nL_v \geq B\}}\biggr] + \mathbf{E}\biggl[\sum_{r \in R} \Psi(T^{n_r}) + n_r\biggr], \tag{17}$$

since given n_r, the total path length of T_r, $r \in R$, is distributed like T^{n_r}. [The term n_r needs to be added since the cardinality of the root of a tree T is not taken into account from our definition of $\Psi(T)$.] The following two propositions gather the asymptotics for the two terms in (17) above that will enable us to prove Theorem 3.1. In the following, we let

$$d = \sup\{a \geq 0 : \mathbf{P}(\ln V \in a\mathbb{Z}) = 1\}.$$

Indeed, as we already mentioned (it will become clear soon), the arithmetic properties of $\ln V$ influence the asymptotics.

PROPOSITION 4.1. *There exists a constant K such that, for all n large enough, and all B, we have*

$$\left|\mathbf{E}\biggl[\sum_{v \neq \varnothing} n_v \mathbf{1}_{\{nL_v \geq B\}}\biggr] - \frac{1}{\mu} n \ln\left(\frac{n}{B}\right) - n\phi_1\left(\ln \frac{n}{B}\right)\right| \leq K \frac{n}{B},$$

where μ is the constant in (2) *and ϕ_1 is a continuous d-periodic function; in particular, ϕ_1 is constant when $d = 0$.*

PROPOSITION 4.2. *There exists a constant K such that, for all n large enough, all $\varepsilon > 0$ small enough and $B = \varepsilon^{-8}$, we have*

$$\left|\mathbf{E}\biggl[\sum_{r \in R} \Psi(T^{n_r}) + n_r\biggr] - n\varphi_B\left(\ln \frac{n}{B}\right)\right| \leq K\varepsilon n \tag{18}$$

for some φ_B, a d-periodic function that depends on B. Furthermore, there exists a constant K' (independent of B) such that, for $\varepsilon > 0$ small enough,

$$\sup_{|q-q'|\le\varepsilon^3} |\varphi_B(q) - \varphi_B(q')| \le K'\varepsilon\ln(1/\varepsilon). \tag{19}$$

The proofs of Propositions 4.1 and 4.2 both rely on renewal theory: first, the sum $S_{n,B}$ is easily approximated by a function of sums of i.i.d. random variables; second, the sizes n_r in the second contribution can be estimated using overshoot arguments. The necessary technical lemmas are introduced in the following section. Then, we prove Propositions 4.1 and 4.2 in Sections 4.3 and 4.4, respectively.

Before we proceed to the proofs of Propositions 4.1 and 4.2, we prove that they indeed imply Theorem 3.1. The nonlattice case should be rather clear, but the lattice case requires a little care.

PROOF OF THEOREM 3.1. We have been precise in the statements of Propositions 4.1 and 4.2; we now take the liberty to use $O(\cdot)$ notation to simplify the discussion. It is understood that the hidden constants do not depend on n, ε or B.

(i) First assume that $\ln V$ is nonlattice ($d = 0$). Let $n, \widehat{n}$ be integers such that $n \le \widehat{n}$. Fix $\varepsilon > 0$, and choose $B = \varepsilon^{-20}$. Then by the triangle inequality and Propositions 4.1 and 4.2,

$$\left|\left(\frac{\mathbf{E}[\Psi(T^n)]}{n} - \mu^{-1}\ln n\right) - \left(\frac{\mathbf{E}[\Psi(T^{\widehat{n}})]}{\widehat{n}} - \mu^{-1}\ln\widehat{n}\right)\right| = O(\varepsilon)$$

as $n \to \infty$. Thus, the sequence $(n^{-1}\mathbf{E}[\Psi(T^n)] - \mu^{-1}\ln n, n \ge 0)$ is Cauchy, hence the result.

(ii) If $\ln V$ is lattice, the situation is different since we cannot directly invoke similar arguments. In particular, we need to prove the existence and continuity of the function ϖ. Fix $\beta \in [0, d)$ and consider $\Omega_\beta = \{n \ge 1 : \exists k \in \mathbb{N}, |\ln n - kd + \beta| \le n^{-1}\}$, the set of integers such that $\ln n \bmod d$ is close to β. Then, by the triangle inequality and Propositions 4.1 and 4.2, we have

$$\begin{aligned}
&\left|\left(\frac{\mathbf{E}[\Psi(T^n)]}{n} - \mu^{-1}\ln n\right) - \left(\frac{\mathbf{E}[\Psi(T^{\widehat{n}})]}{\widehat{n}} - \mu^{-1}\ln\widehat{n}\right)\right| \\
&\quad\le \left|\phi_1\left(\ln\frac{n}{B}\right) - \phi_1\left(\ln\frac{\widehat{n}}{B}\right)\right| + \left|\varphi_B\left(\ln\frac{n}{B}\right) - \varphi_B\left(\ln\frac{\widehat{n}}{B}\right)\right| \\
&\qquad + O(\varepsilon) + O(1/B) \\
&\quad= |\phi_1(\ln n) - \phi_1(\ln\widehat{n})| + |\varphi_B(\ln n) - \varphi_B(\ln\widehat{n})| + O(\varepsilon),
\end{aligned}$$

if we choose ε in such a way that $B = \varepsilon^{-20} = \beta \bmod d$. Now, ϕ_1 is continuous and d-periodic so that there exists n_0 (independent of β) such that $|\phi_1(\ln n) -$

$\phi_1(\ln \widehat{n})| \le \varepsilon$ when $n, \widehat{n} \ge n_0$ inside Ω_β. On the other hand, for $n, \widehat{n} \in \Omega_\beta$ such that $n, \widehat{n} \ge 2\varepsilon^{-3}$, we have

$$|\varphi_B(\ln n) - \varphi_B(\ln \widehat{n})| \le K'\varepsilon \ln(1/\varepsilon).$$

Note that the bounds obtained are all uniform in β. It follows that for every $\varepsilon > 0$, there exists $n_1 = \max\{n_0, \varepsilon^{-3}\}$ such that for $n, \widehat{n} \in \Omega_\beta$ satisfying $n, \widehat{n} \ge n_1$, we have

$$\left|\left(\frac{\mathbf{E}[\Psi(T^n)]}{n} - \mu^{-1} \ln n\right) - \left(\frac{\mathbf{E}[\Psi(T^{\widehat{n}})]}{\widehat{n}} - \mu^{-1} \ln \widehat{n}\right)\right| \le O(\varepsilon) + K'\varepsilon \ln(1/\varepsilon).$$

Therefore, the subsequences $(n^{-1}\mathbf{E}[\Psi(T^n)] - \mu^{-1} \ln n, n \in \Omega_\beta)$, $\beta \in [0, d)$, are uniformly Cauchy (in β). It follows that there exists a fixed function ϖ defined on $[0, d)$ such that, for every β and $n \in \Omega_\beta$,

$$\mathbf{E}[\Psi(T^n)] = \frac{1}{\mu} n \ln n + n\varpi(\beta) + o(n).$$

Furthermore, the function ϖ is continuous. This is easily seen using the same arguments with $n \in \Omega_\beta$, $\widehat{n} \in \Omega_{\beta'}$ and $|\beta - \beta| < \varepsilon$. Once the definition of ϖ is extended by periodicity, the continuity ensures that we can write the asymptotics for $\mathbf{E}[\Psi(T^n)]$ in the form claimed in (7). This completes the proof in the lattice case. $\square$

4.2. *The renewal structure of split trees.* Renewal theory has already been used for studying random trees in [26, 28, 32, 42, 43]. The present paper is another example of its wide applicability. We start by quantifying the deviation between n_v and nL_v for fixed nodes $v \in \mathcal{U}$.

LEMMA 4.1. *For any node v, we have for all x large enough*

$$\mathbf{P}\big(|n_v - nL_v| > (nL_v)^{2/3} \mid nL_v > x\big) \le x^{-1/4}.$$

PROOF. First note that by the triangle inequality

$$\begin{aligned} &\mathbf{P}\big(|n_v - nL_v| > (nL_v)^{2/3} \mid nL_v > s\big) \\ &\quad \le \mathbf{P}\big(2|n_v - \mathrm{Bin}(n, L_v)| > (nL_v)^{2/3} \mid nL_v > x\big) \\ &\qquad + \mathbf{P}\big(2|\mathrm{Bin}(n, L_v) - nL_v| > (nL_v)^{2/3} \mid nL_v > x\big). \end{aligned}$$

Suppose that $|v| = d$ and let $\mathscr{G}_d$ be the σ-field generated by the random variables V_u for $|u| \le d$. Conditioning on $\mathscr{G}_d$, the recursive splits of the cardinalities n_v defined in (1) give in a stochastic sense the following bound for n_v:

$$|n_v - \mathrm{Bin}(n, L_v)| \le_{st} \sum_{u \preceq v} \mathrm{Bin}(s, L_v/L_u). \tag{20}$$

Now, by (20), Chebyshev's inequality and Chernoff's bound for binomials (see, e.g., [11, 25, 33]) we obtain

$$\begin{aligned}
&\mathbf{P}\big(|n_v - nL_v| > (nL_v)^{2/3} \mid nL_v > x\big) \\
&\quad \le 2x^{-2/3}\mathbf{E}\biggl[\sum_{u \preceq v} \mathrm{Bin}(s, L_v/L_u)\biggr] \\
&\qquad + \mathbf{E}\biggl[\exp\biggl(\frac{-(nL_v)^{4/3}}{8(nL_v + (nL_v)^{2/3}/6)}\biggr) \Bigm| nL_v > x\biggr] \\
&\quad \le 2sx^{-2/3}\sum_{k\ge 0} b^{-k} + e^{-x^{1/4}} \le x^{-1/4}
\end{aligned}$$

for all x large enough. $\square$

When the cardinalities n_v are close to the product nL_v, renewal theory allows us to get approximations suitable to prove Propositions 4.1 and 4.2. It is convenient to introduce the additive form $S_v = -\ln L_v$. For $|v| = k$,

$$S_v \stackrel{d}{=} S_k = \sum_{i=1}^{|v|} -\ln V_i,$$

where V_i, $i \ge 1$, are i.i.d. copies of V. We define the exponential renewal function

$$U(t) := \sum_{k=1}^{\infty} b^k \mathbf{P}(S_k \le t), \tag{21}$$

which satisfies the following renewal equation with $\nu(t) = b\mathbf{P}(-\ln V \le t)$:

$$U(t) = \nu(t) + (U * d\nu)(t) \qquad \text{where } (U * d\nu)(t) = \int_0^t U(t-z)\,d\nu(z). \tag{22}$$

The measure $d\nu(t)$ is not a probability measure. To work with more convenient renewal equations, involving probability measures, we introduce the tilted measure $d\omega(t) = e^{-t}\,d\nu(t)$. It is easily seen that $d\omega(t)$ is a probability measure, and defines a random variable X by $\mathbf{P}(X \in dt) = d\omega(t)$. In fact ω is the distribution function of $-\ln\Delta$, where Δ is the size-biased random variable in (2): writing I for a random variable that is i with probability V_i given $(V_1, \dots, V_b)$, we have

$$\begin{aligned}
\mathbf{P}(-\ln\Delta \le x) &= \mathbf{E}\mathbf{E}\big[\mathbf{1}_{\{-\ln V_I \le x\}} \mid (V_1, \dots, V_b)\big] \\
&= \mathbf{E}\biggl[\sum_{i=1}^{b} \mathbf{1}_{\{-\ln V_i \le x\}} V_i\biggr] \\
&= b\mathbf{E}\big[\mathbf{1}_{\{-\ln V \le x\}} e^{-\ln V}\big] = \omega(x).
\end{aligned}$$

Then, from (2), X obviously satisfies

$$\mathbf{E}[X] = \mathbf{E}[-\ln\Delta] = \mu \quad \text{and} \quad \mathbf{E}[X^2] = \sigma^2 + \mu^2.$$

The renewal equation (22) can then be rewritten as

$$\widehat{U}(t) = \widehat{\nu}(t) + (\widehat{U} * d\omega)(t), \tag{23}$$

where $\widehat{U}(t) := e^{-t}U(t)$ and $\widehat{\nu}(t) := e^{-t}\nu(t)$. The first-order asymptotics for $U(t)$ as $t \to \infty$ follows from the standard renewal theorem applied to $\widehat{U}(t)$ (see also Theorem 7.1, Chapter V of [1] or Lemma 3.1 of [26] for a formal proof):

$$U(t) = \widehat{U}(t)e^t = \mu^{-1}e^t + o(e^t), \qquad t \to \infty. \tag{24}$$

We will need some information about the second-order behavior of $U(t)$. The following lemma will be sufficient for us.

LEMMA 4.2. *Let $d = \sup\{a \ge 0 : \mathbf{P}(\ln V \in a\mathbb{Z}) = 1\}$, so that $d = 0$ if $\ln V$ is nonlattice. Then, as $x \to \infty$*

$$\begin{aligned} &\int_0^x e^{-t}\big(U(t) - \mu^{-1}e^t\big)\,dt \\ &\quad = \begin{cases} \dfrac{\sigma^2 - \mu^2}{2\mu^2} - \mu^{-1} + o(1), & \text{if } d = 0, \\ \dfrac{\sigma^2 - \mu^2}{2\mu^2} - \mu^{-1} + \phi(x) + o(1), & \text{if } d > 0, \end{cases} \end{aligned} \tag{25}$$

where $\phi(x)$ is a bounded continuous periodic function with period d.

PROOF. Let X_k be i.i.d. copies of a random variable X defined by $\mathbf{P}(X \in dt) = e^{-t}\,d\nu(t)$. Define the (standard) renewal function

$$F(t) := \sum_{n\ge 0} \mathbf{P}\left(\sum_{k=1}^{n} X_k \le t\right). \tag{26}$$

Then the renewal theorem (Theorem V.2.4 of [1]) applied to (23) yields

$$e^{-t}U(t) = \widehat{U}(t) = \int_0^t \widehat{\nu}(t-u)\,dF(u) = \int_0^\infty \widehat{\nu}(u)\,dF(t-u). \tag{27}$$

[Note that $dF(t)$ includes a term $d\mathbf{P}(0 \le t) = \delta_0(t)$.] By Fubini's theorem we obtain

$$\begin{aligned} \int_0^x e^{-t}\big(U(t) - \mu^{-1}e^t\big)\,dt &= \int_0^\infty \widehat{\nu}(u)\int_0^x dF(t-u)\,du - \frac{x}{\mu} \\ &= \int_0^\infty \widehat{\nu}(u)F(x-u)\,du - \frac{x}{\mu}. \end{aligned} \tag{28}$$

Recall that $\widehat{\nu}(x) = \nu(x)e^{-x}$. Integration by parts gives

$$\int_0^\infty \widehat{\nu}(x)\,dx = b[-e^{-t}\mathbf{P}(-\ln V \le t)]_0^\infty + \int_0^\infty e^{-t}\,d\nu(t) = b\mathbf{E}[e^{-\ln V}] = 1. \tag{29}$$

Rewriting (28) as a single integral, it follows that

$$
\begin{aligned}
\int_0^x & e^{-t}\big(U(t)-\mu^{-1}e^t\big)\,dt \\
&= \int_0^\infty \widehat{\nu}(u)\Big(F(x-u)-\frac{x}{\mu}\Big)\,du \\
&= -\frac{1}{\mu}\int_0^x \widehat{\nu}(u)u\,du - \frac{1}{\mu}\int_x^\infty \widehat{\nu}(u)x\,du \\
&\quad + \int_0^x \widehat{\nu}(u)\Big(F(x-u)-\frac{x-u}{\mu}\Big)\,du.
\end{aligned}\tag{30}
$$

We start with the first two terms in (30). Using again integration by parts and applying (29) yields

$$
\begin{aligned}
\int_0^\infty \widehat{\nu}(u)u\,du &= \int_0^\infty e^{-u}\nu(u)u\,du \\
&= \int_0^\infty \widehat{\nu}(u)\,du + \int_0^\infty ue^{-u}\,d\nu(u) \\
&= 1 + b\mathbf{E}[-V\ln V] = 1+\mu,
\end{aligned}\tag{31}
$$

where the last equality follows from the definition of μ in (2). Finally, note that for all x,

$$
\int_x^\infty \widehat{\nu}(u)x\,du \le \int_x^\infty \widehat{\nu}(u)u\,du \to 0 \tag{32}
$$

as $x\to\infty$ since $\int_0^\infty |\widehat{\nu}(u)u|\,du<\infty$.

So it only remains to estimate the third term in (30). This is related to the asymptotics for the renewal function $F(t)$, which are different depending on whether $\ln V$ is lattice or not. Write $\{x\}$ for the fractional part of a real number x, that is, $\{x\}=x-\lfloor x\rfloor$. Then, by Theorem 5.1 in [22] we have, as $t\to\infty$,

$$
F(t)-\frac{t}{\mu}=\frac{\sigma^2+\mu^2}{2\mu^2}+o(1) \quad\text{and}\quad F(t)-\frac{t}{\mu}=\frac{\sigma^2+\mu^2}{2\mu^2}+\frac{d}{\mu}\Big(\frac{1}{2}-\Big\{\frac{t}{d}\Big\}\Big)+o(1)
$$

in the nonlattice and the d-lattice case, respectively. Furthermore, by Lorden's inequality ([37], Theorem 1),

$$
0\le F(t)-\frac{t}{\mu}\le\frac{\sigma^2+\mu^2}{\mu^2}.
$$

(i) We now first assume that $\ln V$ is nonlattice. The dominated convergence theorem applied to the last integral in (30), and (29), yield

$$
\begin{aligned}
\lim_{x\to\infty}\int_0^\infty \widehat{\nu}(u)\Big(F(x-u)-\frac{x-u}{\mu}\Big)\mathbf{1}_{\{u\le x\}}\,du &= \int_0^\infty \widehat{\nu}(u)\frac{\sigma^2+\mu^2}{2\mu^2}\,du \\
&= \frac{\sigma^2+\mu^2}{2\mu^2}.
\end{aligned}\tag{33}
$$

Putting (33) together with (30), (31) and (32) we obtain, as $x \to \infty$,

$$\int_0^x e^{-t}\big(U(t) - \mu^{-1}e^t\big)\,dt = -\frac{1}{\mu} - 1 + \frac{\sigma^2 + \mu^2}{2\mu^2} + o(1),$$

which proves the claim in (25) in the nonlattice case.

(ii) Similarly in the lattice case with span d, from the dominated convergence theorem we obtain

$$\begin{aligned}
&\int_0^x \widehat{\nu}(u)\left(F(x-u) - \frac{x-u}{\mu}\right)du \\
&\quad = \frac{\sigma^2+\mu^2}{2\mu^2} + \frac{d}{\mu}\int_0^x \left(\frac{1}{2} - \left\{\frac{x-u}{d}\right\}\right)\widehat{\nu}(u)\,du + o(1) \\
&\quad = \frac{\sigma^2+\mu^2}{2\mu^2} + \frac{d}{\mu}\int_0^\infty \left(\frac{1}{2} - \left\{\frac{x-u}{d}\right\}\right)\widehat{\nu}(u)\,du + o(1)
\end{aligned} \tag{34}$$

by (32). The function ϕ defined for $x \geq 0$ by

$$\phi(x) = \frac{d}{\mu}\int_0^\infty \left(\frac{1}{2} - \left\{\frac{x-u}{d}\right\}\right)\widehat{\nu}(u)\,du$$

is clearly d-periodic. Furthermore, the function $\phi(\cdot)$ is continuous. Indeed, for any x, y such that $|x-y| < \varepsilon$ we have

$$\begin{aligned}
\phi(y) &= \frac{d}{\mu}\int_0^\infty \left(\frac{1}{2} - \left\{\frac{y-u}{d}\right\}\right)\widehat{\nu}(u)\,du \\
&= \frac{d}{\mu}\int_0^\infty \left(\frac{1}{2} - \left\{\frac{y-u}{d}\right\}\right)\mathbf{1}_{\{y-u \bmod d \in [\varepsilon, 1-\varepsilon]\}}\widehat{\nu}(u)\,du \\
&\quad + \frac{d}{\mu}\int_0^\infty \left(\frac{1}{2} - \left\{\frac{y-u}{d}\right\}\right)\mathbf{1}_{\{y-u \bmod d \notin [\varepsilon, 1-\varepsilon]\}}\widehat{\nu}(u)\,du.
\end{aligned}$$

It follows that

$$\begin{aligned}
|\phi(y) - \phi(x)| &\leq \frac{2}{\mu}\varepsilon + 2\sup_{z\in\{x,y\}}\frac{d}{\mu}\int_0^\infty \left|\frac{1}{2} - \left\{\frac{z-u}{d}\right\}\right|\mathbf{1}_{\{z-u \bmod d \notin [\varepsilon, 1-\varepsilon]\}}\widehat{\nu}(u)\,du \\
&\leq \frac{2}{\mu}\varepsilon + 2\sup_{z\in\{x,y\}}\frac{d}{\mu}\int_0^\infty \mathbf{1}_{\{z-u \bmod d \notin [\varepsilon, 1-\varepsilon]\}}\widehat{\nu}(u)\,du.
\end{aligned}$$

Since $|\widehat{\nu}(u)| = e^{-u}b\mathbf{P}(-\ln V \leq t) \leq b$, the dominated convergence theorem implies that $|\phi(y) - \phi(x)| \to 0$ as $\varepsilon \to 0$.

Finally, putting (34) together with (30), (31) and (32) as before proves the lattice case in (25). $\square$

4.3. *Contribution of the top of the tree*. In this section, we prove Proposition 4.1. For the top of the tree, the sizes n_v are well approximated by $\mathrm{Bin}(n, L_v)$. This suggests that the main contribution of the top of the tree should be

$$\mathbf{E}\Big[\sum_{v\neq\varnothing} n_v \mathbf{1}_{\{nL_v\geq B\}}\Big] = \mathbf{E}\Big[\sum_{v\neq\varnothing} \mathrm{Bin}(n, L_v)\mathbf{1}_{\{nL_v\geq B\}}\Big] + R_{n,B} \tag{35}$$

for a remainder $R_{n,B}$ that should be small. We first estimate the main contribution; we will then quantify $R_{n,B}$ using (20).

LEMMA 4.3. *Let $d = \sup\{a : \mathbf{P}(\ln V \in a\mathbb{Z}) = 1\}$, so that $d = 0$ if $\ln V$ is nonlattice. Then, as $n/B \to \infty$,*

$$\begin{aligned}
&\mathbf{E}\Big[\sum_{v\neq\varnothing} \mathrm{Bin}(n, L_v)\mathbf{1}_{\{nL_v\geq B\}}\Big] \\
&\quad = \begin{cases} \dfrac{1}{\mu} n \ln\Big(\dfrac{n}{B}\Big) + n\dfrac{\sigma^2-\mu^2}{2\mu^2} + o(n), & \textit{if } d = 0,\\[2ex] \dfrac{1}{\mu} n \ln\Big(\dfrac{n}{B}\Big) + n\dfrac{\sigma^2-\mu^2}{2\mu^2} + n\phi\Big(\ln\dfrac{n}{B}\Big) + o(n), & \textit{if } d > 0,\end{cases}
\end{aligned}$$

where μ and σ are the constants in (2) and $\phi(\cdot)$ is a bounded continuous d-periodic function.

PROOF. Let V_i, $i \geq 1$ be i.i.d. copies of V, and define $L_k = \prod_{i=1}^k V_i$ and $S_k = -\ln L_k$. Then, we have

$$\begin{aligned}
\mathbf{E}\Big[\sum_{v\neq\varnothing} \mathrm{Bin}(n, L_v)\mathbf{1}_{\{nL_v\geq B\}}\Big] &= n\mathbf{E}\Big[\sum_{k\geq 1} b^k L_k \mathbf{1}_{\{nL_k\geq B\}}\Big] \\
&= n\mathbf{E}\Big[\sum_{k\geq 1} b^d e^{-S_k}\mathbf{1}_{\{S_k\leq \ln n-\ln B\}}\Big] \\
&= n\int_0^{\ln(n/B)} \sum_{k\geq 1} b^k e^{-t}\, d\mathbf{P}(S_k\leq t) \\
&= n\int_0^{\ln(n/B)} e^{-t}\, dU(t),
\end{aligned}$$

where $U(t)$ is the renewal function defined in (21). Using integration by parts we obtain, if $-\ln V$ is nonlattice,

$$\begin{aligned}
&\int_0^{\ln(n/B)} e^{-t}\, dU(t) \\
&\quad = [e^{-t}U(t)]_0^{\ln(n/B)} + \int_0^{\ln(n/B)} e^{-t}U(t)\, dt
\end{aligned}$$

$$= \frac{B}{n} U\big(\ln(n/B)\big) + \int_0^{\ln(n/B)} e^{-t} \big(U(t) - \mu^{-1} e^t\big)\, dt$$
$$+ \mu^{-1} \ln(n/B)$$
$$= \mu^{-1} + o(1) + \frac{\sigma^2 - \mu^2}{2\mu^2} - \mu^{-1}$$
$$+ \mu^{-1} \ln(n/B) + o(1)$$

by Lemma 4.2 and (24). Similarly if $-\ln V$ is lattice with span d, Lemma 4.2 and (24) yield

$$\int_0^{\ln(n/B)} e^{-t}\, dU(t) = \mu^{-1} + o(1) + \frac{\sigma^2 - \mu^2}{2\mu^2} - \mu^{-1} + \mu^{-1} \ln(n/B)$$
$$+ \phi\big(\ln(n/B)\big) + o(1),$$

where $\phi(t)$ is a continuous periodic function with period d. $\square$

We now deal with the remainder $R_{n,B}$ introduced in (35). The difference between n_v and the binomial is bounded in (20) and we have

$$|R_{n,B}| \le \mathbf{E}\bigg[\sum_{v \neq \varnothing} \mathbf{1}_{\{nL_v \ge B\}} \sum_{u \preceq v} \mathrm{Bin}(s, L_v/L_u)\bigg].$$

LEMMA 4.4. *The following estimate holds*: *there exist a constant and n_0 such that*, *for every fixed B and $n \ge n_0$*, *we have*

$$\mathbf{E}\bigg[\sum_{v \neq \varnothing} \mathbf{1}_{\{nL_v \ge B\}} \sum_{u \preceq v} \mathrm{Bin}(s, L_v/L_u)\bigg] = O\bigg(\frac{n}{B}\bigg).$$

PROOF. In the following, $|v| = d$, $|u| = k \le d$, and we write $\ell = d - k$. Then L_v is distributed as $L_d = L_k \cdot L_\ell$, where the two factors are products of k and ℓ copies of V, respectively; all of them are independent. Swapping the sums over u and v, we obtain

$$\mathbf{E}\bigg[\sum_{v \neq \varnothing} \mathbf{1}_{\{nL_v \ge B\}} \sum_{u \preceq v} \mathrm{Bin}(s, L_v/L_u)\bigg]$$
$$= \mathbf{E}\bigg[\sum_{u} \sum_{v\,:\, u \preceq v, v \neq \varnothing} s \frac{L_v}{L_u} \mathbf{1}_{\{nL_v \ge B\}}\bigg] \le s\mathbf{E}\bigg[\sum_{k \ge 0} b^k \sum_{\ell \ge 0} b^\ell L_\ell \mathbf{1}_{\{nL_k L_\ell \ge B\}}\bigg] \tag{36}$$
$$= s\mathbf{E}\bigg[\sum_{k \ge 0} b^k \sum_{\ell \ge 0} b^\ell e^{-S_\ell} \mathbf{1}_{\{e^{S_k + S_\ell} \le n/B\}}\bigg].$$

First conditioning on S_k in each term of the sum above, and recalling the renewal function $U(t)$ defined in (21), we see that

$$\mathbf{E}\biggl[\sum_{\ell\geq 0} b^\ell e^{-S_\ell}\mathbf{1}_{\{e^{S_k+S_\ell}\leq n/B\}}\biggm| S_k\biggr] = \int_0^{\ln(n/B)-S_k} e^{-t}\,dU(t) + b\mathbf{1}_{\{e^{S_k}\leq n/B\}}.$$

However, there exists a constant C such that, for any real number x,

$$\int_0^x e^{-t}\,dU(t) \leq Cx\mathbf{1}_{\{x\geq 0\}}.$$

Going back to (36) and choosing $x = \ln(n/B) - S_k$, it follows that

$$\begin{aligned}
&\mathbf{E}\biggl[\sum_{v\neq\varnothing}\mathbf{1}_{\{nL_v\geq B\}}\sum_{u\preceq v}\mathrm{Bin}(s, L_v/L_u)\biggr]\\
&\quad\leq C\mathbf{E}\biggl[\sum_{k\geq 0} b^k\bigl(\ln(n/B) - S_k + b\bigr)\mathbf{1}_{\{S_k\leq\ln(n/B)\}}\biggr]\\
&\quad= C\int_0^{\ln(n/B)}\bigl(\ln(n/B) - t + b\bigr)\,dU(t)\\
&\quad= C\bigl[\bigl(\ln(n/B) - t\bigr)U(t)\bigr]_0^{\ln(n/B)}\\
&\qquad+ C'\int_0^{\ln(n/B)} U(t)\,dt,
\end{aligned}$$

where the last line follows by integration by parts and we wrote $C' = C(1+b)$. The claim then follows from (24). $\square$

4.4. *Contribution of the fringe*: *Proof of Proposition* 4.2. Finally, we prove Proposition 4.2 that deals with the contribution of the fringe of the tree. Recall that from (17), we have to estimate

$$\mathbf{E}\biggl[\sum_{r\in R}\widetilde{\Psi}(T^{n_r})\biggr] := \mathbf{E}\biggl[\sum_{r\in R}\Psi(T^{n_r}) + n_r\biggr], \tag{37}$$

where, for convenience, we introduced $\widetilde{\Psi}(T^k) := \Psi(T^k) + k$. The proofs here get quite technical at times, and the reader should bear in mind that we will essentially express the expected value in (37) as a mixture of the expected values of $\mathbf{E}[\widetilde{\Psi}(T^k)]$, for k lower than B.

For a node r, define the conditional expectation $\Gamma_r = \mathbf{E}[\widetilde{\Psi}(T^{n_r}) \mid n_r]$. First, the first asymptotic order of the expected total path length implies that

$$\Gamma_r = O(n_r \ln n_r). \tag{38}$$

The next lemma is used to get an error bound for the sum of the expected total path lengths of the subtrees $T_r, r\in R$, with cardinalities n_r that differ from nL_r by at least $B^{2/3}$ items, so that we only have to bother about the subtrees $T_r, r \in R$, with cardinalities n_r that are close to nL_r.

LEMMA 4.5. *The following error bound holds*:

$$\mathbf{E}\left[\sum_{r\in R} n_r \ln n_r \mathbf{1}_{\{|n_r - nL_r| \geq B^{2/3}\}}\right] = \mathcal{O}\left(\frac{n \ln B}{B^{1/4}}\right).$$

We omit the proof; it follows by a simple modification of the proof of Lemma 4.3 of [26]. By Lemma 4.5, we have

$$\mathbf{E}\left[\sum_{r\in R} \widetilde{\Psi}(T^{n_r})\right] = \mathbf{E}\left[\sum_{r\in R} \Gamma_r \mathbf{1}_{\{|n_r - nL_r| \leq B^{2/3}\}}\right] + O\left(\frac{n \ln B}{B^{1/4}}\right).$$

Define $R' \subseteq R$ to be the set of "good" nodes in R:

$$R' := \{r \in R : |n_r - nL_r| \leq B^{2/3}\} \tag{39}$$

and let $R'' \subseteq R'$ be the subset of nodes $r \in R'$ that also satisfy $nL_r > \varepsilon^2$.

We will now explain that it is enough to consider the nodes $r \in R''$. The approximation of $U(t)$ in (24) implies that the expected number of nodes v such that $nL_v \geq B$ is $O(n/B)$; thus, since each node has at most b children,

$$\mathbf{E}[|R|] = O(n/B) \tag{40}$$

as well. Hence, it follows from (39) that the expected number of nodes in the T_r, $r \in R'$, with $nL_r \leq \varepsilon^2 B$ is bounded by $O(\varepsilon^2 n)$. Using this fact yields

$$\mathbf{E}\left[\sum_{r\in R} \widetilde{\Psi}(T^{n_r})\right] = \mathbf{E}\left[\sum_{r\in R''} \Gamma_r\right] + \mathcal{O}(\varepsilon^2 n \ln B) + O\left(\frac{n \ln B}{B^{1/4}}\right). \tag{41}$$

Because of the concentration of n_r around nL_r, the cardinalities n_r of the nodes $r \in R$ are naturally related to the behavior of the "overshoot" of the renewal process $(-\ln L_k, k \geq 0)$, when it crosses the line $\ln(n/B)$. Estimating the empirical distribution of the cardinalities of the nodes $r \in R$ will allow us to approximate the right-hand side above. So we further subdivide the nodes $r \in R$ into smaller classes according to the values of $nL_r, r \in R$.

Let $Z = \{B, B - \gamma B, B - 2\gamma B, \ldots, \varepsilon^2 B\}$, where we let $\gamma = \varepsilon^3$. We write $R_z \subseteq R, z \in Z$, for the set of nodes $r \in R$, such that $nL_r \in [z - \gamma B, z)$. Then (41) can be rewritten as

$$\mathbf{E}\left[\sum_{r\in R} \widetilde{\Psi}(T^{n_r})\right] = \mathbf{E}\left[\sum_{z\in Z} \sum_{r\in R'\cap R_z} \Gamma_r\right] + O(\varepsilon^2 n \ln B) + O\left(\frac{n \ln B}{B^{1/4}}\right). \tag{42}$$

Even in a fixed class R_z, not all the nodes have the same cardinality n_r. So, in order to estimate the expected value in (42) we need the following lemma that quantifies the discrepancy of $\mathbf{E}[\Psi(T^n)]$ under small variations of n.

LEMMA 4.6. *There exists a constant C such that, for any natural numbers n and K, we have*

$$|\mathbf{E}[\widetilde{\Psi}(T^{n+K})] - \mathbf{E}[\widetilde{\Psi}(T^n)]| \leq CK \ln(n+K).$$

PROOF. From the iterative construction, we clearly have $\mathbf{E}[\widetilde{\Psi}(T^{n+K})] \geq \mathbf{E}[\widetilde{\Psi}(T^n)]$; so it suffices to bound the increase in path length when adding K extra items to the tree T^n. Thinking again of the iterative construction, every ball trickles down until it finds a leaf. Then, either it sits there if there is room left, or it triggers a growth of the tree. It is important to notice that only these $s+1$ balls may move. Furthermore, the increase in depth of any of the $s+1$ items (the last one, plus the s that were already sitting at the leaf) is at most the height of the final tree H_{n+K}. Hence, upon adding K items, the path length increases by $K(s+1)H_{n+K} \leq CK \ln(n+K)$, by the results of [13] on the height of split trees. □

Write $f_x = \mathbf{E}[\widetilde{\Psi}(T^{\lfloor x \rfloor})]$. Then Lemma 4.6 ensures that, for any node $r \in R' \cap R_z$, we have $\Gamma_r = f_z + O(\gamma B \ln B)$. By using (39) and Lemma 4.6, from (42) we obtain

$$\begin{aligned}
\mathbf{E}\Big[\sum_{r\in R} \widetilde{\Psi}(T^{n_r})\Big] &= \sum_{z\in Z} \mathbf{E}[|R' \cap R_z|]\big(f_z + O(\gamma B \ln B)\big) \\
&\quad + O(\varepsilon^2 n \ln B) + O\Big(\frac{n \ln B}{B^{1/4}}\Big) \\
&= \sum_{z\in Z} \mathbf{E}[|R' \cap R_z|] f_z + O(\gamma n \ln B) \\
&\quad + O(\varepsilon^2 n \ln B) + O\Big(\frac{n \ln B}{B^{1/4}}\Big),
\end{aligned} \tag{43}$$

since $\mathbf{E}[|R|] = O(n/B)$ by (40).

So the contribution of the fringe is essentially a mixture of the f_z, $z \in Z$. To complete the proof of Proposition 4.2, it suffices to estimate the mixing measure $\mathbf{E}[|R' \cap R_z|]$, $z \in Z$. We first focus on the asymptotics for $\mathbf{E}[|R_z|]$, $z \in Z$. The following result is obtained by an application of the key renewal theorem.

LEMMA 4.7. *Fix $\varepsilon > 0$ and let $S := \{1, 1-\gamma, 1-2\gamma, \dots, \varepsilon^2\}$, where $\gamma = \varepsilon^3$. Let $d = \sup\{a : \mathbf{P}(\ln V \in a\mathbb{Z}) = 1\}$. If $d > 0$, we suppose that $\ln B \in d\mathbb{N}$. Then for any $\alpha \in S$ we have, as $n \to \infty$,*

$$\frac{\mathbf{E}[|R_{\alpha B}|]}{n/B} = \begin{cases} c_\alpha + o(1), & \text{if } \ln V \text{ is nonlattice } (d = 0), \\ \psi_\alpha(\ln n) + o(1), & \text{if } \ln V \text{ is } d\text{-lattice } (d > 0), \end{cases} \tag{44}$$

for a constant c_α (only depending on α and γ), $\psi_\alpha(\cdot)$ is the d-periodic function given in (48) below.

PROOF. Let V_j, $j \geq 1$, be i.i.d. copies of V. For an integer k, write $S_k = -\sum_{j=1}^{k} \ln V_j$. Then, by definition, for $\alpha \in S$, we have

$$\begin{aligned}
\mathbf{E}[|R_{\alpha B}|] &= \sum_{u \in U} \mathbf{P}(u \in R_{\alpha B}) \\
&= \sum_{k=0}^{\infty} b^{k+1} \bigl(\mathbf{P}\bigl(S_k - \ln V_{k+1} > \ln(n/B) - \ln\alpha \text{ and } S_k \leq \ln(n/B)\bigr) \\
&\qquad\qquad - \mathbf{P}\bigl(S_k - \ln V_{k+1} > \ln(n/B) - \ln(\alpha - \gamma) \\
&\qquad\qquad\qquad\qquad \text{and } S_k \leq \ln(n/B)\bigr)\bigr) \\
&= \int_0^{\ln(n/B)} b\mathbf{P}\bigl(\ln(n/B) - t - \ln\alpha \\
&\qquad\qquad < -\ln V_{k+1} \leq \ln(n/B) - t - \ln(\alpha - \gamma)\bigr)\, dU_0(t),
\end{aligned}$$

where $U_0(t) = U(t) + 1$ is a simple modification of the renewal $U(t) = \sum_{k \geq 1} b^k \mathbf{P}(S_k \leq t)$ defined in (21). Thus, seeing $\mathbf{E}[|R_{\alpha B}|]$ as a function of $\ln(n/B)$ and writing

$$H(q) := \int_0^q b\mathbf{P}\bigl(q - t - \ln\alpha < -\ln V_{k+1} \leq q - t - \ln(\alpha - \gamma)\bigr)\, dU_0(t), \tag{45}$$

we have $\mathbf{E}[|R_{\alpha B}|] = H(\ln(n/B))$. So we are after the asymptotics for $H(q)$, as $q \to \infty$. It is convenient to use a change of measure to relate $H(q)$ to a renewal function associated to a *probability* measure. We have

$$\begin{aligned}
\widehat{H}(q) &:= e^{-q} H(q) \\
&= \int_0^q e^{-(q-t)} G(q - t) e^{-t}\, dU_0(t) \qquad (46)\\
&= \int_0^q b e^{-(q-t)} \mathbf{P}\bigl(q - t - \ln\alpha < -\ln V_{k+1} \leq q - t - \ln(\alpha - \gamma)\bigr)\, dF(t),
\end{aligned}$$

where $F(t)$ is the standard renewal function already introduced in (26). The asymptotics for the integral above are then easily obtained by using the key renewal theorem. In particular, they depend on whether $\ln V$ is lattice or not.

(i) If $\ln V$ is nonlattice, by the key renewal theorem ([22], Theorem II.4.3), we obtain

$$\lim_{q \to \infty} \widehat{H}(q) = c_\alpha := \frac{b}{\mu} \int_0^\infty e^{-t} \mathbf{P}\bigl(t - \ln\alpha < -\ln V \leq t - \ln(\alpha - \gamma)\bigr)\, dt. \tag{47}$$

Note that the constant c_α only depends on α (and γ) and that $\sum_{\alpha \in S} c_\alpha \leq b/\mu$. Thus, since $\widehat{H}(x) = e^{-x} H(x)$ it follows immediately that $\mathbf{E}[|R_{\alpha B}|] = \frac{n}{B} c_\alpha + o(\frac{n}{B})$ which proves the nonlattice case in (44).

(ii) Similarly, if $\ln V$ is lattice with span d, the key renewal theorem (see [22], Theorem II.4.3, or [32], Theorem A.7) implies that

$$\begin{aligned}\widehat{H}(q) &\sim \psi_\alpha(q)\\ &:= \frac{bd}{\mu}\sum_{k\,:\,kd\le q} e^{kd-q}\mathbf{P}\big(q-kd-\ln\alpha < -\ln V \le q-kd-\ln(\alpha-\gamma)\big)\end{aligned} \tag{48}$$

as $q\to\infty$. Note that ψ_α is a (positive) d-periodic function. Observe also that for fixed α, the function $\psi_\alpha(\cdot)$ is not continuous since $\ln V \in d\mathbb{Z}$ almost surely. Since $\widehat{H}(x) = e^{-x}H(x)$, it follows from (48) that $\mathbf{E}[|R_{\alpha B}|] \sim \frac{n}{B}\psi_\alpha(\ln(n/B))$. This proves the lattice case in (44), and completes the proof. □

With Lemma 4.7 in hand, we can now deduce the asymptotics for $\mathbf{E}[|R'\cap R_z|]$, $z\in Z$ and use them in (43) to complete the proof of Proposition 4.2. Recall that $R' = \{r\in R : |n_r - nL_r| \le B^{2/3}\}$. Clearly, $\mathbf{E}[|R'\cap R_{\alpha B}|] \le \mathbf{E}[|R_{\alpha B}|]$. Furthermore,

$$\begin{aligned}\mathbf{E}[|R'\cap R_{\alpha B}|] &= \sum_{r\in R}\mathbf{P}\big(|n_r - nL_r| \le B^{2/3}, (\alpha-\gamma)B \le nL_r < \alpha B\big)\\ &= \sum_{r\in R}\mathbf{P}\big((\alpha-\gamma)B \le nL_r < \alpha B\big)\\ &\qquad\times \mathbf{P}\big(|n_r - nL_r| \le B^{2/3} \mid (\alpha-\gamma)B \le nL_r < \alpha B\big)\\ &\ge \mathbf{E}[|R_{\alpha B}|]\big(1 - O(B^{-1/4})\big)\end{aligned}$$

by Lemma 4.1. We now choose $B = \varepsilon^{-20}$ so that $B^{-1/4} = \varepsilon^5$.

(i) If $\ln V$ is nonlattice, it follows from Lemma 4.7 that for each choice of γ there is a constant K_γ such that for all $\alpha\in S$ and some constant c_α (that of Lemma 4.7) we have

$$\left|\frac{\mathbf{E}[|R'\cap R_{\alpha B}|]}{n/B} - c_\alpha\right| \le \gamma^2 + O(B^{-1/4}) = \gamma^2 + O(\varepsilon^5) = O(\varepsilon^5),$$

whenever $n/B \ge K_\gamma$. So for all n large enough, since $f_x = O(x\ln x)$, we have

$$\begin{aligned}\mathbf{E}\bigg[\sum_{r\in R}\widetilde{\Psi}(T^{n_r})\bigg] &= \sum_{\alpha\in S} c_\alpha \frac{n}{B} f_{\alpha B} + \frac{n}{B}\sum_{\alpha\in S} O(f_{\alpha B}\varepsilon^5) + O(n\gamma\ln B) + O(\varepsilon^2 n\ln B)\\ &= n\sum_{\alpha\in S}\frac{f_{\alpha B}}{B}c_\alpha + O(\varepsilon n).\end{aligned}$$

This proves Proposition 4.2 when $\ln V$ is nonlattice.

(ii) Similarly, if $\ln V$ is d-lattice, for any choice of γ, there is a K_γ such that for any $\alpha\in S$ and some continuous d-periodic function $\psi_\alpha(t)$ [that of Lemma 4.7

defined in (48)], we have

$$\left|\frac{\mathbf{E}[|R' \cap R_{\alpha B}|]}{n/B} - \psi_\alpha(\ln n)\right| \le \gamma^2 + O(B^{-1/4}) = \gamma^2 + O(\varepsilon^5),$$

whenever $n/B \ge K_\gamma$. It follows that

$$\begin{aligned}\mathbf{E}\Big[\sum_{r\in R}\widetilde{\Psi}(T^{n_r})\Big] &= \sum_{\alpha\in S}\psi_\alpha(\ln n)\frac{n}{B}f_{\alpha B} + \frac{n}{B}\sum_{\alpha\in S}O(f_{\alpha B}\varepsilon^5)\\ &\quad + O(n\gamma\ln B) + O(\varepsilon^2 n\ln B) \qquad (49)\\ &= n\sum_{\alpha\in S}\frac{f_{\alpha B}}{B}\psi_\alpha(\ln n) + O(\varepsilon n).\end{aligned}$$

This proves the claim in the lattice case with φ_B defined by

$$\varphi_B(q) := \sum_{\alpha\in S}\frac{f_{\alpha B}}{B}\psi_\alpha(q). \qquad (50)$$

It now only remains to prove that, although the functions $\psi_\alpha(\cdot)$, $\alpha \in S$, are not continuous, the d-periodic function φ_B satisfies the bound in (19).

LEMMA 4.8. *The function φ_B defined in (50) satisfies*

$$\sup_{|q-q'|\le\varepsilon^3}|\varphi_B(q) - \varphi_B(q')| \le K\varepsilon\ln(1/\varepsilon).$$

PROOF. From the expresssion for ψ_α in (48), we have

$$\begin{aligned}\varphi_B(q) &= \frac{bd}{\mu}\sum_{\alpha\in S}\frac{f_{\alpha B}}{B}\sum_{k:kd\le q}e^{kd-q}\mathbf{P}\big(q - kd + \ln V \in [\ln(\alpha-\gamma),\ln\alpha)\big)\\ &= \frac{bd}{\mu}\sum_{k:kd\le q}e^{kd-q}\sum_{\alpha\in S}\frac{f_{\alpha B}}{B}\mathbf{P}\big(q - kd + \ln V \in [\ln(\alpha-\gamma),\ln\alpha)\big).\end{aligned}$$

Note that, since $\gamma = \varepsilon^3$ and $\alpha \ge \varepsilon^2$,

$$|\ln(\alpha-\gamma) - \ln\alpha| \sim \frac{\gamma}{\alpha}$$

as $\varepsilon \to 0$. As a consequence, for all $\varepsilon > 0$ small enough, the intervals involved in the definition of ψ_α satisfy, uniformly in $\alpha \in S$,

$$\frac{\varepsilon^3}{2} < |\ln(\alpha-\gamma) - \ln\alpha| \le \varepsilon.$$

In particular, since $\ln V \in d\mathbb{Z}$ almost surely, there is at most one atom in the interval as soon as $\varepsilon < d$. It follows that, if we choose $\delta = \varepsilon^3/2$, we have for any q, q' such that $|q - q'| < \delta$

$$\mathbf{P}\big(q' - kd + \ln V \in [\ln(\alpha-\gamma),\ln\alpha)\big) = \mathbf{P}\big(q - kd + \ln V \in [\ln(\alpha'-\gamma),\ln\alpha')\big)$$

for some α' in $\{\alpha+\gamma, \alpha, \alpha-\gamma\}$. We adopt the following point of view: for fixed k and q, S induces a partition into the intervals $[q-kd-\ln(\alpha), q-kd-\ln(\alpha-\gamma))$, $\alpha \in S$. Each interval contains at most one atom of $-\ln V$. Changing q into q' as above modifies the partition, but each atom may only move to an adjacent interval. All atoms of $\ln V$ appear in both sums, except if one is so far that it escapes the range of the partition (recall that $\alpha \ge \varepsilon^2$). So following the atoms of $-\ln V$ rather than the intervals in one or the other partition yields

$$\begin{aligned}
&\frac{\mu}{bd}|\varphi_B(q)-\varphi_B(q')| \\
&\qquad \le \max_{x\in\{q,q'\}} \sum_{k:\,kd\le x} e^{kd-x+\delta} \sum_{\alpha\in S} \max_{|\alpha'-\alpha|\le\gamma} \left|\frac{f_{\alpha B}}{B}-\frac{f_{\alpha' B}}{B}\right| \\
&\qquad\qquad\qquad \times \mathbf{P}\big(x-kd+\ln V \in \big[\ln(\alpha-\gamma), \ln\alpha\big)\big) \\
&\qquad\quad + \max_{x\in\{q,q'\}} \sum_{k:\,kd\le x} e^{kd-x}\frac{f_{\varepsilon^2 B}}{B},
\end{aligned}$$

where the second term accounts for the escape of one atom. It follows that

$$\begin{aligned}
&\frac{\mu}{bd}|\varphi_B(q)-\varphi_B(q')| \\
&\qquad \le \max_{x\in\{q,q'\}} \sum_{k:\,kd\le x} e^{kd-x+\delta} \sum_{\alpha\in S} K\gamma \ln B \cdot \mathbf{P}\big(x-kd+\ln V \in \big[\ln(\alpha-\gamma), \ln\alpha\big)\big) \\
&\qquad\quad + K\varepsilon^2 \ln B
\end{aligned}$$

for some constant K, by Lemma 4.6 and the asymptotics for f_z. Swapping the sums once again to recover the functions $\psi_\alpha(\cdot)$, it follows that

$$|\varphi_B(q)-\varphi_B(q')| \le \frac{bd}{\mu} K\gamma e^{\delta} \ln B \cdot \sup_x \sum_{\alpha\in S} \psi_\alpha(x).$$

However, since every summand is nonnegative, we have for any x

$$\begin{aligned}
0 \le \sum_{\alpha\in S}\psi_\alpha(x) &= \frac{bd}{\mu} \sum_{k:\,kd\le x} e^{kd-x} \sum_{\alpha\in S} \mathbf{P}\big(x-kd+\ln V \in \big[\ln(\alpha-\gamma), \ln\alpha\big)\big) \\
&\le \frac{bd}{\mu} \sum_{k:\,kd\le x} e^{kd-x} \le \frac{be^d}{\mu}.
\end{aligned} \tag{51}$$

The desired bound follows: for any q, q' such that $|q-q'| < \varepsilon^3/2$, we have

$$|\varphi_B(q)-\varphi_B(q')| \le K''\varepsilon \ln(1/\varepsilon)$$

for some constant K'' independent of q, q' or ε. $\square$

5. Extensions and concluding remarks.

5.1. *An alternative notion of path length.* The notion of path length we have considered so far is the sum of the depths of the *items* in the tree. This is most natural when one thinks about performance measures for algorithms or sorted data structures. However, for some applications, it is sometimes important to introduce a related notion of path length $\Upsilon(T)$, that is the sum of the depths of *nodes*:

$$\Upsilon(T) := \sum_{u \in \mathcal{U}} |u| \mathbf{1}_{\{u \in T\}} = \sum_{u \neq \sigma} N_u,$$

where N_u denotes the number of nodes in the subtree rooted at u. This notion of path length appears, for instance, in the analysis of cutting-down processes. Suppose that you are given a rooted tree T. Initially, the process starts with T. At each time step, a uniformly random edge is cut, the portion of the tree that is disconnected from the root is lost, and the process continues with the portion containing the root. How many random cuts does it take to isolate the root? The question originates in the seminal work of Meir and Moon [40, 41]. Recently, the subject has regained interest, and new results have been proved about the weak limit of the number of cuts when the initial tree is randomly picked according to various distributions. See [16, 27–29, 31] for more references and details about the precise models and results.

For instance, Holmgren [28] has proved that, when the initial tree is a split tree satisfying two general conditions (one on $\mathbf{E}[\Upsilon(T^n)]$ and one on the number of nodes), the normalized number of cuttings converges in distribution to a weakly 1-stable law (Theorem 1.1 there). Our Theorem 3.1 allows us to prove that one of the conditions assumed in [28] actually implies the other. More precisely, the conditions assumed in [28] are that $\Upsilon(T^n)$ (the path length of nodes) satisfies

$$\mathbf{E}[\Upsilon(T^n)] = \frac{\alpha}{\mu} n \ln n + \zeta n + o(n),$$

and that the number of nodes $N = |T^n|$ verifies, for some constants $\alpha > 0$ and $\varepsilon > 0$,

$$\mathbf{E}[N] = \alpha n + f(n) \qquad \text{where } f(n) = O\left(\frac{n}{\ln^{1+\varepsilon} n}\right). \tag{52}$$

We deduce from Theorem 3.1:

COROLLARY 5.1. *Suppose that* $\ln V$ *is nonlattice, and assume that* (52) *holds true*; *then, as* $n \to \infty$,

$$\mathbf{E}[\Upsilon(T^n)] = \frac{\alpha}{\mu} n \ln n + \zeta n + o(n).$$

REMARKS. The assumption in (52) is just slightly stronger than the estimate proved by Holmgren [26], that is, that for split tree with nonlattice $\ln V$, we have $f(n) = o(n)$. Moreover, the assumption in (52) does make sense, since it is known to hold, for instance, for m-ary search trees [2, 10, 36, 39]: for such random trees, $f(n)$ is $o(\sqrt{n})$ when $m \le 26$ and is $O(n^{1-\varepsilon})$ when $m \ge 27$. On the other hand, it is also known that the condition in (52) does not always hold. For instance, Flajolet et al. [20] proved that, in the case of binary tries generated by a memoryless source with probabilities p_1, p_2 such that $(\log p_1)/(\log p_2)$ is a Liouville number, then the error term $f(n)$ can come arbitrarily close to $O(n)$ [but of course, stays $o(n)$]. See [20], page 249, and the monograph by Baker [3] for more information about Liouville numbers.

SKETCH OF PROOF. Define $q(n)$ and $r(n)$ by

$$\mathbf{E}[\Psi(T^n)] = \frac{1}{\mu} n \ln n + nq(n) \quad \text{and} \quad \mathbf{E}[\Upsilon(T^n)] = \frac{\alpha}{\mu} n \ln n + nr(n).$$

Let $\Delta_n := \alpha nq(n) - nr(n)$, and note that

$$\Delta_n = \alpha \mathbf{E}[\Psi(T^n)] - \mathbf{E}[\Upsilon(T^n)]. \tag{53}$$

Since, by Theorem 3.1, $q(n)$ converges as $n \to \infty$, it suffices to prove that Δ_n/n also converges to some constant. From (53) and the assumption in (52) we obtain

$$\begin{aligned} \Delta_n &= \alpha \mathbf{E}\biggl[\sum_{v \ne \sigma} n_v\biggr] - \mathbf{E}\biggl[\sum_{v \ne \sigma}\biggl(\alpha n_v + O\biggl(\frac{n_v}{\ln^{1+\varepsilon} n_v}\biggr)\biggr)\biggr] \\ &= \mathbf{E}\biggl[\sum_{v} O\biggl(\frac{n_v}{\log^{1+\varepsilon} n_v}\biggr)\biggr]. \end{aligned} \tag{54}$$

[The constants hidden in the $O(\cdot)$ above are the same for every term.]

Consider the subtrees T_r, $r \in R$, introduced in the course of the proof of Theorem 3.1. Recall that a node r is in R if it is the first on its path from the root such that $nL_r \le B$, for some parameter B. In the following, we take $B = \delta^{-8}$, for $\delta > 0$. We now show that the main contribution to Δ_n is accounted for by the nodes in the subtrees T_r, $r \in R$; in other words $\Delta_n = \mathbf{E}[\sum_{r \in R} \Delta_{n_r}] + o(n)$, where

$$\Delta_{n_r} = \alpha \mathbf{E}[\Psi(T_r) | n_r] - \mathbf{E}[\Upsilon(T_r) | n_r].$$

To see this, observe that we deduce from (54) and (52) that

$$\begin{aligned} \Delta_n - \mathbf{E}\biggl[\sum_{r \in R} \Delta_{n_r}\biggr] &= \mathbf{E}\biggl[\sum_{\substack{v \notin T_r, r \in R, \\ v \ne \sigma}} O\biggl(\frac{n_v}{\log^{1+\varepsilon} n_v}\biggr)\biggr] \\ &= \mathbf{E}\biggl[\sum_{k \ge 0} \sum_{\substack{v \notin T_r, r \in R, \\ 2^k \le n_v < 2^{k+1}}} O\biggl(\frac{n_v}{\log^{1+\varepsilon} n_v}\biggr)\biggr] + O\biggl(\frac{n}{\log n}\biggr). \end{aligned}$$

We split the sum in k above at some constant K to be chosen later. By Lemma 4.1 and since the expected number of nodes $v \in T^n$ with $nL_v \geq B$ is $O(n/B)$, we obtain

$$\begin{aligned}\Delta_n - \mathbf{E}\Bigg[\sum_{r\in R}\Delta_{n_r}\Bigg] &= \sum_{k>K} O\left(\frac{n}{2^k}\cdot\frac{2^k}{k^{1+\varepsilon}}\right) + \sum_{0\leq k\leq K} O\left(\frac{n}{B}\cdot\frac{2^k}{k^{1+\varepsilon}}\right) + o(n)\\ &= O(nK^{-\varepsilon}) + O(nK2^K/B) + o(n).\end{aligned}$$

We choose $K = \lfloor a\ln(1/\delta)\rfloor$, for some small constant $a > 0$. Since $\delta > 0$ was arbitrary, the claim follows.

Now since $\Delta_{n_r} = O(n_r \ln n_r)$, the proof of Proposition 4.2 (in the nonlattice case) may be extended to show that $\mathbf{E}[\sum_{r\in R}\Delta_{n_r}] = n\zeta + o(n)$ for some constant ζ. The details are omitted. $\square$

5.2. *Beyond split trees and multinomial partitions*. To conclude, we indicate the lines of the arguments to extend the applicability of our main theorem to a greater family of random trees. The model of split trees [13] supposes that the distribution of the subtree cardinalities $n_1, n_2, \dots, n_b$ of a node of cardinality n is *exactly* of the form

$$(n_1, n_2, \dots, n_b) = \mathrm{Mult}(n - s_0 - bs_1, V_1, V_2, \dots, V_b) + (s_1, s_1, \dots, s_1) \tag{55}$$

for a random vector $(V_1, \dots, V_b)$; in particular, the vector $(V_1, \dots, V_b)$ cannot depend on n. Although many important data structures satisfy this property, some other more combinatorial examples do not; see, for instance, the case of increasing trees [5].

Also, the reader might have noticed that our proof does not quite use the full strength of the assumption in (55). Indeed, our proof mainly uses two facts: first, that the sequence of subtree sizes along a branch is well approximated by the product form $nL_u = n\prod_{v\preceq u} V_v$, which modulo some details about $C(\mathcal{V})$, implies that

$$X \stackrel{d}{=} \sum_{k=1}^{b} V_k X^{(k)} + C(\mathcal{V});$$

and second, that the addition of some items to the tree only modifies moderately $\mathbf{E}[\Psi(T)]$ (see Lemma 4.6).

The two requirements are satisfied when the items are distributed in subtrees according to (55). We now indicate why our result would still hold under the much weaker condition that there exists a vector $\mathcal{V} = (V_1, \dots, V_b)$ such that the cardinalities $n_1, \dots, n_b$ of the children of a node of cardinality n satisfy

$$\left(\frac{n_1}{n}, \frac{n_2}{n}, \dots, \frac{n_b}{n}\right) \to (V_1, V_2, \dots, V_b) \qquad \text{in distribution} \tag{56}$$

as $n \to \infty$. Of course, the copies of the limit vectors $\mathcal{V}$ at distinct nodes should be independent. The general shape of trees under this model has recently been completed by work by Broutin et al. [8] (see also Drmota [15] who treats the model of increasing trees by Bergeron et al. [5] more directly).

One should be easily convinced that the relaxed condition in (56) should be sufficient for the result to hold:

- Proposition 4.1 may be extended using the coupling arguments already used in [8], proving that the contribution of the top of the tree to the path length may be estimated using renewal functions associated to the limit vector $\mathcal{V}$.
- Similarly, the extension of Proposition 4.2 relies on the same coupling argument (the overshoot there is still approximated by that of the limit vector). Here, it is important to note that the proof of smoothness of the path length (Lemma 4.6) requires the existence of a fixed function g such that the size $|T^n|$ of a "generalized" split tree of cardinality n satisfies $|T^n| \leq g(n)$ with probability 1 (at least our proof does). This was already necessary for the results on the shape of the trees in [8] to hold. The constraint is not too strong, since it holds as soon as s_0 or s_1 is nonzero, and any function would do, regardless of its growth. (This is another reason why the case of digital trees should be treated separately: for such trees, the size of a tree containing two items can be arbitrarily large.)
- As already noted in Section 3, the part of the proof relative to the contraction method in [45, 46] will go through as long as the coefficients $C_n(\overline{n})$ converge, and the expansion for mean implies their convergence.

Acknowledgments. We would like to warmly thank Svante Janson and Ralph Neininger for very helpful discussions. We are also grateful to an anonymous referee for his helpful comments.

REFERENCES

[1] ASMUSSEN, S. (2003). *Applied Probability and Queues*. Springer, New York.

[2] BAEZA-YATES, R. A. (1987). Some average measures in m-ary search trees. *Inform. Process. Lett.* **25** 375–381. MR0905782

[3] BAKER, A. (1990). *Transcendental Number Theory*, 2nd ed. Cambridge Univ. Press, Cambridge. MR1074572

[4] BELL, C. J. (1965). An investigation into the principles of the classification and analysis of data on an automatic digital computer. Ph.D. thesis, Leeds Univ.

[5] BERGERON, F., FLAJOLET, P. and SALVY, B. (1992). Varieties of increasing trees. In *CAAP'92* (*Rennes*, 1992). *Lecture Notes in Computer Science* **581** 24–48. Springer, Berlin. MR1251994

[6] BROUTIN, N. and DEVROYE, L. (2006). Large deviations for the weighted height of an extended class of trees. *Algorithmica* **46** 271–297. MR2291957

[7] BROUTIN, N., DEVROYE, L. and MCLEISH, E. (2008). Weighted height of random trees. *Acta Inform.* **45** 237–277. MR2403296

[8] BROUTIN, N., DEVROYE, L., MCLEISH, E. and DE LA SALLE, M. (2008). The height of increasing trees. *Random Structures Algorithms* **32** 494–518. MR2422392

[9] BRUHN, V. (1996). Eine methode zur asymptotischen behandlung einer klasse von rekursionsgleichungen mit einer anwendung in der stochastischen analyse des quicksort-algorithmus. Ph.D. thesis, Univ. Kiel.

[10] CHAUVIN, B. and POUYANNE, N. (2004). m-ary search trees when $m \geq 27$: A strong asymptotics for the space requirements. *Random Structures Algorithms* **24** 133–154. MR2035872

[11] CHERNOFF, H. (1952). A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations. *Ann. Math. Statist.* **23** 493–507. MR0057518

[12] COFFMAN, E. G. and EVE, J. (1970). File structures using hashing functions. *Communications of the ACM* **13** 427–436.

[13] DEVROYE, L. (1998). Universal limit laws for depths in random trees. *SIAM J. Comput.* **28** 409–432. MR1634354

[14] DOBROW, R. P. and FILL, J. A. (1999). Total path length for random recursive trees. *Combin. Probab. Comput.* **8** 317–333. MR1723646

[15] DRMOTA, M. (2009). The height of increasing trees. *Ann. Comb.* **12** 373–402. MR2496124

[16] DRMOTA, M., IKSANOV, A., MOEHLE, M. and ROESLER, U. (2009). A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures Algorithms* **34** 319–336. MR2504401

[17] FILL, J. A. and JANSON, S. (2001). Approximating the limiting Quicksort distribution. *Random Structures Algorithms* **19** 376–406. MR1871560

[18] FILL, J. A. and JANSON, S. (2002). Quicksort asymptotics. *J. Algorithms* **44** 4–28. MR1932675

[19] FINKEL, R. A. and BENTLEY, J. L. (1974). Quad trees, a data structure for retrieval on composite keys. *Acta Inform.* **4** 1–19.

[20] FLAJOLET, P., ROUX, M. and VALLÉE, B. (2010). Digital trees and memoryless sources: From arithmetics to analysis. In 21*st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms* (*AofA*'10). *Discrete Mathematics and Theoretical Computer Science Proceedings* **AM** 233–260. Assoc. Discrete Math. Theor. Comput. Sci., Nancy. MR2735344

[21] FREDKIN, E. (1960). Trie memory. *Communications of the ACM* **3** 490–499.

[22] GUT, A. (2009). *Stopped Random Walks: Limit Theorems and Applications*, 2nd ed. Springer, New York. MR2489436

[23] HIBBARD, T. N. (1962). Some combinatorial properties of certain trees with applications to searching and sorting. *J. Assoc. Comput. Mach.* **9** 13–28. MR0152155

[24] HOARE, C. A. R. (1962). Quicksort. *Comput. J.* **5** 10–15. MR0142216

[25] HOEFFDING, W. (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** 13–30. MR0144363

[26] HOLMGREN, C. (2012). Novel characteristics of split trees by use of renewal theory. *Electron. J. Probab.* **17** 1–27.

[27] HOLMGREN, C. (2010). Random records and cuttings in binary search trees. *Combin. Probab. Comput.* **19** 391–424. MR2607374

[28] HOLMGREN, C. (2011). A weakly 1-stable distribution for the number of random records and cuttings in split trees. *Adv. in Appl. Probab.* **43** 151–177. MR2761152

[29] IKSANOV, A. and MÖHLE, M. (2007). A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree. *Electron. Commun. Probab.* **12** 28–35. MR2407414

[30] JACQUET, P. and RÉGNIER, M. (1988). Normal limiting distribution for the size and the external path length of tries. Technical Report 827, INRIA-Rocquencourt.

[31] JANSON, S. (2006). Random cutting and records in deterministic and random trees. *Random Structures Algorithms* **29** 139–179. MR2245498

[32] JANSON, S. (2010). Renewal theory for the analysis of tries and strings: Extended abstract. In *Proceedings of the International Conference on Probabilistic*, *Combinatorial*, *and Asymptotic Methods in the Analysis of Algorithms* (*AofA*). *Discrete Mathematics and Theoretical Computer Science Proceedings* **AM** 427–438. Assoc. Discrete Math. Theor. Comput. Sci., Nancy.

[33] JANSON, S., ŁUCZAK, T. and RUCINSKI, A. (2000). *Random Graphs*. Wiley, New York. MR1782847

[34] KIRSCHENHOFER, P., PRODINGER, H. and SZPANKOWSKI, W. (1989). On the variance of the external path length in a symmetric digital trie. *Discrete Appl*. *Math*. **25** 129–143. MR1031267

[35] KIRSCHENHOFER, P., PRODINGER, H. and SZPANKOWSKI, W. (1994). Digital search trees again revisited: The internal path length perspective. *SIAM J*. *Comput*. **23** 598–616. MR1274646

[36] KNUTH, D. E. (1973). *The Art of Computer Programming*: *Sorting and Searching*, *Vol*. 3. Addison-Wesley, Reading, MA.

[37] LORDEN, G. (1970). On excess over the boundary. *Ann*. *Math*. *Statist*. **41** 520–527. MR0254981

[38] MAHMOUD, H. M. (1991). Limiting distributions for path lengths in recursive trees. *Probab*. *Engrg*. *Inform*. *Sci*. **5** 53–59. MR1183165

[39] 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

[40] MEIR, A. and MOON, J. W. (1970). Cutting down random trees. *J*. *Aust*. *Math*. *Soc*. **11** 313–324. MR0284370

[41] MEIR, A. and MOON, J. W. (1974). Cutting down recursive trees. *Math*. *Biosci*. **21** 173–181.

[42] MOHAMED, H. and ROBERT, P. (2005). A probabilistic analysis of some tree algorithms. *Ann*. *Appl*. *Probab*. **15** 2445–2471. MR2187300

[43] MOHAMED, H. and ROBERT, P. (2010). Dynamic tree algorithms. *Ann*. *Appl*. *Probab*. **20** 26–51. MR2582641

[44] MUNSONIUS, G. O. (2011). On the asymptotic internal path length and the asymptotic Wiener index of random split trees. *Electron*. *J*. *Probab*. **16** 1020–1047. MR2820068

[45] NEININGER, R. and RÜSCHENDORF, L. (1999). On the internal path length of d-dimensional quad trees. *Random Structures Algorithms* **15** 25–41. MR1698407

[46] NEININGER, R. and RÜSCHENDORF, L. (2004). A general limit theorem for recursive algorithms and combinatorial structures. *Ann*. *Appl*. *Probab*. **14** 378–418. MR2023025

[47] PYKE, R. (1965). Spacings (with discussion). *J*. *Roy*. *Statist*. *Soc*. *Ser*. *B* **27** 395–449. MR0216622

[48] RACHEV, S. T. and RÜSCHENDORF, L. (1995). Probability metrics and recursive algorithms. *Adv*. *in Appl*. *Probab*. **27** 770–799. MR1341885

[49] RÉGNIER, M. (1989). A limiting distribution for quicksort. *RAIRO Inform*. *Théor*. *Appl*. **23** 335–343. MR1020478

[50] RÖSLER, U. (1991). A limit theorem for "Quicksort". *RAIRO Inform*. *Théor*. *Appl*. **25** 85–100. MR1104413

[51] RÖSLER, U. (1992). A fixed point theorem for distributions. *Stochastic Process*. *Appl*. **42** 195–214. MR1176497

[52] RÖSLER, U. (2001). On the analysis of stochastic divide and conquer algorithms. *Algorithmica* **29** 238–261. MR1887306

[53] SCHACHINGER, W. (2004). Concentration of size and path length of tries. *Combin*. *Probab*. *Comput*. **13** 763–793. MR2096302

[54] SZPANKOWSKI, W. (2001). *Average Case Analysis of Algorithms on Sequences*. Wiley, New York. MR1816272

[55] TAN, K. H. and HADJICOSTAS, P. (1995). Some properties of a limiting distribution in Quicksort. *Statist. Probab. Lett.* **25** 87–94. MR1364822

PROJET ALGORITHMS
INRIA PARIS-ROCQUENCOURT
DOMAINE DE VOLUCEAU
78153 LE CHESNAY
FRANCE
E-MAIL: nicolas.broutin@inria.fr

DPMMS
CENTRE FOR MATHEMATICAL SCIENCES
WILBERFORCE ROAD
CAMBRIDGE CB3 0WA
UNITED KINGDOM
E-MAIL: cecilia@math.uu.se