ALEA, Lat. Am. J. Probab. Math. Stat. 19, 665–700 (2022)
DOI: 10.30757/ALEA.v19-26

The fluctuations of the giant cluster for percolation on random split trees

Gabriel Berzunza Ojeda, Xing Shi Cai and Cecilia Holmgren

Department of Mathematical Sciences, University of Liverpool, United Kingdom
E-mail address: `Gabriel.Berzunza-Ojeda@liverpool.ac.uk`

Duke Kunshan University, China
E-mail address: `xingshi.cai@dukekunshan.edu.cn`

Department of Mathematics, Uppsala University, Sweden
E-mail address: `cecilia.holmgren@math.uu.se`

Abstract. A split tree of cardinality n is constructed by distributing n "balls" in a subset of vertices of an infinite tree which encompasses many types of random trees such as m-ary search trees, quad trees, median-of-$(2k+1)$ trees, fringe-balanced trees, digital search trees and random simplex trees. In this work, we study Bernoulli bond percolation on arbitrary split trees of large but finite cardinality n. We show for appropriate percolation regimes that depend on the cardinality n of the split tree that there exists a unique giant cluster, the fluctuations of the size of the giant cluster as $n \to \infty$ are described by an infinitely divisible distribution that belongs to the class of stable (asymmetric) Cauchy laws. This work generalizes the results for the random m-ary recursive trees by Berzunza (2015). Our approach is based on a remarkable decomposition of the size of the giant percolation cluster as a sum of essentially independent random variables which may be useful for studying percolation on other trees with logarithmic height; for instance in this work we study also the case of regular trees.

1. Introduction

Consider a tree T_n of large but finite size $n \in \mathbb{N}$ and perform Bernoulli bond-percolation with parameter $p_n \in [0,1]$ that depends on the size of the graph. This means that we remove each edge in T_n with probability $1-p_n$, independently of the other edges, inducing a partition of the set of vertices into connected clusters. In particular, we are interested in the supercritical percolation regime, in the sense that with high probability, there exists a giant cluster, that is of size comparable to that of the entire tree. Bertoin (2013) established for several families of trees with n vertices that the supercritical regime corresponds to percolation parameters of the form $1-p_n = c/\ell(n) + o(1/\ell(n))$ as $n \to \infty$, where $c > 0$ is fixed and $\ell(n)$ is an estimate of the height of a typical vertex in the tree

Received by the editors October 4th, 2020; accepted February 17th, 2022.

2010 *Mathematics Subject Classification.* 60C05; 60F05; 60K35; 68P05; 05C05; 05C80.

Key words and phrases. split trees; random trees, percolation; giant cluster; fluctuations.

This work was supported by the Knut and Alice Wallenberg Foundation, a grant from the Swedish Research Council and The Swedish Foundations' starting grant from Ragnar Söderbergs Foundation.

structure[1]. More precisely, Bertoin (2013) showed that under the previous regime the size Γ_n of the cluster containing the root satisfies $\lim_{n\to\infty} n^{-1}\Gamma_n = \Gamma(c)$ in law, for some random variable $\Gamma(c) \not\equiv 0$. This includes, important families of random trees with logarithmic height, such as random recursive trees, preferential attachment trees, binary search trees where it is well-known that $\ell(n) = \ln n$; see Drmota (2009), Durrett (2010, Section 4.4). In those cases the random variable $\Gamma(c)$ is a constant; see Bertoin (2014b), Bertoin and Uribe Bravo (2015), Berzunza (2015). A different example is the Cayley tree where $\ell(n) = \sqrt{n}$ and $\Gamma(c)$ is not a constant; see Pitman (1999).

More recently, some authors have considered analysing the fluctuations of the size of the largest percolation cluster as $n \to \infty$ for different families of trees with logarithmic height; see Schweinsberg (2012) and Bertoin (2014a) for random recursive trees, Berzunza (2015) for m-ary random increasing trees (these include binary search trees) and preferential attachment trees. The motivation stems from the feature that the size of the giant cluster resulting from supercritical bond percolation on those trees has non-Gaussian fluctuations. Instead, they are described by an infinitely divisible distribution that belongs to the class of stable (asymmetric) Cauchy laws. This contrasts with analogous results on other random graphs where the asymptotic normality of the size of the giant clusters on supercritical percolation is established. We refer for instance to the works of Stepanov (1970), Bollobás and Riordan (2012) and Seierstad (2013).

The main purpose of this work is to investigate analogously the case of random split trees which were introduced by Devroye (1999). The class of random split trees includes many families of trees that are frequently used in algorithm analysis, e.g., binary search trees (Hoare (1962)), m-ary search trees (Pyke (1965)), quad trees (Finkel and Bentley (1974)), median-of-$(2k+1)$ trees (Walker and Wood (1976)), fringe-balanced trees (Devroye (1993)), digital search trees (Coffman and Eve (1970)) and random simplex trees (Devroye (1999, Example 5)). Informally, a random split tree T_n^{sp} of "size" (or cardinality) n is constructed as follows. Consider a rooted infinite b-ary tree with $b \in \mathbb{N}$ and where each vertex is a bucket of finite capacity $s \in \mathbb{N}$. We place n balls at the root, and the balls individually trickle down the tree in a random fashion until no bucket is above capacity. Each vertex draws a split vector $\mathcal{V} = (V_1, \ldots, V_b)$ from a common distribution, where V_i describes the probability that a ball passing through the vertex continues to the i-th child. A precise description of this algorithm is given in Section 1.1. Finally, any vertex u such that the sub-tree rooted as u contains no balls is then removed, and we consider the resulting tree T_n^{sp}. An important peculiarity of the split tree T_n^{sp} is that the number of vertices is random in general which makes the study of split trees usually challenging. It must also be pointed out that later we assume that $b < \infty$. However, we believe that our approach can be applied to cases when $b = \infty$ with a little extra effort. The case $b = \infty$ includes uniform recursive trees and preferential attachment trees for which recently Janson (2019) has shown that they can be viewed as special split trees.

Loosely speaking, our main result shows that in the supercritical percolation regime the size of the giant cluster has also non-Gaussian fluctuations where the "size" of T_n^{sp} can be defined as either the number of vertices or the number of balls. We then show that the supercritical regime corresponds to $1 - p_n = c/\ln n$ with $c > 0$ fixed which agrees with the fact that split trees belong to the family of trees with logarithmic height; see Devroye (1999). Essentially, this is Bertoin (2013) criterion. Then, our main contribution establishes that the fluctuations of the "size" (either number of vertices or balls) of the giant cluster are described by an infinitely divisible distribution, the so-called continuous Luria-Delbrück law. Finally, we show that the approach developed in this work may be useful for studying percolation on other classes of trees, such as for instance regular trees (see Section 5 below).

We next introduce formally the family of random split trees and relevant background, which will enable us to state our main result in Section 1.2.

[1]For two sequences of real numbers $(A_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ such that $B_n > 0$, we write $A_n = o(B_n)$ if $\lim_{n\to\infty} A_n/B_n = 0$. We also write $A_n = O(B_n)$ if $\limsup_{n\to\infty} |A_n|/B_n < \infty$

1.1. *Random split trees.* In this section, we introduce the split tree model with parameters b, s, s_0, $s_1, \mathcal{V}$ and n introduced by Devroye (1999). Some of the parameters are the branch factor $b \in \mathbb{N}$, the vertex capacity $s \in \mathbb{N}$, and the number of balls (or cardinality) $n \in \mathbb{N}$. The additional integers s_0 and s_1 are needed to describe the ball distribution process. They satisfy the inequalities

$$0 < s, \quad 0 \leq s_0 \leq s, \quad 0 \leq bs_1 \leq s + 1 - s_0. \tag{1.1}$$

The so-called random split vector $\mathcal{V} = (V_1, \dots, V_b)$ is a random non-negative vector with $\sum_{i=1}^{b} V_i = 1$ and $V_i \geq 0$, for $i = 1, \dots, b$.

Consider an infinite rooted b-ary tree $\mathbb{T}$, i.e., every vertex has b children. We view each vertex of $u \in \mathbb{T}$ as a bucket with capacity s and assign to it an independent copy $\mathcal{V}_u = (V_{u,1}, \dots, V_{u,b})$ of the random split vector $\mathcal{V}$. The split tree T_n^{sp} is constructed by distributing n balls among the vertices of $\mathbb{T}$. For a vertex u, let n_u be the number of balls stored in the sub-tree rooted at u. The tree T_n^{sp} is then defined as the largest sub-tree of $\mathbb{T}$ such that $n_u > 0$ for all $u \in T_n^{\text{sp}}$. Let $u_1, \dots, u_b$ be the child vertices of u. Conditioning on n_u and $\mathcal{V}_u$, if $n_u \leq s$, then $n_{u_i} = 0$ for all $i = 1, \dots, b$; if $n_u > s$, then the cardinalities $(n_{u_1}, \dots, n_{u_b})$ of the b sub-trees rooted at $u_1, \dots, u_b$ are distributed as

$$\text{Mult}(n_u - s_0 - bs_1, V_{u,1}, \dots, V_{u,b}) + (s_1, \dots, s_1),$$

where Mult denotes the multinomial distribution, and b, s, s_0, s_1 are integers satisfying (1.1).

It would be convenient to recall one more equivalent description of T_n^{sp} where one inserts data items into an initially empty data structure $\mathbb{T}$. Let $C(u)$ denote the number of balls in vertex u, initially setting $C(u) = 0$ for all u. We call u a leaf if $C(u) > 0$ and $C(v) = 0$ for all children v of u, and internal if $C(v) > 0$ for some strict descendant v of u. Then T_n^{sp} is constructed recursively by distributing n balls one at time to generate a subset of vertices of $\mathbb{T}$. The balls are labelled using the set $\{1, 2, \dots, n\}$ in the order of insertion. The j-th ball is added by the following procedure.

(1) Insert j to the root.
(2) While j is at an internal vertex $u \in \mathbb{T}$, choose child i with probability $V_{u,i}$ and move j to child i.
(3) If j is at a leaf u with $C(u) < s$, then j stays at u and $C(u)$ increases by 1. If j is at a leaf with $C(u) = s$, then the balls at u are distributed among u and its children as follows. We select $s_0 \leq s$ of the balls uniformly at random to stay at u. Among the remaining $s+1-s_0$ balls, we uniformly at random distribute s_1 balls to each of the b children of u. Each of the remaining $s + 1 - s_0 - bs_1$ balls is placed at a child vertex chosen independently at random according to the split vector assigned to u. This splitting process is repeated for any child which receives more than s balls.

We stop once all n balls have been placed in $\mathbb{T}$ and obtain T_n^{sp} by deleting all vertices $u \in \mathbb{T}$ such that the sub-tree rooted at u contains no balls; an internal vertex of T_n^{sp} contains exactly s_0 balls, while a leaf contains a random number in $\{1, ..., s\}$. This description will be used in Appendix A.

Remark 1.1. The number of vertices N of T_n^{sp} is a random variable in general although the number of balls n is deterministic. This is one of the main challenges in the study of split trees.

Remark 1.2. Depending on the choice of the parameters, several important data structures may be modelled. For instance, the binary search trees where $b = 2$, $s = s_0 = 1$, $s_1 = 0$ and $\mathcal{V}$ is distributed as $(U, 1-U)$ for U a random variable uniform on $[0, 1]$. In this case $N = n$. Some other relevant (and more complicated) examples of split trees are m-ary search trees, median-of-$(2k+1)$ trees, quad trees, simplex trees; see the original work of Devroye (1999) for details.

Remark 1.3. We can and will assume without loss of generality that the components of the split vector $\mathcal{V}$ are identically distributed by the random permutations explained by Devroye (1999). In particular, $\mathbb{E}[V_1] = 1/b$.

Two quantities deeply related to the structure of split trees are

$$\mu := b\mathbb{E}[-V_1 \ln V_1] \quad \text{and} \quad \sigma^2 := b\mathbb{E}[V_1 \ln^2 V_1] - \mu^2. \tag{1.2}$$

Note that $\mu \in (0, \ln b)$ and $\sigma < \infty$. They were introduced first by Devroye (1999) to study the height of T_n^{sp} as the number of balls increases.

In the study of split trees, the following condition is often assumed as this is satisfied by all types of split trees used in applications:

Condition 1. *Assume that* $\mathbb{P}(V_1 = 1) = \mathbb{P}(V_1 = 0) = 0$.

In the present work, we use the so-called total path length of T_n^{sp} defined by $\Psi(T_n^{\mathrm{sp}}) := \sum_{i=1}^{n} D_n(i)$, where $D_n(j)$ denotes the height (or depth) of the ball labeled j when all n balls have been inserted in T_n^{sp}. Broutin and Holmgren (2012, Theorem 3.1) have shown that under Condition 1,

$$\mathbb{E}[\Psi(T_n^{\mathrm{sp}})] = \mu^{-1} n \ln n + \varpi(\ln n) n + o(n), \tag{1.3}$$

where $\varpi : \mathbb{R} \to \mathbb{R}$ is a continuous periodic function of period

$$d := \sup\{a \geq 0 : \mathbb{P}(\ln V_1 \in a\mathbb{Z}) = 1\}. \tag{1.4}$$

In particular, if the random variable $\ln V_1$ is non-lattice[2], then $d = 0$ and the function ϖ is a constant and we write $\varsigma \equiv \varpi$.

We point out that the proof of Broutin and Holmgren (2012, Theorem 3.1) is missing some details for the case when $\ln V_1$ is lattice. The issue there is that the convergence (24) in Broutin and Holmgren (2012) only holds when the distribution of $\ln V_1$ is non-lattice. Nevertheless, a close look to the proof of Broutin and Holmgren (2012, Lemma 4.2) and Lemma 1.7 (ii) below show that the result by Broutin and Holmgren (2012, Theorem 3.1) is correct also in the lattice case.

Remark 1.4. In binary search trees the function ϖ equals to $2\gamma - 4$ where γ is the Euler's constant; see Hibbard (1962). A similar result has been proven for random m-ary search trees (Mahmoud (1986)), quad trees (Neininger and Rüschendorf (1999)), the random median of a $(2k+1)$-tree (Rösler (2001)), tries, and Patricia tries (Bourdon (2001)).

An alternative notion of path length is the sum of all the heights of the vertices in T_n^{sp}, i.e., $\Upsilon(T_n^{\mathrm{sp}}) := \sum_{u \in T_n^{\mathrm{sp}}} d_n(u)$, where $d_n(u)$ denotes the height of the vertex $u \in T_n^{\mathrm{sp}}$. Recall that the height of a vertex is defined as the minimal number of edges of T_n^{sp} which are needed to connect it to the root.

Condition 2. *Suppose that* $\ln V_1$ *is non-lattice. Furthermore, for some* $\alpha > 0$ *and* $\varepsilon > 0$, $\mathbb{E}[N] = \alpha n + O(n(\ln n)^{-1-\varepsilon})$.

Assuming that Condition 2 holds, Broutin and Holmgren (2012, Corollary 5.1) showed that

$$\mathbb{E}[\Upsilon(T_n^{\mathrm{sp}})] = \alpha\mu^{-1} n \ln n + \zeta n + o(n), \quad \text{for some constant} \quad \zeta \in \mathbb{R}. \tag{1.5}$$

Remark 1.5. Holmgren (2012, Theorem 1.1) showed that if $\ln V_1$ is non-lattice, i.e., $d = 0$, then there exists a constant $\alpha > 0$ such that $\mathbb{E}[N] = \alpha n + o(n)$ and furthermore $Var(N) = o(n^2)$. However, this result is not enough to deduce (1.5) from (1.3) and the extra control in $\mathbb{E}[N]$ is needed; see Broutin and Holmgren (2012, Section 5.1). On the one hand, Condition 2 is satisfied in many interesting cases. For instance, it holds for m-ary search trees (Mahmoud and Pittel (1989)). Moreover, Flajolet et al. (2010) showed that for most tries (where $s = 1$ and $s_0 = 0$ and as long as $\ln V_1$ is non-lattice) Condition 2 holds. On the other hand, there are some special cases of random split trees that do not satisfy Condition 2. For instance, tries with a fixed split vector $(1/b, \dots, 1/b)$, in which case $\ln V_1$ is lattice with $d = b$.

[2]The random variable $\ln V_1$ is non-lattice when there is not $a \in \mathbb{R}$ such that $\ln V_1 \in a\mathbb{Z}$ almost surely. The constant d is called the span of the lattice when $d > 0$ and $\ln V_1$ is non-lattice when $d = 0$.

Remark 1.6. One can use Condition 2 to improve the result by Holmgren (2012, Theorem 1.1) and obtain that $Var(N) = o(n^2 \ln^{-2-2\varepsilon} n)$. We refer to Holmgren (2012, Theorem 1.1) and Holmgren (2012, Remark 3.1) for a proof.

Finally, we recall and extend some results by Holmgren (2012, Section 2) and Broutin and Holmgren (2012, Section 4.2) related to the application of renewal theory in the study of split-trees. For $k \geq 1$, set $S_k := \sum_{j=1}^{k} -\ln V_j'$ where $(V_j', j \geq 1)$ is a sequence of i.i.d. copies of V_1. Following the presentation by Holmgren (2012) (or Broutin and Holmgren (2012, Section 4.2)), for $k \geq 1$ and $t \in \mathbb{R}$, let $\vartheta_k(t) := b^k \mathbb{P}(S_1 \leq t)$ and define the renewal function

$$U(t) = \sum_{k=1}^{\infty} \vartheta_k(t).$$

Observe that $U(t) = 0$, for $t < 0$. For $t \in \mathbb{R}$, let $\vartheta(t) = \vartheta_1(t)$ and observe that U satisfies the following renewal equation

$$U(t) = \vartheta(t) + (U * \mathrm{d}\vartheta)(t), \quad \text{where} \quad (U * \mathrm{d}\vartheta)(t) = \int_0^t U(t-z)\mathrm{d}\vartheta(z), \quad \text{for } t \geq 0. \tag{1.6}$$

Lemma 1.7. *Suppose that Condition 1 holds. The renewal function U satisfies the following.*

(i) *If* $\ln V_1$ *is non-lattice, then*

$$U(t) = \left(\frac{1}{\mu} + o(1)\right) e^t, \quad \text{as } t \to \infty.$$

(ii) *If the distribution of* $\ln V_1$ *is lattice with span d defined in (1.4), then*

$$U(d\lfloor t \rfloor) = \left(\frac{d}{\mu}\frac{1}{1-e^{-d}} + o(1)\right) e^{d\lfloor t \rfloor}, \quad \text{as } t \to \infty.$$

Proof: Part (i) follows from Holmgren (2012, Lemma 2.1). To prove part (ii), we use the lattice version of the key renewal theorem. Observe that $\mathrm{d}\vartheta(t)$ is not a probability measure. Following Holmgren (2012) (or Broutin and Holmgren (2012, Section 4.2)), one can define another ("tilted") measure $\mathrm{d}\omega(t) = e^{-t}\mathrm{d}\vartheta(t)$ which indeed is a probability measure. Furthermore, $\mathrm{d}\omega(t)$ is lattice with period d. The renewal equation (1.6) can then be written as

$$\hat{U}(t) = \hat{\vartheta}(t) + (\hat{U} * \mathrm{d}\omega)(t), \quad \text{where} \quad \hat{U}(t) = e^{-t}U(t) \quad \text{and} \quad \hat{\vartheta}(t) = e^{-t}\vartheta(t),$$

for $t \geq 0$. On the other hand, $\sum_{k=0}^{\infty} \hat{\vartheta}(kd) = (1-e^{-d})^{-1}$. Therefore, (ii) follows from Asmussen (2003, Proposition 4.1, Chapter V). □

In Broutin and Holmgren (2012, Section 4.2), the second-order behaviour of the renewal function U is also studied. More precisely, Broutin and Holmgren (2012, Lemma 4.2) establishes that under Condition 1 (and even for degenerate V_1) one has that

$$\int_0^t e^{-z}\left(U(z) - \mu^{-1}e^z\right)\mathrm{d}z = \frac{\sigma^2 - \mu^2}{2\mu^2} - \mu^{-1} + \phi(t) + o(1), \quad \text{as } t \to \infty, \tag{1.7}$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is a continuous d-periodic function with d defined in (1.4). Moreover, if $d = 0$ then $\phi \equiv 0$; see Holmgren (2012, Corollary 2.2) for the non-lattice case.

1.2. *Main results.* In this section, we present the main results of this work. Let T_n^{sp} be a split tree with n balls. We then perform Bernoulli bond percolation with parameter

$$p_n = 1 - \frac{c}{\ln n}, \tag{1.8}$$

where $c > 0$ is fixed. We write $\hat{G}_n$ for the size, i.e., the number of balls, of the percolation cluster that contains the root. Our first result shows that this choice of the percolation parameter corresponds precisely to the supercritical regime we are interested in.

Lemma 1.8. *Suppose that Condition 1 holds. In the regime (1.8), we have that*

$$\lim_{n\to\infty} n^{-1}\hat{G}_n = e^{-\frac{c}{\mu}}, \qquad \text{in probability.}$$

Moreover, the root cluster is the unique giant component, i.e., $\lim_{n\to\infty} n^{-1}\hat{G}_n^{\mathrm{2nd}} = 0$ *in probability, where* $\hat{G}_n^{\mathrm{2nd}}$ *denotes the number of balls of the second largest percolation cluster.*

Alternatively, let G_n be the number of vertices in the root cluster. Then we have the similar result:

Lemma 1.9. *Suppose that Conditions 1 and 2 hold. In the regime (1.8), we have that*

$$\lim_{n\to\infty} n^{-1}G_n = \alpha e^{-\frac{c}{\mu}}, \qquad \text{in probability,} \tag{1.9}$$

where $\alpha > 0$ *is the constant in Condition 2. Moreover, the root cluster is the unique giant component, i.e.,* $\lim_{n\to\infty} n^{-1}G_n^{\mathrm{2nd}} = 0$ *in probability, where* G_n^{2nd} *denotes the number of vertices of the second largest percolation cluster.*

Lemma 1.8 and Lemma 1.9 are a consequence of the results of Bertoin (2013) which provides a simple characterization of tree families and percolation regimes which yield giant clusters; their proofs are given in Section 2. Lemma 1.8 and Lemma 1.9 can be viewed as the law of large numbers for the "size" of the giant cluster, and it is then natural to investigate the fluctuations of $\hat{G}_n$ and G_n. To give a precise statement, recall that a real-valued random variable Z has the so-called continuous Luria-Delbrück law[3] when its characteristic function is given by

$$\mathbb{E}\left[e^{itZ}\right] = \exp\left(-\frac{\pi}{2}|t| - it\ln|t|\right), \qquad t \in \mathbb{R}.$$

This distribution arises in limit theorems for sums of positive i.i.d. random variables in the domain of attraction of a completely asymmetric Cauchy process; see e.g., Geluk and de Haan (2000). In the context of percolation on large trees, it was observed first by Schweinsberg (2012) (see also Bertoin (2014a) for an alternative approach) in relation with the fluctuations of the size (number of vertices) of the giant cluster for supercritical percolation on random recursive trees. More precisely, let T_n^{rec} be a random recursive tree with n vertices and denote by G_n^{rec} the size (number of vertices) of the largest percolation cluster after performing percolation with parameter p_n as in (1.8); In Bertoin (2014b), it has been proven that this yields also to the supercritical regime in T_n^{rec}, i.e., $\lim_{n\to\infty} n^{-1}G_n^{\mathrm{rec}} = e^{-c}$ in probability. Then,

$$\left(n^{-1}G_n^{\mathrm{rec}} - e^{-c}\right)\ln n - ce^{-c}\ln\ln n \xrightarrow{d} -ce^{-c}(Z + \ln c),$$

where $\xrightarrow{d}$ means convergence in distribution as $n \to \infty$. More recently, Berzunza (2015) has shown for preferential attachment trees and m-ary random increasing trees (the latter includes the case of binary search trees) that the fluctuations of the size of the giant component in the percolation regime (1.8) are also described by the continuous Luria-Delbrück distribution.

On the other hand, the continuous Luria-Delbrück distribution has been further observed in several weak limit theorems for the number of cuts required to isolate the root of a tree; see the original work of Meir and Moon (1970). For random recursive tree (Drmota et al. (2009), Iksanov and Möhle (2007)), random binary search tree (Holmgren (2010b)) and split trees (Holmgren (2011)).

[3]The name of this distribution had its origin in a series of classic experiments in evolutionary biology pioneered by Luria and Delbrück (1943) in order to study "random mutation" versus "directed adaptation" in the context of bacteria becoming resistant to a previously lethal agent. We refer also to Möhle (2005).

We refer to Cai et al. (2019) and Cai and Holmgren (2019) for a generalization of the Meir and Moon cutting model where similar results appear.

We now state the central results of this work.

Theorem 1.10. *Suppose that Condition 1 holds and that* $\ln V_1$ *is non-lattice. As* $n \to \infty$*, there is the convergence in distribution*

$$\left(\frac{\hat{G}_n}{n} - e^{-\frac{c}{\mu}}\right)\ln n - \frac{c}{\mu}e^{-\frac{c}{\mu}}\ln\ln n \xrightarrow{d} -\frac{c}{\mu}e^{-\frac{c}{\mu}}\left(Z + \ln\left(\frac{c}{\mu}\right) + \varsigma\mu + \frac{(\mu^2-\sigma^2)(c+\mu)}{2\mu^2} - \gamma + 1\right),$$

where μ *and* σ^2 *are the constants defined in (1.2),* $\varpi \equiv \varsigma$ *(a constant) is defined in (1.3),* γ *is the Euler constant and the variable* Z *has the continuous Luria-Delbrück distribution.*

Similarly, we obtain that the fluctuations of G_n are also described by Z.

Theorem 1.11. *Suppose that Condition 1 and 2 hold. As* $n \to \infty$*, there is the convergence in distribution*

$$\left(\frac{G_n}{n} - \alpha e^{-\frac{c}{\mu}}\right)\ln n - \frac{c\alpha}{\mu}e^{-\frac{c}{\mu}}\ln\ln n \xrightarrow{d} -\frac{c\alpha}{\mu}e^{-\frac{c}{\mu}}\left(Z + \ln\left(\frac{c}{\mu}\right) + \frac{\zeta\mu}{\alpha} + \frac{(\mu^2-\sigma^2)(c+\mu)}{2\mu^2} - \gamma + 1\right),$$

where μ *and* σ^2 *are the constants defined in (1.2),* α *is defined in Condition 2,* ζ *is defined in (1.5),* γ *is the Euler constant and the variable* Z *has the continuous Luria-Delbrück distribution.*

We also show that Theorem 1.10 can essentially be extended to the case when $\ln V_1$ is lattice. Consider the following additional condition. Write $y = \lfloor y \rfloor + \{y\}$ for the decomposition of a real number y as the sum of its integer and fractional parts.

Condition 3. *Let* T_n^{sp} *be a split tree with cardinality* n *and span* $d > 0$ *defined in (1.4). Furthermore, suppose that* $\{d^{-1}\ln\ln n\} \to \varrho \in [0,1)$*, as* $n \to \infty$*.*

We introduce for every $\varrho \in [0,1)$ and $c, d, x > 0$,

$$\bar{\Xi}_{\varrho}^{c,d}(x) = \frac{c}{\mu}\frac{d}{1-e^{-d}}e^{d\lfloor \varrho - d^{-1}\ln x - d^{-1}c/\mu\rfloor - d\varrho},$$

where μ is the constant defined in (1.2). The function $\bar{\Xi}_{\varrho}^{c,d}$ decreases as $x \to \infty$ and it can be viewed as the tail of a measure $\Xi_{\varrho}^{c,d}$ on $(0,\infty)$. Note also that this measure fulfils the integral condition $\int_{(0,\infty)}(1 \wedge x^2)\Xi_{\varrho}^{c,d}(\mathrm{d}x) < \infty$. This enables us to introduce a Lévy process without negative jumps $Z_{\varrho}^{c,d} = (Z_{\varrho}^{c,d}(t))_{t\geq 0}$ with Laplace exponent

$$\Phi_{\varrho}^{c,d}(a) = \int_{(0,\infty)}(e^{-ax} - 1 + ax\mathbb{1}_{\{x<1\}})\Xi_{\varrho}^{c,d}(\mathrm{d}x),$$

i.e., $\mathbb{E}[e^{-aZ_{\varrho}^{c,d}(t)}] = e^{t\Phi_{\varrho}^{c,d}(a)}$, for $a \geq 0$.

Theorem 1.12. *Suppose that Condition 1 holds and that* T_n^{sp} *satisfies Condition 3. For any constant* $\theta > 0$*, as* $n \to \infty$*, there is the convergence in distribution*

$$\left(\frac{\hat{G}_n}{n} - e^{-\frac{c}{\mu}}\right)\ln n - \frac{c}{\mu}e^{-\frac{c}{\mu}}\ln\ln n + ce^{-\frac{c}{\mu}}\left(\varpi(\ln n) - \phi\left(\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right)\right)$$
$$\xrightarrow{d} -Z_{\varrho}^{c,d}(1) - \frac{c}{\mu}e^{-\frac{c}{\mu}}\left(\frac{c}{\mu} + \frac{(\mu^2-\sigma^2)(c+\mu)}{2\mu^2}\right),$$

where μ *and* σ^2 *are the constants defined in (1.2),* ϖ *is the function defined in (1.3),* ϕ *is the function defined in (1.7),* ϱ *is defined in Condition 3 and* γ *is the Euler constant.*

Remark 1.13. Following Bertoin (2013), we point out that Lemmas 1.8 and 1.9 still hold whenever the percolation parameter satisfies $p_n = 1 - c\ln^{-1} n + o(\ln^{-1} n)$, where $c > 0$ is fixed, which still falls in the supercritical regime; see Bertoin (2013, Theorem 1). However, to obtain similar results to those in Theorems 1.10, 1.11 and 1.12 one needs to know more information of the $o(\ln^{-1} n)$ term.

The constants appearing in our main results can be computed explicitly for some types of trees. For example, if T_n^{bst} is a binary search tree with n vertices (recall Remark 1.2), then $N = n$, $\alpha = 1$, $\mu = 1/2$, $\sigma^2 = 1/4$, $\zeta = \varsigma = 2\gamma - 4$ and $\phi \equiv 0$; see for example Hibbard (1962). Moreover, the result in Theorem 1.10 (or Theorem 1.11) applied to T_n^{bst} coincides with Berzunza (2015, Theorem 1.1). The value of the constant can also be computed, for instance, for quad trees or for m-ary search trees; we refer to Neininger and Rüschendorf (1999) and Mahmoud (1986), respectively, for details.

The approach used by Schweinsberg (2012) for recursive trees relies on its connection with the Bolthausen-Sznitman coalescent found by Goldschmidt and Martin (2005) and the estimation of the rate of decrease of the number of blocks in such coalescent process. The alternative approach of Bertoin (2014a) makes use of the special properties of recursive trees (namely the splitting property) and more specifically of a coupling due to Iksanov and Möhle (2007) connecting the Meir and Moon (1970) algorithm for the isolation of the root with a certain random walk in the domain of attraction of the completely asymmetric Cauchy process. This clearly fails for split-trees. On the other hand, the basic idea of Berzunza (2015) for the case of m-ary random increasing trees and preferential attachment trees is based in the close relation of these trees with Markovian branching processes and the dynamical incorporation of percolation as neutral mutations. Roughly speaking, this yields to the analysis of the asymptotic behaviour of branching processes subject to rare neutral mutations. The relationship between percolation on trees and branching process with mutations was first observed by Bertoin and Uribe Bravo (2015). Recently, Holmgren and Janson (2017) have shown that some kinds of split trees (but not all) can be related to genealogical trees of general age-dependent branching processes (or Crump-Mode-Jagers processes), for instance, m-ary search trees and median-of-$(2\ell+1)$ trees. Furthermore, Berzunza (2020) has proven the existence of a giant percolation cluster for appropriate regimes of such genealogical trees via a similar relationship with a general branching process with mutations. However, the branching processes with mutations in Berzunza (2020) is in general not Markovian due to the nature of the Crump-Mode-Jagers processes; see Jagers (1975). This makes the idea of Berzunza (2015) difficult to implement since there the Markov property is crucial. We thus have to use here a fairly different route.

The method used here is inspired in the original technique developed by Janson (2004) to study the number of cuts needed to isolate the root of complete binary trees with the cutting-down procedure of Meir and Moon (1970). Holmgren (2010b, 2011) has successfully extended this method to study the same quantity as in Janson (2004) for split trees. Informally speaking, we approximate $\hat{G}_n$ (resp. G_n) by the sum of the "sizes" of the percolation clusters of the sub-trees rooted at vertices that are at a distance around $\ln\ln n$ from the root. There are approximately $b^{\ln\ln n}$ clusters, but we only consider those that are still connected to the root of T_n^{sp} after performing percolation for the regime p_n as in (1.8). The number of balls (or number of vertices) between the root of T_n^{sp} and the the vertices at height $\ln\ln n$ is equal to $O(\ln n)$ and thus they do not contribute to the fluctuations of $\hat{G}_n$ (resp. G_n). We then analyse carefully the "sizes" of percolation clusters at distances close to $\ln\ln n$ from the root, and essentially, we view $\hat{G}_n$ (resp. G_n) as a sum of independent random variables. This will allow us to apply a classical limit theorem for the convergence of triangular arrays to get our main result. Therefore, we conclude that most of the random fluctuations can be explained by the "sizes" of percolation clusters at distances close to $\ln\ln n$ from the root of T_n^{sp} and that they are still connected to the root. Indeed, this phenomenon has also been observed by Bertoin (2014a, Section 3) who studied the fluctuations of the number of vertices at height $\ln\ln n$ which has been disconnected from the root in b-regular trees after performing supercritical percolation.

In this setting, the fluctuations are described by a Lévy process without negative jumps that also appears in Janson (2004).

The rest of this paper is organized as follows: We start by proving Lemma 1.8 and Lemma 1.9 in Section 2. In Section 3, we then focus on the proof of Theorem 1.10 and Theorem 1.12. Section 4 is devoted to the proof of Theorem 1.11 which follows essentially from Theorem 1.10. In Section 5, we briefly point out that the present approach also applies to study the fluctuations of the size of the giant cluster for percolation on regular trees. The appendices provide details on some technical results that are used in the proofs of the main result but that we decided to postpone for a better understanding of our approach. In particular, Appendix A is dedicated to investigate the asymptotic behaviour of distances between uniformly chosen vertices and uniformly chosen balls in T_n^{sp} which may be of independent interest.

2. Proof of Lemma 1.8 and Lemma 1.9

Lemma 1.8 and Lemma 1.9 are a merely consequence of the results of Bertoin (2013) after mild modifications.

Proof of Lemma 1.8: The result follows from exactly the same argument as the proof of Bertoin (2013, Corollary 1 and Proposition 1) by using Lemma A.2, Corollary A.6 in Appendix A and by taking into account that the size is defined as the number of balls instead of the number of vertices. □

Proof of Lemma 1.9: The result follows from Bertoin (2013, Corollary 1). Note that conditions $(\mathbf{H}_k)$ and $(\mathbf{H}'_k)$, for $k = 1, 2$, in Bertoin (2013, Corollary 1) are verified in Lemma A.3 and Corollary A.6 in Appendix A. Therefore, in the percolation regime (1.8), we have that $\lim_{n\to\infty} N^{-1}G_n = e^{-\frac{c}{\mu}}$, in probability. On the other hand, Conditions 1 and 2 imply that $\lim_{n\to\infty} N/n = \alpha$, in probability. This establishes (1.9) in Lemma 1.9. The uniqueness of the giant component follows from Bertoin (2013, Proposition 1) by noticing that the condition there is satisfied as a consequence of Lemma A.3 and Corollary A.6 in Appendix A, that is,

$$\lim_{n\to\infty} \frac{1}{\ln n}(d_n(u_1), d_n(u_1, u_2)) = (1/\mu, 2/\mu)\,, \qquad \text{in probability,}$$

where u_1, u_2 are two i.i.d. uniform random vertices in T_n^{sp}, $d_n(u_1)$ denotes the height of u_1 and $d_n(u_1, u_2)$ is the number of edges of T_n^{sp} which are needed to connect the root and the vertices u_1 and u_2. □

3. Proof of Theorem 1.10

This section is devoted to the proofs of Theorem 1.10 and Theorem 1.12 along the lines explained at the end of Section 1.2. The starting point is Lemma 3.1 where we estimate the number of balls of the percolation clusters of sub trees rooted at vertices that are around height $\ln\ln n$. We continue with Lemmas 3.2, 3.3 and 3.4 that allow us to approximate $\hat{G}_n$ as essentially a sum of independent random variables. Finally, we establish Theorem 3.5 that shows that the conditions of Kallenberg (2002, Theorem 15.28), a classical limit theorem for triangular arrays, are fulfilled which allow us to conclude with the proof of Theorem 1.10.

For a vertex $v \in T_n^{\mathrm{sp}}$ that is at height $d_n(v) = j$, it is not difficult to see from the definition of random split trees in Section 1.1 that conditioning on the split vectors, we have

$$\text{binomial}\left(n, \prod_{k=1}^{j} W_{v,k}\right) - sj \leq_{\mathrm{st}} n_v \leq_{\mathrm{st}} \text{binomial}\left(n, \prod_{k=1}^{j} W_{v,k}\right) + s_1 j, \tag{3.1}$$

where $\leq_{\mathrm{st}}$ denotes *stochastically dominated by and* $(W_{v,k}, k = 1, \dots, j)$ are i.i.d. random variables on $[0,1]$ given by the split vectors associated with the vertices in the unique path from v to the root; This property has been used by Devroye (1999) and Holmgren (2012). In particular $W_{v,k} = V_1$ in distribution. We deduce the following important estimates.

$$E[n_v] \leq n \prod_{k=1}^{j} \mathbb{E}[W_{v,k}] + s_1 j = nb^{-j} + s_1 j, \tag{3.2}$$

where we have used $\mathbb{E}[W_{v,k}] = \mathbb{E}[V_1] = 1/b$. Moreover,

$$E[n_v^2] \leq n^2 \prod_{k=1}^{j} \mathbb{E}[W_{v,k}^2] + n \left(\prod_{k=1}^{j} \mathbb{E}[W_{v,k}] - \prod_{k=1}^{j} \mathbb{E}[W_{v,k}^2] \right) + 2s_1 jn \prod_{k=1}^{j} \mathbb{E}[W_{v,k}] + s_1^2 j^2. \tag{3.3}$$

Note that $\mathbb{E}[W_{v,k}^2] = \mathbb{E}[V_1^2] < 1/b$.

We use the notation $\log_b x = \ln x / \ln b$ for the logarithm with base b of $x > 0$, and we write $m_n = \lfloor \beta \log_b \ln n \rfloor$ for some constant $\beta > -2/(1 + \log_b \mathbb{E}[V_1^2])$. We further assume that n is large enough such that $0 < m_n < \ln n$. For $1 \leq i \leq b^{m_n}$, let v_i be a vertex in T_n^{sp} at height m_n and let n_i be the number of balls stored at the sub-tree rooted at v_i. In particular, for an arbitrary $k \geq 0$,

$$E[n_i^2] = n^2 \mathbb{E}^{m_n}[V_1^2] + o(n^2 \ln^{-k} n). \tag{3.4}$$

We denote by $\hat{C}_{n,i}$ the number of balls of the sub-tree of T_n^{sp} rooted at v_i after Bernoulli bond-percolation with parameter p_n. Clearly, $(\hat{C}_{n,i}, 1 \leq i \leq b^{m_n})$ are conditionally independent random variables given $(n_i, 1 \leq i \leq b^{m_n})$. We write $\mathbb{E}_{n_i}[\hat{C}_{n,i}] \coloneqq \mathbb{E}[\hat{C}_{n,i} | n_i]$, i.e., it is the conditional expected value of $\hat{C}_{n,i}$ given n_i.

We use the notation $A_n = B_n + o_{\mathrm{p}}(f(n))$, where A_n and B_n are two sequences of real random variables and $f : \mathbb{N} \to (0, \infty)$ a function, to indicate that $(A_n - B_n)/f(n) \to 0$ in probability.

Lemma 3.1. *Suppose that Condition 1 is fulfilled. For $1 \leq i \leq b^{m_n}$, we have that*

$$\mathbb{E}_{n_i}[\hat{C}_{n,i}] = n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - c\frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o\left(\frac{n_i}{\ln n}\right),$$

where $\varpi : \mathbb{R} \to \mathbb{R}$ is the function in (1.3).

Proof: For $1 \leq i \leq b^{m_n}$, let T_i be the sub-tree of T_n^{sp} rooted at the vertex v_i at height m_n. Let b_i be an uniformly chosen ball in T_i. Let $D_{n_i}(b_i)$ be the height of b_i in the sub-tree T_i. We use the following observation made by Bertoin (2013, Proof of Theorem 1),

$$\mathbb{E}_{n_i}\left[n_i^{-1}\hat{C}_{n,i}\right] = \mathbb{E}_{n_i}\left[p_n^{D_{n_i}(b_i)}\right]. \tag{3.5}$$

In words, the left-hand side can be interpreted as the probability that b_i belongs to the percolation cluster containing the root of T_i, i.e., v_i, while the right-hand side can be interpreted as the probability that no edge has been removed in the path between b_i and v_i.

We assume for a while that

$$\begin{aligned}\mathbb{E}_{n_i}&\left[p_n^{D_{n_i}(b_i)}\right] \\ &= \mathbb{E}_{n_i}\left[p_n^{\frac{\ln n_i}{\mu}}\left(1 + \left(D_{n_i}(b_i) - \frac{\ln n_i}{\mu}\right)\ln p_n + \frac{1}{2}\left(D_{n_i}(b_i) - \frac{\ln n_i}{\mu}\right)^2 \ln^2 p_n\right)\right] + o\left(\frac{1}{\ln n}\right).\end{aligned} \tag{3.6}$$

By our assumption (1.8) in the percolation parameter,

$$\ln p_n = -\frac{c}{\ln n} + o\left(\frac{1}{\ln n}\right) \quad \text{and} \quad p_n^{\frac{\ln n_i}{\mu}} = e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \frac{c^2}{2\mu}\frac{\ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o\left(\frac{1}{\ln n}\right). \tag{3.7}$$

We have used that $\ln n_i \leq \ln n$. Then it follows from Lemma A.2 (i)-(ii) in Appendix A and a couple of lines of calculations that

$$\mathbb{E}_{n_i}\left[p_n^{D_{n_i}(b_i)}\right] = e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3}\frac{\ln n_i}{\ln^2 n}e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - c\frac{\varpi(\ln n_i)}{\ln n}e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o\left(\frac{1}{\ln n}\right).$$

Therefore, the result in Lemma 3.1 follows from the identity (3.5) and the above estimation.

Now, we focus on establishing (3.6). From the inequality

$$\left|p_n^{D_{n_i}(b_i)} - p_n^{\frac{\ln n_i}{\mu}}\left(1 + \left(D_{n_i}(b_i) - \frac{\ln n_i}{\mu}\right)\ln p_n + \frac{1}{2}\left(D_{n_i}(b_i) - \frac{\ln n_i}{\mu}\right)^2 \ln^2 p_n\right)\right| \leq \left|\left(D_{n_i}(b_i) - \frac{\ln n_i}{\mu}\right)\ln p_n\right|^3,$$

we conclude that it is enough to show that

$$\mathbb{E}_{n_i}\left[\left|\left(D_{n_i}(b_i) - \frac{\ln n_i}{\mu}\right)\ln p_n\right|^3\right] = o\left(\frac{1}{\ln n}\right)$$

in order to obtain (3.6). But this follows from Lemma A.2 (iii) in Appendix A and (3.7). □

Let $\eta_{n,i}$ be the total number of edges on the branch from v_i to the root which have been deleted after percolation with parameter p_n. The random variable $\eta_{n,i}$ has the binomial distribution with parameters $(m_n, 1-p_n)$. But the random variables $(\eta_{n,i}, 1 \leq i \leq b^{m_n})$ are not independent. On the other hand, $\eta_{n,i} = 0$ if and only if the vertex v_i is still connected to the root.

Lemma 3.2. *Suppose that Condition 1 is fulfilled. We have for $\beta > -2/(1+\log_b \mathbb{E}[V_1^2])$ that*

$$\hat{G}_n = \sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}[\hat{C}_{n,i}]\mathbb{1}_{\{\eta_{n,i}=0\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

Proof: We denote by $\hat{C}_{n,0}$ the number of balls in the vertices of T_n^{sp} at height less or equal to $m_n - 1$ that are connected to the root after percolation with parameter p_n. Then, it should be plain that

$$\hat{G}_n = \hat{C}_{n,0} + \sum_{i=1}^{b^{m_n}} \hat{C}_{n,i}\mathbb{1}_{\{\eta_{n,i}=0\}}.$$

The sequences of random variables $(\eta_{n,i}, 1 \leq i \leq b^{m_n})$ and $(\hat{C}_{n,i}, 1 \leq i \leq b^{m_n})$ are independent. Furthermore, the sequence of random variables $(\eta_{n,i}, 1 \leq i \leq b^{m_n})$ and $(n_i, 1 \leq i \leq b^{m_n})$ are also independent. Let $\mathcal{F}_n$ be the σ-field generated by $(\eta_{n,i}, 1 \leq i \leq b^{m_n})$ and $(n_i, 1 \leq i \leq b^{m_n})$. Note also that $\mathbb{E}[\hat{C}_{n,i}|\mathcal{F}_n] = \mathbb{E}_{n_i}[\hat{C}_{n,i}]$. By conditioning on the σ-field $\mathcal{F}_n$ and taking expectation, we obtain that

$$\mathbb{E}\left[\left(\hat{G}_n - \hat{C}_{n,0} - \sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}[\hat{C}_{n,i}]\mathbb{1}_{\{\eta_{n,i}=0\}}\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}\left[\left(\hat{C}_{n,i} - \mathbb{E}_{n_i}[\hat{C}_{n,i}]\right)^2\right]\mathbb{1}_{\{\eta_{n,i}=0\}}\right]$$
$$= \sum_{i=1}^{b^{m_n}} \mathbb{E}\left[\left(\hat{C}_{n,i} - \mathbb{E}_{n_i}[\hat{C}_{n,i}]\right)^2\right]\mathbb{P}\left(\eta_{n,i}=0\right).$$

Since $\mathbb{P}(\eta_{n,i}=0) \leq 1$ and $\mathbb{E}[(\hat{C}_{n,i} - \mathbb{E}_{n_i}[\hat{C}_{n,i}])^2] \leq 2\mathbb{E}[n_i^2]$, because $\hat{C}_{n,i} \leq n_i$, we deduce that

$$\mathbb{E}\left[\left(\hat{G}_n - \hat{C}_{n,0} - \sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}[\hat{C}_{n,i}]\mathbb{1}_{\{\eta_{n,i}=0\}}\right)^2\right] \leq 2\sum_{i=1}^{b^{m_n}} \mathbb{E}[n_i^2].$$

Since $\beta > -2/(1+\log_b \mathbb{E}[V_1^2])$, we obtain from the estimate (3.4) that

$$\mathbb{E}\left[\left(\hat{G}_n - \hat{C}_{n,0} - \sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}[\hat{C}_{n,i}]\mathbb{1}_{\{\eta_{n,i}=0\}}\right)^2\right] = o\left(\frac{n^2}{\ln^2 n}\right).$$

The above implies together with Chebyshev's inequality that

$$\hat{G}_n = \hat{C}_{n,0} + \sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}[\hat{C}_{n,i}]\mathbb{1}_{\{\eta_{n,i}=0\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

Finally, the statement follows easily after noticing that $0 \le \hat{C}_{n,0} < b^{m_n+1} = o\left(\frac{n}{\ln n}\right)$. □

Next, we combine Lemma 3.1 and 3.2.

Lemma 3.3. *Suppose that Condition 1 is fulfilled. We have for $\beta > -2/(1+\log_b \mathbb{E}[V_1^2])$ that*

$$\hat{G}_n = -e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} n_i \mathbb{1}_{\{\eta_{n,i}\ge 1\}} + \sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - ce^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} - \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3} e^{-\frac{c}{\mu}} \frac{n}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

where $\varpi : \mathbb{R} \to \mathbb{R}$ is the function in (1.3).

Proof: The two sequences of random variables $(\eta_{n,i}, 1 \le i \le b^{m_n})$ and $(n_i, 1 \le i \le b^{m_n})$ are independent. Recall that the random variable $\eta_{n,i}$ has the binomial distribution with parameters $(m_n, 1-p_n)$. Hence

$$1 - \mathbb{P}(\eta_{n,i} = 0) = \mathbb{P}(\eta_{n,i} \ge 1) = 1 - p_n^{m_n} = O\left(\frac{\ln\ln n}{\ln n}\right). \tag{3.8}$$

Since $\sum_{i=1}^{b^{m_n}} n_i \le n$ and $\mathbb{P}(\eta_{n,i} = 0) \le 1$, we obtain that

$$\mathbb{E}\left[\sum_{i=1}^{b^{m_n}} \frac{n_i}{\ln n} \mathbb{1}_{\{\eta_{n,i}=0\}}\right] = \frac{1}{\ln n} \sum_{i=1}^{b^{m_n}} \mathbb{E}[n_i]\mathbb{P}(\eta_{n,i} = 0) \le \frac{n}{\ln n}.$$

Thus Lemma 3.1 and Lemma 3.2 imply that

$$\hat{G}_n = \sum_{i=1}^{b^{m_n}} \left(n_i - \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3} \frac{n_i \ln n_i}{\ln^2 n} - c\frac{n_i \varpi(\ln n_i)}{\ln n}\right) e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.9}$$

By the estimation (3.8) and the fact that $\sum_{i=1}^{b^{m_n}} n_i \le n$, we get that

$$\mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} - \sum_{i=1}^{b^{m_n}} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}}\right|\right] \le \frac{1}{\ln n} \sum_{i=1}^{b^{m_n}} \mathbb{E}[n_i]\mathbb{P}(\eta_{n,i} \ge 1) = o\left(\frac{n}{\ln n}\right)$$

and

$$\mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} - \sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}}\right|\right] \le \frac{K}{\ln n} \sum_{i=1}^{b^{m_n}} \mathbb{E}[n_i]\mathbb{P}(\eta_{n,i} \ge 1) = o\left(\frac{n}{\ln n}\right),$$

for some constant $K > 0$ such that $|\varpi(x)| \leq K$ for $x \in \mathbb{R}$; recall that ϖ in (1.3) is a continuous function with period $d \geq 0$. The previous two estimates together with Markov's inequality imply that

$$\sum_{i=1}^{b^{m_n}} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} = \sum_{i=1}^{b^{m_n}} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right), \tag{3.10}$$

and

$$\sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} = \sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.11}$$

For large enough $k \geq 1$,

$$\sum_{i=1}^{b^{m_n}} n_i \mathbb{1}_{\{n_i \leq n b^{-k m_n}\}} \leq b^{-m_n(k-1)} n = o\left(\frac{n}{\ln^{k-1} n}\right).$$

Then, by using the inequality $|e^{-x} - e^{-y}| \leq |x-y|$ for $x, y \in \mathbb{R}_+$, we have that

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}}\right|\right] &\leq \frac{c}{\mu}\frac{1}{\ln n}\mathbb{E}\left[\sum_{i=1}^{b^{m_n}} n_i(\ln n - \ln n_i)\right] \\
&= \frac{c}{\mu}\frac{1}{\ln n}\mathbb{E}\left[\sum_{i=1}^{b^{m_n}} n_i(\ln n - \ln n_i)\mathbb{1}_{\{n_i > n b^{-k m_n}\}}\right] + o\left(\frac{n}{\ln^{k-1} n}\right) \\
&= O\left(\frac{n \ln\ln n}{\ln n}\right),
\end{aligned} \tag{3.12}$$

where we have used that $\sum_{i=1}^{b^{m_n}} n_i \leq n$ in order to obtain the last estimation. The above implies

$$\frac{1}{\ln n}\sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} = e^{-\frac{c}{\mu}} \frac{1}{\ln n}\sum_{i=1}^{b^{m_n}} n_i + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.13}$$

Similarly, we deduce from (3.8) and (3.12)

$$\begin{aligned}
\mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}\geq 1\}} - \sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}} \mathbb{1}_{\{\eta_{n,i}\geq 1\}}\right|\right] &\leq \mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}}\right|\right] \mathbb{P}(\eta_{n,1} \geq 1) \\
&= o\left(\frac{n}{\ln n}\right),
\end{aligned}$$

$$\mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \frac{1}{\ln n}\sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}}\right|\right] = o\left(\frac{n}{\ln n}\right)$$

and

$$\mathbb{E}\left[\left|\sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}}\right|\right] = o\left(\frac{n}{\ln n}\right).$$

As a consequence, the previous three estimates and the Markov's inequality imply that

$$\sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}\geq 1\}} = e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} n_i \mathbb{1}_{\{\eta_{n,i}\geq 1\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right), \tag{3.14}$$

$$\sum_{i=1}^{b^{m_n}} \frac{n_i \ln n_i}{\ln^2 n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} = e^{-\frac{c}{\mu}} \frac{1}{\ln n}\sum_{i=1}^{b^{m_n}} n_i + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right), \tag{3.15}$$

and

$$\sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} = e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.16}$$

Plugging the estimations (3.10), (3.11), (3.13), (3.14), (3.15) and (3.16) into the expression in (3.9) yields that

$$\begin{aligned}\hat{G}_n = &-e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} n_i \mathbb{1}_{\{\eta_{n,i}\geq 1\}} + \sum_{i=1}^{b^{m_n}} n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \\ &- \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3} e^{-\frac{c}{\mu}} \frac{1}{\ln n} \sum_{i=1}^{b^{m_n}} n_i - ce^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} \frac{n_i \varpi(\ln n_i)}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right);\end{aligned}$$

note also that $\mathbb{1}_{\{\eta_{n,i}\geq 1\}} = 1 - \mathbb{1}_{\{\eta_{n,i}=0\}}$. Finally, our claim in Lemma 3.3 follows by showing that

$$\frac{1}{\ln n} \sum_{i=1}^{b^{m_n}} n_i = \frac{n}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.17}$$

Note that $\sum_{i=1}^{b^{m_n}} n_i = n - \hat{C}(n)$, where $\hat{C}(n)$ denotes the number of balls of the vertices of T_n^{sp} at distance less or equal to $m_n - 1$ from the root. Since $0 \leq \hat{C}(n) < \max(s, s_0) b^{m_n+1} = o(n)$, we deduce (3.17). □

We refine the result of Lemma 3.3.

Lemma 3.4. *Suppose that Condition 1 is fulfilled. We have for $\beta > -2/(1 + \log_b \mathbb{E}[V_1^2])$ that*

$$\begin{aligned}\hat{G}_n = &-e^{-\frac{c}{\mu}} \sum_{1\leq d_n(v)\leq m_n} n_v \varepsilon_v + \sum_{d_n(v)=m_n} n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} \\ &- ce^{-\frac{c}{\mu}} \sum_{d_n(v)=m_n} \frac{n_v \varpi(\ln n_v)}{\ln n} - \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3} e^{-\frac{c}{\mu}} \frac{n}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).\end{aligned}$$

where $\varpi : \mathbb{R} \to \mathbb{R}$ is the function in (1.3) and $(\varepsilon_v, 1 \leq d_n(v) \leq m_n)$ is a sequence of i.i.d. Bernoulli random variables with parameter $1 - p_n$.

Proof: Our claim follows from Lemma 3.3 by showing that

$$e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} n_i \mathbb{1}_{\{\eta_{n,i}\geq 1\}} = e^{-\frac{c}{\mu}} \sum_{1\leq d_n(v)\leq m_n} n_v \varepsilon_v + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.18}$$

Recall that the sequences of random variables $(\eta_{n,i}, 1 \leq i \leq b^{m_n})$ and $(n_i, 1 \leq i \leq b^{m_n})$ are independent. It should be obvious that

$$\mathbb{E}\left[e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} n_i \mathbb{1}_{\{\eta_{n,i}\geq 1\}}\right] = (1 - p_n^{m_n})\, e^{-\frac{c}{\mu}} \sum_{i=1}^{b^{m_n}} \mathbb{E}[n_i]. \tag{3.19}$$

Next consider the vertices $v_{i,0}, v_{i,1}, \dots, v_{i,m_n} = v_i$ along the path from the root $v_{i,0}$ of T_n^{sp} to the vertex v_i at height m_n. For $j = 1, \dots, m_n$, we associate to each consecutive pair of vertices $(v_{i,j-1}, v_{i,j})$ the edge that is between them (where $v_{i,j}$ is a vertex at height j on T_n^{sp}). Define the event $E_{i,j} \coloneqq \{\text{the edge } (v_{i,j-1}, v_{i,j}) \text{ has been removed after percolation}\}$ and write $\varepsilon_{i,j} \coloneqq \mathbb{1}_{E_{i,j}}$. So, $(\varepsilon_{i,j}, 1 \leq j \leq m_n)$ is a sequence of i.i.d. Bernoulli random variables with parameter $1 - p_n$ and

$$\eta_{n,i} = \sum_{j=1}^{m_n} \varepsilon_{i,j}. \tag{3.20}$$

Then

$$\mathbb{E}\left[e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i\eta_{n,i}\right] = m_n\left(1-p_n\right)e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}}\mathbb{E}\left[n_i\right]. \tag{3.21}$$

Since

$$e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i\mathbb{1}_{\{\eta_{n,i}\geq 1\}} \leq e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i\eta_{n,i},$$

we deduce from (3.19) and (3.21) that

$$\mathbb{E}\left[e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i\eta_{n,i} - e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i\mathbb{1}_{\{\eta_{n,i}\geq 1\}}\right] \leq \left(m_n(1-p_n)-(1-p_n^{m_n})\right)e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}}\mathbb{E}[n_i] = o\left(\frac{n}{\ln n}\right),$$

where we have used that $\sum_{i=1}^{b^{m_n}} n_i \leq n$ and our assumption (1.8). Therefore, the identity (3.20) implies that

$$e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i\mathbb{1}_{\{\eta_{n,i}\geq 1\}} = e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}}\sum_{j=1}^{m_n} n_i\varepsilon_{i,j} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right). \tag{3.22}$$

Finally, let $P(v_i)$ denote the unique path from the root $v_{i,0}$ of T_n^{sp} to v_i, i.e., the unique sequence of vertices $v_{i,0}, v_{i,1}, \dots, v_{i,m_n} = v_i$. For $v = v_{i,j} \in P(v_i)\setminus\{v_{i,0}\}$, write ε_v instead of $\varepsilon_{i,j}$. Note that

$$\begin{aligned} e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}}\sum_{j=1}^{m_n} n_i\varepsilon_{i,j} &= e^{-\frac{c}{\mu}}\sum_{i=1}^{b^{m_n}} n_i \sum_{v\in P(v_i)\setminus\{v_{i,0}\}}\varepsilon_v = e^{-\frac{c}{\mu}}\sum_{1\leq d_n(v)\leq m_n}\varepsilon_v\sum_{i:v\in P(v_i)\setminus\{v_{i,0}\}} n_i \\ &= e^{-\frac{c}{\mu}}\sum_{1\leq d_n(v)\leq m_n} n_v\varepsilon_v + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right), \end{aligned} \tag{3.23}$$

because $n_v - sb^{m_n} \leq \sum_{i:v\in P(v_i)\setminus\{v_{i,0}\}} n_i \leq n_v$.

Therefore, the estimation (3.18) follows by combining (3.22) and (3.23). □

Following the idea of Janson (2004) and subsequently used by Holmgren (2010b, 2011) (where the number of random cuts required to isolate the root of a tree was studied), we express $\hat{G}_n$ as a sum of triangular arrays. We write

$$\xi_v \coloneqq e^{-\frac{c}{\mu}}\frac{\ln n}{n}n_v\varepsilon_v, \quad \text{for} \quad v\in T_n^{\mathrm{sp}} \quad \text{such that} \quad d_n \leq m_n, \tag{3.24}$$

where $(\varepsilon_v, 1\leq d_n(v)\leq m_n)$ is a sequence of i.i.d. Bernoulli random variables with parameter $1-p_n$. We also write $\xi'_i \coloneqq -\alpha_n/n$ for $i\in\mathbb{N}$, where

$$\begin{aligned} \alpha_n \coloneqq \frac{\ln n}{n}&\sum_{d_n(v)=m_n} n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - ce^{-\frac{c}{\mu}}\sum_{d_n(v)=m_n}\frac{n_v\varpi(\ln n_v)}{n} \\ &- e^{-\frac{c}{\mu}}\ln n - \frac{c}{\mu}e^{-\frac{c}{\mu}}\ln\ln n + ce^{-\frac{c}{\mu}}\varpi(\ln n) - ce^{-\frac{c}{\mu}}\phi\left(\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) - \frac{c^2\mu^2-c^2\sigma^2}{2\mu^3}e^{-\frac{c}{\mu}} \end{aligned}$$

for any constant $\theta > 0$. By normalizing $\hat{G}_n$, Lemma 3.4 gives that

$$\begin{aligned} \left(n^{-1}\hat{G}_n - e^{-\frac{c}{\mu}}\right)\ln n - c\mu^{-1}e^{-\frac{c}{\mu}}\ln\ln n + ce^{-\frac{c}{\mu}}\left(\varpi(\ln n) - \phi\left(\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right)\right) \\ = -\sum_{1\leq d_n(v)\leq m_n}\xi_v - \sum_{i=1}^{n}\xi'_i + o_{\mathrm{p}}(1). \end{aligned}$$

Recall that the cardinalities $(n_v, 1 \leq d_n(v) \leq m_n)$ are not independent random variables and thus the sequence $(\xi_v, 1 \leq d_n(v) \leq m_n) \cup (\xi_i', i \in \mathbb{N})$ is not a triangular array. However, conditional on $\mathcal{F}_{m_n}$, the σ-field generated by $(n_v, 1 \leq d_n(v) \leq m_n)$, the sequence $(\xi_v, 1 \leq d_n(v) \leq m_n) \cup (\xi_i', i \in \mathbb{N})$ is a triangular array where $(\xi_i', i \in \mathbb{N})$ is a deterministic sequence.

Finally, the proofs of Theorem 1.10 and Theorem 1.12 are going to be completed via a classical theorem for convergence of sums of triangular arrays to infinitely divisible distributions; see e.g. Kallenberg (2002, Theorem 15.28). In this direction, we need the following result. For the sake of simplicity, we introduce the following notation. For any constants $\theta, x > 0$,

$$\Delta_{n,1} := \sum_{1 \leq d_n(v) \leq m_n} \mathbb{P}(\xi_v \geq x | \mathcal{F}_{m_n}), \qquad \Delta_{n,2} := \sum_{1 \leq d_n(v) \leq m_n} \mathbb{E}\left[\xi_v \mathbb{1}_{\{\xi_v \leq \theta\}} | \mathcal{F}_{m_n}\right] - \alpha_n,$$

$$\text{and} \qquad \Delta_{n,3} := \sum_{1 \leq d_n(v) \leq m_n} Var\left(\xi_v \mathbb{1}_{\{\xi_v \leq \theta\}} | \mathcal{F}_{m_n}\right).$$

For $\theta > 0$ and $x \geq 0$, we also define the function

$$\psi_\theta(x) = 1 - \frac{\theta x}{1 - e^{-x}} e^{x\lfloor \varrho - x^{-1}\ln\theta - x^{-1}c/\mu \rfloor - x\varrho + c/\mu}$$

such that $\psi_\theta(0) = 0$.

Theorem 3.5. *Recall that $m_n = \lfloor \beta \log_b \ln n \rfloor$. Suppose that Condition 1 holds. Furthermore, if $\ln V_1$ is lattice with span d defined in (1.4), we also assume that Condition 3 holds for some $\varrho \in [0,1)$. For any constant $\theta > 0$ and large enough β, the following statements hold as $n \to \infty$,*

(i) $\sup_{1 \leq d_n(v) \leq m_n} \mathbb{P}(\xi_v \geq x | \mathcal{F}_{m_n}) \xrightarrow{a.s.} 0$, *for every $x > 0$.*

(ii) *For every $x > 0$,*

$$\Delta_{n,1} \xrightarrow{\mathbb{P}} \nu([x,\infty)) := \begin{cases} \frac{c}{\mu} e^{-\frac{c}{\mu}} \frac{1}{x} & \text{if} \quad \ln V_1 \quad \text{is non-lattice,} \\ \frac{c}{\mu} \frac{d}{1-e^{-d}} e^{d\lfloor \varrho - d^{-1}\ln x - d^{-1}c/\mu \rfloor - d\varrho} & \text{if} \quad \ln V_1 \quad \text{is lattice.} \end{cases}$$

(iii) $\Delta_{n,2} \xrightarrow{\mathbb{P}} \left(\frac{2c\mu + c\mu^2 - c\sigma^2 - \mu\sigma^2 + \mu^3}{2\mu^2} + \ln\theta + \psi_\theta(d) \right) \frac{c}{\mu} e^{-\frac{c}{\mu}}$.

(iv) $\Delta_{n,3} \xrightarrow{\mathbb{P}} \theta\left(1 + \psi_\theta(d)\right) \frac{c}{\mu} e^{-\frac{c}{\mu}}$.

The proof of this theorem is rather technical and postponed until the Appendix B.

Proof of Theorem 1.10: We apply Kallenberg (2002, Theorem 15.28) with the constants

$$a = 0 \quad \text{and} \quad b = \left(\frac{2c\mu + c\mu^2 - c\sigma^2 - \mu\sigma^2 + \mu^3}{2\mu^2} \right) \frac{c}{\mu} e^{-\frac{c}{\mu}}$$

to the sequence $\left(Z_n := \sum_{1 \leq d_n(v) \leq m_n} \xi_v + \sum_{i=1}^n \xi_i', n \geq 1\right)$ conditioned on $\mathcal{F}_{m_n}$. We observe that $\alpha_n/n \to 0$ as $n \to \infty$. Thus, Theorem 3.5 (i) implies that conditioned on $\mathcal{F}_{m_n}$ the variables $(\xi_v, 1 \leq d_n(v) \leq m_n) \cup (\xi_i', i \geq 1)$ form a null array. Theorem 3.5 (ii) shows that $\nu(\mathrm{d}x) = c\mu^{-1} e^{-\frac{c}{\mu}} x^{-2}$, for $x > 0$. Hence

$$\int_0^\theta x^2 \nu(\mathrm{d}x) = c\mu^{-1} e^{-\frac{c}{\mu}} \theta \quad \text{and} \quad \int_\theta^1 x \nu(\mathrm{d}x) = -c\mu^{-1} e^{-\frac{c}{\mu}} \ln\theta \quad \text{for} \quad \theta > 0.$$

Thus the right-hand side of Theorem 3.5 (iii) and (iv) can be written as

$$b - \int_\theta^1 x\nu(\mathrm{d}x) \quad \text{and} \quad a + \int_0^\theta x^2 \nu(\mathrm{d}x), \quad \text{for} \quad \theta > 0,$$

respectively. Therefore Kallenberg (2002, Theorem 15.28) implies that there is the convergence in distribution $Z_n \xrightarrow{d} W$ conditioned on $\mathcal{F}_{m_n}$, where W has a weakly 1-stable distribution with characteristic function given by

$$\mathbb{E}[e^{itW}] = \exp\left(ibt + \int_0^\infty \left(e^{itx} - 1 - itx\mathbb{1}_{\{x<1\}}\right)\nu(\mathrm{d}x)\right).$$

This expression can be simplified to show that W is equal in distribution to

$$\frac{c}{\mu}e^{-\frac{c}{\mu}}\left(Z + \ln\left(\frac{c}{\mu}\right) + \frac{(\mu^2-\sigma^2)(c+\mu)}{2\mu^2} - \gamma + 1\right),$$

where γ is the Euler constant and the variable Z has the continuous Luria-Delbrück distribution; see, e.g., Feller (1971, Section XVII.3). Finally, note that the conditioning does not affect the distribution of W. Then it follows that the convergence $Z_n \xrightarrow{d} W$ holds also unconditioned; We refer to Holmgren (2010b, pages 407-409) for a formal proof of this fact where a general argument is provided for a sequence with a similar structure as $(Z_n, n \geq 1)$. Therefore, the proof of Theorem 1.10 is completed. □

Proof of Theorem 1.12: It follows along the lines of the proof of Theorem 1.10. Details are left to the reader. □

4. Proof of Theorem 1.11

In this section, we deduce Theorem 1.11 from Theorem 1.10 by showing that $\frac{n}{\ln n}G_n$ and $\frac{\alpha n}{\ln n}\hat{G}_n$ are close enough as $n \to \infty$. We start by recalling some notation from Section 3. Remember that we write $m_n = \lfloor \beta \log_b \ln n \rfloor$, for some constant $\beta > 0$, and that we assume that n is large enough such that $0 < m_n < \ln n$. For $1 \leq i \leq b^{m_n}$, recall also that we let v_i be a vertex in T_n^{sp} at height m_n and we let n_i be the number of balls stored at the sub-tree rooted at v_i. We further let N_i be the (random) number of vertices at the sub-tree rooted at v_i.

We denote by $C_{n,i}$ the number of vertices of the sub-tree of T_n^{sp} rooted at v_i after percolation with parameter p_n. Clearly, $(C_{n,i}, 1 \leq i \leq b^{m_n})$ are conditionally independent random variables given $(n_i, 1 \leq i \leq b^{m_n})$. Write $\mathbb{E}_{n_i}[C_{n,i}] := \mathbb{E}[C_{n,i}|n_i]$, i.e., it is the conditional expected value of $C_{n,i}$ given n_i. We have the following estimation of $C_{n,i}$ that corresponds to Lemma 3.1.

Lemma 4.1. *Suppose that Condition 1 and 2 are fulfilled. For $1 \leq i \leq d^{m_n}$, we have that*

$$\mathbb{E}_{n_i}[C_{n,i}] = \alpha n_i e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \alpha\frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3}\frac{n_i \ln n_i}{\ln^2 n}e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - c\zeta\frac{n_i}{\ln n}e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o\left(\frac{n_i}{\ln n}\right),$$

where $\zeta \in \mathbb{R}$ is the constant in (1.5).

Proof: For $1 \leq i \leq d^{m_n}$, let T_i be the sub-tree of T_n^{sp} rooted at the vertex v_i at height m_n. Let u_i be a vertex in T_i with the uniform distribution on the set of vertices of the sub-tree T_i. Let $d_{n_i}(u_i)$ be the height of u_i. Recall the observation made by Bertoin (2013, Proof of Theorem 1),

$$\mathbb{E}_{n_i}\left[N_i^{-1}C_{n,i}\right] = \mathbb{E}_{n_i}\left[p_n^{d_{n_i}(u_i)}\right]. \tag{4.1}$$

In words, the left-hand side can be interpreted as the probability that u_i belongs to the percolation cluster containing the root of T_i, i.e., v_i, while the right-hand side can be interpreted as the probability that no edge has been removed in the path between u_i and v_i. Then a similar computation as in the proof of Lemma 3.1 together with Lemma A.3 (i)-(iii) in Appendix A shows that

$$\mathbb{E}_{n_i}\left[p_n^{d_{n_i}(u_i)}\right] = e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \frac{c^2\mu^2 - c^2\sigma^2}{2\mu^3}\frac{\ln n_i}{\ln^2 n}e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} - \frac{c\zeta}{\alpha}\frac{1}{\ln n}e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} + o\left(\frac{1}{\ln n}\right). \tag{4.2}$$

On the other hand, we note that $C_{n,i} \leq N_i$. Hence Condition 2 and Remark 1.6 imply that

$$\left|\mathbb{E}_{n_i}\left[N_i^{-1} C_{n,i}\right] - \mathbb{E}_{n_i}\left[\mathbb{E}_{n_i}^{-1}[N_i] C_{n,i}\right]\right| \leq \mathbb{E}_{n_i}^{-1}[N_i]\mathbb{E}\left[|N_i - \mathbb{E}_{n_i}[N_i]|\right] = o\left(\ln^{-1} n\right).$$

By making use of Condition 2 one more time, we deduce that

$$\mathbb{E}_{n_i}\left[N_i^{-1} C_{n,i}\right] = \alpha^{-1}\mathbb{E}_{n_i}\left[n_i^{-1} C_{n,i}\right] + o\left(\ln^{-1} n\right). \tag{4.3}$$

Therefore, our claim follows from the combination of (4.1), (4.2) and (4.3). □

Recall that $\eta_{n,i}$ denotes the total number of edges on the branch from v_i to the root which has been deleted after percolation with parameter p_n. The next result is analogous of Lemma 3.2.

Lemma 4.2. *Suppose that Condition 1 and 2 holds. We have for $\beta > -2/(\log_b \mathbb{E}[V_1^2] + 1)$ that*

$$G_n = \sum_{i=1}^{b^{m_n}} \mathbb{E}_{n_i}[C_{n,i}] \mathbb{1}_{\{\eta_{n,i}=0\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

Proof: The proof follows from a very similar argument as the proof of Lemma 3.2. □

Finally, we show that $\frac{n}{\ln n} G_n$ and $\frac{\alpha n}{\ln n}\hat{G}_n$ possess the same asymptotic behaviour.

Lemma 4.3. *Suppose that Condition 1 and 2 holds. We have for $\beta > -2/(1 + \log_b \mathbb{E}[V_1^2] + 1)$ that*

$$G_n = \alpha \hat{G}_n + c\alpha(\varsigma - \zeta\alpha^{-1}) e^{-\frac{c}{\mu}} \frac{n}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

where $\zeta \in \mathbb{R}$ is defined in (1.5) and $\varsigma \in \mathbb{R}$ is the constant value of the function ϖ in (1.3) when $d = 0$.

Proof: We deduce from Lemma 4.1, Lemma 4.2 and equation (3.9) that

$$G_n = \alpha \hat{G}_n - c\alpha \sum_{i=1}^{b^{m_n}} \left(\frac{\zeta}{\alpha}\frac{n_i}{\ln n} - \frac{n_i \varpi(\ln n_i)}{\ln n}\right) e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

By Condition 2, the random variable $\ln V_1$ is non-lattice and thus the function ϖ is a constant equal to ς. Hence

$$G_n = \alpha \hat{G}_n - c\alpha(\zeta\alpha^{-1} - \varsigma) \sum_{i=1}^{b^{m_n}} \frac{n_i}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

Furthermore, the estimations (3.11), (3.13) and (3.17) allow us to deduce that

$$\sum_{i=1}^{b^{m_n}} \frac{n_i}{\ln n} e^{-\frac{c}{\mu}\frac{\ln n_i}{\ln n}} \mathbb{1}_{\{\eta_{n,i}=0\}} = e^{-\frac{c}{\mu}} \frac{n}{\ln n} + o_{\mathrm{p}}\left(\frac{n}{\ln n}\right).$$

Therefore, the result follows clearly by combining the previous two estimates. □

We are now in the position to prove Theorem 1.11.

Proof of Theorem 1.11: By normalizing G_n, Lemma 4.3 gives that

$$\begin{aligned}&\left(n^{-1} G_n - \alpha e^{-\frac{c}{\mu}}\right) \ln n - \frac{c\alpha}{\mu} e^{-\frac{c}{\mu}} \ln \ln n \\ &\qquad = \alpha\left(n^{-1}\hat{G}_n - e^{-\frac{c}{\mu}}\right) \ln n - \frac{c\alpha}{\mu} e^{-\frac{c}{\mu}} \ln\ln n + c\alpha(\varsigma - \zeta\alpha^{-1}) e^{-\frac{c}{\mu}} + o_{\mathrm{p}}(1).\end{aligned}$$

Therefore, the result in Theorem 1.11 follows from a simple application of Theorem 1.10. □

5. Percolation on b-regular trees

In this section, we point out that the approach developed in the proof of Theorem 1.10 can be also applied to study percolation on other classes of trees. We focus here on the case of rooted complete regular b-ary trees T_h^{reg} with height $h \in \mathbb{N}$ and $b \geq 2$ a fixed integer (i.e., each vertex has exactly out-degree b). We note that there are b^k vertices at distance $k = 0, 1, \ldots, h$ from the root and a total of $n_h = (b^{h+1} - 1)/(b-1)$ vertices. We perform Bernoulli bond percolation with parameter

$$p_h = e^{-c/h},$$

where $c > 0$ is fixed. Indeed, this choice of the percolation parameter corresponds precisely to the supercritical regime, i.e., there exists a (unique) giant cluster such that $\lim_{h\to\infty} n_h^{-1} G_h^{\mathrm{reg}} = e^{-c}$, in probability, where G_h^{reg} denotes the size (i.e., the number of vertices) of the cluster that contains the root. We refer to Bertoin (2013, Section 3) for details. We are interested in the fluctuations of G_h^{reg}. We introduce for every $\rho \in [0, 1)$ and $x > 0$,

$$\bar{\Lambda}_\rho(x) = \frac{b^{-\rho + \lfloor \rho - \log_b x \rfloor + 1}}{b-1}.$$

This function decreases as $x \to \infty$ and it can be viewed as the tail of a measure Λ_ρ on $(0, \infty)$. Furthermore, this measure fulfils the integral condition $\int_{(0,\infty)} (1 \wedge x^2) \Lambda_\rho(\mathrm{d}x) < \infty$. This enables us to introduce a Lévy process without negative jumps $L_\rho = (L_\rho(t))_{t \geq 0}$ with Laplace exponent

$$\Psi_\rho(a) = \int_{(0,\infty)} (e^{-ax} - 1 + ax\mathbb{1}_{\{x<1\}}) \Lambda_\rho(\mathrm{d}x), \quad \text{for } a \geq 0.$$

Bertoin (2014a, Theorem 3.1) has proven that the fluctuations of the number of vertices at height h which has been disconnected from the root after percolation are described by L_ρ. Indeed, L_ρ also appears in relation with limit theorems for the number of random records on a complete binary tree; see Janson (2004).

We state the following analogue of Theorem 1.11.

Theorem 5.1. *In the regime where $h \to \infty$ with $\{\log_b h\} \to \rho \in [0, 1)$, we have that*

$$\left(\frac{G_h^{\mathrm{reg}}}{n_h} - e^{-c} \right) h - ce^{-c} \log_b h \xrightarrow{d} -e^{-c} \left(L_\rho(c) + c\rho - \frac{c}{b-1} \right).$$

The proof strategy is the similar as the one used in the proof of Theorem 1.10. We write $m_h = 2\lfloor \log_b h \rfloor$ and assume that h is large enough such that $0 < m_h < h$. For $1 \leq i \leq b^{m_h}$, let v_i be the b^{m_h} vertices at height m_h. Note that the number of vertices of the sub-tree of T_h^{reg} rooted at v_i is given by $n_{h,i} = (b^{h-m_h+1} - 1)/(b-1)$. Denote by $C_{h,i}$ the number of vertices of the sub-tree of T_h^{reg} rooted at v_i after percolation with parameter p_h. Clearly, $(C_{h,i}, 1 \leq i \leq b^{m_h})$ is a sequence of i.i.d. random variables.

Lemma 5.2. *For $1 \leq i \leq b^{m_h}$, we have that*

$$\mathbb{E}[C_{h,i}] = n_{h,i}e^{-c} + n_{h,i}h^{-1}(b-1)^{-1}ce^{-c} + n_{h,i}m_h h^{-1}ce^{-c} + o(n_{h,i}h^{-1}).$$

Proof: For $1 \leq i \leq b^{m_h}$, let $T_{h,i}$ be the sub-tree of T_h^{reg} rooted at the vertex v_i. Let u_i denote a uniform chosen vertex in $T_{h,i}$ and write $d_h(u_i)$ for its height in $T_{h,i}$. Note that $\mathbb{P}(d_h(u_i) = k) = b^k n_{h,i}^{-1}$, for $k \in \{0, 1, \ldots, h - m_h\}$. By the key observation made by Bertoin (2013, Proof of Theorem 1), we have that

$$\mathbb{E}\left[n_{h,i}^{-1}C_{h,i}\right] = \mathbb{E}\left[e^{-ch^{-1}d_h(u_i)}\right] = \sum_{k=0}^{h-m_h} e^{-ch^{-1}k}\mathbb{P}(d_h(u_i)=k) = \frac{b^{h-m_h}}{n_{h,i}}e^{-c\frac{h-m_h}{h}}\sum_{k=0}^{h-m_h} e^{ch^{-1}k}b^{-k}$$
$$= \frac{b^{h-m_h}}{n_{h,i}}e^{-c\frac{h-m_h}{h}}\left(\frac{b}{b-1}+\frac{cb}{h(b-1)^2}+o(h^{-1})\right).$$

Recall that $n_{h,i} = (b^{h-m_h+1}-1)/(b-1)$. Therefore, after some simple computations we obtain that

$$\mathbb{E}\left[n_{h,i}^{-1}C_{h,i}\right] = e^{-c\frac{h-m_h}{h}}\left(1+c(b-1)^{-1}h^{-1}\right)+o(h^{-1})$$

from which our claim follows. □

Let $\eta_{h,i}$ be the total number of edges on the branch from v_i to the root which have been deleted after percolation with parameter p_h. Note that the random variable $\eta_{h,i}$ has the binomial distribution with parameters $(m_h, 1-p_h)$. But the random variables $(\eta_{h,i}, 1 \leq i \leq b^{m_h})$ are not independent. On the other hand, $\eta_{h,i}=0$ if and only if the vertex v_i is still connected to the root of T_h^{reg}.

Lemma 5.3. *We have that*

$$G_h^{\mathrm{reg}} = -n_{h,1}e^{-c}\sum_{i=1}^{b^{m_h}}\mathbb{1}_{\{\eta_{h,i}\geq 1\}} + n_h e^{-c} + n_h h^{-1}(b-1)^{-1}ce^{-c} + n_h m_h h^{-1}ce^{-c} + o_{\mathrm{p}}(n_h h^{-1}).$$

Proof: We denote by $C_{h,0}$ the number of vertices of the tree T_h^{reg} at height less or equal to m_h-1 that are connected to the root after percolation with parameter p_h. Then, it should be plain that

$$G_h^{\mathrm{reg}} = C_{h,0} + \sum_{i=1}^{b^{m_h}} C_{h,i}\mathbb{1}_{\{\eta_{h,i}=0\}}.$$

The sequences of random variables $(\eta_{h,i}, 1 \leq i \leq b^{m_h})$ and $(C_{h,i}, 1 \leq i \leq b^{m_h})$ are independent. By conditioning first on the value of the random variables $(\eta_{h,i}, 1 \leq i \leq b^{m_h})$ and then taking expectation, we obtain that

$$\mathbb{E}\left[\left(G_h^{\mathrm{reg}} - C_{h,0} - \sum_{i=1}^{b^{m_h}}\mathbb{E}\left[C_{h,i}\right]\mathbb{1}_{\{\eta_{h,i}=0\}}\right)^2\right] = \mathbb{E}\left[\sum_{i=1}^{b^{m_h}}\mathbb{E}\left[(C_{h,i}-\mathbb{E}[C_{h,i}])^2\right]\mathbb{1}_{\{\eta_{h,i}=0\}}\right]$$
$$= \sum_{i=1}^{b^{m_h}}\mathbb{E}\left[(C_{h,i}-\mathbb{E}[C_{h,i}])^2\right]\mathbb{P}\left(\eta_{h,i}=0\right).$$

On the one hand, $\mathbb{P}(\eta_{h,i}=0)\leq 1$. On the other hand, Bertoin (2013, Section 3) has proven in Bertoin (2013, Proof of Corollary 1) that $\mathbb{E}[(C_{h,i}-\mathbb{E}[C_{h,i}])^2] = o(n_{h,i}^2)$. Thus,

$$\mathbb{E}\left[\left(G_h^{\mathrm{reg}} - C_{h,0} - \sum_{i=1}^{b^{m_h}}\mathbb{E}\left[C_{h,i}\right]\mathbb{1}_{\{\eta_{h,i}=0\}}\right)^2\right] = \sum_{i=1}^{b^{m_h}} o(n_{h,i}^2) = o(n_h^2 h^{-2}).$$

The above estimate and Chebyshev's inequality imply that

$$G_h^{\mathrm{reg}} = C_{h,0} + \mathbb{E}[C_{h,1}]\sum_{i=1}^{b^{m_h}}\mathbb{1}_{\{\eta_{h,i}=0\}} + o_{\mathrm{p}}(n_h h^{-1})$$

since $(C_{h,i}, 1 \leq i \leq b^{m_h})$ is a sequence of i.i.d. random variables. Moreover, we notice that $0 \leq C_{h,0} < b^{m_h+1} = o(n_h h^{-1})$. Hence

$$G_h^{\mathrm{reg}} = \mathbb{E}[C_{h,1}]\sum_{i=1}^{b^{m_h}}\mathbb{1}_{\{\eta_{h,i}=0\}} + o_{\mathrm{p}}(n_h h^{-1}). \tag{5.1}$$

We note that

$$\sum_{i=1}^{b^{m_h}} \mathbb{1}_{\{\eta_{h,i}=0\}} = b^{m_h} - \sum_{i=1}^{b^{m_h}} \mathbb{1}_{\{\eta_{h,i}\geq 1\}}. \tag{5.2}$$

Finally, our claim follows by combining (5.2) and Lemma 5.2 into (5.1). □

We can now complete the proof of Theorem 5.1.

Proof of Theorem 5.1: From Lemma 5.3 we deduce that

$$\left(\frac{G_h^{\text{reg}}}{n_h} - e^{-c}\right) h - ce^{-c}\log_b h$$
$$= -\frac{n_{h,1}h}{n_h}e^{-c}\sum_{i=1}^{b^{m_h}} \mathbb{1}_{\{\eta_{h,i}\geq 1\}} + ce^{-c}\lfloor \log_b h\rfloor - ce^{-c}\{\log_b h\} + \frac{c}{b-1}e^{-c} + o_{\text{p}}(1).$$

Since $n_h^{-1}n_{h,1} = b^{-m_h} + o(b^{-m_h})$ and

$$\mathbb{E}\left[\sum_{i=1}^{b^{m_h}} \mathbb{1}_{\{\eta_{h,i}\geq 1\}}\right] = \sum_{i=1}^{b^{m_h}} \mathbb{P}\left(\eta_{h,i}\geq 1\right) = b^{m_h}(1-e^{-cm_h h^{-1}}),$$

we conclude by the Markov inequality that

$$\left(\frac{G_h^{\text{reg}}}{n_h} - e^{-c}\right) h - ce^{-c}\log_b h$$
$$= -hb^{-m_h}e^{-c}\sum_{i=1}^{b^{m_h}} \mathbb{1}_{\{\eta_{h,i}\geq 1\}} + ce^{-c}\lfloor \log_b h\rfloor - ce^{-c}\{\log_b h\} + \frac{c}{b-1}e^{-c} + o_{\text{p}}(1).$$

Our claim follows by Bertoin (2014a, Corollary 3.4) that establishes the convergence in distribution

$$hb^{-m_h}\sum_{i=1}^{b^{m_h}} \mathbb{1}_{\{\eta_{h,i}\geq 1\}} - c\lfloor \log_b h\rfloor \xrightarrow{d} L_\rho(c),$$

in the regime where $h\to\infty$ with $\{\log_b h\}\to\rho\in[0,1)$. □

Remark 5.4. One could have finished the proof of Theorem 5.1 along the same lines as for Theorem 1.10, i.e., by using a classical limit result for triangular arrays. But for the sake of avoiding repetition, we decided to directly apply a result proven by Bertoin (2014a) which is enough for our purpose.

Appendix A. Distances in split trees

The purpose of this section is to establish some general results on the distribution of the distances between uniform chosen vertices and uniformly chosen balls in T_n^{sp} when $n\to\infty$. The results can be seen as a complement (or extension) of those of Devroye (1999) and Holmgren (2012). Let H_n be the height of T_n^{sp}, i.e., the maximal distance between the root and any leaf in T_n^{sp}. We deduce the following moment estimate for H_n. For $y\in\mathbb{R}$, recall that $\lceil y\rceil$ denotes the least integer greater than or equal to y. Similarly, $\lfloor y\rfloor$ denotes the greatest integer less than or equal to y

Lemma A.1. *If Condition 1, then* $\sup_{n\geq 1}\mathbb{E}[H_n^r]\ln^{-r} n<\infty$*, for all* $r>0$.

Proof: We claim that for all $r>0$ there exists $c_r>0$ such that

$$\lim_{n\to\infty} n^r\mathbb{P}(H_n\geq (3s_1+4)\lfloor c_r\ln n\rfloor) = 0. \tag{A.1}$$

Then, the bound $H_n\leq n$ implies that

$$\mathbb{E}[H_n^r]\leq (3s_1+4)c_r\ln^r n + n^r\mathbb{P}(H_n\geq (3s_1+4)\lfloor c_r\ln n\rfloor)$$

which combined with (A.1) allows us to conclude with the proof of Lemma A.1. Therefore, it only remains to prove the claim in (A.1). Devroye (1999) has shown that for integers $0 \le k' \le k$ and $l = k'(s_1+1)$ such that $s_1 k' < l$, and real numbers $t, t' > 0$, we have that

$$\mathbb{P}(H_n \ge k + 3l) \le 2b^{-k} + b^k (ne)^t b^{2kt/l} m(t)^k + b^k (s_1(k-k'+1)e)^{t'} b^{2kt'/l} m(t')^{k'}, \tag{A.2}$$

where $m(t) = \mathbb{E}[V_1^t]$ for $t > 0$; see proof of Devroye (1999, Theorem 1) for details. Then consider the estimate in (A.2) with $k = k' = \lfloor c_r \ln n \rfloor$ and $l = k'(s_1+1)$. Then choose $t, t' \ge 0$ large enough such that $bm(t) < 1$ and $bm(t') < 1$. This is possible because $\mathbb{P}(V_1 = 1) = 0$ by Condition 1, and thus, $m(t) \to 0$ as $t \to \infty$; see Devroye (1999, Lemma 1). Finally, (A.1) follows immediately by taking $c_r > \max(r/\ln b, -(r+t)/\ln(bm(t)), -r/\ln(bm(t')))$. □

For each fixed $n \in \mathbb{N}$, let b_1 be a uniformly distributed ball on the set $\{1, \dots, n\}$ of balls in T_n^{sp}. Recall that we denote by $D_n(b_1)$ the height (or depth) of the ball b_1 in T_n^{sp}, i.e., the number of edges of T_n^{sp} which are between the root and the vertex where the ball b_1 is stored.

Lemma A.2. *Assume that Condition 1 is fulfilled.*

(i) $\mathbb{E}[D_n(b_1)] = \mu^{-1}\ln n + \varpi(\ln n) + o(1)$, *where* $\varpi : \mathbb{R} \to \mathbb{R}$ *denotes the function in (1.3).*
(ii) $\mathbb{E}[(D_n(b_1) - \mu^{-1}\ln n)^2] = \mu^{-3}\sigma^2 \ln n + o(\ln n)$.
(iii) $\mathbb{E}[|D_n(b_1) - \mu^{-1}\ln n|^3] = O(\ln^{\frac{3}{2}} n)$.
(iv) $\lim_{n\to\infty} D_n(b_1)(\ln n)^{-1} = 1/\mu$, *in probability.*

Proof: We observe that $\mathbb{E}[D_n(b_1)] = n^{-1}\mathbb{E}[\sum_{i=1}^n D_n(i)] = n^{-1}\mathbb{E}\left[\Psi(T_n^{\mathrm{sp}})\right]$. Then (i) follows immediately from the result in (1.3). Turning our attention to the proof of (ii), we write

$$\mathbb{E}[(D_n(b_1) - \mu^{-1}\ln n)^2] = n^{-1}\mathbb{E}\Big[\sum_{i=1}^n (D_n(i) - \mu^{-1}\ln n)^2\Big]. \tag{A.3}$$

By Holmgren (2012, Proposition 1.1), $D_n(j) \le_{\mathrm{st}} D_n(j')$ for $j \le j'$. Moreover, $D_j(j) \le D_n(j)$, for $n \ge j$, since a ball with label j only move downward during the splitting process when new balls are added to the tree. Furthermore, it follows from Holmgren (2012, Theorem 1.3) that

$$\mathbb{E}\left[(D_n(j) - \mu^{-1}\ln n)^2\right] = \mu^{-3}\sigma^2\ln n + o(\ln n), \qquad \text{uniformly for} \qquad \lceil n\ln^{-1} n\rceil \le j \le n. \tag{A.4}$$

Since $D_n(j)$ can be stochastically dominated from above and below by $D_n(n)$ and $D_j(j)$, for $1 \le j \le n$, respectively, we deduce that

$$\begin{aligned}\mathbb{E}\left[(D_n(j) - \mu^{-1}\ln n)^2\right] &\le \mathbb{E}\left[(D_n(n) - \mu^{-1}\ln n)^2\right] + \mathbb{E}\left[(D_j(j) - \mu^{-1}\ln n)^2\right] \\ &\le \mathbb{E}\left[(D_n(n) - \mu^{-1}\ln n)^2\right] + 4\mathbb{E}\left[(D_j(j) - \mu^{-1}\ln j)^2\right] + 4\mu^{-2}\left|\ln j - \ln n\right|^2 \\ &= o(\ln^2 n),\end{aligned} \tag{A.5}$$

uniformly for $\lceil n\ln^{-2} n\rceil \le j < \lceil n\ln^{-1} n\rceil$; We have used the inequality $|x-y|^2 \le 4x^2 + 4y^2$ for $x, y \ge 0$. On the other hand, Lemma A.1 implies that

$$\mathbb{E}\left[(D_n(j) - \mu^{-1}\ln n)^2\right] \le 4\mathbb{E}\left[H_n^2\right] + 4\mu^{-2}\ln^2 n = o(\ln^3 n) \tag{A.6}$$

uniformly for $1 \le j < \lceil n\ln^{-2} n\rceil$. Then the combination (A.3), (A.4), (A.5) and (A.6) imply (ii).

We now prove (iii). We observe that

$$\mathbb{E}\left[\left|D_n(b_1) - \mu^{-1}\ln n\right|^3\right] = n^{-1}\mathbb{E}\Big[\sum_{i=1}^n \left|D_n(i) - \mu^{-1}\ln n\right|^3\Big]. \tag{A.7}$$

We also observe that

$$\begin{aligned}\mathbb{E}\left[\left|D_n(j) - \mu^{-1}\ln n\right|^3\right] &\le \mathbb{E}\left[\left|D_n(n) - \mu^{-1}\ln n\right|^3\right] + \mathbb{E}\left[\left|D_j(j) - \mu^{-1}\ln n\right|^3\right] \\ &\le \mathbb{E}\left[\left|D_n(n) - \mu^{-1}\ln n\right|^3\right] + 8\mathbb{E}\left[\left|D_j(j) - \mu^{-1}\ln j\right|^3\right] + 8\mu^{-3}\left|\ln j - \ln n\right|^3,\end{aligned}$$

for $1 \leq j \leq n$; we have used the inequality $|x-y|^3 \leq 8x^3 + 8y^3$ for $x, y \geq 0$. From Holmgren (2012, equation (3.62)) we deduce that

$$\mathbb{E}\left[\left|D_n(j) - \mu^{-1}\ln n\right|^3\right] = O\left(\ln^{\frac{3}{2}} n\right), \quad \text{uniformly for} \quad \lceil n \ln^{-2} n \rceil \leq j \leq n. \tag{A.8}$$

Observe that $\mathbb{E}\left[\left|D_n(j) - \mu^{-1}\ln n\right|^3\right] \leq 8\mathbb{E}\left[H_n^3\right] + 8\mu^{-3}\ln^3 n$, uniformly for $1 \leq j < \lceil n \ln^{-2} n \rceil$. Then Lemma A.1 implies that

$$\mathbb{E}\left[\left|D_n(j) - \mu^{-1}\ln n\right|^3\right] = O\left(\ln^3 n\right), \quad \text{uniformly for} \quad 1 \leq j < \lceil n \ln^{-2} n \rceil. \tag{A.9}$$

Therefore, (iii) follows from (A.7), (A.8) and (A.9).

The point (iv) follows immediately from (ii) and a standard application of Chebyshev's inequality. □

We turn our attention to the height of a random chosen vertex in T_n^{sp}. For each fixed $n \in \mathbb{N}$, let u_1 be a uniformly distributed vertex on the random split tree T_n^{sp} with n balls. Recall that we denote by $d_n(u_1)$ the height of the vertex u_1 in T_n^{sp}, i.e., the minimal number of edges of T_n^{sp} which are needed to connect the root and u_1.

Lemma A.3. *Assume that Conditions 1 and 2 are fulfilled.*

(i) *Recall that $\zeta \in \mathbb{R}$ is the constant in Condition 2. Then $\mathbb{E}[d_n(u_1)] = \mu^{-1}\ln n + \zeta\alpha^{-1} + o(1)$.*
(ii) *We also have $\mathbb{E}[(d_n(u_1) - \mu^{-1}\ln n)^2] = \mu^{-3}\sigma^2 \ln n + o(\ln n)$.*
(iii) *Furthermore, for $\delta > 1/2 - \varepsilon$, $\mathbb{E}[\left|d_n(u_1) - \mu^{-1}\ln n\right|^3] = O(\ln^{\frac{3}{2}+\delta} n)$, where $\varepsilon > 0$ is the constant that appears in Condition 2.*
(iv) *As a consequence, we conclude that $\lim_{n\to\infty} d_n(u_1)(\ln n)^{-1} = 1/\mu$, in probability.*

Proof: Observe that

$$\mathbb{E}[d_n(u_1)] = \mathbb{E}\left[\frac{1}{N}\sum_{u \in T_n^{\text{sp}}} d_n(u)\right] = \frac{1}{\mathbb{E}[N]}\mathbb{E}\left[\Upsilon(T_n^{\text{sp}})\right] + \mathbb{E}\left[\left(\frac{1}{N} - \frac{1}{\mathbb{E}[N]}\right)\Upsilon(T_n^{\text{sp}})\right].$$

It should be clear that (i) follows from Condition 2 and the result in (1.5) by showing that

$$\mathbb{E}\left[\left(\frac{1}{N} - \frac{1}{\mathbb{E}[N]}\right)\Upsilon(T_n^{\text{sp}})\right] = o(1). \tag{A.10}$$

Therefore, we focus on the proof of (A.10).

Note that

$$\left|\frac{1}{N} - \frac{1}{\mathbb{E}[N]}\right|\Upsilon(T_n^{\text{sp}}) = \left|\frac{N - \mathbb{E}[N]}{N\mathbb{E}[N]}\right|\sum_{u \in T_n^{\text{sp}}} d_n(u) \leq \frac{|N - \mathbb{E}[N]|\, H_n}{\mathbb{E}[N]},$$

where H_n denotes the height of T_n^{sp}. An application of the Cauchy–Schwarz inequality shows that

$$\mathbb{E}\left[\left(\frac{1}{N} - \frac{1}{\mathbb{E}[N]}\right)\Upsilon(T_n^{\text{sp}})\right] \leq \mathbb{E}^{-1}[N](Var(N))^{\frac{1}{2}}\mathbb{E}^{1/2}[H_n^2] = o(1),$$

where in the last step we used Remark 1.6, Condition 2 and Lemma A.1.

We turn our attention to the proof of (ii). Note that

$$\begin{aligned}\mathbb{E}[(d_n(u_1)-\mu^{-1}\ln n)^2] &= \mathbb{E}\left[\frac{1}{N}\sum_{u\in T_n^{\mathrm{sp}}}(d_n(u)-\mu^{-1}\ln n)^2\right]\\ &= \frac{1}{\mathbb{E}[N]}\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}}(d_n(u)-\mu^{-1}\ln n)^2\right]\\ &\quad+\mathbb{E}\left[\left(\frac{1}{N}-\frac{1}{\mathbb{E}[N]}\right)\sum_{u\in T_n^{\mathrm{sp}}}(d_n(u)-\mu^{-1}\ln n)^2\right].\end{aligned}$$

Holmgren (2010a, Corollary 2.1) has shown that

$$\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}}(d_n(u)-\mu^{-1}\ln n)^2\right]=\alpha n\mu^{-3}\sigma^2\ln n+o(n\ln n).$$

Then (ii) follows from Condition 2 and Remark 1.6 by providing that

$$\mathbb{E}\left[\left(\frac{1}{N}-\frac{1}{\mathbb{E}[N]}\right)\sum_{u\in T_n^{\mathrm{sp}}}(d_n(u)-\mu^{-1}\ln n)^2\right]=o(\ln n).$$

This is proved from similar arguments as in the proof of (A.10). The details are omitted.

We continue with the proof of (iii). We have that

$$\begin{aligned}\mathbb{E}\left[|d_n(u_1)-\mu^{-1}\ln n|^3\right] &= \mathbb{E}\left[\frac{1}{N}\sum_{u\in T_n^{\mathrm{sp}}}|d_n(u)-\mu^{-1}\ln n|^3\right]\\ &= \frac{1}{\mathbb{E}[N]}\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}}|d_n(u)-\mu^{-1}\ln n|^3\right]\\ &\quad+\mathbb{E}\left[\left(\frac{1}{N}-\frac{1}{\mathbb{E}[N]}\right)\sum_{u\in T_n^{\mathrm{sp}}}|d_n(u)-\mu^{-1}\ln n|^3\right].\end{aligned}$$

Suppose that we have proven that

$$\mathbb{E}\left[\frac{1}{n}\sum_{u\in T_n^{\mathrm{sp}}}|d_n(u)-\mu^{-1}\ln n|^3\right]=O\left(\ln^{\frac{3}{2}+\delta}n\right), \tag{A.11}$$

for $\delta>1/2-\varepsilon$. Then (iii) follows from Condition 2 and by showing that

$$\mathbb{E}\left[\left(\frac{1}{N}-\frac{1}{\mathbb{E}[N]}\right)\sum_{u\in T_n^{\mathrm{sp}}}|d_n(u)-\mu^{-1}\ln n|^3\right]\ln^{-\frac{3}{2}-\delta}n=o(1),\quad \text{for}\quad \delta>1/2-\varepsilon.$$

This can be proved by using similar arguments as in the proof of (A.10) and the details are omitted.

Finally, we check that (A.11) holds. For $\delta > 1/2 - \varepsilon$ and $C > 0$, we notice that

$$\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}} |d_n(u) - \mu^{-1}\ln n|^3 \mathbb{1}_{\left\{|d_n(u)-\mu^{-1}\ln n| > \ln^{\frac{1}{2}+\frac{\delta}{3}} n\right\}}\right]$$
$$\le 8\mathbb{E}\left[\left(H_n^3 + \mu^{-3}\ln^3 n\right)\sum_{u\in T_n^{\mathrm{sp}}} \mathbb{1}_{\left\{|d_n(u)-\mu^{-1}\ln n| > \ln^{\frac{1}{2}+\frac{\delta}{3}} n\right\}}\right]$$
$$\le 8(C^3 + \mu^{-3})(\ln^3 n)\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}} \mathbb{1}_{\left\{|d_n(u)-\mu^{-1}\ln n| > \ln^{\frac{1}{2}+\frac{\delta}{3}} n\right\}}\right] + 8n^4\mathbb{P}(H_n \ge C\ln n).$$

On the one hand, Holmgren (2012, Theorem 1.2) has shown that

$$(C^3 + \mu^{-3})(\ln^3 n)\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}} \mathbb{1}_{\left\{|d_n(u)-\mu^{-1}\ln n| > \ln^{\frac{1}{2}+\frac{\delta}{3}} n\right\}}\right] = o\left(n\ln^{\frac{3}{2}+\delta} n\right)$$

(The sum inside the expectation is what Holmgren (2012, Theorem 1.2) calls the number of bad vertices). On the other hand, by (A.1), we can choose $C > 0$ such that $8n^4\mathbb{P}(H_n \ge C\ln n) = o(n\ln^{\frac{3}{2}+\delta} n)$. Hence,

$$\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}} |d_n(u) - \mu^{-1}\ln n|^3 \mathbb{1}_{\left\{|d_n(u)-\mu^{-1}\ln n| > \ln^{\frac{1}{2}+\frac{\delta}{3}} n\right\}}\right] = o\left(n\ln^{\frac{3}{2}+\delta} n\right). \tag{A.12}$$

We also note that

$$\mathbb{E}\left[\sum_{u\in T_n^{\mathrm{sp}}} |d_n(u) - \mu^{-1}\ln n|^3 \mathbb{1}_{\left\{|d_n(u)-\mu^{-1}\ln n| \le \ln^{\frac{1}{2}+\frac{\delta}{3}} n\right\}}\right] = O\left(n\ln^{\frac{3}{2}+\delta} n\right),$$

which combined with (A.12) implies (A.11).

The point (iv) follows immediately from (ii) and a standard application of Chebyshev's inequality. □

Recall the labelling of the balls induced by the split tree generating algorithm explained in Section 1.1. Let v and v' be the vertices in T_n^{sp} where the balls labelled j and j' are located, respectively. We call the vertex $v \wedge v'$ at which the paths in T_n^{sp} from the vertices v and v' to the root intersect the last common ancestor of the balls with labels j and j'. For simplicity, we denote by $j \wedge j'$ a last common ancestor of the balls j and j' (notice that $j \wedge j'$ is not necessary unique). Let $D_n(j \wedge j')$ be the height of $j \wedge j'$ when all n balls have been inserted.

Lemma A.4. *Assume that Condition 1 is fulfilled. For $n \in \mathbb{N}$ fixed, let b_1 and b_2 denote two independent uniformly distributed random ball labels in T_n^{sp}. Let $h : \mathbb{N} \to \mathbb{R}_+$ be some function such that $\lim_{n\to\infty} h(n) = \infty$. We have that $D_n(b_1 \wedge b_2)h(n)^{-1} \to 0$, as $n \to \infty$, in probability.*

Proof: For $\delta > 0$, note that $D_n(b_1 \wedge b_2) \ge \delta h(n)$ when both balls b_1 and b_2 lie in the same sub-tree and the height of the last common ancestor related to this sub-tree has to be greater than $\delta h(n)$. For $1 \le i \le b^{\lceil \delta h(n)\rceil}$, let v_i be a vertex in T_n^{sp} at height $\lceil \delta h(n)\rceil$ and let n_i be the number of balls stored at the sub-tree rooted at v_i; note that those balls have depth greater than $\delta h(n)$. Since b_1 and b_2 denote two independent uniformly distributed random ball in T_n^{sp}, we have that

$$\mathbb{P}(D_n(b_1 \wedge b_2) \ge \delta h(n)) \le \mathbb{E}\left[\sum_{i=1}^{b^{\lceil \delta h(n)\rceil}} \left(\frac{n_i}{n}\right)^2\right] = n^{-2}\sum_{i=1}^{b^{\lceil \delta h(n)\rceil}} \mathbb{E}\left[n_i^2\right]. \tag{A.13}$$

On the other hand, Condition 1 and the inequality by Holmgren (2011, equation (1.10)) for subtrees sizes in split-trees (we refer to the estimation (3.3) for a formal proof) imply that

$$\mathbb{E}\left[n_i^2\right] = n^2 \mathbb{E}^{\lceil \delta h(n) \rceil}\left[V_1^2\right] + o(n^2 \ln^{-k} n), \tag{A.14}$$

for an arbitrary $k \geq 0$ and where $\mathbb{E}[V_1^2] < 1/b$. Then, (A.14) combined with (A.13) implies our claim. □

Let v and v' be two vertices in the split tree T_n^{sp}. We denote by $d_n(v \wedge v')$ the height of the last common ancestor $v \wedge v'$ of the vertices v and v' in the tree T_n^{sp}.

Lemma A.5. *Assume that Conditions 1 and 2 are fulfilled. For $n \in \mathbb{N}$ fixed, let u_1 and u_2 denote two independent uniformly distributed random vertices in T_n^{sp}. Let $h : \mathbb{N} \to \mathbb{R}_+$ be some function with $\lim_{n\to\infty} h(n) = \infty$. We have that $d_n(u_1 \wedge u_2)h(n)^{-1} \to 0$, as $n \to \infty$, in probability.*

Proof: We follow a similar argument as in the proof Lemma A.4. For $\delta > 0$, note that $d_n(u_1 \wedge u_2) \geq \delta h(n)$ when both vertices lie in the same sub-tree and the height of the last common ancestor related to this sub-tree has to be greater than $\delta h(n)$. For $1 \leq i \leq b^{\lceil \delta h(n) \rceil}$, let v_i be a vertex in T_n^{sp} at height $\lceil \delta h(n) \rceil$ and let N_i be the number of vertices of the sub-tree rooted at v_i. Since u_1 and u_2 are two independent uniformly distributed random vertices in T_n^{sp}, we have that

$$\begin{aligned}
\mathbb{P}(d_n(u_1 \wedge u_2) \geq \delta h(n)) &\leq \mathbb{E}\left[\sum_{i=1}^{b^{\lceil \delta h(n) \rceil}} \left(\frac{N_i}{N}\right)^2\right] \\
&= \mathbb{E}\left[\frac{N^2 - \mathbb{E}^2[N]}{N^2 \mathbb{E}^2[N]} \sum_{i=1}^{b^{\lceil \delta h(n) \rceil}} N_i^2\right] + \mathbb{E}\left[\sum_{i=1}^{b^{\lceil \delta h(n) \rceil}} \left(\frac{N_i}{\mathbb{E}[N]}\right)^2\right].
\end{aligned} \tag{A.15}$$

We analyse the first term at the right-hand side of (A.15). Note that $\sum_{i=1}^{b^{\lceil \delta h(n) \rceil}} N_i^2 \leq N^2$. Then Condition 2 and Remark 1.6 imply that

$$\mathbb{E}\left[\frac{N^2 - \mathbb{E}^2[N]}{N^2 \mathbb{E}^2[N]} \sum_{i=1}^{b^{\lceil \delta h(n) \rceil}} N_i^2\right] \leq \frac{Var(N)}{\mathbb{E}^2[N]} = o(1). \tag{A.16}$$

We now focus in the second term at the right-hand side of (A.15). Note that Condition 2 and Remark 1.6 imply that $\mathbb{E}[N_i^2] = \mathbb{E}\left[Var(N_i|n_i) + \mathbb{E}^2[N_i|n_i]\right] = O(\mathbb{E}[n_i^2])$, where we have used the well-known formula $Var(N_i) = \mathbb{E}[Var(N_i|n_i)] + Var(\mathbb{E}[N_i|n_i])$. Hence the previous estimate, the inequality (A.14) and Condition 2 allow us to conclude that

$$\mathbb{E}\left[\sum_{i=1}^{b^{\lceil \delta h(n) \rceil}} \left(\frac{N_i}{\mathbb{E}[N]}\right)^2\right] = o(1). \tag{A.17}$$

Finally, our claim follows by applying (A.16) and (A.17) into (A.15). □

We complete this section by stating a corollary of the previous lemmas. Let u_1 and u_2 be two independent uniformly chosen vertices in T_n^{sp}. We write $d_n(u_1, u_2)$ for the number of edges of T_n^{sp} which are needed to connect the root, u_1 and u_2. Similarly, let b_1 and b_2 be two independent uniformly chosen balls in T_n^{sp}. We write $D_n(b_1, b_2)$ for the number of edges of T_n^{sp} which are needed to connect the root, and vertices where the balls b_1 and b_2 are stored.

Corollary A.6. *Assume that Condition 1 is fulfilled. We have that $D_n(b_1, b_2)(\ln n)^{-1} \to 2/\mu$, as $n \to \infty$, in probability. If we further assume that Condition 2 is also satisfied. We have that $d_n(b_1, b_2)(\ln n)^{-1} \to 2/\mu$, as $n \to \infty$, in probability.*

Proof: We note that $D_n(b_1, b_2) = D_n(b_1) + D_n(b_2) - D_n(b_1 \wedge b_2)$, where $D_n(b_1)$ has the same distribution as $D_n(b_2)$. Therefore, the first result is a direct consequence of Lemma A.2 and Lemma A.4. The proof of the second claim follows from a similar argument by using Lemma A.3 and Lemma A.5. □

Appendix B. Proof of Theorem 3.5

In this section, we prove Theorem 3.5 which is an important ingredient in the proof of Theorem 1.10. For $1 \leq i \leq m_n$, we denote by $\mathcal{F}_i$ the σ-field generated by $(n_v, d_n(v) \leq i)$. Recall from the beginning of Section 3 that for a vertex $v \in T_n^{\mathrm{sp}}$ that is at height $d_n(v) = i$, we write $(W_{v,k}, k = 1, \dots, i)$ for a sequence of i.i.d. random variables on $[0,1]$ given by the split vectors associated with the vertices on the unique path from v to the root. We denote by $\mathcal{G}_i$ the σ-field generated by $((W_{v,k}, k = 1, \dots, i) : d_n(v) = i)$. Recall the notation ε_v in (3.24) and write

$$\hat{n}_v \coloneqq n \prod_{k=1}^{i} W_{v,k}, \quad \text{and} \quad \hat{\xi}_v \coloneqq e^{-\frac{c}{\mu}} \frac{\ln n}{n} \hat{n}_v \varepsilon_v. \tag{B.1}$$

Note that $\mathcal{G}_i$ is equivalent to the σ-field generated by $(\hat{n}_v, d_n(v) \leq i)$.

We present now some crucial lemmas that are used in the proof of Theorem 3.5. Recall the notation $m_n = \lfloor \beta \log_b \ln n \rfloor$ for $\beta > 0$. Furthermore, through this section we assume that β is large enough. For the sake of simplicity, we introduce the following notation. For any constants $\theta, x > 0$,

$$\alpha'_n \coloneqq \frac{\ln n}{n} \sum_{d_n(v)=m_n} \hat{n}_v e^{-\frac{c}{\mu}\frac{\ln \hat{n}_v}{\ln n}} - c e^{-\frac{c}{\mu}} \sum_{d_n(v)=m_n} \frac{\hat{n}_v}{n} \varpi(\ln \hat{n}_v), \qquad \Delta'_{n,1} \coloneqq \sum_{1 \leq d_n(v) \leq m_n} \mathbb{P}(\hat{\xi}_v \geq x | \mathcal{G}_{m_n}),$$

$$\Delta'_{n,2} \coloneqq \sum_{1 \leq d_n(v) \leq m_n} \mathbb{E}\left[\hat{\xi}_v \mathbb{1}_{\{\hat{\xi}_v \leq \theta\}} | \mathcal{G}_{m_n}\right] \qquad \text{and} \qquad \Delta'_{n,3} \coloneqq \sum_{1 \leq d_n(v) \leq m_n} Var\left(\hat{\xi}_v \mathbb{1}_{\{\hat{\xi}_v \leq \theta\}} | \mathcal{G}_{m_n}\right).$$

Recall also the notation $\Delta_{n,i}$, for $i = 1, 2, 3$, in the statement of Theorem 3.5.

Lemma B.1. *Suppose that Condition 1 holds. Furthermore, if $\ln V_1$ is lattice with span d defined in (1.4), we also assume that Condition 3 holds for some $\varrho \in [0, 1)$. We have that*

(i) $\Delta_{n,1} = \Delta'_{n,1} + o_{\mathrm{p}}(1)$.

(ii) $\displaystyle\sum_{1 \leq d_n(v) \leq m_n} \mathbb{E}\left[\xi_v \mathbb{1}_{\{\xi_v \leq \theta\}} | \mathcal{F}_{m_n}\right] = \Delta'_{n,2} + o_{\mathrm{p}}(1)$.

(iii) $\displaystyle\frac{\ln n}{n} \sum_{d_n(v)=m_n} n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - c e^{-\frac{c}{\mu}} \sum_{d_n(v)=m_n} \frac{n_v \varpi(\ln n_v)}{n} = \alpha'_n + o_{\mathrm{p}}(1)$.

(iv) $\Delta_{n,3} = \Delta'_{n,3} + o_{\mathrm{p}}(1)$.

Lemma B.2. *Suppose that Condition 1 holds. Furthermore, if $\ln V_1$ is lattice with span d defined in (1.4), we also assume that Condition 3 holds for some $\varrho \in [0, 1)$. We have that*

(i) $\mathbb{E}[\Delta'_{n,1}] = \nu([x, \infty)) + o(1)$ *for every* $x > 0$.

(ii) $\displaystyle\mathbb{E}[\Delta'_{n,2}] = \left(\mu m_n + \frac{2c - \sigma^2 + \mu^2}{2\mu} + \ln \theta - \mu \phi\left(\ln\left(\theta^{-1} e^{-\frac{c}{\mu}} \ln n\right)\right) + \psi_\theta(d) - \ln \ln n\right) \frac{c}{\mu} e^{-\frac{c}{\mu}} + o(1)$.

(iii) $\mathbb{E}[\alpha'_n] = e^{-\frac{c}{\mu}} \ln n + c e^{-\frac{c}{\mu}} m_n - c e^{-\frac{c}{\mu}} \varpi(\ln n) + o(1)$.

(iv) $\displaystyle\mathbb{E}[\Delta'_{n,3}] = \theta\left(1 + \psi_\theta(d)\right) \frac{c}{\mu} e^{-\frac{c}{\mu}} + o(1)$.

For any constants $\theta, x > 0$ and β large enough, we define $m'_n \coloneqq \lfloor \frac{1}{2} \log_b \ln n \rfloor$ and we write

$$\Delta''_{n,1} \coloneqq \sum_{m'_n \leq d_n(v) \leq m_n} \mathbb{P}(\hat{\xi}_v \geq x | \mathcal{G}_{m_n}), \qquad \Delta''_{n,2} \coloneqq \sum_{m'_n \leq d_n(v) \leq m_n} \mathbb{E}\left[\hat{\xi}_v \mathbb{1}_{\{\hat{\xi}_v \leq \theta\}} | \mathcal{G}_{m_n}\right] - \alpha'_n$$

and

$$\Delta''_{n,3} := \sum_{m'_n \leq d_n(v) \leq m_n} Var\left(\hat{\xi}_v \mathbb{1}_{\{\hat{\xi}_v \leq \theta\}} | \mathcal{G}_{m_n}\right).$$

Lemma B.3. *Suppose that Condition 1 holds. Furthermore, if $\ln V_1$ is lattice with span d defined in (1.4), we also assume that Condition 3 holds for some $\varrho \in [0,1)$. We have that $Var\Big(\mathbb{E}\left[\Delta''_{n,i}\middle|\mathcal{G}_{m'_n}\right]\Big) = o(1)$, for $i = 1, 2, 3$.*

Lemma B.4. *Suppose that Condition 1 holds. Furthermore, if $\ln V_1$ is lattice with span d defined in (1.4), we also assume that Condition 3 holds for some $\varrho \in [0,1)$. We have that $E\Big[Var\Big(\Delta''_{n,i}\Big|\mathcal{G}_{m'_n}\Big)\Big] = o(1)$, for $i = 1, 2, 3$.*

Proof of Theorem 3.5: For $v \in T_n^{\mathrm{sp}}$ such that $1 \leq d_n(v) \leq m_n$, observe that

$$\mathbb{P}\left(\xi_v \geq x | \mathcal{F}_{m_n}\right) = \mathbb{P}\left(\varepsilon_v \geq x e^{\frac{c}{\mu}} \frac{n}{\ln n} \frac{1}{n_v} \middle| \mathcal{F}_{m_n}\right) = (1 - p_n)\, \mathbb{1}_{\left\{x e^{\frac{c}{\mu}} \frac{n}{\ln n} \frac{1}{n_v} \leq 1\right\}} \leq \frac{c}{\ln n}, \tag{B.2}$$

for $x > 0$. Thus,

$$\lim_{n\to\infty} \sup_{1 \leq d_n(v) \leq m_n} \mathbb{P}\left(\xi_v \geq x | \mathcal{F}_{m_n}\right) = 0, \quad \text{almost surely},$$

for every $x > 0$, which proves (i).

We deduce from Lemma B.1 that $\Delta_{n,1} = \Delta'_{n,1} + o_{\mathrm{p}}(1)$,

$$\Delta_{n,2} = \Delta'_{n,2} - \alpha'_n + e^{-\frac{c}{\mu}} \ln n + \frac{c}{\mu} e^{-\frac{c}{\mu}} \ln\ln n - c e^{-\frac{c}{\mu}} \varpi(\ln n) + \left(\frac{c\mu^2 - c\sigma^2}{2\mu^2} + \mu\phi\left(\ln\left(\theta^{-1} e^{-\frac{c}{\mu}} \ln n\right)\right)\right) \frac{c}{\mu} e^{-\frac{c}{\mu}} + o_{\mathrm{p}}(1)$$

and $\Delta_{n,3} = \Delta'_{n,3} + o_{\mathrm{p}}(1)$. Furthermore, Lemma B.2 shows that the expected value of the previous quantities converge to the right-hand sides of Theorem 3.5 (ii), (iii) and (iv). We complete the proof of Theorem 3.5 by showing that

$$Var(\Delta'_{n,1}) = o(1) \quad \text{for every} \quad x > 0, \quad Var(\Delta'_{n,2} - \alpha'_n) = o(1) \quad \text{and} \quad Var(\Delta'_{n,3}) = o(1). \tag{B.3}$$

Then an application of the Chebyshev's inequality implies Theorem 3.5 (ii), (iii) and (iv).

Thus, we prove (B.3). A similar argument as in the proof of Lemma B.1 implies that

$$\Delta'_{n,1} = \Delta''_{n,1} + o(1), \quad \Delta'_{n,2} - \alpha'_n = \Delta''_{n,2} + o(1) \quad \text{and} \quad \Delta'_{n,3} = \Delta''_{n,3} + o(1).$$

Recall the well-known variance formula $Var(X) = \mathbb{E}[Var(X|\mathcal{G})] + Var(\mathbb{E}[X|\mathcal{G}])$, where X is a random variable and $\mathcal{G}$ is a sub-σ-field. Consequently, a combination of the variance formula with $\mathcal{G} = \mathcal{G}_{m'_n}$, Lemma B.3 and Lemma B.4 show (B.3). This concludes our proof. □

Finally, it only remains to prove Lemmas B.1, B.2, B.3 and B.4. Their proofs are close to those of Lemmas 2.5, 2.6, 2.7 and 2.8 in Holmgren (2011), respectively. However, they are not exactly the same due to the nature of the problem. Therefore, we have decided to give only complete proofs of Lemmas B.1 and B.2, where the main differences appear and moreover, the key estimations for the proofs of Lemmas B.3 and B.4 are developed. Then, to avoid unnecessary repetitions, the interested reader can verify that Lemmas B.3 and B.4 follows along the lines of the proofs of Lemma 2.7 and 2.8 in Holmgren (2011) (see also Holmgren (2010c, Lemmas 2.7 and 2.8)) together with estimations used in the proof of Lemma B.2.

B.1. *Proof of Lemma B.1*. Recall the definition of $(\hat{n}_v, 1 \leq d_n(v) \leq m_n)$ in (B.1). The following result shows that n_v is close to $\hat{n}_v$.

Proposition B.5. *Suppose that Condition 1 holds. For* $0 \leq i \leq m_n$, *let* $v \in T_n^{\mathrm{sp}}$ *such that* $d_n(v) = i$. *For large enough* n, *we have that* $\mathbb{P}\left(|n_v - \hat{n}_v| > n^{0.6}\right) \leq n^{-0.19}$.

Proof: See Holmgren (2012, Lemma 1.1) (which holds also in the lattice case). □

Recall the definition of $(\hat{\xi}_v, 1 \leq d_n(v) \leq m_n)$ in (B.1). It is not difficult to deduce that

$$\mathbb{P}\left(\hat{\xi}_v \geq x | \mathcal{G}_{m_n}\right) = (1-p_n)\, \mathbb{1}_{\left\{xe^{\frac{c}{\mu}} \frac{n}{\ln n} \frac{1}{\hat{n}_v} \leq 1\right\}}, \qquad x > 0. \tag{B.4}$$

Proof of Lemma B.1: We first show (i) for the non-lattice case. The lattice case follows from exactly the same argument. From (B.2), (B.4) and the triangle inequality, we notice that

$$\begin{aligned}\mathbb{E}\left[|\Delta_{n,1} - \Delta'_{n,1}|\right] &\leq (1-p_n) \sum_{i=1}^{m_n} \sum_{d_n(v)=i} \mathbb{E}\left[\left| \mathbb{1}_{\left\{n_v \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right\}} - \mathbb{1}_{\left\{\hat{n}_v \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right\}} \right|\right] \\ &= \frac{c}{\ln n} \sum_{i=1}^{m_n} \sum_{d_n(v)=i} \mathbb{P}\left(n_v \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n}, \hat{n}_v < xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right) \\ &\quad + \frac{c}{\ln n} \sum_{i=1}^{m_n} \sum_{d_n(v)=i} \mathbb{P}\left(n_v < xe^{\frac{c}{\mu}} \frac{n}{\ln n}, \hat{n}_v \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right).\end{aligned}$$

Denote the first term on the right-hand side by I_n^1 and the second term by I_n^2. We first deal with I_n^1 and show that $I_n^1 = o(1)$. For $\delta_1 \in (0,1)$, we observe that

$$\begin{aligned}I_n^1 &\leq \frac{c}{\ln n} \sum_{i=1}^{m_n} \sum_{d_n(v)=i} \mathbb{P}\left(n_v \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n}, \hat{n}_v < \delta_1 xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right) \\ &\quad + \frac{c}{\ln n} \sum_{i=1}^{\infty} \sum_{d_n(v)=i} \mathbb{P}\left(\delta_1 xe^{\frac{c}{\mu}} \frac{n}{\ln n} \leq \hat{n}_v < xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right).\end{aligned} \tag{B.5}$$

If $d_n(v) = i$ for $1 \leq i \leq m_n$, the relationship (3.1) implies that

$$\begin{aligned}&\mathbb{P}\Big(n_v \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n}, \hat{n}_v < \delta_1 xe^{\frac{c}{\mu}} \frac{n}{\ln n}\Big) \\ &\qquad \leq \mathbb{P}\left(\mathrm{binomial}(n, \hat{n}_v/n) \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n} - s_1 i, \hat{n}_v < \delta_1 xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right) \\ &\qquad \leq \mathbb{P}\left(\mathrm{binomial}\left(n, \delta_1 xe^{\frac{c}{\mu}} \frac{1}{\ln n}\right) \geq xe^{\frac{c}{\mu}} \frac{n}{\ln n} - s_1 i\right) \\ &\qquad = \mathbb{P}\left(\mathrm{binomial}\left(n, \delta_1 xe^{\frac{c}{\mu}} \frac{1}{\ln n}\right) - xe^{\frac{c}{\mu}} \frac{\delta_1 n}{\ln n} \geq xe^{\frac{c}{\mu}} \frac{(1-\delta_1)n}{\ln n} - s_1 i\right) \\ &\qquad \leq C_1 (\ln n)/n,\end{aligned} \tag{B.6}$$

for $t \geq 0$ and some constant $C_1 > 0$; where we have used Chebyshev's inequality and the fact that the variance of a binomial(m, q) is $mq(1-q)$, for the last inequality. On the other hand, Lemma 1.7 (i) implies that

$$\lim_{n \to \infty} \frac{c}{\ln n} \sum_{i=1}^{\infty} \sum_{d_n(v)=i} \mathbb{P}\left(\delta_1 xe^{\frac{c}{\mu}} \frac{n}{\ln n} \leq \hat{n}_v < xe^{\frac{c}{\mu}} \frac{n}{\ln n}\right) = (\delta_1^{-1} - 1) c \mu^{-1} x^{-1} e^{-\frac{c}{\mu}}.$$

By combining the previous limit and the estimate (B.6) into (B.5), we obtain that

$$\limsup_{n\to\infty} I_n^1 = (\delta_1^{-1} - 1)c\mu^{-1}x^{-1}e^{-\frac{c}{\mu}}.$$

By the arbitrariness of $\delta_1 \in (0,1)$, we deduce that $I_n^1 = o(1)$. We complete the proof of (i) by showing that $I_n^2 = o(1)$. For $\delta_2 > 1$, we observe that

$$I_n^2 \leq \frac{c}{\ln n}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{P}\left(n_v < xe^{\frac{c}{\mu}}\frac{n}{\ln n}, \hat{n}_v \geq \delta_2 xe^{\frac{c}{\mu}}\frac{n}{\ln n}\right)$$
$$+ \frac{c}{\ln n}\sum_{i=1}^{\infty}\sum_{d_n(v)=i}\mathbb{P}\left(xe^{\frac{c}{\mu}}\frac{n}{\ln n} \leq \hat{n}_v < \delta_2 xe^{\frac{c}{\mu}}\frac{n}{\ln n}\right).$$

But one can show similarly that $I_n^2 = o(1)$; details are left to the reader. Then, an application of the Markov's inequality combined with the previous estimates concludes the proof of (i).

We next establish (ii). Observe that

$$\mathbb{E}\left[\xi_v \mathbb{1}_{\{\xi_v\leq\theta\}}|\mathcal{F}_{m_n}\right] = (1-p_n)\frac{\ln n}{n}e^{-\frac{c}{\mu}}n_v\mathbb{1}_{\left\{n_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}$$

and

$$\Delta'_{n,2} = \mathbb{E}\left[\hat{\xi}_v \mathbb{1}_{\{\hat{\xi}_v\leq\theta\}}|\mathcal{G}_{m_n}\right] = (1-p_n)\frac{\ln n}{n}e^{-\frac{c}{\mu}}\hat{n}_v\mathbb{1}_{\left\{\hat{n}_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}.$$

Then triangle inequality implies that

$$\mathbb{E}\Big[\Big|\sum_{1\leq d_n(v)\leq m_n}\mathbb{E}\left[\xi_v\mathbb{1}_{\{\xi_v\leq\theta\}}|\mathcal{F}_{m_n}\right] - \Delta'_{n,2}\Big|\Big]$$
$$\leq \frac{c}{n}e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{E}\left[\left|n_v\mathbb{1}_{\left\{n_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}} - \hat{n}_v\mathbb{1}_{\left\{\hat{n}_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\right|\right]$$
$$\leq \frac{c}{n}e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{E}\left[|n_v - \hat{n}_v|\,\mathbb{1}_{\left\{n_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\right]$$
$$+ \frac{c}{n}e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{E}\left[\hat{n}_v\left|\mathbb{1}_{\left\{n_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}} - \mathbb{1}_{\left\{\hat{n}_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\right|\right].$$

On the one hand, Proposition B.5 implies that

$$\frac{c}{n}e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{E}\left[|n_v - \hat{n}_v|\,\mathbb{1}_{\left\{n_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\right] \leq \frac{c}{n}e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{E}\left[|n_v - \hat{n}_v|\right] = o(1).$$

On the other hand, a similar computation as in the proof of point (i) shows that

$$\frac{c}{n}e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\sum_{d_n(v)=i}\mathbb{E}\left[\hat{n}_v\left|\mathbb{1}_{\left\{n_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}} - \mathbb{1}_{\left\{\hat{n}_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\right|\right] = o(1).$$

Thus, a combination of the previous estimates with the Markov inequality shows (ii).

We continue with the proof of (iii). An application of the triangle inequality implies that

$$\mathbb{E}\left[\left|\frac{\ln n}{n}\sum_{d_n(v)=m_n} n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - ce^{-\frac{c}{\mu}}\sum_{d_n(v)=m_n}\frac{n_v\varpi(\ln n_v)}{n} - \alpha_n'\right|\right]$$
$$\leq \frac{\ln n}{n}b^{m_n}\mathbb{E}\left[\left|n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - \hat{n}_v e^{-\frac{c}{\mu}\frac{\ln \hat{n}_v}{\ln n}}\right|\right] + \frac{c}{n}e^{-\frac{c}{\mu}}b^{m_n}\mathbb{E}\left[\left|n_v\varpi(\ln n_v) - \hat{n}_v\varpi(\ln \hat{n}_v)\right|\right]. \quad \text{(B.7)}$$

By using Proposition B.5, a similar argument as in the proof of point (ii) shows that

$$\frac{\ln n}{n}b^{m_n}\mathbb{E}\left[\left|n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - \hat{n}_v e^{-\frac{c}{\mu}\frac{\ln \hat{n}_v}{\ln n}}\right|\right]$$
$$\leq \frac{\ln n}{n}b^{m_n}\left(\mathbb{E}\left[\left|n_v - \hat{n}_v\right|\right] + \mathbb{E}\left[\hat{n}_v\left|e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - e^{-\frac{c}{\mu}\frac{\ln \hat{n}_v}{\ln n}}\right|\right]\right) = o(1). \quad \text{(B.8)}$$

On the other hand, the triangle inequality and Proposition B.5 imply that

$$\begin{aligned}\mathbb{E}\left[\left|n_v\varpi(\ln n_v) - \hat{n}_v\varpi(\ln \hat{n}_v)\right|\right] &\leq \mathbb{E}\left[\varpi(\ln n_v)\left|n_v - \hat{n}_v\right|\right] + \mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\right]\\ &= \mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\right] + o(nb^{-m_n}),\end{aligned}$$

where we have used that ϖ is a continuous d-periodic function, with d defined in (1.4), and thus it is bounded. Note that

$$\begin{aligned}\mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\right] &= \mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\mathbb{1}_{\{|n_v-\hat{n}_v|\leq \hat{n}_v^{2/3}\}}\right]\\ &\quad + \mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\mathbb{1}_{\{|n_v-\hat{n}_v|> \hat{n}_v^{2/3}\}}\right]. \end{aligned} \quad \text{(B.9)}$$

It is not difficult to see that in the event $\{|n_v - \hat{n}_v| < \hat{n}_v^{2/3}\}$, we can make $|\ln n_v - \ln \hat{n}_v|$ arbitrary small by taking n large enough. Hence the continuity of the function ϖ allows us to deduce that

$$\mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\right] = \mathbb{E}\left[\hat{n}_v\left|\varpi(\ln n_v) - \varpi(\ln \hat{n}_v)\right|\mathbb{1}_{\{|n_v-\hat{n}_v|> \hat{n}_v^{2/3}\}}\right] + o(nb^{-m_n}). \quad \text{(B.10)}$$

Recall that a binomial random variable with parameters (n, q) has expected value nq and variance $nq(1-q)$. Following the same reasoning as in the proof of Proposition B.5, we deduce from an application of (3.1) and the conditional version of Chebyshev's inequality that

$$\mathbb{E}\left[\hat{n}_v\mathbb{1}_{\{|n_v-\hat{n}_v|> \hat{n}_v^{2/3}\}}\right] = 4\mathbb{E}[\hat{n}_v^{2/3}] = o(nb^{-m_n}). \quad \text{(B.11)}$$

By recalling that the function ϖ is continuous and thus bounded, the estimations (B.9), (B.10) and (B.11) imply that

$$b^{m_n}\mathbb{E}\left[\left|n_v\varpi(\ln n_v) - \hat{n}_v\varpi(\ln \hat{n}_v)\right|\right] = o(n). \quad \text{(B.12)}$$

Therefore, the combination of (B.8) and (B.12) into (B.7) implies

$$\mathbb{E}\left[\left|\frac{\ln n}{n}\sum_{d_n(v)=m_n} n_v e^{-\frac{c}{\mu}\frac{\ln n_v}{\ln n}} - ce^{-\frac{c}{\mu}}\sum_{d_n(v)=m_n}\frac{n_v\varpi(\ln n_v)}{n} - \alpha_n'\right|\right] = o(1)$$

which together with the Markov inequality proves (iii).

Finally, (iv) follows from a similar argument as in the proof of (ii) by using Proposition B.5. □

B.2. *Proof of Lemma B.2.* We observe that $(n_v, d_n(v) = i)$ is a sequence of identically distributed random variables, for $1 \leq i \leq m_n$. Moreover, the distribution of n_v for $v \in T_n^{\mathrm{sp}}$ such that $d_n(v) = i$ is determined by the sequence $(W_{v,k}, k = 1, \dots, i)$ of i.i.d. random variables on $[0, 1]$ given by the split vectors associated with the vertices on the unique path from v to the root. We introduce the notation $Y_{v,i} := -\sum_{k=1}^{i}\ln W_{v,k}$. We sometimes omit the vertex index of $(W_{v,k}, k = 1, \dots, i)$ and we just write $(W_k, k = 1, \dots, i)$ when it is free of ambiguity. Similarly, we write Y_i instead of $Y_{v,i}$.

Proof of Lemma B.2: Recall our assumption (1.8) in the percolation parameter, i.e., $p_n = 1-c/\ln n$, where $c>0$ is fixed. We first show (i) in the non-lattice case. From the identity (B.4), we see that

$$\mathbb{E}\left[\Delta'_{n,1}\right] = (1-p_n)\sum_{i=1}^{m_n}\mathbb{E}\Big[\sum_{d_n(v)=i}\mathbb{1}_{\left\{xe^{\frac{c}{\mu}}\frac{n}{\ln n}\frac{1}{\hat{n}_v}\leq 1\right\}}\Big] = (1-p_n)\sum_{i=1}^{m_n} b^i\mathbb{P}\left(Y_i \leq \ln\left(x^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right).$$

By Lemma 1.7, we obtain that

$$\sum_{i=1}^{\infty} b^i\mathbb{P}\left(Y_i \leq \ln\left(x^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) = \left(\mu^{-1}+o(1)\right)x^{-1}e^{-\frac{c}{\mu}}\ln n = \mu^{-1}e^{-\frac{c}{\mu}}x^{-1}\ln n + o(\ln n). \quad \text{(B.13)}$$

Thus (i) follows from (B.13) by providing that

$$(1-p_n)\sum_{i=m_n+1}^{\infty} b^i\mathbb{P}\left(Y_i \leq \ln\left(x^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) = o(1). \quad \text{(B.14)}$$

Choose an arbitrary $t>0$. By an application of the Markov inequality and the fact that $(W_k, k = 1,\dots,i)$ are i.i.d. random variables, we obtain that

$$\mathbb{P}\left(Y_i \leq \delta\right) = \mathbb{P}\left(e^{-tY_i} \geq e^{-\delta t}\right) \leq m(t)^i e^{\delta t}, \quad \text{(B.15)}$$

for $\delta > 0$, where we define $m(t) \coloneqq \mathbb{E}[V_1^t]$ for $t>0$. Then,

$$(1-p_n)\sum_{i=m_n+1}^{\infty} b^i\mathbb{P}\left(Y_i \leq \ln\left(x^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) \leq cx^{-t}e^{-\frac{c}{\mu}}(\ln n)^{t-1}\sum_{i=m_n+1}^{\infty}(m(t)b)^i. \quad \text{(B.16)}$$

Thus our claim (B.14) follows after some computations by taking $t>0$ such that $bm(t)<1$ (this is possible by Condition 1) and $\beta > \max((1-t)/\log_b(bm(t)), -2/(1+\log_b\mathbb{E}[V_1^2]))$.

In the lattice case, we see that (B.13) becomes

$$\begin{aligned}\sum_{i=1}^{\infty} b^i\mathbb{P}\left(Y_i \leq \ln\left(x^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) &= \left(\frac{d}{\mu}\frac{1}{1-e^{-d}}+o(1)\right)e^{d\lfloor d^{-1}\ln(x^{-1}e^{-\frac{c}{\mu}}\ln n)\rfloor} \\ &= \frac{d}{\mu}\frac{1}{1-e^{-d}}e^{d\lfloor d^{-1}\ln(x^{-1}e^{-\frac{c}{\mu}})+\{d^{-1}\ln\ln n\}\rfloor - d\{d^{-1}\ln\ln n\}}\ln n + o(\ln n),\end{aligned}$$

and the results follows exactly as in the non-lattice case.

We next establish (ii) only in the non-lattice case. The lattice case is similar. Observe that

$$\begin{aligned}\mathbb{E}[\Delta'_{n,2}] &= \frac{\ln n}{n}(1-p_n)e^{-\frac{c}{\mu}}\sum_{i=1}^{m_n}\mathbb{E}\Big[\sum_{d_n(v)=i}\hat{n}_v\mathbb{1}_{\left\{\hat{n}_v\leq\theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\Big] \\ &= ce^{-\frac{c}{\mu}}\sum_{i=1}^{m_n} b^i\mathbb{E}\left[e^{-Y_i}\mathbb{1}_{\left\{Y_i\geq\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right\}}\right].\end{aligned}$$

By noticing that $\mathbb{E}[e^{-Y_i}] = b^{-i}$, we use integration by parts to obtain that

$$\begin{aligned}\mathbb{E}[\Delta'_{n,2}] &= ce^{-\frac{c}{\mu}}m_n - \frac{c\theta}{\ln n}\sum_{i=1}^{m_n} b^i\mathbb{P}\left(Y_i \leq \ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) \\ &\quad - ce^{-\frac{c}{\mu}}\int_0^{\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)} e^{-z}\sum_{i=1}^{m_n} b^i\mathbb{P}(Y_i\leq z)\mathrm{d}z \\ &= ce^{-\frac{c}{\mu}}m_n - \frac{c}{\mu}e^{-\frac{c}{\mu}} - ce^{-\frac{c}{\mu}}\int_0^{\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)} e^{-z}\sum_{i=1}^{m_n} b^i\mathbb{P}(Y_i\leq z)\mathrm{d}z + o(1),\end{aligned}$$

where we have used (B.13) and (B.14), with $t > 0$ such that $bm(t) < 1$ and $\beta > \max((1 - t)/\log_b(bm(t)), -2/(1+\log_b \mathbb{E}[V_1^2]))$, in order to get the last equality.

On the other hand, we deduce from (B.16) that

$$\int_0^{\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)} e^{-z}\sum_{i=m_n+1}^{\infty} b^i \mathbb{P}\left(Y_i \le z\right) \mathrm{d}z \le \theta^{-t} e^{-\frac{c}{\mu}} (\ln n)^t \left(\sum_{i=m_n+1}^{\infty} (m(t)b)^i\right) \ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right) = o(1),$$

when $t > 0$ such that $bm(t) < 1$ (this is possible by Condition 1) and $\beta > \max(-t/\log_b(bm(t)), (1 - t)/\log_b(bm(t)), -2/(1+\log_b \mathbb{E}[V_1^2]))$. Hence

$$\mathbb{E}[\Delta'_{n,2}] = ce^{-\frac{c}{\mu}} m_n - \frac{c}{\mu}e^{-\frac{c}{\mu}} - ce^{-\frac{c}{\mu}} \int_0^{\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)} e^{-z}\sum_{i=1}^{\infty} b^i \mathbb{P}(Y_i \le z)\mathrm{d}z + o(1).$$

By the result in (1.7), we know that

$$\int_0^{\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)} e^{-z}\left(\sum_{i=1}^{\infty} b^i\mathbb{P}(Y_i \le z) - \mu^{-1}e^z\right)\mathrm{d}z = \frac{\sigma^2-\mu^2}{2\mu^2} - \mu^{-1} + \phi\left(\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) + o(1),$$

where $\phi : \mathbb{R} \to \mathbb{R}$ is the d-periodic continuous function in (1.7). Therefore,

$$\begin{aligned}\mathbb{E}[\Delta'_{n,2}] = ce^{-\frac{c}{\mu}} m_n + \frac{2c^2 - c\sigma^2 + c\mu^2}{2\mu^2} e^{-\frac{c}{\mu}} - ce^{-\frac{c}{\mu}}\phi\left(\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right)\\ + \frac{c}{\mu}e^{-\frac{c}{\mu}}\ln\theta - \frac{c}{\mu}e^{-\frac{c}{\mu}}\ln\ln n + o(1)\end{aligned}$$

which proves point (ii).

We continue with the proof of (iii). Recall that $m(t) = \mathbb{E}[V_1^t]$ for $t > 0$. From the definition of $\hat{n}_v$ in (B.1), we deduce that

$$\begin{aligned}\mathbb{E}[\alpha'_n] &= b^{m_n}e^{-\frac{c}{\mu}} m\left(1 - \frac{c}{\mu}\frac{1}{\ln n}\right)^{m_n} \ln n - ce^{-\frac{c}{\mu}} b^{m_n}\mathbb{E}\left[\prod_{k=1}^{m_n} W_k \varpi\left(\ln n + \sum_{k=1}^{m_n}\ln W_k\right)\right]\\ &= b^{m_n}e^{-\frac{c}{\mu}} m\left(1 - \frac{c}{\mu}\frac{1}{\ln n}\right)^{m_n} \ln n - ce^{-\frac{c}{\mu}}\varpi(\ln n),\end{aligned}$$

since ϖ is d-periodic, with d defined in (1.4), and $\ln W_k \in d\mathbb{Z}$. We notice that $m(1) = \mathbb{E}[V_1] = 1/b$ and $m'(1) = \mathbb{E}[V_1 \ln V_1] = -\mu/b$. Then a simple Taylor's expansion calculation shows that

$$m\left(1 - \frac{c}{\mu}\frac{1}{\ln n}\right) = \frac{1}{b} + \frac{c}{b\ln n} + o\left(\frac{1}{b\ln^2 n}\right)$$

which implies that

$$\mathbb{E}[\alpha'_n] = e^{-\frac{c}{\mu}}\ln n + ce^{-\frac{c}{\mu}} m_n - ce^{-\frac{c}{\mu}}\varpi(\ln n) + o(1),$$

and completes the proof of (iii).

We finally show (iv) only in the non-lattice case. The lattice case follows from exactly the same argument. Note that

$$\begin{aligned}\mathbb{E}[\Delta'_{n,3}] &= \frac{\ln^2 n}{n^2}(1-p_n)p_n e^{-2\frac{c}{\mu}} \sum_{i=1}^{m_n}\mathbb{E}\left[\sum_{d_n(v)=i} \hat{n}_v^2 \mathbb{1}_{\left\{\hat{n}_v \le \theta e^{\frac{c}{\mu}}\frac{n}{\ln n}\right\}}\right]\\ &= ce^{-2\frac{c}{\mu}}(\ln n)p_n \sum_{i=1}^{m_n} b^i \mathbb{E}\left[e^{-2Y_i}\mathbb{1}_{\left\{Y_i \ge \ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right\}}\right].\end{aligned}$$

By integration by parts, we obtain that

$$\begin{aligned}\mathbb{E}[\Delta'_{n,3}] &= -\frac{c\theta^2 p_n}{\ln n}\sum_{i=1}^{m_n} b^i \mathbb{P}\left(Y_i \le \ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)\right) \\ &\quad + 2ce^{-2\frac{c}{\mu}}(\ln n)p_n \int_{\ln\left(\theta^{-1}e^{-\frac{c}{\mu}}\ln n\right)}^{\infty} e^{-2z}\sum_{i=1}^{m_n} b^i \mathbb{P}(Y_i \le z)\mathrm{d}z \\ &= c\mu^{-1}e^{-\frac{c}{\mu}}\theta + o(1),\end{aligned}$$

where we have used (B.13) and (B.14), with $t > 0$ such that $bm(t) < 1$ and

$$\beta > \max(-t/\log_b(bm(t)), (1-t)/\log_b(bm(t)), -2/(1+\log_b \mathbb{E}[V_1^2])),$$

to get the last equality. This concludes the proof of (iv). □

Acknowledgements

The authors would like to thank an anonymous referees for the relevant remarks they made on the paper.

References

Asmussen, S. *Applied probability and queues*, volume 51 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition (2003). ISBN 0-387-00211-1. Stochastic Modelling and Applied Probability. MR1978607.

Bertoin, J. Almost giant clusters for percolation on large trees with logarithmic heights. *J. Appl. Probab.*, **50** (3), 603–611 (2013). MR3102504.

Bertoin, J. On the non-Gaussian fluctuations of the giant cluster for percolation on random recursive trees. *Electron. J. Probab.*, **19**, no. 24, 15 (2014a). MR3174836.

Bertoin, J. Sizes of the largest clusters for supercritical percolation on random recursive trees. *Random Structures Algorithms*, **44** (1), 29–44 (2014b). MR3143589.

Bertoin, J. and Uribe Bravo, G. Supercritical percolation on large scale-free random trees. *Ann. Appl. Probab.*, **25** (1), 81–103 (2015). MR3297766.

Berzunza, G. Yule processes with rare mutation and their applications to percolation on b-ary trees. *Electron. J. Probab.*, **20**, no. 43, 23 (2015). MR3339863.

Berzunza, G. The existence of a giant cluster for percolation on large Crump-Mode-Jagers trees. *Adv. in Appl. Probab.*, **52** (1), 266–290 (2020). MR4092814.

Bollobás, B. and Riordan, O. Asymptotic normality of the size of the giant component in a random hypergraph. *Random Structures Algorithms*, **41** (4), 441–450 (2012). MR2993129.

Bourdon, J. Size and path length of Patricia tries: dynamical sources context. *Random Structures Algorithms*, **19** (3-4), 289–315 (2001). MR1871557.

Broutin, N. and Holmgren, C. The total path length of split trees. *Ann. Appl. Probab.*, **22** (5), 1745–1777 (2012). MR3025680.

Cai, X. S. and Holmgren, C. Cutting resilient networks—complete binary trees. *Electron. J. Combin.*, **26** (4), Paper No. 4.43, 28 (2019). MR4045395.

Cai, X. S., Holmgren, C., Devroye, L., and Skerman, F. k-cut on paths and some trees. *Electron. J. Probab.*, **24**, Paper No. 53, 22 (2019). MR3968715.

Coffman, E. G. and Eve, J. File Structures Using Hashing Functions. *Commun. ACM*, **13** (7), 427–432 (1970). DOI: 10.1145/362686.362693.

Devroye, L. On the expected height of fringe-balanced trees. *Acta Inform.*, **30** (5), 459–466 (1993). MR1236537.

Devroye, L. Universal limit laws for depths in random trees. *SIAM J. Comput.*, **28** (2), 409–432 (1999). MR1634354.

Drmota, M. *Random trees. An interplay between combinatorics and probability.* SpringerWien-NewYork, Vienna (2009). ISBN 978-3-211-75355-2. MR2484382.

Drmota, M., Iksanov, A., Moehle, M., and Roesler, U. A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures Algorithms*, **34** (3), 319–336 (2009). MR2504401.

Durrett, R. *Random graph dynamics*, volume 20 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge (2010). ISBN 978-0-521-15016-3. MR2656427.

Feller, W. *An introduction to probability theory and its applications. Vol. II.* John Wiley & Sons, Inc., New York-London-Sydney, second edition (1971). MR0270403.

Finkel, R. A. and Bentley, J. L. Quad trees a data structure for retrieval on composite keys. *Acta Inform.*, **4** (1), 1–9 (1974). DOI: 10.1007/BF00288933.

Flajolet, P., Roux, M., and Vallée, B. Digital trees and memoryless sources: from arithmetics to analysis. In *21st International Meeting on Probabilistic, Combinatorial, and Asymptotic Methods in the Analysis of Algorithms (AofA'10)*, Discrete Math. Theor. Comput. Sci. Proc., AM, pp. 233–260. Assoc. Discrete Math. Theor. Comput. Sci., Nancy (2010). MR2735344.

Geluk, J. L. and de Haan, L. Stable probability distributions and their domains of attraction: a direct approach. *Probab. Math. Statist.*, **20** (1, Acta Univ. Wratislav. No. 2246), 169–188 (2000). MR1785245.

Goldschmidt, C. and Martin, J. B. Random recursive trees and the Bolthausen-Sznitman coalescent. *Electron. J. Probab.*, **10**, no. 21, 718–745 (2005). MR2164028.

Hibbard, T. N. Some combinatorial properties of certain trees with applications to searching and sorting. *J. Assoc. Comput. Mach.*, **9**, 13–28 (1962). MR152155.

Hoare, C. A. R. Quicksort. *Comput. J.*, **5**, 10–15 (1962). MR142216.

Holmgren, C. Novel characteristic of split trees by use of renewal theory. *ArXiv Mathematics e-prints* (2010a). arXiv: 1005.4594.

Holmgren, C. Random records and cuttings in binary search trees. *Combin. Probab. Comput.*, **19** (3), 391–424 (2010b). MR2607374.

Holmgren, C. A Weakly 1-Stable Limiting Distribution for the Number of Random Records and Cuttings in Split Trees. *ArXiv Mathematics e-prints* (2010c). arXiv: 1005.4590.

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

Holmgren, C. Novel characteristic of split trees by use of renewal theory. *Electron. J. Probab.*, **17**, no. 5, 27 (2012). MR2878784.

Holmgren, C. and Janson, S. Fringe trees, Crump-Mode-Jagers branching processes and m-ary search trees. *Probab. Surv.*, **14**, 53–154 (2017). MR3626585.

Iksanov, A. and Möhle, M. A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree. *Electron. Comm. Probab.*, **12**, 28–35 (2007). MR2407414.

Jagers, P. *Branching processes with biological applications.* Wiley Series in Probability and Mathematical Statistics—Applied Probability and Statistics. Wiley-Interscience [John Wiley & Sons], London-New York-Sydney (1975). ISBN 0-471-43652-6. MR0488341.

Janson, S. Random records and cuttings in complete binary trees. In *Mathematics and computer science. III*, Trends Math., pp. 241–253. Birkhäuser, Basel (2004). MR2090513.

Janson, S. Random recursive trees and preferential attachment trees are random split trees. *Combin. Probab. Comput.*, **28** (1), 81–99 (2019). MR3917907.

Kallenberg, O. *Foundations of modern probability.* Probability and its Applications (New York). Springer-Verlag, New York, second edition (2002). ISBN 0-387-95313-2. MR1876169.

Luria, S. E. and Delbrück, M. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics*, **28** (6), 491 (1943). DOI: 10.1093/genetics/28.6.491.

Mahmoud, H. M. On the average internal path length of m-ary search trees. *Acta Inform.*, **23** (1), 111–117 (1986). MR845626.

Mahmoud, H. M. and Pittel, B. Analysis of the space of search trees under the random insertion algorithm. *J. Algorithms*, **10** (1), 52–75 (1989). MR987097.

Meir, A. and Moon, J. W. Cutting down random trees. *J. Austral. Math. Soc.*, **11**, 313–324 (1970). MR0284370.

Möhle, M. Convergence results for compound Poisson distributions and applications to the standard Luria-Delbrück distribution. *J. Appl. Probab.*, **42** (3), 620–631 (2005). MR2157509.

Neininger, R. and Rüschendorf, L. On the internal path length of d-dimensional quad trees. *Random Structures Algorithms*, **15** (1), 25–41 (1999). MR1698407.

Pitman, J. Coalescent random forests. *J. Combin. Theory Ser. A*, **85** (2), 165–193 (1999). MR1673928.

Pyke, R. Spacings. (With discussion.). *J. Roy. Statist. Soc. Ser. B*, **27**, 395–449 (1965). MR216622.

Rösler, U. On the analysis of stochastic divide and conquer algorithms. *Algorithmica*, **29** (1-2), 238–261 (2001). MR1887306.

Schweinsberg, J. Dynamics of the evolving Bolthausen-Sznitman coalecent. *Electron. J. Probab.*, **17**, no. 91, 50 (2012). MR2988406.

Seierstad, T. G. On the normality of giant components. *Random Structures Algorithms*, **43** (4), 452–485 (2013). MR3124692.

Stepanov, V. E. Phase transitions in random graphs. *Teor. Verojatnost. i Primenen.*, **15**, 200–216 (1970). MR0270407.

Walker, A. and Wood, D. Locally balanced binary trees. *Comput. J.*, **19** (4), 322–325 (1976). DOI: 10.1093/comjnl/19.4.322.