Largest Clusters for Supercritical Percolation on Split Trees

Gabriel Berzunza
Department of Mathematics, Uppsala University, Sweden
gabriel.berzunza-ojeda@math.uu.se

Cecilia Holmgren
Department of Mathematics, Uppsala University, Sweden
cecilia.holmgren@math.uu.se

Abstract

We consider the model of random trees introduced by Devroye [13], the so-called random split trees. The model encompasses many important randomized algorithms and data structures. We then perform supercritical Bernoulli bond-percolation on those trees and obtain a precise weak limit theorem for the sizes of the largest clusters. The approach we develop may be useful for studying percolation on other classes of trees with logarithmic height, for instance, we have also studied the case of complete d-regular trees.

2012 ACM Subject Classification Mathematics of computing → Probabilistic algorithms

Keywords and phrases Split trees, random trees, supercritical bond-percolation, cluster size, Poisson measures

Digital Object Identifier 10.4230/LIPIcs.AofA.2020.6

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

1 Introduction

In this extended abstract, we investigate the asymptotic behaviour of the sizes of the largest clusters created by performing Bernoulli bond-percolation on random split trees. Split trees were first introduced by Devroye [13] to encompass many families of trees that are frequently used to model efficient data structures or sorting algorithms (we will be more precise shortly). Some important examples of split trees are binary search trees [18], m-ary search trees [25], quad trees [16], median-of-$(2k+1)$ trees [27], fringe-balanced trees [12], digital search trees [11] and random simplex trees [13, Example 5].

To be more precise, we consider trees 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 tree (i.e., one removes 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 going to be interested in the supercritical regime, in the sense that with high probability, there exists a giant cluster, that is of a size comparable to that of the entire tree.

Bertoin [2] established a simple characterization of tree families with n vertices and percolation regimes which results in giant clusters. Roughly speaking, Bertoin [2] showed 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 approximation of the height of a typical vertex in the tree structure[1]. Then the size Γ_n of the cluster containing the root satisfies $\lim_{n\to\infty} n^{-1}\Gamma_n = \Gamma(c)$ in distribution to some random variable $\Gamma(c) \not\equiv 0$. In several examples the supercritical percolation parameter satisfies

$$p_n = 1 - c/\ln n + o\left(1/\ln n\right), \tag{1}$$

for some fixed parameter $c > 0$. For example, this happens for some important families of random trees with logarithmic height, such as random recursive trees, preferential attachment trees, binary search trees; see [14], [15, Section 4.4]. In those cases the random variable $\Gamma(c)$ is an (explicit) constant and the giant cluster is unique.

A natural problem in this setting is then to estimate the size of the next largest clusters. Concerning trees with logarithmic height, Bertoin [3] proved that in the supercritical regime, the sizes of the next largest clusters of a uniform random recursive tree, normalized by a factor $\ln n/n$, converge to the atoms of some Poisson random measure; see also [1]. This result was extended by Bertoin and Bravo [4] to preferential attachment trees. A different example is the uniform Cayley trees where $\ell(n) = \sqrt{n}$ and $\Gamma(c)$ is not constant. But unlike the previous examples, the number of giant components is unbounded as $n \to \infty$; see [24, 23].

As a motivation, it is important to point out that supercritical Bernoulli bond-percolation on large but finite connected graphs is an ongoing subject of research in statistical physics and mathematics. Furthermore, the estimation of the size of the next largest clusters is a relevant question in this setting. An important example where the graph is not a tree is the case of a complete graph with n vertices. A famous result due to Erdös and Rényi (see [9]) shows that Bernoulli bond-percolation with parameter $p_n = c/n + o(1/n)$ for $c > 1$ fixed, produces with high probability as $n \to \infty$, a unique giant cluster of size close to $\theta(c)n$, where $\theta(c)$ is the unique solution to the equation $x + e^{-cx} = 1$, while the second, third, etc. largest clusters have only size of order $\ln n$.

The main purpose of this work is to investigate the case of random split trees which belong to the family of random trees with logarithmic heights; see Devroye [13]. Informally speaking, a random split tree T_n^{sp} of “size” (or cardinality) n is constructed by first distributing n balls (or keys) among the vertices of an infinite b-ary tree ($b \in \mathbb{N}$) and then removing all sub-trees without balls. Each vertex in the infinite b-ary tree is given a random non-negative split vector $\mathcal{V} = (V_1, \dots, V_b)$ such that $\sum_{i=1}^{b} V_i = 1$ and $V_i \geq 0$, are drawn independently from the same distribution. These vectors affect how balls are distributed. Its exact definition is somewhat lengthy and we postpone it to Section 1.1. An important peculiarity is that the number of vertices of T_n^{sp} is often random which makes the study of split trees usually challenging.

Recently, we have shown in [7, Lemma 1 and Lemma 2] that the supercritical percolation regime in split trees of cardinality n corresponds precisely to parameters fulfilling (1). Notice that here n corresponds to the number of balls (or keys) and not to the number of vertices. More precisely, let C_n^0 (resp. $\hat{C}_n^0$) be the number of balls (resp. number of vertices) in the percolation cluster that contains the root. Then, in the regime (1) and under some mild conditions on the split tree, it holds that

$$n^{-1}C_n^0 \xrightarrow{d} e^{-c/\mu} \qquad \Big(\text{resp. } n^{-1}\hat{C}_n^0 \xrightarrow{d} \alpha e^{-c/\mu}\Big), \qquad \text{as } n \to \infty, \tag{2}$$

where $\mu = b\mathbb{E}[-V_1 \ln V_1]$ ($\alpha > 0$ is some constant depending on the split tree) and $\xrightarrow{d}$ denotes convergence in distribution. Furthermore, the giant cluster is unique. These results agree with that of Bertoin [2] even when the number of vertices in split trees is random and the cluster sizes can be defined as either the number of balls or the number of vertices.

[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 $\lim_{n\to\infty} A_n/B_n = 0$.

Loosely speaking, our main result shows that in the supercritical regime (1) the next largest clusters of a split tree T_n^{sp} have a size of order $n/\ln n$. Moreover, we obtain a limit theorem in terms of certain Poisson random measures. A more precise statement will be given in Theorems 1 and 2 below. These results exhibit that cluster sizes, in the supercritical regime, of split-trees, uniform recursive trees and preferential attachment trees present similar asymptotic behaviour. Finally, we point out that our present approach also applies to study the size of the largest clusters for percolation on complete regular trees (see Theorem 3).

The approach developed in this work differs from that used to study the cases of uniform random recursive trees (RRT) in [3] and preferential attachment trees in [4]. The method of [3] is based on a coupling of Iksanov and Möhle [20] connecting the Meir and Moon [22] algorithm for the isolation of the root in a RRT and a certain random walk. This makes use of special properties of recursive trees (the so-called randomness preserving property, i.e., if one removes an edge from a RRT, then the two resulting subtrees, conditionally on their sizes, are independent RRT's) which fail for split-trees. The basic idea of [4] is based on the close relation of preferential attachment trees with Markovian branching processes and the dynamical incorporation of percolation as neutral mutations. The recent work of Berzunza [5] shows that one can also relate percolation on some types of split trees (but not all) with general age-dependent branching processes (or Crump-Mode-Jagers processes) with neutral mutations. However, the lack of the Markov property in those general branching processes makes the idea of [4] difficult to implement.

A common feature in these previous works, namely [3] and [4], is that, even though one addressed a static problem, one can consider a dynamical version in which edges are removed, respectively vertices inserted, one after the other in a certain order as time passes. Here we use a fairly different route and view percolation on split trees as a static problem.

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

1.1 Random split trees

In this section, we introduce the split tree generating algorithm with parameters $b, s, s_0, s_1, \mathcal{V}$ and n introduced by Devroye [13]. 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 $0 < s$, $0 \leq s_0 \leq s$, $0 \leq bs_1 \leq s + 1 - s_0$. 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 $\mathbb{T}$ as a bucket with capacity s and we assign to each vertex $u \in \mathbb{T}$ an independent copy $\mathcal{V}_u = (V_{u,1}, \dots, V_{u,b})$ of the random split vector $\mathcal{V}$. 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. The split tree 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 labeled 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 we obtain T_n^{sp} by deleting all vertices $u \in \mathbb{T}$ such that the sub-tree rooted at u contains no balls. Note that an internal vertex of T_n^{sp} contains exactly s_0 balls, while a leaf contains a random amount in $\{1, ..., s\}$. Notice also that in general the number N of vertices of T_n^{sp} is a random variable while the number of balls n is deterministic.

It is important to mention that depending on the choice of the parameters b, s, s_0, s_1 and the distribution of $\mathcal{V}$, several important data structures may be modeled. For instance, binary search trees correspond to $b = 2$, $s = s_0 = 1$, $s_1 = 0$ and $\mathcal{V}$ distributed as $(U, 1-U)$, where U is an uniform random variable 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 tree; see [13, 19, 10], for details and more examples.

In the present work, we assume without loss of generality that the components of the split vector $\mathcal{V}$ are identically distributed; this can be done by using random permutations as explained in [13]. In particular, we have that $\mathbb{E}[V_1] = 1/b$. We frequently use the following notation. Set

$$\mu \coloneqq b\mathbb{E}[-V_1 \ln V_1]. \tag{3}$$

Note that $\mu \in (0, \ln b)$. The quantity was first introduced by Devroye [13] 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:

▶ **Condition 1.** Assume that $\mathbb{P}(V_1 = 1) = \mathbb{P}(V_1 = 0) = 0$ and that V_1 is not monoatomic, that is, $V_1 \neq 1/b$.

We sometimes consider the following condition:

▶ **Condition 2.** Suppose that $\ln V_1$ is non-lattice. Furthermore, for some $\alpha > 0$ and $\varepsilon > 0$,

$$\mathbb{E}[N] = \alpha n + O\left(\frac{n}{\ln^{1+\varepsilon} n}\right).$$

Recall that for two sequences of real numbers $(A_n)_{n\geq 1}$ and $(B_n)_{n\geq 1}$ such that $B_n > 0$, one writes $A_n = O(B_n)$ if $\limsup_{n\to\infty} |A_n|/B_n < \infty$. Condition 2 first appears in [10, equation (52)] for the study of the total path length of split trees.

Holmgren [19, Theorem 1.1] showed that if $\ln V_1$ is non-lattice 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 for our purpose since an extra control in $\mathbb{E}[N]$ is needed (see Theorem 2 below). On the other hand, Condition 2 is satisfied in many interesting cases. For instance, it holds for m-ary search trees [21]. Moreover, Flajolet et al. [17] showed that for most tries (as long as $\ln V_1$ is non-lattice) Condition 2 holds. However, there are some special cases of random split trees that do not satisfy Condition 2. For instance, tries (where $s = 1$ and $s_0 = 0$) with a fixed split vector $(1/b, \dots, 1/b)$, in which case $\ln V_1$ is lattice.

1.2 Main results

In this section, we present the main results of this work. We consider Bernoulli bond-percolation with supercritical parameter p_n satisfying (1) on T_n^{sp}. We denote by C_0 (resp. $\hat{C}_0$) the number of balls (resp. the number of vertices) of the cluster that contains the root and by $C_1 \geq C_2 \geq \cdots$ (resp. $\hat{C}_1 \geq \hat{C}_2 \geq \cdots$) the sequence of the number of balls (resp. the number of vertices) of the remaining clusters ranked in decreasing order. For the sake of simplicity, we have decided to remove the parameter n from our notation of C_i and $\hat{C}_i$.

We now state the central results of this work. The first result corresponds to the size being defined as the number of balls in the cluster.

▶ **Theorem 1.** *Let T_n^{sp} be a split tree that satisfies Condition 1 and suppose that p_n fulfills (1). Then,*

$$n^{-1}C_0 \xrightarrow{d} e^{-c/\mu}, \qquad as \;\; n \to \infty,$$

where μ is the constant defined in (3) and c is defined in (1). Furthermore, for every fixed $i \in \mathbb{N}$, we have the convergence in distribution

$$\left(\frac{\ln n}{n} C_1, \ldots, \frac{\ln n}{n} C_i \right) \xrightarrow{d} (\mathrm{x}_1, \ldots, \mathrm{x}_i), \qquad as \;\; n \to \infty,$$

where $\mathrm{x}_1 > \mathrm{x}_2 > \cdots$ denotes the sequence of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $c\mu^{-1}e^{-c/\mu}x^{-2}\mathrm{d}x$.

The second result corresponds to the size being defined as the number of vertices in the cluster.

▶ **Theorem 2.** *Let T_n^{sp} be a split tree that satisfies Conditions 1-2 and suppose that p_n fulfills (1). Then,*

$$n^{-1}\hat{C}_0 \xrightarrow{d} \alpha e^{-c/\mu}, \qquad as \;\; n \to \infty,$$

where μ is the constant defined in (3), α is defined in Condition 2 and c is defined in (1). Furthermore, for every fixed $i \in \mathbb{N}$, we have the convergence in distribution

$$\left(\frac{\ln n}{n} \hat{C}_1, \ldots, \frac{\ln n}{n} \hat{C}_i \right) \xrightarrow{d} (\mathrm{x}_1, \ldots, \mathrm{x}_i), \qquad as \;\; n \to \infty,$$

where $\mathrm{x}_1 > \mathrm{x}_2 > \cdots$ denotes the sequence of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $c\alpha\mu^{-1}e^{-c/\mu}x^{-2}\mathrm{d}x$.

Alternatively, the law of the limiting sequence in Theorems 1 and 2 can be described as follows: for $i \in \mathbb{N}$, $1/\mathrm{x}_1, 1/\mathrm{x}_2 - 1/\mathrm{x}_1, \ldots, 1/\mathrm{x}_i - 1/\mathrm{x}_{i-1}$ are i.i.d. exponential random variables with parameter $c\mu^{-1}e^{-c/\mu}$ in Theorem 1, while in Theorem 2 they are exponential random variables with parameter $c\alpha\mu^{-1}e^{-c/\mu}$.

It is important to point out the similarity with the results for uniform random recursive trees in [3] and preferential attachment trees in [4]. More precisely, the size of the second largest cluster, and more generally, the size of the i-th largest cluster (for $i \geq 2$) in the supercritical regime is of order $n/\ln n$ as in [3] and [4]. Moreover, their sizes are described by the atoms of a Poisson random measure on $(0, \infty)$ whose intensity measure only differ by a constant factor. For example, for uniform random recursive trees [3] the intensity is $ce^{-c}x^{-2}\mathrm{d}x$.

As we mentioned in the introduction, we shall follow a different route to that used in [3] and [4]. The approach developed in this work is based on a remark made in [2, Section 3] about the behavior of the second largest cluster created by performing (supercritical) Bernoulli bond-percolation on complete regular trees. More precisely, consider a rooted complete regular d-ary tree T_h^{d} of height $h \in \mathbb{N}$, where $d \geq 2$ is some integer (i.e., each vertex has exactly out-degree d). Notice that there are d^k vertices at distance $k = 0, 1, \dots, h$ from the root and a total of $(d^{h+1} - 1)/(d - 1)$ vertices. We then perform Bernoulli bond-percolation with parameter

$$q_h = 1 - ch^{-1} + o(h^{-1}), \tag{4}$$

where $c > 0$ is some fixed parameter. It has been shown in [2, Section 3] that this choice of the percolation parameter corresponds precisely to the supercritical regime, that is, the root cluster is the unique giant component. Because the subtree rooted at a vertex at height $i \leq h$ is again a complete regular d-ary tree with height $h - i$, [2, Corollary 1] essentially shows that the size (number of vertices) G_h^1 of the largest cluster which does not contain the root is close to

$$e^{-c} d^{h - \tau_1(h) + 1}/(d - 1),$$

where $\tau_1(h)$ is the smallest height at which an edge has been removed. Notice that there are $d(d^i - 1)/(d - 1)$ edges with height at most i, so the distribution of $\tau_1(h)$ is given by

$$\mathbb{P}(\tau_1(h) > i) = q_h^{d(d^i - 1)/(d-1)}, \qquad i = 1, \dots, h.$$

We use the notation $\log_d x = \ln x / \ln d$ for the logarithm with base d of $x > 0$, and $y = \lfloor y \rfloor + \{y\}$ for the decomposition of a real number y as the sum of its integer and fractional parts. It follows that in the regime (4) and as soon as one assumes $\{\log_d h\} \to \rho \in [0, 1)$, as $h \to \infty$, then $\tau_1(h) - \log_d h$ converges in distribution, and therefore, $hd^{-h} G_h^1$ also converges in distribution.

Our strategy is then to adapt and improve the above argument to study the size of the i-th largest cluster, for $i \geq 2$, in a random split tree with n balls. We also show that this approach can be used to obtain a result similar as Theorem 1 or Theorem 2 for supercritical percolation on complete d-regular trees of height $h \in \mathbb{N}$. More precisely, write G_0 for the number of vertices of the cluster that contains the root and $G_1 \geq G_2 \geq \cdots$ for the sequence of the number vertices of the remaining clusters ranked in decreasing order; for simplicity, we omit the parameter h from our notation. We introduce for every $\rho \in [0, 1)$ a measure Λ_ρ on $(0, \infty)$ by letting

$$\Lambda_\rho([x, \infty)) := d^{-\rho + \lfloor \rho - \log_d x \rfloor + 1}/(d - 1), \qquad x > 0.$$

▶ **Theorem 3.** *Let T_h^{d} be a complete regular d-ary tree of height $h \in \mathbb{N}$ such that $\{\log_d h\} \to \rho \in [0, 1)$, as $h \to \infty$, and suppose that q_h fulfills (4). Then,*

$$d^{-h} G_0 \xrightarrow{d} de^{-c}/(d - 1), \qquad \text{as } h \to \infty,$$

where the constant c is defined in (4). Furthermore, for every fixed $i \in \mathbb{N}$, we have the convergence in distribution

$$(hd^{-h} G_1, \dots, hd^{-h} G_i) \xrightarrow{d} (\mathrm{x}_1, \dots, \mathrm{x}_i), \qquad \text{as } h \to \infty,$$

where $\mathrm{x}_1 \geq \mathrm{x}_2 \geq \cdots$ denotes the sequence of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $c \frac{d}{d-1} e^{-c} \Lambda_\rho(\mathrm{d}x)$.

We conclude this extended abstract by providing in Section 2 a fair enough guideline of the proof of Theorem 1. The proofs of Theorem 2 and Theorem 3 follows by an adaptation of the arguments used in the proof of Theorem 1. An important ingredient in the proof of Theorem 1 is Lemma 5 that establishes a law of large number for the number of sub-trees in T_n^{sp} with cardinality (number of balls) larger than $n/\ln n$, which may be of independent interest. Detailed proofs of all our results are going to be given in the complete version [6].

2 Proof of Theorem 1

We split the proof of Theorem 1 in two parts. We start by studying the sizes of percolated sub-trees that are close to the root. One could refer to these percolated sub-trees as the early clusters since their distance to the root is the smallest. Then we show that the largest percolation clusters can be found amongst those (early) percolated sub-trees.

2.1 Sizes of early clusters

For $i \in \mathbb{N}$, let $\mathbf{e}_{i,n}$ be the edge with the i-th smallest height (we break ties by ordering the edges from left to right, however, the order is not relevant in the proofs) that has been removed and $\mathbf{v}_{i,n}$ the endpoint (vertex) of $\mathbf{e}_{i,n}$ that is the furthest away from the root of T_n^{sp}. Let $T_{i,n}$ be the sub-tree of T_n^{sp} that is rooted at $\mathbf{v}_{i,n}$ and let $n_{i,n}$ be the number of balls stored in the sub-tree $T_{i,n}$. For $t \in [0, \infty)$, we write

$$N_n(0) := 0 \quad \text{and} \quad N_n(t) := \sum_{i \geq 1} \mathbb{1}_{\left\{n_{i,n} \geq \frac{n}{t \ln n}\right\}} = \sum_{i \geq 1} \mathbb{1}_{\left\{(n/n_{i,n}) \frac{1}{\ln n} \leq t\right\}}$$

for the number of sub-trees $T_{i,n}$ that store more than $\lfloor n/(t \ln n) \rfloor$ balls.

▶ **Theorem 4.** *Suppose that Condition 1 holds and that p_n fulfills (1). Then, the following convergence holds in the sense of weak convergence of finite dimensional distributions,*

$$(N_n(t), t \geq 0) \xrightarrow{d} (N(t), t \geq 0), \qquad \text{as } n \to \infty,$$

where $(N(t), t \geq 0)$ is a (classical) Poisson process with intensity $c\mu^{-1}$.

We stress that the convergence in Theorem 4 can be improved in order to show convergence in distribution of the process $(N_n(t), t \geq 0)$ for the Skorohod topology on the space $\mathbb{D}([0, \infty), \mathbb{R})$ of right-continuous functions with left limits to a Poisson process with intensity $c\mu^{-1}$; see, for instance, [8, Theorem 12.6, Chapter 3].

The proof of Theorem 4 uses the following result which provides a law of large number for the number of sub-trees in T_n^{sp} with cardinality larger than $n/\ln n$. More precisely, for a vertex $v \in T_n^{\mathrm{sp}}$ that is not the root $\circ$, let n_v denote the number of balls stored in the sub-tree of T_n^{sp} rooted at v. Define

$$M_n(t) := \# \left\{ v \in T_n^{\mathrm{sp}} : v \neq \circ \quad \text{and} \quad n_v \geq \frac{n}{t \ln n} \right\}, \qquad \text{for } t \in [0, \infty).$$

▶ **Proposition 5.** *Suppose that Condition 1 holds. Then, for every fixed $t \in [0, \infty)$, we have that $(\ln n)^{-1} M_n(t) \to \mu^{-1} t$, in probability, as $n \to \infty$.*

The proof of Proposition 5 is rather technical and it is given in the complete version [6].

Proof of Theorem 4. For a vertex $v \in T_n^{\mathrm{sp}}$ that is not the root $\circ$, let $\mathbf{e}_v$ be the edge that connects v with its parent. Define the event $E_v :=$ {the edge $\mathbf{e}_v$ has been removed after percolation} and write $\xi_v := \mathbb{1}_{E_v}$. So, $(\xi_v)_{v \neq \circ}$ is a sequence of i.i.d. Bernoulli random variables with parameter $1 - p_n$ (that is, the probability of removing an edge). Then, it should be clear that

$$N_n(t) = \sum_{v \neq \circ} \mathbb{1}_{\left\{n_v \geq \frac{n}{t \ln n}\right\}} \xi_v, \qquad t \in [0, \infty).$$

Let Ω be the σ-algebra generated by $(n_v)_{v \neq \text{o}}$. Conditioning on Ω, we have that $(N_n(t), t \geq 0)$ has independent increments and that for $0 \leq s \leq t$, $N_n(t) - N_n(s) \overset{d}{=} \operatorname{Bin}(M_n(t) - M_n(s), 1-p_n)$, where $\operatorname{Bin}(m,q)$ denotes a binomial (m,q) random variable. Therefore, our claim follows from Proposition 5 by appealing to [8, Theorem 12.6, Chapter 3]. ◀

▶ **Corollary 6.** *Suppose that Condition 1 holds and that p_n fulfills (1). Then, for every fixed $i \in \mathbb{N}$, we have the convergence in distribution*

$$\left(\frac{\ln n}{n} n_{1,n}, \dots, \frac{\ln n}{n} n_{i,n} \right) \xrightarrow{d} (\mathrm{x}_1, \dots, \mathrm{x}_i), \qquad as \ \ n \to \infty,$$

where $\mathrm{x}_1 > \mathrm{x}_2 > \cdots$ denotes the sequence of the atoms of a Poisson random measure on $(0,\infty)$ with intensity $c\mu^{-1}x^{-2}\mathrm{d}x$.

Proof. Notice that $(n/n_{1,n})\frac{1}{\ln n} \leq (n/n_{2,n})\frac{1}{\ln n} \leq \cdots$ is the sequence of atoms (or occurrence times) of the counting process $(N_n(t), t \geq 0)$ ranked in increasing order. Then our claim follows directly from Theorem 4, the mapping theorem ([8, Theorem 2.7, Chapter 1]) and basic properties of Poisson random measures (see [26, Proposition 3.7, Chapter 3]). ◀

2.2 Asymptotic sizes of the largest percolation clusters

Recall that, for $i \in \mathbb{N}$, we let $\mathbf{e}_{i,n}$ be the edge with the i-th smallest height that has been removed and $\mathbf{v}_{i,n}$ the endpoint (vertex) of $\mathbf{e}_{i,n}$ that is the furthest away from the root of T_n^{sp}. Recall also that $T_{i,n}$ denotes the sub-tree of T_n^{sp} that is rooted at $\mathbf{v}_{i,n}$ and that we write $n_{i,n}$ for the number of balls stored in the sub-tree $T_{i,n}$. We denote by $\tilde{C}_i$ the size (number of balls) of the root-cluster of $T_{i,n}$ after performing percolation (where here of course root means $\mathbf{v}_{i,n}$). We also write $\tilde{C}_i^*$ for the size (number of balls) of the second largest cluster of $T_{i,n}$ that does not contain its root. In the sequel, we shall use the following 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 $\lim_{n\to\infty} |A_n - B_n|/f(n) = 0$ in probability.

▶ **Proposition 7.** *Suppose that Condition 1 holds and that p_n fulfills (1). For every fixed $i \in \mathbb{N}$, $\tilde{C}_i^* = o_{\mathrm{p}}(n/\ln n)$. Furthermore, we have the convergence in distribution*

$$\left(\frac{\tilde{C}_1}{n_{1,n}}, \dots, \frac{\tilde{C}_i}{n_{i,n}} \right) \xrightarrow{d} (e^{-c/\mu}, \dots, e^{-c/\mu}), \qquad as \ \ n \to \infty.$$

Proof. Notice that it is enough to show our claim for $i = 1$ since convergence in distribution to a constant is equivalent to convergence in probability, and thus, one can easily deduce the joint convergence for every fixed $i \in \mathbb{N}$. Given $n_{1,n}$, we see that $T_{1,n}$ is a split tree with $n_{1,n}$ balls. Notice that supercritical Bernoulli bond-percolation in $T_{1,n}$ corresponds to a percolation parameter satisfying

$$1 - p_{n_{1,n}} = c/\ln n_{1,n} + o\left(1/\ln n_{1,n}\right),$$

where $c > 0$ is fixed. Notice also that Corollary 6 implies that $(\ln n_{1,n})/\ln n \to 1$, in probability, as $n \to \infty$. Hence $1 - p_{n_{1,n}} = 1 - p_n + o_{\mathrm{p}}\left(1/\ln n\right)$. Therefore, a simple application of [7, Lemma 2] shows that $\tilde{C}_1/n_{1,n} \to e^{-c/\mu}$, in distribution, as $n \to \infty$, which proves the second assertion. Moreover, [7, Lemma 2] also shows that $\tilde{C}_1^*/n_{1,n} \to 0$, in distribution, as $n \to \infty$, and by Corollary 6, we conclude that $\tilde{C}_1^* = o_{\mathrm{p}}(n/\ln n)$. This completes the proof. ◀

▶ **Corollary 8.** *Suppose that Condition 1 holds and that p_n fulfills (1). Then, for every fixed $i \in \mathbb{N}$, we have the convergence in distribution*

$$\left(\frac{\ln n}{n}\tilde{C}_1, \ldots, \frac{\ln n}{n}\tilde{C}_i\right) \xrightarrow{d} (\mathrm{x}_1, \ldots, \mathrm{x}_i), \qquad \textit{as } n \to \infty,$$

where $\mathrm{x}_1 > \mathrm{x}_2 > \cdots$ *denotes the sequence of the atoms of a Poisson random measure on* $(0, \infty)$ *with intensity* $c\mu^{-1}e^{-c/\mu}x^{-2}\mathrm{d}x$*.*

Proof. This follows from Corollary 6, Proposition 7, the mapping theorem ([8, Theorem 2.7, Chapter 1]) and basic distributional properties of the atoms of Poisson random measures. ◀

The last ingredient in the proof of Theorem 1 consists in verifying that for every fixed $i \in \mathbb{N}$, one can choose $\ell \in \mathbb{N}$ large enough such that with probability tending to 1, as $n \to \infty$, the i-th largest percolation cluster of T_n^{sp} can be found amongst the root-clusters of the percolated tree-components $T_{1,n}, \ldots, T_{\ell,n}$. Rigorously, denote by

$$\tilde{C}_{1,\ell} \geq \tilde{C}_{2,\ell} \geq \cdots \geq \tilde{C}_{\ell,\ell}$$

the rearrangement in decreasing order of the $\tilde{C}_i$ for $i = 1, \ldots, \ell$. We then adapt the idea of [3, Lemma 6] (details are given in the complete version [6]).

▶ **Lemma 9.** *Suppose that Condition 1 holds and that p_n fulfills (1). Then for each fixed $i \in \mathbb{N}$,*

$$\lim_{\ell \to \infty} \liminf_{n \to \infty} \mathbb{P}\left(\tilde{C}_{k,\ell} = C_k \quad \textit{for every } \ k = 1, \ldots, i\right) = 1.$$

We can now finish the proof of Theorem 1.

Proof of Theorem 1. We have already proven the first claim in [7, Lemma 2]. We then only prove the second claim. For every fixed $i \in \mathbb{N}$, consider a continuous function $f : [0, \infty)^i \to [0, 1]$ and fix $\varepsilon > 0$. According to Lemma 9, we may choose $\ell \in \mathbb{N}$ sufficiently large so that there exists $n_\varepsilon \in \mathbb{N}$ such that the upper bound

$$\mathbb{E}\left[f\left(\frac{\ln n}{n}C_1, \ldots, \frac{\ln n}{n}C_i\right)\right] \leq \mathbb{E}\left[f\left(\frac{\ln n}{n}\tilde{C}_{1,\ell}, \ldots, \frac{\ln n}{n}\tilde{C}_{i,\ell}\right)\right] + \varepsilon$$

holds for all $n \geq n_\varepsilon$. We now deduce from Corollary 8 and the previous bound that

$$\limsup_{n \to \infty} \mathbb{E}\left[f\left(\frac{\ln n}{n}C_1, \ldots, \frac{\ln n}{n}C_i\right)\right] \leq \mathbb{E}\left[f\left(\mathrm{x}_1, \ldots, \mathrm{x}_i\right)\right] + \varepsilon.$$

Because ε can be arbitrary small and f replaced by $1 - f$, this establishes Theorem 1. ◀

References

1 Erich Baur. Percolation on random recursive trees. *Random Structures Algorithms*, 48(4):655–680, 2016. doi:10.1002/rsa.20603.

2 Jean Bertoin. Almost giant clusters for percolation on large trees with logarithmic heights. *J. Appl. Probab.*, 50(3):603–611, 2013. doi:10.1239/jap/1378401225.

3 Jean Bertoin. Sizes of the largest clusters for supercritical percolation on random recursive trees. *Random Structures Algorithms*, 44(1):29–44, 2014. doi:10.1002/rsa.20448.

4 Jean Bertoin and Gerónimo Uribe Bravo. Supercritical percolation on large scale-free random trees. *Ann. Appl. Probab.*, 25(1):81–103, 2015. doi:10.1214/13-AAP988.

5 Gabriel Berzunza. The existence of a giant cluster for percolation on large crump-mode-jagers trees. *To appears in Advances in Applied Probability, arXiv preprint*, 2020. arXiv:1806.10686.

6 Gabriel Berzunza and Cecilia Holmgren. The asymptotic distribution of cluster sizes for supercritical percolation on random split trees. *Preprint*, 2020.

7 Gabriel Berzunza, Xing Shi Cai, and Cecilia Holmgren. The asymptotic non-normality of the giant cluster for percolation on random split trees. *arXiv e-prints*, February 2019. arXiv:1902.08109.

8 Patrick Billingsley. *Convergence of probability measures.* Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons, Inc., New York, second edition, 1999. A Wiley-Interscience Publication. doi:10.1002/9780470316962.

9 Béla Bollobás. *Random graphs*, volume 73 of *Cambridge Studies in Advanced Mathematics.* Cambridge University Press, Cambridge, second edition, 2001. doi:10.1017/CBO9780511814068.

10 Nicolas Broutin and Cecilia Holmgren. The total path length of split trees. *Ann. Appl. Probab.*, 22(5):1745–1777, 2012. doi:10.1214/11-AAP812.

11 Edward G Coffman Jr and James Eve. File structures using hashing functions. *Communications of the ACM*, 13(7):427–432, 1970.

12 Luc Devroye. On the expected height of fringe-balanced trees. *Acta Inform.*, 30(5):459–466, 1993. doi:10.1007/BF01210596.

13 Luc Devroye. Universal limit laws for depths in random trees. *SIAM J. Comput.*, 28(2):409–432, 1999. doi:10.1137/S0097539795283954.

14 Michael Drmota. *Random trees.* SpringerWienNewYork, Vienna, 2009. An interplay between combinatorics and probability. doi:10.1007/978-3-211-75357-6.

15 Rick Durrett. *Random graph dynamics*, volume 20 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge, 2010.

16 Raphael A. Finkel and Jon Louis Bentley. Quad trees a data structure for retrieval on composite keys. *Acta informatica*, 4(1):1–9, 1974.

17 Philippe Flajolet, Mathieu Roux, and Brigitte Vallée. 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, pages 233–260. Assoc. Discrete Math. Theor. Comput. Sci., Nancy, 2010.

18 C. A. R. Hoare. Quicksort. *Comput. J.*, 5:10–15, 1962. doi:10.1093/comjnl/5.1.10.

19 Cecilia Holmgren. Novel characteristic of split trees by use of renewal theory. *Electron. J. Probab.*, 17:no. 5, 27, 2012. doi:10.1214/EJP.v17-1723.

20 Alex Iksanov and Martin Möhle. 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. doi:10.1214/ECP.v12-1253.

21 Hosam M. Mahmoud and Boris Pittel. Analysis of the space of search trees under the random insertion algorithm. *J. Algorithms*, 10(1):52–75, 1989. doi:10.1016/0196-6774(89)90023-0.

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

23 J. Pitman. *Combinatorial stochastic processes*, volume 1875 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin, 2006. Lectures from the 32nd Summer School on Probability Theory held in Saint-Flour, July 7–24, 2002, With a foreword by Jean Picard.

24 Jim Pitman. Coalescent random forests. *J. Combin. Theory Ser. A*, 85(2):165–193, 1999. doi:10.1006/jcta.1998.2919.

25 R. Pyke. Spacings. (With discussion.). *J. Roy. Statist. Soc. Ser. B*, 27:395–449, 1965. URL: http://links.jstor.org/sici?sici=0035-9246(1965)27:3<395:S>2.0.CO;2-C&origin=MSN.

26 Sidney I. Resnick. *Extreme values, regular variation, and point processes*, volume 4 of *Applied Probability. A Series of the Applied Probability Trust.* Springer-Verlag, New York, 1987. doi:10.1007/978-0-387-75953-1.

27 A Walker and Derick Wood. Locally balanced binary trees. *The Computer Journal*, 19(4):322–325, 1976.