Non-Gaussian fluctuations of the giant cluster for percolation on random recursive trees

Jean Bertoin

▶ **To cite this version:**

Jean Bertoin. Non-Gaussian fluctuations of the giant cluster for percolation on random recursive trees. 2013. hal-00819320v1

HAL Id: hal-00819320

https://hal.science/hal-00819320v1

Preprint submitted on 30 Apr 2013 (v1), last revised 21 May 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Non-Gaussian fluctuations of the giant cluster for percolation on random recursive trees

Jean Bertoin*†

Abstract

We consider a Bernoulli bond percolation on a random recursive tree of size $n \gg 1$, with supercritical parameter $p_n = 1 - c/\ln n$ for some $c > 0$ fixed. It is known that with high probability, there exists then a unique giant cluster of size $G_n \sim \mathrm{e}^{-c}$. We show here that G_n has non-gaussian fluctuations. The approach relies on the analysis of the effect of percolation on different phases of the growth of recursive trees.

Key words: Random recursive tree, giant cluster, fluctuations, super-critical percolation.

1 Introduction and main result

A famous result due to Erdös and Rényi shows that Bernoulli bond percolation on the complete graph with n vertices and with parameter c/n for $c > 1$ fixed, produces with high probability as $n \to \infty$, a unique giant cluster of size $\Gamma_n \sim \theta(c)n$, where $\theta(c)$ is the strictly positive solution to the equation $x + \mathrm{e}^{-cx} = 1$. It has been known since the work of Stepanov [20] that the fluctuations of Γ_n are normal, in the sense that

$$\frac{\Gamma_n - \theta(c)n}{\sqrt{n}} \Longrightarrow \mathcal{N}(0, \sigma_c^2),$$

where as usual $\mathcal{N}(0, \sigma_c^2)$ denotes a centered Gaussian variable with variance σ_c^2, and $\Rightarrow$ means convergence in law. See also e.g. Pittel [18] and Barraez *et al.* [1] for alternative proofs and refinements. Since then, several results have appeared in the literature, establishing the

*Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich
†email: jean.bertoin@math.uzh.ch

asymptotic normality of giant components for various random graph models. We refer in particular to Behrisch *et al.* [2], Bollobàs and Riordan [6], and Seierstad [19].

The purpose of the present work is to point at a much different asymptotic behavior for the giant cluster resulting from supercritical bond percolation on a large random recursive tree. Recall that a tree on an ordered set of vertices, say $[n] = \{1, \ldots, n\}$, is called *recursive* if when rooted at 1, the sequence of vertices along any branch from the root to a leaf increases. The terminology stems from the fact that such trees can be constructed recursively, incorporating each vertex one after the other in the natural order to built a growing tree. See Drmota [9] for background and further references.

We denote by T_n a recursive tree picked uniformly at random amongst the $(n-1)!$ recursive trees on $[n]$. Equivalently, T_n can be constructed recursively by creating for $\ell = 1, \ldots, n-1$ an edge between the vertices $\ell + 1$ and u_ℓ, where u_ℓ has the uniform distribution on $[\ell]$ and $u_1, \ldots, u_{n-1}$ are independent. Given T_n, we then perform a Bernoulli bond percolation with parameter

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

where $c > 0$ is some fixed parameter. It is easy to show that this choice of the percolation parameter corresponds precisely to the supercritical regime, in the sense that with high probability for $n \gg 1$, the cluster containing the root is giant with size $G_n \sim \mathrm{e}^{-c} n$. At this point, it may be interesting to briefly sketch the proof of this result, referring to [4] for details.

Pick a vertex u_n uniformly at random in $[n]$, and denote its distance to the root by h_n. Then it is well known that $h_n \sim \ln n$, and since the first moment of $n^{-1}G_n$ coincides with the probability that u_n is connected to the root, one gets

$$\mathbb{E}(n^{-1}G_n) = \mathbb{E}\left(\left(1 - \frac{c}{\ln n}\right)^{h_n}\right) \sim \mathrm{e}^{-c}.$$

Similarly, if v_n denotes a second uniform vertex chosen independently of the first, then the easy fact that the height of the branch point of u_n and v_n remains stochastically bounded yields the second moment estimate $\mathbb{E}((n^{-1}G_n)^2) \sim \mathrm{e}^{-2t}$, from which the law of large numbers for G_n follows.

Since it is also well-known that h_n is asymptotically normal (see Devroye [8]), this might suggest that the same could also hold for G_n. The main result of this work shows that, as a matter of fact, this is not the case. In order to give a precise statement, recall that a real-valued random variable Z has the so-called continuous Luria-Delbrück law when its characteristic

function is given by

$$\mathbb{E}(\mathrm{e}^{i\theta Z}) = \exp\left(-\frac{\pi}{2}|\theta| - i\theta \ln|\theta|\right), \qquad \theta \in \mathbb{R}.$$

This distribution arises in limit theorem for sums of positive i.i.d. variables in the domain of attraction of a completely asymmetric Cauchy process; see e.g. Geluk and de Haan [11], Möhle [16], and further references therein. Its role in the context of large random recursive trees was observed first by Drmota *et al.* [10] and Iksanov and Möhle [12], in relation with a random algorithm for the isolation of the root. See also the comments (a) and (b) in the forthcoming Section 4.

Theorem 1 *There is the weak convergence*

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

where the variable Z has the continuous Luria-Delbrück distribution.

The details of the proof of Theorem 1 combined with earlier results in the literature provide explanations for the anomalous fluctuations of G_n. In the sequel, it will be convenient to agree that the edges are enumerated naturally in the order induced by the construction, i.e. the ℓ-th edge refers to the edge linking the vertex ℓ to its parent $u_{\ell-1}$. Roughly speaking, one can then distinguish three phases.

Because in percolation, each edge is removed with probability $c/\ln n$, the first edges which are removed correspond to an early phase when the growing tree has size of order $\ln n$. During this phase, only a stochastically bounded number of edges are removed, and it has been shown in [5] that the percolation clusters corresponding to those edges will eventually have size of order $n/\ln n$ when the construction is completed. Informally, this is the source of the random fluctuations involving the Luria-Delbrück variable in Theorem 1.

There is then an intermediate phase when the tree grows from a size of order $\ln n$ to the size $\lfloor \ln^4 n \rfloor$, during which about $c\ln^3 n$ edges are removed. Each of the percolation cluster born during this phase has only size $o(n/\ln n)$ at the end of the process. However, the cumulative effect of these clusters is nonetheless visible and yields the deterministic correction involving the iterated logarithm factor in Theorem 1.

In the final phase when the recursive tree grows from size $\lfloor \ln^4 n \rfloor$ to size n, the root cluster grows essentially regularly, i.e. without inducing further fluctuations. We point out that the threshold $\ln^4 n$ appearing in this work is somewhat arbitrary, and $\ln^\alpha n$ with α close to 4 would

work just as well. It is however crucial to choose a threshold which is both sufficiently high so that fluctuations are already visible and spread afterwards quite regularly, and also sufficiently low so that one can estimate the germ of fluctuations with the desired accuracy.

The rest of this paper is mainly devoted to the proof of Theorem 1. The starting point of our analysis is that it is useful to incorporate percolation during the recursive construction of T_n, rather than first constructing completely T_n and then performing percolation. In Section 2, we interrupt the construction of the random recursive tree when it attains the size $k = \lfloor \ln^4 n \rfloor$ and perform a percolation on T_k with parameter p_n. We obtain a precise estimate of the number Δ_k of vertices of T_k which are disconnected from the root; this can be viewed as the germ of the anomalous fluctuations for Δ_n. In Section 3, we resume the construction of the random recursive tree from the size $k = \lfloor \ln^4 n \rfloor$ to the size n. Using the basic connexion between random recursive trees and Yule processes, we show that the germ of the anomalous fluctuations Δ_k spread regularly. Finally, Section 4 contains some miscellaneous comments.

2 The germ of anomalous fluctuations

Imagine that we interrupt the construction of the random recursive tree when it reaches size $k = \lfloor \ln^4 n \rfloor$; plainly this yields a random recursive tree of size k which we denote by T_k. Our purpose in this section is to estimate precisely the number of vertices which are disconnected from the root when one performs a bond percolation on T_k with parameter p_n. It is convenient to work with the parameter k rather than n, that is we write $p_n = q_k$ and note the change in the asymptotic regime of the percolation parameter as a function of the size of the tree :

$$q_k = 1 - ck^{-1/4} + o(k^{-1}). \tag{1}$$

We shall establish the following limit theorem in law.

Proposition 1 *As $k \to \infty$, there is the weak convergence*

$$k^{-3/4}\Delta_k - \frac{3}{4}c \ln k \implies c\,(Z + \ln c)$$

where Z has the continuous Luria-Delbrück distribution

The rest of this section is devoted to the proof of Proposition 1. Our guiding line is similar to that in [3], although the percolation parameter there had a different asymptotic behavior. Namely we shall work with a continuous-time version of percolation in which edges are removed

independently one of the others at a given rate, and consider the process that counts the number of vertices which are disconnected from the root as time passed. It suffices to focus on the cuts made to the root-cluster, and we interpret the latter as a continuous-time version of a random algorithm introduced by Meir and Moon [14, 15] for the isolation of the root. In turn, this enables us to use a coupling due to Iksanov and Möhle [12] and reduces the problem to the analysis of the asymptotic behavior of a remarkable random walk in the domain of attraction of a completely asymmetric Cauchy distribution.

We shall follow the route sketched above, but in the reverse order for an easier articulation of the argument. To start with, we recall briefly an asymptotic result on a random walk which plays an important role in the study of the isolation of the roof for random recursive trees. Let ξ denote an integer-valued random variable with law

$$\mathbb{P}(\xi = j) = \frac{1}{j(j+1)}\,, \qquad j = 1, 2, \ldots \tag{2}$$

We consider the random walk

$$S_\ell = \xi_1 + \cdots + \xi_\ell\,, \qquad \ell \in \mathbb{N}\,,$$

where the ξ_i are independent copies of ξ. According for instance to [11], there is the weak convergence

$$\ell^{-1} S_\ell - \ln \ell \implies Z \tag{3}$$

where Z has the continuous Luria-Delbrück law.

Iksanov and Möhle have pointed at a useful coupling which connects the preceding random walk and an algorithm of Meir and Moon [14, 15] for the isolation of the root. Following Meir and Moon, we define recursively a decreasing sequence of subtrees

$$T_k = T_k(0) \supset T_k(1) \supset \ldots$$

as follows. We pick a first edge uniformly at random amongst the $k-1$ edges of T_k and remove it, which disconnects T_k into two subtrees. We denote by $T_k(1)$ the subtree which contains the root and imagine that the subtree which does not contain the root is set aside. Then we pick second edge uniformly at random in $T_k(1)$, remove it. We write $T_k(2)$ for the subtree which contains the root, set the other subtree aside, and iterate in an obvious way until the root is finally isolated. We write

$$D_k(\ell) = k - |T_k(\ell)|$$

for the number of vertices which have been disconnected from the root after ℓ steps, that is the sum of the sizes of the subtrees which have been set aside.

Iksanov and Möhle [12] have observed that one may couple the random walk S and the algorithm of isolation of the root described above in such a way that there is the identity

$$D_k(\ell) = S_\ell \qquad \text{for all } \ell < N(k)\,, \tag{4}$$

where $N(k) = \min\{\ell : S_\ell \geq k\}$ denotes the first passage time of the random walk above level k. See also Lemma 2 in [3] for a statement tailored for our needs. It follows easily that, provided that the number of removed edges is relatively small, $D_k(\ell)$ grows nearly linearly. Here is a rather crude bound which will be however sufficient for our purpose.

Lemma 1 *Suppose that $\ell = \ell(k)$ fulfills $1 \ll \ell \leq k \ln^{-2} k$. Then*

$$\lim_{k\to\infty} \frac{D_k(\ell)}{\ell \ln^2 \ell} = 0 \qquad \textit{in probability}\,.$$

Proof: Indeed, it follows immediately from (3) that

$$\lim_{\ell\to\infty} \frac{S_\ell}{\ell \ln^2 \ell} = 0 \qquad \text{in probability.} \tag{5}$$

In particular the assumption $\ell \leq k \ln^{-2} k$ ensures that $\ell < N(k)$ with high probability, so we can use the coupling (4) of Iksanov and Möhle. Then (5) is precisely our statement. □

We now turn our attention to a continuous time version of bond percolation on T_k. We equip each of its $k-1$ edges with an independent exponential variable with parameter $k^{-1/4}$, and remove each edge at the time given by this variable. We define

$$t_k = -k^{1/4} \ln q_k\,,$$

so that the probability that a given edge has not yet been removed at time t_k is $\exp(-k^{-1/4} t_k) = q_k$, and the configuration observed at time t_k is thus precisely that resulting from a bond percolation on T_k with parameter q_k. Note also from (10) that

$$t_k = c + O(k^{-1/4})\,. \tag{6}$$

As we are only interested in the number of vertices which have been disconnected from the root at time t_n, we may focus on the evolution of the cluster which contains the root. Plainly,

if we write $\rho_k(\ell)$ for the instant when the ℓ-th edge is removed from the root-cluster in this continuous-time percolation, then the root-cluster at time $\rho_k(\ell)$ can be identified as $T_k(\ell)$, the subtree obtained by the isolation of the root algorithm after ℓ steps. We will need the following bounds.

Lemma 2 *Take any $\alpha \in (1/2, 3/4)$. Then*

$$\lim_{k\to\infty} \mathbb{P}\left(\rho_k\left(ck^{3/4} - k^\alpha\right) \leq t_k \leq \rho_k\left(ck^{3/4} + k^\alpha\right)\right) = 1.$$

Proof: It should be clear from the dynamics of continuous-time percolation and elementary properties of independent exponential variables that $\rho_k(\ell)$ can be expressed in the form

$$\rho_k(\ell) = \sum_{j=0}^{\ell-1} \frac{k^{1/4}}{k - D_k(j) - 1}\varepsilon_j$$

where $\varepsilon_0, \varepsilon_1, \ldots$ is a sequence of i.i.d. standard exponential variables, which is further independent of the algorithm of isolation of the root (the denominator in the fraction above is the number of edges of $T_k(j)$, and for j exceeding the number of steps needed to isolate the root, the general term of the series becomes infinite by convention).

We take first $\ell = ck^{3/4} - k^\alpha$ and use the obvious lower-bound

$$\rho_k(ck^{3/4} - k^\alpha) \geq k^{-3/4} \sum_{j=0}^{ck^{3/4}-k^\alpha-1} \varepsilon_j.$$

Elementary arguments based on the computation of first moment and variance show that the right-hand side can be bounded from below by $c - 2k^{\alpha-3/4}$ with high probability as $k \to \infty$ (note that $\alpha > 3/8$).

Similarly, we then take $\ell = ck^{3/4} + k^\alpha$ and use Lemma 1 to see that with high probability for $k \gg 1$, there is the upper-bound

$$\rho_k(ck^{3/4} + k^\alpha) \leq \frac{k^{1/4}}{k - (ck^{3/4} + k^\alpha)\ln^2 k} \sum_{j=0}^{ck^{3/4}+k^\alpha-1} \varepsilon_j.$$

Again, the sum in the right hand side is easily bounded from above by $ck^{3/4} + 2k^\alpha$ with high probability. On the other hand, the quotient is bounded from above by $k^{-3/4}(1 + 2ck^{-1/4}\ln^2 k)$ whenever k is sufficiently large. Putting the pieces together and recalling that $3/4 - \alpha < 1/4$,

we get that

$$\rho_k(ck^{3/4} + k^\alpha) \leq c + 3k^{\alpha - 3/4}$$

with high probability for $k \gg 1$. We conclude the proof with an appeal to (6). □

We are now able to establish Proposition 1.

Proof: Lemma 2 and an argument of monotonicity show that for any $\alpha \in (1/2, 3/4)$, the bounds

$$D_k(ck^{3/4} - k^\alpha) \leq \Delta_k \leq D_k(ck^{3/4} + k^\alpha)$$

hold with high probability. On the other hand, we see from (5) that

$$\lim_{k\to\infty} k^{-3/4} S_{k^\alpha} = 0 \qquad \text{in probability,}$$

and we also know from (3) that

$$k^{-3/4} S_{ck^{3/4}} - \frac{3}{4} c \ln k \implies c\,(Z + \ln c)\,.$$

It follows that

$$k^{-3/4} S_{ck^{3/4} \pm k^\alpha} - \frac{3}{4} c \ln k \implies c\,(Z + \ln c)\,,$$

and an appeal to the coupling (4) completes the proof. □

3 The spread of anomalous fluctuations

We now recall the well-known connexion between random recursive trees and the Yule process. Imagine that once the tree T_ℓ of size $\ell = 1, \ldots, n-1$ has been constructed, the vertex $\ell + 1$ is incorporated after an exponential time with parameter ℓ, and then connected by a new edge to some vertex $u_\ell \in [\ell]$ which is picked uniformly at random and independently of the exponential waiting time. We further mark that edge with probability $1 - p_n = c/\ln n$, independently of the preceding events. A mark on an edge indicates that this edge will be removed when percolation is performed; equivalently it can also be interpreted as a mutation occurring in the population. The dynamics described above are those of a Yule process with unit rate of birth per individual and with rare neutral mutations which affect each birth event with probability $c/\ln n$, independently of the other birth events.

In this section, we begin our observation of this process with rare neutral mutations once it has reached the size $k = \lfloor \ln^4 n \rfloor$. We thus write $Y = (Y_t)_{t\geq 0}$ for a standard Yule process started

from $Y_0 = k$, and consider the time

$$\tau(n) = \inf\{t \geq 0 : Y_t = n\}$$

at which it hits n. Equivalently, $\tau(n)$ is the time needed to complete the construction of T_n from T_k. We shall first estimate this quantity.

In the sequel, we shall often use the notation

$$A_n = B_n + o(f(n)) \qquad \text{in probability,}$$

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.

Lemma 3 *We have*

$$\mathrm{e}^{\tau(n)} = \frac{n}{\ln^4 n} + o(1/\ln n) \qquad \textit{in probability.}$$

Proof: Elementary properties of Yule processes (see, e.g., Equation (6) in [7]) show that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|n\mathrm{e}^{-\tau(n)} - k\right| > k^\alpha\right) = 0\,.$$

for all $\alpha > 1/2$. In particular, for $\alpha < 3/4$, this yields

$$n\mathrm{e}^{-\tau(n)} = \left(1 + o(\ln^{-1} n)\right) \ln^4 n \qquad \text{in probability.}$$

Our claim follows. □

An individual in the population corresponds to a vertex on the tree, and vice-versa. It is called a mutant if its ancestral lineage (i.e. its branch to the root) contains at least one mark, and a clone otherwise. The population of size k at the time when we start our observation consists in Δ_k mutants and $k - \Delta_k$ clones. We focus on the clone population and write $Y^{(\mathrm{c})} = (Y^{(\mathrm{c})}_t, t \geq 0)$ for the process that counts the number of clones as time passes. Because each clone gives birth to a clone child at rate p_n and independently of the other clones, $Y^{(\mathrm{c})}$ is a Yule process with reproduction rate p_n per individual, and started from $Y^{(\mathrm{c})}_0 = k - \Delta_k$.

In this framework, it should be plain that the size G_n of the root-cluster of T_n after percolation with parameter p_n, coincides with the number of clone individuals at the time when the total population generated by the Yule process reaches n, i.e. there is the identity

$$G_n = Y^{(\mathrm{c})}_{\tau(n)}\,. \tag{7}$$

One readily get the following estimate.

Lemma 4 *We have*

$$G_n = \mathrm{e}^{p_n\tau(n)}\left(\ln^4 n - \Delta_k\right) + o(n/\ln n) \qquad \textit{in probability,}$$

Proof: Again from the basic estimate of Equation (6) in [7] and the fact that $Y_0^{(\mathrm{c})} \le k$, we have for any $\alpha > 1/2$ that

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\mathrm{e}^{-p_n\tau(n)}Y^{(\mathrm{c})}_{\tau(n)} - Y_0^{(\mathrm{c})}\right| > k^{\alpha}\right) = 0\,.$$

We choose $\alpha < 3/4$ and deduce that

$$Y^{(\mathrm{c})}_{\tau(n)} = \mathrm{e}^{p_n\tau(n)}\left(\ln^4 n - \Delta_k\right) + \mathrm{e}^{p_n\tau(n)}o(\ln^3 n)\,, \qquad \text{in probability.}$$

Since $p_n \le 1$, we see from Lemma 3 that

$$\mathrm{e}^{p_n\tau(n)}o(\ln^3 n) = o(n/\ln n)\,,$$

and our claim follows from (7). □

We have now all the ingredients to establish Theorem 1

Proof of Theorem 1: First, it is convenient to apply Skorokhod's representation theorem and assume that the weak convergence in Proposition 1 holds in fact almost surely. This enables us to write

$$\ln^4 n - \Delta_k = \ln^4 n - \ln^3 n\,(3c\ln\ln n + c(Z + \ln c)) + o(\ln^3 n)\,, \qquad \text{almost surely,}$$

and then to re-express Lemma 4 in the form

$$G_n = \mathrm{e}^{p_n\tau(n)}\left(\ln^4 n - \ln^3 n\,(3c\ln\ln n + c(Z + \ln c))\right) + o(n/\ln n)\,, \qquad \text{in probability.}$$

We next note from Lemma 3 that

$$\mathrm{e}^{p_n\tau(n)} = \left(\frac{n}{\ln^4 n} + o(1/\ln n)\right)^{1-c/\ln n}, \qquad \text{in probability,}$$

and it follows from a couple of lines of calculations that

$$\mathrm{e}^{p_n\tau(n)} = \mathrm{e}^{-c}\frac{n}{\ln^4 n} + 4c\mathrm{e}^{-c}n\frac{\ln\ln n}{\ln^5 n} + o(\ln^{-1} n)\,, \qquad \text{in probability.}$$

Another line of calculation yields

$$G_n = \mathrm{e}^{-c}n + c\mathrm{e}^{-c}n\frac{\ln\ln n}{\ln n} - c\mathrm{e}^{-c}\frac{n}{\ln n}\,(Z + \ln c) + o(n/\ln n)\,, \qquad \text{in probability,}$$

which completes the proof. □

It may be interesting to point out that the same technique can be applied to estimate the descent of the initial mutant population. Specifically, the sub-population that stems from the Δ_k mutants at the initial time is described by a Yule process $Y^{(\mathrm{m})}$ with unit birth rate per individual and started from $Y_0^{(\mathrm{m})} = \Delta_k$. It should be plain that $Y^{(\mathrm{m})}_{\tau(n)}$ coincides with $\Delta_{k,n}$, the number of vertices $i \in [n]$ such that, on the branch from i to the root 1, at least one edge with label at most k is removed when performing percolation. From the same argument as in Lemma 4, one can check that for any $1/2 < \alpha < 1$,

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\Delta_{k,n} - \mathrm{e}^{\tau(n)}\Delta_k\right| > \mathrm{e}^{\tau(n)}k^\alpha\right) = \lim_{n\to\infty} \mathbb{P}\left(\left|\mathrm{e}^{-\tau(n)}Y^{(\mathrm{m})}_{\tau(n)} - Y_0^{(\mathrm{m})}\right| > k^\alpha\right) = 0\,.$$

It follows that

$$\Delta_{k,n} = \mathrm{e}^{\tau(n)}\Delta_k + o(n/\ln n) \qquad \text{in probability.} \tag{8}$$

On the one hand, recall from Lemma 3 that $\mathrm{e}^{\tau(n)}k^\alpha = o(n/\ln n)$ in probability. On the other hand, combining Proposition 1 and Lemma 3, we get

$$\frac{\ln n}{n}\mathrm{e}^{\tau(n)}\Delta_k - 3c\ln\ln n \Longrightarrow c\,(Z + \ln c)\,,$$

and we conclude that

$$\frac{\ln n}{n}\Delta_{k,n} - 3c\ln\ln n \Longrightarrow c\,(Z + \ln c)\,. \tag{9}$$

More precisely, this weak convergence holds jointly with that in Theorem 1, as we can see from Lemma 4 and (8)

4 Miscellaneous comments

For the purpose of this section, it is convenient to rewrite Theorem 1 in terms of $\Delta_n = n - G_n$, the number of vertices in T_n which are disconnected from the root after performing a bond

percolation with parameter p_n. We have then

$$\left(n^{-1}\Delta_n - (1-\mathrm{e}^{-c})\right)\ln n + c\mathrm{e}^{-c}\,\ln\ln n \implies c\mathrm{e}^{-c}\,(Z+\ln c)\;. \tag{10}$$

We also introduce a standard Luria-Delbrück variable Z_m with parameter $m>0$, which has generating function

$$\mathbb{E}\left(s^{Z_m}\right) = (1-s)^{m(1-s)/s}\,, \qquad 0\le s\le 1\,.$$

Recall that as $m\to\infty$, there is the weak convergence

$$\frac{Z_m}{m} - \ln m \implies Z \tag{11}$$

where Z has the continuous Luria-Delbrück distribution. See Pakes [17] or Theorem 4.1 in Möhle [16].

(a) It has been argued that for certain populations models with a small rate of neutral mutation, the number of mutants has a Luria-Delbrück law ; see Section 2 in Kemp [13] and references therein. In this setting, the parameter is given by $m=gN(a+g)$ where N is the total population size, a the rate of birth of clones, and g the rate of birth of new mutants. We stress however that, as pointed out by Kemp, the models leading to these Luria-Delbrück laws '*involve simplifying assumptions that leave realism somewhat in doubt*'.

In our framework, interpreting the recursive construction of T_n as a Yule process and percolation as rare neutral mutations, this suggests that the number Δ_n of vertices disconnected from the root might have a distribution close to the Luria-Delbrück law with parameter $m=cn/\ln n$. If we write $\Delta'_n = Z_m$ for the latter, then (11) yields the weak convergence

$$\left(n^{-1}\Delta'_n - c\right)\ln n + c\ln\ln n \implies c\,(Z+\ln c)\;.$$

This resembles (10), but with fundamental discrepancies. Note in particular that for $c>1$, the estimation above would imply that for $n\gg 1$, the mutant population is close to cn, a quantity strictly larger than the total population! It is therefore unlikely that Theorem 1 could established rigorously from such arguments.

(b) If we write $C_{1,n}\ge C_{2,n}\ge\ldots$ for the sequence of the sizes of the percolation clusters disconnected from the root and ranked in the decreasing order, then there is clearly the identity

$$\Delta_n = \sum_i C_{i,n}\,. \tag{12}$$

Theorem 1 in [3] states that for every fixed integer j,

$$\left(\frac{\ln n}{n} C_{1,n}, \ldots, \frac{\ln n}{n} C_{j,n}\right) \Longrightarrow (\mathtt{x}_1, \ldots, \mathtt{x}_j) \tag{13}$$

where $\mathtt{x}_1 > \mathtt{x}_2 > \ldots$ denotes the sequence of the atoms of a Poisson random measure on $(0, \infty)$ with intensity $c\mathrm{e}^{-c}x^{-2}\mathrm{d}x$. It is certainly tempting to expect that the finite dimensional convergence (13) might be reinforced and then yield (10) via (12).

An obvious obstacle is that the series $\sum \mathtt{x}_i$ diverges a.s.; however this can be circumvented by considering

$$X_n := \sum_i \left\lfloor \frac{n}{\ln n} \mathtt{x}_i \right\rfloor .$$

The reason for taking integer parts above is of course because cluster sizes are integers. Note that this limits *de facto* the sum to atoms such that $\mathtt{x}_i \geq n^{-1} \ln n$, and then $X_n < \infty$ a.s. More precisely, by the elementary mapping theorem for Poisson random measures, $\lfloor \mathtt{x}_1 n \ln^{-1} n \rfloor, \ldots$ can be viewed as the sequence of atoms of a Poisson random measure on $\mathbb{N}$ with intensity $m\mu$ where $m = c\mathrm{e}^{-c} n \ln^{-1} n$ and μ is the probability measure given by $\mu(k) = k^{-1} - (k+1)^{-1}$. Thus $X_n = Z_m$ has the Luria-Delbrück law with parameter m, see Section 3 in Möhle [16].

As a consequence of (11), there is the weak convergence

$$\left(n^{-1} X_n - c\mathrm{e}^{-c}\right) \ln n + c\mathrm{e}^{-c} \ln \ln n \Longrightarrow c\mathrm{e}^{-c} \left(Z + \ln c - c\right) .$$

This again resembles (10), in particular one captures the deterministic correction involving the iterated logarithm, and the random fluctuations are the same up-to a constant. This corroborates the fact that the fluctuations for the size of the giant component are chiefly due to the largest percolation clusters. However, this fails to give the correct first order ($c\mathrm{e}^{-c}$ instead of $1 - \mathrm{e}^{-c}$), showing that Theorem 1 cannot be derived from weak limits theorems as (13).

(c) We now conclude this work by pointing out that the problem considered in this work could also have been formulated in terms of an urn model *à la* Polya-Hoppe. Indeed, the recursive construction of T_n together with marks on edges corresponding to percolation can also be described as follows. We start with an urn containing a single red ball (the root), and at each step, we add either a red ball or a black ball according to the following random algorithm. With probability $c/\ln n$, we add a black ball to the urn, and with probability $p_n = 1 - c/\ln n$, we pick a ball uniformly at random in the current contain of the urn, and then replace it into the urn together with a new ball of the same color. Then Δ_n is the number of black balls when the urn contains exactly n balls, and (13) thus gives then a precise limit

theorem for the proportion of black balls.

References

[1] Barraez, D., Boucheron, S. and Fernandez de la Vega, W. On the fluctuations of the giant component. *Combin. Probab. Comput.* **9** (2000), 287-304.

[2] Behrisch, M., Coja-Oghlan, A. and Kang, M. The order of the giant component of random hypergraphs. *Random Structures Algorithms* **36** (2010), 149-184.

[3] Bertoin, J. Sizes of the largest clusters for supercritical percolation on random recursive trees. To appear in *Random Structures Algorithms*.

[4] Bertoin, J. Almost giant clusters for percolation on large trees with logarithmic heights. To appear in *J. Appl. Probab.*

[5] Bertoin, J. and Uribe Bravo, G. Supercritical percolation on large scale-free random trees. Submitted.

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

[7] de La Fortelle, A. Yule process sample path asymptotics. *Electron. Comm. Probab.* **11** (2006), 193-199.

[8] Devroye, L. Applications of the theory of records in the study of random trees. *Acta Inform.* **26** (1988), 123-130.

[9] Drmota, M. *Random trees*. Springer. NewYork, Vienna, 2009

[10] Drmota, M., Iksanov, A., Möhle, M. and Rösler, U. A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures Algorithms* **34-3** (2009), 319-336.

[11] Geluk, J. L. and de Haan, L. Stable probability distributions and their domains of attraction: a direct approach. *Probab. Math. Statist.* **20** (2000), 169-188.

[12] 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** (2007), 28-35.

[13] Kemp, A. W. Comments on the Luria-Delbrück distribution. *J. Appl. Probab.* **31** (1994), 822-828.

[14] Meir, A. and Moon, J. W. Cutting down random trees. *J. Austral. Math. Soc.* **11** (1970), 313-324.

[15] Meir, A. and Moon, J. W. Cutting down recursive trees. *Mathematical Biosciences* **21** (1974), 173-181.

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

[17] Pakes, A. G. Remarks on the Luria-Delbrück distribution. *J. Appl. Probab.* **30** (1993), 991-994.

[18] Pittel, B. On tree census and the giant component in sparse random graphs. *Random Structures Algorithms* **1** (1990), 311-342.

[19] Seierstad, T.G. On the normality of giant components. To appear in *Random Structures Algorithms*.

[20] Stepanov, V.E. The probability of the connectedness of a random graph $\mathcal{G}_m(t)$. *Teor. Verojatnost. i Primenen* **15** (1970), 58-68.