Adv. Appl. Prob. **49,** 877–902 (2017)
*doi:*10.1017/apr.2017.25
© *Applied Probability Trust* 2017

WEAK LIMITS FOR THE LARGEST SUBPOPULATIONS IN YULE PROCESSES WITH HIGH MUTATION PROBABILITIES

ERICH BAUR,* *ENS de Lyon*
JEAN BERTOIN,** *Universität Zürich*

Abstract

We consider a Yule process until the total population reaches size $n \gg 1$, and assume that neutral mutations occur with high probability $1 - p$ (in the sense that each child is a new mutant with probability $1 - p$, independently of the other children), where $p = p_n \ll 1$. We establish a general strategy for obtaining Poisson limit laws and a weak law of large numbers for the number of subpopulations exceeding a given size and apply this to some mutation regimes of particular interest. Finally, we give an application to subcritical Bernoulli bond percolation on random recursive trees with percolation parameter p_n tending to 0.

Keywords: Branching process; mutation; percolation; random increasing tree

2010 Mathematics Subject Classification: Primary 60J27
Secondary 60J80; 60K35

1. Introduction

We consider a system of branching processes with mutations specified as follows. The underlying total population process is modeled by a standard Yule process Z, that is a continuous-time birth process started from one individual with unit birth rate per unit population size. We superpose independent mutations by declaring that a newborn child is a clone of its parent with probability $p \in (0, 1)$, and a mutant otherwise. Being a mutant means that the individual obtains a new genetic type which was not present before. We observe the process Z at the instant when the nth individual is born and group individuals of the same genetic type into subpopulations.

In this paper we are interested in questions concerning the (asymptotic) sizes of these subpopulations under strong mutations, in the sense that $p = p_n \to 0$. By approximating the population system from below and above by two different processes, where subpopulations are independent and have an explicit distribution, we develop a general strategy to obtain nontrivial (Poisson) weak limits and a law of large numbers for the number of subpopulations exceeding a given size x_n (which might grow with n as well).

We then discuss our strategy in the context of three qualitatively different mutation regimes. For a given $\ell \in \mathbb{N}$, we identify first $p_n \sim an^{-1/\ell}$, $a > 0$ fixed, as the regime in which, in the limit $n \to \infty$, the largest subpopulations have size $\ell + 1$. For their number, we obtain a

Received 22 March 2016; revision received 20 February 2017.
* Current address: Bern University Of Applied Sciences, Quellgasse 21, 2501 Biel, Switzerland.
Email address: erich.baur@bfh.ch
** Postal address: Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland.
Email address: jean.bertoin@math.uzh.ch

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Poisson limit law, while for the number of subpopulations of size j when $j \in \{1, \dots, \ell\}$, we prove a weak law of large numbers (Theorem 4.1).

Further, we discuss the regime $p_n \sim a/\ln n$. Since the size of the subpopulation containing the ancestor is of order n^{p_n}, see Proposition 2.1, this is the borderline case between a bounded and an unbounded size for the ancestral subpopulation. We show that the sizes of the largest subpopulations are concentrated around $c_1 \ln n + c_2 \ln\ln n$, where c_1 and c_2 are positive constants depending on a (Theorem 4.3). For the exact choice $p_n = a/\ln n$ and $\lambda > 0$, we find a correction $c_3 = c_3(a, \lambda) \in \mathbb{R}$ such that, with $y_n = c_1 \ln n + c_2 \ln\ln n + c_3$, the number of subpopulations greater than y_n converges along subsequences $(y_{n(m)}) \subset (y_n)$ with converging fractional part to a Poisson(Λ)-distributed random variable, where Λ is expressed in terms of a, λ, and $(y_{n(m)})$ (Theorem 4.2).

Finally, we study the $1/\ln n \ll p_n \ll 1$ case. Here, it turns out that the sizes of the largest subpopulations are, to the first order, given by $\mathrm{e}^{-1} p_n^{-1} n^{p_n}$. For $p_n \geq \ln\ln n/\ln n$ and given $\lambda > 0$, we compute a precise barrier such that the number of subpopulations exceeding this barrier follows in the limit the Poisson law with parameter λ (Theorem 4.4).

This work originates from questions about Bernoulli bond percolation on so-called random recursive trees, when their size n tends to ∞ and the percolation parameter $p = p_n$ satisfies $p_n \to 0$. The connection to branching processes stems from the fact that the genealogical tree built from the first n individuals in a standard Yule process can be interpreted as a random recursive tree T_n on $\{1, \dots, n\}$. Mutations in the Yule process are naturally modeled on its genealogical tree by cutting the edges that connect mutants to its parent. Then the connected subsets of vertices form the subpopulations of the same genetic type. To put it differently, the connected components (clusters) on a random recursive tree T_n that arise from a Bernoulli bond percolation, where each edge is erased with probability $1-p$ independently of each other, can be viewed as the subpopulations in a Yule process with mutation rate $1-p$, observed at the instant when there are n individuals in total in the system. The strategy we develop here in terms of Yule processes allows a concise analysis of cluster sizes, for any choice of p_n tending to 0. For sequences of p_n such that $p_n \to 1$ or $p_n = p \in (0, 1)$ remains constant, similar connections between systems of branching processes and percolation on increasing tree families have been utilized before in, e.g. [4], [5], [8], and [9]. The precise definition of a random recursive tree, its connection to Yule processes, and more references to existing results on percolation will be discussed in Section 5.

The rest of this paper is organized as follows. In Section 2 we properly define the population system and provide some heuristics as to why and in which regimes Poisson limits should appear. Then, in Section 3, we explain our strategy for obtaining Poisson limit laws for the number of subpopulations (or clusters) greater than a given size. Section 4 contains our main results; we exemplify our strategy by proving limit results for certain mutation rates of particular interest. Finally, in Section 5 we establish the link to percolation on random recursive trees. Appendix A contains some (standard) estimates on Yule processes, which we use in our analysis.

Notation. We let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$. If $(a_n : n \in \mathbb{N})$ and $(b_n : n \in \mathbb{N})$ are two sequences of real numbers, we write $a_n \sim b_n$ if $a_n/b_n \to 1$ as $n \to \infty$, and we write $a_n \ll b_n$ or $b_n \gg a_n$ if and only if $a_n/b_n \to 0$ as $n \to \infty$. Moreover, if $f(n)$ and $g(n)$ are two positive functions, we say $g(n) = O(f(n))$ if there exists $M > 0$ such that $g(n) \leq M f(n)$ for all $n \in \mathbb{N}$, and we write $f(n) = o(g(n))$ if $g(n)/f(n) \to 0$ as $n \to \infty$. We will use the letters c or C for small or large generic constants that do not depend on n. Their values may change from line to line.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

2. Yule processes with mutations

In this section we introduce the population system of interest, and present some basic results and heuristics. For background on Yule processes, we refer the reader to Appendix A, and for more information to [3, Chapter 3].

2.1. A population system with infinitely many types

Let $Z = (Z(t)\colon t \geq 0)$ be a standard Yule process with unit birth rate per individual and started from one single particle, that is, $Z(0) = 1$. Let $p \in [0, 1]$. We superpose mutations on Z as follows. When a new child is born, we declare it to be a clone of its parent with probability p, and a mutant with a new genetic type different from all the previous types with probability $1-p$. We let $b_1^{(p)} = 0$ and write $0 < b_2^{(p)} < b_3^{(p)} < \cdots$ for the sequence of consecutive birth times of individuals which are mutants. More specifically, the ancestor present at time $b_1^{(p)} = 0$ has genetic type 1, as well as its clone children. The first mutant is born at time $b_2^{(p)}$ and has genetic type 2. The next mutant appears in the system at time $b_3^{(p)}$ and receives genetic type 3; it is a mutant of an individual of either type 1 or of type 2, and so on. We group individuals of the same genetic type into subpopulations, so that at time t, the population system consists of at most $Z(t)$ many subpopulations. In contrast to [8], here we are not interested in the genealogical structure, but only in the sizes of the subpopulations.

For $i \in \mathbb{N}$, we let $(Y_i^{(p)}(t)\colon t \geq 0)$ denote the subpopulation process counting the individuals of type i, with $Y_i^{(p)}(t) = 0$ if $t < b_i^{(p)}$. It should be clear that the processes $(Y_i^{(p)}(b_i^{(p)} + t)\colon t \geq 0)$, $i \in \mathbb{N}$, form independent Yule processes with birth rate p per individual and started from one single individual each. The number of subpopulations of different genetic types present at time $t \geq 0$ is denoted by

$$T^{(p)}(t) = \max\{i \in \mathbb{N}_0\colon b_i^{(p)} \leq t\}.$$

Viewed as a process in t, $T^{(p)} = (T^{(p)}(t),\ t \geq 0)$ is a counting process started from 1 which grows at rate $(1-p)Z(t)$. Its predictable compensator is absolutely continuous with density $(1-p)Z(t)$, that is,

$$T^{(p)}(t) - \int_0^t (1-p)Z(s)\,\mathrm{d}s, \qquad t \geq 0,$$

is a martingale; see, e.g. [16, Theorem 9.15].

We next build a process $\boldsymbol{Y}^{(p)} = (\boldsymbol{Y}^{(p)}(t)\colon t \geq 0)$ by setting

$$\boldsymbol{Y}^{(p)}(t) = (Y_1^{(p)}(t), Y_2^{(p)}(t), \dots), \qquad t \geq 0.$$

Note that $\boldsymbol{Y}^{(p)}(0) = (1, 0, \dots)$. Clearly, we can retrieve the total population size $Z(t)$ at time t from $\boldsymbol{Y}^{(p)}$,

$$Z(t) = \sum_{i=1}^{\infty} Y_i^{(p)}(t).$$

It follows from its construction that the process $\boldsymbol{Y}^{(p)}$ is Markovian with transition rates at time $t \geq 0$ for $y_1, \dots, y_k \in \mathbb{N}$ given by

$$\begin{aligned}(y_1, \dots, y_k, 0, \dots) &\mapsto (y_1, \dots, y_k, 1, 0, \dots) \quad \text{at rate } (1-p)(y_1 + \cdots + y_k),\\ (y_1, \dots, y_k, 0, \dots) &\mapsto (y_1, \dots, y_{i-1}, y_i + 1, y_{i+1}, \dots, y_k, 0, \dots) \quad \text{at rate } py_i.\end{aligned}$$

Mostly, we will stop $\boldsymbol{Y}^{(p)}$ at the instant when the nth particle is born,

$$\tau_n = \inf\{t \geq 0\colon Z(t) = n\}.$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Obviously, we have $\tau_i \le b_i^{(p)}$ for all $i \in \mathbb{N}$ and all $p \in [0,1]$. Moreover, since $\mathrm{e}^{-t}Z(t)$ is a martingale converging almost surely to a standard exponential random variable $\mathcal{E}$ as t tends to ∞, see Appendix A,

$$\lim_{n\to\infty} (\tau_n - \ln n) = -\ln \mathcal{E} \quad \text{almost surely.} \tag{2.1}$$

2.2. The size of the ancestral subpopulation

In this section we point at a simple limit theorem in distribution for the size $Y_1^{(p)}(\tau_n)$ of the subpopulation at time τ_n having the same genetic type as the ancestor individual. We obtain the following characterization when $p = p_n \ll 1$.

Proposition 2.1. *For $\lambda > 0$, denote by* geom(λ) *a geometrically distributed random variable of parameter λ, and by* Exp(1) *a standard exponential random variable. Then the following hold as $n \to \infty$:*

(i) *if $0 \le p_n \ll 1/\ln n$ then $Y_1^{(p_n)}(\tau_n) \xrightarrow{\mathrm{D}} 1$;*

(ii) *if $p_n \sim a/\ln n$ for some $a > 0$ then $Y_1^{(p_n)}(\tau_n) \xrightarrow{\mathrm{D}}$ geom(e^{-a});*

(iii) *if $1/\ln n \ll p_n \ll 1$ then $n^{-p_n} Y_1^{(p_n)}(\tau_n) \xrightarrow{\mathrm{D}}$ Exp(1),*

where '$\xrightarrow{\mathrm{D}}$' denotes convergence in distribution.

Proof. Assume that $p_n \to 0$, and let $G_n = \{|\tau_n - \ln n| \le p_n^{-1/2}\}$. From (2.1), we know that $\mathbb{P}(G_n) \to 1$ as $n \to \infty$. For (i) and (ii), it thus suffices to show that, for $x \in \mathbb{N}$,

$$\mathbb{P}(Y_1^{(p_n)}(\tau_n) > x;\ G_n) \to 1 - F(x) \quad \text{as } n \to \infty,$$

where F is the distribution function of the stated limit in (i) or (ii), respectively. Since $Y_1^{(p_n)}(t)$ is monotone increasing in t, we have

$$\begin{aligned} \mathbb{P}(Y_1^{(p_n)}(\ln n - p_n^{-1/2}) > x;\ G_n) &\le \mathbb{P}(Y_1^{(p_n)}(\tau_n) > x;\ G_n) \\ &\le \mathbb{P}(Y_1^{(p_n)}(\ln n + p_n^{-1/2}) > x). \end{aligned} \tag{2.2}$$

Since $Y^{(p_n)}(t)$ has a geometric law with success probability $\exp(-p_n t)$, we obtain

$$\mathbb{P}(Y_1^{(p_n)}(\ln n + p_n^{-1/2}) > x) = (1 - \exp(-(p_n \ln n + p_n^{1/2})))^x.$$

It is readily checked that when $x \in \mathbb{N}$ and $n \to \infty$, the right-hand side converges for $p_n \ll 1/\ln n$ to 0, whereas for $p_n \sim a/\ln n$, it converges to $(1 - \mathrm{e}^{-a})^x$. Writing

$$\mathbb{P}(Y_1^{(p_n)}(\ln n - p_n^{-1/2}) > x;\ G_n) = \mathbb{P}(Y_1^{(p_n)}(\ln n - p_n^{-1/2}) > x) + \mathbb{P}(G_n^c),$$

we see that the same convergence holds for the probability on the left-hand side in (2.2). This shows (i) and (ii). The proof of (iii) is entirely similar and left to the reader. □

Remark 2.1. (i) The above statements remain true if $Y_1^{(p)}(\tau_n)$ is replaced by any $Y_i^{(p)}(\tau_n)$ for $i \in \mathbb{N}$ fixed.

(ii) For $p \in (0,1)$ fixed, it has been shown independently in [5] and [20] that $n^{-p} Y_1^{(p)}(\tau_n)$ converges to a Mittag–Leffler-distributed random variable with parameter p. In turn, the latter converges to a standard exponential random variable when $p \to 0$. If $p_n \to 1$ as $n \to \infty$ then $n^{-p_n} Y_1^{(p_n)}(\tau_n) \to 1$ almost surely; see [4].

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

(iii) We note that for each $p \in [0, 1]$ the mean of $Y_1^{(p)}(\tau_n)$ can be computed exactly. Indeed, we have

$$\begin{aligned}\mathbb{E}[Y_1^{(p)}(\tau_n)] &= \sum_{i=1}^{n} \mathbb{P}(\text{the } i\text{th individual belongs to the ancestral subpopulation}) \\ &= \sum_{i=1}^{n} \mathbb{E}[p^{D(i)}],\end{aligned}$$

where $D(i)$ is the insertion depth of i, that is, $D(i)$ is the height of i in the genealogical tree of the Yule process stopped at time τ_n (with $D(1) = 0$). From [12] we know that $D(i) \overset{\mathrm{D}}{=} \sum_{j=1}^{i-1} \xi_j$, where ξ_j are independent Bernoulli random variables of parameter $1/j$. Consequently,

$$\begin{aligned}\sum_{i=1}^{n} \mathbb{E}[p^{D(i)}] &= 1 + p \sum_{i=2}^{n} \prod_{j=2}^{i-1} \mathbb{E}[p^{\xi(j)}] \\ &= 1 + p \sum_{i=2}^{n} \prod_{j=2}^{i-1} \left(\frac{p}{j} + 1 - \frac{1}{j} \right) \\ &= 1 + p \sum_{i=1}^{n-1} \frac{1}{i!} \prod_{j=2}^{i} (j - 1 + p).\end{aligned}$$

2.3. Poisson heuristic for the number of large subpopulations

Clearly, when $p_n \ll 1/n$, the first n individuals are all mutants and all subpopulations at time τ_n are singletons with high probability. When $p_n \sim a/n$, standard Poisson approximation to the binomial law (and the fact that the genealogical tree of $Z(\tau_n)$ is not star-like) shows that the number of subpopulations of size 2 is asymptotically Poisson(a)-distributed. For general $\ell > 1$, it is, however, *a priori* not obvious in which regime subpopulations of size $\ell + 1$ will emerge, and, if so, how many of them. Following the tradition of Aldous [1], we will now present a heuristic argument to determine this regime, and will check later that our informal approach can actually be made rigorous.

We are interested in the number of subpopulations of size $> \ell$ at time τ_n,

$$N^{n,p}(\ell) = \sum_{i=1}^{\infty} \mathbf{1}_{\{Y_i^{(p)}(\tau_n) > \ell\}}.$$

For $i \gg 1$ and $p \ll 1$, the birth time $b_i^{(p)}$ of the ith mutant is close to the birth time τ_i of the ith individual, and, by (2.1), we have $b_i^{(p)} - \ln i \sim -\ln \mathcal{E}$ almost surely. Thus, the distribution of $Y_i^{(p)}(\tau_n)$ should be close to that of a Yule process with birth rate p evaluated at time $\tau_n - b_i^{(p)} \sim \ln(n/i)$, that is, to a geometric distribution with parameter $\exp(-p \ln(n/i)) = (i/n)^p$. Therefore, the Bernoulli variable $\mathbf{1}_{\{Y_i^{(p)}(\tau_n) > \ell\}}$ has parameter

$$\mathbb{P}(Y_i^{(p)}(\tau_n) > \ell) \sim \left(1 - \left(\frac{i}{n}\right)^p\right)^{\ell} \sim \left(p \ln \frac{n}{i}\right)^{\ell}.$$

Let us now recall Le Cam's inequality for the reader's convenience.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

The inequality of Le Cam [18]. Let $(a_n)_n \subset \mathbb{N}$ be a sequence of integers. For each $n \in \mathbb{N}$, let $\xi_{n,i}$, $1 \le i \le a_n$, be independent Bernoulli random variables with parameters $q_{n,i}$. Set $S_n = \xi_{n,1} + \cdots + \xi_{n,a_n}$, and let $\lambda_n = \mathbb{E}[S_n] = \sum_{i=1}^{a_n} q_{n,i}$. Then

$$\sum_{j=1}^{\infty} \left| \mathbb{P}(S_n = j) - \frac{\lambda_n^j \mathrm{e}^{-\lambda_n}}{j!} \right| \le 2 \sum_{i=1}^{a_n} q_{n,i}^2.$$

In particular, if $\lim_{n\to\infty} \lambda_n = \lambda \in (0, \infty)$ and $\lim_{n\to\infty} \sum_{i=1}^{a_n} q_{n,i}^2 = 0$ then we have $S_n \overset{\mathrm{D}}{\to} \mathrm{Poi}(\lambda)$, where $\mathrm{Poi}(\lambda)$ is a Poisson(λ)-distributed random variable.

Even though the variables $Y_i^{(p)}(\tau_n)$, $i \in \mathbb{N}$, which describe the sizes of the subpopulations, are clearly not independent (obviously, they add up to n), let us pretend for the purpose of this section that $\mathbf{1}_{\{Y_i^{(p)}(\tau_n) > \ell\}}$ for $i = 1, \ldots, n$, form a sequence of independent Bernoulli variables. Since

$$\sum_{i=1}^{n} \left(p \ln \frac{n}{i} \right)^{\ell} \sim p^{\ell} n \int_0^1 \left(\ln \frac{1}{t} \right)^{\ell} \mathrm{d}t = p^{\ell} n \ell!,$$

we infer that, for $a > 0$,

$$p = p_n \sim \frac{a}{n^{1/\ell}}$$

should be the regime in which the largest subpopulations have size $\ell + 1$, and, more precisely, then

$$N^{n,p_n}(\ell) \overset{\mathrm{D}}{\to} \mathrm{Poi}(\ell! a^{\ell}).$$

These informal calculations are, of course, far from a rigorous proof; nonetheless we will show in the forthcoming Theorem 4.1 that the above weak convergence actually holds. For this, we will first develop a general strategy to analyze the asymptotic behavior of the number of large subpopulations in the next section.

3. The number of subpopulations exceeding a given size

We will establish limit laws for the number of subpopulations of size $> x \in \mathbb{N}$ at time τ_n,

$$N^{n,p}(x) = \sum_{i=1}^{\infty} \mathbf{1}_{\{Y_i^{(p)}(\tau_n) > x\}}, \tag{3.1}$$

when $p = p_n \to 0$. The threshold x may be fixed or $x = x_n \to \infty$ (depending on the choice of $(p_n)_n$). Roughly speaking, the key step consists in approximating the population system $\boldsymbol{Y}^{(p)}$ at time τ_n by systems $\underline{\boldsymbol{Y}}^{(p)}$ and $\bar{\boldsymbol{Y}}^{(p)}$ in which mutants are born at deterministic times, and such that $\boldsymbol{Y}^{(p)}$ is bounded by $\underline{\boldsymbol{Y}}^{(p)}$ from below and by $\bar{\boldsymbol{Y}}^{(p)}$ from above. The advantage of working with deterministic times comes from the fact that we can in this way decouple populations sizes from birth times and then deal with independent Bernoulli variables. In the new systems, we can compute moments of the quantities corresponding to $N^{n,p}(x)$ more easily. Provided we construct $\underline{\boldsymbol{Y}}^{(p)}$ and $\bar{\boldsymbol{Y}}^{(p)}$ such that the expected numbers of subpopulations of a size larger than x match asymptotically in the two systems, this will ultimately lead to limit statements for $N^{n,p}(x)$ in the original system.

In fact, it will be sufficient to control $\boldsymbol{Y}^{(p)}$ on a set of large probability. In this regard, recall that convergence in distribution of a sequence of random variables $(U_n : n \in \mathbb{N})$ to some random variable U is equivalent to convergence in distribution of $U_n \mathbf{1}_{E_n}$ to U, provided $(E_n : n \in \mathbb{N})$ is a sequence of events with $\mathbb{P}(E_n) \to 1$. Concerning limit results for $N^{n,p}$, we can therefore restrict ourselves to certain 'good' events E_n of large probability, which we specify next.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Lemma 3.1. *Assume that $p_n \to 0$, and let $(k_n)_n \subset \mathbb{N}$ be a sequence of integers such that $k_n \to \infty$ and $k_n p_n \to 0$ as $n \to \infty$. There exists an event E_n (depending on n, p_n, and k_n) of probability $\mathbb{P}(E_n) \to 1$, on which the following hold:*

(i) $\tau_{k_n} = b_{k_n}^{(p)} \le (\ln k_n)^2$;

(ii) $(1 - k_n^{-1/3})k_n e^t \le Z(\tau_{k_n} + t) \le (1 + k_n^{-1/3})k_n e^t$ *for all* $t \ge 0$;

(iii) $(1 - 3k_n^{-1/3})k_n e^t \le T^{(p_n)}(b_{k_n}^{(p)} + t) \le (1 + 3k_n^{-1/3})k_n e^t$ *for all* $t \ge 0$.

We call E_n the *good event*, and for establishing our limit results for $N^{n,p_n}(x_n)$, we will work on E_n. The proof of the lemma uses standard estimates on Yule processes and is given in Appendix A.

Note that the event $\{\tau_{k_n} = b_{k_n}^{(p)}\}$ under (i) is precisely the event that the first $k_n - 1$ born individuals (discounting the ancestor individual) are all mutants. The choice of k_n will depend on our applications and will be specified later. Roughly speaking, we choose k_n in such a way that with high probability, the first $2k_n$ born individuals do not contribute to $N^{n,p_n}(x_n)$.

In the following, we will always implicitly assume that n is large enough such that $k_n > 27$. We define, for $i \in \mathbb{N}$,

$$\bar{\beta}(i) = \ln i - \ln(1 - 3k_n^{-1/3}) - \ln k_n, \qquad \underline{\beta}(i) = \ln i - \ln(1 + 3k_n^{-1/3}) - \ln k_n,$$

so that

$$(1 - 3k_n^{-1/3})k_n \exp(\bar{\beta}(i)) = (1 + 3k_n^{-1/3})k_n \exp(\underline{\beta}(i)) = i. \tag{3.2}$$

Lemma 3.1 immediately implies the following control over birth times on the event E_n.

Corollary 3.1. *On the good event E_n, we have, for $i \ge 2k_n$, the bounds*

$$\tau_{k_n} + \underline{\beta}(i) \le \tau_i \le \tau_{k_n} + \bar{\beta}(i) \quad \text{and} \quad b_{k_n}^{(p)} + \underline{\beta}(i) \le b_i^{(p)} \le b_{k_n}^{(p)} + \bar{\beta}(i).$$

In the next two sections, we show how to upper and lower bound $N^{n,p_n}(x_n)$ on the good event E_n for $(x_n)_n$ a sequence of positive integers (possibly constant). For ease of notation, we write p instead of p_n, and x instead of x_n. We will always work on the event E_n, with the properties stated in Lemma 3.1.

3.1. An upper bound

Recall the notation (3.1). We treat the summands i with $i \le 2k_n - 1$ and $i \ge 2k_n$ separately and first consider the $i \ge 2k_n$ case. From Corollary 3.1, we deduce that, for $i \ge 2k_n$,

$$\tau_n \le b_{k_n}^{(p)} + \bar{\beta}(n) \le b_i^{(p)} + \bar{\beta}(n) - \underline{\beta}(i). \tag{3.3}$$

In the sum (3.1), only the terms with $b_i^{(p)} \le \tau_n$ contribute. Thus, we can restrict ourselves to summands with $2k_n \le i \le n^*$, where

$$n^* = \max\{i \in \mathbb{N} \colon \underline{\beta}(i) \le \bar{\beta}(n)\}. \tag{3.4}$$

Setting $U_i^{n,p} = Y_i^{(p)}(b_i^{(p)} + \bar{\beta}(n) - \underline{\beta}(i))$, we obtain, from (3.3) and the fact that $Y_i^{(p)}(t)$ is monotone increasing in t, the almost sure upper bound

$$\sum_{i=2k_n}^{n^*} \mathbf{1}_{\{Y_i^{(p)}(\tau_n) > x\}} \mathbf{1}_{E_n} \le \sum_{i=2k_n}^{n^*} \mathbf{1}_{\{U_i^{n,p} > x\}}.$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Note that the variables $U_i^{n,p}$, $2k_n \le i \le n^*$, are independent of each other, and $U_i^{n,p}$ has the law of the population size of a Yule process with birth rate p at time $\bar\beta(n) - \underline\beta(i)$ when started from a single individual.

The summands with $1 \le i \le 2k_n - 1$ are all bounded in the same (rough) way by disregarding the individual birth times of the corresponding subpopulations. More specifically, we remark that on E_n,

$$\tau_n \le \tau_{k_n} + \bar\beta(n) + \le \bar\beta(n) + \ln^2(k_n).$$

so that, defining $U_i^{n,p} = Y_i^{(p)}(b_i^{(p)} + \bar\beta(n) + \ln^2(k_n))$ and using again the monotonicity of $Y_i^{(p)}(t)$, we have almost surely

$$\sum_{i=1}^{2k_n-1} \mathbf{1}_{\{Y_i^{(p)}(\tau_n)>x\}} \mathbf{1}_{E_n} \le \sum_{i=1}^{2k_n-1} \mathbf{1}_{\{U_i^{n,p}>x\}}.$$

For $1 \le i \le 2k_n - 1$, the $U_i^{n,p}$ are independent and identically distributed according to the law of the population size of a Yule process with birth rate p at time $\bar\beta(n)+\ln^2(k_n)$ when started from a single individual. Moreover, the families $(U_i^{n,p} : 1 \le i \le 2k_n - 1)$ and $(U_i^{n,p} : 2k_n \le i \le n^*)$ are independent of each other. Letting

$$\bar S^{n,p}(x) = \sum_{i=1}^{n^*} \mathbf{1}_{\{U_i^{n,p}>x\}},$$

we note that our above estimates yield

$$N^{n,p}(x)\, \mathbf{1}_{E_n} \le \bar S^{n,p}(x) \quad \text{almost surely.} \tag{3.5}$$

3.2. A lower bound

Our lower bound on $N^{n,p}(x)\, \mathbf{1}_{E_n}$ begins with the trivial estimate

$$N^{n,p}(x) \ge \sum_{i=2k_n}^{\infty} \mathbf{1}_{\{Y_i^{(p)}(\tau_n)>x\}},$$

that is, we will consider only summands with an index $i \ge 2k_n$. From Corollary 3.1, we see that on the good event E_n,

$$\tau_n \ge \tau_{k_n} + \underline\beta(n) = b_{k_n}^{(p)} + \underline\beta(n) \ge b_i^{(p)} + \underline\beta(n) - \bar\beta(i). \tag{3.6}$$

Now let

$$n_* = \max\{i \in \mathbb{N} : \bar\beta(i) \le \underline\beta(n)\}.$$

Defining $V_i^{n,p} = Y_i^{(p)}(b_i^{(p)} + \underline\beta(n) - \bar\beta(i))$, we arrive with (3.6) at the almost sure lower bound

$$\sum_{i=2k_n}^{\infty} \mathbf{1}_{\{Y_i^{(p)}(\tau_n)>x\}} \mathbf{1}_{E_n} \ge \sum_{i=2k_n}^{n_*} \mathbf{1}_{\{V_i^{n,p}>x\}} \mathbf{1}_{E_n}.$$

The variables $V_i^{n,p}$, $2k_n \le i \le n_*$, are independent, and $V_i^{n,p}$ is distributed as the population size of a Yule process with birth rate p at time $\underline\beta(n) - \bar\beta(i)$ when started from a single individual. With

$$\underline S^{n,p}(x) = \sum_{i=2k_n}^{n_*} \mathbf{1}_{\{V_i^{n,p}>x\}},$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

we have shown that

$$N^{n,p}(x)\,\mathbf{1}_{E_n} \geq \underline{S}^{n,p}(x)\,\mathbf{1}_{E_n} \quad \text{almost surely.} \tag{3.7}$$

Note that the $V_i^{n,p}$ are not independent of the event E_n. By construction, we have $V_i^{n,p} \leq U_i^{n,p}$ almost surely for $2k_n \leq i \leq n_*$.

3.3. Bounds on the mean

The crucial step of our approach is to obtain control over the means $\mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{E_n}]$ and $\mathbb{E}[\bar{S}^{n,p}(x)]$. In this section we develop bounds valid for all regimes. The exact asymptotic analysis depends on the choice of $(p_n)_n$ and $(x_n)_n$ and will be postponed to Section 4.

The behavior of both expectations will be dominated by the following integral.

Lemma 3.2. *Let $p > 0$, $x > 0$. Then*

$$\int_0^\infty (1 - \mathrm{e}^{-ps})^x \mathrm{e}^{-s}\,\mathrm{d}s = \frac{\Gamma(1/p+1)\Gamma(x+1)}{\Gamma(1/p+x+1)},$$

where Γ denotes the gamma function.

Proof. We first let $u = \mathrm{e}^{-s}$ and then $v = u^p$ to obtain

$$\begin{aligned}\int_0^\infty (1 - \mathrm{e}^{-ps})^x \mathrm{e}^{-s}\,\mathrm{d}s &= \int_0^1 (1-u^p)^x\,\mathrm{d}u \\ &= \frac{1}{p}\int_0^1 (1-v)^x v^{1/p-1}\,\mathrm{d}v \\ &= \frac{1}{p}\,\mathrm{beta}\left(\frac{1}{p}, x+1\right) \\ &= \frac{\Gamma(1/p+1)\Gamma(x+1)}{\Gamma(1/p+x+1)},\end{aligned}$$

where the last two identities follow from well-known expressions for the beta and gamma functions, respectively. □

From now on, we let for $n, x \in \mathbb{N}$ and $p > 0$,

$$I(n,p,x) = n\frac{\Gamma(1/p+1)\Gamma(x+1)}{\Gamma(1/p+x+1)}.$$

The following asymptotics as $n \to \infty$ are immediately derived from Stirling's formula.

Lemma 3.3. *When $p = p_n \to 0$ and $x = x_n \to \infty$,*

$$I(n,p,x) \sim \sqrt{2\pi}\,n\frac{(1/p)^{1/p+1/2}x^{x+1/2}}{(1/p+x)^{1/p+x+1/2}},$$

whereas if $p = p_n \to 0$ and $x \in \mathbb{N}$ is fixed, then $I(n,p,x) \sim x!\,np^x$.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

3.3.1. *An upper bound on the mean.* In order to estimate $\mathbb{E}[\bar{S}^{n,p}(x)]$, recall the definition of $U_i^{n,p}$ for $1 \le i \le n^*$. When evaluated at time $t \ge 0$, a Yule process with birth rate p, started from a single individual, follows the geometric law with success probability $\exp(-pt)$. For $1 \le i \le 2k_n - 1$, this implies that

$$\mathbb{P}(U_i^{n,p} > x) = (1 - \exp(-p(\bar{\beta}(n) + \ln^2(k_n))))^x.$$

Letting $\Sigma_1^{n,p}(x) = \mathbb{E}[\sum_{i=1}^{2k_n-1} \mathbf{1}_{\{U_i^{n,p}>x\}}]$, we thus have

$$\Sigma_1^{n,p}(x) = 2k_n(1 - \exp(-p(\bar{\beta}(n) + \ln^2(k_n))))^x.$$

In our main applications, we will choose k_n such that $\Sigma_1^{n,p}(x) \to 0$ as $\to \infty$, that is, the first $2k_n - 1$ summands are negligible. For the summands with $i \ge 2k_n$, we obtain

$$\mathbb{E}\Big[\sum_{i=2k_n}^{n^*} \mathbf{1}_{\{U_i^{n,p}>x\}}\Big] = \sum_{i=2k_n}^{n^*} (1 - \exp(-p(\bar{\beta}(n) - \underline{\beta}(i))))^x,$$

and it remains to analyze the sum on the right-hand side. We set $f(n,t) = (1 + 3k_n^{-1/3})k_n \mathrm{e}^t$ and obtain

$$\begin{aligned}
&\sum_{i=2k_n}^{n^*} (1 - \exp(-p(\bar{\beta}(n) - \underline{\beta}(i))))^x \\
&\quad = \sum_{i=2k_n}^{n^*} (1 - \exp(-p(\bar{\beta}(n) - \underline{\beta}(i))))^x \int_{\underline{\beta}(i-1)}^{\underline{\beta}(i)} f(n,t)\,\mathrm{d}t \\
&\quad \le \int_{\underline{\beta}(2k_n-1)}^{\underline{\beta}(n^*)} (1 - \exp(-p(\bar{\beta}(n) - t)))^x f(n,t)\,\mathrm{d}t \\
&\quad \le \int_0^{\bar{\beta}(n)} (1 - \exp(-p(\bar{\beta}(n) - t)))^x f(n,t)\,\mathrm{d}t. \qquad (3.8)
\end{aligned}$$

In the first step, we used the fact that $\int_{\underline{\beta}(i-1)}^{\underline{\beta}(i)} f(n,t)\,\mathrm{d}t = 1$ by (3.2), and for the last equality that $\underline{\beta}(n^*) \le \bar{\beta}(n)$ by definition of n^*; see (3.4). With the change of variables $s = \bar{\beta}(n) - t$ and again (3.2), the last integral is equal to

$$\frac{1 + 3k_n^{-1/3}}{1 - 3k_n^{-1/3}} n \int_0^{\bar{\beta}(n)} (1 - \mathrm{e}^{-ps})^x \mathrm{e}^{-s}\,\mathrm{d}s \le (1 + 10k_n^{-1/3}) n \int_0^{\infty} (1 - \mathrm{e}^{-ps})^x \mathrm{e}^{-s}\,\mathrm{d}s,$$

where the bound on the right-hand side holds for sufficiently large n, recalling that $k_n \to \infty$. For evaluating the integral on the right-hand side, we use Lemma 3.2. We have shown that, for large enough n,

$$\mathbb{E}[N^{n,p}(x)\,\mathbf{1}_{E_n}] \le \mathbb{E}[\bar{S}^{n,p}(x)] \le \Sigma_1^{n,p}(x) + (1 + 10k_n^{-1/3}) I(n,p,x). \qquad (3.9)$$

3.3.2. *A lower bound on the mean.* We turn to $\mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{E_n}]$ and write

$$\mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{E_n}] = \mathbb{E}[\underline{S}^{n,p}(x)] - \mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{(E_n)^c}]. \qquad (3.10)$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Set $\Sigma_2^{n,p}(x) = \mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{(E_n)^c}]$. The term $\Sigma_2^{n,p}(x)$ is small compared to $\mathbb{E}[\underline{S}^{n,p}(x)]$: indeed, with Remark 3.1 below and the Cauchy–Schwarz inequality,

$$\Sigma_2^{n,p}(x) \le 2(1-\mathbb{P}(E_n))^{1/2}\mathbb{E}[\underline{S}^{n,p}(x)].$$

Concerning the first term on the right-hand side of (3.10), we have

$$\mathbb{E}[\underline{S}^{n,p}(x)] = \sum_{i=2k_n}^{n_*} (1-\exp(-p(\underline{\beta}(n)-\bar{\beta}(i))))^x.$$

Then, with $g(n,t) = (1-3k_n^{-1/3})k_n\mathrm{e}^t$, a similar calculation as for (3.8) shows that

$$\sum_{i=2k_n}^{n_*} (1-\exp(-p(\underline{\beta}(n)-\bar{\beta}(i))))^x \ge \int_{\bar{\beta}(2k_n)}^{\underline{\beta}(n)} (1-\exp(-p(\underline{\beta}(n)-t)))^x g(n,t)\,\mathrm{d}t.$$

Setting $s = \underline{\beta}(n) - t$, the last integral is bounded from below by

$$(1-10k_n^{-1/3})n\int_0^{\underline{\beta}(n)-\bar{\beta}(2k_n)} (1-\mathrm{e}^{-ps})^x\mathrm{e}^{-s}\,\mathrm{d}s.$$

Now let

$$\Sigma_3^{n,p}(x) = (1-10k_n^{-1/3})n\int_{\underline{\beta}(n)-\bar{\beta}(2k_n)}^{\infty} (1-\mathrm{e}^{-ps})^x\mathrm{e}^{-s}\,\mathrm{d}s.$$

The term $\Sigma_3^{n,p}(x)$ will be negligible in our applications. Using again Lemma 3.2, we have shown that

$$\mathbb{E}[\underline{S}^{n,p_n}(x)] \ge (1-10k_n^{-1/3})I(n,p,x) - \Sigma_3^{n,p}(x), \tag{3.11}$$

and, therefore,

$$\mathbb{E}[N^{n,p}(x)\,\mathbf{1}_{E_n}] \ge \mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{E_n}] \ge (1-10k_n^{-1/3})I(n,p,x) - (\Sigma_2^{n,p}(x)+\Sigma_3^{n,p}(x)). \tag{3.12}$$

In what follows, we refer to the quantities $\Sigma_1^{n,p}(x)$ and $\Sigma_2^{n,p}(x)$, $\Sigma_3^{n,p}(x)$ as the error terms of $\mathbb{E}[\bar{S}^{n,p}(x)]$ and $\mathbb{E}[\underline{S}^{n,p}(x)\,\mathbf{1}_{E_n}]$, respectively, whereas the term $I(n,p,x)$ involving the gamma function (which is the same for both expectations) is referred to as the main term.

3.4. General strategy

Here we outline our general program of how to obtain nondegenerate (Poisson) limit laws for $N^{n,p}(x)$ when $p = p_n \to 0$ and $x = x_n$ may depend on n as well. We will control $N^{n,p}(x)$ on the good event E_n in terms of the lower and upper bounds $\underline{S}^{n,p}(x)$ and $\bar{S}^{n,p}(x)$.

Our strategy is based on the following general observation.

Lemma 3.4. *Let $(U_n : n \in \mathbb{N})$ and $(V_n : n \in \mathbb{N})$ be sequences of $\mathbb{R}$-valued uniformly integrable random variables with $U_n \le V_n$ almost surely for all $n \in \mathbb{N}$. Assume that $V_n \xrightarrow{\mathrm{D}} V$ for some random variable V. Then, if $\mathbb{E}[U_n] \to \mathbb{E}[V]$, we have $U_n \xrightarrow{\mathrm{D}} V$ as $n \to \infty$.*

Proof. Writing $U_n = V_n + (U_n - V_n)$, the claim follows if we show that $U_n - V_n \to 0$ in probability. Since $U_n \le V_n$ almost surely, we have $\mathbb{E}[|V_n - U_n|] = \mathbb{E}[V_n] - \mathbb{E}[U_n]$. Uniform integrability and the fact that $V_n \to V$ in distribution implies $\mathbb{E}[V_n] \to \mathbb{E}[V]$; see, e.g. [10, Theorem 3.5]. Since $\mathbb{E}[U_n] \to \mathbb{E}[V]$ by assumption, the proof is complete. □

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

We stress that instead of assuming $V_n \xrightarrow{\mathrm{D}} V$ and $\mathbb{E}[U_n] \to \mathbb{E}[V]$, one could also *assume* $U_n \xrightarrow{\mathrm{D}} V$ and $\mathbb{E}[V_n] \to \mathbb{E}[V]$ and then *deduce* $V_n \xrightarrow{\mathrm{D}} V$ (in fact, for positive integrable random variables, this can be proved by Fatou's lemma via Skorokhod's representation theorem, without assuming uniform integrability). However, for us it is more natural to work under the assumption of Lemma 3.4.

Remark 3.1. Uniform integrability will not pose any problem in our setting: if V_n is a sum of indicators of independent events then

$$\mathbb{E}[V_n^2] \le \mathbb{E}[V_n] + \mathbb{E}[V_n]^2,$$

implying that if $\sup_{n\in\mathbb{N}} \mathbb{E}[V_n] < \infty$ then $(V_n : n \in \mathbb{N})$ has a uniformly bounded second moment as well and is, therefore, uniformly integrable.

We now assume that the sequence $(p_n)_n$ is fixed and satisfies $p_n \to 0$. We ask for nondegenerate limits of $N^{n,p_n}(x_n)$. Recall the definition of $\bar{S}^{n,p_n}(x_n)$. Our strategy consists of the following steps. We use the notation from the preceding sections.

Step 1. Find a sequence $(x_n)_n \subset \mathbb{N}$ such that, for a suitable choice of $(k_n)_n$ with $p_n k_n \to 0$,

$$\liminf_{n\to\infty} \mathbb{E}[\underline{S}^{n,p_n}(x_n)\,\mathbf{1}_{E_n}] \ge \lambda \quad \text{and} \quad \limsup_{n\to\infty} \mathbb{E}[\bar{S}^{n,p_n}(x_n)] \le \lambda$$

for some strictly positive constant $\lambda > 0$. By construction, this implies that

$$\mathbb{E}[\underline{S}^{n,p_n}(x_n)\,\mathbf{1}_{E_n}] \sim \mathbb{E}[N^{n,p_n}(x_n)\,\mathbf{1}_{E_n}] \sim \mathbb{E}[\bar{S}^{n,p_n}(x_n)] \sim \lambda.$$

Step 2. Show that for $(k_n)_n$ as under step 1, with $a_n = n^*$ and $q_{n,i} = \mathbb{P}(U_i^{n,p_n} > x_n)$,

$$\lim_{n\to\infty} \sum_{i=1}^{a_n} q_{n,i}^2 = 0.$$

Step 3. Le Cam's inequality applied to the sum $\bar{S}^{n,p_n}(x_n)$ yields $\bar{S}^{n,p_n}(x_n) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda)$. Apply Lemma 3.4 to deduce that $N^{n,p_n}(x_n)\,\mathbf{1}_{E_n} \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda)$ and, hence, $N^{n,p_n}(x_n) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda)$ as $n \to \infty$.

Note that the constant λ in step 1 does not depend on the sequence $(k_n)_n$. Of course, depending on the point of view, one can also *fix* a sequence of thresholds $(x_n)_n$ and then *ask* for a choice of $(p_n)_n$ such that nondegenerate limits of $N^{n,p_n}(x_n)$ appear.

We remark that Le Cam's inequality can be used to obtain quantitative bounds on the rate of convergence. In our case, since we regard $N^{n,p_n}(x_n)$ only on the good event E_n, one would have to optimize the choice of this event for establishing good bounds. This will not be our concern here.

A modification of our strategy will allow us to prove a weak law of large numbers for $N^{n,p_n}(x_n)$ when $(x_n)_n$ is chosen such that $N^{n,p_n}(x_n) \to \infty$ in probability.

4. Limit results for subpopulation sizes

We will now work out our strategy explained in the last section. Even though it can be applied to any choice of p_n such that $p_n \to 0$, we will restrict ourselves to discussing three regimes of particular interest. Each of them corresponds to a different limiting behavior of the ancestral subpopulation, as discussed in Proposition 2.1.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

4.1. The regime of bounded subpopulations

In Section 2.3 we argued heuristically that in the regime

$$p_n \sim \frac{a}{n^{1/\ell}}, \quad a > 0 \text{ and } \ell \in \mathbb{N} \text{ fixed},$$

there should be a Poissonian number of subpopulations of size $> \ell$ when $n \to \infty$. We will now prove this rigorously, together with the fact that there are no subpopulations of a size strictly larger than $\ell+1$, and of order $O(n^{(\ell+1-j)/\ell})$ many subpopulations of size j for each $1 \le j \le \ell$. We point to Theorem 5.4 of [11] for similar results for percolation on the complete graph, that is, the Erdős–Rényi random graph model.

It will be convenient to use the notation $\Delta^{n,p_n}(j) = N^{n,p_n}(j-1) - N^{n,p_n}(j)$ for the number of subpopulations of size equal to j. Recall that c and C denote generic constants whose values may change from line to line.

Theorem 4.1. *Fix $\ell \in \mathbb{N}$ and $a > 0$, and assume that $p_n \sim an^{-1/\ell}$. Then, as $n \to \infty$,*

$$\Delta^{n,p_n}(\ell+1) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\ell!\, a^\ell).$$

Moreover, $N^{n,p_n}(\ell+1) \xrightarrow{\mathbb{P}} 0$, and, for each $1 \le j \le \ell$, we have the weak law of large numbers

$$\frac{\Delta^{n,p_n}(j)}{n^{(\ell+1-j)/\ell}} \xrightarrow{\mathbb{P}} (j-1)!\, a^{j-1}.$$

Proof. Our first two claims follow if we show that $N^{n,p_n}(\ell) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\ell!\, a^\ell)$ and $N^{n,p_n}(\ell+1) \xrightarrow{\mathbb{P}} 0$ as n tends to ∞. We follow the strategy outlined in Section 3. We choose $k_n = \lfloor \ln n \rfloor$ and first bound $\mathbb{E}[\bar{S}^{n,p_n}(\ell)]$ from above via (3.9). The error term $\Sigma_1^{n,p_n}(\ell)$ is estimated by

$$\Sigma_1^{n,p_n}(\ell) \le \frac{3a^\ell \ln^{\ell+1} n}{n} = o(1).$$

For the main term in (3.9), we obtain, with Lemma 3.3, as $n \to \infty$,

$$I(n, p_n, \ell) = n\frac{\Gamma(1/p_n+1)\Gamma(\ell+1)}{\Gamma(1/p_n+\ell+1)} \sim \ell!\, np_n^\ell = \ell!\, a^\ell + o(1).$$

The last two displays imply that

$$\mathbb{E}[\bar{S}^{n,p_n}(\ell)] \le \ell!\, a^\ell + o(1).$$

We turn to a lower bound on $\mathbb{E}[\underline{S}^{n,p_n}(\ell)\, \mathbf{1}_{E_n}]$, which we will establish via (3.12). We first show that the error terms $\Sigma_2^{n,p_n}(\ell)$ and $\Sigma_3^{n,p_n}(\ell)$ tend to 0. For that purpose, recall that, by construction,

$$\mathbb{E}[\underline{S}^{n,p_n}(\ell)] \le \mathbb{E}[\bar{S}^{n,p_n}(\ell)] \le C.$$

Hence, the second moment of $\underline{S}^{n,p_n}(\ell)$ is uniformly bounded as well, see Remark 3.1, so that, by the Cauchy–Schwarz inequality,

$$\Sigma_2^{n,p_n}(\ell) \le C(1 - \mathbb{P}(E_n))^{1/2} = o(1).$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

For the error term $\Sigma_3^{n,p_n}(\ell)$ in (3.12), we note that $\underline{\beta}(n) - \bar{\beta}(2k_n) \geq \ln n - 3\ln\ln n$, whence

$$\begin{aligned}\Sigma_3^{n,p_n}(\ell) &\leq n\int_{\underline{\beta}(n)-\bar{\beta}(2k_n)}^{\infty}(1-\mathrm{e}^{-p_n s})^{\ell}\mathrm{e}^{-s}\,\mathrm{d}s\\ &\leq \frac{1}{n} + n\int_{\ln n-3\ln\ln n}^{2\ln n}(1-\mathrm{e}^{-p_n s})^{\ell}\mathrm{e}^{-s}\,\mathrm{d}s\\ &\leq C\frac{\ln^{4\ell} n}{n}\\ &= o(1).\end{aligned}$$

Since the main term of (3.12) agrees with that of (3.9), we obtain $\mathbb{E}[\underline{S}^{n,p_n}(\ell)\,\mathbf{1}_{E_n}] \geq \ell!\,a^{\ell} - o(1)$, and, consequently,

$$\mathbb{E}[\underline{S}^{n,p_n}(\ell)\,\mathbf{1}_{E_n}] \sim \mathbb{E}[N^{n,p_n}(\ell)\,\mathbf{1}_{E_n}] \sim \mathbb{E}[\bar{S}^{n,p_n}(\ell)] \sim \ell!\,a^{\ell} \quad \text{for } n\to\infty.$$

Step 1 of the strategy is therefore established (with the constant λ, there given by $\ell!\,a^{\ell}$), and so is step 2, since

$$\begin{aligned}\sum_{i=1}^{n^*}\mathbb{P}(U_i^{n,p_n} > \ell)^2 &\leq \max_i \mathbb{P}(U_i^{n,p_n} > \ell)\mathbb{E}[\bar{S}^{n,p_n}(\ell)]\\ &\leq C(1-\exp(-p_n(\bar{\beta}(n)+\ln^2(k_n))))^{\ell}\\ &= o(1).\end{aligned}$$

We follow step 3 and obtain $N^{n,p_n}(\ell) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\ell!\,a^{\ell})$ as n tends to ∞.

For the second part of the theorem, we use Markov's inequality to obtain

$$\mathbb{P}(N^{n,p_n}(\ell+1)\geq 1) \leq \mathbb{E}[N^{n,p_n}(\ell+1)\,\mathbf{1}_{E_n}] + \mathbb{P}(E_n^c) \leq \mathbb{E}[\bar{S}^{n,p_n}(\ell+1)] + o(1).$$

Using again the bound (3.9) for $\mathbb{E}[\bar{S}^{n,p_n}(\ell+1)]$, we first estimate the error term $\Sigma_1^{n,p_n}(\ell+1)$ and then the main term similarly as above and obtain $\mathbb{E}[\bar{S}^{n,p_n}(\ell+1)] \leq Cn^{-1/\ell}$. This proves $N^{n,p_n}(\ell+1) \xrightarrow{\mathbb{P}} 0$ as $n\to\infty$.

It remains to show the weak law of large numbers for $\Delta^{n,p_n}(j)$ when $1\leq j\leq \ell$. Since $\Delta^{n,p_n}(j) = N^{n,p_n}(j-1) - N^{n,p_n}(j)$ and $N^{n,p_n}(\ell)$ is stochastically bounded by the first part of the theorem, it suffices to show that, for $0\leq j\leq \ell-1$,

$$\frac{N^{n,p_n}(j)}{n^{(\ell-j)/\ell}} \xrightarrow{\mathbb{P}} j!\,a^{j}. \tag{4.1}$$

We therefore fix $0\leq j\leq \ell-1$ and note that by (3.5) and (3.7), almost surely

$$\underline{S}^{n,p_n}(j)\,\mathbf{1}_{E_n} \leq N^{n,p_n}(j)\,\mathbf{1}_{E_n} \leq \bar{S}^{n,p_n}(j).$$

Since $\mathbb{P}(E_n)\to 1$, (4.1) follows if we show that, for $n\to\infty$,

$$\frac{\underline{S}^{n,p_n}(j)}{n^{(\ell-j)/\ell}} \xrightarrow{\mathbb{P}} j!\,a^{j} \quad \text{and} \quad \frac{\bar{S}^{n,p_n}(j)}{n^{(\ell-j)/\ell}} \xrightarrow{\mathbb{P}} j!\,a^{j}. \tag{4.2}$$

Recall (3.9) and (3.11). For the error, we have the bounds $\Sigma_1^{n,p_n}(j) \leq 2\ln n$, $\Sigma_2^{n,p_n}(j) \leq C(\ln n)^3$, and $\Sigma_3^{n,p_n}(j) = o(\mathbb{E}[\underline{S}^{n,p_n}(j)])$. Therefore, Lemma 3.3 shows that, for $n\to\infty$,

$$\frac{\mathbb{E}[\underline{S}^{n,p_n}(j)]}{n^{(\ell-j)/\ell}} \sim \frac{\mathbb{E}[\bar{S}^{n,p_n}(j)]}{n^{(\ell-j)/\ell}} \sim j!\,a^{j}.$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Now let $\varepsilon > 0$. By the definition of $\bar{S}^{n,p_n}$ and Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}(|\bar{S}^{n,p_n}(j) - \mathbb{E}[\bar{S}^{n,p_n}(j)]| > \varepsilon n^{(\ell-j)/\ell}) &\le \frac{1}{\varepsilon^2 n^{2(\ell-j)/\ell}} \operatorname{var}(\bar{S}^{n,p_n}(j)) \\ &\le \frac{1-p_n}{\varepsilon^2 n^{2(\ell-j)/\ell}} \mathbb{E}[\bar{S}^{n,p_n}(j)] \\ &= o(1),\end{aligned}$$

and an entirely similar estimate holds for $\underline{S}^{n,p_n}(j)$. This proves (4.2) and, hence, (4.1). □

The following corollary of Theorem 4.1 characterizes the regime where unbounded subpopulations appear as $n \to \infty$. For the sake of clarity, we write $\mathbb{P}_{p_n}$ for the law of the system $\boldsymbol{Y}^{(p_n)}$.

Corollary 4.1. *Let* $(p_n)_n \subset [0, 1]$. *Then*

$$\lim_{K\to\infty} \liminf_{n\to\infty} \mathbb{P}_{p_n}(\textit{there exists } i \in \mathbb{N} \textit{ such that } Y_i^{(p_n)}(\tau_n) > K) = 1$$

if and only if $p_n n^{1/\ell} \to \infty$ *for each* $\ell \in \mathbb{N}$. *If one of the statements fails,*

$$\lim_{K\to\infty} \liminf_{n\to\infty} \mathbb{P}_{p_n}(\textit{there exists } i \in \mathbb{N} \textit{ such that } Y_i^{(p_n)}(\tau_n) > K) = 0.$$

Proof. The proof is a direct application of Theorem 4.1 and is left to the reader. □

4.2. The regime $p_n \sim a \ln^{-1} n$

Parts (i) and (ii) of Proposition 2.1 identify

$$p_n \sim \frac{a}{\ln n}, \quad a > 0 \text{ fixed},$$

as the regime in which the ancestral subpopulation becomes nontrivial. Its size, however, remains bounded. What are the sizes of the largest subpopulations that do appear? As a consequence of Theorem 4.3, we will see that if we shift the subpopulation sizes by $-(c_1 \ln n + c_2 \ln\ln n)$ for some explicit constants $c_1, c_2 > 0$ then for any $\ell \in \mathbb{N}$ and any $\varepsilon > 0$, we find $C = C(\ell, \varepsilon)$ such that the ℓ largest (shifted) sizes are contained in $[-C, C]$ with probability at least $1 - \varepsilon$, provided n is large enough.

While Theorem 4.3 holds whenever $p_n \sim a \ln^{-1} n$, we will first prove Theorem 4.2, which provides a stronger result valid for the $p_n = a \ln^{-1} n$ case. Here we will compute a correction c_3 of order one to the above shift, such that the number of subpopulations greater than $y_n = c_1 \ln n + c_2 \ln\ln n + c_3$ converges to a Poisson limit along all subsequences $(y_{n(m)})_m$ of $(y_n)_n$, whose fractional part has a limit as m tends to ∞.

Before we state the precise formulation of our results, we need some preparation. For $a > 0$, define

$$f_a(t) = 1 + t \ln(at) - \frac{1+at}{a} \ln(1+at), \qquad t > 0. \tag{4.3}$$

On $(0, \infty)$, f_a is a smooth function with $f_a(t) \to 1$ as $t \to 0$ and $f_a(t) \to -\infty$ for $t \to \infty$. Moreover, since on $(0, \infty)$,

$$f_a'(t) = \ln(at) - \ln(1+at) < 0,$$

the function f_a is strictly decreasing, and there is a unique $t^* = t^*(a) \in (0, \infty)$ for which $f_a(t^*) = 0$; see Figure 1.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

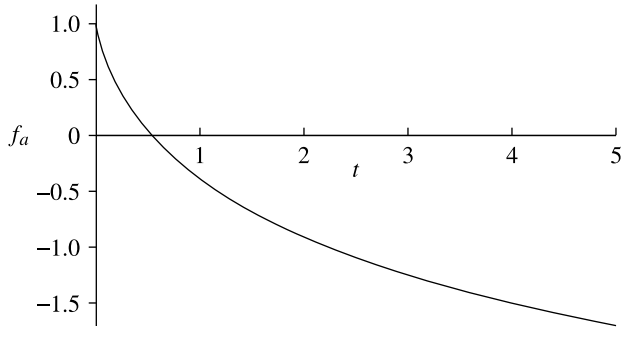


FIGURE 1: The function f_a for $a = 1$.

Now, for $a > 0$ and $\lambda > 0$ fixed and t^* the root of f_a, set

$$y_n = t^* \ln n - \frac{1}{2 f_a'(t^*)} \left(\ln \ln n + \ln \left(\frac{2\pi t^*}{\lambda^2 (1 + a t^*)} \right) \right). \tag{4.4}$$

We use the standard notation $\{y\} = y - \lfloor y \rfloor$ to denote the fractional part of a real $y \in \mathbb{R}$. As we will see in the following theorem, the barrier y_n defines a precise threshold in the following sense. Whenever $(y_{n(m)})_m$ is a subsequence of $(y_n)_n$ such that $\lim_{m\to\infty} \{y_{n(m)}\} =: b \in [0, 1)$ exists, then the number of subpopulations exceeding size $y_{n(m)}$ converges weakly to a Poisson$(\Lambda(a, b, \lambda))$-distributed random variable with intensity given by

$$\Lambda(a, b, \lambda) = \lambda \left(1 + \frac{1}{a t^*} \right)^b. \tag{4.5}$$

We will see in Corollary 4.2 that the restriction to subsequences with converging fractional part is actually necessary.

Theorem 4.2. *Assume that $p_n = a \ln^{-1} n$ for some $a > 0$. Let $\lambda > 0$, and define $y_n = y_n(a, \lambda)$ as in (4.4). Let $(y_{n(m)})_m$ be a subsequence of $(y_n)_n$ such that $\{y_{n(m)}\} \to b$ for some $b \in [0, 1)$. Then, with $\Lambda(a, b, \lambda)$ as above, for $m \to \infty$,*

$$N^{n(m), p_{n(m)}}(\lfloor y_{n(m)} \rfloor) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\Lambda(a, b, \lambda)).$$

Proof. We again follow the strategy explained in Section 3. Recall that we require $p_n k_n \to 0$, and we will here choose $k_n = \lfloor \ln^{1/2} n \rfloor$. Setting $t_n = y_n / \ln n$, the first part of Lemma 3.3 and a small calculation show that as n tends to ∞,

$$I(n, p_n, \lfloor y_n \rfloor) \sim \sqrt{\frac{2\pi t_n}{1 + a t_n}} n \sqrt{\ln n} \frac{(a t_n)^{t_n \ln n}}{(1 + a t_n)^{(a^{-1} + t_n) \ln n}} \left(1 + \frac{1}{a t_n} \right)^{\{t_n \ln n\}}.$$

Taking the logarithm of the right-hand side, we arrive at

$$f_a(t_n) \ln n + \frac{1}{2} \left(\ln \ln n + \ln \left(\frac{2\pi t_n}{1 + a t_n} \right) \right) + \{t_n \ln n\} \ln \left(1 + \frac{1}{a t_n} \right) \tag{4.6}$$

for f_a as defined under (4.3). Obviously, $t_n = t^* + O(\ln \ln n / \ln n)$, and, by Taylor's formula,

$$f_a(t_n) = f_a'(t^*)(t_n - t^*) + O\left(\left(\frac{\ln \ln n}{\ln n} \right)^2 \right),$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

so that

$$f_a(t_n)\ln n = -\frac{1}{2}\left(\ln\ln n + \ln\left(\frac{2\pi t_n}{\lambda^2(1+at_n)}\right)\right) + o(1).$$

Taking exponentials, (4.6) and the last display imply that, for m tending to ∞,

$$I(n(m), p_{n(m)}, \lfloor y_{n(m)}\rfloor) \sim \Lambda(a, b, \lambda).$$

We next control the error terms Σ_1^{n,p_n}, Σ_2^{n,p_n}, and Σ_3^{n,p_n}. First, recalling that $\bar\beta(n) \le \ln n + o(1)$, we have, for sufficiently large n,

$$\begin{aligned}\Sigma_1^{n,p_n}(\lfloor y_n\rfloor) &= 2k_n(1-\exp(-p_n(\bar\beta(n)+\ln^2(k_n))))^{\lfloor y_n\rfloor}\\ &\le 2(\ln^{1/2} n)(1-\exp(-2a))^{\lfloor y_n\rfloor}\\ &= o(1).\end{aligned}$$

In particular,

$$\mathbb{E}[\bar S^{n(m),p_{n(m)}}(\lfloor y_{n(m)}\rfloor)] \le \Lambda(a,b,\lambda) + o(1),$$

and, by the Cauchy–Schwarz inequality and Remark 3.1, $\Sigma_2^{n(m),p_{n(m)}}(\lfloor y_{n(m)}\rfloor) = o(1)$. Finally,

$$\begin{aligned}\Sigma_3^{n,p_n}(\lfloor y_n\rfloor) &\le n\int_{\underline\beta(n)-\bar\beta(2k_n)}^{\infty}(1-\mathrm{e}^{-p_n s})^{\lfloor y_n\rfloor}\mathrm{e}^{-s}\,\mathrm{d}s\\ &\le \frac{1}{n} + n(1-\mathrm{e}^{-2a})^{\lfloor y_n\rfloor}\int_{\ln n-2\ln\ln n}^{2\ln n}\mathrm{e}^{-s}\,\mathrm{d}s\\ &= o(1),\end{aligned}$$

so that, as $m\to\infty$,

$$\begin{aligned}\mathbb{E}[\underline S^{n(m),p_{n(m)}}(\lfloor y_{n(m)}\rfloor)\,\mathbf{1}_{E_{n(m)}}] &\sim \mathbb{E}[N^{n(m),p_{n(m)}}(\lfloor y_{n(m)}\rfloor)\,\mathbf{1}_{E_{n(m)}}]\\ &\sim \mathbb{E}[\bar S^{n(m),p_{n(m)}}(\lfloor y_{n(m)}\rfloor)]\\ &\sim \Lambda(a,b,\lambda).\end{aligned}$$

The condition under step 2 is fulfilled as well, see the estimate for Σ_1^{n,p_n}. Performing step 3, it follows that $N^{n(m),p_{n(m)}}(\lfloor y_{n(m)}\rfloor)$ converges to a Poisson$(\Lambda(a,b,\lambda))$ random variable. □

Theorem 4.2 implies the following remarkable weak convergence along subsequences for the size C_*^{n,p_n} of the largest subpopulation at time τ_n. For $a>0$, $r\in\mathbb{R}$, and t^* the root of f_a, set

$$\lambda_{a,r} = \left(\frac{2\pi t^*}{1+at^*}\right)^{1/2}\mathrm{e}^{-r/2}.$$

Corollary 4.2. *Let $a>0$, $p_n = a\ln^{-1} n$, and $r\in\mathbb{R}$. Define $\lambda_{a,r}$ as above, and $y_n = y_n(a,\lambda_{a,r})$ in terms of $\lambda_{a,r}$ as in (4.4). Let $(y_{n(m)})_m$ be any subsequence of $(y_n)_n$ such that $\{y_{n(m)}\}\to b$ for some $b\in[0,1)$. Then, for $m\to\infty$, with $\mu = \ln(2\pi/a^{2b}) + (1-2b)\ln(t^*/(1+at^*))$,*

$$\mathbb{P}\left(C_*^{n(m),p_{n(m)}} - t^*\ln n(m) + \frac{1}{2f_a'(t^*)}\ln\ln n(m) \le r\right) \to \exp(-\mathrm{e}^{-(r-\mu)/2}).$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Proof. The probability on the left-hand side is equal to

$$\mathbb{P}(C_*^{n(m),p_{n(m)}} \leq \lfloor y_{n(m)} \rfloor) = \mathbb{P}(N^{n(m),p_{n(m)}}(\lfloor y_{n(m)} \rfloor) = 0) \sim \mathbb{P}(\mathrm{Poi}(\Lambda(a,b,\lambda_{a,r})) = 0),$$

with $\Lambda(a,b,\lambda_{a,r})$ defined as in (4.5). The last asymptotics hold thanks to Theorem 4.2. The claim follows. □

Remark 4.1. The limit expression on the right-hand side in the above corollary is the distribution function of the Gumbel law with location parameter μ and scale parameter 2. In view of our heuristics explained in Section 2.3 and of what is known about the maximum of n independent geometrically distributed random variables, it should not come as a surprise that a Gumbel distribution appears in the limit for the recentered size of the largest subpopulation. Since C_*^{n,p_n} is a discrete random variable and the Gumbel law is a continuous distribution, convergence along the full sequence cannot hold.

It is instructive to compare our previous two results with Theorem 5.10 and Corollary 5.11 of [11] for the Erdős–Rényi random graph model.

From the proof of Theorem 4.2, we easily derive that in the general case $p_n \sim a \ln^{-1} n$, the sizes of the largest subpopulations are concentrated around

$$b_n = t^* \ln n - \frac{1}{2 f_a'(t^*)} \ln \ln n. \tag{4.7}$$

Recall the definition of $I(n, p_n, x_n)$ and its asymptotics given in Lemma 3.3. We will again write $\Delta^{n,p_n}(j)$ for the number of subpopulations of size equal to j.

Theorem 4.3. *Assume that $p_n \sim a \ln^{-1} n$ for some $a > 0$. Define $b_n = b_n(a)$ as in (4.7), and let $(x_n)_n$ be a sequence of positive integers. Then the following holds as $n \to \infty$:*

(i) *if $(x_n - b_n) \gg 1$ then $N^{n,p_n}(x_n) \xrightarrow{\mathbb{P}} 0$;*

(ii) *If $(b_n - x_n) \gg 1$ then $N^{n,p_n}(x_n) \xrightarrow{\mathbb{P}} \infty$. More precisely, we have the weak law of large numbers*

$$\frac{N^{n,p_n}(x_n)}{I(n,p_n,x_n)} \xrightarrow{\mathbb{P}} 1.$$

In particular, for $j \in \mathbb{N}$,

$$\frac{\Delta^{n,p_n}(j)}{(\ln n)^{j-1} n} \xrightarrow{\mathbb{P}} (j-1)!\, a^{j-1}.$$

Proof. We have to adapt only the estimates given in the proofs of Theorems 4.1 and 4.2, so we will sketch only the necessary modifications. We again choose $k_n = \lfloor \ln^{1/2} n \rfloor$. Assume first that $(x_n - b_n) \gg 1$. With y_n everywhere replaced by x_n, we deduce from (4.6) that $I(n,p_n,x_n) \to 0$. The error term $\Sigma_1^{n,p_n}(x_n)$ is seen to be of order $o(1)$ with the same argument as in the proof of Theorem 4.2. Therefore, we have by Markov's inequality and (3.9),

$$\mathbb{P}(N^{n,p_n}(x_n) \geq 1) \leq \mathbb{E}[\bar{S}^{n,p_n}(x_n)] + o(1) = o(1).$$

This proves (i). Now assume that $(b_n - x_n) \gg 1$. Then we see from (4.6) that $I(n,p_n,x_n) \to \infty$. The rest is now entirely similar to the proof of the weak law of large numbers in Theorem 4.1: since $\underline{S}^{n,p_n}(x_n)\,\mathbf{1}_{E_n} \leq N^{n,p_n}(x_n)\,\mathbf{1}_{E_n} \leq \bar{S}^{n,p_n}(x_n)$ and $\mathbb{P}(E_n) \to 1$, we have $N^{n,p_n}(x_n) \sim I(n,p_n,x_n)$ in probability if, for $n \to \infty$,

$$\frac{\underline{S}^{n,p_n}(x_n)}{I(n,p_n,x_n)} \xrightarrow{\mathbb{P}} 1 \quad \text{and} \quad \frac{\bar{S}^{n,p_n}(x_n)}{I(n,p_n,x_n)} \xrightarrow{\mathbb{P}} 1.$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

To see why the last display holds, we note that the error terms $\Sigma_1^{n,p_n}(x_n)$, $\Sigma_2^{n,p_n}(x_n)$, and $\Sigma_3^{n,p_n}(x_n)$ are all bounded by $o(I(n, p_n, x_n))$, so that (3.11) together with (3.9) yield

$$\frac{\mathbb{E}[\underline{S}^{n,p_n}(x_n)]}{I(n, p_n, x_n)} \sim \frac{\mathbb{E}[\bar{S}^{n,p_n}(x_n)]}{I(n, p_n, x_n)} \sim 1.$$

Again, Chebyshev's inequality allows us to conclude. Concerning the last statement under (ii), we write $\Delta^{n,p_n}(j) = N^{n,p_n}(j-1) - N^{n,p_n}(j)$. Choosing the sequence $(x_n)_n$ to be constant means that $j - 1 \in \mathbb{N}_0$, the claim then follows from the part we just proved and from the asymptotics of $I(n, p_n, j-1)$ specified in Lemma 3.3. □

4.3. The regime $\ln^{-1} n \ll p_n \ll 1$.

In the regime $p_n \gg 1/\ln n$, the ancestral subpopulation grows as n^{p_n}; see Proposition 2.1(iii). As we will prove in the following theorem, when $1/\ln n \ll p_n \ll 1$, the sizes of the largest subpopulations are to first order given by $\mathrm{e}^{-1} p_n^{-1} n^{p_n}$. For the case $\ln\ln n/\ln n \le p_n \ll 1$ and $\lambda > 0$, we will show that the number of subpopulations greater than

$$x_n = x_n(p_n, \lambda) = \left\lfloor \left(\lambda^{-1}\sqrt{2\pi p_n^{-1}}\right)^{p_n} \mathrm{e}^{-1} p_n^{-1} n^{p_n} \right\rfloor \tag{4.8}$$

is asymptotically Poisson(λ)-distributed. Note that, as $n \to \infty$,

$$g_\lambda(p_n) = \left(\lambda^{-1}\sqrt{2\pi p_n^{-1}}\right)^{p_n} \sim 1;$$

see Figure 2.

As it will become clear from the proof, we require $p_n \ge \ln\ln n/\ln n$ (and not merely $p_n \gg 1/\ln n$) only to simplify the calculation; without this assumption, the exact behavior of p_n has to be taken into account more carefully.

Theorem 4.4. *Assume that* $\ln^{-1} n \ll p_n \ll 1$. *Then the following hold as* $n \to \infty$*:*

(i) *if* $u > \mathrm{e}^{-1}$ *then* $N^{n,p_n}(\lfloor u p_n^{-1} n^{p_n} \rfloor) \xrightarrow{\mathbb{P}} 0$*; if* $u < \mathrm{e}^{-1}$ *then* $N^{n,p_n}(\lfloor u p_n^{-1} n^{p_n} \rfloor) \xrightarrow{\mathbb{P}} \infty$*;*

(ii) *assume, additionally, that* $p_n \ge \ln\ln n/\ln n$ *for large* n*. Then, with* x_n *as in (4.8),*

$$N^{n,p_n}(x_n) \xrightarrow{\mathrm{D}} \mathrm{Poi}(\lambda).$$

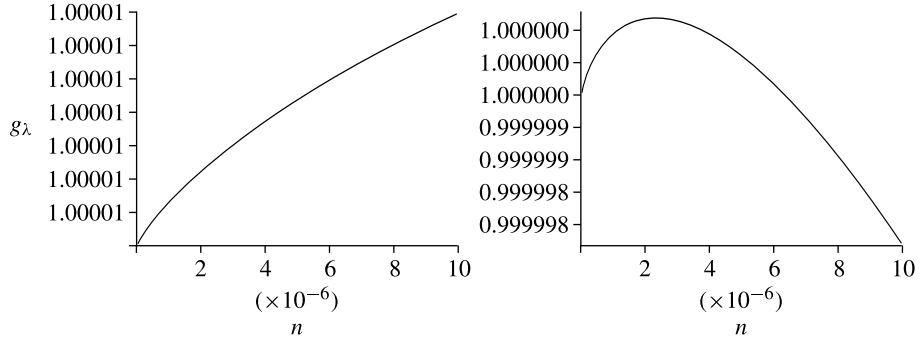


FIGURE 2: The function $g_\lambda(p)$ on $[0, 10^{-5}]$ for $\lambda = 200$ (*left*) and $\lambda = 1000$ (*right*).

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Proof. We work with the choice $k_n = \lfloor p_n^{-1/2} \rfloor$ and follow the steps of the strategy presented in Section 3. Since our proof is similar to the proofs of Theorems 4.1 and 4.2, we do not provide every detail.

We fix a real number $u > 0$ and let $y_n = \lfloor u p_n^{-1} n^{p_n} \rfloor$, $n \in \mathbb{N}$. For the main term $I(n, p_n, y_n)$, Lemma 3.3 shows that

$$I(n, p_n, y_n) \sim \sqrt{2\pi} n \frac{(1/p_n)^{1/p_n+1/2} y_n^{y_n+1/2}}{(1/p_n + y_n)^{1/p_n+y_n+1/2}}. \tag{4.9}$$

Note that since $p_n \gg 1/\ln n$, we have $p_n y_n \to \infty$ as $n \to \infty$. By Taylor's formula,

$$\ln\left(\frac{1}{p_n} + y_n\right) = \ln(y_n) + \frac{1}{p_n y_n} + O\left(\frac{1}{p_n^2 y_n^2}\right).$$

From the last display, we deduce, after a short calculation, that the logarithm of the right-hand side of (4.9) behaves asymptotically as

$$\begin{aligned} &\frac{1}{2}\ln(2\pi) + \ln n - \left(\frac{1}{p_n} + \frac{1}{2}\right)\ln p_n - \frac{1}{p_n}\ln(y_n) - \frac{1}{p_n} + O\left(\frac{1}{p_n^2 y_n}\right) \\ &\quad = -\frac{1}{p_n}(1 + \ln(u) + O(n^{-p_n})) - \frac{1}{2}\ln p_n + \frac{1}{2}\ln(2\pi). \end{aligned} \tag{4.10}$$

In particular, if $u > \mathrm{e}^{-1}$ then the right-hand side diverges to $-\infty$, whereas if $u < \mathrm{e}^{-1}$, it diverges to $+\infty$. Consequently, $I(n, p_n, y_n) \to 0$ in the first case and $I(n, p_n, y_n) \to \infty$ in the second case. In order to finish the proof of (i), it remains to convince ourselves that the error terms $\Sigma_1^{n,p_n}(y_n)$ and $\Sigma_3^{n,p_n}(y_n)$ do not contribute when n tends to ∞. For the first one, we have

$$\Sigma_1^{n,p_n}(y_n) \le 2p_n^{-1/2}(1 - \exp(-p_n(\bar{\beta}(n) + \ln^2(\lfloor p_n^{-1/2} \rfloor))))^{y_n}.$$

Since $\bar{\beta}(n) \le \ln n + o(1)$, we see from taking logarithms that $\Sigma_1^{n,p_n}(y_n) = o(1)$, as required. The error term $\Sigma_3^{n,p_n}(y_n)$ is readily seen to be of order $o(1)$ as well, and the claims under (i) now follow from the same arguments as in the proof of Theorem 4.3. (In fact, if $u < \mathrm{e}^{-1}$, the reasoning there shows the stronger statement $N^{n,p_n}(y_n)/I(n, p_n, y_n) \to 1$ in probability.)

We turn to (ii) and fix $\lambda > 0$. If $p_n \ge \ln\ln n / \ln n$ for n large then, with x_n defined as in (4.8),

$$O\left(\frac{1}{p_n n^{p_n}}\right) = o(1).$$

Performing the above calculation with y_n replaced by x_n and u by

$$u_n = \left(\lambda^{-1}\sqrt{2\pi p_n^{-1}}\right)^{p_n} \mathrm{e}^{-1},$$

we deduce from (4.10) that the logarithm of the right-hand side in (4.9) behaves as $\ln\lambda + o(1)$, and, thus,

$$I(n, p_n, x_n) \sim \lambda.$$

Part (ii) now follows from the same line of reasoning as in the proof of Theorem 4.2, using the fact that all the error terms $\Sigma_i^{n,p_n}(x_n)$, $i = 1, 2, 3$, are of order $o(1)$. □

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Remark 4.2. Let us end this section by noting that our methods allow for various extensions and specifications to other regimes. For instance, independently of the specific regime, we obtain the following law of large numbers for the number $\Delta^{n,p_n}(j)$ of subpopulations of a fixed size j.

Corollary 4.3. *Assume that* $0 < p_n \ll 1$. *Let* $j \in \mathbb{N}$. *If* $p_n^{j-1} n \to \infty$ *as* $n \to \infty$ *then*

$$\frac{\Delta^{n,p_n}(j)}{p_n^{j-1} n} \xrightarrow{\mathbb{P}} (j-1)!.$$

From this perspective, the law of large numbers formulated in Theorems 4.1 and 4.3 are just special cases of this more general form.

5. Subcritical percolation on random recursive trees

In this final section we make the precise link between the population system $\boldsymbol{Y}^{(p)}$ defined in Section 2 and percolation on random recursive trees.

A recursive tree with vertex set $[n] = \{1, \dots, n\}$ is a tree rooted at 1 with the property that the labels along the unique path from the root to any other vertex form an increasing sequence. A tree chosen uniformly at random among all these $(n-1)!$ recursive trees is called a random recursive tree and denoted T_n. The study of Bernoulli bond percolation on large random recursive trees can be traced back to the analysis of an algorithm for isolating the root by Meir and Moon [19]; see also [13], [15], and [17] for more recent developments in this direction. It further plays a key role in the construction of the Bolthausen–Sznitman coalescent by Goldschmidt and Martin [14].

We choose $p = p_n \in (0, 1)$ and write $C_1^{n,p}, C_2^{n,p}, \dots$ for the percolation clusters of a Bernoulli bond percolation on T_n with parameter p. That is, we erase each edge of T_n with probability $1 - p$ and independently of each other, and enumerate the connected components according to the increasing order of the label of their root vertex (that is, their vertex with the smallest label).

We use the convention that $C_i^{n,p} = \varnothing$ if there are less than i connected components after percolation on T_n. With our ordering, $C_1^{n,p}$ always represents the root cluster containing 1; $C_2^{n,p}$ is the cluster rooted at the smallest vertex which does not belong to the root cluster $C_1^{n,p}$, and so on. We write $|C_i^{n,p}|$ for the number of vertices of $C_i^{n,p}$. Then the following holds.

Lemma 5.1. *Let* $p \in [0, 1]$. *The families* $(|C_1^{n,p}|, |C_2^{n,p}|, \dots)$ *and* $(Y_1^{(p)}(\tau_n), Y_2^{(p)}(\tau_n), \dots)$ *have the same law.*

Proof. We construct percolation on a random recursive tree as a growth process in continuous time as follows. At time $t = 0$, we start from the singleton $\{1\}$. Given percolation with parameter p on a random recursive tree on $[k]$, $k \geq 1$, has been constructed, we equip each vertex $i \in [k]$ with an independent exponential clock $\mathcal{E}_i$ of parameter 1. At time $\min_{i=1,\dots,k} \mathcal{E}_i$, we flip a coin with heads probability p. If head shows up, we attach a vertex labeled $k+1$ to the vertex with label $\arg\min_{i=1,\dots,k} \mathcal{E}_i$. Otherwise, we add vertex $k+1$ to the system without connecting it to any other vertex. It follows from the construction of random recursive trees and the independence of the coin flips that we observe at the instant when the nth vertex is incorporated, a Bernoulli bond percolation on T_n with parameter p. Moreover, if we keep track of the sizes $\boldsymbol{C}^{(p)}(t) = (|C_1^{(p)}(t)|, |C_2^{(p)}(t)|, \dots)$ of the growing percolation clusters, where $|C_i^{(p)}(t)|$ stores the size of the ith cluster at time t and clusters are ordered according to their

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

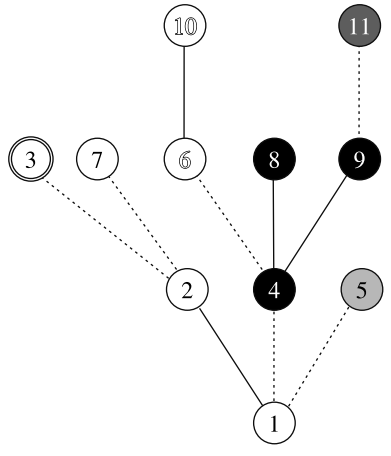


FIGURE 3: The genealogical tree of a Yule process with mutation, stopped after the birth of the 11th individual. Subpopulations of the same genetic type are shown as the same style of numbered circle, and the edges connecting mutants to their parent are represented by dashed lines. Alternatively, the subpopulations can be viewed as the clusters of a Bernoulli bond percolation on a recursive tree on $\{1, \dots, 11\}$.

birth times, we obtain a Markov chain with initial state $\boldsymbol{C}^{(p)}(0) = (1, 0, \dots)$ and transition rates at time $t \geq 0$ for $c_1, \dots, c_k \in \mathbb{N}$ given by

$$(c_1, \dots, c_k, 0, \dots) \mapsto (c_1, \dots, c_k, 1, 0, \dots) \quad \text{at rate } (1-p)(c_1 + \cdots + c_k),$$
$$(c_1, \dots, c_k, 0, \dots) \mapsto (c_1, \dots, c_{i-1}, c_i + 1, c_{i+1}, \dots, c_k, 0, \dots) \quad \text{at rate } pc_i.$$

From the very construction of $\boldsymbol{Y}^{(p)}$, it follows that the processes $(\boldsymbol{Y}^{(p)}(t) : t \geq 0)$ and $(\boldsymbol{C}^{(p)}(t) : t \geq 0)$ grow according to the same dynamics, and the statement follows. □

If we think of the individuals of $\boldsymbol{Y}^{(p)}$ as being labeled $1, 2, \dots$, according to the increasing order of their birth times, there is in fact a one-to-one correspondence between subpopulations and clusters that respects the genealogy, as illustrated in Figure 3.

With the above lemma at hand, the results from Section 4 have a direct interpretation in terms of clusters stemming from a percolation on T_n with parameter p_n. For example, Corollary 4.1 gives a necessary and sufficient condition on the percolation parameter p_n such that in the limit $n \to \infty$, clusters of unbounded size appear.

More generally, the strategy developed in Section 3 gives a concise tool to decide whether, for a given percolation sequence $p_n \to 0$ and thresholds x_n, there are clusters of a size of order x_n, and, if so, how many. Indeed, as exemplified in Section 4, basically one has to check only the asymptotic behavior of the expression $I(n, p_n, x_n)$ introduced below Lemma 3.2.

This work completes the study of percolation on random recursive trees initiated in [7], where percolation with supercritical parameter $p_n \sim 1 - a/\ln n$, $a > 0$ fixed, was studied. In [6], non-Gaussian fluctuations of the root cluster were proved, and the analysis of percolation clusters was extended in [4] to all regimes $p_n \to 1$. The works [4] and [7] contain additional information on the genealogy of clusters (and not merely on their sizes), a type of question we did not investigate here.

As a common feature when $p_n \to 1$, the root cluster containing 1 has size $\sim n^{p_n}$, while the sizes of the next largest clusters are of order $(1 - p_n)n^{p_n}$. This motivated a subclassification of regions into weakly supercritical $[1/\ln n \ll 1 - p_n \ll 1]$ when $|C_1^{n,p_n}| \ll n$, supercritical

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

$[1 - p_n \sim t/\ln n,\ t > 0$ fixed] when $|C_1^{n,p_n}| \sim \mathrm{e}^{-t}n$, and strongly supercritical $[1 - p_n \ll 1/\ln n]$ when $|C_1^{n,p_n}| \sim n$.

The regime of constant $p \in (0, 1)$ may best be termed 'critical', since in this case, the root cluster loses its dominating role (with respect to the size). Some results can be found in [5] and [20], although the motivation there is somehow different. For similar reasons, we term the regime $p = p_n \to 0$ considered here 'subcritical'. The classification into different regimes draws on the terminology which is usually used to describe the Erdős–Rényi random graph model; see, e.g. [2].

We stress that there are other natural families of trees which can be grown according to a probabilistic evolution algorithm, such as scale-free random trees or b-ary random increasing trees. Indeed, branching systems with mutations were used in [8] to study percolation on scale-free random trees when $p_n \sim a/\ln n$, and in a similar fashion by Berzunza in [9] for b-ary random increasing trees. In the case of random recursive trees, the underlying population system is particularly simple, so we restricted our discussion of the subcritical regime to these trees, but we certainly expect similar results to hold for other classes of increasing tree families.

Appendix A. Proof of Lemma 3.1

In this appendix we will construct an event E_n with the 'good' properties specified in Lemma 3.1. We first collect some properties of Yule processes.

A.1. Some properties of Yule processes

For $p \in [0, 1]$, denote by $Y^{(p)} = (Y^{(p)}(t) : t \geq 0)$ a Yule process with birth rate p per unit population size. Given $y \in \mathbb{N}$, write $\mathbb{P}_y$ for its law under which $\mathbb{P}_y(Y^{(p)}(0) = y) = 1$, and, similarly, $\mathbb{E}_y$ for its expectation. It is well known that $W^{(p)}(t) = \mathrm{e}^{-pt}Y^{(p)}(t), t \geq 0$, is a square-integrable martingale. Under $\mathbb{P}_y$ and for $p > 0$, its terminal value $W^{(p)}(\infty)$ is distributed as a sum of y standard exponential random variables and thus follows the gamma$(y, 1)$ law. Note that Doob's inequality applied to the square-integrable martingale $W^{(p)}(t) - y$ yields

$$\mathbb{E}_y\Big[\sup_{t\geq 0} |W^{(p)}(t) - W^{(p)}(0)|^2\Big] \leq 4\mathbb{E}_y[|W^{(p)}(\infty) - W^{(p)}(0)|^2] = 4y. \tag{A.1}$$

We now work in the setting of Section 2. Recall the definition of the process $T^{(p)} = (T^{(p)}(t) : t \geq 0)$, counting the number of different genetic types in the population system $\boldsymbol{Y}^{(p)}$, starting from $T^{(p)}(0) = 1$. For $i \in \mathbb{N}$, $\tau_i = \inf\{t \geq 0 : Z(t) = i\}$ denotes the birth time of the ith individual in the system $\boldsymbol{Y}^{(p)}$ (counting both clones and mutants), with Z denoting the underlying standard Yule process. The time $b_i^{(p)}$ denotes the first time when an individual of genetic type i appears, that is, the $b_i^{(p)}$ are the jump times of $T^{(p)}(t)$. We need the following control over $T^{(p)}$.

Lemma A.1. *Let $p \in (0, 1)$. For $k \in \mathbb{N}$, let $\mathbb{P}_{|k}$ be the conditional law given $\{\tau_k = b_k^{(p)}\}$ (that is, the first $k - 1$ children are all mutants), and let $\mathbb{E}_{|k}$ be its expectation. For every $t \geq 0$, we have*

$$\mathbb{E}_{|k}\left[\sup_{0\leq s\leq t} \left|T^{(p)}(\tau_k + s) - \left(k + \int_{\tau_k}^{\tau_k+s} (1 - p)Z(r)\,\mathrm{d}r\right)\right|^2\right] \leq 4(1 - p)k(\mathrm{e}^t - 1)$$

and

$$\mathbb{E}_{|k}\left[\sup_{s\geq t} \mathrm{e}^{-5s/3}\left|T^{(p)}(\tau_k + s) - \left(k + \int_{\tau_k}^{\tau_k+s} (1 - p)Z(r)\,\mathrm{d}r\right)\right|^2\right] \leq 324(1 - p)k\mathrm{e}^{-2t/3}.$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Proof. Let $\mathcal{F}_t$ denote the natural filtration generated by the process $(\boldsymbol{Y}^{(p)}(t)\colon t \geq 0)$. Both processes $T^{(p)}$ and Z are $\mathcal{F}_t$-adapted, and τ_k is a stopping time. We work now under the probability measure $\mathbb{P}_{|k}$. From the strong Markov property and the dynamics of $\boldsymbol{Y}^{(p)}$ described above, we see that $M^{(p)}(s) = T^{(p)}(\tau_k + s) - (k + \int_{\tau_k}^{\tau_k+s}(1-p)Z(r)\,\mathrm{d}r)$ is a martingale with $M^{(p)}(0) = 0$. Its quadratic variation is given by $[M^{(p)}](s) = T^{(p)}(\tau_k + s) - k$, and its second moment takes the form

$$\begin{aligned}\mathbb{E}_{|k}[|M^{(p)}(s)|^2] &= \mathbb{E}_{|k}[[M^{(p)}](s)]\\ &= \mathbb{E}_{|k}[T^{(p)}(\tau_k + s)] - k\\ &= \mathbb{E}_{|k}\left[\int_{\tau_k}^{\tau_k+s}(1-p)Z(r)\,\mathrm{d}r\right]\\ &= \int_0^s (1-p)\mathbb{E}[Z(\tau_k + r)]\,\mathrm{d}r\\ &= (1-p)k(\mathrm{e}^t - 1).\end{aligned}$$

For the last equality, we used the strong Markov property, which entails that $(Z(\tau_k + r)\colon r \geq 0)$ is independent of $\mathcal{F}_{\tau_k}$ and distributed as a standard Yule process started from k individuals. An appeal to Doob's inequality yields

$$\mathbb{E}_{|k}\Big[\sup_{0\leq s\leq t}|M^{(p)}(s)|^2\Big] \leq 4\mathbb{E}_{|k}[|M^{(p)}(t)|^2] = 4(1-p)k(\mathrm{e}^t - 1).$$

For the second statement, we bound, for every $n \in \mathbb{N}$,

$$\sup_{n\leq s<n+1}\mathrm{e}^{-5s/6}|M^{(p)}(s)| \leq \mathrm{e}^{-5n/6}\sup_{n\leq s<n+1}|M^{(p)}(s)|.$$

By Doob's inequality, see the first part, the L^2-norm of the right-hand side (with respect to the conditional measure $\mathbb{P}_{|k}$) is bounded by

$$\mathrm{e}^{-5n/6}\Big\|\sup_{n\leq s<n+1}|M^{(p)}(s)|\Big\|_2 \leq 2\mathrm{e}^{-5n/6}\sqrt{(1-p)k\mathrm{e}^{n+1}} \leq 2\mathrm{e}^{1/2}\sqrt{(1-p)k}\,\mathrm{e}^{-n/3}.$$

Thus, summing over n and applying the triangle inequality,

$$\Big\|\sup_{s\geq t}\mathrm{e}^{-5s/6}|M^{(p)}(s)|\Big\|_2 \leq 2\mathrm{e}^{1/2}\sqrt{(1-p)k}\sum_{n\geq\lfloor t\rfloor}\mathrm{e}^{-n/3} \leq 18\sqrt{(1-p)k}\,\mathrm{e}^{-t/3}.$$

Squaring both sides, the claim follows. □

We finally turn to the construction of an event $E_n = E_{n,p_n,k_n}$ that satisfies the properties of Lemma 3.1.

Proof of Lemma 3.1. We assume that $p_n \to 0$ and fix an arbitrary sequence of integers $k_n \in \mathbb{N}$ satisfying $k_n \to \infty$ and $k_n p_n \to 0$ as $n \to \infty$. Concerning Lemma 3.1(i), we first note that since each newborn child is a mutant with probability $1 - p_n$, we have

$$\mathbb{P}(\tau_{k_n} = b_{k_n}^{(p_n)}) = (1-p_n)^{k_n-1} = 1 - o(1).$$

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

Moreover, by (2.1), we also have $\mathbb{P}(\tau_{k_n} < (\ln k_n)^2) = 1 - o(1)$, so that the event

$$E_n^1 = \{\tau_{k_n} = b_{k_n}^{(p_n)}\} \cap \{\tau_{k_n} < (\ln k_n)^2\}$$

has probability $1 - o(1)$ as n tends to ∞.

We turn to Lemma 3.1(ii). Since conditionally on τ_{k_n}, $(Z(t + \tau_{k_n}) : t \geq 0)$ is a Yule process started from k_n individuals, we obtain, from (A.1) and Chebyshev's inequality,

$$\mathbb{P}\Big(\sup_{t\geq 0} |\mathrm{e}^{-t} Z(t + \tau_{k_n}) - k_n| > k_n^{2/3}\Big) \leq \frac{4}{k_n^{1/3}}.$$

In particular, if we let

$$E_n^2 = \{(1 - k_n^{-1/3})k_n \mathrm{e}^t \leq Z(\tau_{k_n} + t) \leq (1 + k_n^{-1/3})k_n \mathrm{e}^t \text{ for all } t \geq 0\},$$

then $\mathbb{P}(E_n^2) \geq 1 - 4/k_n^{1/3} = 1 - o(1)$.

Finally, for Lemma 3.1(iii), we need to control $(T^{(p)}(t) : t \geq 0)$. In this regard, Lemma A.1 in combination with Chebyshev's inequality shows that, conditionally on $\{\tau_{k_n} = b_{k_n}^{(p_n)}\}$, the event

$$E_n^3 = \left\{ \left| T^{(p_n)}(\tau_{k_n} + t) \left(k_n + \int_{\tau_{k_n}}^{\tau_{k_n}+t} (1 - p_n) Z(r)\, \mathrm{d}r \right) \right| \leq \mathrm{e}^{5t/6} k_n^{2/3} \text{ for all } t \geq 0 \right\}$$

has probability at least $1 - 324/k_n^{1/3}$. On $E_n^2 \cap E_n^3$, we note that for sufficiently large n, recalling that $p_n k_n \to 0$,

$$T^{(p_n)}(\tau_{k_n} + t) \geq k_n + (1 - p_n)(1 - k_n^{-1/3})k_n \int_0^t \mathrm{e}^r\, \mathrm{d}r - \mathrm{e}^{5t/6} k_n^{2/3} \geq (1 - 3k_n^{-1/3})k_n \mathrm{e}^t,$$

and, similarly,

$$T^{(p_n)}(\tau_{k_n} + t) \leq (1 + 3k_n^{-1/3})k_n \mathrm{e}^t.$$

Setting

$$E_n = E_n^1 \cap E_n^2 \cap E_n^3,$$

we have constructed an event of probability $1 - o(1)$ on which, for sufficiently large n, Lemma 3.1(i)–(iii) is fulfilled. □

Acknowledgements

The research of E.B. was supported by the Swiss National Science Foundation (grant number P300P2_161011) and performed within the framework of the LABEX MILYON (grant number ANR-10-LABX-0070) of the Université de Lyon, within the program 'Investissements d'Avenir' (grant number ANR-11-IDEX-0007) operated by the French National Research Agency (ANR).

References

[1] Aldous, D. (1989). *Probability Approximations Via the Poisson Clumping Heuristic*. Springer, New York.
[2] Alon, N. and Spencer, J. H. (2008). *The Probabilistic Method,* 3rd edn. John Wiley, Hoboken, NJ.
[3] Athreya, K. B. and Ney, P. E. (2004). *Branching Processes.* Dover, Mineola, NY.
[4] Baur, E. (2016). Percolation on random recursive trees. *Random Structures Algorithms* **48,** 655–680.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press

[5] Baur, E. and Bertoin, J. (2015). The fragmentation process of an infinite recursive tree and Ornstein–Uhlenbeck type processes. *Electron. J. Prob.* **20,** 98.
[6] Bertoin, J. (2014). On the non-Gaussian fluctuations of the giant cluster for percolation on random recursive trees. *Electron. J. Prob.* **19,** 24.
[7] Bertoin, J. (2014). Sizes of the largest clusters for supercritical percolation on random recursive trees. *Random Structures Algorithms* **44,** 29–44.
[8] Bertoin, J. and Uribe Bravo, G. (2015). Supercritical percolation on large scale-free random trees. *Ann. Appl. Prob.* **25,** 81–103.
[9] Berzunza, G. (2015). Yule processes with rare mutation and their applications to percolation on b-ary trees. *Electron. J. Prob.* **20,** 43.
[10] Billingsley, P. (1999). *Convergence of Probability Measures,* 2nd edn. John Wiley, New York.
[11] Bollobás, B. (2001). *Random Graphs,* 2nd edn. Cambridge University Press.
[12] Dobrow, R. P. and Smythe R. T. (1996). Poisson approximations for functionals of random trees. *Random Structures Algorithms* **9,** 79–92.
[13] Drmota, M., Iksanov, A., Moehle, M. and Roesler, U. (2009). A limiting distribution for the number of cuts needed to isolate the root of a random recursive tree. *Random Structures Algorithms* **34,** 319–336.
[14] Goldschmidt, C. and Martin, J. B. (2005). Random recursive trees and the Bolthausen–Sznitman coalescent. *Electron. J. Prob.* **10,** 718–745.
[15] Iksanov, A. and Möhle, M. (2007). A probabilistic proof of a weak limit law for the number of cuts needed to isolate the root of a random recursive tree. *Electron. Commun. Prob.* **12,** 28–35.
[16] Klebaner, F. C. (2005). *Introduction to Stochastic Calculus with Applications,* 2nd edn. Imperial College Press, London.
[17] Kuba, M. and Panholzer, A. (2014). Multiple isolation of nodes in recursive trees. *Online J. Analysis Comb.* **9**.
[18] Le Cam, L. (1960). An approximation theorem for the Poisson binomial distribution. *Pacific J. Math.* **10,** 1181–1197.
[19] Meir, A. and Moon, J. W. (1974). Cutting down recursive trees. *Math. Biosci.* **21,** 173–181.
[20] Möhle, M. (2015). The Mittag–Leffler process and a scaling limit for the block counting process of the Bolthausen–Sznitman coalescent. *ALEA Latin Amer. J. Prob. Math. Statist.* **12,** 35–53.

https://doi.org/10.1017/apr.2017.25 Published online by Cambridge University Press