J. Appl. Prob. **50,** 741–759 (2013)
Printed in England

THE ANCESTRAL PROCESS OF LONG-RANGE SEED BANK MODELS

JOCHEN BLATH,* *TU Berlin*
ADRIÁN GONZÁLEZ CASANOVA,** *Berlin Mathematical School*
NOEMI KURT,* *** *TU Berlin*
DARIO SPANÒ,**** *University of Warwick*

Abstract

We present a new model for seed banks, where direct ancestors of individuals may have lived in the near as well as the very far past. The classical Wright–Fisher model, as well as a seed bank model with bounded age distribution considered in Kaj, Krone and Lascoux (2001) are special cases of our model. We discern three parameter regimes of the seed bank age distribution, which lead to substantially different behaviour in terms of genetic variability, in particular with respect to fixation of types and time to the most recent common ancestor. We prove that, for age distributions with finite mean, the ancestral process converges to a time-changed Kingman coalescent, while in the case of infinite mean, ancestral lineages might not merge at all with positive probability. Furthermore, we present a construction of the forward-in-time process in equilibrium. The mathematical methods are based on renewal theory, the urn process introduced in Kaj, Krone and Lascoux (2001) as well as on a paper by Hammond and Sheffield (2013).

Keywords: Wright–Fisher model; seed bank; renewal process; long-range interaction; Kingman coalescent

2010 Mathematics Subject Classification: Primary 92D15
Secondary 60K05

1. Introduction

In this paper we discuss a new mathematical model for the description of the genetic variability of neutral haploid populations of fixed size under the influence of a general *seed bank* effect. In contrast to previous models, such as the Kaj *et al.* [6] model, we are particularly interested in situations where ancestors of individuals of the present generation may have lived in the rather remote past.

Seed banks are of significant evolutionary importance, and come in various guises. Typical situations range from plant seeds which fall dormant for several generations during unfavourable ecological circumstances [11], [12], fruit tissue preserved in Siberian permafrost [13], to bacteria turning into *endospores* if the concentration of nutrients in the environment falls below a certain threshold. Such endospores may in principle persist for an unlimited amount of time before they become active again (see, e.g. [2]). Seed bank related effects can be viewed as sources of genetic novelty [7] and are generally believed to increase observed genetic variability.

Received 5 April 2012; revision received 22 October 2012.
* Postal address: Institut für Mathematik, TU Berlin, Sekr. MA 7-5, Strasse des 17. Juni 136, D-10623 Berlin, Germany.
** Current address: TU Berlin, Strasse des 17. Juni 136, D-10623 Berlin, Germany.
*** Email address: kurt@math.tu-berlin.de
**** Postal address: Department of Statistics, University of Warwick, Coventry CV4 7AL, UK.

In [6], a mathematical model for a (weak) seed bank effect is investigated, with the number of generations backwards in time that may influence the current population being bounded by a constant m and being small when compared to the total population size (respectively during passage to a scaling limit). Under such circumstances, it is then shown that the ancestral process of the population can be approximately described by a time-changed Kingman coalescent, where the (constant) time change leads to a linear decrease of the coalescence rates of ancestral lineages depending on the square of the expected seed bank age distribution. Overall, genetic variability is thus increased (in particular if mutation is taken into account), but the qualitative features of the ancestral history of the population remain unchanged.

In the present paper we consider the ancestral process of a neutral seed bank model with Wright–Fisher-type dynamics, assuming a constant population size N. However, the distance measured in generations between the direct ancestor and potential offspring will not be assumed to be bounded, but rather sampled according to some (potentially unbounded) age distribution μ on $\mathbb{N}$. For $\mu = \delta_1$, we recover the ancestral process of the classical Wright–Fisher model, and scaling by the population size yields a Kingman coalescent as the limiting ancestral process. For μ with bounded support, say with a maximum value m, independent of N, we are in the setup of [6], and obtain a time change of Kingman's coalescent appearing in the limit (again after classical scaling).

Yet, some species suggest (i.e. bacteria transforming into endospores) that μ could be effectively unbounded, in particular nonnegligible when compared to the population size. This can lead to entirely different regimes.

Our first result is that if μ has finite expectation, we again obtain a time-changed Kingman's coalescent after classical rescaling. The behaviour of the model however changes completely if we assume μ to have infinite expectation. A natural example for age distributions is a discrete measure μ with a power-law decay, that is,

$$\mu(\{n, n+1, \ldots\}) = n^{-\alpha} L(n)$$

for some $\alpha > 0$ and some slowly varying function L. Depending on the choice of α, we investigate the time to the most recent common ancestor (MRCA) of two individuals, if it exists. It turns out (Theorem 2) that, for $\alpha > \frac{1}{2}$, there is always a common ancestor, but the expected time to the MRCA is finite if $\alpha > 1$ and infinite if $\alpha < 1$. If $\alpha < \frac{1}{2}$, any two ancestral lineages never meet at all with positive probability.

In the following section we construct our model and present the main results. The proofs are given in Section 3.

2. Model and main results

We work in discrete time (measured in units of nonoverlapping generations) and with fixed finite population size $N \in \mathbb{N}$. Time in generations is indexed by $\mathbb{Z}$. The dynamics of the population forwards in time are given in the following way. Each individual chooses the generation of its father according to a law μ on $\mathbb{N}$, meaning that $\mu(n)$ gives the probability that the immediate ancestor of an individual of generation i has lived in generation $i - n$. We call μ the seed bank age distribution. To avoid technicalities, we will always assume that $\mu(\{1\}) > 0$. After having chosen the generation, the individual picks the father uniformly among the N possible ancestors from that generation.

For concreteness, we will often assume that the age distribution μ is of the form $\mu = \mu_\alpha$, with

$$\mu_\alpha(\{n, n+1, \ldots\}) = n^{-\alpha} L(n), \qquad n \in \mathbb{N},$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

for some $\alpha \in (0, \infty)$ and some slowly varying function L. Let $\Gamma_\alpha := \{\mu_\alpha\}$, $\alpha \in (0, \infty)$, denote the set of all measures μ of this form. We are interested in the question of whether or not in such a population a genetic type eventually fixates, and if this happens in finite time almost surely. In the backward picture, this is related to asking if a finite set of individuals has an MRCA and when it lived.

It turns out that in the above construction an ancestral line can be described by a renewal process with interarrival law μ. The question of existence of a common ancestor and the time to the MRCA can therefore be investigated via classical results of Lindvall [8] on coupling times of discrete renewal processes, which are controlled in the power-law case via applications of Karamata's Tauberian theorem for power series; see, e.g. [1]. This leads to three different regimes; see Theorem 2. If, on the other hand, one is interested in the forward-in-time process, mathematical modelling problems arise. In order to obtain a new generation of such a population, one requires information about the whole history, i.e. it is necessary to start sampling at '$-\infty$'. In Subsection 2.2 we present a construction of such a population *in equilibrium*, which allows us to study the correlations of the allele frequency process. This construction can be formalized in terms of Gibbs measures, following the paper of Hammond and Sheffield [4], where the case $N = 1$ is considered in order to construct a discrete process with long-range correlations that converges to fractional Brownian motion. This is sketched in Appendix A.

2.1. Renewal construction of ancestral lineages and the time to the MRCA

We start with a description of the ancestral lineages of samples in our model in terms of renewal theory. Fix $N \in \mathbb{N}$ and a probability measure μ on the natural numbers. Let $v \in V_N := \mathbb{Z} \times \{1, \dots, N\}$ denote an individual of our population. For $v \in V_N$, we write $v = (i_v, k_v)$ with $i_v \in \mathbb{Z}$, and $1 \le k_v \le N$; hence, i_v indicates the generation of the individual in $\mathbb{Z}$, and k_v the label among the N individuals alive in this generation.

The ancestral line $A(v) = \{v_0 = v, v_1, v_2, \dots\}$ of our individual v is a set of sites in V_N, where $i_{v_0}, i_{v_1}, \dots \downarrow -\infty$ is a strictly decreasing sequence of generations, with independent decrements $i_{v_l} - i_{v_{l-1}} =: \eta_l$, $l \ge 1$, with distribution μ, and where the $k_{v_0}, k_{v_1}, \dots$ are independent and identically distributed Laplace random variables with values in $\{1, \dots, N\}$, independent of $\{i_{v_l}\}_{l \in \mathbb{N}_0}$. Letting

$$S_n := \sum_{l=0}^{n} \eta_l,$$

where we assume that $S_0 = \eta_0 = 0$, we obtain a discrete renewal process with interarrival law μ. In the language of [9], we say that a renewal takes place at each of the times S_n, $n \ge 0$, and we write $(q_n)_{n \in \mathbb{N}_0}$ for the renewal sequence, that is, q_n is the probability that n is a renewal time.

It is now straightforward to give a formal construction of the full ancestral process starting from N individuals at time 0 in terms of a family of N independent renewal processes with interarrival law μ and a sequence of independent uniform random variables $U^r(i)$, $i \in -\mathbb{N}$, $r \in \{1, \dots, N\}$, with values in $\{1, \dots, N\}$ (independent also of the renewal processes). Indeed, let the ancestral processes pick previous generations according to their respective renewal times, and then among the generations pick labels according to their respective uniform random variables. As soon as at least two ancestral lineages hit a joint ancestor, their renewal processes couple, i.e. follow the same realization of one of their driving renewal processes (chosen arbitrarily, and discarding those remaining parts of the renewal processes and renewal times which are no longer needed). In other words, their ancestral lines merge.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Denote by $\mathbb{P}_N^\mu$ the law of the above ancestral process. For $v \in V_N$ with $i_v = 0$, we have

$$q_n = \mathbb{P}_N^\mu(A(v) \cap (\{-n\} \times \{1, \dots, N\}) \neq \varnothing),$$

and the probability that $w \in V_N$ is an ancestor of v, for $i_w < i_v$, is given by

$$\mathbb{P}_N^\mu(w \in A(v)) = \frac{1}{N} q_{i_v - i_w}.$$

For notational convenience, let us extend q_n to $n \in \mathbb{Z}$ by setting $q_n = 0$ if $n < 0$. Note that $q_0 = 1$.

In [6] it was proved that if μ has finite support then the ancestral process, rescaled by the population size, converges to a time-changed Kingman coalescent. Our first result shows that this remains true with the same classical scaling for μ with infinite support, as long as it has finite expectation. We consider the ancestral process of a sample of $n \leq N$ individuals labelled $v_1, \dots, v_n$ sampled from generation $k = 0$. We define the equivalence relation '$\sim_k$' on the set $\{1, \dots, n\}$ by

$$i \sim_k j \quad \Longleftrightarrow \quad A(v_i) \cap A(v_j) \cap (\{-k, \dots, 0\} \times \{1, \dots, N\}) \neq \varnothing,$$

that is, $i \sim_k j$ if and only if v_i and v_j have a common ancestor at most k generations back. Let $A_{N,n}(k)$ denote the set of equivalence classes with respect to '$\sim_k$', which is a stochastic process taking values in the partitions of $\{1, \dots, n\}$. Let $E := \{1, \dots, n\}$, and let $D_E[0, \infty)$ denote the space of càdlàg functions from $[0, \infty)$ to E with the Skorokhod topology.

Theorem 1. *Assume that* $\mathbb{E}_\mu[\eta_1] < \infty$. *Let* $\beta := 1/\mathbb{E}_\mu[\eta_1]$. *As* $N \to \infty$, *the process* $(A_{N,n}(\lfloor Nt/\beta^2 \rfloor))_{t \geq 0}$ *converges weakly in* $D_E[0, \infty)$ *to Kingman's n-coalescent.*

Two individuals $v, w \in V_N$ have a common ancestor if and only if $A(v) \cap A(w) \neq \varnothing$. If this is the case, and if v and w belong to the same generation, we denote by τ the time to the MRCA:

$$\tau := \inf\{n \geq 0 \colon A(v) \cap A(w) \cap (\{-n\} \times \{1, \dots, N\}) \neq \varnothing\}.$$

Clearly, the law of τ is the same for all v, w with $i_v = i_w$.

Theorem 1 implies that if μ has finite expectation, two randomly sampled individuals have a common ancestor with probability 1, and the expected time to this ancestor is of order N. If the expectation does not exist, this changes completely. Let us now assume that $\mu \in \Gamma_\alpha$, which means that the tails of μ follow a power law. Our second result distinguishes three regimes.

Theorem 2. (Existence and expectation of the time to the MRCA.) *Let* $\mu \in \Gamma_\alpha$, *and let* $v, w \in V_N$, $v \neq w$.

(a) *If* $\alpha \in (0, \frac{1}{2})$ *then* $\mathbb{P}_N^\mu(A(v) \cap A(w) \neq \varnothing) < 1$ *for all* $N \in \mathbb{N}$.

(b) *If* $\alpha \in (\frac{1}{2}, 1)$ *then* $\mathbb{P}_N^\mu(A(v) \cap A(w) \neq \varnothing) = 1$ *and* $\mathbb{E}_N^\mu[\tau] = \infty$ *for all* $N \in \mathbb{N}$.

(c) *If* $\alpha > 1$ *then* $\mathbb{P}_N^\mu(A(v) \cap A(w) \neq \varnothing) = 1$ *for all* $N \in \mathbb{N}$, *and* $\lim_{N \to \infty} \mathbb{E}_N^\mu[\tau]/N = 1/\beta^2$, *with* $\beta = 1/\mathbb{E}_\mu[\eta_1]$.

In other words, for $\alpha > \frac{1}{2}$, two individuals almost surely share a common ancestor, but the expected time to the MRCA is finite for $\alpha > 1$ and infinite if $\alpha \in (\frac{1}{2}, 1)$. Hence, in real-world populations observed over realistic time scales, for $\alpha \in (\frac{1}{2}, 1)$ (or even for $\alpha \in (1, 2)$ where the mean, but not the variance, of μ exists), the assumption that a population is in equilibrium has to be treated with care.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Remark 1. In the boundary case $\alpha = 1$, the choice of the slowly varying function L becomes relevant. If we choose $L =$ constant then it is easy to see from the proof that $\mathbb{E}_N^\mu[\tau] = \infty$. The case $\alpha = \frac{1}{2}$ also depends on L and requires further investigation.

2.2. Forward-in-time process

Having obtained a good idea about the ancestral process, we would now like to study the forward picture. For this, it is useful to construct the whole bi-infinite genealogy of the whole population at once, which can be done as a spanning forest of a suitable vertex set. We consider graphs—in fact, trees—with vertex set $V_N = \mathbb{Z}^N$ and a set of bonds E_N which will be a (random) subset of $B_N := \{(v, w) : v, w \in V_N\}$ where the edges are *directed*. For $v \in V_N$, we write, as before, $v = (i_v, k_v)$ with $i_v \in \mathbb{Z}$ and $1 \le k_v \le N$. We consider the set of directed spanning forests of V_N, which we can write down as follows. Let

$$\mathcal{T}_N := \{G = (V_N, E_N) : E_N \subset B_N \text{ such that, for all } v \in V_N, \\ \text{there exists } w \in V_N,\ i_w < i_v,\ \text{with } e = (w, v) \in E_N\}.$$

This means that we consider trees where each vertex v has exactly one outgoing (to the past) edge, which we denote by e_v. This unique outgoing edge, or, equivalently, the unique ancestor of v, is determined as follows. Let $\{\eta_v\}_{v \in V_N}$ be a countable family of independent μ-distributed random variables, and let $\{U_v\}_{v \in V_N}$ denote independent uniform random variables with values in $\{1, \dots, N\}$ independent of the η_v. This infinite product measure induces a law on $\mathcal{T}_N$ if we define

$$e_v := ((i_v - \eta_v, U_v), v).$$

We denote this probability measure by $\hat{\mathbb{P}}_N^\mu$. In words, the ancestor of v is found by sampling the generation according to μ, and then choosing the individual uniformly. We see that

$$\hat{\mathbb{P}}_N^\mu(e_v = (w, v) \in E_N) = \frac{1}{N}\mu(i_v - i_w).$$

Comparing this to our previous construction of the ancestral process, we realise that $\mathbb{P}_N^\mu$ can be considered as being the restriction of $\hat{\mathbb{P}}_N^\mu$ to situations regarding the ancestry of a sample, and, hence, with a slight abuse of notation, we will identify the two measures, dropping the notation $\hat{\mathbb{P}}_N^\mu$. A tree $G \in \mathcal{T}_N$ is interpreted as the ancestral tree of the whole bi-infinite population.

Remark 2. Note that, for $\mu \in \Gamma_\alpha$, it follows from Theorem 2 that $G \in \mathcal{T}_N$ has only one connected component almost surely if $\alpha > \frac{1}{2}$, since two individuals belong to the same connected component if and only if their ancestral lines meet. If $\alpha < \frac{1}{2}$ then G has infinitely many connected components almost surely, since in that case any two individuals belong to two disjoint components with positive probability by Theorem 2(a).

Having obtained a construction of the genealogy of the population for all times, we can now, for example, introduce genetic types. We take the simplest situation of just two types. Let individual $v \in V_N$ have type $X_v \in \{a, A\}$, and assume a neutral Wright–Fisher reproduction, that is, types are inherited from the parent. This means that in the above construction, individuals belonging to the same component of the tree have the same type. In particular, in the case $\alpha > \frac{1}{2}$ everyone in the population has the same type. This is clear since constructing the whole tree at once means that we are talking about a population in equilibrium, meaning that fixation of one of the two types has already occurred. However, in the case $\alpha < \frac{1}{2}$ the tree has infinitely many

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

components almost surely, and, therefore, both types can persist for all times. We can assign to each component independently type a with probability $p \in [0, 1]$ and type A otherwise. For each $p \in [0, 1]$, this procedure defines a probability measure on $\{a, A\}^{V_N}$.

Definition 1. Let λ_N^p denote the probability measure on $\{a, A\}^{V_N}$ which, given $G \in \mathcal{T}_N$, assigns each connected component of G independently type a with probability p, and type A otherwise.

Remark 3. It can be shown, following [4], that in a certain sense the measures λ_N^p are the only relevant probability measures on $\{a, A\}^{V_N}$ consistent with the dynamics of our population model. We make this precise in Appendix A. For now, we just assume that the type distribution of our population is given by λ_N^p.

We can now introduce the frequency process. Let $v := (i, k) \in V_N$, that is, i denotes the generation of the individual, and k its label among the N individuals of generation i. Let

$$Y_N(i) := \frac{1}{N} \sum_{k=1}^{N} \mathbf{1}_{\{X_{i,k}=a\}}.$$

Our construction allows us to easily compute some correlations for the frequency process of the seed bank model. Recall q_n from the last section.

Theorem 3. *Let* $\lambda = \lambda_N^p$.

(a) $\mathbb{E}_\lambda[Y_N(i)] = p$ *for all* $i \in \mathbb{Z}$.

(b) *If* $\mu \in \Gamma_\alpha$ *with* $\alpha > \frac{1}{2}$, $\operatorname{cov}_\lambda(Y_N(0), Y_N(i)) = p(1-p)$ *for all* $i \in \mathbb{Z}$ *and all* $N \in \mathbb{N}$.

(c) *If* $\mu \in \Gamma_\alpha$ *with* $\alpha \in (0, \frac{1}{2})$, *we have* $\lim_{N\to\infty} \operatorname{cov}_\lambda(Y_N(0), Y_N(i)) = 0$,

$$C(i) := \lim_{N\to\infty} \operatorname{corr}_\lambda(Y_N(0), Y_N(i)) \in (0, 1) \quad \textit{for all } i \in \mathbb{Z},$$

and, as $i \to \infty$, *for some constant* c *and some slowly varying function* L,

$$C(i) \sim \frac{(1-\alpha)^2 p(1-p)}{\Gamma(2-\alpha)^2 \Gamma(2\alpha)(\sum_{n=0}^{\infty} q_n^2 + 1)} i^{2\alpha-1} L(i),$$

where '$\sim$*' means that the ratio of the two sides tends to* 1, *and the sum occurring in the denominator is finite.*

Remark 4. If $\alpha > \frac{1}{2}$, we have $\operatorname{corr}_\lambda(Y_N(0), Y_N(i)) = 1$. This is clear since in this case all individuals have the same type, and $\mathbb{E}_\lambda[Y_N(i)] = p$, $\operatorname{var}_\lambda(Y_N(i)) = p(1-p)$, and $\operatorname{corr}_\lambda(Y_N(0), Y_N(i)) = 1$.

3. Proofs

3.1. Proof of Theorem 1

The proof of Theorem 1 follows ideas from [6], which we combine with a coupling argument relying on renewal theory. In certain steps we have to take particular care of the unboundedness of the support of the measure μ; these steps are carried out with particular care in Lemmas 1 and 2. Recall that in Theorem 1 we assumed that the expectation of the renewal process exists, i.e. $\mathbb{E}_\mu[\eta_1] < \infty$, which in the case $\mu \in \Gamma_\alpha$ holds for $\alpha > 1$. For the case $\alpha = 1$, finiteness of the expectation depends on the choice of the slowly varying function L.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

We first introduce an 'urn process' similar to that introduced in [6] for measures μ with potentially unbounded support. The point is that our ancestral process A_N can then be realised as a simple function of this urn process.

Keep N fixed. For $1 \le n \le N$, let

$$S_n := \left\{(x_1, x_2, \ldots),\ x_i \in \mathbb{N}_0,\ \sum_{i=1}^{\infty} x_i = n\right\}.$$

For $n \in \mathbb{N}$, we construct a discrete-time Markov chain $\{X^n(k)\}_{k\in\mathbb{N}_0}$ with values in S_n that we will refer to as the n-sample process. Let $X^n(0) = (X_1^n(0), X_2^n(0), \ldots)$ be such that $|X^n(0)| = n$. We think of $X_i^n(0) \in \{0, \ldots, n\}$ as the number of balls currently placed in urn number i. Later, urns will correspond to generations, balls to individuals. The transition from time k to time $k+1$ is made by relocating the $X_1^n(k)$ balls in the first urn in a way that is consistent with the ancestral process of our seed bank model, and shift the other urns including their contained balls one step to the left. Let $\sigma : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}} : (x_1, x_2, \ldots) \mapsto (x_2, x_3, \ldots)$ denote the one-step shift operator, and, for $l \in \mathbb{N}$, let $R(l)$ be an S_l-valued random variable which is multinomially distributed with infinitely many parameters:

$$R(l) \sim \mathrm{Mult}(l; \mu(1), \mu(2), \ldots);$$

that is, $R(i)$ is a random vector of infinite length, and $R_i(l)$ counts the number of outcomes that take value i in l independent trials distributed according to μ. Define

$$X^n(k+1) = \sigma(X^n(k)) + R(X_1^n(k)), \qquad k = 0, 1, \ldots.$$

By definition, $X^n = \{X^n(k)\}_{k\in\mathbb{N}_0}$ is a Markov chain with a (countably infinite) state space S_n (see Figure 1). It provides a construction of n independent renewal processes with interarrival law μ, if one keeps track of the balls. For our purpose, it suffices to note that $X_1^n(k)$ gives, for each k, the number of renewal processes that have a renewal after k steps, which is equal in law to the number of original individuals in our seed bank model that have an ancestor in generation $-k$. Now recall our ancestral process $\{A_{N,n}(k)\}$ from Section 2, which was constructed using coalescing renewal processes. In terms of the X^n-process it can be described

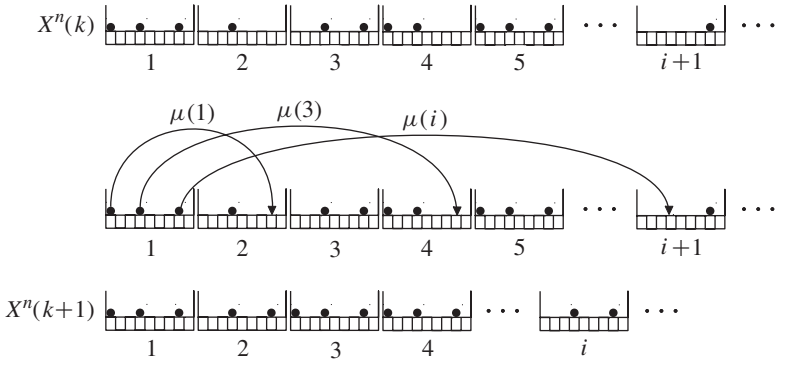

FIGURE 1: Transition from $X^n(k)$ (*top*) to $X^n(k+1)$ (*bottom*). All the balls in urn number 1 are relocated independently according to μ.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

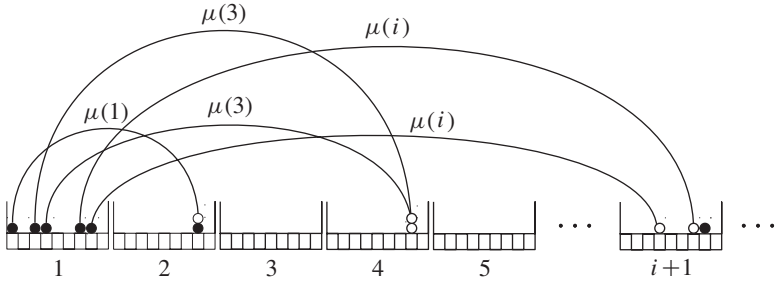

FIGURE 2: The possible types of coalescence events in the $X^{N,n}$-process: a coalescent event in urn 2 induced by a ball landing in an occupied place, a coalescent event in urn 4 due to two balls landing in the same empty place, and no coalescence in urn $i+1$ although it holds several balls.

as follows. Think for the moment of each of the urns as being subdivided into N sections. We start with n balls and run the X^n-process. At each relocation step, each ball which is relocated to urn $i+1$ is put with equal probability into one of the N sections in urn $i+1$. All balls that end up in the same section within an urn are merged into a single ball (see Figure 2).

Since this results in a decrease in the total number of balls, say from n to $n' < n$, after a merger event, we continue to run according to a Markov process with law $\mathcal{L}(X^{n'})$ with n' balls, and so on. Denote by $\{X^{N,n}(k)\}_{k\in\mathbb{N}}$ the well-defined process obtained by this procedure. Define by $Z^{N,n}$ the process obtained from $X^{N,n}$ by not merging the balls the moment they fall into the same urn and same section, but keeping them both at first, and only merging them at the moment they again reach urn 1, which happens automatically since we shift the configuration at each step. The number of balls present at time k in this process is equal in law to the block-counting process of our ancestral process started with n sampled individuals:

$$|Z^{N,n}(k)| \stackrel{\mathrm{D}}{=} |A_{N,n}(k)|.$$

Unlike A_N, the process $X^{N,n} = \{X^{N,n}(k)\}_{k\in\mathbb{N}}$ is a Markov chain in discrete time with countable state space $\bigcup_{i=1}^{n} S_i$. Of course, it is also possible to define an exchangeable partition-valued process as a function of $X^{N,n}$, where balls correspond to blocks (we refrain from a formal definition in order to keep the notational effort reasonable).

An important step is to observe that, for each n, the corresponding urn process X^n has a unique invariant distribution. Indeed, let

$$\beta_i := \frac{\mu\{i, i+1, \ldots\}}{\mathbb{E}_\mu[\eta]}.$$

This fraction is well defined since we assumed that $\mathbb{E}_\mu[\eta] < \infty$. Denote by $\nu^n := \mathrm{Mult}(n, \beta_1, \beta_2, \ldots)$ the multinomial distribution with success probabilities β_i. We claim that this is the stationary distribution for the n-sample process X^n. From classical renewal theory, we know that ν^1 is the stationary distribution in the case $n = 1$ (see [9]). For n independent renewal processes, we have the following result (cf. [6]).

Lemma 1. *If $\mathbb{E}_\mu[\eta] < \infty$ then ν^n is the stationary distribution for X^n, and X^n is positive recurrent for all $n \in \mathbb{N}$.*

Proof. We reduce the proof to the finite case discussed in [6]. For each $j \in \mathbb{N}$, we define

$$\mu_j(\{i\}) := \frac{1}{\sum_{l=1}^{j} \mu(\{l\})} \mathbf{1}_{\{i\le j\}} \mu(\{i\}), \qquad i \in \mathbb{N}.$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

This defines a probability measure μ_j with support $\{1, \dots, j\}$. Clearly, $\lim_{j\to\infty} \mu_j(i) = \mu(i)$ for all i, and $\lim_{j\to\infty} \mathbb{E}_{\mu_j}[\eta] = \mathbb{E}_{\mu}[\eta]$ by monotone convergence.

Let $Y^{n,j} = (Y^{n,j}(k))_{k\in\mathbb{N}_0}$ be the Markov chain constructed in the same way as X^n, but with relocation measure μ_j instead of μ, that is, $Y^{n,j}(k+1) = \sigma(Y^{n,j}(k)) + R^j(Y_1^{n,j}(k))$, where $R^j(l) \sim \text{Mult}(l; \mu_j(1), \dots, \mu_j(j))$, and with $Y^{n,j}(0) = X^n(0)$. Now define

$$\beta_i^j := \frac{\mu_j\{i, i+1, \dots\}}{\mathbb{E}_{\mu_j}[\eta]}.$$

Clearly, $\lim_{j\to\infty} \beta_i^j = \beta_i$ for all $i \in \mathbb{N}$. Let $\nu_j^n := \text{Mult}(n; \beta_1^j, \beta_2^j, \dots)$ denote the multinomial distributions on S_n with success probabilities β_i^j. By Lemma 1 of [6] we know that ν_j^n is the stationary distribution for $Y^{n,j}$. Fix $x, y \in S_n$. By construction,

$$\begin{aligned}\mathbb{P}(X^n(1) = y \mid X^n(0) = x) &= \mathbb{P}(R(x_1) = y - \sigma(x)) \\ &= \lim_{j\to\infty} \mathbb{P}(R^j(x_1) = y - \sigma(x)) \\ &= \lim_{j\to\infty} \mathbb{P}(Y^{n,j}(1) = y \mid Y^{n,j}(0) = x).\end{aligned}$$

For $x \in S_n$, let $j_x := \max\{j : x_j \neq 0\}$. Note that $\mathbb{P}(X^n(1) = y \mid X^n(0) = x) = 0$ for all x such that $j_x > j_y + 1$. We write $\mathbb{P}_{\nu^n}$ for the distribution of $(X^n(k))_{k\in\mathbb{N}}$ with initial distribution ν^n. Then, for every $y \in S_n$,

$$\begin{aligned}\mathbb{P}_{\nu^n}(X^n(1) = y) &= \sum_{x\in S_n,\, j_x \le j_y+1} \nu^n(x)\mathbb{P}(X^n(1) = y \mid X^n(0) = x) \\ &= \lim_{j\to\infty} \sum_{x\in S_n,\, j_x \le j_y+1} \nu_j^n(x)\mathbb{P}(Y^n(1) = y \mid Y^{n,j}(0) = x) \\ &= \lim_{j\to\infty} \nu_j^n(y) \\ &= \nu^n(y).\end{aligned}$$

So $\text{Mult}(n; \beta_1, \beta_2, \dots)$ is a stationary distribution for X^n. By irreducibility, it is unique, and X^n is positive recurrent.

Recall the dynamics of the process $X^{N,n} = (X^{N,n}(k))_{k\in\mathbb{N}_0}$ from above. We first compute the probability of a coalescence given that we are in a fixed configuration. Define the events

$$B_{l,k} := \{\text{exactly } l \text{ mergers at time } k \text{ in } X^{N,n}\}$$

and

$$B_{\ge l,k} := \{\text{at least } l \text{ mergers at time } k \text{ in } X^{N,n}\}$$

for $1 \le l \le n$ and $k \in \mathbb{N}$.

Lemma 2. *Fix $N \in \mathbb{N}$, $n < N$, and μ such that $\mathbb{E}_\mu[\eta] < \infty$. With the notation of the last section,*

$$\mathbb{P}(B_{1,k+1} \mid X^{N,n}(k) = (x_1, x_2, \dots)) = \frac{1}{N}\sum_{i=1}^{\infty}\left(x_1 x_{i+1}\mu(i) + \binom{x_1}{2}\mu(i)^2\right) + O(N^{-2}),$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

and there exists $0 < c(n) < \infty$*, depending on* $X^{N,n}$ *only via* n*, such that*

$$\mathbb{P}(B_{\geq 2,k+1} \mid X^{N,n}(k) = (x_1, x_2, \ldots)) \leq \frac{c(n)}{N^2}.$$

Proof. We start by computing the probability of a coalescence in a fixed urn $i \in \mathbb{N}$ given $X^{N,n}(k) = (X_1^{N,n}(k), X_2^{N,n}(k), \ldots)$ and $R(X_1^{N,n}(k)) = (R_1(X_1^{N,n}(k)), R_2(X_1^{N,n}(k)), \ldots)$. The probability of having *exactly* one coalescence occurring in urn i (note that from k to $k+1$ we shift all urns by 1) is

$$\frac{1}{N} X_{i+1}^{N,n}(k) R_i(X_1^{N,n}(k)) + \frac{1}{N} \binom{R_i(X_1^{N,n}(k))}{2} - p(i),$$

where $p(i) = p(i, X^{N,n}(k), R(X_1^{n,N}(k)))$ is the probability that more than one coalescence happens in urn i. Here, the first term is the probability that we see at least one coalescence due to one of the relocated balls falling into an already occupied section of urn i, and the second term is the probability of seeing at least one coalescence due to two relocated balls falling into the same section of urn i. Observe that $p(i)$ is $O(N^{-2})$. More precisely, writing

$$M_i := X_{i+1}^{N,n}(k) R_i(X_1^{N,n}(k)) + \binom{R_i(X_1^{N,n}(k))}{2},$$

it is easy to see that, because each ball being moved to urn i has a probability of at most n/N to merge at all,

$$p(i) \leq \frac{n^4}{N^2},$$

and, therefore, since, *given* $X^{N,n}(k)$ and $R(X_1^{N,n}(k))$, there are at most n occupied urns,

$$\sum_{i=1}^{\infty} p(i) \leq \frac{n^5}{N^2}.$$

Furthermore, given $X^{N,n}(k)$ and $R(X_1^{N,n}(k))$, the probability of having at least two mergers at step $k+1$, which occur in two different urns i and j, is

$$\frac{1}{N^2} M_i M_j.$$

Moreover, for fixed $X^{N,n}(k)$ and $R(X_1^{N,n}(k))$, we have the trivial bound $\sum_{j=1}^{\infty} M_j \leq 2n^3$. This implies that

$$\frac{1}{N^2} \sum_{i=1}^{\infty} \sum_{\{j:\, j \neq i\}} M_i M_j \leq \frac{4n^6}{N^2}.$$

Thus, the probability of seeing exactly one coalescence in step $k+1$, *given* $X^{N,n}(k)$ and $R(X_1^{N,n}(k))$, is

$$\sum_{i=1}^{\infty} \left(\frac{1}{N} M_i - p(i) \right) - \frac{1}{N^2} \sum_{\substack{i,j=1 \\ j \neq i}}^{\infty} M_i M_j = \frac{1}{N} \sum_{i=1}^{\infty} M_i + O(N^{-2}).$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Computing $R(X_1^{N,n}(k))$ given $X^{N,n}(k)$ using the multinomial distribution, we obtain

$$\begin{aligned}
&\mathbb{P}(B_{1,k+1} \mid X^{N,n}(k) = x) \\
&\quad = \sum_{r \in S_n} \mathbb{P}(B_{1,k+1} \mid X^{N,n}(k) = x,\ R(x) = r)\mathbb{P}(R(x) = r \mid X^{N,n}(k) = x) \\
&\quad = \frac{1}{N} \sum_{r \in S_n} \left[\sum_{i=1}^{\infty} \left(x_{i+1} r_i + \binom{r_i}{2} \right) + O(N^{-2}) \right] \mathbb{P}(R(x) = r \mid X^{N,n}(k) = x) \\
&\quad = \frac{1}{N} \sum_{i=1}^{\infty} \left(x_{i+1} x_1 \mu(i) + \binom{x_1}{2} \mu(i)^2 \right) + O(N^{-2}),
\end{aligned}$$

where we have used the fact that

$$\sum_{r \in S_n} O(N^{-2}) \mathbb{P}(R(X_1^{N,n}) = r \mid X^{N,n}(k) = x) = O(N^{-2}),$$

since the $O(N^{-2})$ term is bounded uniformly in $r \in S_n$ by some $c(n)/N^2$, and we average with respect to a probability measure. This proves the first claim. We have seen that

$$\mathbb{P}(B_{\geq 2,k+1} \mid X^{N,n}(k), R(X_1^{N,n}(k))) = \sum_{i=1}^{\infty} p(i) + \frac{1}{N^2} \sum_{\substack{i,j=1 \\ j \neq i}}^{\infty} M_i M_j \leq \frac{c(n)}{N^2}.$$

This proves the second part.

We now have the ingredients to prove convergence to Kingman's coalescent.

Proof of Theorem 1. Fix $n \in \mathbb{N}$. We will first study the process started in the stationary distribution ν. Then we will extend the result to arbitrary initial distributions using an adaptation of Doeblin's coupling method. To prove convergence in the stationary case, we just need to prove that the intercoalescence times for binary mergers are distributed asymptotically exponential with rate $\beta_1^2 \binom{n}{2}$, and that multiple coalescences are negligible. Starting from the stationary distribution, the probability of seeing a coalescence in the next step given that we have currently n balls is obtained as in [6], using Lemma 2, i.e.

$$\begin{aligned}
\mathbb{P}(B_{1,k+1} \mid X^{N,n}(k) \sim \nu^n) &= \mathbb{E}_{\nu^n}[\mathbb{P}(B_{1,k+1} \mid X^{N,n}(k))] \\
&= \frac{\beta_1^2}{N} \binom{n}{2} \left(2 \sum_{i=1}^{\infty} \frac{\beta_{i+1}}{\beta_1} \mu(i) + \sum_{i=1}^{\infty} \mu(i)^2 \right) + O(N^{-2}) \\
&= \frac{\beta_1^2}{N} \binom{n}{2} + O(N^{-2}), \qquad (1)
\end{aligned}$$

where we have computed the expectations with respect to the multinomial distribution ν^n and used $2 \sum_{i=1}^{\infty} \beta_{i+1} \mu(i)/\beta_1 + \sum_{i=1}^{\infty} \mu(i)^2 = 1$.

We claim now that, for the process $Z^{N,n}$, which, as we recall, describes our ancestral process, we have

$$\lim_{N \to \infty} \mathbb{P}(\text{no coalescence in } Z^{N,n} \text{ before time } Nt) = \exp\left[-\beta_1^2 \binom{n}{2} t \right]. \qquad (2)$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Note that this does not follow immediately from (1), since after an unsuccessful attempt at coalescence, the process is slightly out of the stationary distribution. We see that (1) implies that while we are in the stationary distribution, the probability for a coalescence in the next step is $\beta_1^2 N^{-1}\binom{n}{2} + O(N^{-2})$. This is equivalent to saying that, for large N, we consider $\binom{n}{2}$ independent geometric clocks with parameter $1/N$, and each time a clock rings, a fixed pair of balls coalesces with probability β_1^2. These geometric clocks give exactly the number of jumps any pair of balls needs to perform before they jump into the same small section, but not necessarily the same urn. Therefore, they give the number of attempts to coalesce, which are only successful if both balls are in urn 1. If the process is slightly out of the stationary distribution, the probability of two given balls being in urn 1 is not exactly β_1^2 anymore. But, since $\mathbb{E}_\mu[\eta] < \infty$, by classical renewal theory (see, e.g. [9, Theorem 3.1]), for

$$\tilde{\nu}^n := \mathbb{P}(X^{N,n}(1) \mid X^{N,n}(0) \sim \nu^n, \text{ no coalescence in step 1}),$$

we obtain

$$\|\mathbb{P}(X^n(k) \in \cdot \mid X^n(0) \sim \tilde{\nu}^n) - \mathbb{P}(X^n(k) \in \cdot \mid X^n(0) \sim \nu^n)\|_{\mathrm{TV}} \to 0 \quad \text{as } k \to \infty,$$

which implies that, as $k \to \infty$,

$$\mathbb{P}_{\nu^n}(\text{two fixed balls in urn 1 at time } k \mid \text{no coalescence at time 1}) = \beta_1^2 + o(1).$$

Therefore, as $N \to \infty$, if $\tau(N)$ is a geometric random variable with parameter $\binom{n}{2}N^{-1}$ independent of the lengths of the jumps, we obtain

$$\mathbb{P}_{\nu^n}(\text{two fixed balls in urn 1 at time } \tau(N) \mid \text{no coalescence at time 1}) = \beta_1^2 + o(1).$$

By the memorylessness of the geometric distribution, and the fact that the choice of the section and the jump lengths of the urn process are independent, we obtain

$$\begin{aligned}\lim_{N\to\infty} \mathbb{P}_{\nu^n}&(\text{no coalescence in } Z^{N,n} \text{ before time } Nt) \\ &= \lim_{N\to\infty}\left(1 - \frac{\beta_1^2 + o(1)}{N}\binom{n}{2} + O(N^{-2})\right)^{Nt},\end{aligned}$$

which proves (2). Moreover, we have seen in Lemma 2 that multiple coalescences are negligible.

For the coupling argument, we now consider a process $\tilde{X}^{N,n}$ which runs as follows. Start with n balls in the stationary distribution ν^n, and let it evolve according to the n-sample dynamics. After each coalescence event, sample a new starting configuration according to $\nu^{n'}$, where n' is the number of balls present after the coalescence, and run the process according to the n'-sample dynamics. Assume now that $X^{N,n}$ starts in a given initial distribution. Define

$$T^{(N)} := \inf\{t > 0 \colon X^{N,n}(t) = \tilde{X}^{N,n}(t)\}.$$

We couple $X^{N,n}$ and $\tilde{X}^{N,n}$ as follows. Colour the balls of $X^{N,n}$ red and the balls of $\tilde{X}^{N,n}$ blue. Label both the red and the blue balls $1, \ldots, n$. Recall that the dynamics of our urn process just consist in moving balls from urn 1 independently from each other to a new urn according to μ, and merging balls in the same urn with probability N^{-1} per pair. Run the red and the blue processes independently. Let us first assume that no coalescences occur in either of the processes.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Now if at some time k the red ball number i and the blue ball number i happen to be in the same urn (but not necessarily in the same section), we couple them and let them move together from this time onwards. Denote by σ_i the time of this coupling. Note that σ_i is finite almost surely, since it is the coupling time of two renewal processes. Then we continue running our processes until all the balls have coupled. Let $T_{\text{coup}} := \max\{\sigma_i,\ 1 \le i \le n\}$. Note that this time is independent of N. Since n is fixed, and the different balls move independently, we have $\mathbb{P}(T_{\text{coup}} < \infty) = 1$ no matter which initial distributions we choose (see [9, Chapter II]), and, hence,

$$\lim_{t\to\infty} \mathbb{P}(T_{\text{coup}} \ge t) = 0.$$

Speeding up time by N, the coupling happens much faster than the coalescence. Let $T_{\text{coal}}^{(N)}$ be the time of the first coalescence in either the red or the blue process. At each time step, the probability of having a coalescence in the next step is bounded from above by the crude uniform estimate n^2/N. Hence,

$$\lim_{N\to\infty} \mathbb{P}(T_{\text{coal}}^{(N)} \ge \sqrt{N}) \ge \lim_{N\to\infty} \left(1 - \frac{n^2}{N}\right)^{\sqrt{N}} = 1.$$

Since

$$\lim_{N\to\infty} \mathbb{P}(T_{\text{coup}} \le \sqrt{N}) = 1,$$

we get

$$\lim_{N\to\infty} \mathbb{P}(T_{\text{coal}}^{(N)} \ge T_{\text{coup}}) \ge \lim_{N\to\infty} \mathbb{P}(T_{\text{coal}}^{(N)} \ge \sqrt{N},\ T_{\text{coup}} \le \sqrt{N}) = 1.$$

This implies that

$$\lim_{N\to\infty} \mathbb{P}(T^{(N)} \ne T_{\text{coup}}) = \lim_{N\to\infty} \mathbb{P}(T_{\text{coal}}^{(N)} < T_{\text{coup}}) = 0,$$

from which we see that

$$\lim_{N\to\infty} \mathbb{P}(T^{(N)} \ge Nt) = \lim_{N\to\infty} \mathbb{P}(T_{\text{coup}} \ge Nt) = 0.$$

Hence we can restart our process $\tilde{X}^{N,n}$ after each coalescence event, and the two processes will couple with probability 1 before the next coalescence takes place, and indeed on the coalescent time scale (time sped up by N) the coupling happens instantaneously. Using (2), for the intercoalescence times of the process started in an arbitrary but fixed initial configuration, we thus obtain

$$\lim_{N\to\infty} \mathbb{P}(\text{no coalescence in } Z^{N,n} \text{ before } Nt) = \exp\left[-\beta_1^2 \binom{n}{2} t\right]. \tag{3}$$

This implies as before by standard arguments (see [3]) that $|Z^{N,n}(Nt)|$ converges weakly as $N \to \infty$ to the block-counting process of Kingman's coalescent. Since $|Z^{N,n}(Nt)| \overset{\text{D}}{=} |A_{N,n}(Nt)|$, and the fact that we obviously have exchangeability of the ball configurations, we even obtain the convergence to Kingman's n-coalescent in the obvious sense. This completes the proof of Theorem 1.

Remark 5. It appears remarkable that $\mathbb{E}_\mu[\eta] < \infty$ is sufficient for this result. If $\mathbb{E}_\mu[\eta^2] = \infty$, and Y denotes the label of the urn that a ball is placed in, then $\mathbb{E}_{\nu^n}[Y] = \infty$ and, by [8], $\mathbb{E}[T_{\text{coup}}] = \infty$. However, due to the time rescaling, the fact that $\mathbb{P}(T_{\text{coup}} < \infty) = 1$ is enough for our purpose.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Remark 6. In order to show convergence to Kingman's coalescent, we could also follow the approach of [6], which uses Möhle's lemma [10] to show convergence of finite-dimensional distributions. Note however that in our case the state space of the Markov chain is infinite; hence, the transition matrices are infinite. Indeed, denoting the transition matrix of $X^{N,n}$ by $\Pi_N = \{\Pi_N(x, y)\}_{x,y\in\bigcup_{j=1}^{\infty} S_j}$, we can decompose Π_N as $\Pi_N = A + N^{-1}B + O(N^{-2})$, where A is given by the transitions of the X^n-processes without coalescence, and B contains adjustments that need to be made to the X^n-process in the case of a single coalescence event (compare [6]). The higher-order coalescences are $O(N^{-2})$ by Lemma 2. To apply Möhle's lemma, it is sufficient to show that $P := \lim_{m\to\infty} A^m$ and $G := PBP$ exist. We first take care of the part without coalescence. Let A be defined by $A(x, y) := \sum_{j=1}^{n} \mathbf{1}_{\{x,y\in S_n\}} A_n(x, y)$, where $(A_n(x, y))_{(x,y)\in S^n}$ denotes the transition matrix of X^n. Then Lemma 1 yields

$$\lim_{k\to\infty} A_n^k(x, y) = \nu^n(y)$$

for all $x, y \in S_n$. Thus, we obtain $\lim_{m\to\infty} A^m = P$, where $P = (P(x, y))_{x,y\in S}$ with $P(x, y) = \sum_{j=1}^{n} \mathbf{1}_{\{x,y\in S_j\}} \nu^j(y)$. We can now define B as the matrix of the single coalescence events as in [6]. That is, if $x \in S_i$ and $y \in S_{i-1}$, then $B(x, y)$ is the probability that the balls from configuration x are relocated according to the matrix A_i, and that exactly one pair of them coalesces, so that we end up with configuration y. If $x \in S_i$ then $B(x, y) = 0$ if $y \notin S_i \cup S_{i-1}$. If x and y are in S_i, then $B(x, y)$ gives the correction for the X^n-process in the case of a coalescence; therefore, $B(x, y) \geq -A(x, y)$ in this case. Hence, B has the same block form as in [6]; however, the single blocks are of infinite size. Furthermore, $\|B\| = \max_{x\in\bigcup_{i=1}^{n} S_i} \sum_y |B(x, y)| \leq 2$. Since P is a projection, $G = PBP$ is a bounded operator, and, therefore, e^{tG}, $t \in \mathbb{R}$, exists as a convergent series. Now the computations proceed exactly as in the case of bounded support; hence, we obtain the convergence to Kingman's coalescent following the proof of [6].

Remark 7. Note that Möhle's result allows the following heuristic interpretation of our limiting process $X^{N,n}$ as $N \to \infty$. First, the process, for each number of 'active' balls $n' \leq n$, mixes rapidly and essentially instantaneously enters its stationary distribution on the configuration with n' balls. Note that as long as there is no coalescence event, any future evolution does not affect either the block counting process $A_{N,n}$, or the corresponding partition-valued process, where each 'active' ball denotes a block in a partition of $\{1, \ldots, n\}$ consisting of all labels of balls that have merged into this active ball. Now, in each 'infinitesimal time step', our limiting process picks an entirely new state from its stationary distribution, independent of its 'previous' state (this is the effect of the projection operator P). In a way it can be regarded as a 'white noise' process on the space of stationary samples. While this process obviously has no càdlàg modification, both the block counting process and the partition-valued process remain constant until there is a new merger, and are thus well defined (recalling that such mergers, that is, transitions from n' active balls to $n' - 1$ active balls, happen at a finite positive rate in the limit).

3.2. Proof of Theorem 2

Recall from Section 2 that the time to the MRCA is related to the coupling time of two versions of the renewal process. Recall that

$$q_n = \mathbb{P}_N^{\mu}(A(v) \cap (\{-n\} \times \{1, \ldots, N\}) \neq \varnothing).$$

We will need some bounds on the q_n that can be obtained via Tauberian theorems.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Lemma 3. *Let* $\mu \in \Gamma_\alpha$.

(a) *Let* $\alpha \in (0, 1)$. *Then*

$$\sum_{n=0}^{i} q_n \sim \frac{1-\alpha}{\Gamma(2-\alpha)\Gamma(1+\alpha)} i^\alpha L(i)^{-1} \quad \text{as } i \to \infty.$$

(b) *The sum*

$$\sum_{n=0}^{\infty} q_n^2$$

is finite if $\alpha \in (0, \frac{1}{2})$ *and infinite if* $\alpha > \frac{1}{2}$.

(c) *Let* $\alpha \in (0, \frac{1}{2})$. *Then*

$$\sum_{n=0}^{\infty} q_n q_{n-i} \sim \frac{(1-\alpha)^2}{\Gamma(2-\alpha)^2\Gamma(2\alpha)} i^{2\alpha-1} L(i) \quad \text{as } i \to \infty.$$

Proof. The proof of this lemma can be found in [4, Lemma 5.1].

Proof of Theorem 2. We first prove part (c), which corresponds to the case where we have convergence to Kingman's coalescent. Without loss of generality, assume that $i_v = i_w = 0$. Denote by (R_n) and (R'_n) the sequences of renewal times of the renewal processes corresponding to v and w, respectively, that is, $R_n = \mathbf{1}_{\{n \in \{S_0, S_1, \dots\}\}}$. In other words, $R_n = 1$ if and only if v has an ancestor in generation $-n$, and $q_n = \mathbb{P}(R_n = 1)$. Let

$$T := \inf\{n \colon R_n = R'_n = 1\}$$

denote the coupling time of the two renewal processes. Since each time v and w have an ancestor in the same generation, these ancestors are the same with probability N, we get

$$\mathbb{E}[\tau] = N\mathbb{E}[T].$$

But, if $\alpha > 1$, we have $\mathbb{E}_\mu[\eta_1] < \infty$, and, therefore, by Proposition 2 of [8], $\mathbb{E}[T] < \infty$. The result now follows from Theorem 1 and the fact that the expected time to the MRCA of n individuals in Kingman's coalescent with time change β^2 is given by

$$\mathbb{E}[T_{\mathrm{MRCA}}] = \frac{1}{\beta^2} \sum_{k=2}^{n} \frac{1}{\binom{k}{2}} = \frac{2}{\beta^2}\left(1 - \frac{1}{n}\right);$$

hence, for $n = 2$, we get $1/\beta^2$.

(b) For independent samples R and R', the expected number of generations where both individuals have an ancestor is given by

$$\mathbb{E}\left[\sum_{n=0}^{\infty} R_n R'_n\right] = \sum_{n=0}^{\infty} \mathbb{E}[R_n]\mathbb{E}[R'_n] = \sum_{n=0}^{\infty} q_n^2,$$

which is infinite if $\alpha > \frac{1}{2}$ due to Lemma 3(b). Each of these times, the ancestors are the same with probability $1/N$; therefore, with probability 1, $A(v)$ and $A(w)$ eventually meet. However, the expected time until this event is bounded from below by the expectation of the step size:

$$\mathbb{E}^N_\mu[\tau] \geq \mathbb{E}[\eta] = \infty$$

if $\alpha < 1$.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

(a) In this case, $\mathbb{E}[\sum_{n=0}^{\infty} R_n R'_n] = \sum_{n=0}^{\infty} q_n^2 < \infty$, and, therefore,

$$\mathbb{P}\left(\sum_{n=0}^{\infty} R_n R'_n = \infty\right) = 0,$$

which implies that the probability that $A(v)$ and $A(w)$ never meet is positive.

3.3. Proof of Theorem 3

We prove now Theorem 3. We define $Y_v := \mathbf{1}_{\{X_v = a\}}$.

Lemma 4. *Let $\lambda = \lambda_N^p$, and assume that $\mu \in \Gamma_\alpha$.*

(a) *If $\alpha > \frac{1}{2}$,*

$$\mathrm{cov}_\lambda(Y_v, Y_w) = p(1-p).$$

(b) *If $\alpha \in (0, \frac{1}{2})$, $v \neq w$,*

$$\mathrm{cov}_\lambda(Y_v, Y_w) = p(1-p)\frac{\sum_{n=0}^{\infty} q_n q_{n+i_v-i_w}}{N + \sum_{n=1}^{\infty} q_n^2}.$$

Proof. We have

$$\mathbb{E}_\lambda^N[Y_v Y_w] = \lambda(X_v = X_w = a) = p\mathbb{P}_N^\mu(A(v)\cap A(w) \neq \varnothing) + p^2(1 - \mathbb{P}_N^\mu(A(v)\cap A(w) \neq \varnothing))$$

and $\mathbb{E}_\lambda^N[Y_v]\mathbb{E}_\lambda^N[Y_w] = p^2$. This implies that

$$\mathrm{cov}_\lambda(Y_v, Y_w) = p(1-p)\mathbb{P}_\mu^N(A(v) \cap A(w) \neq \varnothing).$$

If $\alpha > \frac{1}{2}$ then $\mathbb{P}_N^\mu(A(v) \cap A(w) \neq \varnothing) = 1$ which proves (a). Hence, we need to compute $\mathbb{P}_N^\mu(A(v) \cap A(w) \neq \varnothing)$ for $\alpha < \frac{1}{2}$. To do this, let S_n and S'_n denote two independent samples of the renewal process, with $S_0 = i_v$ and $S'_0 = i_w$. Note that this implies that, for the times of the renewals,

$$\mathbb{P}(R_n = 1) = q_{n+i_v}.$$

Recall that the renewal process is running forward in time, whence the ancestral lines are traced backwards. Let A_v and A_w denote two independent samples of the ancestral lines of v and w, using the processes S and S', respectively, without coupling the processes. Then the expected number of intersections of A_v and A_w is given by

$$\mathbb{E}[|A_v \cap A_w|] = \frac{1}{N}\mathbb{E}\left[\sum_{n=-i_w}^{\infty} R_n R'_n\right] = \frac{1}{N}\sum_{n=-i_w}^{\infty} q_{n+i_v} q_{n+i_w} = \frac{1}{N}\sum_{n=0}^{\infty} q_n q_{n+i_v-i_w}.$$

On the other hand, conditioning on the event that the ancestral lines meet (which clearly has positive probability), and then restarting the renewal processes in the generation of the first common ancestor, which is the same as sampling two ancestral lines starting at $(0, 0)$,

$$\begin{aligned}\mathbb{E}[|A_v \cap A_w|] &= \mathbb{E}[|A_v \cap A_w| \mid A_V \cap A_w \neq \varnothing]\mathbb{P}(A_v \cap A_w \neq \varnothing) \\ &= \mathbb{P}(A(v) \cap A(w) \neq \varnothing)\mathbb{E}[|A_{(0,0)} \cap A_{(0,0)}|] \\ &= \mathbb{P}(A(v) \cap A(w) \neq \varnothing)\left(q_0 + \frac{1}{N}\sum_{n=1}^{\infty} q_n^2\right).\end{aligned}$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Recalling that $q_0 = 1$, this implies that

$$\mathbb{P}_N^{\mu}(A(v) \cap A(w) \neq \varnothing) = \frac{\sum_{n=0}^{\infty} q_n q_{n+i_v-i_w}}{N + \sum_{n=1}^{\infty} q_n^2},$$

which proves the lemma.

Proof of Theorem 3. Part (a) is obvious and (b) follows from Lemma 4. For (c), let $\alpha \in (0, \frac{1}{2})$. Lemma 3 tells us that $\sum_{n=0}^{\infty} q_n^2 < \infty$. From Lemma 4, it follows that, for $i \neq 0$,

$$\operatorname{cov}_{\lambda}(Y_N(0), Y_N(i)) = p(1-p)\frac{\sum_{n=0}^{\infty} q_n q_{n-i}}{N + \sum_{n=1}^{\infty} q_n^2}.$$

For the variance, we obtain

$$\begin{aligned}\operatorname{var}_{\lambda}(Y_N(i)) &= \frac{1}{N^2}\sum_{k,j=1}^{N} \operatorname{cov}_{\lambda}(Y_{(i,k)}, Y_{(i,j)}) \\ &= \frac{1}{N^2}\left(Np(1-p) + N(N-1)p(1-p)\frac{\sum_{n=0}^{\infty} q_n^2}{N + \sum_{n=1}^{\infty} q_n^2}\right) \\ &= p(1-p)\frac{\sum_{n=0}^{\infty} q_n^2 + 1 - 1/N}{N + \sum_{n=1}^{\infty} q_n^2}.\end{aligned}$$

Hence,

$$\operatorname{corr}_{\lambda}(Y_N(0), Y_N(i)) = \frac{\sum_{n=0}^{\infty} q_n q_{n-i}}{\sum_{n=0}^{\infty} q_n^2 + 1 - 1/N},$$

which converges as $N \to \infty$. The result now follows from Lemma 3(c).

Appendix A. Gibbs measure characterization of the forward process

In Section 2.2 we claimed that the measures λ_N^p are in a certain sense the only measures describing the type distribution which are consistent with the dynamics of our process. In order to make this rigorous, we use a Gibbs measure characterization, which relies on the approach in [4]. In order to construct the Gibbs measure, we start with prescribing the distribution of types conditional on the (infinite) past. Let $S_N := \{a, A\}^N$ denote the finite-dimensional state space. Let $X_v = X_{(i_v,k_v)} \in \{a, A\}$ denote the type of individual v that is the kth individual of generation i. We denote by $\mathcal{C}$ the sigma-algebra of cylinder events, and write σ_n for the σ-algebra generated by cylinder sets contained in $\{\ldots, n\}$. For $i \in \mathbb{Z}$, we define the probability kernel $\lambda_{N,i}(\cdot \mid \cdot)$ from $(S_N^{\mathbb{Z}}, \sigma_i)$ to $(S_N^{\mathbb{Z}}, \mathcal{C})$ by saying that, for any finite set $B \subset \{i+1, \ldots\}^N$, $x_B \in \{a, A\}^B$, and $\xi \in S_N^{\{\ldots,i-1,i\}}$, the conditional probability

$$\lambda_{N,i}^{\xi}(X|_B = x_B) := \lambda_{N,i}(\{X|_B = x_B\} \mid \xi)$$

is obtained by first sampling $G \in \mathcal{T}_N$, tracing back the ancestral line of every $v \in B$ until it first hits $\{\ldots, i\}$, and then assigning the type $\xi_{\cdot}$ of this ancestor to v. This is well defined because, under $\mathbb{P}_N^{\mu}$, the tree until it first hits $\{\ldots, i\}$ is independent of σ_i. The kernels $\lambda_{N,i}^{\xi}$, $i \in \mathbb{Z}$, are then used to construct the Gibbs measures. Due to the construction via product measures, it is clear that they are consistent. If $i < j$ then, for $B \subset \{j+1, \ldots\} \times \{1, \ldots, N\}$,

$$\lambda_{N,i}^{\xi^1}(X_v = x_v,\ v \in B \mid X_w = \xi_w^2,\ i+1 \leq i_w \leq j) = \lambda_{N,j}^{\xi^1 \vee \xi^2}(X_v = x_v,\ v \in B).$$

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Here $\xi^1 \vee \xi^2$ denotes the configuration which is equal to ξ^1 on $\{\dots, i\}$ and equal to ξ^2 on $\{i+1, \dots, j\}$. So we can now define the Gibbs measures for our model.

Definition 2. A probability measure λ_N on $S_N^{\mathbb{Z}}$ is called a *μ-Gibbs measure* if, for all $i \in \mathbb{Z}$, all finite subsets $B \subset \{i+1, \dots\} \times \{1, \dots, N\}$, and all $x_B \in \{a, A\}^B$, the mapping $\xi \mapsto \lambda_{N,i}^{\xi}(x_B)$ is a version of the conditional probability

$$\lambda_N(X|_B = x_B \mid \sigma_i).$$

In other words, to sample from the Gibbs measure *conditional on the past* up to generation i, we first sample a $G \in \mathcal{T}_N$ according to $\mathbb{P}_N^{\mu}$, and assigning each X_v, $i_v \geq i+1$, $1 \leq k_v \leq N$, its type according to the ancestors. It is clear that such measures exist; in fact, λ_N^p defined in Section 2.2 is clearly a μ-Gibbs measure for $p \in \{0, 1\}$, and if $G \in \mathcal{T}_N$ has infinitely many components almost surely then, for all $p \in [0, 1]$, the measures λ_N^p are μ-Gibbs measures. Recall that this is the case if $\mu \in \Gamma_\alpha$ with $\alpha < \frac{1}{2}$. This is the situation where the Gibbs measure characterization is interesting.

A particularly useful feature of our model is that the only relevant Gibbs measures are of the form λ_N^p. Note that the μ-Gibbs measures form a convex set, as can be seen easily, and we can characterise the extremal points of this set generalizing Proposition 1 of [4].

Proposition 1. *Assume that $\mu \in \Gamma_\alpha$.*

(a) *Let $\alpha \in (0, \frac{1}{2})$. For each fixed N and each $p \in [0, 1]$, there is precisely one extremal μ-Gibbs measure λ_N on $S_N^{\mathbb{Z}}$ such that $\lambda_N(X_{i,k} = a) = p$ for all $i \in \mathbb{Z}$, $1 \leq k \leq N$.*

(b) *Let $\alpha \in (\frac{1}{2}, \infty]$. The only extremal Gibbs measures are λ_N^0 and λ_N^1. For $p \in (0, 1)$, the measures λ_N^p are given by $\lambda_N^p = p\lambda_N^0 + (1-p)\lambda_N^1$.*

A.1. Proof of Proposition 1

The proof of Proposition 1 follows closely that of Proposition 1 of [4], and we refer the reader to this work for details. Note that part (b) follows immediately from Theorem 2, as this implies that all individuals have the same type almost surely. The crucial step in the proof of part (a) of the proposition is the following lemma.

Lemma 5. *Let λ be a extremal μ-Gibbs measure. Then there exist $p \in [0, 1]$ such that, for all $v = (i_v, k_v) \in V_N$,*

$$\lim_{m \to \infty} \lambda(X_v = a \mid \sigma_{-m}) = p \quad \lambda\text{-almost surely.}$$

Proof. For fixed v, the existence of the limit follows from the backward martingale convergence theorem (see [5, p. 233]), and the fact that it is constant follows from the tail triviality of extremal Gibbs measures. It remains to prove that it is independent of v. For this, we couple the ancestral lines of two individuals v_1 and v_2 as in [4] in as far as their i-coordinate (the generations) is concerned, and concerning the k-coordinate, that is, the label of the individual among the N individuals per generation, we simply couple them completely, which does not change the law of the process. Hence, the proof of [4] goes through with only minor changes.

For the rest of the proof of Proposition 1, see [4]. The main idea is as follows. For any finite set of individuals, there exists a (random) time T before which the ancestral lines do not meet. This time is finite almost surely, and in view of Lemma 5, there exists $p \in [0, 1]$ such that the ancestors alive just after time T get their types independently with probability between $p - \varepsilon$ and $p + \varepsilon$. This then implies that $\lambda = \lambda_N^p$, which, as we recall, conditional on $G \in \mathcal{T}_N$, is induced by the product Bernoulli measure on the components of G with success parameter p.

https://doi.org/10.1239/jap/1378401233 Published online by Cambridge University Press

Acknowledgements

DS's research was partly supported by CRiSM, an EPSRC-HEFCE UK grant. The authors wish to thank an anonymous referee for making valuable suggestions which improved the presentation of the results considerably, and Julien Berestycki for discussions related to the coupling argument.

References

[1] Bingham, N. H., Goldie, C. M. and Teugels, J. L. (1987). *Regular Variation*. Cambridge Universtiy Press.
[2] Cano, R. J. and Borucki, M. K. (1995). Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. *Science* **268,** 1060–1064.
[3] Ethier, S. and Kurtz, T. G. (1986). *Markov Processes: Characterization and Convergence*. John Wiley, New York.
[4] Hammond, A. and Sheffield, S. (2013). Power law Pólya's urn and fractional Brownian motion. To appear in *Prob. Theory Relat. Fields*.
[5] Jacod, J. and Protter, P. (2003). *Probability Essentials*, 2nd edn. Springer, Berlin.
[6] Kaj, I., Krone, S. M. and Lascoux, M. (2001). Coalescent theory for seed bank models. *J. Appl. Prob.* **38,** 285–300.
[7] Levin, D. A. (1990). The seed bank as a source of genetic novelty in plants. *Amer. Naturalist* **135,** 563–572.
[8] Lindvall, T. (1979). On coupling of discrete renewal processes. *Z. Wahrscheinlichkeitsth.* **48,** 57–70.
[9] Lindvall, T. (1992). *Lectures on the Coupling Method*. John Wiley, New York.
[10] Möhle, M. (1998). A convergence theorem for Markov chains arising in population genetics and the coalescent with selfing. *Adv. Appl. Prob.* **30,** 493–512.
[11] Tellier, A. *et al.* (2011). Inference of seed bank parameters in two wild tomato species using ecological and genetic data. *Proc. Nat. Acad. Sci. USA* **108,** 17052–17057.
[12] Vitalis, R., Glémin, S. and Olivieri, I. (2004). When genes go to sleep: the population genetic consequences of seed dormancy and monocarpic perenniality. *Amer. Naturalist* **163,** 295–311.
[13] Yashina, S. *et al.* (2012). Regeneration of whole fertile plants from 30,000-y-old fruit tissue buried in Siberian permafrost. *Proc. Nat. Acad. Sci. USA* **109,** 4008–4013.