

# A Mean Field Model for Species Abundance

Thomas J. Pfaff<sup>a</sup>

<sup>a</sup>*Math and Comp. Sci. Dept., Ithaca College, 1212 Williams Hall, Ithaca NY  
14850-7284*

---

## Abstract

In this paper, we use the multitype mean field voter model as a model of species interaction, to obtain results about species abundance. Briefly, we start with the complete graph on  $n$  vertices,  $C_n$ , with each site occupied by a particle. Particles are represented by a value in  $(0, 1)$ , where distinct values represent different species. Particles then undergo mutation at rate  $\alpha$ , and are relabelled with a value chosen uniformly from  $(0, 1)$ . Particles also give birth at rate 1, and invade any of the other  $n$  sites randomly. This process has a unique stationary distribution denoted by  $\xi_\infty^n$ , which is given by the Ewens sampling formula. For each value in  $(0, 1)$  that is present in  $\xi_\infty^n$ , we count the number of particles represented by the same value, and call that the patch size of the species. Let  $K_n[a, b]$  denote the number of species with patch size in  $[a, b]$ . We study the limiting distribution of  $K_n[a, b]$ , for certain values of  $a$  and  $b$ , as the mutation rate  $\alpha$  tends to 0, which will in turn force  $n \rightarrow \infty$ .

*Key words:* Species abundance distribution, mean field voter model, Moran model, linear birth process

*1991 MSC:* 60K35, 92D25

---

## 1 Introduction

Bramson, Cox, and Durrett (1998), which we refer to as BCD, use the multitype voter model with mutation to model species abundance. This model is a dynamic model with a “fixed amount of some governing resource combining two previous approaches to modelling species abundance.” Previous models are described in the introduction of BCD. The fixed resource is represented by a grid and species may interact with their nearest neighbor. Here we use the multitype mean field voter model with mutation to model species abundance. The main difference between these two models is that in the mean field

---

*Email address:* [tpfaff@ithaca.edu](mailto:tpfaff@ithaca.edu) (Thomas J. Pfaff).

case the fixed resource is represented by the complete graph on  $n$  vertices and species may interact with all other species in the system.

The mean field voter model with mutation, also known as the infinite alleles model or Moran model, with  $n$  sites is described as follows. Each of the  $n$  sites contains a particle of a specified type, labelled by a real number from  $(0, 1)$ . There may be more than one particle of the same type. Each particle (or site) invades one of the other  $n - 1$  sites chosen at random, at rate 1, and mutation occurs at each site, at rate  $\alpha$ . When a particle invades another site, we think of the particle as producing an offspring, of the same type, that takes over the new site. The invaded site still has one particle, but now of the type of the invader and the source site remains unchanged. When mutation occurs, the particle changes its type to a new type, chosen randomly from  $(0, 1)$ . Formally, let  $C_n$  be the complete graph with  $n$  sites so that the random function  $\xi_t^n : C_n \rightarrow (0, 1)$  gives the state of the system at time  $t$ . For each  $i \in C_n$ ,  $\xi_t^n(i)$  is the type of individual, or species, at site  $i$  at time  $t$ . BCD proved that the multitype mean field voter model with mutation has, independent of the initial configuration, a unique stationary state,  $\xi_\infty^n$ . We define the patch size for the type at site  $i$  at time  $t$  to be the number of sites  $j$  with  $\xi_t^n(i) = \xi_t^n(j)$ . Now define  $K_n[m]$  to be the number of species with patch size equal to  $m$ . Also, for  $1 \leq a < b$ ,  $K_n[a, b]$  is the number of species with patch size in  $[a, b]$ . In other words,  $K_n[a, b] = \sum_{m=a}^b K_n[m]$ , where the sum is over all integers in  $[a, b]$ .

The distribution of species in the stationary state of the multitype mean field voter model with mutation is given by the Ewens Sampling Formula. Namely, the joint distribution of  $K_n[1], K_n[2], \dots, K_n[n]$  is

$$\frac{n!}{\theta^{(n)}} \prod_{i=1}^n \frac{\theta^{J_i}}{i^{J_i} J_i!},$$

where we have  $J_i$  species with patch size  $i$ . A proof of this can be found in Kelly (1979). In Proposition 2 we show that the multitype mean field voter model with mutation is equivalent to a linear birth process with immigration, and that both processes are connected to the Ewens Sampling Formula. In particular, we provide a new proof that directly relates each process to the Ewens Sampling Formula. The Ewens Sampling Formula is well known and it provides the null hypothesis distribution for the neutral theory of evolution. This is seen in our model since each species is equally likely to invade the other  $n - 1$  sites and thus no species has a selective advantage over any other species.

In our model, species enter the system through mutation at rate  $\alpha$ , which can be thought of as migration or genetic mutation. We will investigate the limiting behavior of the species abundance distribution as  $\alpha \rightarrow 0$ , since we want the mutation rate  $\alpha$  to be small. This will require us to simultaneously

send  $n \rightarrow \infty$ , as  $\alpha \rightarrow 0$ . Otherwise, with  $n$  fixed and  $\alpha = 0$ , one species would eventually take over due to the fact that there are only a finite number of sites. We use  $v \wedge 1$  to represent the minimum of  $v$  and 1, and for convenience we will write  $\bar{\alpha}$  for  $\alpha^{-1}$ .

The results about the mean field model are similar to the spatial model given in BCD. Let  $B(L)$  be the cube with side  $L$  centered at the origin in  $\mathbb{Z}^d$ . Then, in BCD,  $N(B(L), I)$  is the number of species in  $B(L)$  with patch size in  $I$ . In other words,  $N(B(L), I)$  is analogous to  $K_n[a, b]$ , with  $I = [a, b]$ , where  $L$  and  $n$  play similar roles. When  $d \geq 3$ , Theorem 1 in BCD holds with  $K_n[1, \bar{\alpha}^v]/n$  playing the role of  $N(B(L), [1, \bar{\alpha}^v])/|B(L)|$ , where in both cases the division is used to get species abundance per unit volume. Namely, if we let  $n = n(\alpha)$  be a function of  $\alpha$  such that  $\alpha n \geq 1$  as  $\alpha \rightarrow 0$ , then for any  $\epsilon$  positive,

$$\lim_{\alpha \rightarrow 0} P \left( \left| \frac{K_n[1, \bar{\alpha}^v]}{\alpha n \log \bar{\alpha}} - (v \wedge 1) \right| > \epsilon \right) = 0, \quad (1)$$

for all  $v$  positive. In the case when  $v = 1$ , this says that the number of species per unit volume with patch size in  $[1, \bar{\alpha}]$  grows like  $\alpha \log \bar{\alpha}$ . In BCD, for  $d \geq 3$ , this would translate to the number of species per unit volume with patch size in  $[1, \bar{\alpha}]$  grows as  $\alpha \log \bar{\alpha} / \gamma_d$ , where  $\gamma_d$  is the probability that the simple symmetric random walk in  $\mathbb{Z}^d$  never returns to the origin.

Another way to motivate (1) is as follows. Let  $Z_m$  be a Poisson random variable with  $EZ_m = \alpha n / m$ . If  $\alpha n$  remains constant, then Theorem 1 in Arratia, Barbour, and Tavaré (1992), implies that  $K_n[m]$  converges in distribution, as  $n \rightarrow \infty$ , to  $Z_m$ . Their work was done in terms  $J_i$ , the number of cycles of size  $i$  in a random permutation. Given this we expect

$$K_n[m] \approx \alpha n / m, \quad (2)$$

where  $\approx$  means approximately equal. Now, for  $v \leq 1$ ,

$$K_n[1, \bar{\alpha}^v] = \sum_{m=1}^{\bar{\alpha}^v} K_n[m] \approx \sum_{m=1}^{\bar{\alpha}^v} \alpha n / m \approx \alpha n v \log \bar{\alpha}. \quad (3)$$

It is also the case that Theorem 3 in BCD, when  $d \geq 3$ , holds with  $K_n[a\bar{\alpha}, b\bar{\alpha}]/n$  playing the role of  $N(B(L), [a\gamma_d\bar{\alpha}, b\gamma_d\bar{\alpha}])/|B(L)|$ . Specifically, if  $\alpha n / \log \log \bar{\alpha} \rightarrow \infty$  as  $\alpha \rightarrow 0$ , then for any  $\epsilon$  positive and  $0 < a < b$ ,

$$\lim_{\alpha \rightarrow 0} P \left( \left| K_n[a\bar{\alpha}, b\bar{\alpha}] - \alpha n \int_a^b z^{-1} e^{-z} dz \right| > \epsilon \alpha n \right) = 0. \quad (4)$$

To motivate this, for  $a$  and  $b$  small we get, from (2), that

$$K_n[a\bar{\alpha}, b\bar{\alpha}] = \sum_{m=a\bar{\alpha}}^{b\bar{\alpha}} K_n[m] \approx \sum_{m=a\bar{\alpha}}^{b\bar{\alpha}} \alpha n/m \approx \alpha n \log(b/a), \quad (5)$$

which gives (4) since for  $a$  and  $b$  small we have that

$$\int_a^b \frac{1}{z} e^{-z} dz \approx \int_a^b \frac{1}{z} dz = \log(b/a). \quad (6)$$

To see that (4) is consistent with (1) we allow greater liberty on how to choose  $a$  and  $b$ . First, note that if  $a=0$ , (6) implies that the integral is infinite. This is expected since, by (1),  $K_n[1, \bar{\alpha}]$  is of order  $\alpha n \log \bar{\alpha}$  and not  $\alpha n$ . We want to “apply” (4) to estimate  $K_n[\bar{\alpha}^{v_1}, \bar{\alpha}^{v_2}]$ , where  $0 < v_1 < v_2 < 1$ . To do this, define  $a$  and  $b$  so that  $\bar{\alpha}^{v_1} = a\bar{\alpha}$  and  $\bar{\alpha}^{v_2} = b\bar{\alpha}$ . Since  $a$  and  $b$  tend to 0 as  $\alpha \rightarrow 0$ , we expect that for  $\alpha$  small,

$$K_n[\bar{\alpha}^{v_1}, \bar{\alpha}^{v_2}] \approx \alpha n \int_a^b \frac{1}{z} e^{-z} dz \approx \alpha n \int_a^b \frac{1}{z} dz = \alpha n \log(b/a) = (v_2 - v_1) \alpha n \log \bar{\alpha}.$$

This matches up with the right hand side of (5) and what is expected from (1), since

$$K_n[\bar{\alpha}^{v_1}, \bar{\alpha}^{v_2}] = K_n[1, \bar{\alpha}^{v_2}] - K_n[1, \bar{\alpha}^{v_1}] \approx (v_2 - v_1) \alpha n \log \bar{\alpha}.$$

Due to the lack of spatial structure in the mean field model, we are able to prove a converse to the analog of Theorem 2 in BCD. We obtain,

**Theorem 1** *Put  $J_\delta^\alpha = \{j : \delta^{-1} \leq r^j \leq \delta \bar{\alpha}\}$ , and let  $r > 1$  and  $\epsilon$  positive. If  $\alpha n / \log \log \bar{\alpha} \rightarrow \infty$  as  $\alpha \rightarrow 0$ , then*

$$\lim_{\delta \rightarrow 0} \limsup_{\alpha \rightarrow 0} P \left( \bigcup_{J_\delta^\alpha} (|K_n[r^j, r^{j+1}] - \alpha n \log r| > \epsilon \alpha n) \right) = 0. \quad (7)$$

Moreover, if  $\alpha n \rightarrow \infty$  and  $\alpha n / \log \log \bar{\alpha} \rightarrow 0$  as  $\alpha \rightarrow 0$ , then

$$\lim_{\delta \rightarrow 0} \liminf_{\alpha \rightarrow 0} P \left( \bigcup_{J_\delta^\alpha} (|K_n[r^j, r^{j+1}] - \alpha n \log r| > \epsilon \alpha n) \right) > 0. \quad (8)$$

Here (8) is a converse of (7). In BCD there is an analogue of (7), but there is not an analogue of (8). Our Theorem 1 gives a more complete picture than Theorem 2 in BCD, by almost obtaining necessary and sufficient conditions.

Theorem 1 organizes the species abundance counts in terms of a histogram. We fix  $r > 1$  and consider  $K_n[r^j, r^{j+1}]$ , the number of species with patch size

in  $[r^j, r^{j+1})$ . We use the half open intervals so that the histogram bins don't overlap. Intuitively, we expect from (2), that

$$K_n[r^j, r^{j+1}) \approx \sum_{m=r^j}^{r^{j+1}} K_n[m] \approx \sum_{m=r^j}^{r^{j+1}} \alpha n/m \approx \alpha n \log r.$$

Although we need  $\alpha n \rightarrow \infty$ , as opposed to just  $\alpha n \geq 1$  as in (1), the result is uniform.

We proceed in the following manner. In section 2 we define the multitype mean field voter model and show that its stationary distribution is also given by a linear birth model. At the same time we show that both distributions are given by the Ewens Sampling Formula, and introduce necessary notation. Sections 3 and 4 contain the proof of Theorem 1 equation (7) and Theorem 1 equation (8), respectively, which will use the same linear birth process used by Donnelly, Kurtz, and Tavaré (1991). The proofs of (1) and (4) are similar to that of Theorem 1 and can be found in Pfaff (1999).

## 2 Constructions and Connections

In this section we will define the mean field voter model with mutation and relate its equilibrium state to a linear birth process with immigration. In fact, we will show that the distribution of species abundance, of the mean field voter model with mutation in equilibrium and linear birth process with immigration, is given by the Ewens sampling formula. We will also record some results that will be used repeatedly in proving the theorem in the following sections. Throughout this section,  $n$  will be fixed and  $C_n$  will denote the complete graph on  $n$  vertices, with the vertices labelled 1 to  $n$ .

### 2.1 Constructions and Notation

The construction of the mean field voter model is similar to the spatial model given in BCD, which we refer the reader to for details. Essentially, the difference between the construction in BCD and the one here is that we replace  $\mathbf{Z}^d$  with  $C_n$ , and particles will be able to interact with all other sites as opposed to nearest neighbors. Informally, we will have times  $\tilde{T}_m^i$  at which the particle at site  $i$  will choose a site  $Z_m^i$ , and the particle at  $Z_m^i$  will adopt the value at  $i$ . Here we allow a particle to choose its own site. We also have times  $S_m^i$  at which the particle at site  $i$  undergoes a mutation event and adopts the value  $U_m^i$ , where  $U_m^i$  is uniform  $(0, 1)$ . Similar to BCD, we let  $\xi_t^{n,j}$  be the number

of species with patch size  $j$  at time  $t$  in  $C_n$ , and let  $\xi_\infty^n$  denote the unique stationary distribution.

Again as in BCD, we have a process “dual” to  $\xi_t^n$ , which is a coalescing random walk with killing, by killing or removing from the system any particle which experiences a mutation. Particles undergo a rate-one random walk, jumping to any of the  $n$  sites uniformly, and are killed at rate  $\alpha$ . The mass of a particle  $i$  at time  $t$  is  $\hat{\zeta}_t^n(i)$ , the number of particles that have coalesced with the particle up to time  $t$ . If  $\hat{\zeta}_t^n(i) = 0$  then there is no particle at site  $i$  at time  $t$ .

We now define the process  $\hat{\mathcal{J}}_t^n = (\hat{\mathcal{J}}_1^t, \hat{\mathcal{J}}_2^t, \dots, \hat{\mathcal{J}}_n^t)$ , where each  $\hat{\mathcal{J}}_r^t$  for  $1 \leq r \leq n$  is a nonnegative integer. At time  $t = 0$ ,  $\hat{\mathcal{J}}_r^t = 0$  for all  $r$ . At every time  $t$  such that  $\hat{\zeta}_t^n$  undergoes a killing, say at site  $i$ , we increase  $\hat{\mathcal{J}}_{\hat{\zeta}_t^n(i)}^t$  by 1. Informally, the vector  $\hat{\mathcal{J}}_t^n$  records the mass of the particles in  $\hat{\zeta}_t^n$  at the time they are killed. Now, since killings happen at rate  $\alpha$  for each site, eventually all particles will be killed. Hence, we let  $\hat{\mathcal{J}}_\infty^n = \lim_{t \rightarrow \infty} \hat{\mathcal{J}}_t^n$ , which is the mass of all the particles killed once there are no particles left. Similarly, let

$$\mathcal{J}_k^n = \left| i : \sum_{j \in C_n} 1\{\xi_\infty^n(i) = \xi_\infty^n(j)\} = k \right| / k, \quad (9)$$

which is the number of species in  $\xi_\infty^n$  with patch size equal to  $k$ , and put

$$\mathcal{J}^n = (\mathcal{J}_1^n, \mathcal{J}_2^n, \dots, \mathcal{J}_n^n). \quad (10)$$

Let  $J^n = (J_1, \dots, J_n)$ , where  $n = |J^n| = \sum_{i=1}^n i J_i$  and each  $J_i$  is a nonnegative integer. The duality equation we get is

$$P(\mathcal{J}^n = J^n) = P(\hat{\mathcal{J}}_\infty^n = J^n). \quad (11)$$

In other words, the process  $\hat{\zeta}_t^n$  is dual to  $\xi_t^n$  in the sense that the distribution of patch sizes in  $\xi_\infty^n$  equals the distribution of masses in  $\hat{\mathcal{J}}_\infty^n$ .

We now define the linear birth process with immigration, which is also used in Donnelly, Kurtz, and Tavaré, (1991). Let  $\{T_m, m \geq 1\}$  be the arrival times of an independent rate  $\alpha n$  Poisson process,  $\mathcal{P}(\cdot)$ . At each time  $T_m$ , a new family or species, is initiated and grows independently of all other events as a linear rate-one birth process. Let  $X_m$  denote the birth process started at time  $T_m$ . The process, with  $I(0) = 0$ ,

$$I(t) = \sum_{m=1}^{(t)} X_m(t - T_m),$$

counts the total number of individuals present at time  $t$ , and was first introduced by Tavaré (1987). Note that if  $t < 0$  then  $X_m(t) = 0$ . We will use  $X(t)$

to denote a generic linear birth process started at time 0. Let

$$\tau_n = \inf\{t : I(t) = n\},$$

the first time the process reaches  $n$  individuals, and note that

$$\tau_n = \sum_{i=0}^{n-1} Y_i, \quad (12)$$

where  $Y_i$  are exponential rate  $\alpha n + i$ .

For  $m$  positive and  $1 \leq a < b$ , put

$$K_n[m] = \sum_{i=0}^{(\tau_n)} 1\{X_i(\tau_n - T_i) = m\}, \quad (13)$$

$$K_n[a, b] = \sum_{i=0}^{(\tau_n)} 1\{X_i(\tau_n - T_i) \in [a, b)\}, \quad (14)$$

which is the number of families or species in the linear birth process with patch size equal to  $m$  and in  $[a, b)$ , respectively.

## 2.2 Equivalence of VM and LBP

The equivalence between the VM and LBP can be found in Tavaré (1987), but here we provide a counting argument and take advantage of the dual process to prove the result. Let  $\beta_{J^n}$  be a  $n$ -permutation with  $J_i$  cycles of size  $i$ , where if  $J_i = 0$  then there is no cycle of size  $i$ . The cycles of  $\beta_{J^n}$  are written in increasing order by the first integer of each cycle. Put

$$B_{J^n} = \{\beta_{J^n} : |J^n| = n\}.$$

We use the permutation  $\beta_{J^n}$  to specify the order of immigration and birth in the birth process, and to specify the order of coalescing and killing in the coalescing random walk with killing.

As an example, suppose that we have the permutation

$$(1\ 5\ 2)(3\ 4\ 6), \quad (15)$$

which is a  $\beta_{J^6}$  with  $J^6 = (0, 0, 2, 0, 0)$ . For the birth process this is read as at time  $\tau_6$  the process had two families of size three. Here, 1 gave birth to 2, then the second family is started with a 3, which then gave birth to 4. The 1 then gave birth to 5 and 4 then gave birth to 6. The first number in a cycle represents an immigration. All other numbers represent a birth by the

first number to the left that is smaller. Given  $J^n$ , it is clear that there is a one to one correspondence with the permutations in  $B_{J^n}$  and the number of ways that the linear birth process can arrive at  $J_i$  species with patch size  $i$ . In particular, the order of  $B_{J^n}$  is

$$n! \prod_{i=1}^n \frac{1}{i^{J_i} J_i!}, \quad (16)$$

the number of  $n$ -permutations with  $J_i$  cycles of size  $i$ .

For the coalescing random walk with killing, we will need to keep track of the particles. We do this by labelling each particle 1 – 6 corresponding to the starting vertices of  $C_6$ . We may now refer to the particles as particle 1, 2, 3, 4, 5, or 6. When two particles coalesce the “new” particle will assume the number of the particle that did not jump. Let  $\pi^6$  be any permutation on  $n$  integers where  $\pi_i^6$  is the image of  $i$  under  $\pi^6$ . We now use the permutation

$$\pi^6 \beta_{J^6} = (\pi_1^6 \pi_5^6 \pi_2^6)(\pi_3^6 \pi_4^6 \pi_6^6), \quad (17)$$

where  $\beta_{J^6}$  is given in (15), to represent the sequence of coalescings and killings. This representation will disregard jumps that don't result in a coalescing. Here (17) is read as  $\hat{\mathcal{J}}_\infty^6 = (0, 0, 2, 0, 0)$ , in other words two particles were killed with mass three, which is represented by the two cycles of size three. The subscripts of  $\pi^6$  will give us the order of coalescings and killings. Now the process arrived at  $\hat{\mathcal{J}}_\infty^6$  in the following manner. First, particle  $\pi_6^6$  jumps to and coalesces with particle  $\pi_4^6$ . Then particle  $\pi_5^6$  jumps to particle  $\pi_1^6$  and the two particles coalesce. Now particle  $\pi_4^6$  jumps to and coalesces with particle  $\pi_3^6$ , and then particle  $\pi_3^6$  is killed. Finally, particle  $\pi_2^6$  jumps and coalesces with particle  $\pi_1^6$ , and then particle  $\pi_1^6$  is killed. If we didn't use the permutation  $\pi^6$  and used only the permutation  $\beta_{J^n}$ , it would always be the case, for example, that particle 1 is the last particle killed, which is certainly not always the case.

Formally, the subscripts of  $\pi^n$  in the permutation are followed in reverse order. When the first number in a cycle is reached, the particle is killed. All other subscripts of  $\pi^n$  represent a jump and coalescing with the particle with the smallest subscripts of  $\pi^n$ , to the left. Now given  $J^n$ , it is clear that there is a one to one correspondence with the permutations  $\pi^n \beta_{J^n}$ , where  $\pi^n$  is any  $n$ -permutation and  $\beta_{J^n}$  in  $B_{J^n}$ , and the number of ways that the random walk can have  $J_i$  particles killed with mass  $i$ . In particular, the number of ways that the random walk can have  $J_i$  particles killed with mass  $i$  is, similar to (16),

$$n! n! \prod_{i=1}^n \frac{1}{i^{J_i} J_i!}, \quad (18)$$

where the extra  $n!$  occurs from counting all the  $n$ -permutations,  $\pi^n$ .

We will now write  $\mathcal{I}(\tau_n) = \beta_{J^n}$  to mean that the birth process grew according

to the order given by the permutation  $\beta_{J^n}$ , and so  $(K_n[1], K_n[2], \dots, K_n[n]) = J^n$ . Similarly,  $\tilde{\mathcal{J}}^n = \pi^n \beta_{J^n}$  means that the random walk underwent the coalescent-killing sequence given by  $\pi^n \beta_{J^n}$ , and so  $\hat{\mathcal{J}}_\infty^n = J^n$ .

Put  $\theta = \alpha n$  and  $\theta_{(n)} = \theta(\theta+1) \cdots (\theta+n-1)$ . Suppose there are  $i > 0$  particles alive in the coalescing random walk with killing on  $C_n$ . Now for  $i \geq 2$  particles alive and any two particles  $x$  and  $y$ , the probability that particle  $x$  jumps onto particle  $y$  before any other event is

$$\frac{n^{-1}}{\alpha i + i(i-1)n^{-1}} = \frac{1}{i(\theta + i - 1)}. \quad (19)$$

Similarly, for  $i \geq 1$  particles alive and a given particle  $x$ , the probability that particle  $x$  is killed before any other event is

$$\frac{\alpha}{\alpha i + i(i-1)n^{-1}} = \frac{\theta}{i(\theta + i - 1)}. \quad (20)$$

For the birth process with  $i \geq 1$  individuals present, the probability that a given individual gives birth before any other event is

$$\frac{1}{\theta + i}, \quad (21)$$

and the probability of an immigration occurring before a birth occurs, with  $i \geq 0$  individuals alive, is

$$\frac{\theta}{\theta + i}. \quad (22)$$

Note that the numerators of equations (19), (20), (21), and (22) don't depend on the number of particles or individuals present.

**Proposition 2** *Let  $J^n = (J_1, \dots, J_n)$  with  $|J| = n$  fixed. The probability that the birth process at time  $\tau_n$  and the voter model in equilibrium,  $\xi_\infty^n$ , have  $J_i$  species with patch size  $i$  is given by the Ewens sampling formula. In other words*

$$P(\mathcal{J}^n = J^n) = \frac{n!}{\theta_{(n)}} \prod_{i=1}^n \frac{\theta^{J_i}}{i^{J_i} J_i!} = P((K_n[1], K_n[2], \dots, K_n[n]) = J^n). \quad (23)$$

**PROOF.** By (19) and (20), we have that

$$P(\tilde{\mathcal{J}}^n = \pi^n \beta_{J^n}) = \prod_{i=1}^n \frac{p_{i,i-1}}{i(\theta + i - 1)} = \frac{1}{n! \theta_{(n)}} \prod_{i=1}^n p_{i,i-1}, \quad (24)$$

where  $p_{i,i-1}$  is either 1 or  $\theta$  depending on whether or not we had a coalescent or killing as we decreased from  $i$  to  $i-1$  particles. The permutation  $\pi^n \beta_{J^n}$

tells us that for each  $1 \leq j \leq n$  there are  $J_j$  particles of mass  $j$  that are killed. Hence,  $\sum_{j=1}^n J_j$  is the number of times that  $p_{i,i-1} = \theta$ , and so

$$\prod_{i=1}^n p_{i,i-1} = \prod_{i=1}^n \theta^{J_i}. \quad (25)$$

Now, the number of ways to get  $J_i$  species with patch size  $i$  is given by (18). Thus, by summing over all  $n$ -permutations  $\pi^n$  and all  $\beta_{J^n} \in B_{J^n}$  we get, by (24), (25), (18), and the duality given in (11), that

$$P(\mathcal{J}^n = J^n) = P(\hat{\mathcal{J}}_\infty^n = J^n) = \frac{n!}{\theta_{(n)}} \prod_{i=1}^n \frac{\theta^{J_i}}{i^{J_i} J_i!}.$$

For the birth process, by (21) and (22), we have that

$$P(\mathcal{I}(\tau_n) = \beta_J) = \prod_{i=0}^{n-1} \frac{p_{i,i+1}}{\theta + i} = \frac{1}{\theta_{(n)}} \prod_{i=0}^{n-1} p_{i,i+1}, \quad (26)$$

where  $p_{i,i+1}$  is either 1 or  $\theta$  depending on whether or not we had an immigration or birth, as we increase from  $i$  to  $i + 1$  individuals. Here the permutation  $\beta_{J^n}$  tells us that for each  $j$  there are  $J_j$  families of size  $j$ . Hence,  $\sum_{j=1}^n J_j$  is the number of times that  $p_{i,i+1} = \theta$ , and so

$$\prod_{i=0}^{n-1} p_{i,i+1} = \prod_{i=1}^n \theta^{J_i}. \quad (27)$$

Now by summing over all  $\beta_{J^n} \in B_{J^n}$  we get, by (26), (27), and (16), that

$$P((K_n[1], K_n[2], \dots, K_n[n]) = J^n) = \frac{n!}{\theta_{(n)}} \prod_{i=1}^n \frac{\theta^{J_i}}{i^{J_i} J_i!}.$$

□

### 2.3 Preliminaries

This section contains some notation and a few results that will be used in the sections that follow. One may skip the section for now and return back as needed. The difficulty with (14) is that it contains the stopping time  $\tau_n$ . We need an event,  $\{l \leq \tau_n \leq u\}$ , with probability close to 1, on which we can

bound  $K_n[a, b]$  by Poisson random variables. Put

$$Y_n^{l,u}[a, b] = \sum_{i=1}^{(l)} 1\{X_i(u - T_i) < b; X_i(l - T_i) \geq a\} \quad (28)$$

$$\bar{Y}_n^{l,u}[a, b] = \sum_{i=1}^{(l)} 1\{X_i(u - T_i) < b\} - \sum_{i=1}^{(l)} 1\{X_i(l - T_i) < a\} \quad (29)$$

$$Z_n^{l,u}[a, b] = \sum_{i=1}^{(u)} 1\{X_i(l - T_i) < b; X_i(u - T_i) \geq a\} \quad (30)$$

$$\bar{Z}_n^{l,u}[a, b] = \sum_{i=1}^{(u)} 1\{X_i(l - T_i) < b\} - \sum_{i=1}^{(u)} 1\{X_i(u - T_i) < a\}. \quad (31)$$

Recall that if  $t < 0$  then  $X_i(t) = 0$ . Now, on  $\{l \leq \tau_n \leq u\}$ ,

$$K_n[a, b] = \sum_{i=1}^{(\tau_n)} 1\{X_i(\tau_n - T_i) < b; X_i(\tau_n - T_i) \geq a\} \geq Y_n^{l,u}[a, b] \geq \bar{Y}_n^{l,u}[a, b], \quad (32)$$

where the last inequality follows since  $1(A \cap B^c) \geq 1(A) - 1(B)$ . In a similar fashion we obtain upper bounds on  $K_n[a, b]$  and so we have that , on  $\{l \leq \tau_n \leq u\}$ ,

$$\bar{Y}_n^{l,u}[a, b] \leq Y_n^{l,u}[a, b] \leq K_n[a, b] \leq Z_n^{l,u}[a, b] = \bar{Z}_n^{l,u}[a, b]. \quad (33)$$

An important fact here is that  $Y$ 's and  $Z$ 's are Poisson random variables, since we are thinning the Poisson process  $\mathcal{P}(\cdot)$ . In particular, from the sums in (29) and (31), we get, see for instance Ross (1998), that

$$E\bar{Y}_n^{l,u}[a, b] = \alpha(n-1) \int_0^l P(X(u-s) < b) ds - \alpha n \int_0^l P(X(l-s) < a) ds \quad (34)$$

and

$$E\bar{Z}_n^{l,u}[a, b] = \alpha(n-1) \int_0^u P(X(l-s) < b) ds - \alpha n \int_0^u P(X(u-s) < a) ds. \quad (35)$$

**Lemma 3** *With  $\bar{Y}_n^{l,u}[a, b]$  and  $\bar{Z}_n^{l,u}[a, b]$  given by (29) and (31), we have that*

$$\begin{aligned} \frac{E\bar{Y}_n^{l,u}[a, b]}{\alpha(n-1)} &= \int_0^u P(X(u-s) \in [a, b)) ds - \int_l^u P(X(u-s) < b) ds \\ &\quad + \int_0^u P(X(u-s) < a) ds - \int_0^l P(X(l-s) < a) ds, \end{aligned} \quad (36)$$

and

$$\begin{aligned} \frac{E\bar{Z}_n^{l,u}[a,b]}{\alpha(n-1)} &= \int_0^l P(X(l-s) \in [a,b)) ds + \int_l^u P(X(l-s) < b) ds \\ &\quad + \int_0^l P(X(l-s) < a) ds - \int_0^u P(X(u-s) < a) ds. \end{aligned} \quad (37)$$

**PROOF.** This follows by rewriting the integrals in (34) and (35).  $\square$

Since we have Poisson random variables and a way to calculate their means we will make use the following Lemma, which is Lemma 2.2 in BCD.

**Lemma 4** *Let  $X$  be a Poisson random variable, and, for  $\lambda > 0$ , let  $c_\lambda = \lambda \log \lambda - \lambda + 1$ . Then,  $c_\lambda > 0$  for  $\lambda \neq 1$  and*

$$\begin{aligned} P(X \geq \lambda EX) &\leq \exp(-c_\lambda EX), & \lambda > 1, \\ P(X \leq \lambda EX) &\leq \exp(-c_\lambda EX), & \lambda < 1. \end{aligned}$$

Recall that  $X(t)$  denotes a generic linear birth process. The next Lemma shows that  $e^{-t}X(t) \approx \xi(1)$  for  $t$  large, where  $\xi(1)$  is an exponential random variable with parameter 1.

**Lemma 5** *Given  $\epsilon_1, \epsilon_2, x$  and  $y$  positive, there exists a  $t_0$  such that for any  $t \geq t_0$*

$$P\left(e^{-t}X(t) \in [xe^{-t}, ye^{-t})\right) \leq P\left(\xi(1) \in [(1+\epsilon_1)^{-1}xe^{-t}, (1-\epsilon_1)^{-1}ye^{-t})\right) + \epsilon_2 \quad (38)$$

$$P\left(e^{-t}X(t) \in [xe^{-t}, ye^{-t})\right) \geq P\left(\xi(1) \in [(1-\epsilon_1)^{-1}xe^{-t}, (1+\epsilon_1)^{-1}ye^{-t})\right) - \epsilon_2 \quad (39)$$

**PROOF.** The proof follows since  $e^{-t}X(t)$  is an  $L_2$  bounded martingale that converges almost surely to  $X$ , where  $X$  is an exponential random variable with parameter 1, and so  $e^{-t}X(t) \rightarrow \xi(1)$  in probability.  $\square$

For small values of  $t$  we are able to control the size of  $e^{-t}X(t)$  with the next Lemma. The proof of the first part of Lemma 6 may be found in Donnelly, Kurtz, and Tavaré (1991), and the second follows from Doob's inequality, since  $e^{-s}X(s)$  is a positive martingale with mean 1. The last Lemma of this section, Lemma 7, provides a calculation involving  $\xi(1)$ .

**Lemma 6** For any  $x$  positive

$$P\left(\inf_{s>0} e^{-s} X(s) \leq x\right) < 2x^{1/2}, \quad (40)$$

and

$$P\left(\sup_{s>0} e^{-s} X(s) > x\right) \leq x^{-1}. \quad (41)$$

**Lemma 7** For  $r \geq 1$ ,

$$\lim_{b \rightarrow \infty} \int_{-b}^b P\left(\xi(1) \in [e^{-t}, re^{-t}]\right) dt = \log r. \quad (42)$$

**PROOF.** First, set  $v = e^t$  and split the domain of integration of the last integral into  $[e^{-b}, 0]$  and  $[0, e^b]$ . Note that the integral over  $[e^{-b}, 0]$  is 0. For the other integral we have that

$$\int_0^{e^b} \frac{\exp(-v^{-1}) - \exp(-rv^{-1})}{v} dv = \int_{e^b/r}^{e^b} \frac{\exp(-v^{-1})}{v} dv.$$

One more change of variables,  $t = v/e^b$ , and noticing that

$$\lim_{b \rightarrow \infty} \int_{1/r}^1 \frac{\exp(-e^{-b}t^{-1})}{t} dt = \int_{1/r}^1 \frac{dt}{t} = \log r,$$

finishes the proof.  $\square$

### 3 Proof of Theorem 1 Equation (7)

Throughout this section  $r > 1$  will be fixed. Put

$$\begin{aligned} d_\alpha &= (\log \log \bar{\alpha})^{-1/4}, \\ J_\delta^\alpha &= \{j : \delta^{-1} \leq r^j \leq \delta \bar{\alpha}\}, \\ \Omega_\alpha &= \{|\tau_n - \log \bar{\alpha}| \leq d_\alpha\}. \end{aligned}$$

Also let

$$l = \log \bar{\alpha} - d_\alpha \quad \text{and} \quad u = \log \bar{\alpha} + d_\alpha,$$

so that  $\Omega_\alpha = \{l \leq \tau_n \leq u\}$ .

The proof of (7) we will need the following two results, which we will prove at the end of this section.

**Lemma 8** *If  $\alpha n / \log \log \bar{\alpha} \rightarrow \infty$  as  $\alpha \rightarrow 0$ , then*

$$\lim_{\alpha \rightarrow 0} P(\Omega_\alpha^c) = 0. \quad (43)$$

**Proposition 9** *Given  $\epsilon$  positive there exists a  $\delta_0$ ,  $\alpha_0$ , and a constant  $C(\epsilon)$ , with  $C(\epsilon) > 0$  and  $\lim_{\epsilon \rightarrow 0} C(\epsilon) = 0$ , such that for any  $\delta \leq \delta_0$  and  $\alpha \leq \alpha_0$*

$$P\left(\left|K_n[r^j, r^{j+1}] - \alpha n \log r\right| > \alpha n \epsilon; \Omega_\alpha\right) \leq \exp(-\alpha n C(\epsilon) \log r), \quad (44)$$

for all  $j \in J_\delta^\alpha$ .

In order to prove (7) we use Proposition 9 to get that

$$\begin{aligned} & P\left(\bigcup_{J_\delta^\alpha} \left(\left|K_n[r^j, r^{j+1}] - \alpha n \log r\right| > \alpha n \epsilon\right)\right) \\ & \leq P(\Omega_\alpha^c) + \frac{\log \bar{\alpha} + 2 \log \delta + \log r}{\log r} \exp(-\alpha n C(\epsilon) \log r), \end{aligned} \quad (45)$$

which we will be able to make small. If we ignore the  $\log r$  and  $\log \delta$ , the key part of this upper bound is

$$\log \bar{\alpha} \exp(-\alpha n C(\epsilon)) = \exp(\log \log \bar{\alpha} (1 - \alpha n C(\epsilon) / \log \log \bar{\alpha})).$$

In order to make this small we need  $\alpha n C(\epsilon) / \log \log \bar{\alpha} > 1$ . Since  $C(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$ , we see the need for  $\alpha n / \log \log \bar{\alpha} \rightarrow \infty$ .

**Proof of 7** We will show that given  $\epsilon$  positive, there exists a  $\delta_0$  such that for any  $\delta \leq \delta_0$

$$\limsup_{\alpha \rightarrow 0} P\left(\bigcup_{J_\delta^\alpha} \left(\left|K_n[r^j, r^{j+1}] - \alpha n \log r\right| > \epsilon \alpha n\right)\right) \leq 2\epsilon.$$

Let  $\delta_0$  satisfy Proposition 9 and fix  $\delta \leq \delta_0$ . Note that for  $\delta$  fixed,

$$\frac{\log \bar{\alpha} + 2 \log \delta + \log r}{\log r} \exp(-\alpha n C(\epsilon) \log r) \approx \exp(\log \log \bar{\alpha} - \alpha n C(\epsilon) \log r). \quad (46)$$

Require  $\alpha_0$  to satisfy Proposition 9 and so that for any  $\alpha \leq \alpha_0$ , by (46) and  $\alpha n / \log \log \bar{\alpha} \rightarrow \infty$ ,

$$\frac{\log \bar{\alpha} + 2 \log \delta + \log r}{\log r} \exp(-\alpha n C(\epsilon) \log r) < \epsilon, \quad (47)$$

and, by Lemma 8,

$$P(\Omega_\alpha^c) < \epsilon. \quad (48)$$

We now have, by (44), (47), and the fact that the cardinality of  $J_\delta^\alpha$  is bounded above by  $((\log \bar{\alpha} + 2 \log \delta) / \log r) + 1$ , that

$$\sum_{J_\delta^\alpha} P \left( \left| K_n[r^j, r^{j+1}] - \alpha n \log r \right| > \alpha n \epsilon; \Omega_\alpha \right) \leq \epsilon. \quad (49)$$

Combining (45), (48) and (49), finishes the proof.  $\square$

### 3.1 Proof of Lemma 8

**Proof of Lemma 8** We will first estimate  $E(\tau_n)$ , which is approximately  $\log \bar{\alpha}$ , and  $Var(\tau_n)$ . Then we will apply Chebyshev's inequality to

$$P(|\tau_n - \log \bar{\alpha}| > d_\alpha), \quad (50)$$

to finish the proof. Note that

$$\log(\bar{\alpha}) \leq \sum_{i=0}^{n-1} \frac{1}{\alpha n + i} \leq (\alpha n)^{-1} + \log(1 + \alpha) + \log \bar{\alpha}. \quad (51)$$

By (12),  $E\tau_n = \sum_{i=0}^{n-1} 1/(\alpha n + i)$ . Hence, from (51)

$$|E\tau_n - \log(\bar{\alpha})| = \sum_{i=0}^{n-1} \frac{1}{\alpha n + i} - \log \bar{\alpha} \leq (\alpha n)^{-1} + \log(1 + \alpha). \quad (52)$$

Since the  $\{X_i\}$  are independent,  $Var(X_i) = (\alpha n + i)^{-2}$ , and  $\alpha n \geq 1$ ,

$$Var(\tau_n) = \sum_{i=0}^{n-1} \frac{1}{(\alpha n + i)^2} \leq \frac{1}{(\alpha n)^2} + \frac{1}{\alpha n}.$$

Moreover, since  $\alpha n / \log \log \bar{\alpha}$  as  $\alpha \rightarrow 0$ , we may assume that  $1 \leq \log \log \bar{\alpha} \leq \alpha n$ , and get that

$$Var(\tau_n) \leq 2(\alpha n)^{-1} \leq 2(\log \log \bar{\alpha})^{-1}. \quad (53)$$

Now by (52), Chebyshev's inequality, and then (53)

$$\begin{aligned} P(|\tau_n - \log \bar{\alpha}| > d_\alpha) &\leq P(|\tau_n - E\tau_n| > d_\alpha - (\alpha n)^{-1} - \log(1 + \alpha)) \\ &\leq 2(\log \log \bar{\alpha})^{-1/2} (1 - (\alpha n d_\alpha)^{-1} - d_\alpha^{-1} \log(1 + \alpha))^{-2}. \end{aligned} \quad (54)$$

It is easy to see that  $d_\alpha^{-1} \log(1 + \alpha)$  and  $(\alpha n d_\alpha)^{-1}$  tend to 0 as  $\alpha \rightarrow 0$ , since  $\alpha n / \log \log \bar{\alpha} \rightarrow \infty$  implies that  $\alpha n \rightarrow \infty$ . Therefore (54) to 0 as  $\alpha \rightarrow 0$ .  $\square$

### 3.2 Proof of Proposition 9

Recall (33) and the related definitions. We will prove Proposition 9 if we prove that

$$P\left(Y_n^{l,u}[r^j, r^{j+1}] < (1 - \epsilon)\alpha n \log r\right) \leq \exp(-\alpha n C(\epsilon)\alpha n \log r), \quad (55)$$

$$P\left(Z_n^{l,u}[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r\right) \leq \exp(-\alpha n C(\epsilon)\alpha n \log r), \quad (56)$$

in place of (44). This will be done by applying Lemma 4 to the probabilities in (55) and (56). Thus we will need to estimate the mean of  $Y_n^{l,u}[r^j, r^{j+1}]$  and  $Z_n^{l,u}[r^j, r^{j+1}]$ , and this is done using (36) and (37). The key to estimating these means is to show that

$$\int_0^u P\left(X(u-s) \in [r^j, r^{j+1}]\right) ds \approx \log r, \quad (57)$$

$$\int_0^l P\left(X(l-s) \in [r^j, r^{j+1}]\right) ds \approx \log r, \quad (58)$$

which are the first integrals in (36) and (37). The proof of (57) is obtained by integrating over  $[u - \log r^j - H, u - \log r^j + H]$ , which will give the necessary lower bound. For this, we need that given any  $\epsilon$  positive there exists an  $H$  such that there exists a  $\delta_0$  so that for any  $\delta \leq \delta_0$

$$\liminf_{\delta \rightarrow 0} \int_{u - \log r^j - H}^{u - \log r^j + H} P\left(X(u-s) \in [r^j, r^{j+1}]\right) ds \geq \log r - \epsilon, \quad (59)$$

for all  $j$  such that  $r^j \geq \delta^{-1}$ , which is proven at the end of the section.

In order to prove (58) we will break up the domain of integration of the integral into three pieces, where  $H$  is chosen so that  $l - \log r^j - H > 0$  and  $\log r^j > H$ ,

$$[0, l - \log r^j - H] \quad [l - \log r^j - H, l - \log r^j + H] \quad [l - \log r^j + H, l],$$

and use the following results, which we will prove at the end of this section. For any  $H$  positive,

$$\int_0^{l - \log r^j - H} P\left(X(l-s) \in [r^j, r^{j+1}]\right) ds \leq 4r^{1/2}e^{-H/2}, \quad (60)$$

$$\int_{l - \log r^j + H}^l P\left(X(l-s) \in [r^j, r^{j+1}]\right) ds \leq e^{-H}. \quad (61)$$

Also given  $\epsilon$  positive there exists an  $H_\epsilon$ , such that for any  $H \geq H_\epsilon$  there exists a  $\delta_0$  such that for any  $\delta \leq \delta_0$

$$\int_{l - \log r^j - H}^{l - \log r^j + H} P\left(X(l-t) \in [r^j, r^{j+1}]\right) dt \leq \log r + \epsilon, \quad (62)$$

for all  $j$  such that  $r^j \geq \delta^{-1}$

Intuitively this is what we expect. The integral over  $[l - \log r^j - H, l - \log r^j + H]$  will contribute the  $\log r$  since arrivals in  $[0, l - \log r^j - H]$  (resp  $[l - \log r^j + H, l]$ ) have had too much (resp not enough) time to end up in  $[r^j, r^{j+1}]$ .

The rest of the section will proceed as follows: Lemmas 10 and 11 will make precise (57) and (58). These results will then be used to prove the necessary estimates on  $EY_n^{l,u}[r^j, r^{j+1}]$  and  $EZ_n^{l,u}[r^j, r^{j+1}]$ , which are found in Lemmas 12 and 13. We will then prove Proposition 9 and finish with the proofs of (59) and (60).

**Lemma 10** *For any  $\epsilon$  positive there exists a  $\delta_0$  such that for any  $\delta \leq \delta_0$*

$$\int_0^u P\left(X(u-t) \in [r^j, r^{j+1}]\right) dt \geq \log r - \epsilon, \quad (63)$$

for any  $\alpha$  positive and all  $j \in J_\delta^\alpha$ .

**PROOF.** Let  $H$  be given by (59) and choose  $\delta_0 < 1$  so that for any  $\delta \leq \delta_0$ ,  $\log \delta^{-1} \geq H$  and, by (59),

$$\int_{u - \log r^j - H}^{u - \log r^j + H} P\left(X(u-t) \in [r^j, r^{j+1}]\right) dt \geq \log r - \epsilon, \quad (64)$$

for all  $j$  such that  $r^j \geq \delta^{-1}$ . In particular, for any  $\alpha$  positive, if  $j \in J_\delta^\alpha$  then  $j$  satisfies  $r^j \geq \delta^{-1}$ . Hence  $\log \delta^{-1} \leq \log r^j \leq \log \delta \bar{\alpha}$  and so

$$\begin{aligned} u - \log r^j - H &\geq \log \bar{\alpha} + d_\alpha - \log \delta \bar{\alpha} - H = \log \delta^{-1} - H + d_\alpha > 0, \\ \log r^j - H &\geq \log \delta^{-1} - H \geq 0, \end{aligned}$$

which shows that

$$[0, u] \supseteq [u - \log r^j - H, u - \log r^j + H].$$

Therefore, by (64),

$$\int_0^u P\left(X(u-t) \in [r^j, r^{j+1}]\right) dt \geq \log r - \epsilon.$$

□

**Lemma 11** *For any  $\epsilon$  positive there exists a  $\delta_0$  such that for any  $\delta \leq \delta_0$*

$$\int_0^l P\left(X(l-t) \in [r^j, r^{j+1}]\right) dt \leq \log r + 3\epsilon, \quad (65)$$

for all  $j$  such that  $r^j \geq \delta^{-1}$

**PROOF.** Let  $H_\epsilon$  be given by (62). Now take  $H \geq H_\epsilon$  so that

$$4r^{1/2}e^{-H/2} \leq \epsilon \quad (66)$$

$$e^{-H} \leq \epsilon. \quad (67)$$

Choose  $\delta_0$  so that for any  $\delta \leq \delta_0$ , by (62),

$$\int_{l-\log r^j-H}^{l-\log r^j+H} P\left(X(l-t) \in [r^j, r^{j+1})\right) dt \leq \log r + \epsilon, \quad (68)$$

for all  $j$  such that  $r^j \geq \delta^{-1}$ . We have, by (60), (68), and (61), that

$$\int_0^l P\left(X(l-t) \in [r^j, r^{j+1})\right) dt \leq 4r^{1/2}e^{-H/2} + \log r + \epsilon + e^{-H}.$$

Hence, for any  $\delta \leq \delta_0$ , by (66) and (67),

$$\int_0^l P\left(X(l-t) \in [r^j, r^{j+1})\right) dt \leq \log r + 3\epsilon,$$

for all  $j$  such that  $r^j \geq \delta^{-1}$ .  $\square$

We now turn to the estimates of  $E\left(Y_n^{l,u}[r^j, r^{j+1})\right)$  and  $E\left(Z_n^{l,u}[r^j, r^{j+1})\right)$ , which are Lemmas 12 and 13 below. Since the two proofs are almost identical we prove only Lemma 12.

**Lemma 12** *Given  $\epsilon$  positive there exists a  $\delta_0$  and  $\alpha_0$  such that for any  $\delta \leq \delta_0$  and  $\alpha \leq \alpha_0$*

$$E\left(Y_n^{l,u}[r^j, r^{j+1})\right) \geq (1 - \epsilon/2)\alpha n \log r, \quad (69)$$

for all  $j \in J_\delta^\alpha$ .

**PROOF.** Since  $\bar{Y}_n[r^j, r^{j+1}) \leq Y_n^{l,u}[r^j, r^{j+1})$  we will prove (69) with  $\bar{Y}_n[r^j, r^{j+1})$ . Using (36), we have that

$$\begin{aligned} \frac{E\left(\bar{Y}_n[r^j, r^{j+1})\right)}{\alpha n} &\geq \int_0^u P\left(X(u-s) \in [r^j, r^{j+1})\right) ds \\ &\quad - \int_l^u P\left(X(u-s) < r^{j+1}\right) ds \\ &\quad + \int_0^u P\left(X(u-s) < r^j\right) ds \\ &\quad - \int_0^l P\left(X(l-s) < r^j\right) ds. \end{aligned} \quad (70)$$

By Lemma 10, there exists a  $\delta_0$  such that for any  $\delta \leq \delta_0$

$$\int_0^u P\left(X(u-s) \in [r^j, r^{j+1})\right) ds \geq (1 - \epsilon/4) \log r, \quad (71)$$

for any  $\alpha$  positive and all  $j \in J_\delta^\alpha$ . Choose  $\alpha_0$  so that for any  $\alpha \leq \alpha_0$ ,  $8d_\alpha \leq \epsilon \log r$ . Hence, for the second integral in (70), we have that

$$\int_l^u P\left(X(u-s) < r^{j+1}\right) ds \leq u-l \leq 2d_\alpha \leq \epsilon \log r/4. \quad (72)$$

For the last two integrals in (70), a change of variables, setting  $t = s + 2d_\alpha$ , shows that

$$\int_0^l P\left(X(l-s) < r^j\right) ds = \int_{2d_\alpha}^u P\left(X(u-t) < r^j\right) dt,$$

and it follows that

$$\int_0^u P\left(X(u-s) < r^j\right) ds - \int_0^l P\left(X(l-s) < r^j\right) ds > 0. \quad (73)$$

Combining (70) with (71), (72), and (73), proves (69).  $\square$

**Lemma 13** *Given  $\epsilon$  positive there exists a  $\delta_0$  and  $\alpha_0$  such that for any  $\delta \leq \delta_0$  and any  $\alpha \leq \alpha_0$*

$$E\left(Z_n^{l,u}[r^j, r^{j+1}]\right) \leq (1 + \epsilon/2)\alpha n \log r, \quad (74)$$

for all  $j \in J_\delta^\alpha$

In preparation for the proof of Proposition 9 put

$$\lambda_1 = \frac{1-\epsilon}{1-\epsilon/2} < 1 \quad \text{and} \quad \lambda_2 = \frac{1+\epsilon}{1+\epsilon/2} > 1.$$

Also let

$$C(\epsilon) = (1 - \epsilon/2) \min(c_{\lambda_1}, c_{\lambda_2}),$$

where  $c_x = x \log x - x + 1$  is defined in Lemma 4. Note that since  $\lambda_1$  and  $\lambda_2$  tend to 1 as  $\epsilon$  tends to 0,  $c_{\lambda_1}$  and  $c_{\lambda_2}$  tend to 0 as  $\epsilon$  tends to 0, and so we have  $C(\epsilon) \rightarrow 0$  as  $\epsilon \rightarrow 0$ .

**Proof of Proposition 9** As we indicated in (55) and (56), we have to show that given  $\epsilon$  positive there exists a  $\delta_0$  and  $\alpha_0$  such that for any  $\delta \leq \delta_0$  and  $\alpha \leq \alpha_0$

$$P\left(Y_n^{l,u}[r^j, r^{j+1}] < (1 - \epsilon)\alpha n \log r\right) \leq \exp(-\alpha n C(\epsilon)\alpha n \log r), \quad (75)$$

$$P\left(Z_n^{l,u}[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r\right) \leq \exp(-\alpha n C(\epsilon)\alpha n \log r), \quad (76)$$

for all  $j \in J_\delta^\alpha$ . Take  $\delta_0$  and  $\alpha_0$  so that they satisfy Lemma 13 and Lemma 12. Hence for any  $\delta \leq \delta_0$  and  $\alpha \leq \alpha_0$

$$(1 - \epsilon/2)\alpha n \log r \leq EY_n^{l,u}[r^j, r^{j+1}] \leq EZ_n^{l,u}[r^j, r^{j+1}] \leq (1 + \epsilon/2)\alpha n \log r, \quad (77)$$

for all  $j \in J_\delta^\alpha$ . To prove (75) we have, by (77),

$$\begin{aligned} P\left(Y_n^{l,u}[r^j, r^{j+1}] < (1 - \epsilon)\alpha n \log r\right) \\ \leq P\left(Y_n^{l,u}[r^j, r^{j+1}] < \lambda_1 EY_n^{l,u}[r^j, r^{j+1}]\right). \end{aligned} \quad (78)$$

Now, by Lemma 4, since  $Y_n^{l,u}[r^j, r^{j+1}]$  is Poisson and  $\lambda_1 < 1$ ,

$$P\left(Y_n^{l,u}[r^j, r^{j+1}] < \lambda_1 EY_n^{l,u}[r^j, r^{j+1}]\right) \leq \exp\left(-c_{\lambda_1} EY_n^{l,u}[r^j, r^{j+1}]\right), \quad (79)$$

and again by (77) and the definition of  $C(\epsilon)$ ,

$$\exp\left(-c_{\lambda_1} EY_n^{l,u}[r^j, r^{j+1}]\right) \leq \exp\left(-\alpha n C(\epsilon) \log r\right). \quad (80)$$

Putting together (78)- (80) yields (75).

To prove (76) we follow essentially the same argument and get, by (77), that

$$\begin{aligned} P\left(Z_n^{l,u}[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r\right) \\ \leq P\left(Z_n^{l,u}[r^j, r^{j+1}] > \lambda_2 EZ_n^{l,u}[r^j, r^{j+1}]\right). \end{aligned} \quad (81)$$

Now by Lemma 4, since  $Z_n^{l,u}[r^j, r^{j+1}]$  is Poisson and  $\lambda_2 > 1$ ,

$$P\left(Z_n^{l,u}[r^j, r^{j+1}] < \lambda_2 EZ_n^{l,u}[r^j, r^{j+1}]\right) \leq \exp\left(-c_{\lambda_2} EZ_n^{l,u}[r^j, r^{j+1}]\right), \quad (82)$$

and again by (77) and the definition of  $C(\epsilon)$ ,

$$\exp\left(-c_{\lambda_2} EZ_n^{l,u}[r^j, r^{j+1}]\right) \leq \exp\left(-\alpha n C(\epsilon) \log r\right). \quad (83)$$

Putting together (81)- (83) yields (76).  $\square$

The proof of Proposition 9 will be complete once we prove equations (59)-(62). We now proceed to prove (60)-(62) and then finish with (59).

**Proof of 60** Note that

$$\int_0^{l - \log r^j - H} P\left(X(l - s) \in [r^j, r^{j+1}]\right) ds \leq \int_0^{l - \log r^j - H} P\left(X(l - s) < r^{j+1}\right) ds.$$

By a change of variables,  $t = l - s$ , and then dividing through by  $e^t$  inside the probability, we get that

$$\int_0^{l - \log r^j - H} P\left(X(l - s) < r^{j+1}\right) ds = \int_{\log r^j + H}^l P\left(e^{-t} X(t) < r^{j+1} e^{-t}\right) dt.$$

Now, using Lemma 6,

$$\begin{aligned} \int_{\log r^j+H}^l P\left(e^{-t}X(t) < r^{j+1}e^{-t}\right) dt &\leq \int_{\log r^j+H}^l P\left(\inf_{s>0} e^{-s}X(s) < r^{j+1}e^{-t}\right) dt \\ &\leq \int_{\log r^j+H}^l 2r^{(j+1)/2}e^{-t/2} dt \leq 4r^{1/2}e^{-H/2}. \end{aligned}$$

□

**Proof of 61** Similar to the proof of equation 60. □

**Proof of 62** By a change of variables,  $t = l - \log r^j - s$ , and then dividing through by  $\exp(t + \log r^j)$  inside the probability, we have that

$$\begin{aligned} \int_{l-\log r^j-H}^{l-\log r^j+H} P\left(X(l-s) \in [r^j, r^{j+1}]\right) ds \\ = \int_{-H}^H P\left(e^{-t-\log r^j}X(t+\log r^j) \in [e^{-t}, re^{-t}]\right) dt. \end{aligned} \tag{84}$$

What we will now make precise is that

$$\begin{aligned} \int_{-H}^H P\left(e^{-t-\log r^j}X(t+\log r^j) \in [e^{-t}, re^{-t}]\right) dt &\approx \int_{-H}^H P\left(\xi(1) \in [e^{-t}, re^{-t}]\right) dt \\ &\approx \log r. \end{aligned} \tag{85}$$

For the last approximation, choose  $\epsilon_1$  so that

$$\log((1 + \epsilon_1)/(1 - \epsilon_1)) \leq \epsilon.$$

Now choose  $H_\epsilon$  so that by Lemma 7 for any  $H \geq H_\epsilon$ ,

$$\begin{aligned} \int_{-H-\log(1+\epsilon_1)}^{H-\log(1+\epsilon_1)} P\left(\xi(1) \in [e^{-t}, ((1 + \epsilon_1)/(1 - \epsilon_1))re^{-t}]\right) dt \\ \leq \log((1 + \epsilon)r/(1 - \epsilon)) + \epsilon \leq \log r + 2\epsilon. \end{aligned} \tag{86}$$

For the first approximation, let  $H \geq H_\epsilon$  be fixed. Choose  $\delta_0$  so that for any  $\delta \leq \delta_0$ , if  $t \geq \log \delta^{-1} - H$  then, by (38),

$$P\left(e^{-t}X(t) \in [e^{-t}, re^{-t}]\right) \leq P\left(\xi(1) \in [(1 + \epsilon_1)^{-1}e^{-t}, r(1 - \epsilon_1)^{-1}e^{-t}]\right) + \epsilon/2H.$$

Now for any  $j$  such that  $r^j \geq \delta^{-1}$ , we have that

$$\begin{aligned}
& \int_{-H}^H P\left(e^{-t-\log r^j} X(t + \log r^j) \in [e^{-t}, re^{-t}]\right) dt \\
& \leq \int_{-H}^H P\left(\xi(1) \in [(1 + \epsilon_1)^{-1}e^{-t}, r(1 - \epsilon_1)^{-1}e^{-t}]\right) dt + \epsilon \\
& = \int_{-H-\log(1+\epsilon_1)}^{H-\log(1+\epsilon_1)} P\left(\xi(1) \in [e^{-s}, ((1 + \epsilon_1)/(1 - \epsilon_1))re^{-s}]\right) ds + \epsilon,
\end{aligned} \tag{87}$$

where the last equality follows by a change of variables,  $s = -t - \log(1 + \epsilon_1)$ . Combining (84), (87), and (86) we get that

$$\int_{l-\log r^j-H}^{l-\log r^j+H} P\left(X(l-s) \in [r^j, r^{j+1}]\right) ds \leq \log r + 3\epsilon.$$

□

**Proof of 59** Similar to the proof of equation (62). □

#### 4 Proof of Theorem 1 Equation (8)

As in section 3,  $r > 1$  will be fixed. We also use the same definition of  $d_\alpha$ ,  $J_\delta^\alpha$ ,  $l$ ,  $u$ , and  $\Omega_\alpha$ . Let  $\lceil y \rceil$  denote the smallest integer greater than or equal to  $y$ . For  $\epsilon$  positive put  $s = (1 + \epsilon)\alpha n \log r$  and

$$x = (b(\epsilon))^{\lceil s \rceil} (2\pi \lceil s \rceil)^{-1/2}, \tag{88}$$

where  $b(\epsilon) = (1 - \epsilon)^2/(1 + \epsilon)$ . We now state Lemma 14 and Lemma 15, which will be proved at the end of the section. After the proof Lemma 14, in Section 4.1, we will point out why we need  $\alpha n/\log \log \bar{\alpha} \rightarrow 0$  in order to prove (8).

**Lemma 14** *Assume that  $\alpha n \rightarrow \infty$  and  $\alpha n/\log \log \bar{\alpha} \rightarrow 0$  as  $\alpha \rightarrow 0$ . Then with  $x$  given by (88) we have that*

$$\lim_{\alpha \rightarrow 0} (1 - x)^{\log \bar{\alpha}} = 0. \tag{89}$$

**Lemma 15** *Given  $\epsilon$  positive, there exists a  $\delta_0$  and  $\alpha_0$  such that for any  $\delta \leq \delta_0$  and  $\alpha \leq \alpha_0$*

$$P\left(Y_n^{l,u}[r^j, r^{j+1}] \leq (1 + \epsilon)\alpha n \log r\right) \leq 1 - x, \tag{90}$$

for all  $j \in J_\delta^\alpha$ .

The key to the proof of (8) is to show that

$$P \left( \bigcup_{J_\delta^\alpha} (Y_n^{l,u}[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r) \right) \quad (91)$$

$$= 1 - \prod_{J_\delta^\alpha} P(Y_n^{l,u}[r^j, r^{j+1}] \leq (1 + \epsilon)\alpha n \log r), \quad (92)$$

and obtain a lower bound for this, which will be done with Lemmas 14 and 15.

**Proof of 8** We will show that given any  $\epsilon$  positive, there exists a  $\delta_0$  such for any  $\delta \leq \delta_0$

$$\liminf_{\alpha \rightarrow 0} P \left( \bigcup_{J_\delta^\alpha} (K_n[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r) \right) \geq 1 - 2\epsilon, \quad (93)$$

for all  $j \in J_\delta^\alpha$ . The first step is that

$$\begin{aligned} & P \left( \bigcup_{J_\delta^\alpha} (K_n[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r) \right) \\ & \geq P \left( \bigcup_{J_\delta^\alpha} (Y_n^{l,u}[r^j, r^{j+1}] > (1 + \epsilon)\alpha n \log r) \right) - P(\Omega_\alpha^c) \end{aligned} \quad (94)$$

We use the half open interval  $[r^i, r^{i+1})$  so that  $Y_n^{l,u}[r^i, r^{i+1})$  and  $Y_n^{l,u}[r^j, r^{j+1})$  are independent Poisson random variables whenever  $i \neq j$ . Since the  $\{Y_n^{l,u}[r^j, r^{j+1})\}$  thin a Poisson process, to see that they are independent, we need to show that arrivals are counted only once. In other words, if

$$1\{X_m(u - T_m) < r^{j+1}; X_m(l - T_m) \geq r^j\} = 1 \quad (95)$$

then

$$1\{X_m(u - T_m) < r^{i+1}; X_m(l - T_m) \geq r^i\} = 0, \quad (96)$$

whenever  $i \neq j$ . Without loss of generality we may assume that  $i < j$ . If (95) holds then  $X_m(u - T_m) < r^{i+1}$ . But now  $X_m(l - T_m) \leq X_m(u - T_m) < r^{i+1} \leq r^j$ , and so  $X_m(l - T_m) \not\geq r^j$ . Hence (96) follows. From the independence of the  $\{Y_n^{l,u}[r^j, r^{j+1})\}$  we get (91).

Let  $\delta_0$  be given by Lemma 15 and fix  $\delta \leq \delta_0$ . Now take  $\alpha_0$  so that for any  $\alpha \leq \alpha_0$ , by Lemma 8,

$$P(\Omega_\alpha^c) \leq \epsilon, \quad (97)$$

by Lemma 15,

$$P(Y_n^{l,u}[r^j, r^{j+1}] \leq (1 + \epsilon)\alpha n \log r) \leq 1 - \epsilon \quad (98)$$

for all  $j \in J_\delta^\alpha$ , and, by Lemma 14,

$$(1-x)^{\log \bar{\alpha} + 2 \log \delta} \leq \epsilon^{\log r}. \quad (99)$$

By (94) and (97), we need to show that (91) is bounded below by  $1 - \epsilon$ , in order to prove (93). But now, by (98) and (99),

$$\begin{aligned} 1 - \prod_{J_\delta^\alpha} P(Y_n^{l,u}[r^j, r^{j+1}]) &\leq (1 + \epsilon) \alpha n \log r \\ &\geq 1 - \prod_{J_\delta^\alpha} (1 - x) \geq 1 - (1 - x)^{(\log \bar{\alpha} + 2 \log \delta) / \log r} \geq 1 - \epsilon. \end{aligned}$$

□

#### 4.1 Proof of Lemmas 14 and 15

Note that  $x$ , given in (88) tends to 0 as  $\alpha$  tends to 0, provide that  $\alpha n$  tends to infinity.

**Proof of Lemma 14** Notice that

$$x \log \bar{\alpha} = \exp \left( \log \log \bar{\alpha} \left[ 1 + \frac{[s] \log b(\epsilon)}{\log \log \bar{\alpha}} - \frac{\log(2\pi [s])}{2 \log \log \bar{\alpha}} \right] \right). \quad (100)$$

Since  $\alpha n / \log \log \bar{\alpha}$  tends to 0 as  $\alpha$  tends to 0, it follows that

$$\lim_{\alpha \rightarrow 0} \frac{[s] \log b(\epsilon)}{\log \log \bar{\alpha}} = 0 \quad \text{and} \quad \lim_{\alpha \rightarrow 0} \frac{\log(2\pi [s])}{2 \log \log \bar{\alpha}} = 0. \quad (101)$$

Hence

$$\lim_{\alpha \rightarrow 0} x \log \bar{\alpha} = \infty. \quad (102)$$

Now for  $x$  close to 0,

$$\log(1-x) \leq x^2 - x.$$

and so

$$(1-x)^{\log \bar{\alpha}} = \exp(\log \bar{\alpha} \log(1-x)) \leq \exp(\log \bar{\alpha} (x^2 - x)) = \exp((x-1)x \log \bar{\alpha}).$$

Therefore, since  $x \rightarrow 0$  as  $\alpha \rightarrow 0$  and by (102),

$$\lim_{\alpha \rightarrow 0} (1-x)^{\log \bar{\alpha}} = 0.$$

□

At this point, we can see why we have  $\alpha n / \log \log \bar{\alpha} \rightarrow 0$  in the hypothesis for (8) of Theorem 2. Notice that in order for (102) to hold, we need

$$1 + \frac{\lceil s \rceil \log b(\epsilon)}{\log \log \bar{\alpha}} - \frac{\log(2\pi \lceil s \rceil)}{2 \log \log \bar{\alpha}} > 0, \quad (103)$$

or, since  $b(\epsilon) < 1$ ,

$$\frac{\lceil s \rceil \log b^{-1}(\epsilon)}{\log \log \bar{\alpha}} + \frac{\log(2\pi \lceil s \rceil)}{2 \log \log \bar{\alpha}} < 1. \quad (104)$$

The numerators in (104) contain a  $\log r$ , which may be arbitrarily large. Thus, in order to assure (103) we take  $\alpha n / \log \log \bar{\alpha} \rightarrow 0$ , which results in (101).

**Proof of Lemma 15** Since  $Y_n^{l,u}[r^j, r^{j+1}]$  is a Poisson random variable,

$$P\left(Y_n^{l,u}[r^j, r^{j+1}] \leq s\right) \leq 1 - \sum_{i=\lceil s \rceil}^{\infty} \frac{e^{-\theta_j} \theta_j^i}{i!}, \quad (105)$$

where  $EY_n^{l,u}[r^j, r^{j+1}] = \theta_j$ . Recall that  $Y_n^{l,u}[r^j, r^{j+1}] \leq Z_n^{l,u}[r^j, r^{j+1}]$ . Chose  $\delta_0$  and  $\alpha_0$  so that for any  $\delta \leq \delta_0$  and  $\alpha \leq \alpha_0$ , by Lemma 13 and Lemma 12,

$$(1 - \epsilon/2)\alpha n \log r \leq \theta_j \leq (1 + \epsilon/2)\alpha n \log r,$$

for all  $j \in J_\delta^\alpha$ , and, by Stirling's formula,

$$\frac{e^{-\lceil s \rceil} \lceil s \rceil^{\lceil s \rceil}}{\lceil s \rceil!} \geq \frac{1 - \epsilon}{(2\pi \lceil s \rceil)^{1/2}}. \quad (106)$$

Note that with the notation here, we have that

$$(1 - \epsilon)s/(1 + \epsilon) \leq \theta_j \leq s. \quad (107)$$

Now, by (107),

$$\sum_{i=\lceil s \rceil}^{\infty} \frac{e^{-\theta_j} \theta_j^i}{i!} \geq \frac{e^{-\theta_j} \theta_j^{\lceil s \rceil}}{\lceil s \rceil!} \geq \frac{e^{-\lceil s \rceil} \theta_j^{\lceil s \rceil}}{\lceil s \rceil!} \geq \frac{e^{-\lceil s \rceil} \lceil s \rceil^{\lceil s \rceil}}{\lceil s \rceil!} \left(\frac{1 - \epsilon}{1 + \epsilon}\right)^{\lceil s \rceil}, \quad (108)$$

and the proof is finished by combining (105), (106), and (108).  $\square$

**Acknowledgment** I am grateful to J. T. Cox for all his time and effort as my advisor.

## References

- [1] Arratia, R., Barbour, A.D., and Tavaré, S. (1992) Poisson process approximations for the Ewens sampling Formula. *Ann. Appl. Probab.* **2**, 519-535

- [2] Billingsley, P. (1995) *Probability and Measure*. Wiley, New York
- [3] Bramson, M., Cox, J.T., and Durrett, R. (1996) Spatial models for species area curves. *Ann. Probab.* **24**, 1727-1751
- [4] Bramson, M., Cox, J.T., and Durrett, R. (1997) A spatially explicit model for the abundance of species. (preprint)
- [5] Bramson, M., Cox, J.T., and Durrett, R. (1998) A spatial model for the abundance of species. *Ann. Probab.* **26**, 658-709
- [6] Donnelly, P., Kurtz, T.G., and Tavaré, S. (1991) On the functional central limit theorem for the Ewens sampling formula. *Ann. Appl. Probab.* **1**, 539-545
- [7] Durrett, R. (1998) *Lecture Notes on Particle Systems and Percolation*. Wadsworth, Belmont, CA
- [8] Durrett, R. (1996) *Probability: Theory and Examples*. Duxbury Press, Belmont, CA
- [9] Griffeath, D. (1979) *Additive and Cancellative Interacting Particle Systems. Lecture Notes in Math.* **724**, Springer, Berlin.
- [10] Hansen, J.C. (1990) A functional central limit theorem for the Ewens sampling formula. *J. Appl. Probab.* **27**, 28-43
- [11] Karlin, S., Taylor, H.M. (1975) *A First Course in Stochastic Processes* 2nd edition. Academic Press, New York
- [12] Kelly, F.P. (1979) *Reversibility and Stochastic Networks*. Wiley, New York
- [13] Kingman, J.F.C. (1982) The coalescent. *Stoch. Proc. Appln.* **13**, 235-248
- [14] Liggett, T.M. (1985) *Interacting Particle Systems*. Springer, New York
- [15] Pfaff, T.J. (1999) *A Mean Field Model for Species Abundance* Ph.D. Thesis, Syracuse University, Syracuse, New York
- [16] Ross, S.M. (1996) *Stochastic Processes*. Wiley, New York
- [17] Tavaré, S. (1987) The birth process with immigration, and the genealogical structure of large populations. *J. Math. Biol.* **25**, 161-168
- [18] Tavaré, S., Ewens, W.J. (1997) Multivariate Ewens distribution. In *Discrete Multivariate Distributions* (N. L. Johnson, S. Kotz, N. Balakrishnan, eds) chp 41. Wiley, New York