Coalescing Times for IID Random Variables with Applications to Population Biology

Ilan Adler,[1] Hyun-Soo Ahn,[1] Richard M. Karp,[2] Sheldon M. Ross[1,*]
[1]*Department of IEOR, University of California, Berkeley, California 94720; e-mail: lastname@ieor.berkeley.edu*
[2]*Department of EECS, University of California, Berkeley, California 94720; e-mail: karp@cs.berkeley.edu.*

Received 26 February 2002; accepted 9 July 2002

DOI 10.1002/rsa.10080

ABSTRACT: We consider a coalescing particle model where particles move in discrete time. At each time period, each remaining ball is independently put in one of n bins according to a probability distribution $\boldsymbol{p} = (p_1, \ldots, p_n)$, and all balls put into the same bin merge into a single ball. Starting with k balls, we are interested in the properties of $E[N(\boldsymbol{p}, k)]$, the expected time until all balls merge into one. We derive both upper and lower bounds for $E[N(\boldsymbol{p}, k)]$, some asymptotic results, and show that $P\{N(\boldsymbol{p}, k) > t\}$, and thus $E[N(\boldsymbol{p}, k)]$, is a Schur concave function of $\boldsymbol{p}$. Applications to population biology are noted. © 2003 Wiley Periodicals, Inc. Random Struct. Alg., 23: 155–166, 2003

1. INTRODUCTION AND SUMMARY

We consider a discrete time stochastic process where balls (particles) coalesce in bins. In round 1, each of k balls is independently put in one of n bins, with each ball being put in bin i with probability p_i, where $\sum_{i=1}^{n} p_i = 1$. All balls put into the same bin coalesce into a single ball. If more than a single ball remains, the process is repeated until all balls are

Correspondence to: R. M. Karp
*Research supported by the National Science Foundation Grant DMI-9901053 with the University of California.
© 2003 Wiley Periodicals, Inc.

coalesced into one ball. Denoting $\boldsymbol{p} = (p_1, \ldots, p_n)$, we are interested in $E[N(\boldsymbol{p}, k)]$, the expected number of rounds needed until only a single ball remains.

Our model has applications in population biology. Most importantly, it relates to a generalized version of the Wright-Fisher model of population genetics, which considers the evolution of a population that has n individuals in each generation. Suppose that the members of each generation randomly arrange themselves in a "pecking order," which is such that each member of the succeeding generation is independently descended from the individual having pecking position i with probability p_i. (The Wright-Fisher model results when all $p_i = 1/n$.) If we consider this process in steady state, then it is easy to see that the number of generations one need go back in time to find the most recent common ancestor (MRCA) of a specified set of k members of the current generation is equal to the time until k balls coalesce into a single ball in our model.

The number of generations one need go back in time to find a common ancestor of a set of k individuals has previously been studied, both for the Wright-Fisher model [5], and also for a generalization that allows for parental selection as a function of the genetic makeup of the current population [6, 9], under the assumption that n is very large. In the Wright-Fisher model, it is shown in [5] that, when n is large, the number of generations is approximately distributed as the sum of $k - 1$ independent exponential random variables, with respective means $\frac{2n}{i(i-1)}$, $i = 2, \ldots, k$. In the selection models of [6] and [9], the numbers of offspring of the members of a generation has a distribution that, conditional on the genetic information for that generation, is multinomial. Conditions are presented in [6] and [9] under which the asymptotic distribution of the MRCA is the same as that just given for the Wright-Fisher model; numerical methods and simulation techniques for analyzing the MRCA in the general case are also presented in these papers. In addition, convergence results for the model considered in the present paper have previously appeared in [7]. For other literature in this field, one should see [3].

The problem considered is also related to the coalescing random-walk model [1, 2, 4, 10, 11]. In its most general setup (which we will call the coalescing Markov chain model), this model is similar to the one previously described except that the particles move among boxes according to a Markov chain having transition probabilities $P_{i,j}$. The particles move independently according to the transition probabilities of the chain, with the exception that whenever two or more particles are in the same bin they are coalesced into a single particle. The process is repeated until all particles are coalesced into one. Part of the motivation for our model is the equivalence of the coalescing time with the backward coupling time in the "coupling from the past" algorithm (see [10, 11]) that enables one to simulate from the stationary distribution of the Markov chain. In [11] an upper bound, in terms of the mixing time of the Markov chain, on the expected time until all particles have been coalesced into a single particle is given. However, applied to our problem, which is the special case of the coalescing Markov chain model in which $P_{i,j} = p_j$, the mixing time is 1 and the upper bound from [11] is $49n$, whereas we are able to establish the upper bound $2n$.

Other variants of the coalescing random-walk model have also been studied. For instance, as surveyed in [4], a number of papers have considered variants of coalescing random-walk models in which the state of the underlying Markov chain consists of the movements of a particle that moves on the vertices of a graph in such a manner that each succeeding vertex is equally likely to be any of the neighbors of the present vertex. These

models differ from ours, however, in assuming that in each round exactly one randomly chosen particle moves, whereas in our model every ball moves in each round.

In Section 2 we derive upper and lower bounds for $E[N(\boldsymbol{p}, k)]$. In Section 3 we establish its asymptotic behavior for the case in which all the probabilities are roughly the same. In Section 4 we consider the supremum value, over all n and probability distributions on the integers $1, \ldots, n$, of $\sum_j p_j^2 E[N(\boldsymbol{p}, k)]$. In Section 5 we show that $P\{N(\boldsymbol{p}, k) > t\}$, and thus $E[N(\boldsymbol{p}, k)]$, is a Schur-concave function of the probabilities $p_1, \ldots, p_n$.

2. BOUNDS FOR *E[N(p, k)]*

Clearly $E[N(\boldsymbol{p}, k)]$ is a nondecreasing function of k. Also, for $k > n$, $E[N(\boldsymbol{p}, k)] \leq E[N(\boldsymbol{p}, n)] + 1$, since the number of balls remaining after the first round is at most n. Hereafter we restrict attention to the case $k \leq n$.

Let us denote $p_i := (\sum_{j=1}^n p_j^i)^{-1}$ and $\rho := \frac{\rho_2}{\rho_3}$. Note that $\frac{1}{n} \leq \rho \leq 1$, where the lower bound is obtained whenever all the bin probabilities are equal, while the upper bound corresponds to the case in which one bin has a probability of 1. Thus ρ can be viewed as a measure of the diversity of the distribution of the bin probabilities. It should also be noted that $1/\rho_2$ is the probability of merging two given balls, and that $E[N(\boldsymbol{p}, k)] \geq E[N(\boldsymbol{p}, 2)] = \rho_2$. We start by constructing an upper bound for $E[N(\boldsymbol{p}, k)]$.

Theorem 1. *For* $3 \leq k \leq n$,

$$E[N(\boldsymbol{p}, k)] \leq \rho_2 \left(H_k - \frac{1}{2} \right),$$

where H_k *is the kth harmonic number.*

Proof. Define q_k as the probability that starting with k balls, a particular ball (say ball 1) is still unmerged when all the other balls had been merged. Thus,

$$E[N(\boldsymbol{p}, k)] = E[N(\boldsymbol{p}, k-1)] + E[N(\boldsymbol{p}, 2)]q_k = E[N(\boldsymbol{p}, k-1)] + \rho_2 q_k. \tag{1}$$

Solving recursively, the preceding equation leads to

$$E[N(\boldsymbol{p}, k)] = \rho_2 \sum_{i=2}^{k} q_i. \tag{2}$$

Clearly, $q_2 = 1$ while for $k > 2$, $q_k \leq 1/k$ because at most one of these balls can be unmerged when all the others have been merged, and that ball is equally likely to be any of the first k. Consequently, using Eq. (2), we have that, for $3 \leq k \leq n$,

$$E[N(\boldsymbol{p}, k)] \leq \rho_2 \left(H_k - \frac{1}{2} \right).$$

∎

Next, we develop tighter lower and upper bounds for the case in which ρ is below some threshold depending on k.

We begin by developing a lower bound for q_k which, combined with Eq. (2), establishes a lower bound for $E[N(\boldsymbol{p}, k)]$. We shall need the following lemma.

Lemma 1. *Let $P_{i,j}$ be the transition probability of the Markov chain in which the state is the number of remaining balls, and each round corresponds to a state transition. Then*

(i) $1 - P_{k,k} \leq \frac{k(k-1)}{2\rho_2}$,

(ii) $\frac{P_{k,k-1}}{1 - P_{k,k}} \geq 1 - \frac{(k-1)(k-2)\rho}{2}$.

Proof.

(i) $1 - P_{k,k} = P\{\text{at least one merge given } k \text{ balls}\} \leq \binom{k}{2} \sum_{j=1}^{n} p_j^2 = \binom{k}{2} \frac{1}{\rho_2}$.

(ii) Let $M(i)$ be the event that, in a round with i balls, some bin receives at least two balls. Let $B_j(i)$ be the event that, in a round with i balls, bin j receives at least one ball.

$$
\begin{aligned}
P_{k,k-1} &= \binom{k}{2} \sum_{j=1}^{n} p_j^2 (1 - P\{M(k-2) \cup B_j(k-2)\}) \\
&\geq \binom{k}{2} \sum_{j=1}^{n} p_j^2 \left(1 - \binom{k-2}{2} \sum_{j=1}^{n} p_j^2 - (k-2)p_j\right) \\
&= \frac{1}{\rho_2} \binom{k}{2} \left(1 - \frac{1}{\rho_2} \binom{k-2}{2} - \frac{(k-2)\rho_2}{\rho_3}\right) \\
&\geq \frac{1}{\rho_2} \binom{k}{2} \left(1 - \frac{\rho_2}{\rho_3} \binom{k-2}{2} - \frac{(k-2)\rho_2}{\rho_3}\right) \\
&= \frac{1}{\rho_2} \binom{k}{2} \left(1 - \frac{(k-1)(k-2)\rho}{2}\right), \qquad (3)
\end{aligned}
$$

where (3) is justified by considering a random variable Y with probability function $P(Y = p_j) = p_j$ and using $\rho_3^{-1} = E(Y^2) \geq E^2(Y) = \rho_2^{-2}$. Dividing the preceding by $1 - P_{k,k}$ and using (**i**) completes the proof. ∎

Lemma 2.

$$
q_k \geq \frac{2}{k(k-1)} \left(1 - \frac{(k-2)(k-1)k\rho}{6}\right). \qquad (4)
$$

Proof. Clearly,

$$
q_k \geq P_{k,k} q_k + \frac{k-2}{k} P_{k,k-1} q_{k-1},
$$

which, by using Lemma 1, leads to

$$q_k \geq \frac{(k-2)P_{k,k-1}}{k(1-P_{k,k})}\, q_{k-1} \geq \frac{k-2}{k}\left(1 - \frac{(k-1)(k-2)\rho}{2}\right) q_{k-1}. \tag{5}$$

Noting that $q_2 = 1$, it is clear that (4) holds for $k = 2$. Assuming the bound holds for $k - 1$, and using (5), we get

$$\begin{aligned} q_k &\geq \frac{k-2}{k}\left(1 - \frac{(k-1)(k-2)\rho}{2}\right)\frac{2}{(k-2)(k-1)}\left(1 - \frac{(k-3)(k-2)(k-1)\rho}{6}\right) \\ &\geq \frac{2}{k(k-1)}\left(1 - \frac{[3(k-1)(k-2) + (k-3)(k-2)(k-1)]\rho}{6}\right) \\ &= \frac{2}{k(k-1)}\left(1 - \frac{(k-2)(k-1)k\rho}{6}\right) \end{aligned}$$

■

Theorem 2. $E[N(\boldsymbol{p}, k)] \geq 2\rho_2\left(1 - \frac{1}{k} - \frac{(k-1)(k-2)\rho}{12}\right)$.

Proof. Noting that $E[N(\boldsymbol{p}, 2)] = \rho_2$, it is clear that the preceding inequality holds for $k = 2$. Assuming the bound holds for $k - 1$, by (1) and by Lemma 2,

$$\begin{aligned} E[N(\boldsymbol{p}, k)] &= E[N(\boldsymbol{p}, k-1)] + \rho_2 q_k \\ &\geq 2\rho_2\left[\left(1 - \frac{1}{k-1} - \frac{(k-2)(k-3)\rho}{12}\right) + \frac{1}{k(k-1)}\left(1 - \frac{(k-2)(k-1)k\rho}{6}\right)\right] \\ &= 2\rho_2\left[\left(1 - \frac{1}{k-1} - \frac{(k-2)(k-3)\rho}{12}\right) + \frac{1}{k-1} - \frac{1}{k} - \frac{2(k-2)\rho}{12}\right] \\ &= 2\rho_2\left(1 - \frac{1}{k} - \frac{(k-1)(k-2)\rho}{12}\right). \end{aligned}$$

■

We now construct a tighter upper bound for $E[N(\boldsymbol{p}, k)]$.

Theorem 3. *If* $\rho < \frac{3}{k}$, *then* $E[N(\boldsymbol{p}, k)] < 2\rho_2\left[1 - \frac{1}{k} + \frac{\rho}{3}\left(H_k - 1 - \ln\left(1 - \frac{k\rho}{3}\right)\right)\right]$.

To prove the theorem, we again consider the process as a Markov chain in which the state is the number of remaining balls, and each round corresponds to a state transition. In each transition the state decreases or remains unchanged. Let $d(k)$ be the expected decrease in a transition from state k. Because $d(k)$ is easily shown to be nondecreasing in k, we obtain the following lemma.

Lemma 3 (4.5.2 on p. 124 of [13]).

$$E[N(\boldsymbol{p}, k)] \leq \sum_{i=2}^{k} \frac{1}{d(i)}.$$

Let us denote $b(k) = \sum_{i=2}^{k} \frac{1}{d(i)}$.

Lemma 4. *If* $\rho < \frac{3}{k}$, *then* $b(k) < 2\rho_2\left(1 - \frac{1}{k} + \frac{\rho}{3}\left(H_k - 1 - \ln\left(1 - \frac{k\rho}{3}\right)\right)\right)$.

Theorem 3 follows immediately from the preceding two lemmas. We now prepare for the proof of Lemma 4. In a round with i balls the probability that bin j is unoccupied is $(1 - p_j)^i$. Therefore,

$$d(i) = i - \left[n - \sum_{j=1}^{n} (1 - p_j)^i\right].$$

To prove Lemma 4, we need the following lemma.

Lemma 5. *For any positive integer* m *and* $0 \le p \le 1$,

$$(1 - p)^m \ge 1 - mp + \binom{m}{2}p^2 - \binom{m}{3}p^3.$$

Proof. Consider the events $A_1, \ldots, A_m$ representing independent Bernoulli trials with probability p of success. It follows from the inclusion–exclusion bounds that

$$P\left\{\bigcup_{1}^{m} A_i\right\} \le \sum_{i} P\{A_i\} - \sum_{i<j} P\{A_iA_j\} + \sum_{i<j<k} P\{A_iA_jA_k\},$$

which implies that

$$1 - (1 - p)^m \le mp - \binom{m}{2}p^2 + \binom{m}{3}p^3.$$

■

Proof of Lemma 4. Applying Lemma 5 to $d(i)$ and simplifying, we get

$$b(k) \le \sum_{i=2}^{k} \frac{2\rho_2}{i(i-1)\left(1 - \frac{(i-2)\rho}{3}\right)}.$$

Since $\frac{k\rho}{3} < 1$, we can expand the denominator as a power series, giving

$$b(k) \le 2\rho_2 \sum_{i=2}^{k} \frac{1}{i(i-1)} \sum_{t=0}^{\infty} \left(\frac{(i-2)\rho}{3}\right)^t$$

Reversing the order of summation, we obtain

$$b(k) \le 2\rho_2 \sum_{t=0}^{\infty} \sum_{i=2}^{k} \frac{1}{i(i-1)} \left(\frac{(i-2)\rho}{3} \right)^t \le 2\rho_2 (T_1 + T_2 + T_3),$$

where

$$T_1 = \sum_{i=2}^{k} \frac{1}{i(i-1)} = 1 - \frac{1}{k},$$

$$T_2 = \sum_{i=2}^{k} \frac{(i-2)\rho}{i(i-1)3} < \frac{\rho}{3} \sum_{i=2}^{k} \frac{1}{i} = \frac{\rho}{3} (H_k - 1),$$

$$T_3 = \sum_{t=2}^{\infty} \sum_{i=2}^{k} \frac{1}{i(i-1)} \left(\frac{(i-2)\rho}{3} \right)^t < \sum_{t=2}^{\infty} \sum_{i=2}^{k} (i-2)^{t-2} \left(\frac{\rho}{3} \right)^t.$$

But

$$\sum_{i=2}^{k} (i-2)^{t-2} < \int_{2}^{k-1} x^{t-2}\, dx < \frac{k^{t-1}}{t-1},$$

giving

$$T_3 < \frac{\rho}{3} \sum_{t=2}^{\infty} \frac{1}{t-1} \left(\frac{k\rho}{3} \right)^{t-1} = \frac{\rho}{3} \left[-\ln\left(1 - \frac{k\rho}{3} \right) \right].$$

■

3. THE ASYMPTOTIC BEHAVIOR OF *E[N(p, k)]* FOR LARGE *n*

We now consider a sequence $\{\boldsymbol{p}(n),\ n = 1, 2, \dots\}$ of distributions, where $\boldsymbol{p}(n)$ has support $\{1, 2, \dots, n\}$. In this setting we write $\boldsymbol{p}(n) = (p_1(n), \dots, p_n(n))$, $\rho_i(n)$, $\rho(n)$, and $E[N(\boldsymbol{p}(n), k(n))]$ instead of $\boldsymbol{p}$, ρ_i, ρ, and $E[N(\boldsymbol{p}, k)]$, respectively. We derive our asymptotic results under a condition that is intended to capture the property that the $p_j(n)$, for $j = 1, 2, \dots, n$, are roughly equal.

Theorem 4. *If there exists a constant* $\alpha < 1$ *such that, for all* n, $\rho(n) < \frac{3\alpha}{n}$, *then, as* $n \to \infty$:

(i) *For a fixed* k, $\frac{E[N(\boldsymbol{p}(n), k)]}{\rho_2(n)}$ *is asymptotic to* $2(1 - \frac{1}{k})$.

(ii) *For* $n^{\epsilon} \le k(n) \le n$ (*for some* $\epsilon < 1$), $\frac{E[N(\boldsymbol{p}(n), k(n))]}{\rho_2(n)}$ *is asymptotic to* 2.

Proof.

(i) Follows directly from Theorems 2 and 3.
(ii) Applying Theorem 2, we have

$$\frac{E[N(\boldsymbol{p}(n), k(n))]}{\rho_2(n)} \geq \frac{E[N(\boldsymbol{p}(n), \lfloor k(n)^{1/3} \rfloor)]}{\rho_2(n)} \geq 2\left(1 - \frac{1}{\lfloor k(n)^{1/3} \rfloor} - \frac{k(n)^{2/3}\rho(n)}{2}\right)$$

$$\geq 2\left(1 - \frac{1}{\lfloor k(n)^{1/3} \rfloor} - \frac{3\alpha k(n)^{2/3}}{2n}\right).$$

Combining the preceding inequality with Theorem 3 completes the proof. ∎

Remarks. Theorem 4 was proven in [5] for the equally likely case $p_i(n) = 1/n$ [e.g., $\rho(n) = 1/n$]. Theorem 4(**i**) was previously proven in [7] for the case $\sup_n \sup_{1 \leq i \leq n} np_i(n) < \infty$. The preprint [8] shows that 4(**i**) holds for a large class of exchangeable models provided that $\lim_{n \to \infty} \rho(n) = 0$.

4. SUPREMUM OF $E[N(\boldsymbol{p}, k)]/\rho_2$

It was established in the preceding section that whenever the $p_j(n)$ are roughly even, $\frac{E[N(\boldsymbol{p}(n), k)]}{\rho_2(n)}$ is much smaller compared to the general upper bound given by Theorem 1. Next we consider the problem of finding

$$s_k := \sup_{\boldsymbol{p}} \frac{E[N(\boldsymbol{p}, k)]}{\rho_2},$$

where k is fixed, and $\boldsymbol{p}$ is a probability distribution whose support is the positive integers.

Theorem 5.

$$s_k \geq .407 H_k.$$

Proof. Consider the distribution where $p_1 = p$, and all other p_i are arbitrarily small. Letting X_i denote the number of rounds that it takes to put ball i in bin 1 for the first time, it follows, because merges can be assumed to occur only in bin 1, that

$$E[N(\boldsymbol{p}, k)] \geq E[\max\{X_1, \ldots, X_k\}]. \tag{6}$$

Noting that the right-hand side of the preceding is the expected value of the maximum of k independent geometric random variables with common parameter p, let $Y_1, \ldots, Y_k$ be independent exponential random variables with mean 1, and let $c = -\ln(1 - p)$ (so $1 - e^{-c} = p$). If we now define W_i, $i = 1, \ldots, k$ by

$$W_i = j \qquad \text{if } (j-1)c \leq Y_i < jc,$$

then the W_i are independent geometric random variables with common parameter p. Hence,

$$\frac{1}{c}E[\max(Y_1, \ldots, Y_k)] \leq E[\max(W_1, \ldots, W_k)] < \frac{1}{c}E[\max(Y_1, \ldots, Y_k)] + 1.$$

Therefore,

$$\frac{H_k}{-\ln(1-p)} \leq E[\max(X_1, \ldots, X_k)] < \frac{H_k}{-\ln(1-p)} + 1.$$

Consequently,

$$s_k \geq \frac{p^2 H_k}{-\ln(1-p)}.$$

The right side of the preceding is maximized by choosing p to satisfy

$$2(1-p)\ln(1-p) + p = 0.$$

A numerical solution is $p \approx .715$, giving the result that $s_k \geq .407 H_k$. ∎

We conclude this section by showing that, for the class of probability distributions considered in the proof of Theorem 5, the inequality (6) is close to being an equality.

Theorem 6. *For the distribution having $p_1 = p$, and arbitrarily small p_i for $i > 1$, there is a constant C such that*

$$E[N(\boldsymbol{p}, k)] \leq E[\max\{X_1, \ldots, X_k\}] + \rho_2(-\ln(p) + C), \tag{7}$$

where $X_1, \ldots, X_k$ are as given in Theorem 5.

Proof. To prove the result, define

- $N_0 :=$ $\max\{X_1, \ldots, X_k\}$, (the first round in which every initial ball—possibly as part of a merged ball—had landed in bin 1 at least once.)
- $K :=$ the random variable corresponding to the number of distinct balls in the process at N_0.
- $I(j) :=$ the indicator variable for the event that at least one ball landed in bin 1 in round $N_0 - j$ and that ball did not again land in bin 1 in any of the rounds $N_0 - j + 1, \ldots, N_0$.

Thus,

$$E(K) = E\left[\sum_{j=0}^{\infty} I(j)\right] \leq \sum_{j=0}^{\infty} (1-p)^j = \frac{1}{p},$$

where the first inequality in the preceding is justified by noting that, given N_0, a ball's future movements after landing in bin 1 are iid according to the p_i.

Considering the preceding, and that Theorem 1 implies that there exists a C such that $E[N(\boldsymbol{p}, k)] \leq \rho_2(\ln(k) + C)$, we have

$$\begin{aligned} E[N(\boldsymbol{p}, k)] &= E[\max\{X_1, \ldots, X_k\}] + E\{E_K[N(\boldsymbol{p}, K)]\} \\ &\leq E[\max\{X_1, \ldots, X_k\}] + E[\rho_2(\ln(K) + C)] \\ &\leq E[\max\{X_1, \ldots, X_k\}] + \rho_2(\ln(E[K]) + C) \\ &\leq E[\max\{X_1, \ldots, X_k\}] + \rho_2[-\ln(p) + C], \end{aligned}$$

where the second inequality is justified by Jensen's Inequality. ∎

5. SCHUR CONCAVITY

For a probability vector $\boldsymbol{p} = (p_1, \ldots, p_n)$, let $p_{[i]}$ be the i^{th} largest value of $p_1, \ldots, p_n$. Say that the probability vector $\boldsymbol{p}$ *majorizes* the probability vector $\boldsymbol{q}$ if

$$\sum_{i=1}^{j} p_{[i]} \geq \sum_{i=1}^{j} q_{[i]}, \qquad \text{for all } j = 1, \ldots, n.$$

The symmetric function $f(\boldsymbol{p})$, defined on probability vectors, is said to be a *Schur concave* (*convex*) function if whenever $\mathbf{p}$ majorizes $\mathbf{q}$

$$f(\mathbf{p}) \leq (\geq) f(\mathbf{q}).$$

Assuming differentiability of f, a necessary and sufficient condition for f to be Schur concave (convex) is that

$$(p_1 - p_2)\left(\frac{\partial f(\boldsymbol{p})}{\partial p_1} - \frac{\partial f(\boldsymbol{p})}{\partial p_2}\right) \leq (\geq)\, 0 \tag{8}$$

Theorem 7. *The function*

$$f(\boldsymbol{p}) = P\{N(\boldsymbol{p}, k) > t\}$$

is a Schur concave function.

Proof. Consider two probability vectors $\boldsymbol{p} = (p_1, \ldots, p_n)$ and $\boldsymbol{p'} = (p'_1, \ldots, p'_n)$, where $p'_1 + p'_2 = p_1 + p_2$, $p'_j = p_j$ $(j = 3, \ldots, n)$. That is, the two probability vectors differ only in their values for p_1 and p_2. Moreover, suppose that $\min(p'_1, p'_2) < \min(p_1, p_2)$. We will now show that if $k' \leq k$, we can simulate $N(\boldsymbol{p}, k)$ and $N(\boldsymbol{p'}, k')$ in such a manner that $N(\boldsymbol{p'}, k') \leq N(\boldsymbol{p}, k)$. This will prove that

$$P\{N(\boldsymbol{p}', k') > t\} \leq P\{N(\boldsymbol{p}, k) > t\},$$

which is sufficient to establish Schur concavity (see [12]).

In doing the simulation, call the k balls $1, \ldots, k$, and call the k' balls $1', \ldots, k'$. Let $a = p_1 + p_2 = p'_1 + p'_2$. The simulation is begun by generating k' random numbers $U_1, \ldots, U_{k'}$. Put balls i and i' in bin j, $j > 2$, whenever

$$\sum_{i=1}^{j-1} p_i < U_i \leq \sum_{i=1}^{j} p_i.$$

Now consider those random numbers that were less than or equal to a; suppose there were r of them. The conditional probability that the r unprimed balls will all go into the same bin (either 1 or 2) is

$$\frac{p_1^r + p_2^r}{a^r},$$

whereas the conditional probability that the r primed balls will all go into the same bin is

$$\frac{(p'_1)^r + (p'_2)^r}{a^r}.$$

Because $p^r + q^r$ is, when $q = 1 - p$, a Schur convex function of p, q (easily checked by the condition 8), it follows that

$$\frac{(p'_1)^r + (p'_2)^r}{a^r} \geq \frac{p_1^r + p_2^r}{a^r}$$

Hence, the number of bins 1 and 2 that contain at least one primed ball is stochastically smaller than the number that contain at least one unprimed ball. Consequently, we can generate the locations of the r primed and unprimed balls so that the number of bins that contain at least one of the r primed balls is no greater than the number that contain at least one of the r unprimed balls. At this point we can generate the locations of the additional $k - k'$ unprimed balls. Now merge all the primed balls that are in the same bin and merge all the unprimed ones that are in the same bin, and then simulate the next round in the same manner. Because, at the beginning of the next round, the number of primed balls will be less than or equal the number of unprimed balls, it follows that this will remain true throughout the simulation, implying that the simulated value of $N(\boldsymbol{p}', k')$ will be less than or equal to that of $N(\boldsymbol{p}, k)$. ∎

Because X stochastically larger than Y implies that $E[X] \geq E[Y]$, we obtain the following corollary to Theorem 7.

Corollary 1. *$E[N(\boldsymbol{p}, k)]$ is a Schur concave function of $\boldsymbol{p}$, and is thus maximized when $p_i = 1/n$, $i = 1, \ldots, n$.*

ACKNOWLEDGMENTS

We thank the referees for pointing out the connection of our model with the coalescent models of population biology.

REFERENCES

[1] D. J. Aldous and J. A. Fill, Reversible Markov chains and random walks on graphs, to appear.

[2] M. Bramson and D. Griffeath, Clustering and dispersion rates for some interacting particle systems on Z_1, Ann Probab 8(2) (1980), 183–213.

[3] Y. X. Fu and W. H. Li, Coalescing into the 21st Century: An overview and prospects of coalescent theory, Theor Pop Biol 56 (1999), 1–10.

[4] D. Griffeath, Frank Spitzer's pioneering work on interacting particle systems, Ann Probab 21(2) (1993), 608–621.

[5] J. F. C. Kingman, "On the genealogy of large populations," Essays in statistical science, Editors Gani and Moran, Applied Probability Trust, Sheffield, J Appl Probab, Special Volume 19A (1982), 27–43.

[6] S. M. Krone and C. Neuheuser, Ancestral processes with selection, Theor Pop Biol 51 (1997), 210–237.

[7] M. Möhle, Robustness results for the coalescent, J Appl Probab 35 (1998), 438–447.

[8] M. Möhle, The time back to the most recent common ancestor in exchangeable population models, unpublished manuscript, 2002.

[9] C. Neuheuser and S. M. Krone, The genealogy of samples in models with selection, Genetics 145 (1997), 519–534.

[10] J. G. Propp and D. B. Wilson, Exact sampling with coupled Markov chains and applications to statistical mechanics, Random Struct Alg 9(2) (1996), 223–252.

[11] J. G. Propp and D. B. Wilson, How to get a perfectly random sample from a generic Markov chain and generate a random spanning tree of a directed graph, J Alg 27 (1998), 170–217.

[12] F. Proschan, "Applications of majorization and Schur functions in reliability and life testing," Reliability and fault tree analysis, 237–258. SIAM, Philadelphia, 1975.

[13] S. M. Ross, Probability models for computer science, Academic Press, New York, 2002.