A CLASSIFICATION OF COALESCENT PROCESSES FOR HAPLOID EXCHANGEABLE POPULATION MODELS

MARTIN MÖHLE, Johannes Gutenberg-Universität, Mainz and
SERIK SAGITOV[1], Chalmers and Göteborgs Universities, Göteborg

Abstract

We consider a class of haploid population models with non-overlapping generations and fixed population size N assuming that the family sizes within a generation are exchangeable random variables. A weak convergence criterion is established for a properly scaled ancestral process as $N \to \infty$. It results in a full classification of the coalescent generators in the case of exchangeable reproduction. In general the coalescent process allows for simultaneous multiple mergers of ancestral lines.

1 Introduction

Consider the class of haploid population models with non-overlapping generations and fixed population size $N \in \mathbb{N} := \{1, 2, \ldots\}$ introduced by Cannings (1974, 1975). Each model in this class is characterized by an

$$\text{exchangeable joint distribution of family sizes } \nu_1, \ldots, \nu_N, \tag{1}$$

where ν_i denotes the number of offspring of the i-th individual. Recall that according to (1) the distribution of the random vector $(\nu_{i_1}, \ldots, \nu_{i_k})$ with pairwise distinct indices depends only upon k and not upon the particular set of indices. As the population size is fixed the condition

$$\nu_1 + \cdots + \nu_N \;=\; N \tag{2}$$

has to be satisfied.

We are interested in the asymptotics of the genealogical structure in such a population in the spirit of Kingman (1982a,b,c). Fix $n \leq N$ and sample n individuals at random from the 0-th generation. Let $\mathcal{R}_r$ denote the equivalence relation which contains the pair (i, j) iff the i-th and the j-th individual of this sample have a common ancestor in the r-th generation backwards in time, $r \in \mathbb{N}_0 := \{0, 1, 2, \ldots\}$. The process $(\mathcal{R}_r)_{r \in \mathbb{N}_0}$ is a time homogeneous Markov chain with the state space

$$\mathcal{E}_n = \text{the set of all equivalence relations on } \{1, \ldots, n\}$$

[1]Supported by the Bank of Sweden Tercentenary Foundation project "Dependence and Interaction in Stochastic Population Dynamics"

AMS 1991 subject classifications. Primary 92D25, 60J70; Secondary 92D15, 60F17.

Key words and phrases. Ancestral processes, coalescent, exchangeability, generator, neutrality, population genetics, weak convergence.

and the initial value $\mathcal{R}_0 = \xi_0$,

$$\xi_0 = \{\text{all equivalence classes are singletons } \{i\},\ i = 1, \ldots, n\}. \tag{3}$$

Since the transition probability $p_{\xi\eta} := P(\mathcal{R}_r = \eta \,|\, \mathcal{R}_{r-1} = \xi)$ is equal to zero for $\xi \not\subseteq \eta$, the focus will be on such pairs $\xi, \eta \in \mathcal{E}_n$ that $\xi \subseteq \eta$. The relation $\xi \subseteq \eta$ implies that every equivalence class of η is either a union of several equivalence classes of ξ or coincides with an equivalence class of ξ. Reflecting this observation write a for the number of η-classes and $b = b_1 + \cdots + b_a$ for the number of ξ-classes, where $b_1 \geq \ldots \geq b_g \geq 2$ are ordered group sizes for merging ξ-classes and $b_{g+1} = \ldots = b_a = 1$. Notice that $g = 0$ if $\xi = \eta$ and $g \geq 1$ if $\xi \subset \eta$. In this notation the transition probability of the ancestral process is given by

$$p_{\xi\eta} = \frac{1}{(N)_b} \sum_{\substack{i_1, \ldots, i_a = 1 \\ \text{all distinct}}}^{N} \mathrm{E}((\nu_{i_1})_{b_1} \cdots (\nu_{i_a})_{b_a}) = \frac{(N)_a}{(N)_b} \mathrm{E}((\nu_1)_{b_1} \cdots (\nu_a)_{b_a}), \tag{4}$$

where $(N)_b := N(N-1)\cdots(N-b+1)$.

Let c_N denote the probability that two individuals, chosen randomly without replacement from some generation, have a common ancestor one generation backwards in time, i.e.

$$c_N := \frac{1}{(N)_2} \sum_{i=1}^{N} \mathrm{E}((\nu_i)_2) = \frac{\mathrm{E}((\nu_1)_2)}{N-1} = \frac{\mathrm{Var}(\nu_1)}{N-1} = 1 - \mathrm{E}(\nu_1\nu_2). \tag{5}$$

This probability, called the *coalescence probability* is of fundamental interest in the coalescent theory as c_N^{-1} is the proper time scale to get convergence to the coalescent (it is only natural to assume that $c_N > 0$ for sufficiently large N because the case $c_N = 0$ corresponds to the trivial reproduction law: $P(\nu_1 = 1, \ldots, \nu_N = 1) = 1$). The coalescence probability is also important as it is directly connected via $c_N = 1 - \lambda_2$ to the eigenvalue $\lambda_2 := \mathrm{E}(\nu_1\nu_2)$ of the transition matrix of the descendant process, i.e. the genealogical process looking forwards in time (see Cannings (1974)).

Kingman (1982b) has shown that given $\sup_N \mathrm{E}(\nu_1^k) < \infty$, $k \geq 2$ (this holds for example for the Moran model and the Wright-Fisher model) the convergence of finite-dimensional distributions

$$(\mathcal{R}_{[t/c_N]})_{t\geq 0} \to (R_t)_{t\geq 0},\ N \to \infty \tag{6}$$

takes place. The limit process $(R_t)_{t\geq 0}$, the so-called (standard) n-coalescent process, is a continuous time Markov process with state space $\mathcal{E}_n$, initial state (3) and infinitesimal generator $Q = (q_{\xi\eta})_{\xi,\eta\in\mathcal{E}_n}$ given by

$$q_{\xi\eta} := \begin{cases} -|\xi|(|\xi|-1)/2 & \text{if } \xi = \eta, \\ 1 & \text{if } \xi \prec \eta, \\ 0 & \text{otherwise,} \end{cases} \tag{7}$$

where $\xi \prec \eta$ means that $\xi \subset \eta$, $g = 1$ and $b_1 = 2$, i.e. during the transition exactly two ancestral lines merge together.

The convergence (6) is based on the asymptotic formula:

$$p_{\xi\eta} \;=\; \delta_{\xi\eta} + c_N\, q_{\xi\eta} + o(c_N),\; N \to \infty \quad \xi, \eta \in \mathcal{E}_n$$

which is often written in matrix notation

$$P_N \;=\; I + c_N\, Q + o(c_N),\; N \to \infty, \tag{8}$$

where $P_N := (p_{\xi\eta})_{\xi,\eta\in\mathcal{E}_n}$ denotes the transition matrix of the ancestral process. In Möhle (1998, 1999) Kingman's result was extended beyond the framework of exchangeable population models and it was shown that (6) holds even in the sense of the weak convergence of stochastic processes.

Recently a richer class of the coalescent generators Q allowing for multiple mergers with $g = 1$ and $b_1 \geq 2$ was found in Sagitov (1999). A member Q of this class is characterized by a probability measure $F(dx)$ on the unit interval [0,1] via the formula

$$q_{\xi\eta} \;=\; \begin{cases} -\displaystyle\int_{[0,1]} \frac{1-(1-x)^{b-1}(1-x+bx)}{x^2}\, F(dx), & \text{if } \xi = \eta, \\ \displaystyle\int_{[0,1]} x^{b_1-2}(1-x)^{b-b_1}\, F(dx), & \text{if } \xi \subset \eta,\; g = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The generator (7) of the standard n-coalescent is recovered from (9) when the probability measure $F = \delta_0$ is concentrated in zero.

The present paper is based upon an instrumental development of Möhle and Sagitov (1998) of the method of Sagitov (1999). We establish a general coalescent structure allowing for simultaneous mergers of ancestral lines ($g \geq 1$). Due to the main result of this paper, Theorem 2.1, in general, a coalescent generator $Q = (q_{\xi\eta})_{\xi,\eta\in\mathcal{E}_n}$ is characterized by a sequence of symmetric measures F_r, $r \in I\!N$, where each F_r is concentrated on the simplex

$$\Delta_r \;:=\; \{(y_1, \ldots, y_r) \in [0,1]^r \,|\, y_1 + \cdots + y_r \leq 1\}$$

with

$$1 = F_1(\Delta_1) \geq F_2(\Delta_2) \geq \ldots \tag{10}$$

If $\xi \subset \eta$, then the corresponding entry has the form

$$q_{\xi\eta} \;=\; \sum_{r=g}^{[(a+g)/2]} \int_{\Delta_r} x_1^{b_1-2} \cdots x_g^{b_g-2}\, T^{(r)}_{g,a-g}(x_1, \ldots, x_r)\, F_r(dx_1, \ldots, dx_r). \tag{11}$$

Here the set of polynomials

$$T^{(r)}_{j,s}(x_1, \ldots, x_r),\; 1 \leq j \leq r,\; r \geq 1,\; s \geq 0 \tag{12}$$

is defined explicitly by the formulae

$$T^{(r)}_{r,s}(x_1,\dots,x_r) = (1-x_1-\cdots-x_r)^s, \tag{13}$$

and

$$\begin{aligned} &T^{(r)}_{r-j,s}(x_1,\dots,x_r) \\ &\quad = (-1)^{j+1}\sum_{i_j=2j-1}^{i_{j+1}-2}\cdots\sum_{i_1=1}^{i_2-2}\prod_{k=0}^{j} i_k(1-\sum_{i=1}^{r-k}x_i)^{i_{k+1}-i_k-2},\ j=1,\dots,r \end{aligned} \tag{14}$$

where $i_0=-1$, $i_{j+1}=s+1$. Note that this implies $T^{(r)}_{r-j,s}\equiv 0$ for $s<2j$.

The diagonal entries of Q are calculated as:

$$q_{\xi\xi} = -\sum_{g=1}^{[b/2]}\sum_{a=g}^{b-g}\sum_{r=g}^{[(a+g)/2]}\int_{\Delta_r} S^{(r)}_{g,a-g}(x_1,\dots,x_r)F_r(dx_1,\dots,dx_r), \tag{15}$$

where

$$S^{(r)}_{g,s}(x_1,\dots,x_r) = \sum_{\substack{b_1\ge\cdots\ge b_g\ge 2\\ b_1+\cdots+b_g=b-s}}\binom{b}{b_1\dots b_g\ s}x_1^{b_1-2}\dots x_g^{b_g-2}T^{(r)}_{g,s}(x_1,\dots,x_r).$$

Observe that in the case when $F_2(\Delta_2)=0$ the formulae (11), (15) and $T^{(1)}_{1,s}(x)=(1-x)^s$ bring us back to (9) with $F=F_1$.

2 A weak convergence criterion

This section presents the main result of the paper, Theorem 2.1, which shows that the formulae (11) and (15) fully describe the class of coalescent generators for the population models with exchangeable reproduction. The central condition of Theorem 2.1 requires the existence of the limits

$$\lim_{N\to\infty}\frac{\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j})}{N^{k_1+\cdots+k_j-j}c_N} = \phi_j(k_1,\dots,k_j),\ k_1\ge\cdots\ge k_j\ge 2 \tag{16}$$

for all $j\in\mathbb{N}$.

To justify the denominator in the LHS of (16) turn to the chain of inequalities

$$\begin{aligned} &\sum_{\substack{i_1,\dots,i_j=1\\ \text{all distinct}}}^{N}(\nu_{i_1})_{k_1}\cdots(\nu_{i_j})_{k_j} \\ &\le \sum_{\substack{i_1,\dots,i_l=1\\ \text{all distinct}}}^{N}(\nu_{i_1})_{m_1}\nu_{i_1}^{k_1-m_1}\cdots(\nu_{i_l})_{m_l}\nu_{i_l}^{k_l-m_l}\sum_{i_{l+1},\dots,i_j=1}^{N}\nu_{i_{l+1}}^{k_{l+1}}\cdots\nu_{i_j}^{k_j} \\ &\le \sum_{\substack{i_1,\dots,i_l=1\\ \text{all distinct}}}^{N}(\nu_{i_1})_{m_1}N^{k_1-m_1}\cdots(\nu_{i_l})_{m_l}N^{k_l-m_l}(\nu_1+\cdots+\nu_N)^{k_{l+1}+\cdots+k_j} \end{aligned} \tag{17}$$

$$= \frac{N^{k_1+\cdots+k_j}}{N^{m_1+\cdots+m_l}} \sum_{\substack{i_1,\ldots,i_l=1 \\ \text{all distinct}}}^{N} (\nu_{i_1})_{m_1} \cdots (\nu_{i_l})_{m_l},$$

where $l \le j,\ k_1 \ge m_1 \ge 1, \ldots, k_l \ge m_l \ge 1$. With $l = 1$ and $m_1 = 2$ it entails that in general

$$\limsup_{N\to\infty} \frac{\mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j})}{N^{k_1+\cdots+k_j-j} c_N} \le 1,\ k_1 \ge \cdots \ge k_j \ge 2.$$

Relation (17) implies also that the set of the limits (16) is monotone:

$$\phi_j(k_1, \ldots, k_j) \le \phi_l(m_1, \ldots, m_l) \text{ whenever } l \le j,\ k_1 \ge m_1, \ldots, k_l \ge m_l. \quad (18)$$

Theorem 2.1 *If the limits (16) exist for all $j \in \mathbb{N}$, then for each sample size $n \in \mathbb{N}$ the asymptotic formula (8) holds with $Q = (q_{\xi\eta})_{\xi,\eta\in\mathcal{E}_n}$ defined by (11) and (15). The corresponding symmetric measures F_r, $r \in \mathbb{N}$ are uniquely determined via their moments*

$$\int_{\Delta_r} x_1^{k_1-2} \cdots x_r^{k_r-2} F_r(dx_1, \ldots, dx_r) = \phi_r(k_1, \ldots, k_r),\ k_1 \ge \cdots \ge k_r \ge 2. \quad (19)$$

If furthermore, the limit $\lim_{N\to\infty} c_N = c$ exists, then convergence (6) holds in the Skorohod sense. The limit coalescent process $(R_t)_{t\ge 0}$ is either (when $c > 0$) a discrete time Markov chain with the initial state (3) and the transition matrix $I + cQ$, or (when $c = 0$) a continuous time Markov chain with the initial state (3) and the transition matrix e^{tQ}.

Conversely, if (8) holds, then all the limits (16), $j \in \mathbb{N}$ exist.

Condition (16) has another two equivalent versions (cf. Section 4 for the proof): one in terms of the central moments

$$\lim_{N\to\infty} \frac{\mathrm{E}((\nu_1 - 1)^{k_1} \cdots (\nu_j - 1)^{k_j})}{N^{k_1+\cdots+k_j-j} c_N} = \phi_j(k_1, \ldots, k_j),\ k_1 \ge \cdots \ge k_j \ge 2 \quad (20)$$

and the other in terms of the tail distributions

$$\lim_{N\to\infty} \frac{N^j}{c_N} P(\nu_1 > Nx_1, \ldots, \nu_j > Nx_j) = \int_{x_1}^1 \cdots \int_{x_j}^1 \frac{F_j(dy_1, \ldots, dy_j)}{y_1^2 \cdots y_j^2}, \quad (21)$$

holding for all points $(x_1, \ldots, x_j)$ of continuity for the limit measure.

Version (21) brings the following picture of the asymptotic coalescent structure. Call *large* every family whose size is of order N. Obviously, every large family with a positive probability embraces two or more sampled ancestral lines (in other words begets a multiple merger). Due to the condition

$$\lim_{N\to\infty} N c_N^{-1} P(\nu_1 > Nx_1) = \int_{x_1}^1 y^{-2} F_1(dy)$$

a finite number of large families is encountered with a positive probability while scanning N generations in the population.

A large family caused a multiple merger might, in a sense, trigger a chain reaction of mergers within the same generation. To see this observe that the total number of families in a generation is equal to N and the relation

$$\lim_{N\to\infty} NP(\nu_2 > Nx_2 | \nu_1 > Nx_1) \;=\; \frac{\int_{x_1}^1 \int_{x_2}^1 y_1^{-2} y_2^{-2} F_2(dy_1, dy_2)}{\int_{x_1}^1 y^{-2} F_1(dy)}$$

indicates to the possibility that we might encounter another large family outside the initial one provided $F_2(\Delta_2) > 0$. Furthermore, if $F_3(\Delta_3) > 0$ the second large family leaves room for the third one:

$$\lim_{N\to\infty} NP(\nu_3 > Nx_3 | \nu_1 > Nx_1, \nu_2 > Nx_2)$$
$$= \frac{\int_{x_1}^1 \int_{x_2}^1 \int_{x_3}^1 y_1^{-2} y_2^{-2} y_3^{-2} F_3(dy_1, dy_2, dy_3)}{\int_{x_1}^1 \int_{x_2}^1 y_1^{-2} y_2^{-2} F_2(dy_1, dy_2)}$$

an so on. This imaginary chain reaction of mergers (there is no real order for the mergers happening within one generation) is bound to stop after a random number of rounds because the population of size N might host only a finite number of large families (given $F_{l+1}(\Delta_{l+1}) = 0$ this number of rounds never exceeds l).

Remark. According to Theorem 2.1 we have $F_r(\Delta_r) = \phi_r(2, \ldots, 2)$ so that (10) follows from (18). Note that $F_{l+1}(\Delta_{l+1}) = 0$ in particular when the random variables $(\nu_1)_2 \cdots (\nu_l)_2$ and $(\nu_{l+1})_2$ are not positively correlated, provided $\lim_{N\to\infty} c_N = 0$. Indeed in this case

$$\mathrm{E}((\nu_1)_2 \cdots (\nu_{l+1})_2) \le \mathrm{E}((\nu_1)_2 \cdots (\nu_l)_2) \cdot \mathrm{E}((\nu_1)_2) \le \frac{N^{2l} c_N}{(N)_l} \cdot N c_N$$

and hence

$$F_{l+1}(\Delta_{l+1}) = \phi_{l+1}(2, \ldots, 2) = \lim_{N\to\infty} \frac{\mathrm{E}((\nu_1)_2 \cdots (\nu_{l+1})_2)}{N^{l+1} c_N} \le \lim_{N\to\infty} c_N = 0.$$

3 The proof of the criterion

Lemma 3.2 *If the limits (16) exist for some $j \in \mathbb{N}$, then there exists a measure F_j uniquely determined on the simplex Δ_j by its moments (19).*

Proof. If $\phi_j(2, \ldots, 2) = 0$ then (19) implies $F_j(\Delta_j) = 0$. In the case $\phi_j(2, \ldots, 2) > 0$ we have $\mathrm{E}((\nu_1)_2 \cdots (\nu_j)_2) > 0$ for sufficiently large N. Let $Y_{1,j}, \ldots, Y_{j,j}$ be random variables with the joint distribution

$$P(Y_{1,j} = i_1, \ldots, Y_{j,j} = i_j) \;:=\; \frac{(i_1)_2 \cdots (i_j)_2}{\mathrm{E}((\nu_1)_2 \cdots (\nu_j)_2)} P(\nu_1 = i_1, \ldots, \nu_j = i_j), \quad (22)$$

where $i_1, \ldots, i_j \in \{2, \ldots, N\}$. The representation

$$\begin{aligned} \mathrm{E}(Y_{1,j}^{k_1} \cdots Y_{j,j}^{k_j}) &= \sum_{i_1,\ldots,i_j} \frac{(i_1)^{k_1} \cdots (i_j)^{k_j} (i_1)_2 \cdots (i_j)_2}{\mathrm{E}((\nu_1)_2 \cdots (\nu_j)_2)} P(\nu_1 = i_1, \ldots, \nu_j = i_j) \\ &= \frac{\mathrm{E}((\nu_1^{k_1+2} - \nu_1^{k_1+1}) \cdots (\nu_j^{k_j+2} - \nu_j^{k_j+1}))}{\mathrm{E}((\nu_1)_2 \cdots (\nu_j)_2)} \end{aligned}$$

in view of the equation $t^k = \sum_{l=1}^{k} (t)_l S_{kl}$, $t \in \mathbb{R}$, $k \geq 1$ (S_{kl} are the Stirling numbers of the second kind) leads to

$$\lim_{N\to\infty} \mathrm{E}((\tfrac{Y_{1,j}}{N})^{k_1} \cdots (\tfrac{Y_{j,j}}{N})^{k_j}) \overset{(16)}{=} \frac{\phi_j(k_1+2, \ldots, k_j+2)}{\phi_j(2, \ldots, 2)}, \; k_1, \ldots, k_j \in \mathbb{N}_0. \quad (23)$$

This convergence of moments implies (see Feller (1971), Chapter 8, Section 1) the weak convergence of the probability distributions on Δ_j:

$$P(\tfrac{Y_{1,j}}{N} \in dy_1, \ldots, \tfrac{Y_{j,j}}{N} \in dy_j) \to P_j(dy_1, \ldots, dy_j), \; N \to \infty. \quad (24)$$

Comparison between (23) with (24) shows that (19) holds with

$$F_j(dx_1, \ldots, dx_j) = \phi_j(2, \ldots, 2) \cdot P_j(dx_1, \ldots, dx_j).$$

The uniqueness of F_j is due to the fact that the limit moments (23) fully characterize the probability measure P_j. □

Definition 3.3 *For $j \in \mathbb{N}$, $k_1, \ldots, k_j \geq 2$ and $s \in \mathbb{N}_0$ define*

$$\psi_{j,s}(k_1, \ldots, k_j) := \lim_{N\to\infty} \frac{\mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j} \nu_{j+1} \cdots \nu_{j+s})}{N^{k_1+\cdots+k_j-j} c_N}$$

as long as this limit exists.

Lemma 3.4 *The following recursion over s holds:*

$$\begin{aligned} &\psi_{j,s+1}(k_1, \ldots, k_j) \\ &\quad = \psi_{j,s}(k_1, \ldots, k_j) - \sum_{i=1}^{j} \psi_{j,s}(k_1, \ldots, k_{i-1}, k_i + 1, k_{i+1}, \ldots, k_j) \\ &\qquad - s\psi_{j+1,s-1}(k_1, \ldots, k_j, 2) \end{aligned}$$

for all $j \in \mathbb{N}$, $k_1, \ldots, k_j \geq 2$ and all $s \in \mathbb{N}_0$.

Proof. Take the LHS and the RHS in the following chain of equalities

$$\begin{aligned} &(N-j-s)\mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j} \nu_{j+1} \cdots \nu_{j+s+1}) \\ &\overset{(1)}{=} \mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j} \nu_{j+1} \cdots \nu_{j+s} (\nu_{j+s+1} + \cdots + \nu_N)) \end{aligned}$$

$$
\begin{aligned}
&\overset{(2)}{=} \mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j}\nu_{j+1}\cdots\nu_{j+s}(N-\nu_1-\cdots-\nu_{j+s}))\\
&= \mathrm{E}\Big((\nu_1)_{k_1}\cdots(\nu_j)_{k_j}\nu_{j+1}\cdots\nu_{j+s}\big(N-(k_1+\cdots+k_j)-s\\
&\qquad\qquad -\sum_{i=1}^{j}(\nu_i-k_i)-\sum_{i=j+1}^{j+s}(\nu_i-1)\\
&= (N-(k_1+\cdots+k_j)-s)\,\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j}\nu_{j+1}\cdots\nu_{j+s})\\
&\qquad -\sum_{i=1}^{j}\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_i)_{k_i+1}\cdots(\nu_j)_{k_j}\nu_{j+1}\cdots\nu_{j+s})\\
&\qquad -s\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j}(\nu_{j+1})_2\nu_{j+2}\cdots\nu_{j+s}),
\end{aligned}
$$

and divide them by $N^{k_1+\cdots+k_j+1-j}c_N$. After letting $N\to\infty$ we get the asserted recursion equation. □

Lemma 3.5 *Polynomials (12) defined by relations (13) and (14) satisfy*

$$
T^{(r)}_{r,s+1}(x_1,\ldots,x_r)=(1-\sum_{i=1}^{r}x_i)\,T^{(r)}_{r,s}(x_1,\ldots,x_r) \tag{25}
$$

and for $j=1,\ldots,r-1$

$$
T^{(r)}_{j,s+1}(x_1,\ldots,x_r)=(1-\sum_{i=1}^{j}x_i)\,T^{(r)}_{j,s}(x_1,\ldots,x_r)-sT^{(r)}_{j+1,s-1}(x_1,\ldots,x_r). \tag{26}
$$

Proof. Formula (25) is obvious in view of (13). To verify (26) rewrite it as

$$
T^{(r)}_{r-j,s+1}(x_1,\ldots,x_r)=(1-\sum_{i=1}^{r-j}x_i)\,T^{(r)}_{r-j,s}(x_1,\ldots,x_r)-sT^{(r)}_{r-j+1,s-1}(x_1,\ldots,x_r)
$$

and apply (14). □

Lemma 3.6 *If the limits (16) exist for all* $j\in\mathbb{N}$, *then*

$$
\psi_{j,s}(k_1,\ldots,k_j) \;=\; \sum_{r\ge j}\int_{\Delta_r}x_1^{k_1-2}\cdots x_j^{k_j-2}T^{(r)}_{j,s}(x_1,\ldots,x_r)\,F_r(dx_1,\ldots,x_r),
$$

for all $j\in\mathbb{N}$, $k_1,\ldots,k_j\ge 2$ *and all* $s\in\mathbb{N}_0$.

Proof. We use induction over s. The case $s=0$ follows from the equality

$$
\psi_{j,0}(k_1,\ldots,k_j) \;=\; \phi_j(k_1,\ldots,k_j) \overset{(19)}{=} \int_{\Delta_j}x_1^{k_1-2}\cdots x_j^{k_j-2}F_j(dx_1,\ldots,dx_j).
$$

Lemma 3.4 and Lemma 3.5 ensure that the induction assumption implies the asserted formula

$$
\psi_{j,s+1}(k_1,\ldots,k_j)
$$

$$\stackrel{L.3.4}{=} \psi_{j,s}(k_1,\dots,k_j) - \sum_{i=1}^{j} \psi_{j,s}(k_1,\dots,k_{i-1},k_i+1,k_{i+1},\dots,k_j)$$
$$-s\psi_{j+1,s-1}(k_1,\dots,k_j,2)$$
$$\stackrel{ind}{=} \sum_{r\ge j}\int_{\Delta_r} x_1^{k_1-2}\cdots x_j^{k_j-2}(1-\sum_{i=1}^{j}x_i)\,T^{(r)}_{j,s}(x_1,\dots,x_r)\,F_r(dx_1,\dots,dx_r)$$
$$-s\sum_{r\ge j+1}\int_{\Delta_r} x_1^{k_1-2}\cdots x_j^{k_j-2}T^{(r)}_{j+1,s-1}(x_1,\dots,x_r)\,F_r(dx_1,\dots,dx_r)$$
$$\stackrel{L.3.5}{=} \sum_{r\ge j}\int_{\Delta_r} x_1^{k_1-2}\cdots x_j^{k_j-2}T^{(r)}_{j,s+1}(x_1,\dots,x_r)\,F_r(dx_1,\dots,dx_r).$$

□

To finish the proof of Theorem 2.1 turn to the equality

$$\begin{aligned} q_{\xi\eta} &= \lim_{N\to\infty}\frac{p_{\xi\eta}}{c_N} \qquad (27)\\ &\stackrel{(4)}{=} \lim_{N\to\infty}\frac{(N)_a}{(N)_b c_N}\mathrm{E}((\nu_1)_{b_1}\cdots(\nu_g)_{b_g}\nu_{g+1}\cdots\nu_a)\\ &= \lim_{N\to\infty}\frac{\mathrm{E}((\nu_1)_{b_1}\cdots(\nu_g)_{b_g}\nu_{g+1}\cdots\nu_a)}{N^{b-a}c_N} = \psi_{g,a-g}(b_1,\dots,b_g)\end{aligned}$$

saying that for any pair $\xi\subset\eta$ the LHS and the RHS exist or do not exist simultaneously and coincide when exist.

Assume that the limits (16) exist for all $j\in I\!N$. Then according to Lemma 3.6 and (27) the asymptotic formula (8) with (11) is valid for all $\xi\subset\eta$. In view of the equality

$$q_{\xi\xi} = -\sum_{\eta:\xi\subset\eta} q_{\xi\eta} = -\sum_{g=1}^{[b/2]}\sum_{a=g}^{b-g}\sum_{\substack{b_1\ge\cdots\ge b_g\ge 2\\ b_1+\cdots+b_g=b-a+g}} q_{\xi\eta}$$

formula (11) entails (15). After the asymptotic formula (8) is proved the weak convergence (6) is obtained as in Möhle (1999). Finally, the "only-if"-part of Theorem 2.1 is a simple corrolary of the equality

$$\phi_j(b_1,\dots,b_a) \;=\; \psi_{g,0}(b_1,\dots,b_g) \;\stackrel{(27)}{=}\; q_{\xi\eta}$$

which holds provided $a=g$. □

4 The equivalence of (16), (20), (21)

Fix some $j\in I\!N$. Here we show the equivalence of conditions (16), (20), (21) with the measures F_j and the limits ϕ_j being linked by (19).

(16)⇔(20). The proof of this equivalence is based on the decomposition

$$(\nu_1)_{k_1}\dots(\nu_j)_{k_j} \;=\; \sum_{i_1=1}^{k_1}\cdots\sum_{i_j=1}^{k_j}\alpha_{i_1,\dots,i_j}\,(\nu_1-1)^{i_1}\cdots(\nu_j-1)^{i_j}, \qquad (28)$$

where $\alpha_{i_1,\dots,i_j}$ are some finite coefficients and $\alpha_{k_1,\dots,k_j}=1$. It suffices to verify that

$$\mathrm{E}((\nu_1-1)^{i_1}\cdots(\nu_j-1)^{i_l}) = o(N^{k_1+\dots+k_j-j}c_N),\ N\to\infty \qquad (29)$$

for all

$$(i_1,\dots,i_l)\in[1,k_1]\times\cdots\times[1,k_l]\setminus\{(k_1,\dots,k_l)\},\ 1\le l\le j,\ k_1\ge\dots\ge k_j\ge 2.$$

To prove (29) notice first that

$$\mathrm{E}(\nu_1-1)=0,\ \mathrm{E}((\nu_1-1)^2)=\mathrm{E}((\nu_1)_2)=(N-1)c_N.$$

Turning to a counterpart of (17) for $\mathrm{E}|(\nu_1-1)^{i_1}\cdots(\nu_j-1)^{i_l}|$ we see that (29) is true when at least one i_r is greater or equal 2. In the remaining case $i_1=\dots=i_l=1$ the equality chain

$$\begin{aligned}(N-l+1)&\mathrm{E}(\nu_1-1)\dots(\nu_l-1)\\ &\overset{(1)}{=}\mathrm{E}(\nu_1-1)\dots(\nu_{l-1}-1)[(\nu_l-1)+\dots+(\nu_N-1)]\\ &\overset{(2)}{=}-\mathrm{E}(\nu_1-1)\dots(\nu_{l-1}-1)[(\nu_1-1)+\dots+(\nu_{l-1}-1)]\\ &=-(l-1)\mathrm{E}(\nu_1-1)^2(\nu_2-1)\dots(\nu_{l-1}-1).\end{aligned}$$

ends with a term of order $o(Nc_N)$ in accordance with the previous argument.

Thus (29) holds and we can conclude from (28) that for any fixed set of indices $k_1\ge\dots\ge k_j\ge 2$ two limits (16) and (20) are equal when exist with the existence of one entailing the existence of the other. This conclusion is slightly stronger than the asserted equivalence. □

(16)⇒(21). To arrive at the weak convergence (21) multiply

$$\begin{aligned}P(\nu_1>Nx_1,\dots,\nu_j>Nx_j) &= \int_{x_1}^1\dots\int_{x_j}^1 P(\frac{\nu_1}{N}\in dy_1\dots\frac{\nu_j}{N}\in dy_j)\\ &\overset{(22)}{=} N^j\mathrm{E}((\nu_1)_2\cdots(\nu_j)_2)\int_{x_1}^1\dots\int_{x_j}^1\frac{P(\frac{Y_1}{N}\in dy_1\dots\frac{Y_j}{N}\in dy_j)}{y_1(y_1-\frac{1}{N})\cdots y_j(y_j-\frac{1}{N})},\end{aligned}$$

by $N^jc_N^{-1}$ and apply (24) and (16). □

(21)⇒(20). Condition (21) implies the weak convergence of measures

$$\lim_{N\to\infty}\frac{N^j}{c_N}P(\frac{\nu_1-1}{N}>x_1,\dots,\frac{\nu_j-1}{N}>x_j) \;=\; \int_{x_1}^1\cdots\int_{x_j}^1\frac{F_j(dy_1,\dots,dy_j)}{y_1^2\cdots y_j^2}$$

which in turn implies the convergence of integrals

$$\begin{aligned}
&\lim_{N\to\infty} \frac{\mathrm{E}((\nu_1-1)^{k_1}\cdots(\nu_j-1)^{k_j})}{N^{k_1+\cdots+k_j-j}c_N} \\
&\quad = \frac{N^j}{c_N}\int_{\Delta_r} x_1^{k_1}\cdots x_r^{k_r} P(\frac{\nu_1-1}{N}\in dx_1,\ldots,\frac{\nu_j-1}{N}\in dx_j) \\
&\quad = \int_{\Delta_r} x_1^{k_1-2}\cdots x_r^{k_r-2} F_r(dx_1,\ldots,dx_r) \stackrel{(19)}{=} \phi_r(k_1,\ldots,k_r).
\end{aligned}$$

□

5 The Wright-Fisher model as a limit

Recall that the Wright-Fisher model describes a population of a fixed size (say l), where every individual chooses its parent at random among l individuals constituting the previous generation. Here we discuss a simple exchangeable population model whose time-scaled ancestral process converges to the ancestral process of the Wright-Fisher model.

Take a fixed constant $1 \le l \le N/2$ and consider such an exchangeable population model that in each generation exactly l families are of size $[N/l]$ while other family sizes are zeros and ones. In this case

$$\begin{aligned}
P(\nu_1 = \ldots = \nu_l = [N/l], \nu_{l+1} = \ldots = \nu_{l+l_1} = 1, \\
\nu_{l+l_1+1} = \ldots = \nu_N = 0) \;=\; \frac{l!l_1!}{(N)_{l+l_1}}
\end{aligned}$$

where $l_1 := N - l[N/l]$. It follows that

$$\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j}) \sim \frac{(l)_j}{(N)_j}(\tfrac{N}{l})_{k_1}\cdots(\tfrac{N}{l})_{k_j}, \quad N\to\infty$$

and hence

$$c \;:=\; \lim_{N\to\infty} c_N \;=\; 1/l.$$

This entails

$$\phi_j(k_1,\ldots,k_j) \;=\; \lim_{N\to\infty} \frac{\mathrm{E}((\nu_1)_{k_1}\cdots(\nu_j)_{k_j})}{N^{k_1+\cdots+k_j-j}c_N} \;=\; (l)_j l^{1-k_1-\cdots-k_j}.$$

Thus for this particular model the limit measure F_j assigns its total mass $\phi_j(2,\ldots,2) = (l)_j l^{1-2j}$ to the single point $(1/l,\ldots,1/l) \in I\!R^j$ being a zero measure for $j > l$. Now using Lemma 3.4 and induction over s we can show that

$$\psi_{j,s}(k_1,\ldots,k_j) \;=\; (l)_{j+s} l^{1-s-k_1-\cdots-k_j}. \tag{30}$$

The case $s = 0$ follows from $\psi_{j,0}(k_1,\ldots,k_j) = \phi_j(k_1,\ldots,k_j) = (l)_j l^{1-k_1-\cdots-k_j}$. The step from s to $s+1$ is given by

$$\begin{aligned}
\psi_{j,s+1}(k_1,\ldots,k_j) &\stackrel{\text{L. 3.4}}{=} (l)_{j+s} l^{1-s-k} - \sum_{i=1}^{j}(l)_{j+s} l^{-s-k} - s(l)_{j+s} l^{-s-k} \\
&= (l)_{j+s} l^{-s-k}(l-j-s) = (l)_{j+s+1} l^{1-(s+1)-k},
\end{aligned}$$

where $k := k_1 + \cdots + k_j$. We conclude that for $\xi \subset \eta$

$$q_{\xi\eta} \stackrel{(27)}{=} \psi_{g,a-g}(b_1, \ldots, b_g) \stackrel{(30)}{=} (l)_a l^{1-(a-g)-b_1-\cdots-b_g} = (l)_a l^{1-b}$$

so that the transition matrix $\Pi = I + cQ$ for the limit Markov chain has entries $\pi_{\xi\eta} = (l)_a l^{-b}$ for $\xi \subset \eta$ and the resulting coalescent process coincides with the ancestral process for the Wright-Fisher model with the population size l.

As l tends to infinity the generator Q converges to the generator of the standard n-coalescent in agreement with the weak convergence of the measure F_1 to the point measure in zero. For $j > 1$ the total mass of F_j converges to zero as l tends to infinity.

To generalize our example take an integer valued random variable L_N with

$$P(1 \leq L_N \leq N/2) = 1, \quad N \in I\!N$$

and conditional on $\{L_N = l\}$, $l \in I\!N$ define a population model as before. Assuming that L_N converges weakly as N tends to infinity to some random variable L, we deduce

$$\begin{aligned}
\phi_j(k_1, \ldots, k_j) &= \lim_{N\to\infty} \frac{\mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j})}{N^{k_1+\cdots+k_j-j} c_N} \\
&= \lim_{N\to\infty} \sum_{l=1}^{N/2} \frac{\mathrm{E}((\nu_1)_{k_1} \cdots (\nu_j)_{k_j} \mid L_N = l)}{N^{k_1+\cdots+k_j-j} c_N} P(L_N = l) \\
&= \sum_{l=1}^{\infty} (l)_j l^{1-k_1-\cdots-k_j} P(L = l) \;=\; \mathrm{E}((L)_j L^{1-k_1-\cdots-k_j})
\end{aligned}$$

and

$$c := \lim_{N\to\infty} c_N \;=\; \mathrm{E}(1/L).$$

Note that the last expectation is positive even if we allow for the possibility $0 \leq P(L = \infty) < 1$. In particular, if $L - 1$ has a Poisson distribution with parameter $\lambda > 0$, then

$$c \;=\; \int_0^1 \mathrm{E}(x^{L-1})\, dx \;=\; \int_0^1 e^{\lambda(x-1)}\, dx \;=\; \frac{1 - e^{-\lambda}}{\lambda}.$$

For the generalized example it follows that the entries of the limit generator Q are given by $q_{\xi\eta} = \mathrm{E}((L)_a L^{1-b})$ and the transition matrix $\Pi = I + cQ$ for the limit Markov chain has entries $\pi_{\xi\eta} = \mathrm{E}(1/L)\mathrm{E}((L)_a L^{1-b})$ for $\xi \subset \eta$. The resulting coalescent process depends on the observed value of the limit random variable L. If $L = l < \infty$, the coalescent is the ancestral process of the Wright-Fisher model with the population size l. When $L = \infty$ the sampled ancestral lines never merge.

References

[1] CANNINGS, C. (1974). The latent roots of certain Markov chains arising in genetics: a new approach, I. Haploid models. Adv. Appl. Prob. 6, 260 – 290.

[2] CANNINGS, C. (1975). The latent roots of certain Markov chains arising in genetics: a new approach, II. Further haploid models. Adv. Appl. Prob. 7, 264 – 282.

[3] FELLER, W. (1971). An Introduction to Probability Theory and Its Applications. Volume I, Second Edition, Wiley.

[4] KINGMAN, J.F.C. (1982a). On the Genealogy of Large Populations. J. Appl. Prob. 19A, 27–43.

[5] KINGMAN, J.F.C. (1982b). Exchangeability and the Evolution of Large Populations. in: KOCH, G. AND SPIZZICHINO, F.: Exchangeability in Probability and Statistics, North–Holland Publishing Company, pp. 97–112.

[6] KINGMAN, J.F.C. (1982c). The Coalescent. Stoch. Process. Appl. 13, 235–248.

[7] MÖHLE, M. (1998). Robustness Results for the Coalescent. J. Appl. Prob. 35, 438 – 447.

[8] MÖHLE, M. (1999). Weak convergence to the coalescent in neutral population models. J. Appl. Prob. 36 (to appear June 1999).

[9] MÖHLE, M. AND SAGITOV, S. (1998). A Characterisation of Ancestral Limit Processes Arising in Haploid Population Genetics Models. Berichte zur Stochastik und verwandten Gebieten, Johannes Gutenberg-Universität Mainz, November 1998, ISSN 0177-0098.

[10] SAGITOV, S. (1999). The General Coalescent with Asynchronous Mergers of Ancestral Lines. J. Appl. Prob. 36 (to appear December 1999)

Martin Möhle
Johannes Gutenberg-University Mainz
Department of Mathematics
Saarstraße 21
55099 Mainz, Germany e-mail: moehle@mathematik.uni-mainz.de

Serik Sagitov
Chalmers and Göteborgs Universities
Department of Mathematical Statistics
41296 Göteborg, Sweden e-mail: serik@math.chalmers.se