Beta coalescents and populations with large family sizes

by Jason Schweinsberg
University of California at San Diego

(joint work with Julien Berestycki and Nathanaël Berestycki)

Outline of Talk

1. Wright-Fisher model and Kingman’s coalescent
2. Large family sizes and coalescents with multiple collisions
3. Segregating sites
4. Site frequency spectrum
Wright-Fisher model and Kingman’s coalescent

$N$ individuals in each generation.
Each individual chooses parent at random.

Sample $n$ individuals from generation 0. Let $\Psi_N(m)$ be the partition of \{1, ..., $n$\} such that $i \sim j$ iff $i$th and $j$th sampled individuals have the same ancestor in generation $-m$.

$P($two individuals have the same parent$) = 1/N$.

As $N \to \infty$, the processes $\Psi_N = (\Psi_N([Nt]), t \geq 0)$ converge to Kingman’s coalescent (Kingman, 1982): only two lineages merge at a time, each pair of lineages merges at rate one.
**Λ-coalescents** (Pitman, 1999; Sagitov, 1999)

Let $\pi$ be a partition of $\{1, \ldots, n\}$ into blocks $B_1, \ldots, B_j$. Let $p \in (0, 1]$. A $p$-merger of $\pi$ is obtained as follows:

- Let $\xi_1, \ldots, \xi_j$ be i.i.d. Bernoulli($p$).
- Merge the blocks $B_i$ such that $\xi_i = 1$.

Coalescents can be described in terms of a finite measure $\Lambda$ on $[0, 1]$. Write $\Lambda = a\delta_0 + \Lambda_0$, where $\Lambda_0(\{0\}) = 0$. Transitions in the $\Lambda$-coalescent are as follows:

- Each pair of blocks merges at rate $a$.
- Construct a Poisson point process on $[0, \infty) \times (0, 1]$ with intensity $dt \times p^{-2}\Lambda_0(dp)$. If $(t, p)$ is a point of this Poisson process, then a $p$-merger occurs at time $t$.

When there are $b$ blocks, let $\lambda_{b,k}$ denote the rate of a transition in which $k$ blocks merge into one. Then, for $2 \leq k \leq b$,

$$
\lambda_{b,k} = \int_0^1 p^{k-2}(1 - p)^{b-k} \Lambda(dp).
$$
Genealogy of Galton-Watson processes

**Theorem** (Schweinsberg, 2003): Assume the numbers of offspring of the $N$ individuals in a generation are i.i.d. Suppose $P(\xi \geq k) \sim Ck^{-\alpha}$ and $E[\xi] > 1$. Obtain the next generation by sampling $N$ of these offspring.

- If $\alpha \geq 2$, limit is Kingman’s coalescent.
- If $1 \leq \alpha < 2$, ancestral processes converge to the $\Lambda$-coalescent,

$$\Lambda(dx) = \frac{1}{\Gamma(\alpha)\Gamma(2 - \alpha)} x^{1-\alpha}(1 - x)^{\alpha-1} dx$$

is the Beta($2 - \alpha, \alpha$) distribution.

If $\alpha \in (0, 1) \cup (1, 2)$, the Beta($2-\alpha, \alpha$)-coalescent gives the genealogy of $\alpha$-stable continuous-state branching process with branching mechanism $\Psi(\lambda) = \lambda^\alpha$, after a time-change (Birkner-Blath-Capaldo-Etheridge-Möhle-Schweinsberg-Wakolbinger, 2005).

Case $\alpha = 1$: Bolthausen-Sznitman coalescent gives genealogy of CSBP with $\Psi(\lambda) = \lambda \log \lambda$ (Bertoin-Le Gall, 2000).
Coalescents with mutations

Assume there are mutations at rate $\theta$ along each lineage.

Segregating sites: $S_n =$ number of sites at which not all members of sample agree. Example: $S_n = 4$.

Allelic partition: blocks represent groups of individuals that got the same mutations. Example: $\Pi_n = \{\{1\}, \{2\}, \{3, 4\}, \{5\}\}$.

Allele frequency spectrum: $N_{k,n} =$ number of blocks of size $k$ in allelic partition. Example: $N_{1,5} = 3$, $N_{2,5} = 1$.

Site frequency spectrum: $M_{k,n} =$ number of mutations affecting $k$ individuals. Example: $M_{1,5} = 2$, $M_{2,5} = 1$, $M_{4,5} = 1$. 
Segregating sites

Kingman’s coalescent:

\[ E[S_n] = \theta \sum_{b=2}^{n} b \left( \frac{b}{2} \right)^{-1} = 2\theta \sum_{b=2}^{n} \frac{1}{b-1} \sim 2\theta \log n. \]

Limiting distribution is normal.

Beta\((2 - \alpha, \alpha)\)-coalescent with \(1 < \alpha < 2\) (Berestycki-Berestycki-Schweinsberg, 2006):

\[ \frac{S_n}{n^{2-\alpha}} \rightarrow_p \frac{\theta\alpha(\alpha - 1)\Gamma(\alpha)}{2 - \alpha}. \]

Limiting distribution unknown, conjectured to be normal if and only if \(\alpha > \sqrt{2}\) (Delmas, Dbersin, Siri-Jegousse, 2007).

Beta\((1,1)\)-coalescent (Drmota-Iksanov-Möhle-Roesler, 2006):

\[ \frac{\log n}{n} S_n \rightarrow_p \theta. \]

Limiting distribution is stable.

\(\int_0^1 p^{-1} \Lambda(dp) < \infty\) (Möhle, 2006): \(S_n/n\theta \rightarrow_d S\).
Proof idea for beta coalescent

Let $\lambda_b = \text{total merger rate when } b \text{ lineages}.$

Let $G_n(b) = P(\text{there are exactly } b \text{ lineages at some time}).$

$$E[S_n] = \theta \sum_{b=2}^{n} b\lambda_b^{-1} G_n(b).$$

Total merger rate when there are $b$ lineages is

$$\lambda_b = \sum_{k=2}^{b} \binom{b}{k} \lambda_{b,k} \sim \frac{1}{\alpha \Gamma(\alpha)} b^\alpha.$$

When there are $b$ lineages, the expected number of lineages that are lost after the next merger converges to $1/(\alpha - 1)$ as $b \to \infty$ (Bertoin-Le Gall, 2005).

A renewal argument gives $G_n(b) \approx \alpha - 1$ for large $n$ and $b.$

$$E[S_n] \approx \theta \alpha(\alpha - 1) \Gamma(\alpha) \int_0^n x^{1-\alpha} dx = \theta \frac{\alpha(\alpha - 1) \Gamma(\alpha)}{2 - \alpha} n^{2-\alpha}.$$
Site and allele frequency spectrum

**Theorem** (Berestycki-Berestycki-Schweinsberg, 2006): For the Beta\((2 - \alpha, \alpha)\)-coalescent with \(1 < \alpha < 2\), we have

\[
\frac{M_{k,n}}{S_n} \rightarrow_p \frac{(2 - \alpha)\Gamma(k + \alpha - 2)}{\Gamma(\alpha - 1)k!} = a_k.
\]

and \(N_{k,n}/S_n \rightarrow_p a_k\).

**Remarks:**

- We have \(a_k \sim Ck^{\alpha - 3}\). Smaller \(\alpha\) means more low frequency mutants.
- We have \(a_1 = 2 - \alpha\).
- For Kingman’s coalescent, \(E[M_{k,n}] = \theta/k\), matching \(\alpha = 2\).
- Basdevant-Goldschmidt (2007) obtained results for Bolthausen-Sznitman coalescent (\(\alpha = 1\)).
Derivation of site frequency spectrum

Consider genealogy of individuals alive at time 1 in $\alpha$-stable continuous-state branching process. For a line that begins at time $t$, the next branchpoint is uniform on $[t, 1]$. At each branchpoint, the number $\chi$ of offspring satisfies (Duquesne-Le Gall, 2002):

$$P(\chi = k) = \frac{\alpha(2 - \alpha) \ldots (k - 1 - \alpha)}{k!}, \quad k \geq 2.$$ 

After time change $t \mapsto -\log(1 - t)$, have Galton-Watson tree with mutations along branches at rate $\theta e^{-t}$. 
Let \((Z_t)_{t \geq 0}\) be a continuous-time Galton-Watson process having offspring distribution \(\chi\), mutations along branches at rate \(\theta e^{-t}\).

Let \(m = E[\chi] = 1 + 1/(\alpha - 1)\). Then \(E[Z_t] = e^{(m-1)t}\).

Let \(c = m - 2\) and \(\tau \sim \text{Exponential}(c)\).

Let \(M_k(t) = \text{number of mutations affecting } k \text{ lineages at time } t\).

\[
E[M_k(t)] = \int_0^t P(Z_{t-s} = k) P(\text{mutation in } ds) \\
= \int_0^t P(Z_{t-s} = k) E[Z_s \theta e^{-s}] \, ds \\
= \int_0^t \theta e^{cs} P(Z_{t-s} = k) \, ds \\
= \frac{\theta}{c} e^{ct} P(Z_{\tau} = k).
\]

Let \(a_k = P(Z_{\tau} = k)\) and \(r_k = P(Z_t = k \text{ for some } t \leq \tau)\).

Then \(a_k = \frac{c}{k+c} r_k\) and \(r_k = \sum_{j=1}^{k-1} r_j \cdot \frac{j}{j+c} \cdot P(\chi = k - j + 1)\).

Solving recursion gives value for \(a_k\).
Block sizes for exchangeable random partitions

Let $\Pi$ be an exchangeable random partition of $\mathbb{N}$.

Let $\Pi_n$ be the restriction of $\Pi$ to $\{1, \ldots, n\}$.

Let $N_n$ be the number of blocks of $\Pi_n$, and let $N_{n,k}$ be the number of blocks of size $k$.

**Theorem** (Karlin, 1967; Gnedenin-Hansen-Pitman, 2007): Suppose $1 < \alpha < 2$. If there exists a random variable $Z_\alpha$ such that $0 < Z_\alpha < \infty$ a.s. and

$$\frac{N_n}{n^{2-\alpha}} \rightarrow Z_\alpha \text{ a.s.}$$

then

$$\lim_{n \rightarrow \infty} \frac{N_{n,k}}{N_n} = \frac{(2 - \alpha)\Gamma(k + \alpha - 2)}{\Gamma(\alpha - 1)k!} = a_k \text{ a.s.}$$

If $\Pi$ is the allelic partition associated with the Beta($2 - \alpha, \alpha$)-coalescent, the convergence in (1) holds in probability (same asymptotics as number of segregating sites).