The nested Kingman coalescent: speed of coming down from infinity

by Jason Schweinsberg
(University of California at San Diego)

Joint work with:

Airam Blancas Benítez (Goethe Universität Frankfurt)
Tim Rogers (University of Bath)
Arno Siri-Jégousse (Universidad Nacional Autónoma de México)

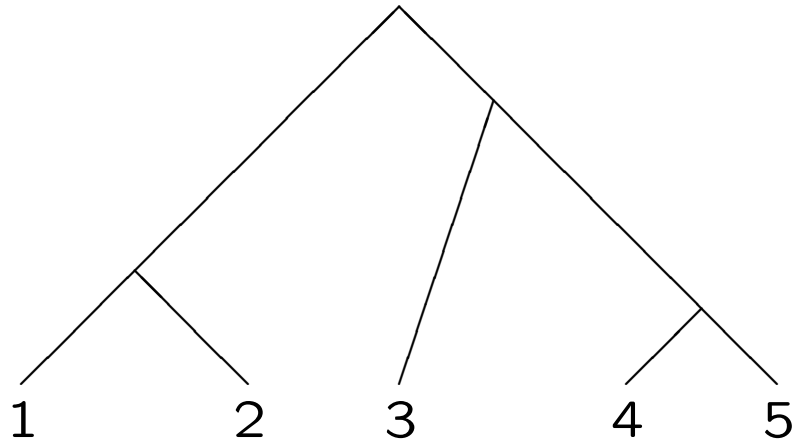
Coalescent Processes

Coalescent processes: describe the genealogy of a collection of individuals, usually from the same species.

Sample n individuals at random from a population. Follow their ancestral lines backwards in time. The lineages coalesce, until they are all traced back to a common ancestor.

Represent by a stochastic process $(\Pi(t), t \geq 0)$ taking its values in the set $\mathcal{P}_n$ of partitions of $\{1, \ldots, n\}$.

Kingman's Coalescent (Kingman, 1982): Continuous-time Markov chain with state space $\mathcal{P}_n$. Only two lineages merge at a time. Each pair of lineages merges at rate one.



$\{1,2,3,4,5\}$

$\{1,2\},\{3,4,5\}$

$\{1,2\},\{3\},\{4,5\}$

$\{1\},\{2\},\{3\},\{4,5\}$

$\{1\},\{2\},\{3\},\{4\},\{5\}$

Number of lineages at small times

Consider Kingman's coalescent started with n lineages.

Let $K_n(t)$ be the number of lineages remaining after time t, which is number of blocks of $\Pi(t)$.

When there are j lineages, time until the next merger has an exponential distribution with rate $\binom{j}{2}$.

For small t and large n, approximate $K_n(t)$ by the solution to

$$x'(t) = -\frac{x(t)^2}{2}, \qquad x(0) = n,$$

which yields

$$K_n(t) \approx \frac{2}{t + 2/n}.$$

If $n = \infty$, we have $K_\infty(t) < \infty$ for all $t > 0$, and (Aldous, 1991):

$$\lim_{t \to 0} tK_\infty(t) = 2 \text{ a.s.}$$

Nested Coalescents

Phylogenetics: the study of the evolutionary relationships among different species.

Nested coalescents: sample individuals from different species and follow their ancestral lines backwards in time.

Nested Kingman's coalescent:

- Sample n individuals from each of s species, with $1 \leq n \leq \infty$, $1 \leq s \leq \infty$.
- Each pair of lineages belonging to the same species merges at rate 1. Lineages from different species do not merge.
- Each pair of species merges into a single species at rate $c > 0$.

Represent by a stochastic process $(\Pi(t), t \geq 0)$ taking its values in the set of labelled partitions of $\{1, \ldots, ns\}$. Each block has a label, indicating the species to which the lineage belongs.

A more general family of nested coalescents was introduced by Blancas Benítez, Duchamps, Lambert, and Siri-Jégousse (2018), allowing for multiple mergers of ancestral lines.

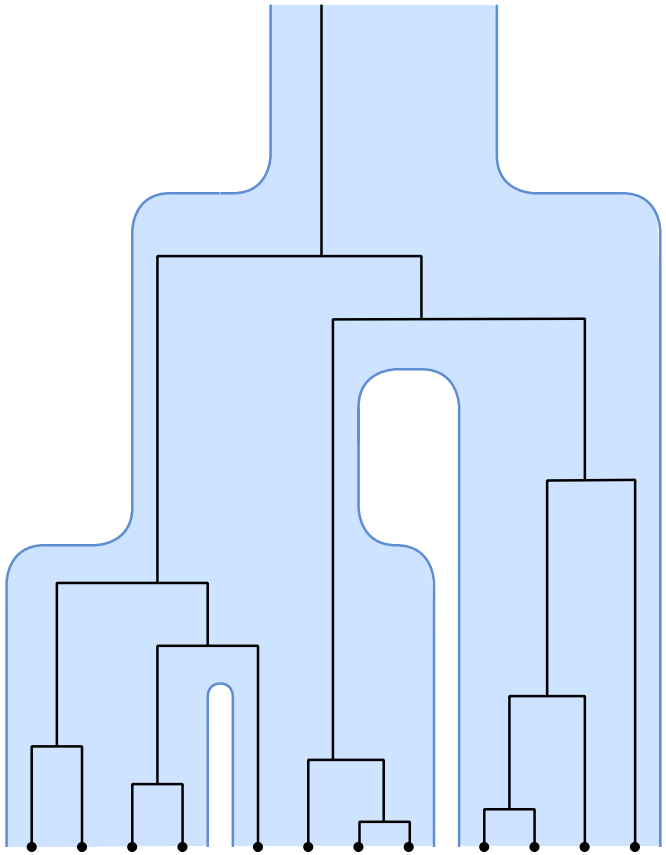


Illustration of the nested Kingman's coalescent with $s = 3$ species and $n = 4$ individuals sampled from each species.

The blue tree that records the mergers is sometimes called the "species tree". The tree that records the genealogical relationship among the 12 sampled individuals is called the "gene tree".

Extensive literature on estimating species tree from gene tree.

Number of lineages at small times

Let $S(t)$ be the number of species remaining at time t.

Then $(S(t), t \geq 0)$ has the same law as the number of blocks in Kingman's coalescent, with time scaled by c, so for small t,

$$S(t) \approx \frac{2}{ct + 2/s}.$$

Let $N(t)$ be the number of individual lineages at time t. We have

$$N(t) = N_1(t) + \cdots + N_{S(t)}(t),$$

where $N_i(t)$ is the number of individual lineages belonging to the ith species.

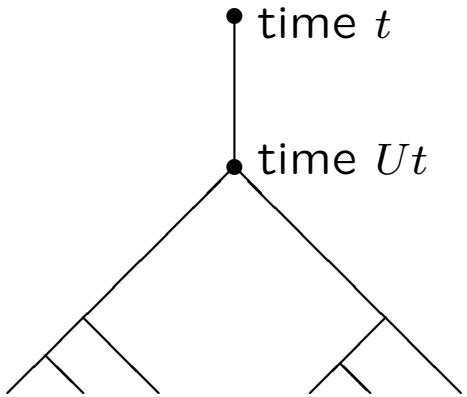
How can we describe the behavior of $N(t)$ when t is small? We focus on $n = s = \infty$.

If there were no species mergers, we would have $N_i(t) \approx 2/t$ and $N(t) \approx 4/ct^2$. The number of lineages belonging to a species makes a large upward jump after a species merger.

The distribution of $N_i(t)$

We try to approximate the distribution of $N_i(t)$ for small t by W/t, where $W \geq 2$ has some unknown distribution.

The portion of the species tree between times 0 and t consists of $S(t)$ subtrees. Consider the subtrees associated with the ith species at time t.



The last branchpoint occurs at approximately time Ut, where U has a uniform$(0, 1)$ distribution. At that time, we merge two species carrying $W_1/(Ut)$ and $W_2/(Ut)$ individual lineages.

A recursive distributional equation

Those $(W_1 + W_2)/Ut$ lineages merge according to Kingman's coalescent for time $(1-U)t$. Recalling that

$$K_n(t) \approx \frac{2}{t + 2/n},$$

the number of these lineages remaining at time t is

$$\frac{2}{(1-U)t + \frac{2Ut}{W_1+W_2}} = \frac{1}{t} \cdot \frac{2}{1 - U\left(1 - \frac{2}{W_1+W_2}\right)}.$$

Thus, W satisfies the recursive distributional equation (RDE)

$$W =_d \frac{2}{1 - U\left(1 - \frac{2}{W_1+W_2}\right)},$$

where W_1, W_2, and U are independent, W_1 and W_2 have the same distribution as W, and U has the uniform$(0,1)$ distribution.

This RDE has unique solution (Aldous and Bandyopadhyay, 2001).

The main result

Let γ be the mean of the distribution satisfying the RDE. Then

$$N(t) = N_1(t) + \cdots + N_{S(t)}(t) \approx \frac{\gamma}{t} S(t) \approx \frac{\gamma}{t} \cdot \frac{2}{ct} = \frac{2\gamma}{ct^2}.$$

Simulations indicate $\gamma \approx 3.45$.

Theorem: Consider the nested Kingman's coalescent started with $n = s = \infty$. We have

$$t^2 N(t) \to_p \frac{2\gamma}{c}, \quad \text{as } t \to 0.$$

Theorem: Consider a sequence of nested Kingman's coalescent processes. The jth process has s_j species and n_j lineages sampled from each species, with $n_j \gg s_j$.

- If $1/n_j \ll t_j \ll 1/s_j$, then $\dfrac{t_j N(t_j)}{s_j} \to_p 2$ as $j \to \infty$
- if $1/s_j \ll t_j \ll 1$, then $t_j^2 N(t_j) \to_p \dfrac{2\gamma}{c}$ as $j \to \infty$.

Another approach

Lambert and Schertzer (2018) obtained our result by different methods. They approximate the distribution of $N_i(t)$ by writing down a partial differential equation for the density.

When $1 \ll s \ll n$, recall that

$$N_i(t) \approx \begin{cases} 2/t & \text{if } 1/n \ll t \ll 1/s \\ W/t & \text{if } 1/s \ll t \ll 1 \end{cases}$$

We do not have results for $t \sim A/s$.

Lambert and Schertzer (2018) have a PDE which should give the density of $N_i(t)$ for $t \sim A/s$, but they have not yet rigorously proved the connection to the discrete model.