# An Introduction to Exchangeable Coalescent Processes

by Jason Schweinsberg

University of California at San Diego

## Outline of Talk

1. Exchangeable random partitions
2. Kingman's Coalescent
3. Coalescents with multiple mergers
4. Biological motivation, problems of interest
5. Poisson process construction
6. Random walk methods
7. Random tree constructions

# Partitions

A *partition* of a set $S$ is a collection of disjoint subsets $B_i$ of $S$ such that

$$\bigcup_i B_i = S.$$

The sets $B_i$ are called *blocks* of the partition. Blocks of a partition of size 1 are called *singletons*.

If $\pi$ is a partition, write $i \sim_\pi j$ if $i$ and $j$ are in the same block. Let $\#\pi$ denote the number of blocks of the partition $\pi$.

$\mathcal{P}_\infty =$ set of partitions of $\mathbb{N}$.

$\mathcal{P}_n =$ set of partitions of $\{1, \ldots, n\}$.

If $\pi \in \mathcal{P}_\infty$, or $\pi \in \mathcal{P}_m$ with $m > n$, then $R_n\pi \in \mathcal{P}_n$ is the restriction of $\pi$ to $\{1, \ldots, n\}$, which means $i \sim_{R_n\pi} j$ if and only if $i \sim_\pi j$.

Example: $\pi = \{\{1,3,4,7,8\},\{2,5,9\},\{6\}\}$

$$R_5\pi = \{\{1,3,4\},\{2,5\}\}.$$

# Exchangeable Random Partitions

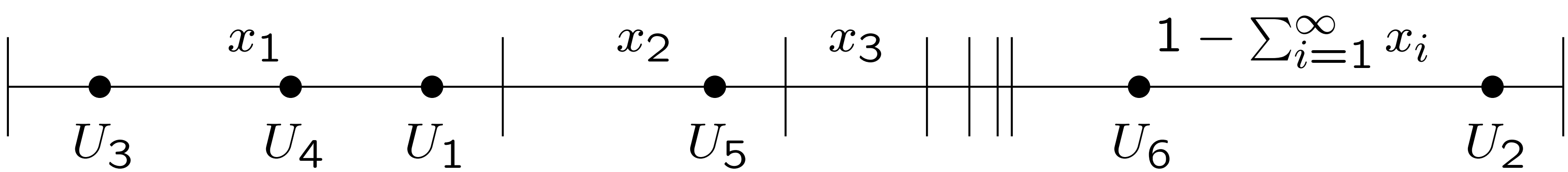
If $\pi \in \mathcal{P}_\infty$ and $\sigma$ is a permutation of $\mathbb{N}$, define $\sigma\pi \in \mathcal{P}_\infty$ such that $\sigma(i) \sim_{\sigma\pi} \sigma(j)$ if and only if $i \sim_\pi j$.

If $\Pi$ is a random partition of $\mathbb{N}$, we say $\Pi$ is *exchangeable* if $\sigma\Pi =_d \Pi$ for all permutations $\sigma$ of $\mathbb{N}$.

Let $\Delta = \left\{(x_1, x_2, \ldots) : x_1 \geq x_2 \geq \ldots \geq 0, \sum_{i=1}^{\infty} x_i \leq 1\right\}$.

Paintbox (stick-breaking) construction: Let $x = (x_1, x_2, \ldots) \in \Delta$. Divide $[0, 1]$ into subintervals of lengths $x_1, x_2, \ldots$ and $1 - \sum_{i=1}^{\infty} x_i$. Let $U_1, U_2, \ldots$ be i.i.d. Uniform(0,1).

Define $\Pi$ such that $i \sim_\Pi j$ if and only if $U_i$ and $U_j$ fall in the same subinterval, other than the last interval of length $1 - \sum_{i=1}^{\infty} x_i$.




$$R_6\Pi = \{\{1, 3, 4\}, \{2\}, \{5\}, \{6\}\}.$$

Given $x \in \Delta$, let $P^x$ denote the distribution of the associated paintbox partition.

**Theorem** (Kingman, 1978): Suppose $\Pi$ is an exchangeable random partition of $\mathbb{N}$. Then there exists a probability measure $\mu$ on $\Delta$ such that

$$P(\Pi \in A) = \int_\Delta P^x(A)\, \mu(dx)$$

for all measurable subsets $A$ of $\mathcal{P}_\infty$. We call $\Pi$ a $\mu$-paintbox partition.

Suppose $B$ is a block of $\Pi$. Then

$$\lim_{n\to\infty} n^{-1} \sum_{i=1}^{n} 1_{\{i \in B\}}$$

exists and is called the *asymptotic frequency* of $B$. The sequence of ranked asymptotic frequencies of blocks has distribution $\mu$.

Note: Every block of $\Pi$ either is a singleton or has positive asymptotic frequency.

## **Kingman's $n$-Coalescent** (Kingman, 1982)

Continuous-time Markov chain $(\Pi_n(t), t \geq 0)$ taking values in $\mathcal{P}_n$.

$\Pi_n(0)$ consists of $n$ singletons.

A transition that involves merging two blocks of the partition into one happens at rate 1. No other transitions are possible.

When there are $k$ blocks, the distribution of the time until the next merger is exponential with rate $k(k-1)/2$. Then two randomly chosen blocks merge.

**Example**:

$$\begin{array}{ll}
\Pi_n(t) = \{1\}, \{2\}, \{3\}, \{4\}, \{5\} & 0 \leq t < \tau_1 \\
\Pi_n(t) = \{1, 3\}, \{2\}, \{4\}, \{5\} & \tau_1 \leq t < \tau_2 \\
\Pi_n(t) = \{1, 3, 5\}, \{2\}, \{4\} & \tau_2 \leq t < \tau_3 \\
\Pi_n(t) = \{1, 3, 5\}, \{2, 4\} & \tau_3 \leq t < \tau_4 \\
\Pi_n(t) = \{1, 2, 3, 4, 5\} & \tau_4 \leq t
\end{array}$$

## Kingman's Coalescent

**Consistency**: if $m > n$, then $(R_n \Pi_m(t), t \geq 0)$ and $(\Pi_n(t), t \geq 0)$ have the same law.

By Kolmogorov's Extension Theorem, there is a continuous-time Markov process $(\Pi_\infty(t), t \geq 0)$ with state space $\mathcal{P}_\infty$ such that $(R_n \Pi_\infty(t), t \geq 0)$ has the same law as $(\Pi_n(t), t \geq 0)$ for all $n$.

The process $(\Pi_\infty(t), t \geq 0)$ is called *Kingman's coalescent*.

## Coalescents with multiple mergers ($\Lambda$-coalescents)

Pitman (1999), Sagitov (1999), Donnelly-Kurtz (1999).

**Definition** (Pitman, 1999): A *coalescent with multiple mergers* is a $\mathcal{P}_\infty$-valued process $\Pi_\infty = (\Pi_\infty(t), t \geq 0)$ such that:

- $\Pi_\infty(0)$ is the partition of $\mathbb{N}$ into singletons.
- For all $n \in \mathbb{N}$, the process $(R_n \Pi_\infty(t), t \geq 0)$ is a continuous-time $\mathcal{P}_n$-valued Markov chain with the property that when $R_n \Pi_\infty(t)$ has $b$ blocks, each $k$-tuple of blocks is merging to form a single block at some fixed rate $\lambda_{b,k}$, and no other transitions are possible.

$$\{1\},\{2\},\{3\},\{4\} \to \{1,2,3\},\{4\} \qquad \text{rate } \lambda_{4,3}$$

$$\{1,2\},\{3,6,7\},\{4\},\{5,8\} \to \{1,2,3,4,6,7\},\{5,8\} \qquad \text{rate } \lambda_{4,3}$$

## Characterization of coalescents with multiple mergers

**Theorem** (Pitman, 1999): For any coalescent with multiple mergers, we have

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\,\Lambda(dx)$$

for some finite measure $\Lambda$ on $[0,1]$.

**Definition**: We call a process with these rates a $\Lambda$-*coalescent*.

**Note**: When $\Lambda = \delta_0$, we get Kingman's coalescent because ($\lambda_{b,2} = 1, \lambda_{b,k} = 0$ for $k > 2$).

# Proof of Pitman's Theorem

Let $(\Pi_\infty(t), t \geq 0)$ be a coalescent with multiple mergers.
Let $T$ be the time when $\{1\}$ and $\{2\}$ merge.

Let $B_1, B_2, \ldots$ be the blocks of $\Pi_\infty(T-)$, ordered by their smallest elements. Assume for now $\#\Pi_\infty(T-) = \infty$.

Let $\xi_i = 1$ if $B_i$ merges with $\{1\}$ and $\{2\}$ at time $T$, and $\xi_i = 0$ otherwise. Then $(\xi_i)_{i=3}^\infty$ is exchangeable. Thus, by de Finetti's Theorem, there exists a probability measure $\Lambda'$ such that

$$P(\xi_3 = \cdots = \xi_k = 1, \xi_{k+1} = \cdots = \xi_b = 0) = \int_0^1 x^{k-2}(1-x)^{b-k}\,\Lambda'(dx).$$

We have $P(\xi_3 = \cdots = \xi_k = 1, \xi_{k+1} = \cdots = \xi_b = 0) = \lambda_{b,k}/\lambda_{2,2}$.

Let $\Lambda = \lambda_{2,2}\Lambda'$. Then

$$\lambda_{b,k} = \int_0^1 x^{k-2}(1-x)^{b-k}\,\Lambda(dx).$$

If $\#\Pi_\infty(T-) < \infty$, then condition $\#\Pi_\infty(T-) \geq k$ and apply Kolmogorov's Extension Theorem.

## Exchangeable Coalescent Processes

(Schweinsberg (2000), Möhle and Sagitov (2001), Bertoin and Le Gall (2003))

One can consider also coalescents that allow for simultaneous multiple mergers.

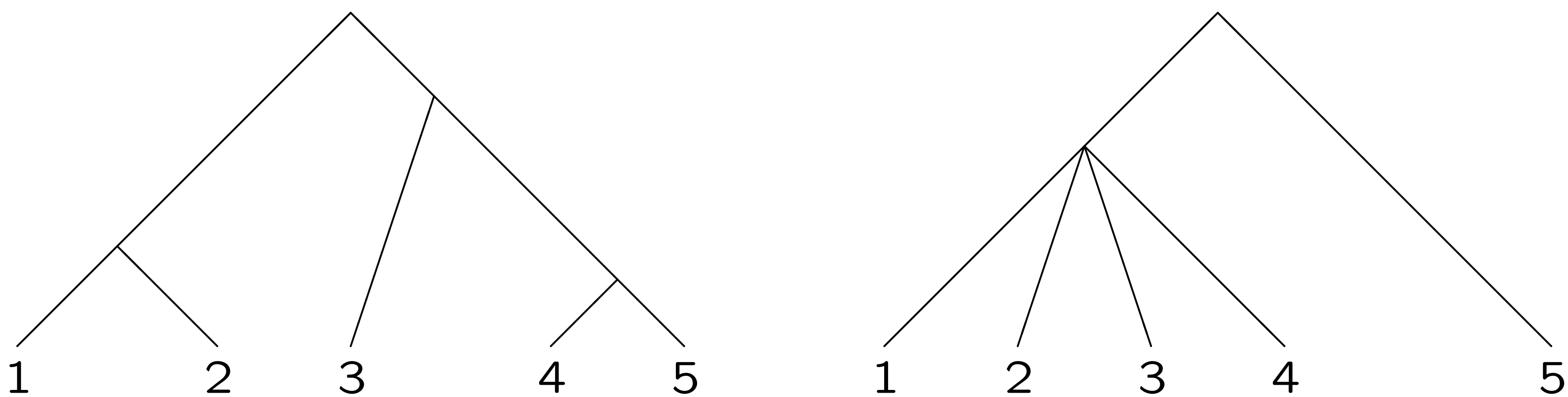
Let $T$ be the time when integers 1 and 2 merge, and let $B_1, B_2, \ldots$ be the blocks of $\Pi(T-)$.

Let $\Psi$ be the partition of $\{3, 4, \ldots\}$ such that $i \sim_\Psi j$ if and only if $B_i$ and $B_j$ are in the same block of $\Pi(T)$. Then $\Psi$ is an exchangeable random partition, thus a $\Xi'$-paintbox partition for some probability measure $\Xi'$ on $\Delta$.

Let $\Xi = \lambda_{2,2}\Xi'$, where $\lambda_{2,2}$ is the rate at which the integers 1 and 2 merge. The associated coalescent process is called the $\Xi$-coalescent.

## Biological motivation

Coalescent processes describe the genealogy of a sample of size $n$ from a population. Here $i \sim_{\Pi_n(t)} j$ if the $i$th and $j$th individuals in the sample have the same ancestor at time $-t$.




Applications of coalescents with multiple mergers:

- Large family sizes (many lineages trace back to individual with large number of offspring).

- Natural selection (many lineages trace back to individual who got a beneficial mutation).

## Questions of Interest

1. Given a model for how a population evolves, determine what coalescent process describes its genealogy.

2. Understand the distribution of the time $T_n$ for $n$ blocks to merge into one. Note that $T_n$ is the height of the tree, and the time back to the most recent common ancestor (MRCA).

3. Understand the distribution of the total tree length $L_n$ when the coalescent starts with $n$ blocks. This should be approximately proportional to the number of mutations in a sample of size $n$ from a population.

4. Understand the distribution of $L_{n,k}$, the total length of the branches that are ancestors of $k$ of the $n$ leaves of the tree. This should be approximately proportional to the number of mutations that appear on $k$ out of $n$ individuals in the population. Note: $L_{n,1}$ is the total length of external branches.

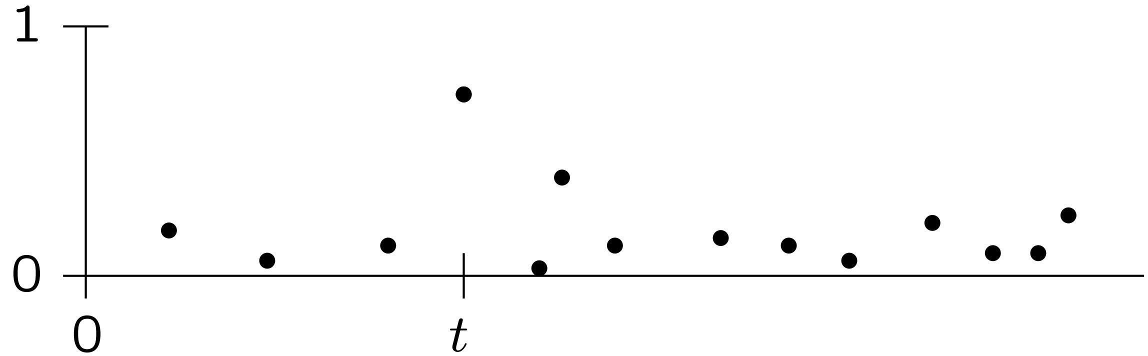
# Poisson process construction of Λ-coalescents

Let $\pi$ be a partition of $\mathbb{N}$ into blocks $B_1, B_2, \ldots$. Let $p \in (0,1]$. A $p$-merger of $\pi$ is obtained as follows:

- Let $\xi_1, \xi_2, \ldots$ be i.i.d. with $P(\xi_i = 1) = p$, $P(\xi_i = 0) = 1 - p$.
- Merge the blocks $B_i$ such that $\xi_i = 1$.

Write $\Lambda = a\delta_0 + \Lambda_0$, where $\Lambda_0(\{0\}) = 0$. Transitions:

- 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 an $p$-merger occurs at time $t$.




When there are $b$ blocks, $\lambda_{b,k} = \int_0^1 p^{k-2}(1-p)^{b-k}\, \Lambda(dp)$.

## **Large family sizes** (Schweinsberg, 2003)

Consider a population in which the number of offspring $\xi$ of an individual satisfies $P(\xi \geq k) \sim Ck^{-\alpha}$, where $1 \leq \alpha < 2$.

If there are $N$ individuals, this reproduction event produces a $p$-merger with $p \geq x$ if and only if

$$\frac{\xi}{\xi + N} \geq x \qquad \Longleftrightarrow \qquad \xi \geq \frac{x}{1-x} \cdot N$$

The probability of such a family in a given generation is

$$NP\left(\xi \geq \frac{x}{1-x} \cdot N\right) \sim NC\left(\frac{x}{1-x} \cdot N\right)^{-\alpha}.$$

The rate of such mergers in the Beta$(2-\alpha, \alpha)$-coalescent is

$$\frac{1}{\Gamma(\alpha)\Gamma(2-\alpha)} \int_x^1 p^{-1-\alpha}(1-p)^{\alpha-1}\,dp = \frac{1}{\alpha\Gamma(\alpha)\Gamma(2-\alpha)}\left(\frac{x}{1-x}\right)^{-\alpha}.$$

If $\alpha \geq 2$, Kingman's coalescent describes the genealogy.

## Populations undergoing selection

In some standard population models involving natural selection, the genealogy is given by the Bolthausen-Sznitman coalescent, the $\Lambda$-coalescent in which $\Lambda$ is the Beta$(1,1)$ distribution.

The Bolthausen-Sznitman coalescent describes the genealogy if, when the population has size $N$, events in which an individual acquires a beneficial mutation and quickly produces at least $Nx$ descendants happen at a rate proportional to $x^{-1}$.

**Non-rigorous work**: Brunet, Derrida, Mueller, Munier (2007)
Desai, Walczak, Fisher (2013)
Neher, Hallatschek (2013)

**Rigorous work**: Berestycki, Berestycki, Schweinsberg (2013)
Schweinsberg (2015)

## **Basic properties of Λ-coalescents** (Pitman, 1999)

Suppose $(\Pi_\infty(t), t \geq 0)$ is a Λ-coalescent. Then:

1. Jump-hold property: let $T = \inf\{t : \Pi_\infty(t) \neq \Pi_\infty(0)\}$. If
$$\int_0^1 p^{-2}\,\Lambda(dp) < \infty,$$
then $P(T > 0) = 1$. Otherwise, $P(T = 0) = 1$.

2. Let $X_1(t) \geq X_2(t) \geq \ldots$ be the asymptotic frequencies of the blocks of the exchangeable random partition $\Pi_\infty(t)$. The coalescent has *proper frequencies* if $P(\sum_{k=1}^\infty X_k(t) = 1) = 1$ for all $t > 0$. This is equivalent to:
$$P(\{1\} \text{ is a block of } \Pi_\infty(t)) = 0 \quad \text{for all } t > 0.$$
Thus, the Λ-coalescent has proper frequencies if and only if
$$\int_0^1 p^{-1}\,\Lambda(dp) = \infty.$$

## Coming Down from Infinity

**Definition**: Suppose $\Pi_\infty$ is a $\Lambda$-coalescent. If $\#\Pi_\infty(t) = \infty$ for all $t > 0$, then we say the process *stays infinite*. If $\#\Pi_\infty(t) < \infty$ for all $t > 0$, then we say the process *comes down from infinity*.

**Theorem** (Pitman, 1999): If $\Lambda(\{1\}) = 0$, then the $\Lambda$-coalescent either comes down from infinity almost surely or stays infinite almost surely.

Let $T_n$ be the first time that $1, \dots, n$ are in the same block. Then $0 < T_2 \leq T_3 \leq \dots \uparrow T_\infty$. If $T_\infty < \infty$, then all positive integers are in the same block after time $T_\infty$.

For Kingman's coalescent, recall that

$$E[T_n] = \sum_{b=2}^{n} \binom{b}{2}^{-1} = 2 - \frac{2}{n},$$

which implies that $E[T_\infty] = 2$ and $T_\infty < \infty$ a.s.

Thus, Kingman's coalescent comes down from infinity.

Let

$$\lambda_b = \sum_{k=2}^{b} \binom{b}{k} \lambda_{b,k}$$

be the total rate of all mergers when the coalescent has $b$ blocks.

It is not true that the $\Lambda$-coalescent comes down from infinity if and only if $\sum_{b=2}^{\infty} \lambda_b^{-1} < \infty$ because $\sum_{b=2}^{n} \lambda_b^{-1}$ overestimates $E[T_n]$.

Let $\gamma_b$ be the rate at which the number of blocks is decreasing:

$$\gamma_b = \sum_{k=2}^{b} (k-1) \binom{b}{k} \lambda_{b,k}.$$

**Theorem** (Schweinsberg, 2000): Suppose $\Lambda(\{1\}) = 0$. Then the $\Lambda$-coalescent comes down from infinity if and only if

$$\sum_{b=2}^{\infty} \gamma_b^{-1} < \infty.$$

The Beta$(2-\alpha, \alpha)$ coalescent comes down from infinity if and only if $\alpha > 1$.

## Random walk methods

For the Beta($2-\alpha, \alpha$)-coalescent with $1 \leq \alpha < 2$. Probability that next merger causes the number of blocks to decrease by $k$ (Bertoin-Le Gall, 2006):

$$\lim_{b \to \infty} \binom{b}{k+1} \frac{\lambda_{b,k+1}}{\lambda_b} = \frac{\alpha \Gamma(k+1-\alpha)}{\Gamma(2-\alpha)\Gamma(k+2)}.$$

Let $V_1, V_2, \ldots$ be independent with

$$P(V_i = k) = \frac{\alpha \Gamma(k+1-\alpha)}{\Gamma(2-\alpha)\Gamma(k+2)} \sim C k^{-\alpha-1}, \qquad E[V_i] = \frac{1}{\alpha - 1}.$$

Suppose we begin with $n$ blocks at time zero.
Let $\tau_n$ be the number of mergers before only one block remains.
Let $X_k$ be the number of blocks remaining after $k$ mergers.

Use the approximation $X_k \approx n - (V_1 + \cdots + V_k)$.

If $n$ and $b$ are large, then $P(X_k = b \text{ for some } k) \approx \alpha - 1$.

## Functions of the block counting process

**Theorem:** (Kersting, Schweinsberg, Wakolbinger, 2014): Let $1 < \alpha < 2$, and suppose $f : (0,1] \to \mathbb{R}$ satisfies $|f'(x)| \leq Cx^{-\gamma}$, where $\gamma < 1 + 1/\alpha$. As $n \to \infty$:

$$n^{-1/\alpha}\left(\sum_{k=0}^{\tau_n - 1} f\left(\frac{X_k}{n}\right) - (\alpha - 1)n \int_0^1 f(x)\,dx\right) \Rightarrow Z,$$

where $Z$ has a stable law of index $\alpha$.

**Example:** (Kersting, 2012): Note that

$$L_n \approx \sum_{k=0}^{\tau_n - 1} \frac{X_k}{\lambda_{X_k}} \approx \alpha\Gamma(\alpha) \sum_{k=0}^{\tau_n - 1} X_k^{1-\alpha} = n^{1-\alpha}\alpha\Gamma(\alpha) \sum_{k=0}^{\tau_n - 1} \left(\frac{X_k}{n}\right)^{1-\alpha}.$$

Take $f(x) = \alpha\Gamma(\alpha)x^{1-\alpha}$ to get that if $1 < \alpha < (1+\sqrt{5})/2$, then

$$n^{-(1-\alpha+1/\alpha)}(L_n - cn^{2-\alpha}) \Rightarrow Z, \qquad c = \frac{\alpha(\alpha-1)\Gamma(\alpha)}{2-\alpha}$$

where $Z$ has a stable law of index $\alpha$.

# Additional Remarks

(Kersting, 2012): The total branch length $L_n$ has an asymptotic stable law when $\alpha = (1+\sqrt{5})/2$, but $L_n - cn^{2-\alpha}$ converges to a nondegenerate limit when $(1+\sqrt{5})/2 < \alpha < 2$.

(Dahmer, Kersting, Wakolbinger, 2014): The external branch length $L_{n,1}$ has an asymptotic stable law for $1 < \alpha < 2$.

(Berestycki, Berestycki, Limic, 2012):

$$\frac{L_{n,k}}{L_n} \to \frac{(2-\alpha)\Gamma(k+\alpha-2)}{\Gamma(\alpha-1)k!} \text{ a.s.}$$

**Theorem** (Drmota, Iksanov, Möhle, and Rösler, 2007): For the Bolthausen-Sznitman coalescent, as $n \to \infty$,

$$\frac{(\log n)^2}{\theta n}\left(L_n - \frac{\theta n}{\log n} - \frac{n \log \log n}{(\log n)^2}\right) \Rightarrow Z,$$

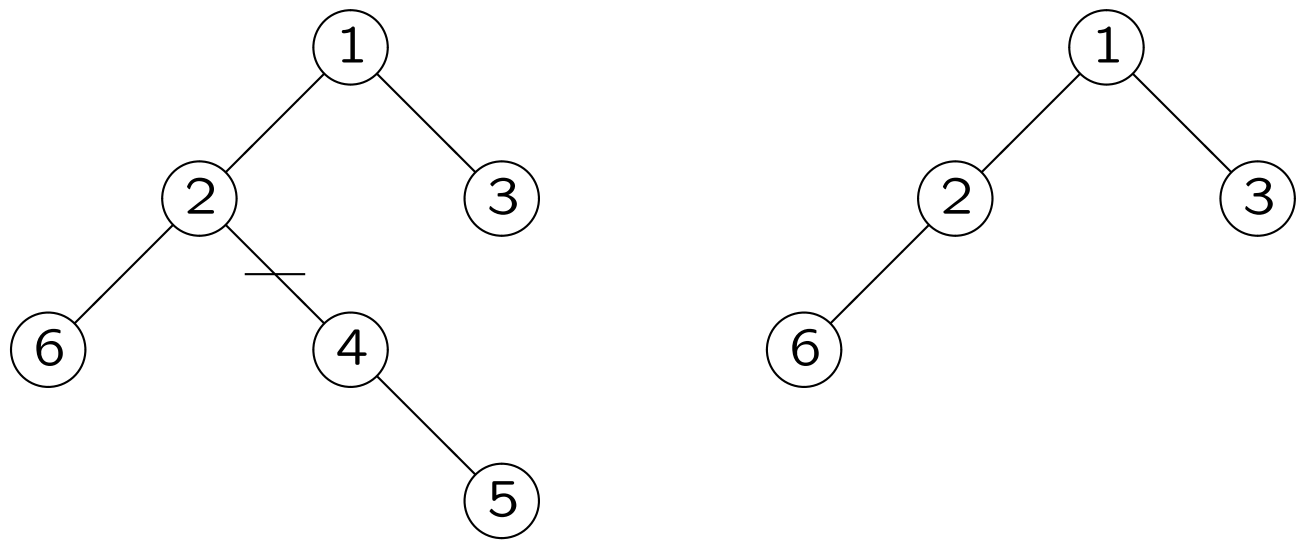
where $Z$ has a stable law of index 1.

See Basdevant and Golschmidt (2008) and Kersting, Pardo, and Siri-Jegousse (2014) for asymptotics about $L_{n,k}$.

## Random recursive trees

**Definition**: A tree on $n$ vertices labeled $1, \ldots, n$ is called a *recursive tree* if the root is labeled 1 and, for $2 \leq k \leq n$, the labels on the path from the root to $k$ are increasing.

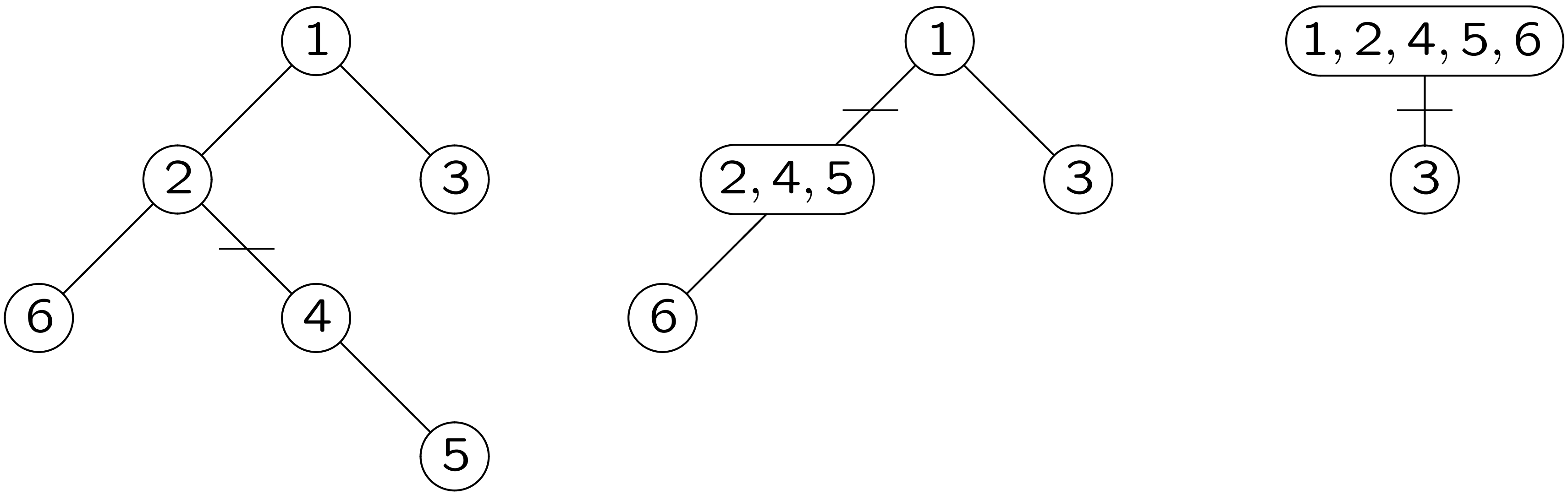
There are $(n-1)!$ recursive trees. To construct a random recursive tree, attach $k$ to one of the previous $k-1$ vertices uniformly at random.




Cutting procedure (Meir and Moon, 1974): Pick an edge at random, and delete it along with the subtree below it. What remains is a random recursive tree on the new label set.

## Connection with Bolthausen-Sznitman coalescent

**Theorem** (Goldschmidt and Martin, 2005): Cut each edge at the time of an exponential(1) random variable, and add the labels below the cut to the vertex above. The labels form a partition of $\{1, \ldots, n\}$ which evolves as a Bolthausen-Sznitman coalescent.




**Proof idea**: Given $\ell_1 < \cdots < \ell_k$, there are $(k-2)!$ recursive trees involving $\ell_2, \ldots, \ell_k$ and $(n-k)!$ recursive trees on the remaining vertices. The probability that $\ell_1, \ldots, \ell_k$ could merge is

$$\frac{(k-2)!(n-k)!}{(n-1)!} = \int_0^1 x^{k-2}(1-x)^{n-k}\, dx = \lambda_{n,k}.$$

## Time for $n$ blocks to merge into one

**Theorem** (Goldschmidt and Martin, 2005): Let $T_n$ be the time required for $n$ blocks in the Bolthausen-Sznitman coalescent to merge into one. For all $x \in \mathbb{R}$,

$$\lim_{n \to \infty} P\Big(T_n - \log\log n \leq x\Big) = e^{-e^{-x}}.$$

**Proof idea**: The last cut must involve one of the edges attached to the root. Because there are approximately

$$\sum_{k=2}^{n} \frac{1}{k-1} \approx \log n$$

such edges, $T_n$ behaves like the maximum of $\log n$ exponential(1) random variables. By extreme value theory, the mean is approximately $\log\log n$, and the asymptotic distribution is Gumbel.

## Other Constructions

Abraham and Delmas (2013) gave a combinatorial construction of Beta$(3/2, 1/2)$-coalescent by pruning a random binary tree.

Abraham and Delmas (2015) constructed the Beta$(2-\alpha,\alpha)$-coalescent for $0 < \alpha \leq 1/2$ by pruning a stable Galton-Watson tree with $n$ leaves.

Preprints listed at web page of Helmut Pitters (not yet available):

- "Lifting linear preferential attachment trees yields the arcsine coalescent"
- "Lifting random trees yields multiple merger coalescents"