# Dynamics of the evolving Bolthausen-Sznitman coalescent

by Jason Schweinsberg
University of California at San Diego

Outline of Talk

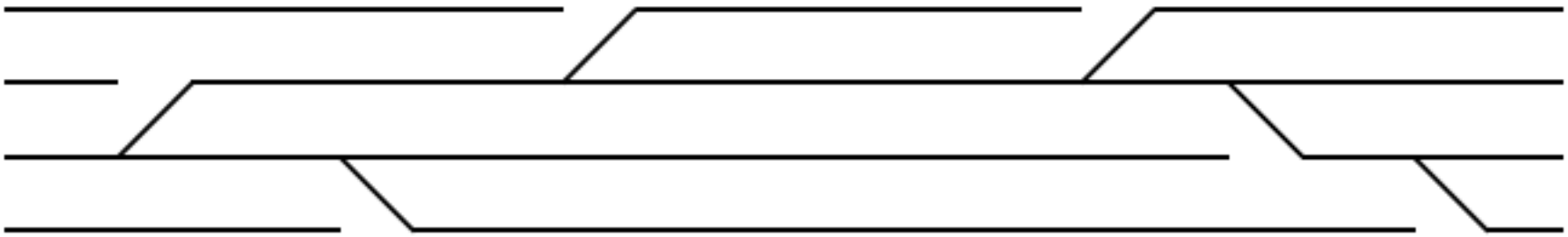
1. The Moran model and Kingman's coalescent
2. The evolving Kingman's coalescent
3. The evolving Bolthausen-Sznitman coalescent
4. Time back to the MRCA
5. Total tree length

## The Moran Model

Introduced by Moran (1958).

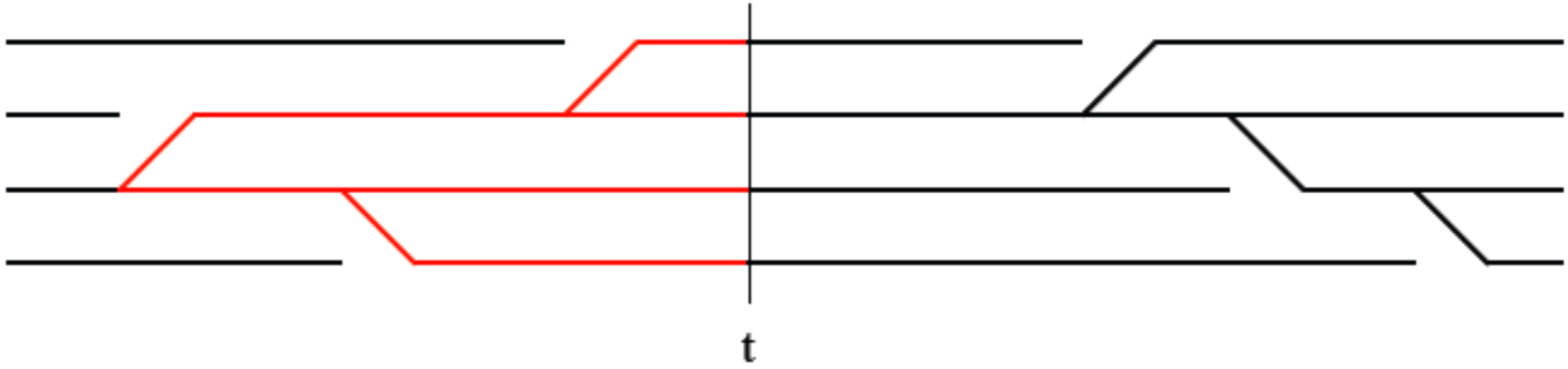
- The population has fixed size $n$.
- Each individual gives birth at times of rate 1 Poisson process.
- When a birth occurs, one individual picked at random to die.

The population can be defined at all times $t \in \mathbb{R}$.



## Genealogy of the Population

Fix $t \in \mathbb{R}$. For $s \geq 0$, let $\Pi_n(s)$ be the partition of $\{1, \ldots, n\}$ such that $i$ and $j$ are in the same block if and only if the $i$th and $j$th individuals at time $t$ have the same ancestor at time $t - s$.



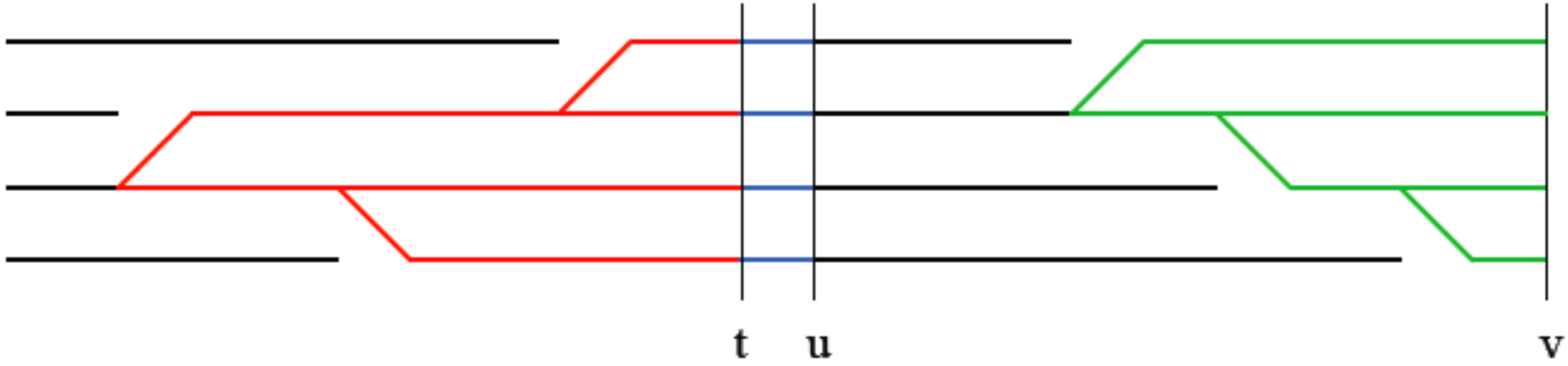

$$\{1\}, \{2\}, \{3\}, \{4\} \to \{1, 2\}, \{3\}, \{4\} \to \{1, 2\}, \{3, 4\} \to \{1, 2, 3, 4\}$$

The process $(\Pi_n(ns/2), s \geq 0)$ is Kingman's (1982) coalescent:

- Continuous-time Markov chain on set of partitions of $\{1, \ldots, n\}$.
- Only two blocks merge at a time.
- Each pair of blocks merges at rate one.

## The evolving Kingman's coalescent

Represent genealogy of the $n$ individuals at time $t$ by a tree $\mathcal{T}_n(t)$.




$\mathcal{T}_n(t) =$ red tree, $\mathcal{T}_n(u) =$ red and blue tree, $\mathcal{T}_n(v) =$ green tree.

**Theorem** (Greven-Pfaffelhuber-Winter, 2008): The processes $(\mathcal{T}_n(nt/2), t \in \mathbb{R})$ converge to a limit $(\mathcal{T}(t), t \in \mathbb{R})$ in the Skorohod topology (when the trees are viewed as random metric measure spaces equipped with the Gromov-weak topology).

**Definition**: $(\mathcal{T}(t), t \in \mathbb{R})$ is the *evolving Kingman's coalescent*.

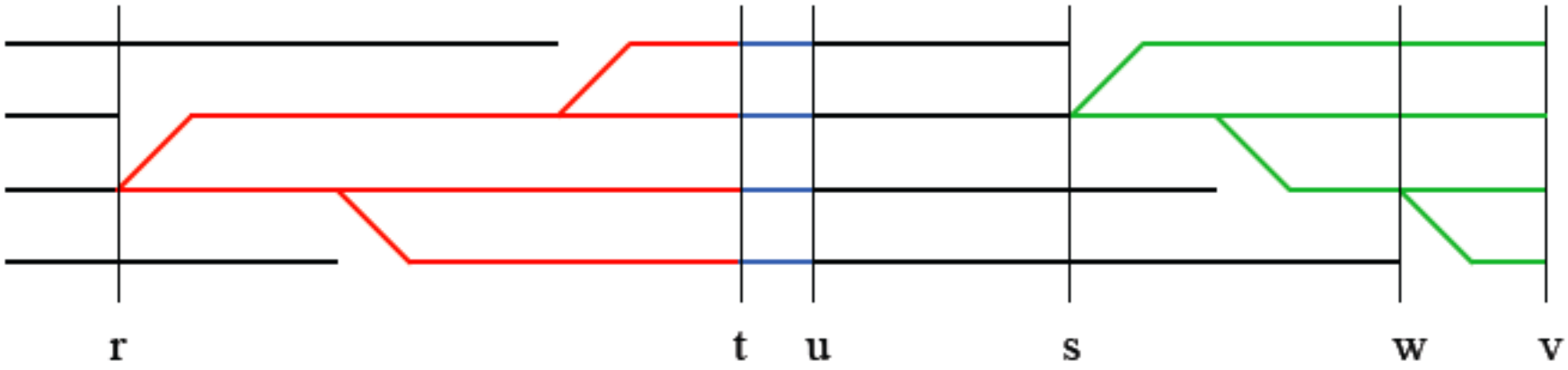
**Remark**: $(\mathcal{T}(t), t \in \mathbb{R})$ can also be constructed directly using the Donnelly-Kurtz (1999) lookdown construction.

## Time back to the MRCA

Let $A(t)$ denote the time back to the MRCA (most recent common ancestor) of the population at time $t$.

The process $(A(t), t \in \mathbb{R})$ is stationary.

The process $(A(t), t \in \mathbb{R})$ increases at speed 1 between jumps, jumps downward when one of the two oldest families dies out.




$$A(t) = t - r,\ A(u) = u - r,\ A(v) = v - s$$
$$A(w) - A(w-) = r - s.$$

## Time back to the MRCA: previous work

Pfaffelhuber and Wakolbinger (2006) showed:

- The process $(A(t), t \in \mathbb{R})$ is not Markovian.
- The jump times of $(A(t), t \in \mathbb{R})$ are a homogeneous Poisson process.
- If $Z(t)$ is the number of individuals that live before time $t$ and will become the MRCA of the population after time $t$, then $P(Z(t) = 0) = 1/3$ and $E[Z(t)] = 1$.
- If $L(t)$ is the number of individuals in the population that will have offspring when the next MRCA is established, then

$$P(L(t) = \ell) = \frac{2}{(\ell + 1)(\ell + 2)}.$$

Delmas, Dhersin, and Siri-Jegousse (2010) considered sizes of two oldest families.

Simon and Derrida (2006) considered correlation between time back to the MRCA and genetic diversity.

## Time back to MRCA: different model

Evans and Ralph (2010) considered population model such that:

- An immortal particle produces offspring at times of a rate one Poisson process.
- All descendants of the offspring die out after a time that has distribution $\mu$.

Example: The $\alpha$-stable CSBP conditioned on nonextinction satisfies these conditions when $1 < \alpha \leq 2$.

The process $(A(t), t \in \mathbb{R})$ is Markov. The jump rates and stationary distribution can be computed explicitly in terms of $\mu$.

## Total branch length

Let $L_n(t)$ be total length of branches of $\mathcal{T}_n(t)$, which is approximately proportional to number of mutations. Then

$$E[L_n(t)] = \sum_{k=2}^{n} k\binom{k}{2}^{-1} = \sum_{k=2}^{n} \frac{2}{k-1} \sim 2\log n$$

and

$$\lim_{n\to\infty} P\left(\frac{L_n(t)}{2} - \log n \le x\right) = e^{-e^{-x}}.$$

**Theorem** (Pfaffelhuber-Wakolbinger-Weisshaupt, 2010): The processes $((L_n(t) - 2\log n), t \in \mathbb{R})$ converge as $n \to \infty$ to a limit $(L(t), t \in \mathbb{R})$ in the Skorohod topology. The limit process is stationary and has infinite infinitesimal variance, with

$$\lim_{t\to 0} \frac{1}{t|\log t|} E\Big[(L(t) - L(0))^2\Big] = 4.$$

## A new population model

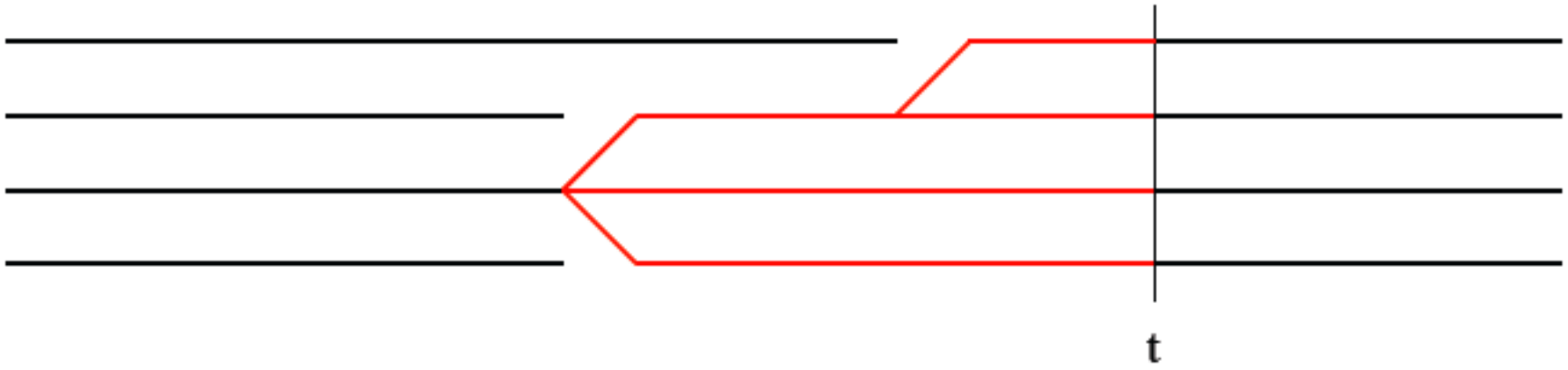
Consider the following population model:

- The population has fixed size $n$.
- Every individual gives birth at times of a Poisson process with rate $(n-1)/n$. The number $\xi$ of offspring produced satisfies

$$P(\xi = k) = \frac{n}{n-1} \cdot \frac{1}{k(k+1)}, \quad k = 1, \ldots, n-1.$$

- When $k$ offspring are produced, $k$ of the individuals previously in the population, other than the parent, are killed.

When $k-1$ individuals are born, $k$ ancestral lines merge at once.



## Bolthausen-Sznitman coalescent

Genealogy given by Bolthausen-Sznitman (1998) coalescent:

- When there are $n$ blocks, each possible merger of $k$ blocks, for $2 \leq k \leq n$, happens at rate

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

- Rate of $k$-mergers:

$$\binom{n}{k}\lambda_{n,k} = \frac{n}{k(k-1)}.$$

- Rate at which blocks are lost:

$$\gamma_n = \sum_{k=2}^{n} (k-1)\binom{n}{k}\lambda_{n,k} \approx n \log n.$$

Example of a $\Lambda$-coalescent: Pitman (1999), Sagitov (1999).

## Alternative construction

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 $\xi_i = 1$.

Construct a Poisson point process on $[0, \infty) \times (0, 1]$ with intensity

$$dt \times p^{-2}\, dp.$$

If $(t, p)$ is a point of this Poisson process, then a $p$-merger occurs at time $t$. For $2 \leq k \leq n$,

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

## Motivation for Bolthausen-Sznitman coalescent

Bertoin and Le Gall (2000): Describes the genealogy of Neveu's continuous-state branching process.

Bovier and Kurkova (2007): Describes genealogical structure in Derrida's GREM.

Brunet, Derrida, Mueller, and Munier (2007): Conjectured to describe the genealogy in population model with selection:

- Population has fixed size $n$.
- Each individual has $k \geq 2$ offspring, whose fitness is the parent's fitness plus an independent random variable.
- The $n$ offspring with the highest fitness survive to form the next generation.

Berestycki, Berestycki, and Schweinsberg (2010): Proved conjecture for branching Brownian motion with absorption.

## Evolving Bolthausen-Sznitman coalescent

Let $\mathcal{T}_n(t)$ be the tree representing the genealogy of the $n$ individuals in the population at time $t$.

Let $A_n(t)$ be the time back to the MRCA at time $t$.

Let $L_n(t)$ be the total length of all branches of $\mathcal{T}_n(t)$.

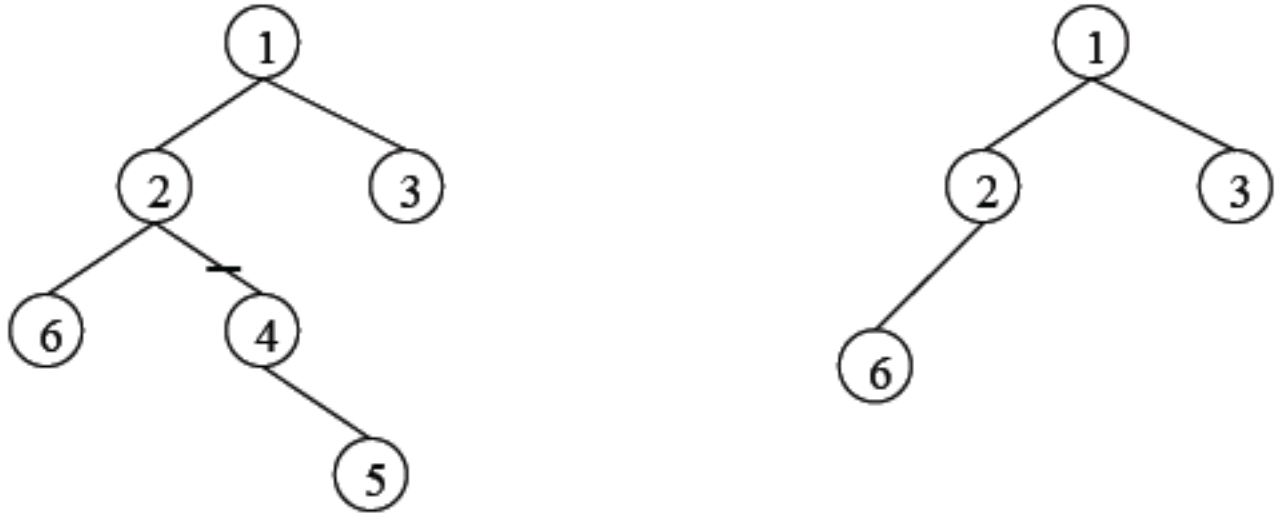
Goal: find the limits of $(A_n(t), t \in \mathbb{R})$ and $(L_n(t), t \in \mathbb{R})$, under suitable scaling, as $n \to \infty$.

Unlike for the evolving Kingman's coalescent, the limits will be Markov processes whose law can be described explicitly.

## 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.

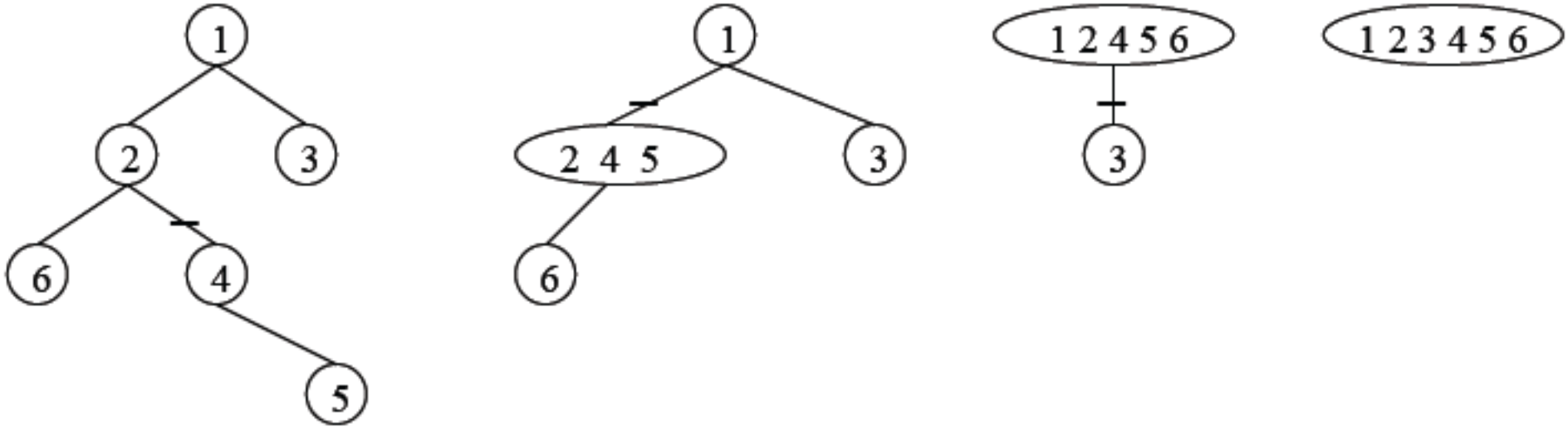
There are $(n-1)!$ recursive trees. To construct a random recursive tree, attach $k$ to one of the previous $k-1$ vertices uniformly at random.




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

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

## Time back to MRCA: non-evolving case

**Theorem** (Goldschmidt and Martin, 2005): For all $x \in \mathbb{R}$,

$$\lim_{n \to \infty} P\Big( A_n(0) - \log\log n > x \Big) = e^{-e^{-x}}.$$

**Proof idea**: The last cut must involve one of the edges attached to the root. Because there are approximately $\log n$ such edges, the time back to the MRCA behaves like the maximum of $\log n$ exponential(1) random variables.

**Remark**: Any limit of $((A_n(t) - \log\log n),\ t \in \mathbb{R})$ must be a stationary process whose stationary distribution is Gumbel.

## Constructing the (backwards) evolving coalescent

Backwards evolving coalescent: similar to ordinary coalescent, except that lineages that disappear get replaced by new ones.

Time evolution for trees on $n$ vertices:

- Begin with a random recursive tree with $n$ vertices and exponential random variables on the edges. Delete vertex labels.
- Edge random variables decrease linearly at speed 1. An edge is cut when its random variable hits zero.
- When $k$ vertices are cut away, attach $k$ new vertices, one at a time, to randomly chosen existing vertices.

Let $R_n(t)$ be the maximum of the exponential variables adjacent to root at time $t$. Then $(R_n(t), t \geq 0)$ evolves like $(A_n(-t), t \geq 0)$.

## Jump rates for time back to MRCA

The rate at which $(R_n(t), t \geq 0)$ jumps above $\log\log n + z$ is approximately the product of:

- The rate at which vertices are lost: $\gamma_n \approx n \log n$.
- The probability that the new vertex attaches to the root: approximately $1/n$.
- The probability that the new edge random variable exceeds $\log\log n + z$, which is $e^{-z}/(\log n)$.

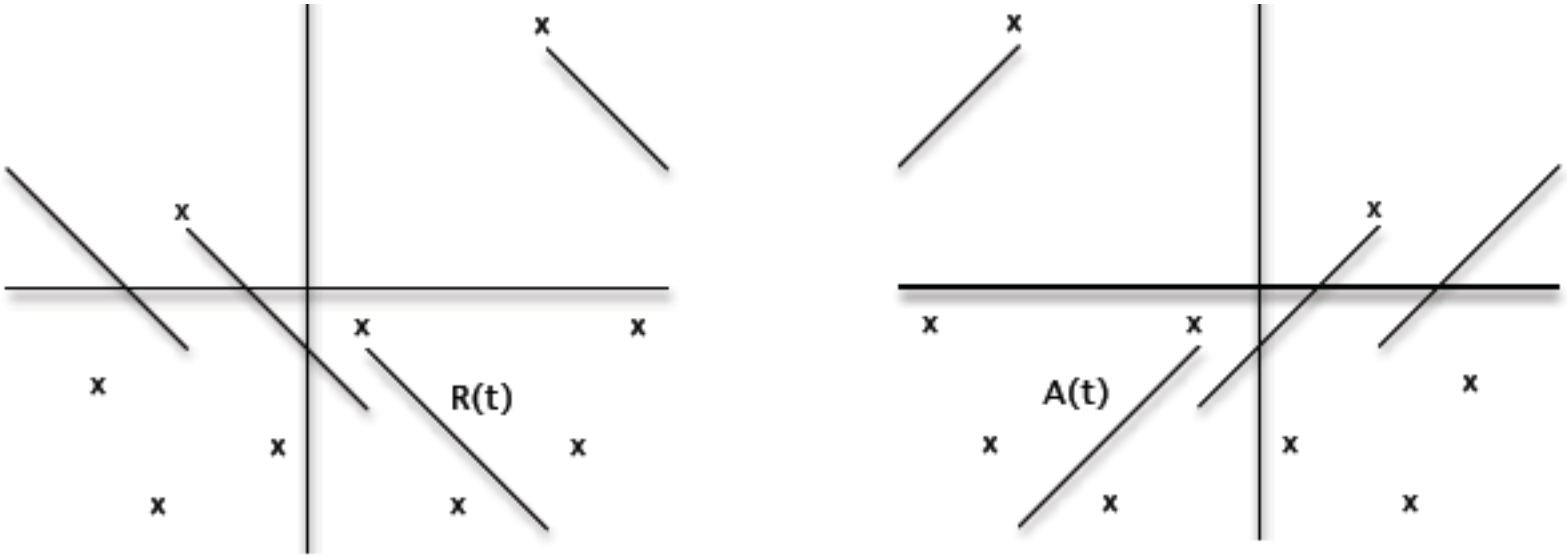
Thus, such a jump occurs at rate approximately $e^{-z}$.

## Limit Theorem for time back to MRCA

Construct a Poisson point process on $\mathbb{R}^2$ with intensity $dt \times e^{-x} dx$.

Define $(R(t), t \in \mathbb{R})$ to decrease at speed 1 between jumps, and jump up to the level of any mark that it encounters.

Define $(A(t), t \in \mathbb{R})$ to be a right-continuous process such that $A(t) = R(-t)$ when $R(t) = R(t-)$.

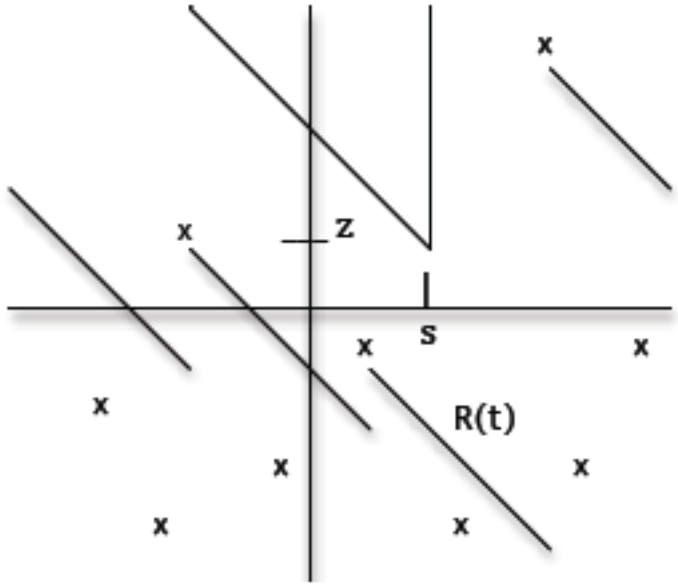


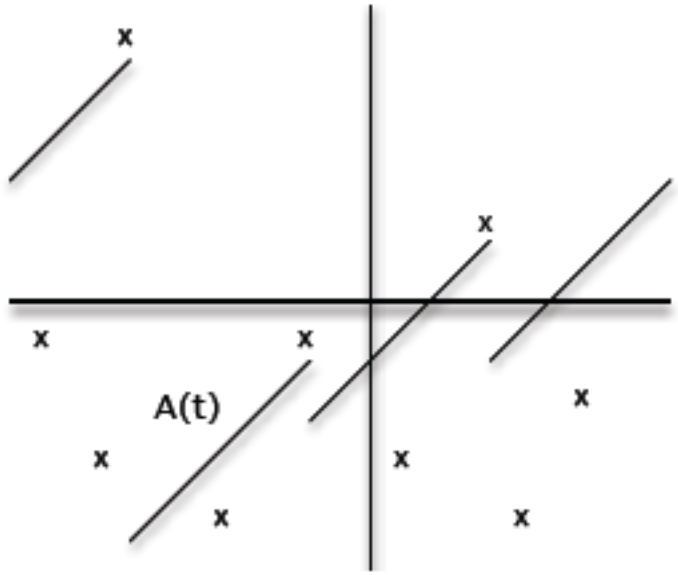

**Theorem** (Schweinsberg, 2011): As $n \to \infty$,

$$((A_n(t) - \log \log n), t \in \mathbb{R}) \Rightarrow (A(t), t \in \mathbb{R})$$

in the Skorohod topology.

## Remarks about the limit process




$(R(t), t \in \mathbb{R})$ and $(A(t), t \in \mathbb{R})$ are piecewise deterministic Markov processes (Davis, 1984) with jump rates

$$q(x,y) = e^{-y}\mathbf{1}_{\{y>x\}}, \qquad r(x,y) = e^{e^{-x}-e^{-y}-y}\mathbf{1}_{\{y<x\}}.$$

Jump times are rate 1 Poisson process: $\int_{-\infty}^{x} r(x,y)\,dy = 1.$

We have $R(s) \leq z$ if and only if there is no point in the triangle, which has probability $e^{-e^{-z}}$.

## Why is the limit Markovian?

Consider the backwards evolving coalescent. What happens to the time back to the MRCA when lineages merge and we replace them by new ones?

Conditional on the coalescent tree of $\{1, \ldots, n\}$, the time at which lineage $n+1$ merges with lineage 1 can be obtained as follows:

- Pick $J$ uniform on $\{1, \ldots, n\}$.
- If $J \geq 2$, $n+1$ merges with 1 at same time as $J$.
- If $J = 1$, they merge at independent Exponential(1) time.

If the time back to the MRCA is $T$, the probability that this time increases when a new lineage is added is $e^{-T}/n$.

These Ghirlanda-Guerra identities are a special property of the Bolthausen-Sznitman coalescent.

## Total tree length: non-evolving case

**Theorem** (Drmota, Iksanov, Möhle, Rösler, 2007): As $n \to \infty$,

$$\frac{(\log n)^2}{n}\left(L_n(0) - \frac{n}{\log n} - \frac{n \log\log n}{(\log n)^2}\right) \Rightarrow X,$$

where, using $\gamma$ to denote Euler's constant,

$$\begin{aligned} E[e^{iuX}] &= \exp\left(-\frac{\pi}{2}|u| + iu \log|u|\right) \\ &= \exp\left(iu(1-\gamma) - \int_{-\infty}^{0} \left(1 - e^{iux} + iux\mathbf{1}_{\{|x|\le 1\}}\right)x^{-2}\,dx\right). \end{aligned}$$

**Remark**: The limit of the tree length processes for the evolving coalescent must be a stationary process whose stationary distribution is stable.

## Generalized Ornstein-Uhlenbeck processes

Suppose $(Y(t), t \geq 0)$ is a Lévy process. The SDE

$$dX(t) = dY(t) - X(t)\, dt$$

has a unique strong solution. The solution is called a generalized Ornstein-Uhlenbeck process.

If $Y$ is Brownian motion, $X$ is the standard Ornstein-Uhlenbeck process, whose stationary distribution is Gaussian.

If $Y$ is a symmetric stable process, the stationary distribution of $X$ is the distribution of $Y(1)$.

In general, $X$ has a stationary distribution if and only if

$$\int_{|x|>2} \log |x|\, \nu(dx) < \infty,$$

where $\nu$ is the Lévy measure of $Y$. In this case, the stationary distribution of $X$ is infinitely divisible.

## Limit Theorem for total tree length

**Theorem** (Schweinsberg, 2011): Let $(Y(t), t \geq 0)$ be the Lévy process such that

$$E[e^{iuY(t)}] = \exp\left(iu(2-\gamma)t - t\int_{-\infty}^{0}\left(1 - e^{iux} + iux\mathbf{1}_{\{|x|\leq 1\}}\right)x^{-2}\,dx\right).$$

Let $(X(t), t \in \mathbb{R})$ be a stationary generalized Ornstein-Uhlenbeck process satisfying

$$dX(t) = dY(t) - X(t)\,dt.$$

Let

$$W_n(t) = \frac{(\log n)^2}{n}\left(L_n\left(\frac{t}{\log n}\right) - \frac{n}{\log n} - \frac{n\log\log n}{(\log n)^2}\right).$$

Then $(W_n(t), t \in \mathbb{R}) \Rightarrow (X(t), t \in \mathbb{R})$ in the Skorohod topology as $n \to \infty$.

**Remarks**:

- The distribution of $X(t)$ is the distribution of $Y(1) - 1$.
- The "mixing time" of $L_n$ is $O(1/(\log n))$.

## Number of Blocks

**Theorem** (Schweinsberg, 2011): Suppose $(\Pi_n(t), t \geq 0)$ is a Bolthausen-Sznitman coalescent started with $n$ blocks. Let $N_n(t)$ be the number of blocks of $\Pi_n(t)$, and let

$$X_n(t) = \frac{\log n}{n}\left(N_n\left(\frac{t}{\log n}\right) - ne^{-t} - \frac{nte^{-t}\log\log n}{\log n}\right).$$

Let $(S(t), t \geq 0)$ be a stable Lévy process satisfying

$$E[e^{iuS(t)}] = \exp\left(-\frac{\pi t}{2}|u| + itu\log|u|\right).$$

As $n \to \infty$, the sequence of processes $(X_n(t), t \geq 0)$ converges in the Skorohod topology to

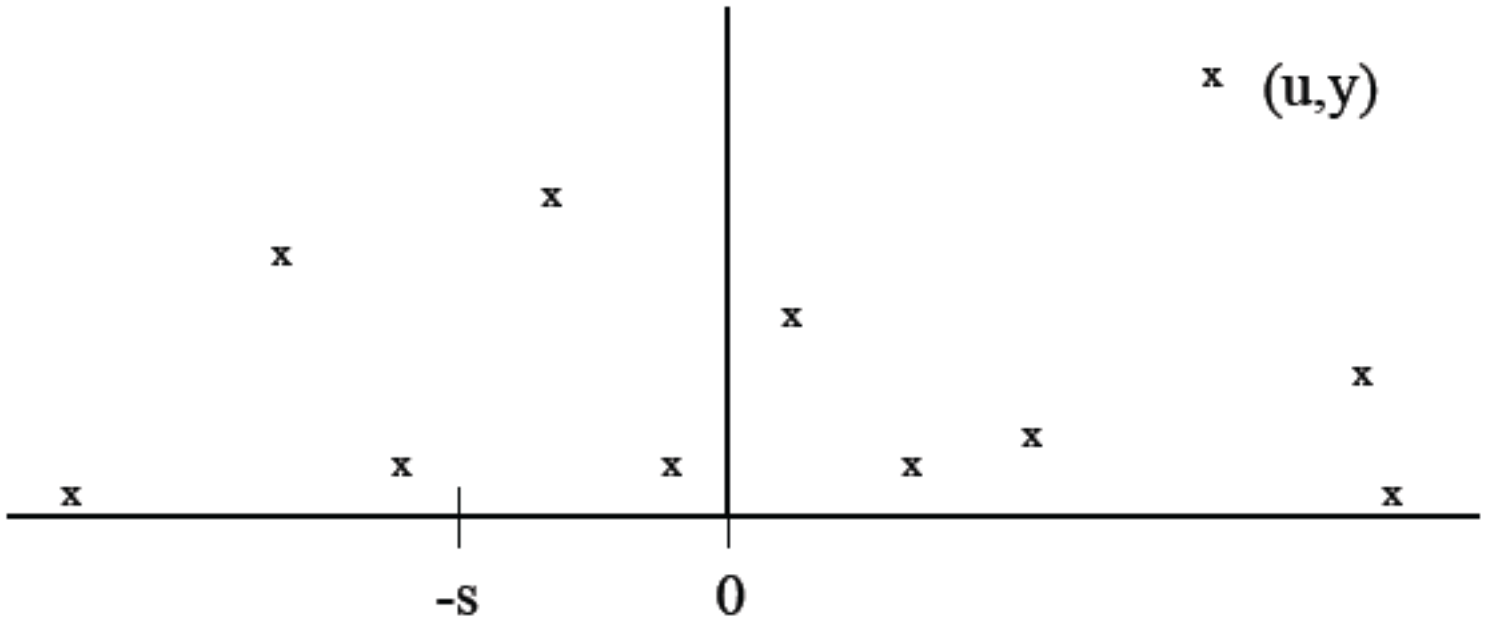
$$\left(e^{-t}S(t) + \frac{e^{-t}t^2}{2},\ t \geq 0\right).$$

## A Family of Stable Processes

Let $\Psi$ be a PPP on $\mathbb{R} \times (0, \infty)$ with intensity $dt \times y^{-2}\, dy$.

For $s \geq 0$, let $(S(s,t), t \geq 0)$ be a stable Lévy process such that if $(u, y)$ is a point of $\Psi$ with $u > -s$, then

$$S(s, u+s) - S(s, (u+s)-) = -y.$$



## Coupling Stable Processes with the Population

Let $\Theta$ be image of $\Psi$ under $(t, y) \mapsto (-t/\log n, y/\log n)$ restricted to $(0, 1]$, which is a PPP on $\mathbb{R} \times (0, 1]$ with intensity $dt \times y^{-2}\, dy$.

If $(t, y)$ is a point of $\Theta$, then at time $t$, flip $y$-coin for each individual. If there are $k \geq 2$ heads, $k$ lineages merge at time $t$.

Let $N_n(s, t)$ be the number of individuals at time $s - t$ with a descendant alive at time $s$, and let

$$X_n(s, t) = \frac{\log n}{n}\left(N_n\left(\frac{s}{\log n}, \frac{t}{\log n}\right) - ne^{-t} - \frac{nte^{-t}\log\log n}{\log n}\right).$$

If $(u, y)$ is a point of $\Psi$ with $u > -s$, then

$$N_n\left(\frac{s}{\log n}, \frac{u+s}{\log n}\right) - N_n\left(\frac{s}{\log n}, \frac{u+s}{\log n}-\right) \approx -\frac{y}{\log n} N_n\left(\frac{s}{\log n}, \frac{u+s}{\log n}-\right).$$

Therefore, $X_n(s, u+s) - X_n(s, (u+s)-) \approx -e^{-(u+s)}y$.

## Coupling Bounds

**Proposition**: Fix $s \geq 0$. Let $T_n = 2 \log \log n$. Then

$$\sup_{0 \leq t \leq T_n} \left| X_n(s,t) - \left( e^{-t} S(s,t) + \frac{e^{-t} t^2}{2} \right) \right| \to_p 0.$$

Let

$$W_n(s) = \int_0^\infty X_n(s,t)\, dt$$

and

$$L(s) = \int_0^\infty \left( e^{-t} S(s,t) + \frac{e^{-t} t^2}{2} \right) dt = 1 + \int_0^\infty e^{-t} S(s,t)\, dt.$$

Then

$$dL(s) = dS(s) - L(s)\, ds,$$

so $(L(s), s \geq 0)$ is a generalized Ornstein-Uhlenbeck process.

**Proposition**: Fix $T > 0$. Then

$$\sup_{0 \leq s \leq T} |W_n(s) - L(s)| \to_p 0.$$