Evolving genealogies in neutral populations

Peter Pfaffelhuber

University of Munich,
Joint work with Anita Winter and Andreas Greven (Erlangen)

May 27th, 2007

Neutral models of constant size

Populations of constant size have been **modelled** by

- **Markov Chains** (Wright-Fisher-model, Moran model)
- **Diffusion approximations** (Fisher-Wright diffusion)

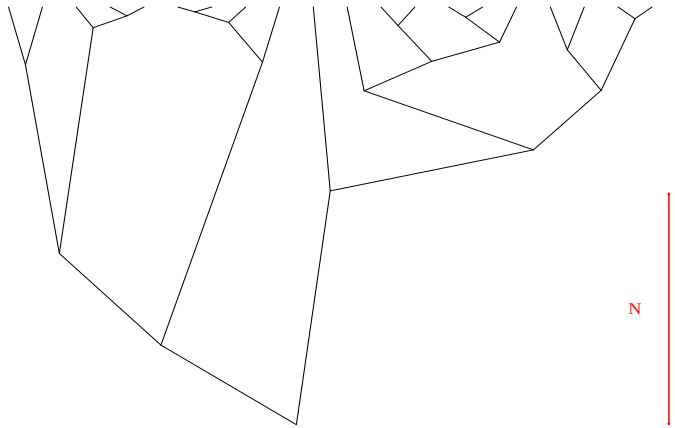
$$dX = \sqrt{X(1-X)}dW.$$

- **Measure-valued diffusions** (Fleming-Viot superprocess)
  Process $(\mu_t)_{t\geq 0}$ with state space $\mathcal{P}(K)$ such that $(\mu_t(A))_{t\geq 0}$ follows a Wright-Fisher diffusion.

Kingman's coalescent

Genealogies relating individuals are known to be distributed according to **Kingman's coalescent**:

- start with n lines
- if there are k lines the coalescence rate is $\binom{k}{2}$.
- stop when reaching one line.

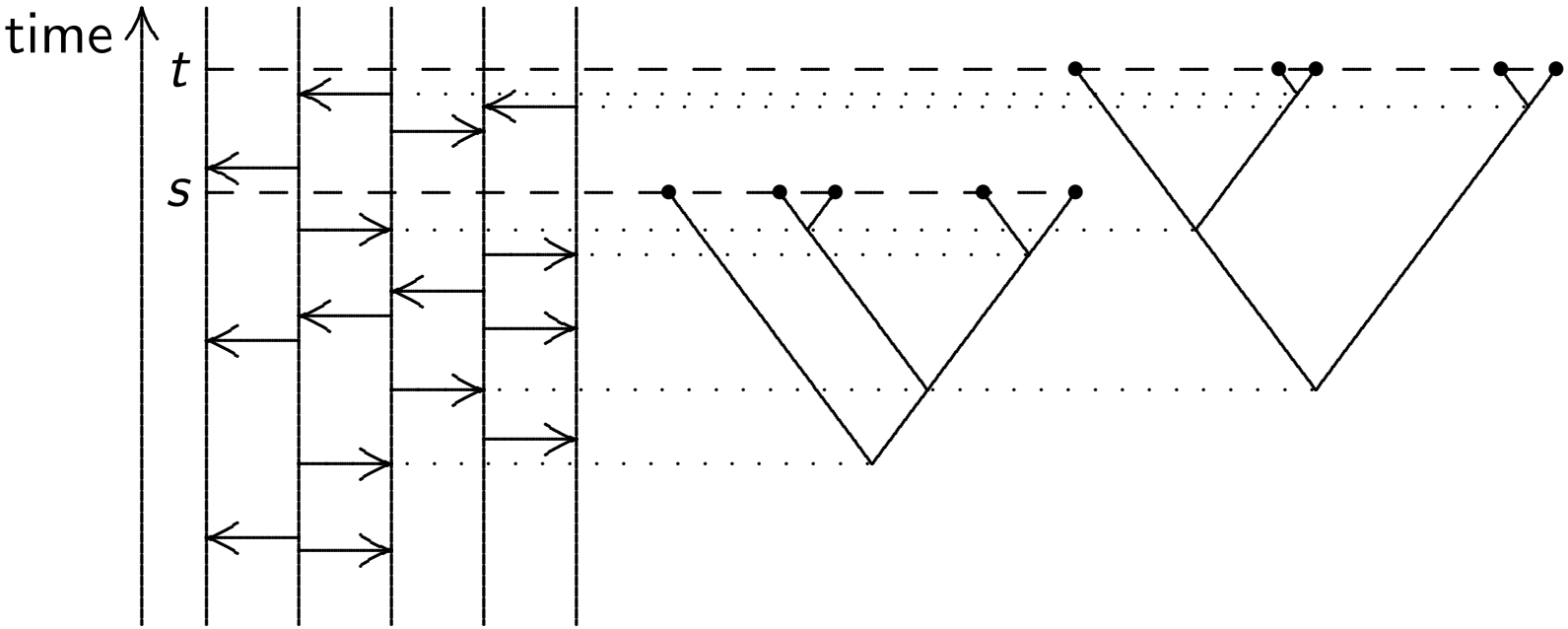


Example: Allelic distribution of a sample of size n at time t is uniquely given by the allelic distribution of ancestors at time 0

Evolving coalescents

Goal: construct a tree-valued stochastic process $\mathcal{U} = (\mathcal{U}_t)_{t \geq 0}$ which

- describes genealogical relationships **dynamically**
- makes **forward** and **coalescent** picture implicit



Formalizing genealogical trees

- **Leaves in Kingman's coalescent** form an ultrametric space.

State space of $\mathcal{U}$:

$\mathbb{U} :=$ {isometry class of (U, r, μ) :
(U, r) complete and separable **ultrametric** space, $\mu \in \mathcal{P}(U)$}.

- $r(u, v)$ defines the genealogical distance of individuals u and v
- μ marks currently living individuals

The Gromov-weak topology on $\mathbb{U}$

- Π: Algebra generated by functions on $\mathbb{U}$ of the form

$$\Phi(U, r, \mu) := \int \varphi(r(\underline{u}, \underline{u}))\mu^{\otimes n}(\mathrm{d}\underline{u})$$

for $\underline{u} = (u_1, \ldots, u_n), \varphi \in \mathcal{C}_b(\mathbb{R}^{\binom{n}{2}})$

Π **separates points** in $\mathbb{U}$.

- **Gromov-weak topology**:

$$(U_n, r_n, \mu_n) \to (U, r, \mu) \iff \forall \Phi \in \Pi : \Phi(U_n, r_n, \mu_n) \to \Phi(U, r, \mu)$$

- **Proposition** (Gromov; Vershik; Greven, P, Winter)
  - The space $(\mathbb{U}, \mathcal{O}_{\mathbb{U}})$ is **Polish**.

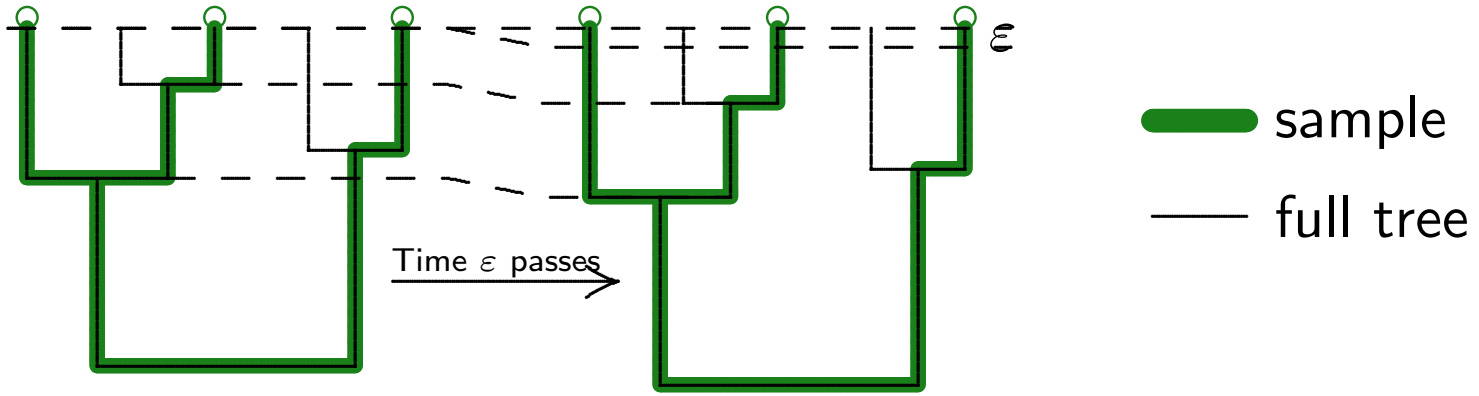
Martingale Problem

- Given: operator Ω on Π
  $\mathcal{U} = (\mathcal{U}_t)_{t\geq 0}$ solution of the (Ω, Π)-**martingale problem** if for all $\Phi \in \Pi$,
  $$\left(\Phi(U_t) - \int_0^t \mathrm{d}s\, \Omega\Phi(U_s)\right)_{t\geq 0}$$
  is a martingale. It is **well-posed** if there is exactly one such process.

- **Program**:
  - Define pregenerator for finite populations.
  - Establish **existence** by generator convergence and tightness.
  - Establish **uniqueness** by duality.

Tree Growth

When no resampling occurs the tree **grows**



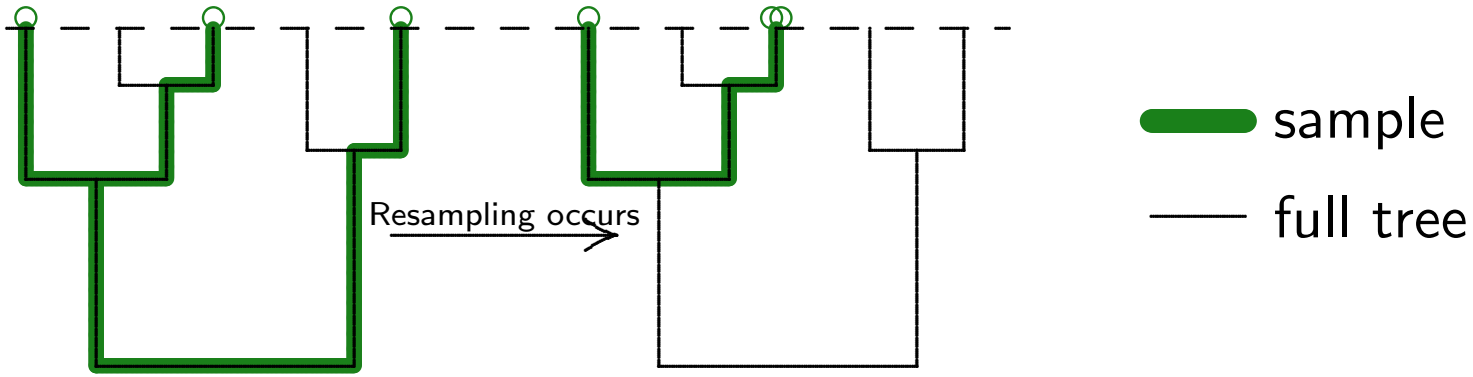
Distances in the sample grow

$$\Omega^N_{\text{grow}}\Phi(U,r,\mu) = \sum_{i,j}\int \frac{\partial\varphi}{\partial r_{ij}}\big(r(\underline{u},\underline{u})\big)\mu^{\otimes n}(d\underline{u}) + \mathcal{O}\big(\tfrac{1}{N}\big)$$

- $\frac{\partial\varphi}{\partial r_{ij}}$: measures the **change of** φ when $r(u_i,u_j)$ grows
- $\mathcal{O}\big(\frac{1}{N}\big)$: due to double sampling of the same individual

Resampling

Resampling does not change the tree



By resampling μ changes, so **different samples are picked**.

$$\Omega_{\text{res}}^N \Phi(U, r, \mu) = \tfrac{1}{2} \sum_{k,l} \int \left(\varphi\left(\theta_{kl} r(\underline{u}, \underline{u})\right) - \varphi\left(r(\underline{u}, \underline{u})\right)\right) \mu^{\otimes n}(d\underline{u}) + \mathcal{O}\left(\tfrac{1}{N}\right)$$

where $$\theta_{kl} r(u_i, u_j) = \begin{cases} r(u_k, u_j), & i = l \\ r(u_i, u_k), & j = l \\ r(u_i, u_j), & \text{else} \end{cases}$$

Martingale Problem

- **Pregenerator** for infinite system:

$$\Omega\Phi(U,r,\mu) = \sum_{i,j} \int \frac{\partial \varphi}{\partial r_{i,j}}\big(r(\underline{u},\underline{u})\big)\mu^{\otimes n}(d\underline{u})$$
$$+ \tfrac{1}{2}\sum_{k,l} \int \big(\varphi\big(\theta_{kl} r(\underline{u},\underline{u})\big) - \varphi\big(r(\underline{u},\underline{u})\big)\big)\mu^{\otimes n}(d\underline{u})$$

- Existence: **Approximation** by finite systems
- Uniqueness: **Duality** to Kingman's coalescent

The tree-valued Fleming-Viot dynamics

Theorem

- Solution $\mathcal{U} = (\mathcal{U}_t)$ of (Ω, Π)-mp exists as limit of tree-valued Moran models and is unique
- Almost surely,
  - $\mathcal{U}$ has **continuous** sample paths
  - $\mathcal{U}_t$ **compact** for all $t > 0$
  - **Quadratic variation** of $\Phi(\mathcal{U})$:

$$\mathrm{d}\langle \Phi(\mathcal{U}) \rangle_t = n^2 \langle \mu_t, \left(\rho - \langle \mu_t, \rho \rangle\right)^2 \rangle \mathrm{d}t,$$

where

$$\rho(u_1) := \int \mu_t^{\otimes(n-1)} \left(\mathrm{d}(u_2, \ldots, u_n) \phi((r(u_i, u_j))_{1 \le i < j \le n}\right)$$

The tree-length process

- Given (U, r, μ) and $u_1, u_2, \ldots$,

$$L_n(\underline{u}) := \text{length of tree spanned by } u_1, \ldots, u_n$$

- **Tree length distribution** of subsequentially sampled points:

$$\Lambda(U, r, \mu) := (L_2, L_3, \ldots)_* \mu^{\otimes \mathbb{N}} \in \mathcal{P}\left(\mathbb{R}_+^{\mathbb{N}}\right)$$

- **Theorem**: The process $(\Lambda(\mathcal{U}_t))_{t \geq 0}$ is **Markov**

Application: number of mutations on the tree

- Real data: **number of mutations** $S_n(t)$ of a sample at time t can be observed
- Length of sample tree $L_n(\underline{u}) \Rightarrow S_n \sim \textbf{Pois}\big(\frac{\theta}{2} L_n(\underline{u})\big)$
- So,

$$\mathbb{E}\big[e^{-\lambda S_n(t)}\big] = g^n\big(t; \theta(1 - e^{-\lambda})\big).$$

for

$$g^n(t;\theta) := \mathbb{E}\Big[\int \mu_t^{\otimes n}(d\underline{u}) \exp\big(-\theta L_n(\underline{u})\big)\Big]$$

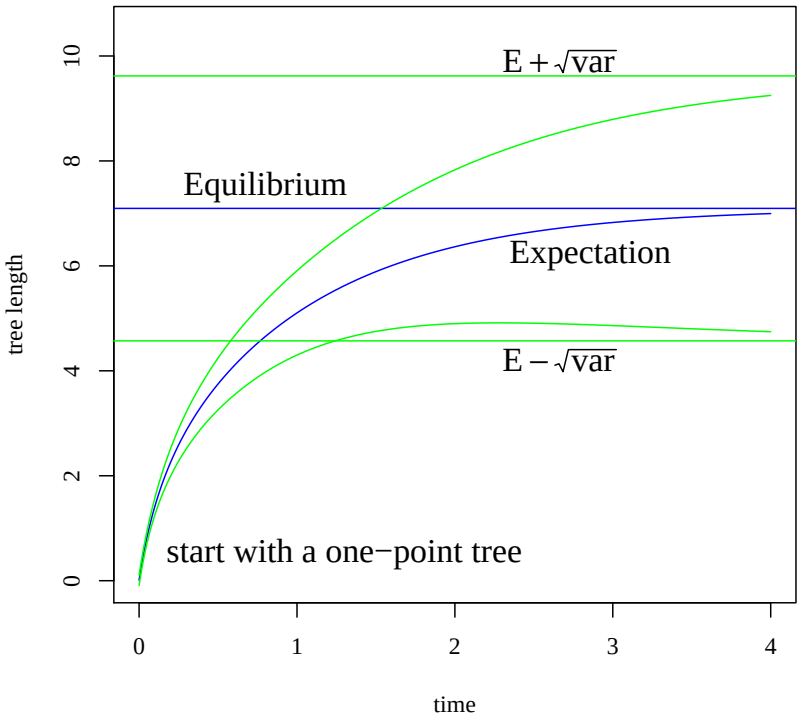
- $(g^2, g^3, \ldots)$ is solution of

$$\frac{d}{dt} g^n = -n\theta g^n + \binom{n}{2} \left(g^{n-1} - g^n \right)$$

- $(g^2, g^3, \ldots)$ given by

$$g^n(t;\theta) = \Gamma(n) \sum_{k=2}^{n} \frac{\binom{n}{k}(-1)^k(\theta+2k-1)}{\Gamma(\theta+n+k)} \cdot$$
$$\cdot \Big\{ e^{-k(\theta+(k-1))t} \sum_{m=2}^{k} \frac{\binom{k}{m}(-1)^m \Gamma(\theta+k+m-1)}{\Gamma(m)} g^m(0;\theta)$$
$$+ \big(1 - e^{-k(\theta+(k-1))t}\big)(k-1)(\theta+k)\Gamma(\theta+k-1) \Big\}.$$

Application: tree lengths



Example: consider a sample of $n = 20$ which starts in $(\{\bullet\}, \delta_\bullet)$. Moments of tree lengths can be calculated.

Outlook

Further questions

- Is it possible to include **mutation**, **selection**, **recombination**?
- What does the mp for tree-valued **branching** processes look like?
- Which tree-evolutions can be defined via **sample evolution**?