Coalescent point process and applications to the size of large families in general branching processes

Nicolas Champagnat[1] Amaury Lambert[2]

[1]IECN & INRIA [2]UPMC, LPMA

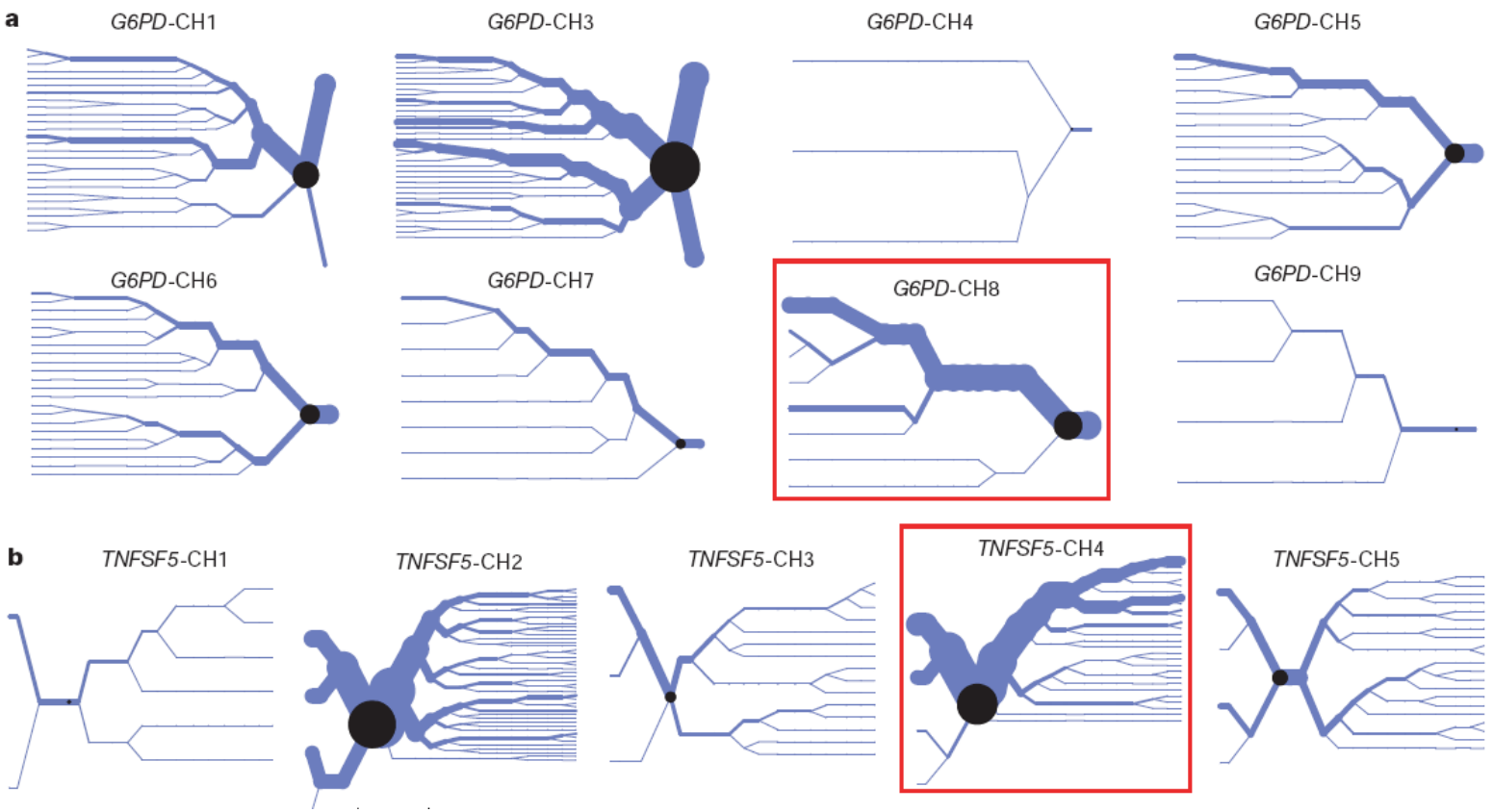
SMEEG conference, Angers, 9 December 2013

Branching processes with neutral mutations: a biological motivation

- In a branching process, each individual behaves independently of the others $\rightsquigarrow$ no interaction between individuals
- Example: assume a new allele of a gene appeared recently, positively selected
  - small, increasing population, with little interaction
  - recombinations may occur on the DNA sequence around the gene $\rightsquigarrow$ no influence on the selected allele, so recombination = neutral mutations
- Biologists might want to detect if a particular allele is currently positiviely selected
  - take a sample of holders of this allele
  - look at the recombination events that can be detected in the sample on the DNA sequence around the gene
  - $\rightsquigarrow$ recombination tree (Sabeti et al., Nature 2002)

Examples of recombination tree (Sabeti et al., Nature 2002)

Try to infer the growth rate of the population from data of the form

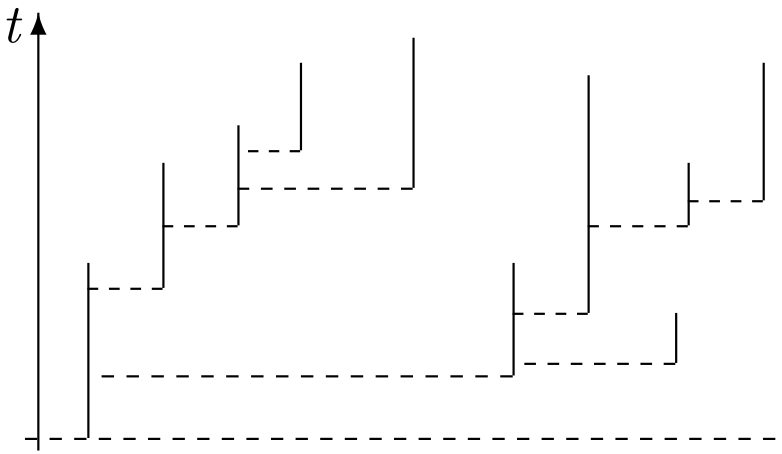


Branching processes with mutations

- Yule (1924): pure-birth process, species and genera
- Griffiths & Pakes (1988): Galton–Watson tree and independent mutations with fixed probability
- Jagers & Nerman (1981–1984), Taïb (1992): general branching process, mutation at birth
- Abraham & Delmas (2007): continuous-state branching processes, all mutants have the same type
- Bertoin (2009, 2010, 2011): Galton–Watson, allelic partition of total descendance
- Sagitov & Serra (2009, 2011): waiting time to n-th mutation

Splitting tree forward in time (Geiger & Kersting 97)

We consider an asexual population where

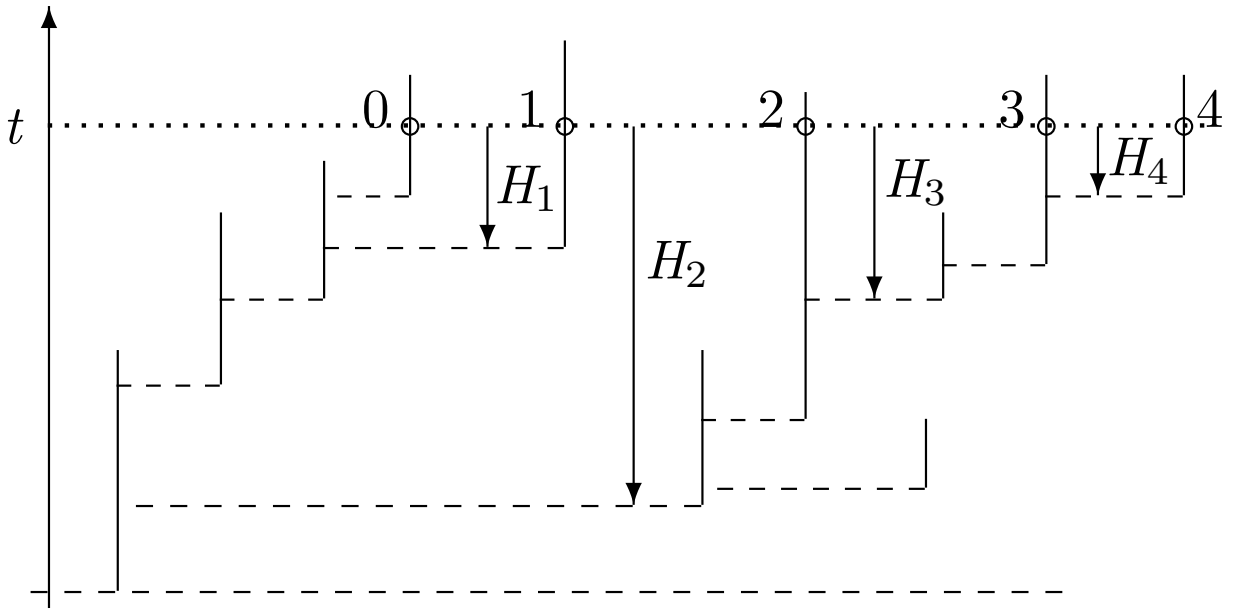


- individuals reproduce independently
- have i.i.d. lifetime durations distributed as some r.v. V
- during which they give birth at constant rate b

- The law of this so-called splitting tree is characterized by the finite measure $\Lambda(dr) := b\mathbb{P}(V \in dr)$
- The population size process $(N_t; t \geq 0)$ is a non-Markovian branching process called (homogeneous, binary) Crump–Mode–Jagers process.

Representation backward in time

Starting from one single individual, the subtree spanned by the individuals alive at time t can be represented as follows

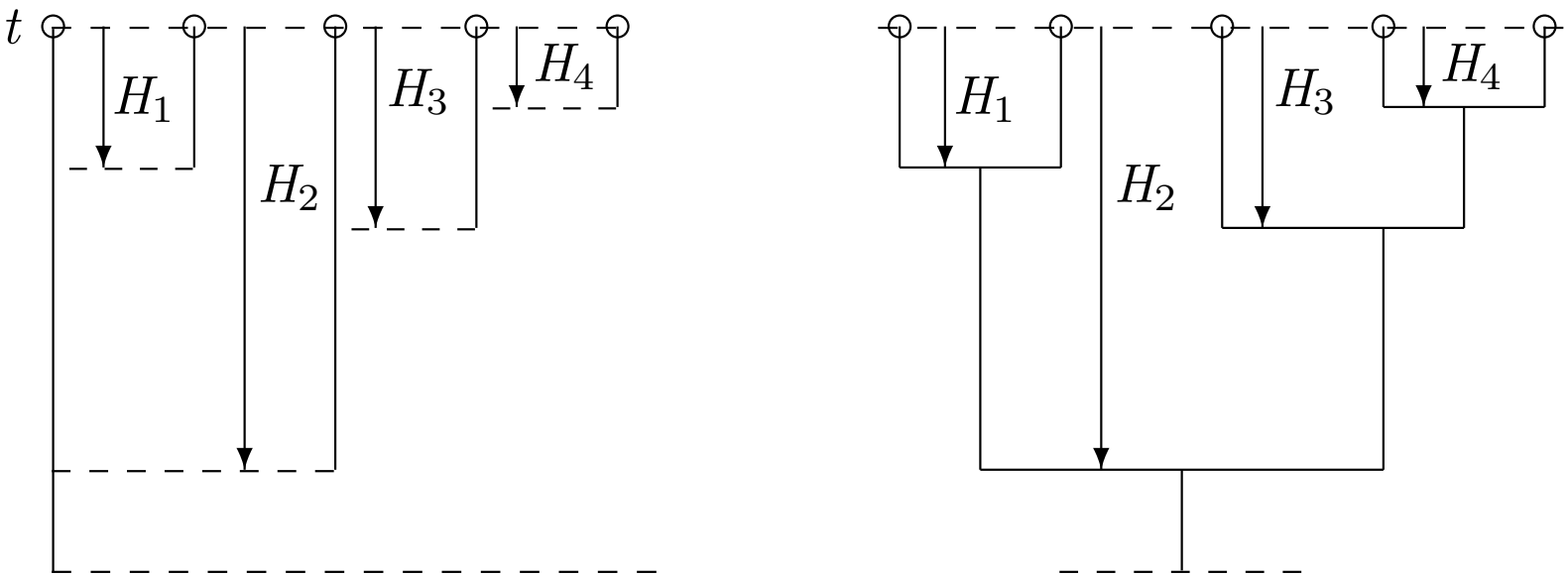


where the times $H_1, H_2, H_3 \ldots$ are called coalescence times.

Representation backward in time

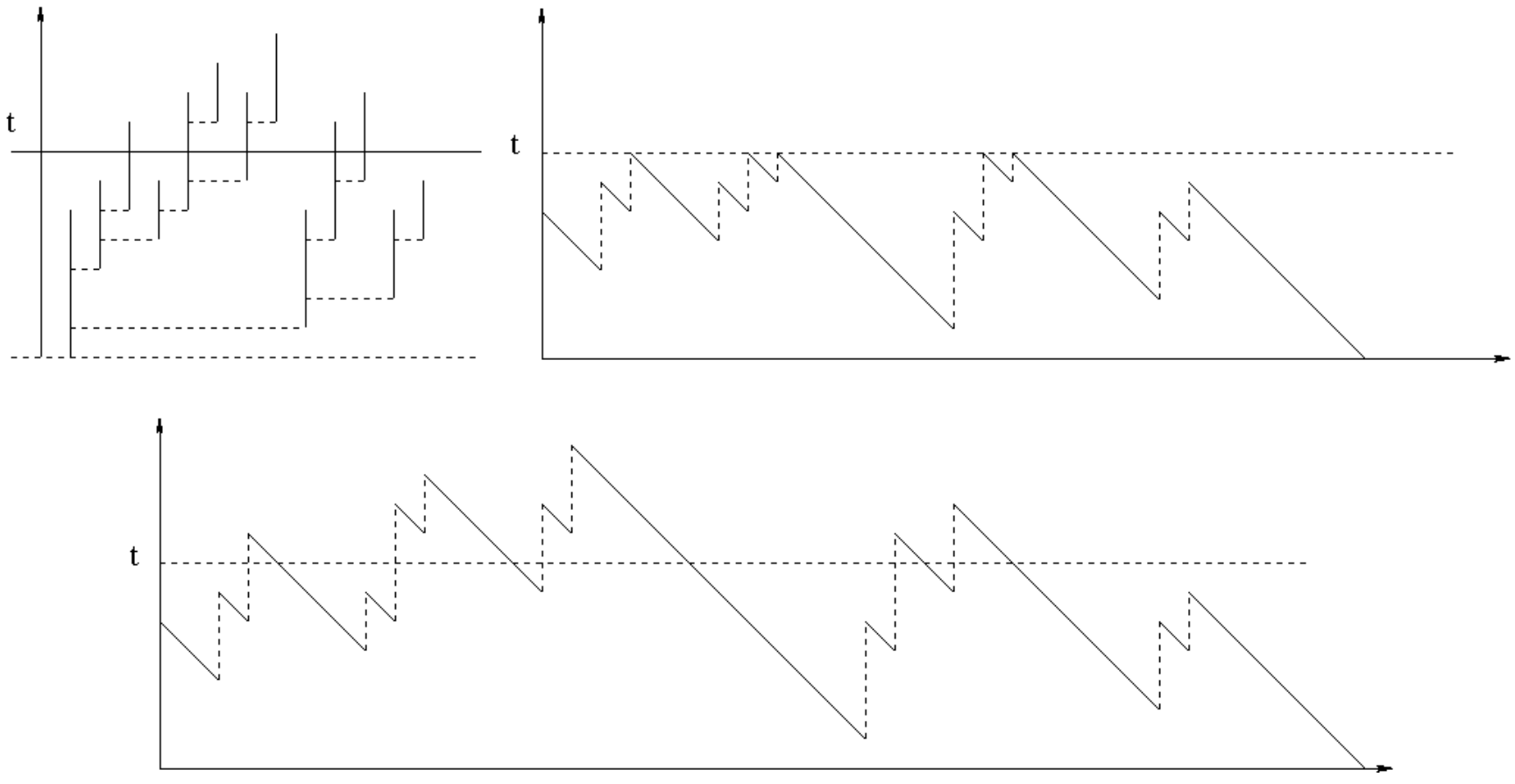
This subtree can also be represented as this...

...or (as usual) this



Contour of a splitting tree

A splitting tree and the jumping contour process of its truncation below time t.



First result

Theorem (Lambert (2010))

The jumping contour of a splitting tree truncated below time t is a strong Markov process. It is composed of successive excursions below t of a Lévy process without negative jumps with Laplace exponent

$$\psi(x) = x - \int_{(0,\infty]} (1 - e^{-rx})\Lambda(dr).$$

As a consequence, conditionally on $N_t \neq 0$, the coalescence times $H_1, H_2, H_3 \ldots$ of the splitting tree form a sequence of i.i.d. positive random variables killed at its first value larger than t.
In addition,

$$\mathbb{P}(H > x) = \frac{1}{W(x)}.$$

where W is the scale function of the Lévy process, positive, increasing, s.t. $W(0) = 1$ and the Laplace transform of W is $1/\psi$.

Examples

- Yule process with (birth) rate b

$$W(x) = e^{bx}$$

- Noncritical birth–death processes with birth rate b, death rate d, growth rate $r := b - d$

$$W(x) = 1 + \frac{b}{r}\left(e^{rx} - 1\right)$$

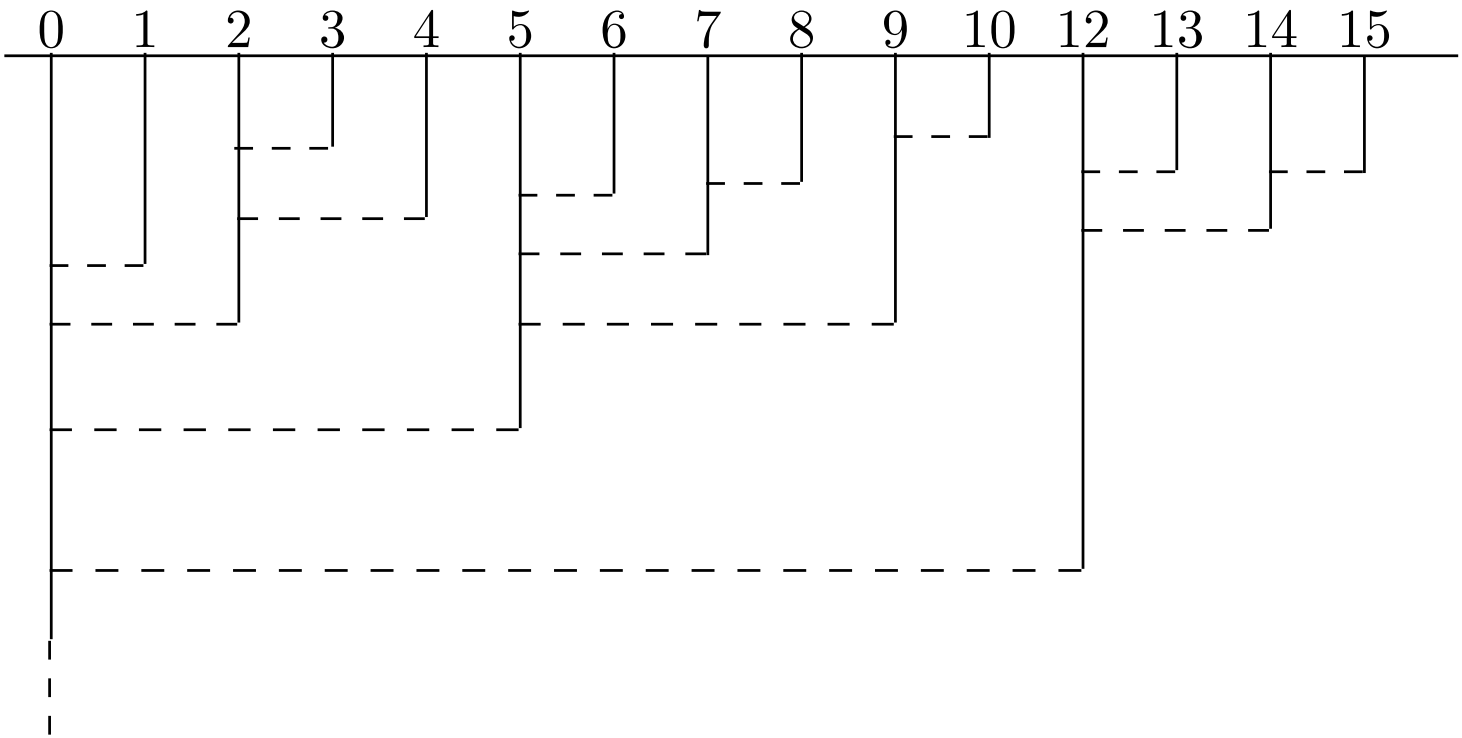
- Critical birth–death processes with birth/death rate b

$$W(x) = 1 + bx$$

Coalescent point process (Popovic 04, Aldous & Popovic 05)

A coalescent point process is the genealogy generated by a sequence of arbitrary i.i.d. positive r.v. $(H_i)_{i\geq 1}$ as below.

Here, we define $W(x)$ as $1/\mathbb{P}(H_1 > x)$.



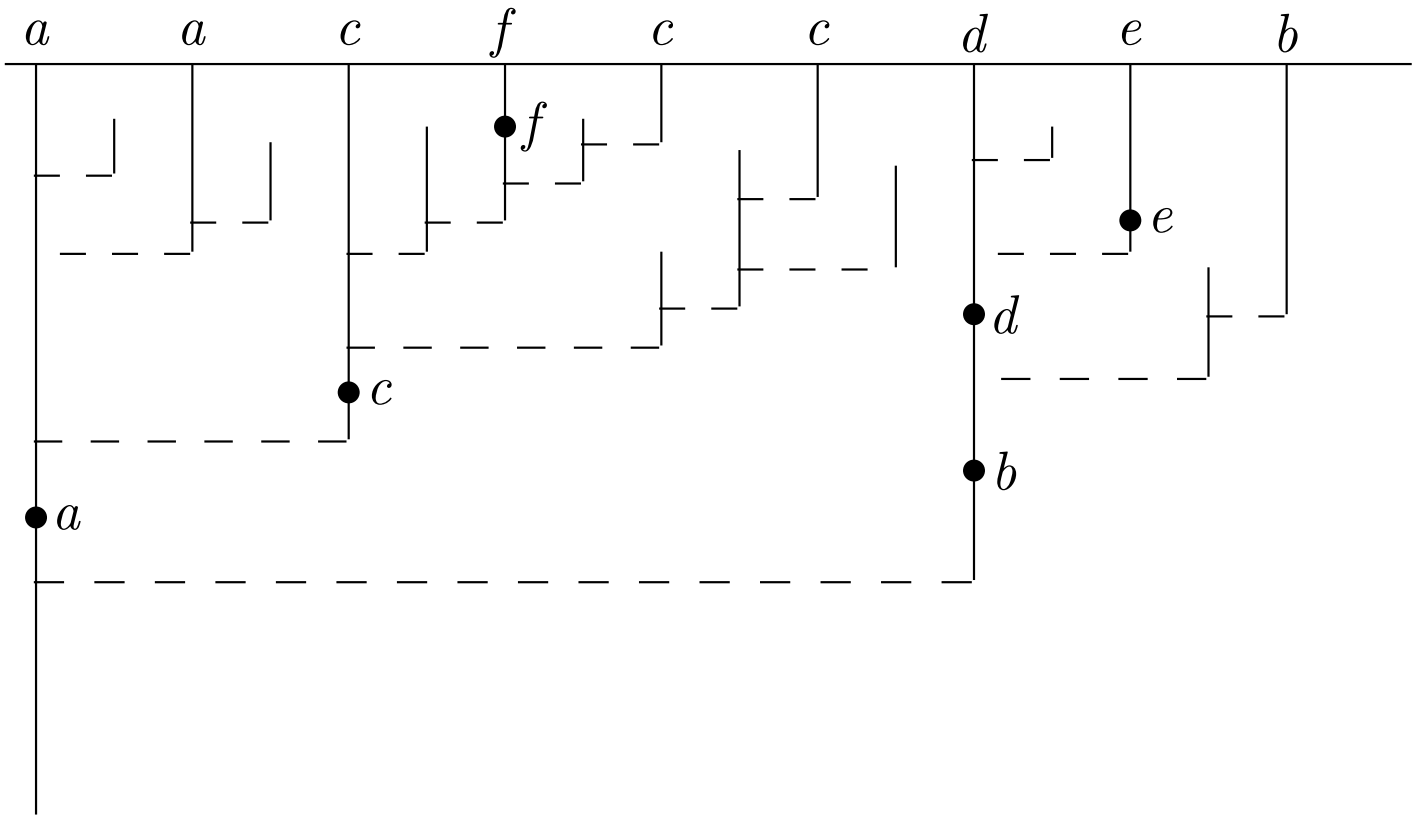
Assumptions on the mutation scheme

Now conditional on the genealogy, point mutations occur randomly.

1. mutations occur at constant rate θ during lifetimes, or, if one only considers the genealogy of individuals alive at time t, on branch lengths of the coalescent point process
2. mutations are neutral: they have no effect on the genealogy (birth rate, lifetimes...)
3. each mutation gives a new type, or allele, to its carrier (infinitely-many alleles model)
4. types are transmitted to the offspring born after this mutation and before the next one.

Mutation at rate θ

$N = 9$ alive individuals at time t, of 6 different types:
4 types of abundance 1, 1 type of abundance 2, and 1 type of abundance 3.



Clonal splitting trees

- the genalogy of *clonal individuals* is a splitting tree with (birth rate b and) lifetime duration distributed as

$$V_\theta := \min(V, E),$$

where E is an exponential variable with parameter θ independent of V.

- to a clonal splitting tree is associated a clonal coalescent point process with i.i.d. branch lengths $H_1^\theta, H_2^\theta, \ldots$ whose inverse of the tail distribution is denoted by W_θ

$$\mathbb{P}(H^\theta > s) =: \frac{1}{W_\theta(s)}.$$

Clonal splitting trees

Proposition (Lambert (2009))

For a coalescent point process with branch lengths $H_1, H_2, \ldots$, we can define H^θ as

$$\max(H_1, \ldots, H_{B^\theta}),$$

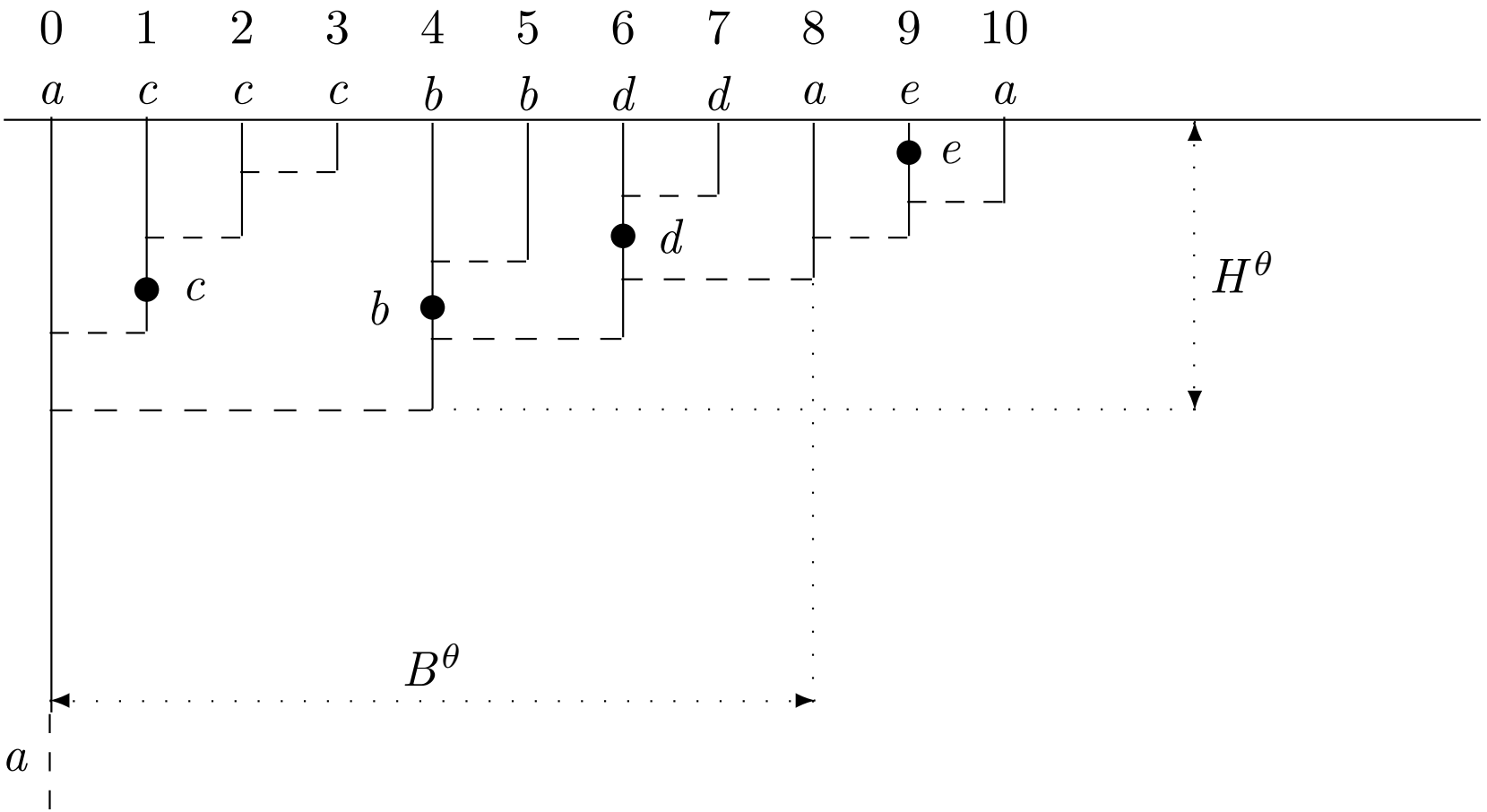
where B^θ is the index of first virgin lineage (i.e., carrying no mutation since it has split from ancestral lineage 0).
The scale function W_θ associated with clonal trees is related to W via

$$W'_\theta(x) = e^{-\theta x} W'(x) \qquad x \geq 0,$$

with $W_\theta(0) = 1$.

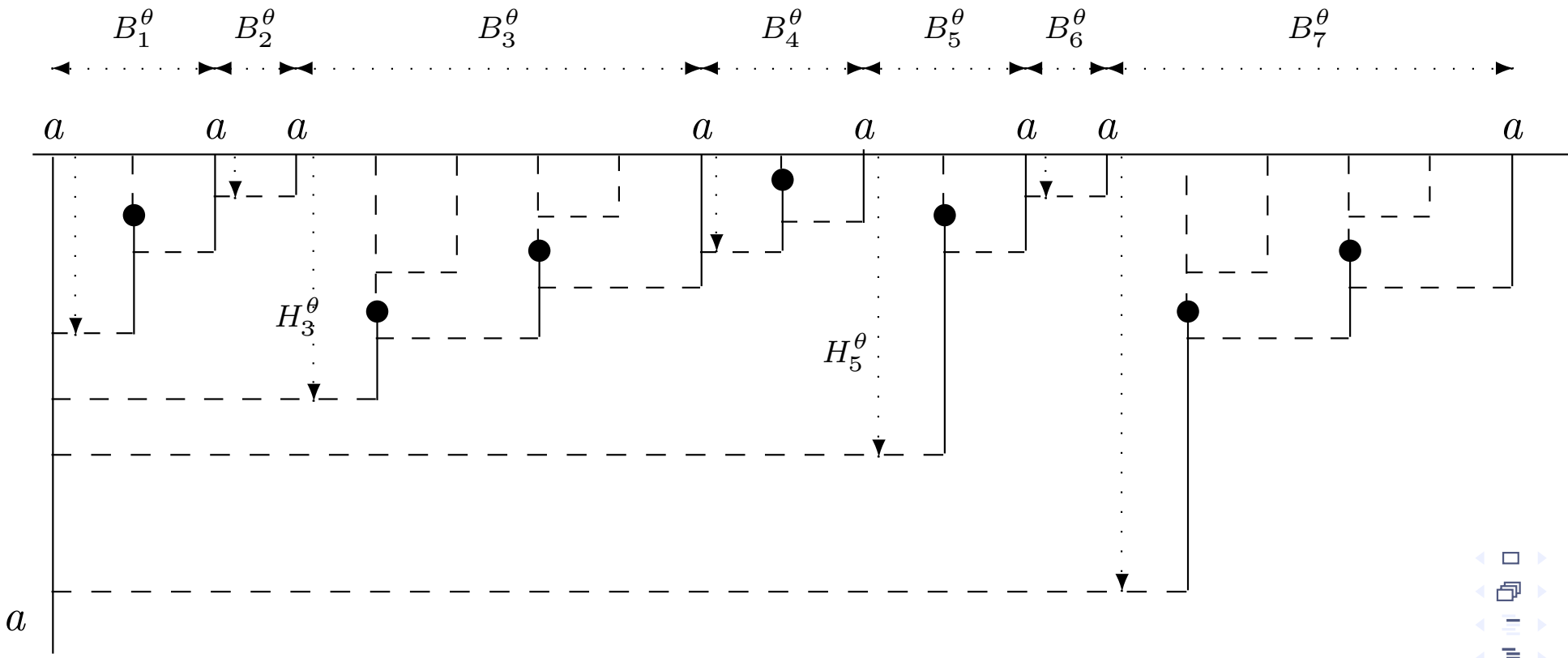
Virgin lineage

Below, the index of the first virgin lineage is 8



Finer result on clonal coalescent point process

B_i^θ = distances between consecutive virgin lineages
H_i^θ = max of branch lengths between consecutive virgin lineages
$\Longrightarrow (B_i^\theta, H_i^\theta)$ are i.i.d.



Finer result on clonal coalescent point process

We are interested in the joint law of H^θ and B^θ. Set

$$W_\theta(x,s) := \frac{1}{1-\mathbb{E}(s^{B^\theta}, H^\theta \leq x)} \qquad x \geq 0, s \in [0,1].$$

In particular, $W_\theta(x,1) = W_\theta(x)$.

Theorem (C. & Lambert 2012)

We have

$$\frac{\partial}{\partial x} W_\theta(x,s) = e^{-\theta x} \frac{\partial}{\partial x} W(x,s) \qquad x \geq 0,$$

with $W_\theta(0,\gamma) = 1$, *where*

$$W(x,s) := \frac{1}{1-s\mathbb{P}(H \leq x)}.$$

In particular, $W(x,1) = W(x)$.

Frequency spectrum

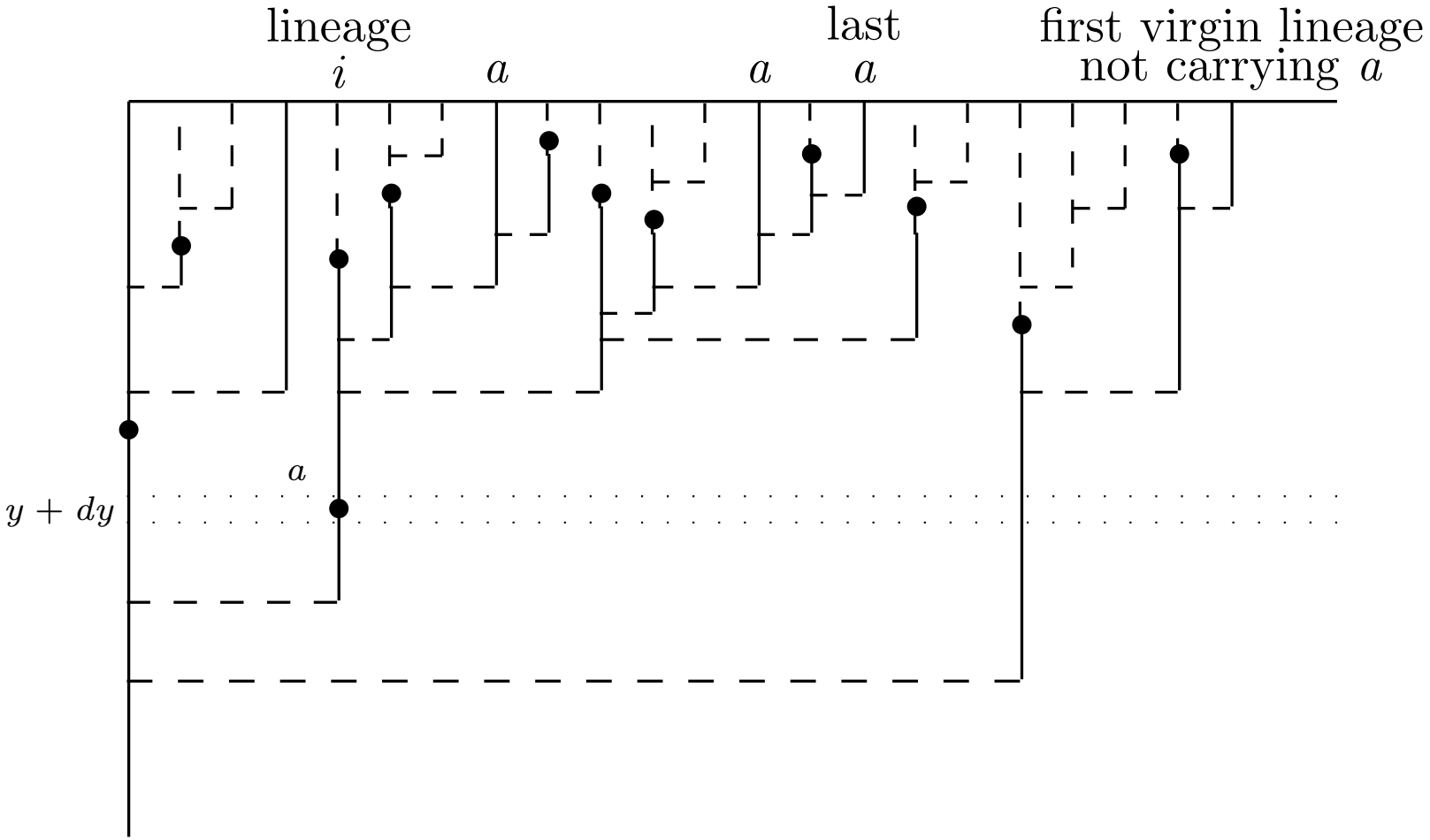
We introduce the notation:

- $A(t) :=$ number of distinct types in the population at time t
- $A(k,t) :=$ number of types represented by k individuals at time t
- then

$$\sum_{k\geq 1} A(k,t) = A(t) \quad \text{and} \quad \sum_{k\geq 1} kA(k,t) = N_t$$

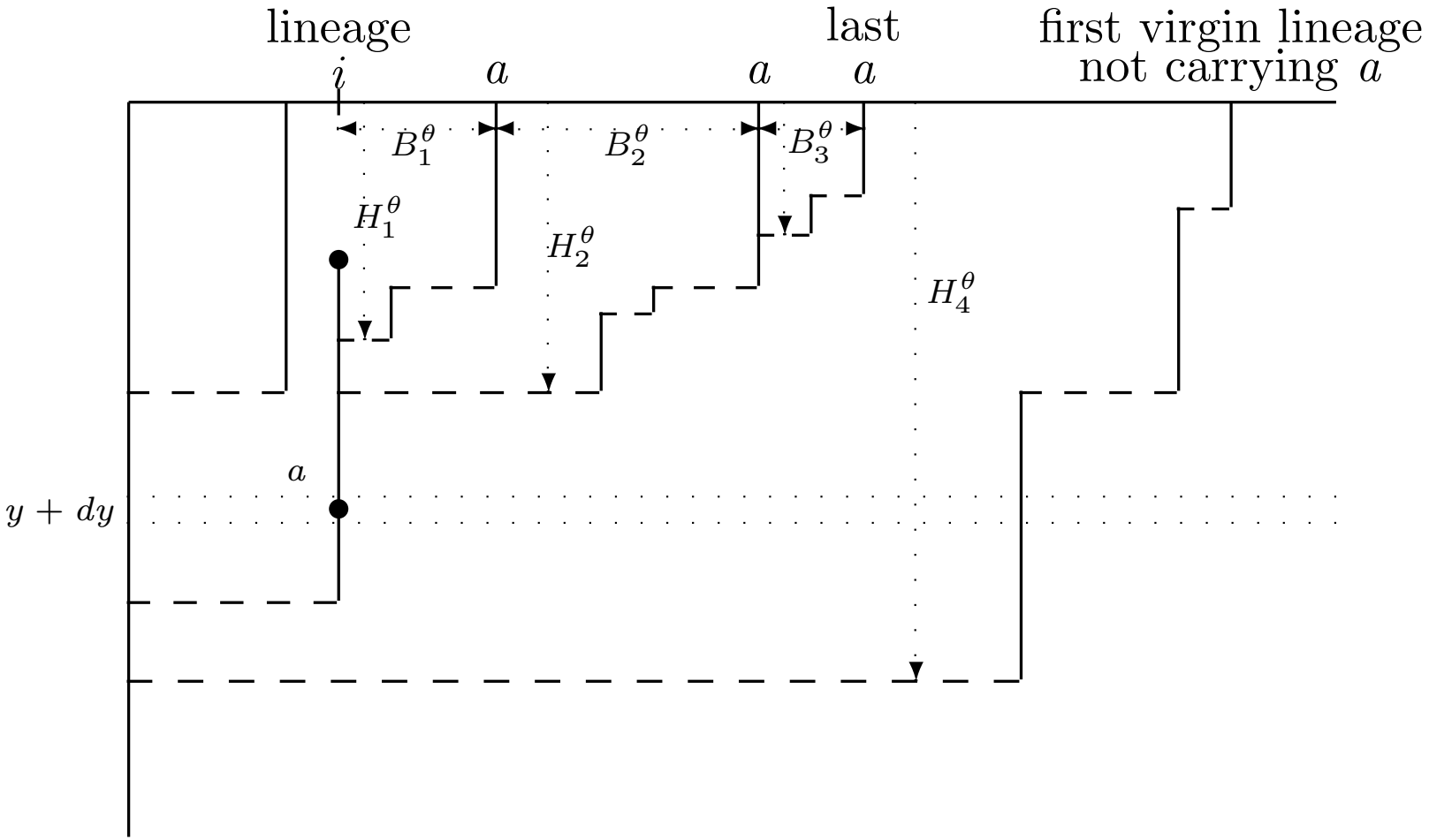
- $(A(k); k \geq 1)$ is called the frequency spectrum

Clonal coalescent point process



Goal. Compute the number of alleles of age in $(y, y + dy)$ and carried by k alive individuals at time t, jointly with N_t.

Clonal coalescent point process



Goal. Compute the number of alleles of age in $(y, y + dy)$ and carried by k alive individuals at time t, jointly with N_t.

Expected frequency spectrum

Recall N_t is the population size at time t.

Theorem (C. & Lambert 2012)

If $A(k,t,dy)$ denotes the number of alleles of age in $(y, y+dy)$ and carried by k alive individuals at time t, then

$$\mathbb{E}\left(s^{N_t-1}A(k,t,dy) \mid N_t \neq 0\right) = \theta\, dy\, \frac{W(t;s)^2}{W(t)}\, \frac{e^{-\theta y}}{W_\theta(y;s)^2}\left(1-\frac{1}{W_\theta(y;s)}\right)^{k-1}$$

Proof ($s = 1$)

Let $C_i(y, dy)$:= $\{i \leq N_t - 1 : H_i \geq y$ and the ith branch has a mutation of age in $(y, y + dy)\}$

$D_i(y)$:= {the ith branch type at time $t - y$ has one alive clone at time t}

$E_i(k, y)$:= {the ith branch type at time $t - y$ has k alive clones at time t}

Then $A_\theta(k, t, dy) = \sum_{i \geq 0} \mathbb{1}_{C_i(y,dy) \cap E_i(k,y)}$. Now

$\mathbb{P}^\star(C_i(y, dy) \cap E_i(k, y)) = \mathbb{P}^\star(C_i(y, dy))\mathbb{P}^\star(D_0(y))\mathbb{P}^\star(E_0(k, y) \mid D_0(y))$
and we claim that

$$\sum_{i \geq 0} \mathbb{P}^\star(C_i(y, dy)) = \theta \, dy \, \frac{W(t)}{W(y)} \tag{1}$$

$$\mathbb{P}^\star(D_0(y) = \frac{W(y)e^{-\theta y}}{W_\theta(y)} \tag{2}$$

$$\mathbb{P}^\star(E_0(k, y) \mid D_0(y)) = \frac{1}{W_\theta(y)} \left(1 - \frac{1}{W_\theta(y)}\right)^{k-1}. \tag{3}$$

Proof

Proof of (1):

$$\mathbb{P}^\star(C_i(y, dy)) = \mathbb{P}^\star(N_t - 1 \geq i)\theta\, dy\, (\mathbb{1}_{i=0} + \mathbb{1}_{i\geq 1}\mathbb{P}(H \geq y \mid H < t))\,.$$

The result follows by expressing $\mathbb{P}(H \geq y \mid H < t)$ in terms of W and summing over i.

Proof of (2): the next mutation on branch i after time $t - y$ occurs after an exponential time of parameter θ. Distinguishing whether this time is larger or smaller than y, we get

$$\mathbb{P}^\star(D_0(y)) = e^{-\theta y} + \int_0^y dx\, \theta e^{-\theta x} \left(1 - \frac{W_\theta(y-x)}{W_\theta(y)}\right).$$

The result then follows from an integration by parts.

The **proof of (3)** is trivial by definition of W_θ.

Applications

The main interest of our result is that we obtain exact formulas for the expected frequency spectrum.

For example, combining this with standard results on Crump–Mode–Jagers process (Jagers & Nerman (1981–1984), Taïb (1992)), we can obtain an exact expression for the

$$\text{a.s. limit of } \frac{A(k,t,a,b)}{N_t},$$

where $A(k,t,a,b)$ denotes the number of alleles of age in (a,b) carried by k alive individuals at time t.

Preliminary remark

We consider a supercritical splitting tree with Malthusian parameter α, so that N_t increases like $e^{\alpha t}$.

Since θ is an additional death rate for clonal families,

$$\text{clonal families are } \begin{cases} \text{supercritical} & \text{if} \quad \alpha > \theta \\ \text{critical} & \text{if} \quad \alpha = \theta \\ \text{subcritical} & \text{if} \quad \alpha < \theta. \end{cases}$$

Notation

We define

- $M_t(x; a, b)$= number of families of size $\geq x$ and of age in $[a, b]$

$$M_t(x; a, b) := \sum_{k \geq x} \int_a^b A(k, t, dy)$$

- $L_t(x)$= number of families of size $\geq x$

$$L_t(x) := M_t(x; 0, \infty)$$

- $O_t(a)$= number of families of age $\geq a$

$$O_t(a) := M_t(0; a, \infty).$$

Goal. Find x_t such that $\mathbb{E}\ L_t(x_t) = O(1)$ and a_t such that $\mathbb{E}\ O_t(a_t) = O(1)$, as $t \to \infty$.

Case $\alpha > \theta$

Assume $\alpha > \theta$

Proposition (C. & Lambert 2013)

For any $c > 0$ and $a < b$,

$$\mathbb{E}M_t\left(ce^{(\alpha-\theta)t}; t-b, t-a\right) = O(1),$$

so that largest families have sizes $cN^{1-\theta/\alpha}$ and are also the oldest ones (born at times $O(1)$).

Case $\alpha < \theta$: largest families

Assume $\alpha < \theta$ and set $\beta := \theta/(\theta - \alpha)$

Proposition (C. & Lambert 2013)

For some other explicit constant b, set

$$x_t := b\,(\alpha t - \beta \log(t))$$

Then for any c

$$\mathbb{E}L_t(x_t + c) \sim \mathbb{E}M_t\left(x_t + c; (1-\epsilon)\frac{\log(t)}{\theta - \alpha}, (1+\epsilon)\frac{\log(t)}{\theta - \alpha}\right) = O(1),$$

so that largest families have sizes $b(\log(N) - \beta \log(\log N)) + c$ and they all have age $\sim \dfrac{\log(t)}{\theta - \alpha}$.

Case $\alpha < \theta$: oldest families

Assume $\alpha < \theta$ and set $\gamma := \alpha/\theta < 1$

Proposition (C. & Lambert 2013)

For any a,

$$\mathbb{E}O_t(\gamma t + b) = O(1).$$

so that oldest families have ages $\gamma t + a$.

Case $\alpha = \theta$: largest families

Assume $\alpha = \theta$ and set $\beta := 1/(2\alpha)$

Proposition (C. & Lambert 2013)

For some explicit constant b, set

$$x_t := b\,(t - \beta \log(t))^2$$

Then for any c

$$\mathbb{E}L_t(x_t + ct) \sim \mathbb{E}M_t\left(x_t + ct; (1-\epsilon)\frac{t}{2}, (1+\epsilon)\frac{t}{2}\right) = O(1),$$

so that largest families have sizes $b(\log(N) - \beta \log(\log N) + c)^2$ and they all have age $\sim t/2$.

Case $\alpha = \theta$: oldest families

Assume $\alpha = \theta$ and set $\gamma := 1/\alpha$

Proposition (C. & Lambert 2013)

For any a,

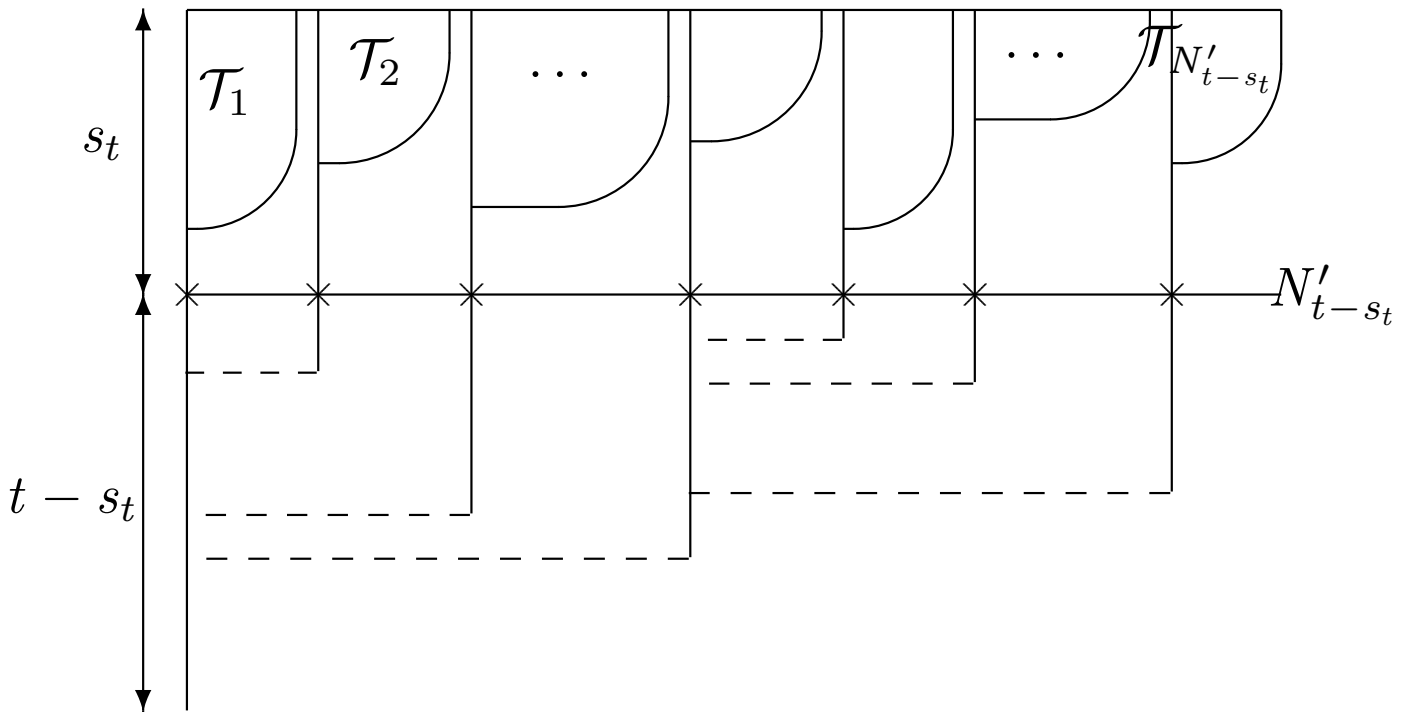
$$\mathbb{E}\, O_t(t - \gamma \log(t) + a) = O(1).$$

so that oldest families have ages $t - \gamma \log(t) + a$.

Convergence in distribution: idea of the method

Take the coalescent point process at time t, fix s_t such that $s_t \to \infty$, and define

N'_{t-s_t} := number of indiv. alive at time $t - s_t$ having alive desc. at time t
= number of subtrees $(\mathcal{T}_i)$ grafted on branch lengths $\geq s_t$



Convergence in distribution: idea of the method

Set

$$X_t^{(k)} := \text{ size of the } k\text{-th largest family in the whole population}$$

$$Y_i := \text{ size of the largest family in subtree } \mathcal{T}_i.$$

When $\alpha \leq \theta$, we choose

$$s_t := \begin{cases} \log(t)^{\frac{1-\varepsilon}{\theta-\alpha}} & \text{if} \quad \alpha < \theta \\ t^{\frac{1-\varepsilon}{2}} & \text{if} \quad \alpha = \theta. \end{cases}$$

This choice entails, conditionally on $N_t \neq 0$,

- $N'_{t-s_t} \to \infty$
- $(X_t^{(1)}, \ldots, X_t^{(k)})$= first k order statistics of $\{ Y_1, \ldots, Y_{N'_{t-s_t}} \}$ with high probability
- $\mathbb{P}(Y \geq x_t + c) = \mathbb{P}(L_{s_t}(x_t + c) \geq 1) \sim \mathbb{E}(L_{s_t}(x_t + c))$

Convergence in distribution: idea of the method

The same results hold with

$$A_t^{(k)} := \text{ age of the } k\text{-th oldest family in the whole population}$$

$$Y_i := \text{ age of the oldest family in subtree } \mathcal{T}_i,$$

and

$$s_t := \begin{cases} \alpha t \ / \ \theta & \text{if} \quad \alpha < \theta \\ t - \log(t) \ / \ \alpha & \text{if} \quad \alpha = \theta. \end{cases}$$

Convergence in distribution: case $\alpha = \theta$

Assume $\alpha = \theta$.

Theorem (C. & Lambert 2013)

There are some explicit constants b, c, u, such that

$$\lim_{t\to\infty} \mathbb{P}(X_t^{(1)} < b(\alpha t^2 - t\log t) + xt \mid N_t \neq 0) = \frac{1}{1 + u.e^{-cx}}.$$

More specifically, $(\frac{X_t^{(k)}}{t} - b(\alpha t - \log t); k \geq 1)$ converge (fdd) to the (ranked) atoms of a mixed Poisson point measure with intensity

$$\mathcal{E}\ e^{-cx}\,dx,$$

where $\mathcal{E}$ is some exponential r.v.

Convergence in distribution: case $\alpha = \theta$

Assume again $\alpha = \theta$.

Theorem (C. & Lambert 2013)

There is some explicit constant $v > 0$ such that

$$\lim_{t\to\infty} \mathbb{P}(A_t^{(1)} < t - \frac{\log t}{\alpha} + a \mid N_t \neq 0) = \frac{1}{1 + v.e^{-\alpha a}}.$$

More specifically, $(A_t^{(k)} - t + \log(t)/\alpha; k \geq 1)$ converge (fdd) to the (ranked) atoms of a mixed Poisson point measure with intensity

$$\mathcal{E}\, e^{-\alpha a}\, da,$$

where $\mathcal{E}$ is some exponential r.v.

Convergence in distribution: case $\alpha < \theta$

Assume $\alpha < \theta$.

Theorem (C. & Lambert 2013)

There are some explicit constants u, c, such that

$$\lim_{t\to\infty} \mathbb{P}(X_t^{(1)} < b(\alpha t - \beta \log(t)) + k \mid N_t \neq 0) = \frac{1}{1 + u.c^k}.$$

More specifically, along some subsequence,
$(X_t^{(k)} - b(\alpha t - \beta \log(t)); k \geq 1)$ *converge (fdd) to the (ranked) atoms of a mixed Poisson point measure with intensity*

$$\mathcal{E} \sum_{j\in\mathbb{Z}} c^j \delta_j,$$

where $\mathcal{E}$ is some exponential r.v.

Convergence in distribution: case $\alpha < \theta$

Assume again $\alpha < \theta$.

Theorem (C. & Lambert 2013)

There is some explicit constant $v > 0$ such that

$$\lim_{t\to\infty} \mathbb{P}(A_t^{(1)} < (\alpha t \ / \ \theta) + a \mid N_t \neq 0) = \frac{1}{1 + v.e^{-\theta a}}.$$

More specifically, $(A_t^{(k)} - (\alpha t \ / \ \theta); k \geq 1)$ converge (fdd) to the (ranked) atoms of a mixed Poisson point measure with intensity

$$\mathcal{E}\, e^{-\theta a}\, da,$$

where $\mathcal{E}$ is some exponential r.v.

Questions and future works

- We have obtained precise results on the size (resp. age) of the largest (resp. oldest) families in the case of (sub)critical clonal families.
- Open questions:
  - Convergence in distribution in the supercritical case?
  - Why an age $t/2$ for the oldest families in the critical case?
- Other question:
  - The case of mutations at birth: Richard (2012), C., Lambert, Richard (2012).
  - To make the link with Sabeti's recombination tree, we should study the the point measure of the sizes of the largest families as a process of the mutation rate θ (= distance to the gene on the DNA sequence).
  - Other questions that can be tackled with coalescent point processes: time to the most recent common ancestor at time t as a process indexed by t... (see also the talk of Amaury on wednesday)