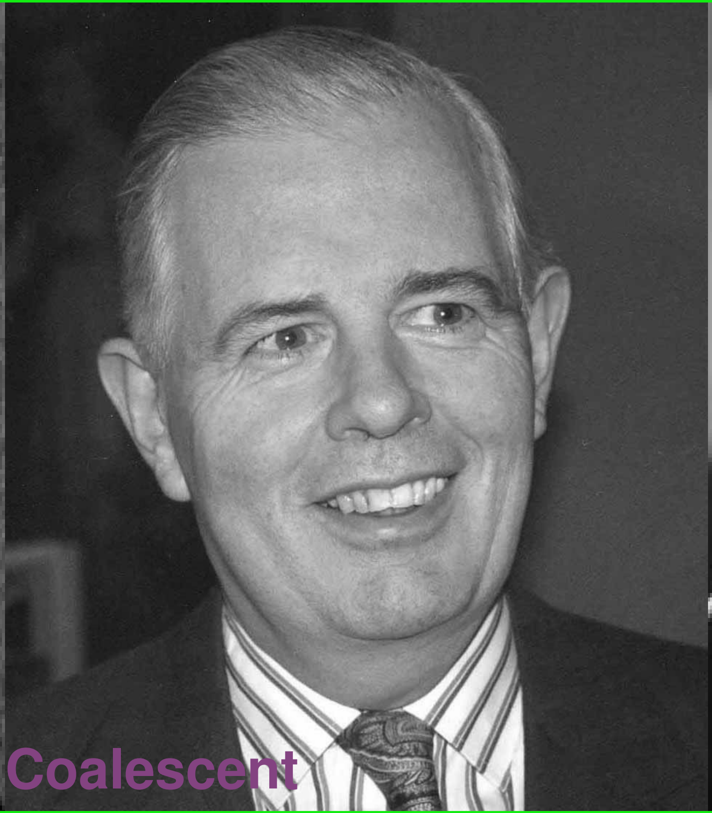
Fractional Coalescent:

A new coalescent theory based on a non-Markovian Poisson process

Exponentials



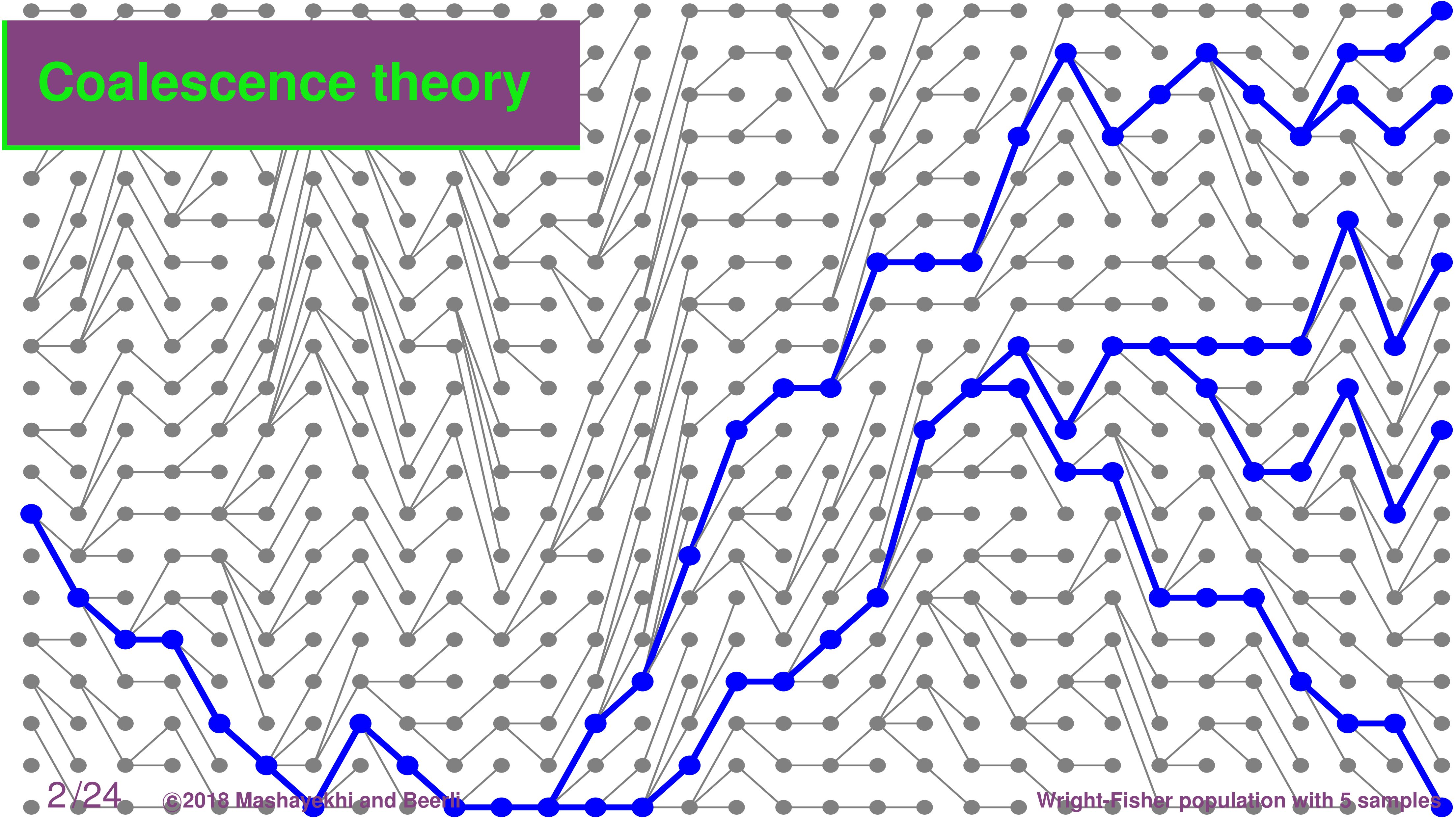
Coalescent

Mittag-Leffler function

Somayeh Mashayekhi and Peter Beerli
Scientific Computing, Florida State University

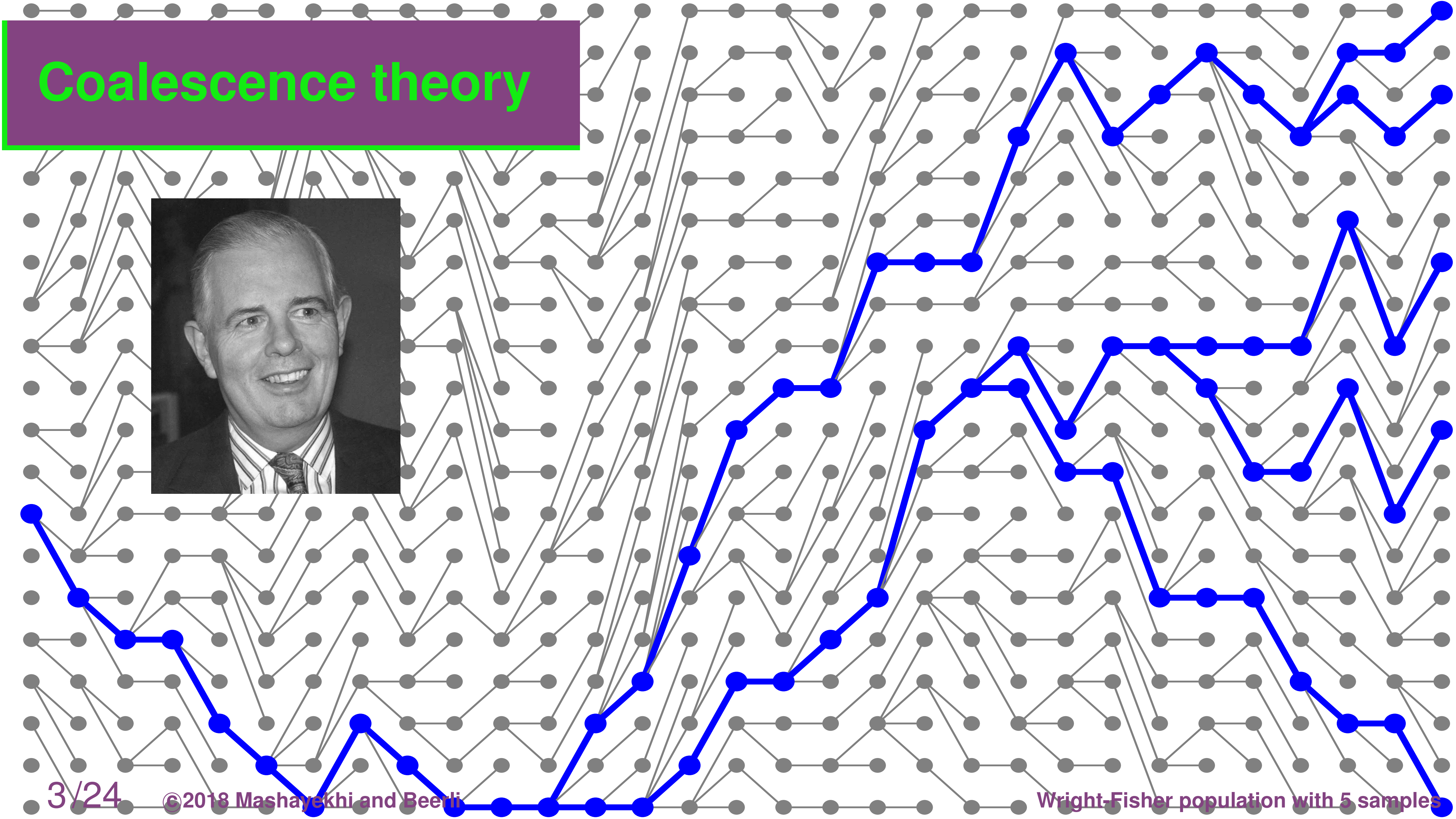
@somayehmashayek
@peterbeerli

Coalescence theory



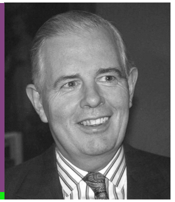
Wright-Fisher population with 5 samples

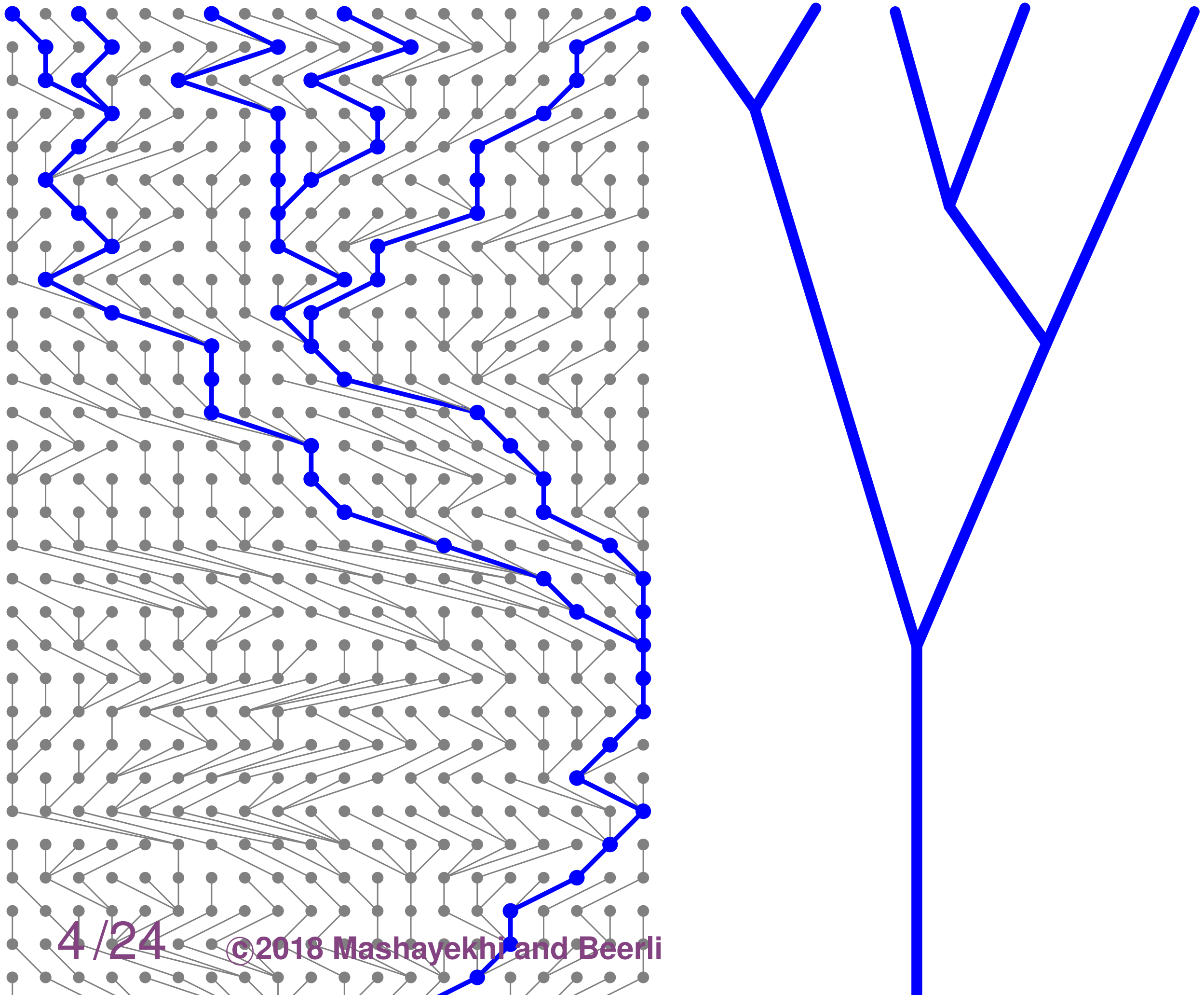
Coalescence theory



Wright-Fisher population with 5 samples

Coalescence theory





Using τ scaled by $1/N$ generations, two lineages remain distinct with probability

$$(1-\frac{1}{N})^{N\tau} \rightarrow e^{-\tau},$$

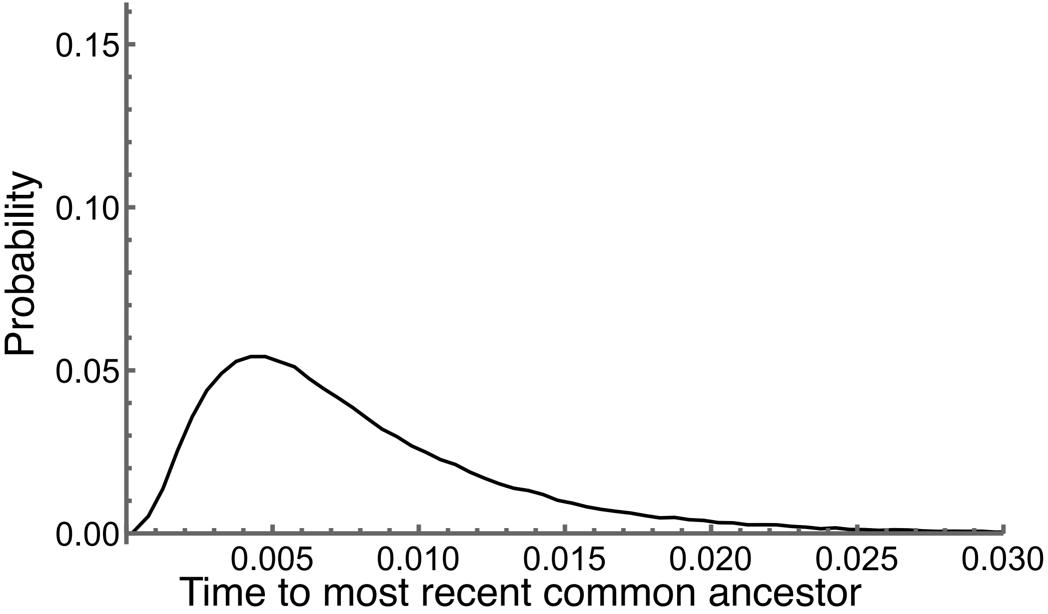
as N goes to ∞. Kingman's n-coalescent generalized the two-lineages framework to k lineages by changing $\tau \rightarrow \tau\binom{k}{2}$:

$$(1-\frac{\binom{k}{2}}{N})^{N\tau} \rightarrow e^{-\binom{k}{2}\tau},$$

as N goes to ∞. With $\tau = t/N$ the familiar coalescence formula $e^{-t\binom{k}{2}/N}$ emerges.

Wright-Fisher population with 5 samples

Coalescent is a convolution of exponential distributions



Sample size $n = 5$

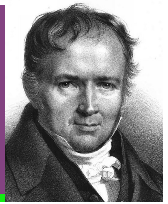
$$p(t|N,k) = e^{-t\binom{k}{2}/N}\frac{\binom{k}{2}}{N}$$

$$p(G|N,n) = \prod_{k=2}^{k=n} p(t|N,k) = \prod_{k=2}^{k=n} e^{-t\binom{k}{2}/N}\frac{\binom{k}{2}}{N}\frac{1}{\binom{k}{2}}$$

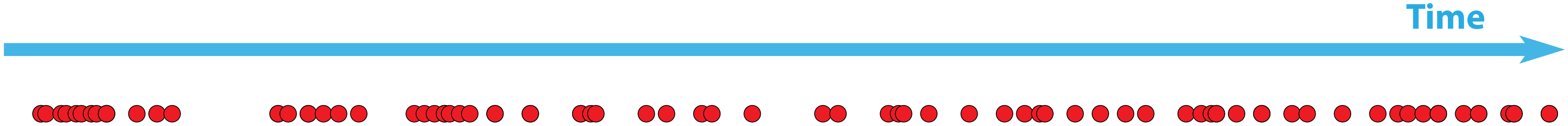
Scaling the time and the population size using mutation we get the familiar

$$p(G|\Theta,n) = \prod_{k=2}^{k=n}\left(e^{-t\frac{k(k-1)}{\Theta}}\frac{2}{\Theta}\right) \quad \text{with} \quad \Theta = 2N\mu$$

Poisson process and waiting time

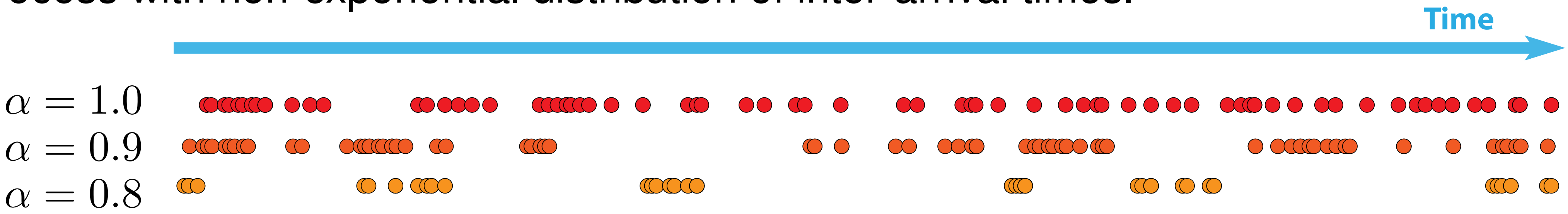


The coalescent rate changes with every coalescent, but with a fixed rate we recognize the randomness in the waiting time more easily! When coalescence events are rare then the Poisson process is a good model for the arrival of the events, the process has no memory and the waiting times between coalescent events are drawn from an exponential distribution.



Fractional Poisson process and waiting time

The fractional Poisson distribution extends the standard Poisson distribution with an additional parameter α adding variability of the waiting times. The time interval between each pair of consecutive counts follows the non-exponential power-law distribution with parameter $0<\alpha\leq1$. Thus, the fractional Poisson process is a non-Markovian counting process with non-exponential distribution of inter-arrival times.



Poisson process and number of offsprings

When the population size is large (and constant), the number of offspring of each individual is approximated by a Poisson distribution.
The probability of having n offspring per x generation

$$P(n, x) = \frac{(\overline{n}x)^n}{n!} e^{-\overline{n}x},$$

the parameter $\overline{n}$ is the average number of offspring per generation.

Fractional Poisson process and number of offspring

The fractional Poisson distribution extends the standard Poisson distribution with an additional parameter α adding variability to the number of offspring per individual, we can calculate the probability of n offsprings in x generations as

$$P(n,x) = \frac{(\bar{\varsigma} x^{\alpha})^n}{n!} \sum_{k=0}^{\infty} \frac{(k+n)!}{k!} \frac{(-\varsigma x^{\alpha})^k}{\Gamma(\alpha(k+n)+1)}$$

Laskin 2003

Fractional Poisson distribution and number of offspring

The fractional Poisson distribution extends the standard Poisson distribution with an additional parameter α adding variability to the number of offspring per individual, we can calculate the probability of n offsprings in x generations as

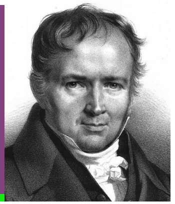
$$P(n,x) = \frac{(\overline{\varsigma} x^{\alpha})^n}{n!} \sum_{k=0}^{\infty} \frac{(k+n)!}{k!} \frac{(-\varsigma x^{\alpha})^k}{\Gamma(\alpha(k+n)+1)}$$

This includes the standard Poisson distribution, with $\overline{\varsigma} = \overline{n}$, for example $\overline{n} = 1$ for the Wright-Fisher model, and $\alpha = 1$ the fractional Poisson becomes

$$P(n,x) = \frac{x^n}{n!} \sum_{k=0}^{\infty} \frac{(k+n)!}{k!} \frac{(-x)^k}{\Gamma(k+n+1)} = \frac{x^n}{n!} \sum_{k=0}^{\infty} \frac{(k+n)!}{k!} \frac{(-x)^k}{(k+n)!} = \frac{x^n}{n!} e^{-x}$$

Laskin 2003

Using the Fractional Poisson distribution



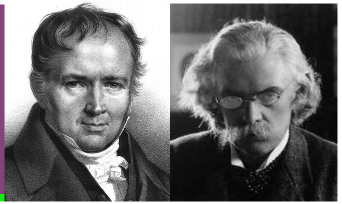
If we assume that the expected number of offspring depends on α, then the probability that two randomly selected individuals have not the same parent in the previous generation is

$$1 - \frac{1}{\Gamma(\alpha+1)N},$$

Using time τ^{α} scaled by $1/N$ generations, two lineages remain distinct with probability

$$\left(1 - \frac{1}{\Gamma(1+\alpha)N}\right)^{N\tau^{\alpha}} \longrightarrow e^{-\frac{\tau^{\alpha}}{\Gamma(\alpha+1)}} \xrightarrow{\alpha=1} e^{-\tau}$$

Using the Fractional Poisson distribution



If we assume that the expected number of offspring depends on α, then the probability that two randomly selected individuals have not the same parent in the previous generation is

$$1-\frac{1}{\Gamma(\alpha+1)N},$$

More interesting are results with $\alpha<1$: Using time τ^{α} scaled by $1/N$ generations, two lineages remain distinct with probability

$$\left(1-\frac{1}{\Gamma(1+\alpha)N}\right)^{N\tau^{\alpha}} \longrightarrow e^{-\frac{\tau^{\alpha}}{\Gamma(\alpha+1)}} \to \mathcal{E}_{\alpha}(-\tau^{\alpha})$$

Mittag-Leffler function and Exponential function

The Mittag-Leffler function

$$\mathcal{E}_\alpha(x) = \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(\alpha n + 1)}, \quad \alpha > 0, x \in \mathcal{C}.$$

generalizes the exponential function, with $\alpha{=}1$ the Mittag-Leffler becomes the Exponential function

$$= \sum_{n=0}^{\infty} \frac{x^n}{\Gamma(n+1)} = \sum_{n=0}^{\infty} \frac{x^n}{n!} = e^x.$$

Mittag-Leffler function and Exponential Distribution

The Mittag-Leffler function can be expressed as a mixture of exponentials

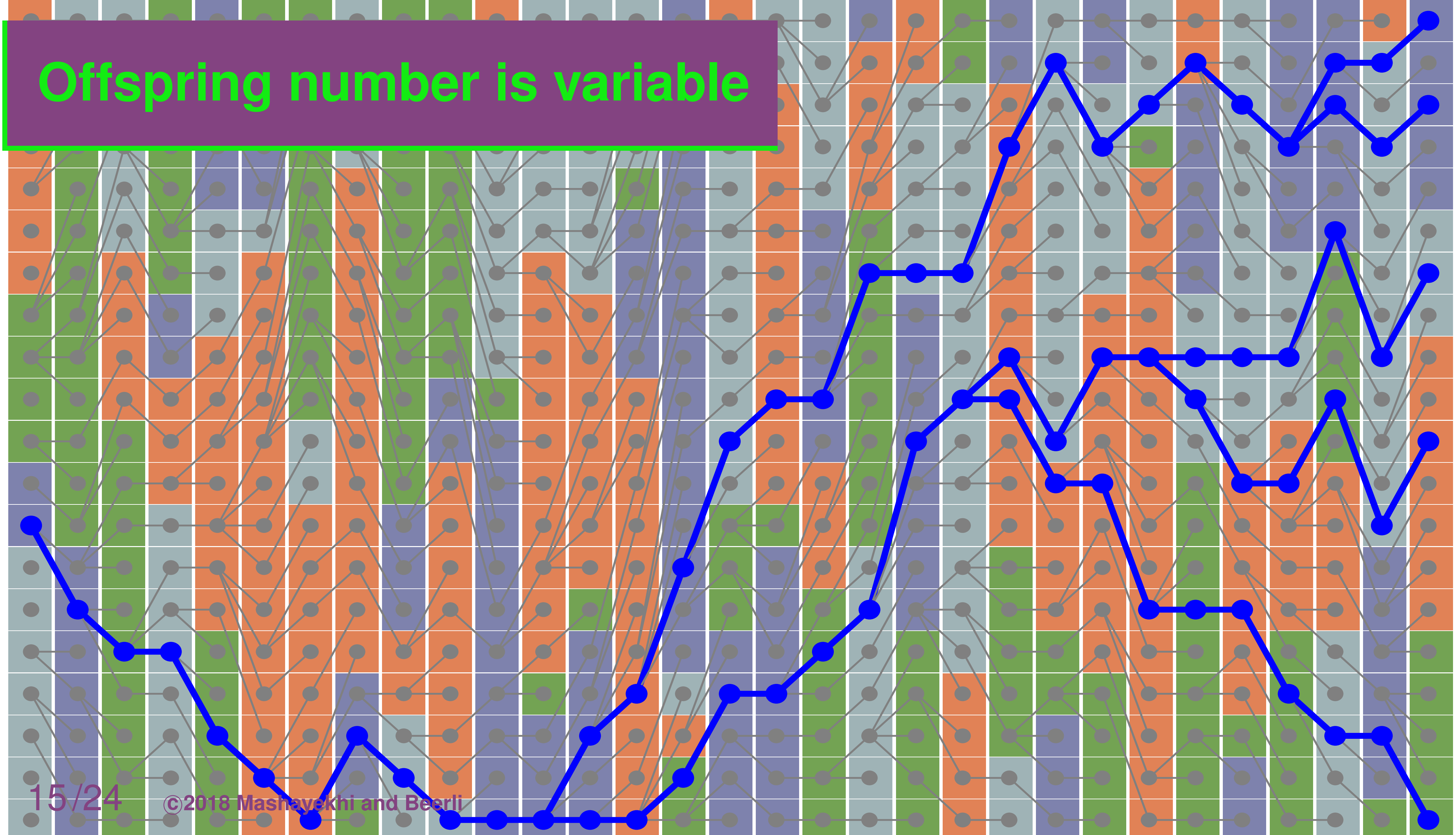
$$\mathcal{E}_\alpha(-\lambda x^\alpha) = \int_0^\infty \omega(\kappa, \alpha) e^{-\kappa x} d\kappa,$$

where $\omega(\kappa, \alpha)$ is a probability density. The discrete form can be written as

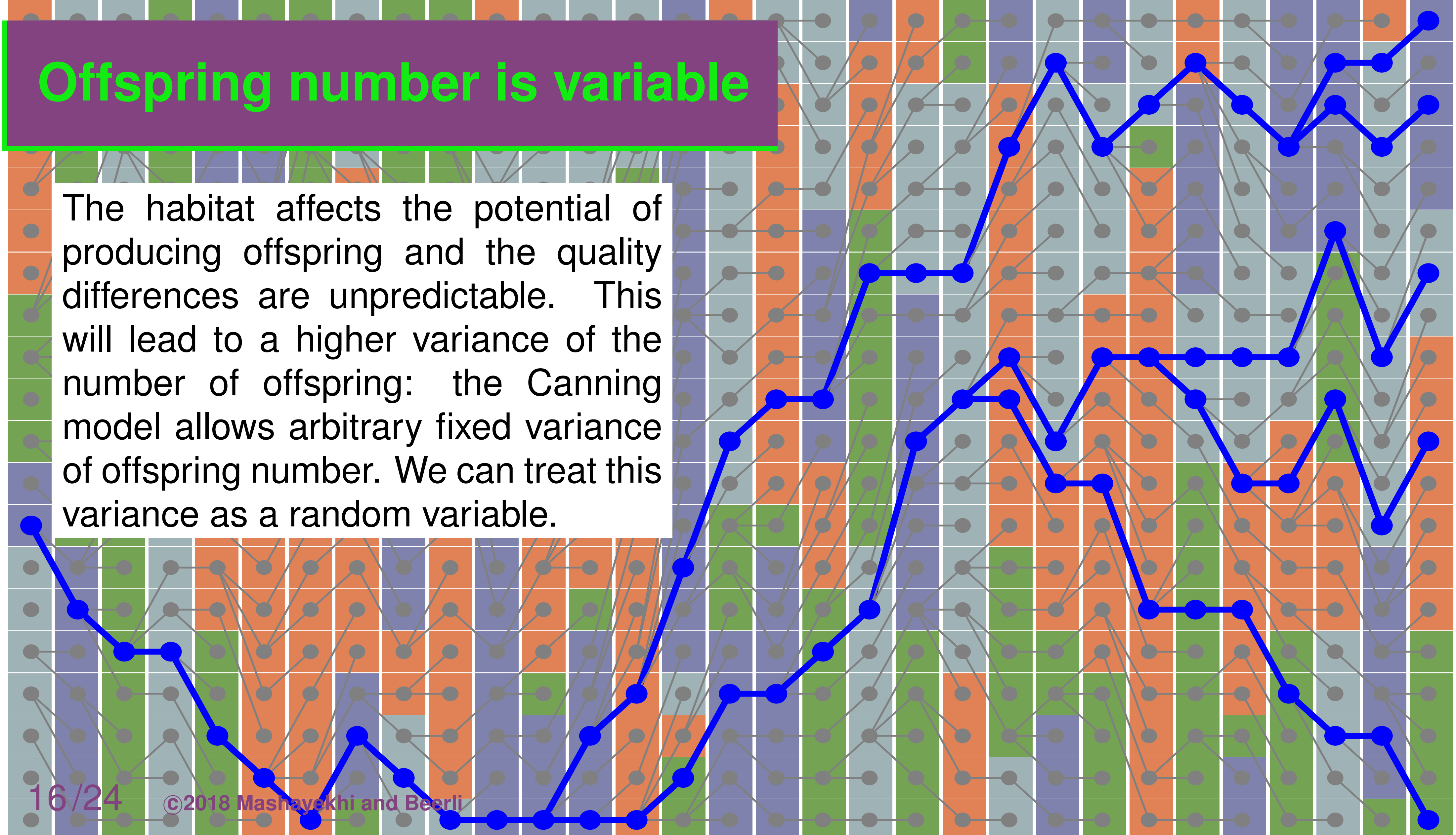
$$\mathcal{E}_\alpha(-\lambda x^\alpha) = \sum_k \omega(\kappa_k, \alpha)(1 - \kappa_k \frac{1}{n})^{xn}.$$

McNamara 2017

Offspring number is variable

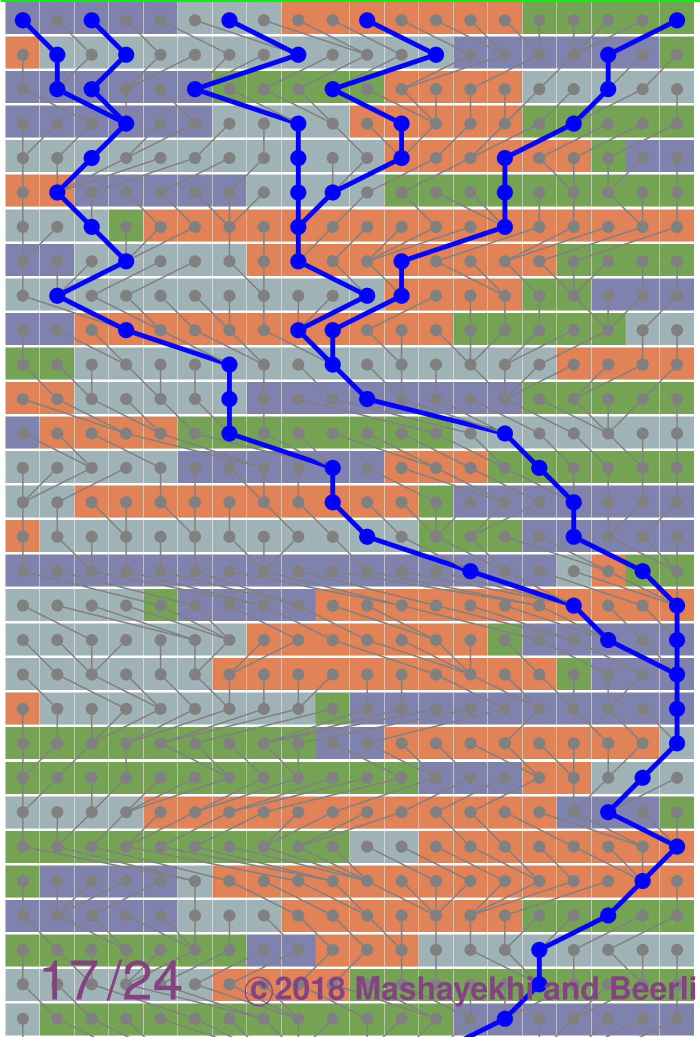


Offspring number is variable



The habitat affects the potential of producing offspring and the quality differences are unpredictable. This will lead to a higher variance of the number of offspring: the Canning model allows arbitrary fixed variance of offspring number. We can treat this variance as a random variable.

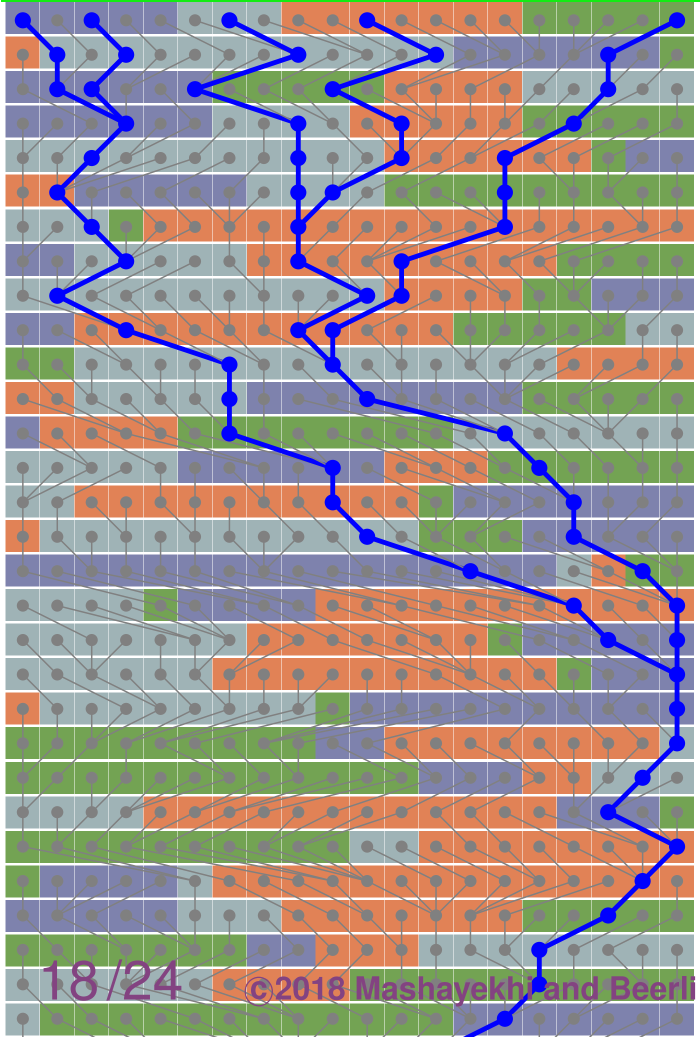
Offspring number is variable



The Canning model allows an arbitrary fixed variance of offspring number σ. We can extend the Canning model and treat this variance as a random variable.

$$P\{\text{not coal } |\sigma^2 = \sigma_j^2\} = \left(1 - \frac{\sigma_j^2}{N}\right)^{N\tau} .$$

Offspring number is variable



The Canning model allows an arbitrary fixed variance of offspring number σ. We can extend the Canning model and treat this variance as a random variable.

$$P\{\text{not coal } |\sigma^2 = \sigma_j^2\} = \left(1 - \frac{\sigma_j^2}{N}\right)^{N\tau}.$$

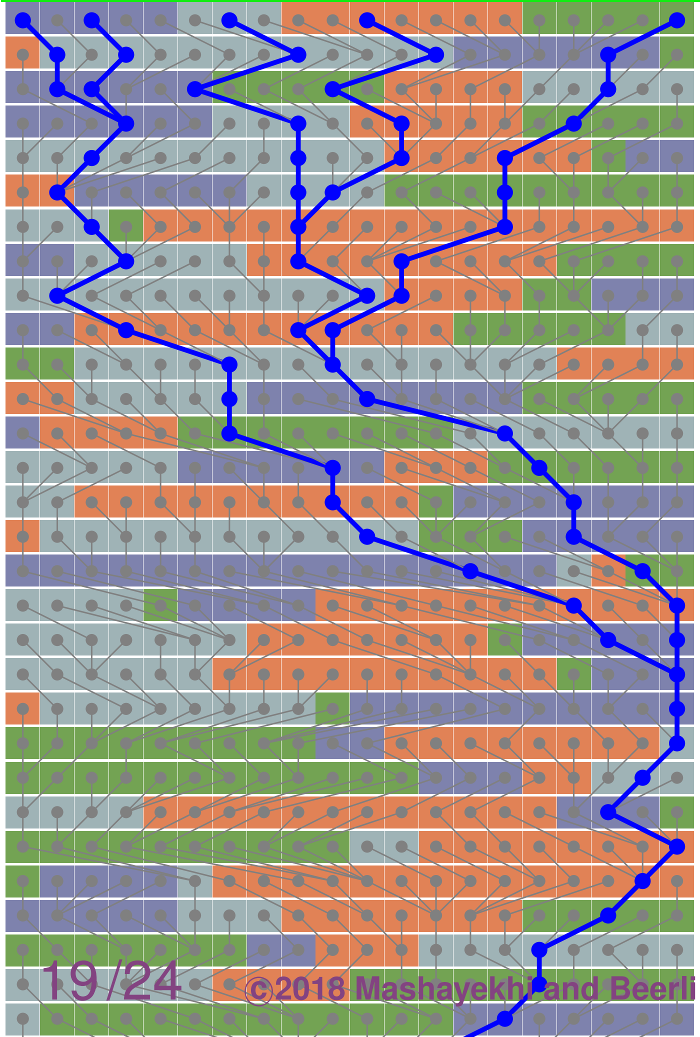
The Mittag-Leffler function with parameter α can be used to approximate this system:

$$\mathcal{E}_\alpha(-\tau^\alpha) = \sum_j \omega(\sigma_j^2, \alpha)\left(1 - \sigma_j^2 \frac{1}{N}\right)^{N\tau}$$

For more than two lineages the waiting time to the next coalescent is

$$\mathcal{E}_\alpha(-\binom{k}{2}\tau^\alpha)$$

Offspring number is variable



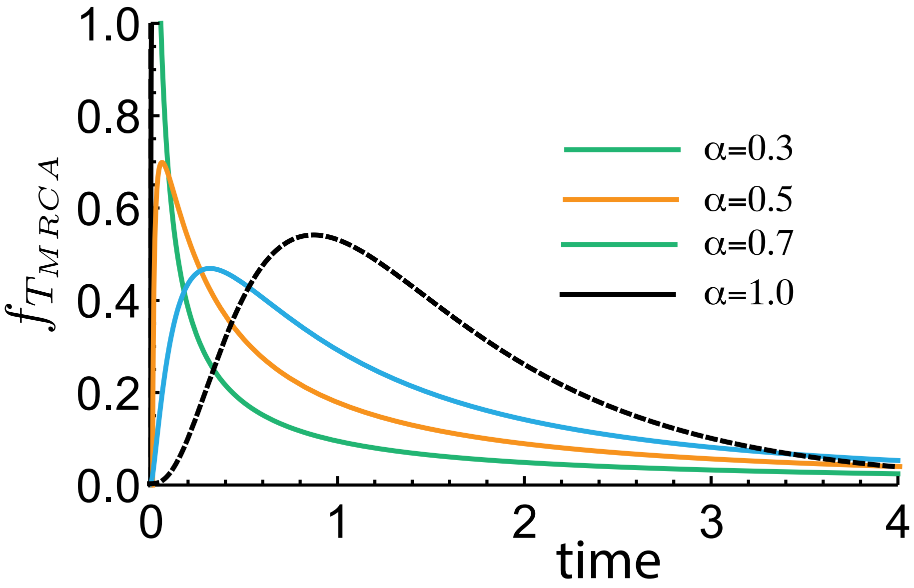
When we replace the scaled time τ with $\frac{u_k}{(N\mu)^{\frac{1}{\alpha}}}$, where u_k has been scaled by mutation ($\mu^{\frac{1}{\alpha}}$), we can rewrite

$$\mathcal{E}_\alpha(-\binom{k}{2}\frac{u_k^\alpha}{N\mu}) \rightarrow \mathcal{E}_\alpha(-\frac{k(k-1)}{\Theta}u_k^\alpha).$$

finally, using $\lambda_k = \frac{k(k-1)}{\Theta}$ with $\Theta = 2N\mu$, we can get the fractional equivalent of Kingman's coalescent"

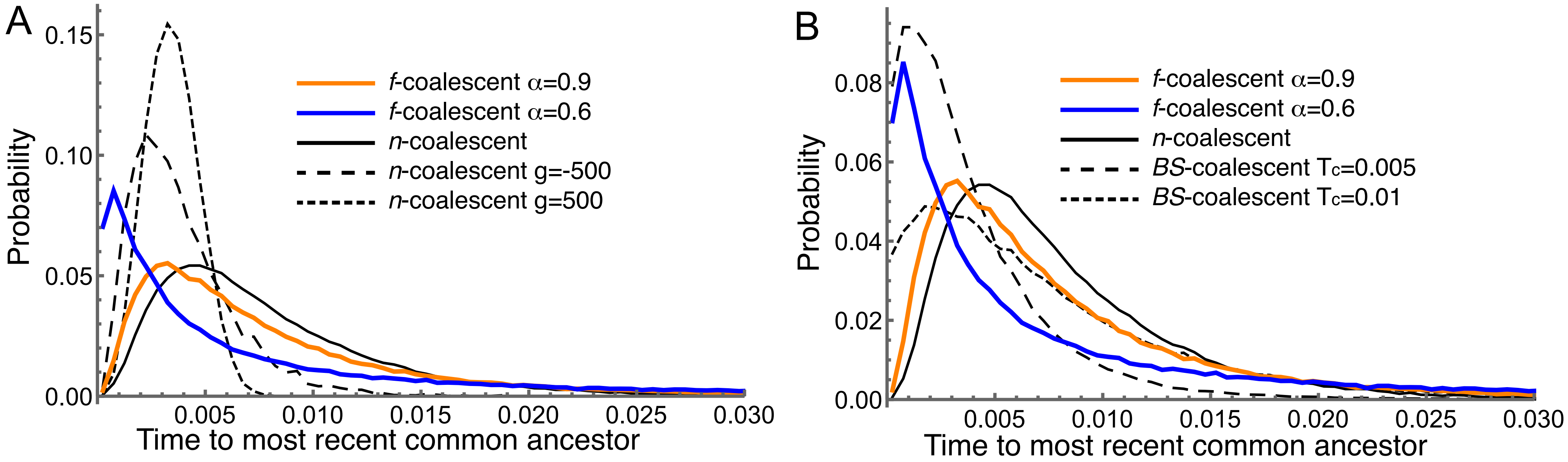
$$p(G|\Theta, n) = \prod_{k=2}^{n} u_k^{\alpha-1}\frac{2}{\Theta}\mathcal{E}_\alpha(-\lambda_k u_k^\alpha)$$

Time to the most recent common ancestor



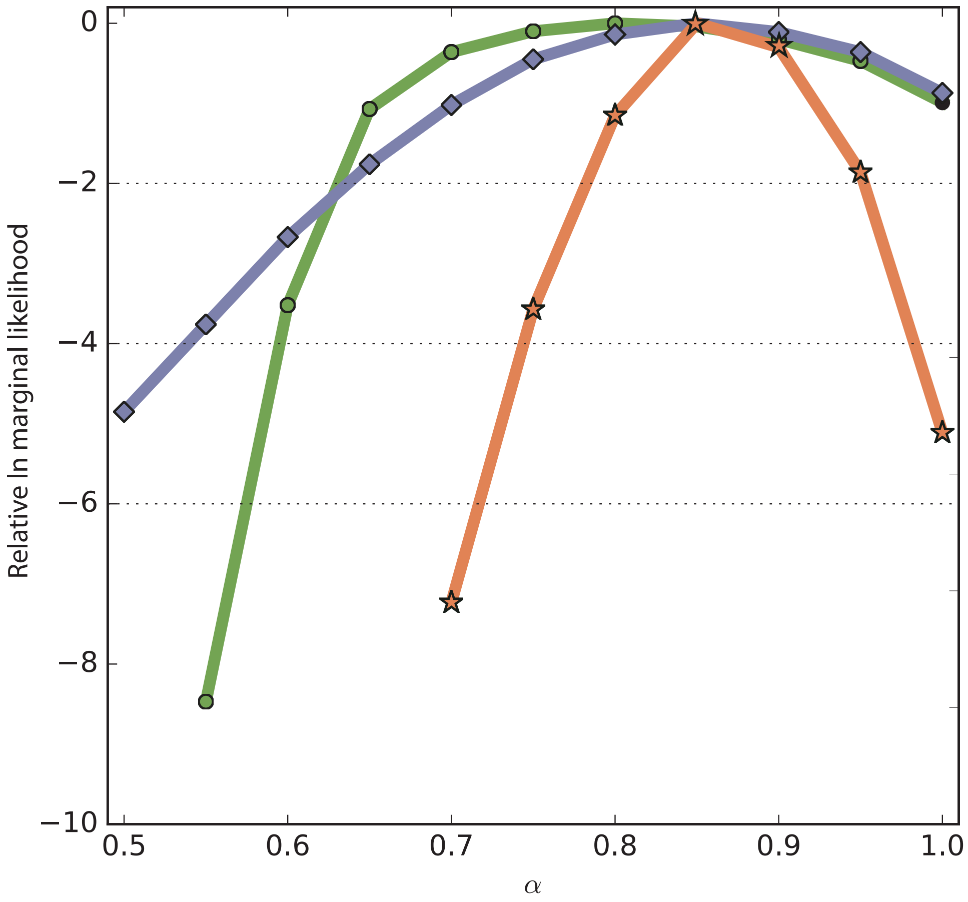
Analytical distribution of the probability of the time to the most recent common ancestor for various α for a samples size of $n = 5$.

Time to the most recent common ancestor: Comparison



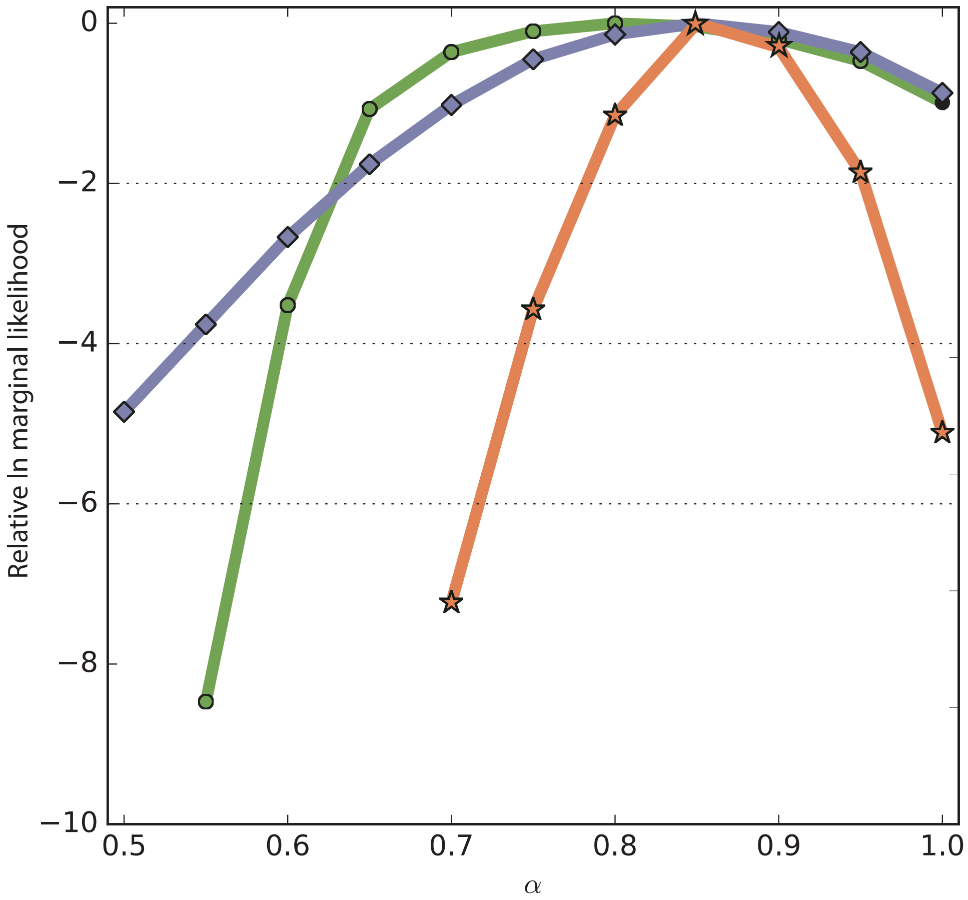
Empirical distribution of the time of the most recent common ancestor for various coalescents: (A) strictly bifurcating; (B) *fractional* coalescent versus n-coalescent and multifurcating Bolthausen-Sznitman coalescent. The x-axis is truncated at 0.03. Each curve represent a histogram of 100,000 draws of the TMRCA.

Different α: model comparison with real data



Model selection using relative marginal likelihoods of a *Plasmodium falciparum (circle)*, a H1N1 influenza (star), and a Humpback whale mtDNA (diamond) dataset.

Different α: model comparison with real data



Model selection using relative marginal likelihoods of a *Plasmodium falciparum (circle)*, a H1N1 influenza (star), and a Humpback whale mtDNA (diamond) dataset.

Influenza model with Θ and growth is 118 ln mL units worse than best model!

Thank you

Lucrezia Bieler

National Science Foundation

Florida State University

http://popgen.sc.fsu.edu