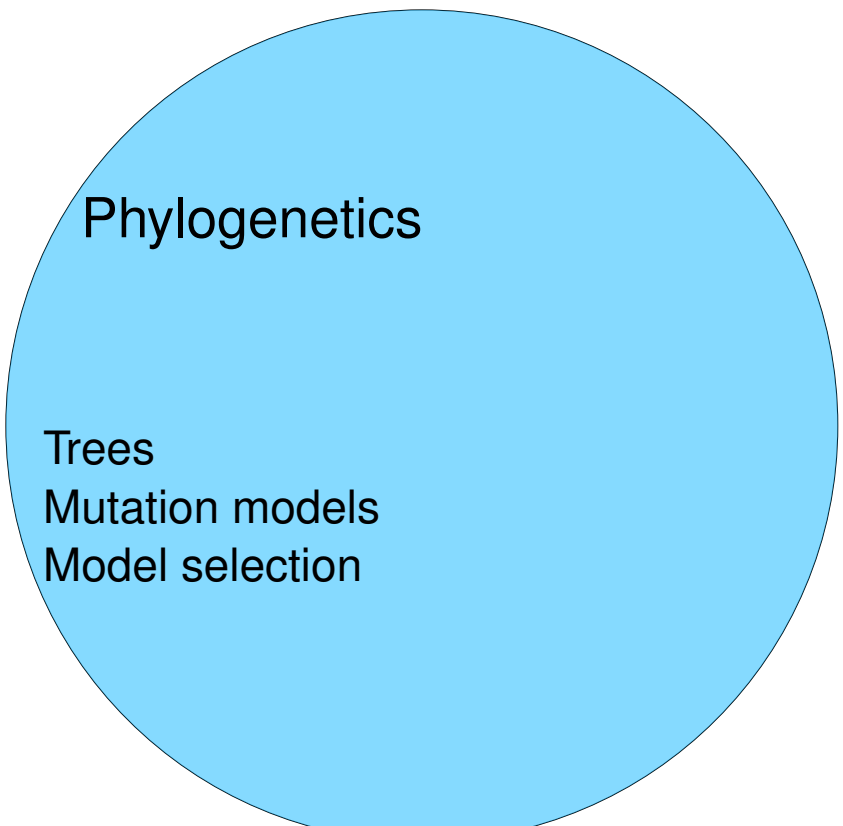
The Coalescent:

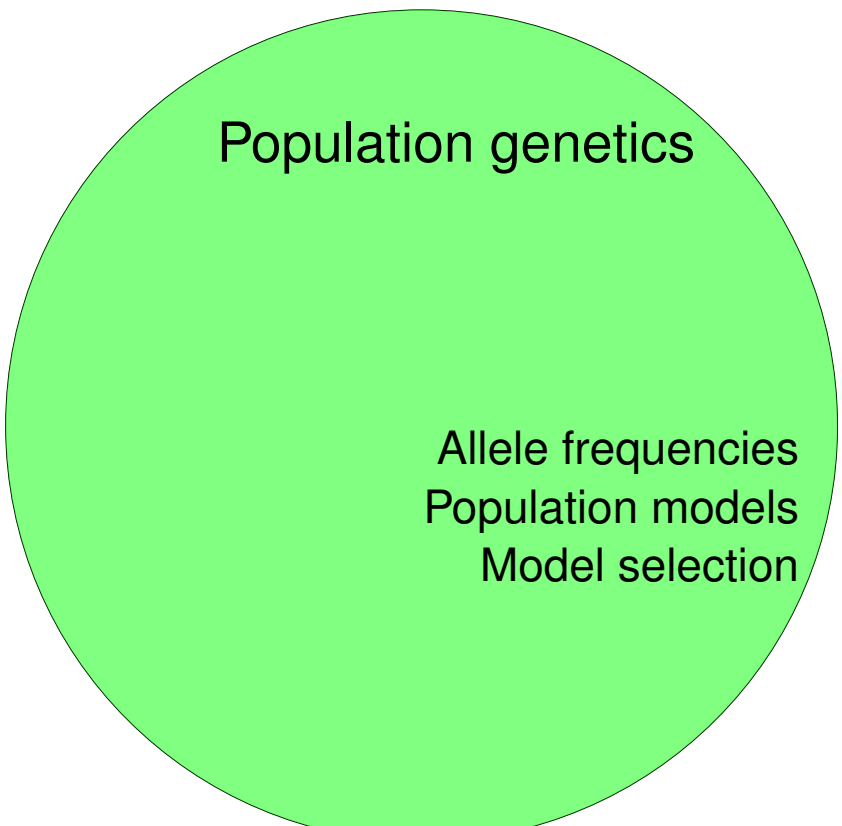
Inference using trees of ‘individuals’

Peter Beerli
Scientific Computing, Florida State University
bluesky: @peterbeerli

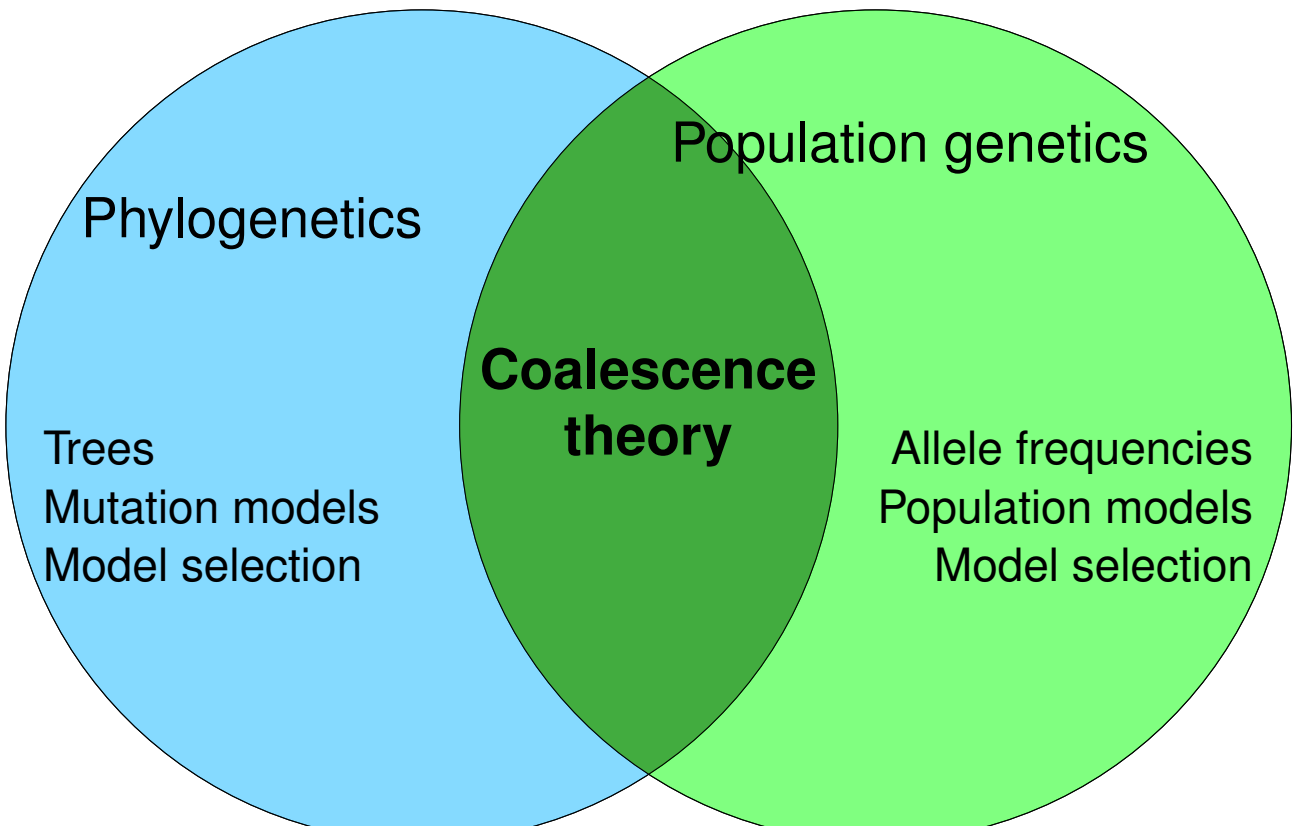
The big overview



The big overview

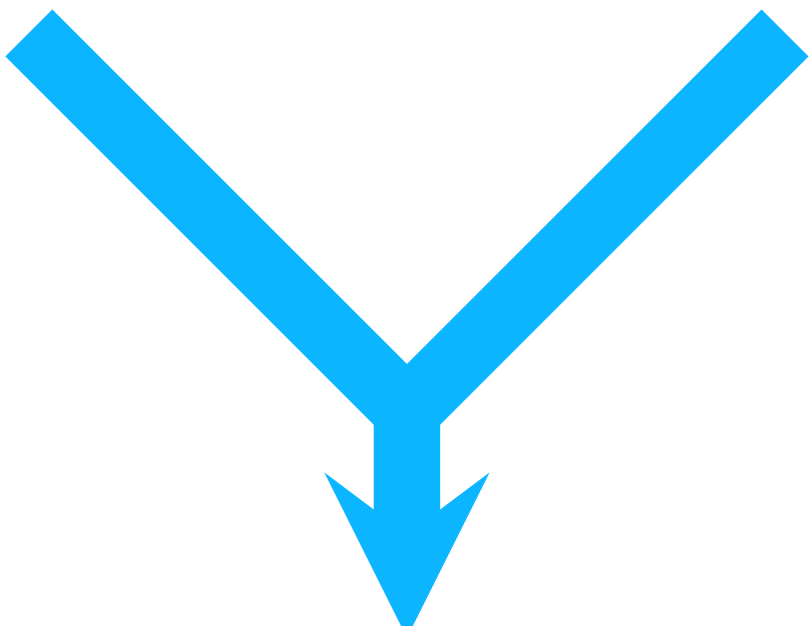


The big overview



Coalescence theory as a tool for population genetics

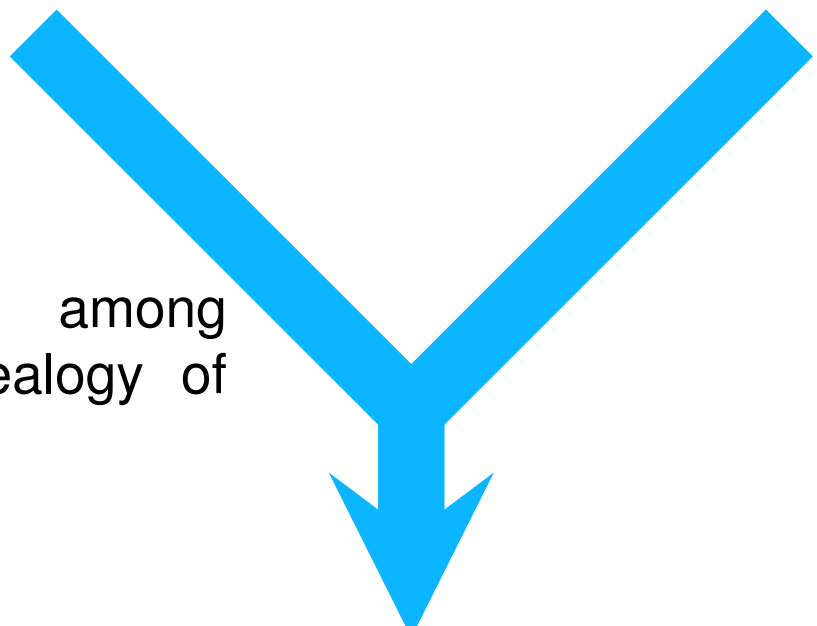
co•a•lesce |ˌkōəˈles|
verb [intrans.]
come together and form one mass or whole *: the puddles had* ***coalesced into*** *shallow streams | the separate details coalesce to form a single body of scientific thought.*
• [trans.] combine (elements) in a mass or whole *: to help coalesce the community, they established an office.*



Coalescence theory as a tool for population genetics

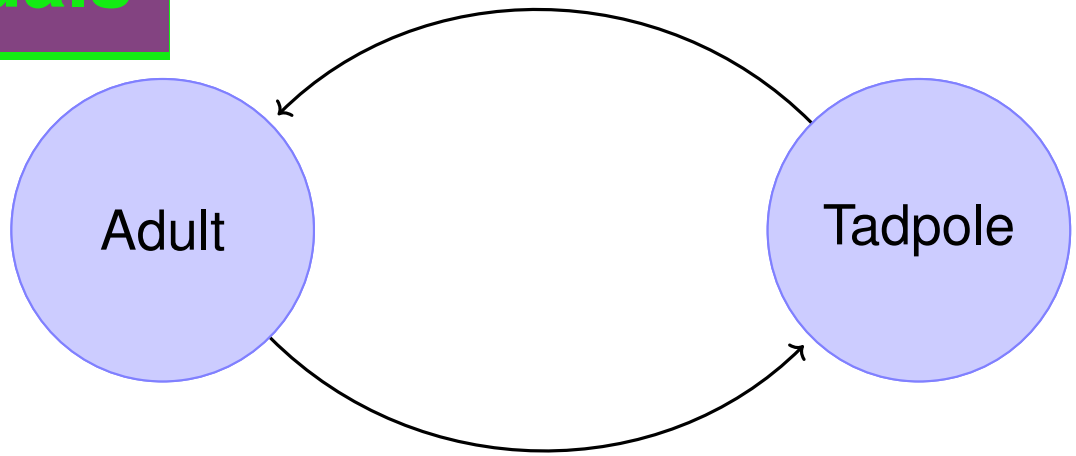
- We have data: for example, microsatellite data, single locus DNA sequences, or genomes
- we need to decide on a model to connect the data with parameters of interest.

The coalescent represent the relationship among individuals and can be expressed as a genealogy of individuals genes (not individuals)



Interaction among individuals

“Interaction” among individuals



Wright-Fisher population model

- All individuals live one generation and get replaced by their offspring
- All have same chance to reproduce, all are equally fit
- The number of individuals in the population is constant

As a result the individuals in generation n are a random draw from the previous generation $n-1$.

Population model

Past

Present

Population model

Past

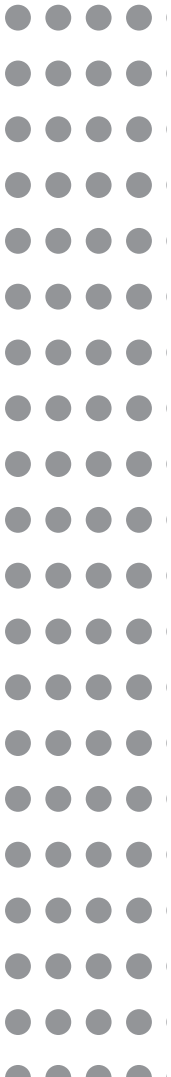
Present

Population model

Past

Present

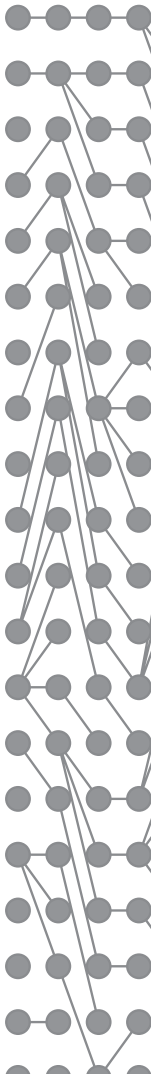
Population model



Past

Present

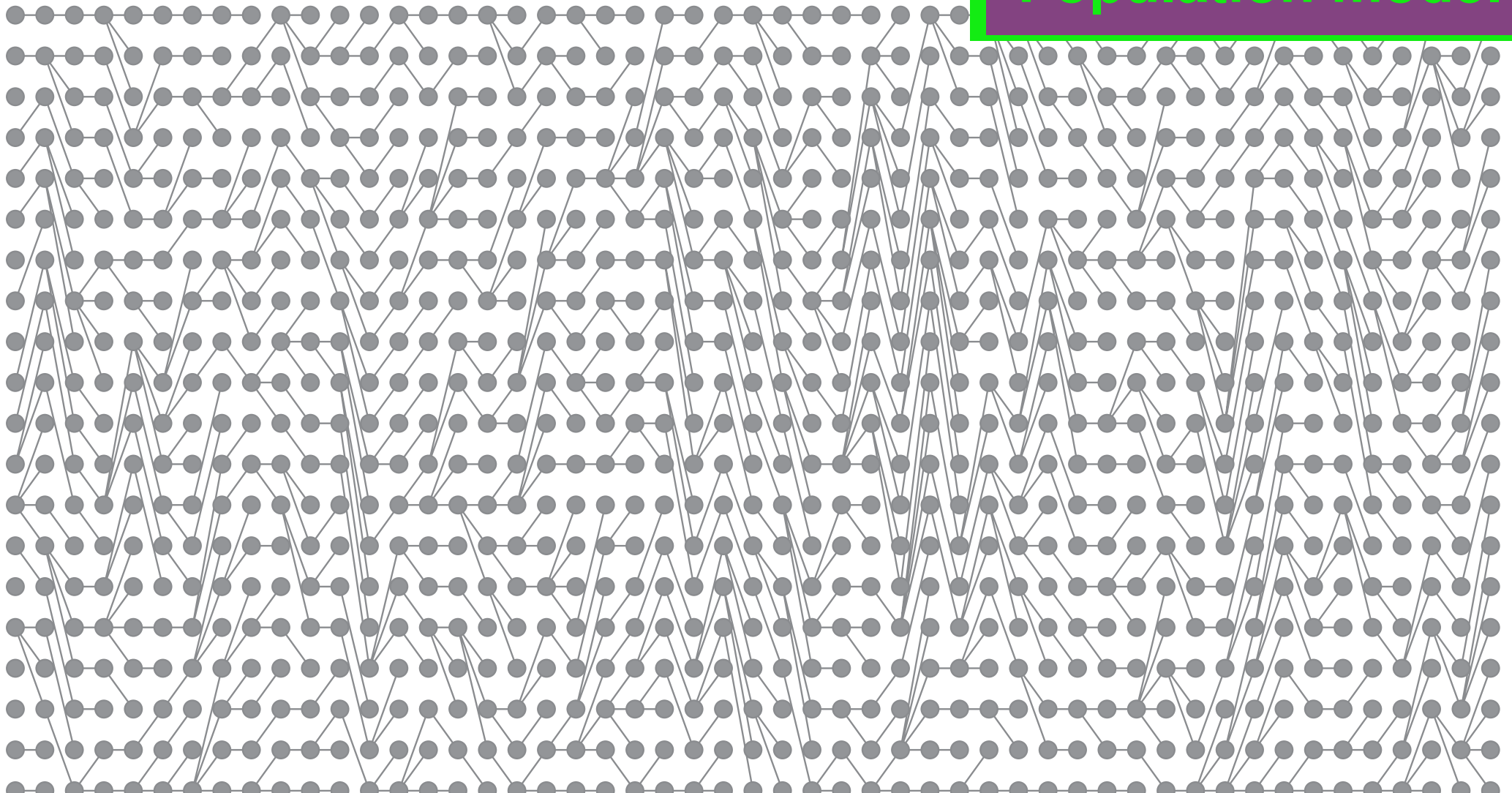
Population model



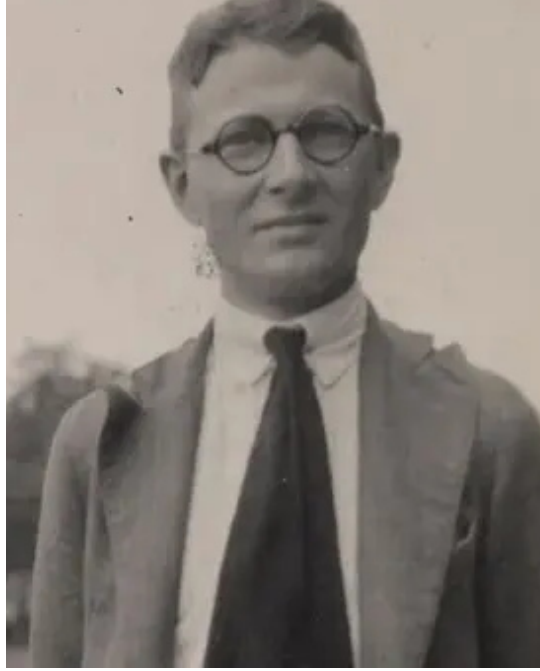
Past

Present

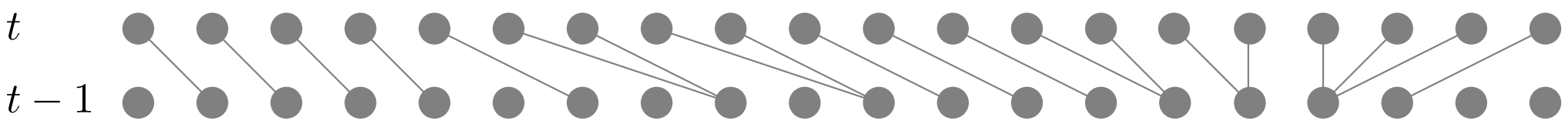
Population model



Population model

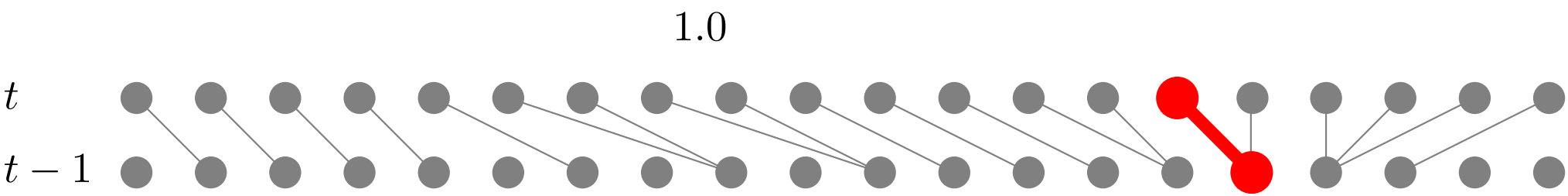


Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation
$t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is



Population model

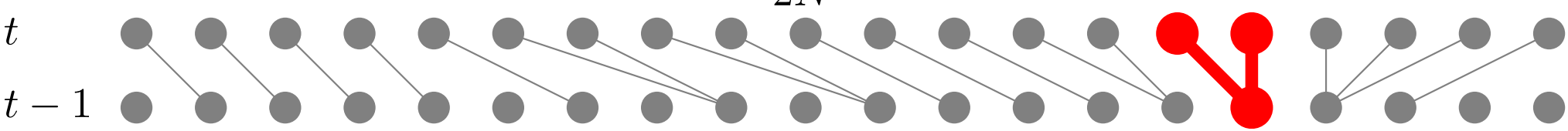
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is



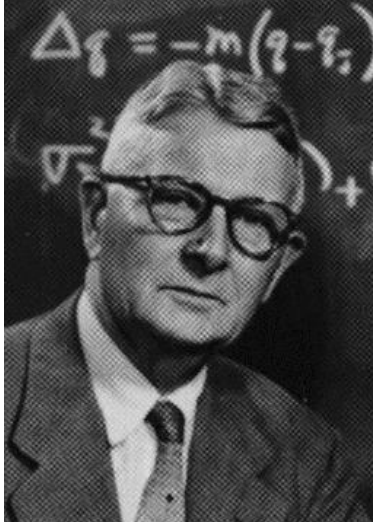
Population model

Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

$$1.0 \times \frac{1}{2N}$$

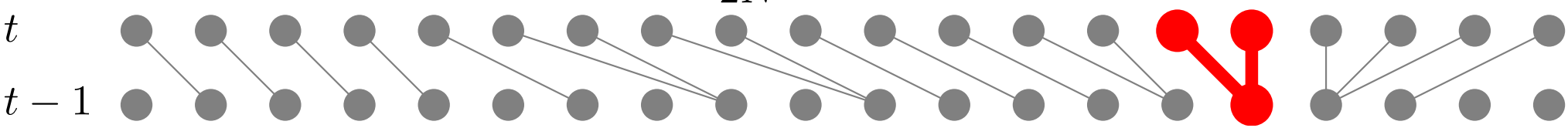


Population model



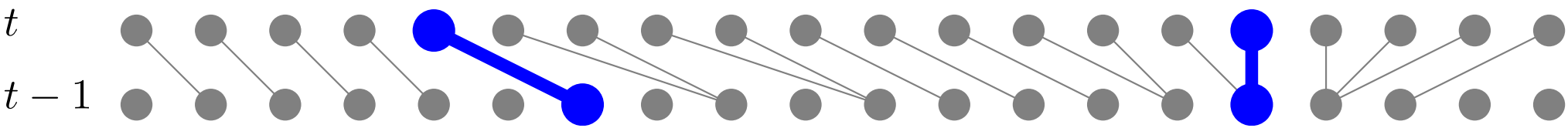
Sewall Wright evaluated the probability that two randomly chosen individuals in generation t have a common ancestor in generation $t-1$. If we assume that there are $2N$ chromosomes then the probability of sharing a common ancestor in the last generation is

$$\frac{1}{2N}$$



The probability that two randomly picked chromosome do not have a common ancestor is

$$1 - \frac{1}{2N}$$



Population model

The probability that two individuals share a common parent after t generations is

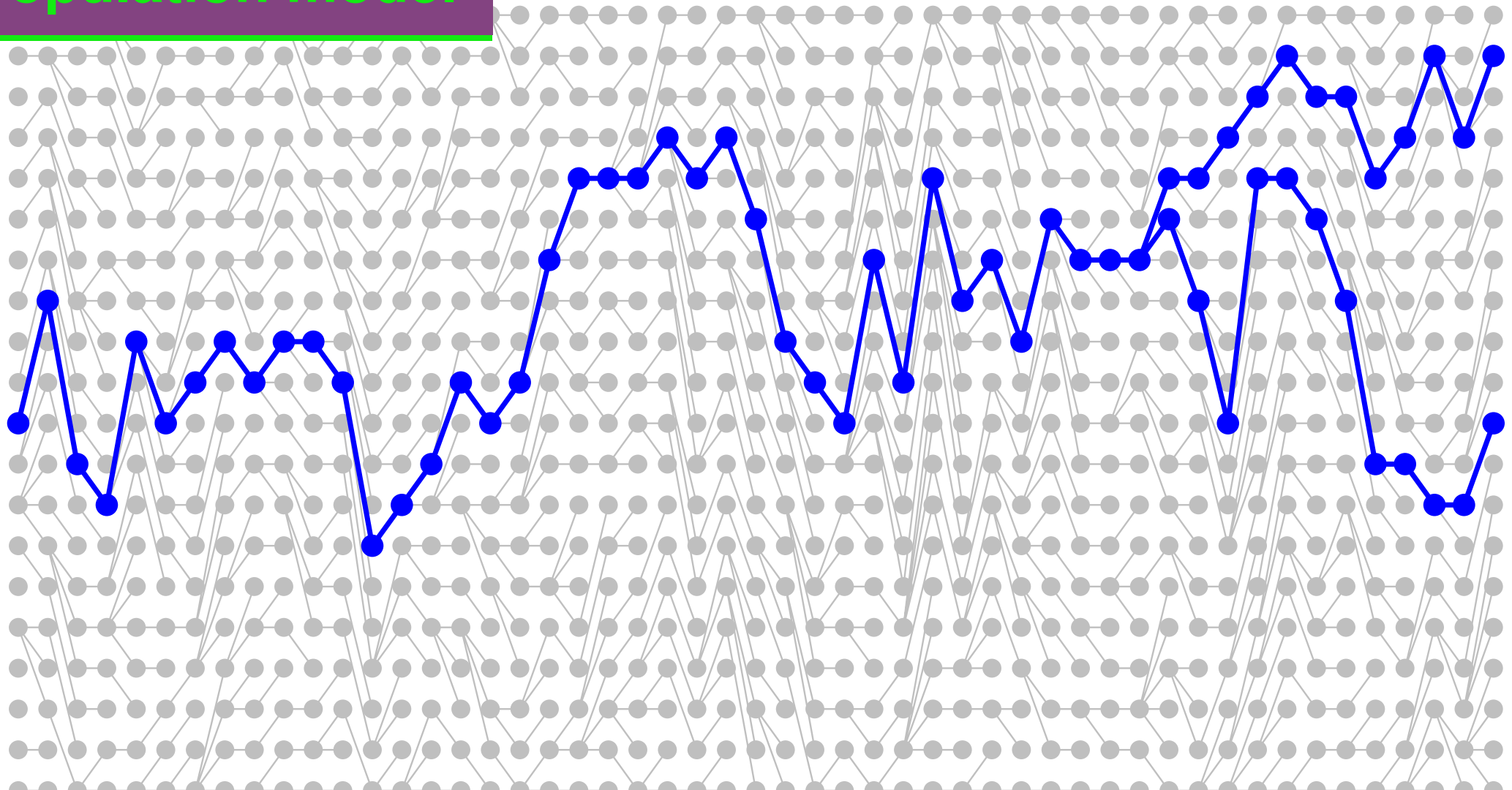
$$\mathrm{P}(t|N) = \underbrace{\left(1-\frac{1}{2N}\right) \times \left(1-\frac{1}{2N}\right) ... \times \left(1-\frac{1}{2N}\right)}_{t \text{ times}} \left(\frac{1}{2N}\right)$$

$$= \left(1-\frac{1}{2N}\right)^{t} \left(\frac{1}{2N}\right)$$

where t is the number of generations with no coalescence. This formula is known as the Geometric Distribution and we can calculate the expectation of the waiting time until two random individuals coalesce as

$$\mathbb{E}(t) = 2N$$

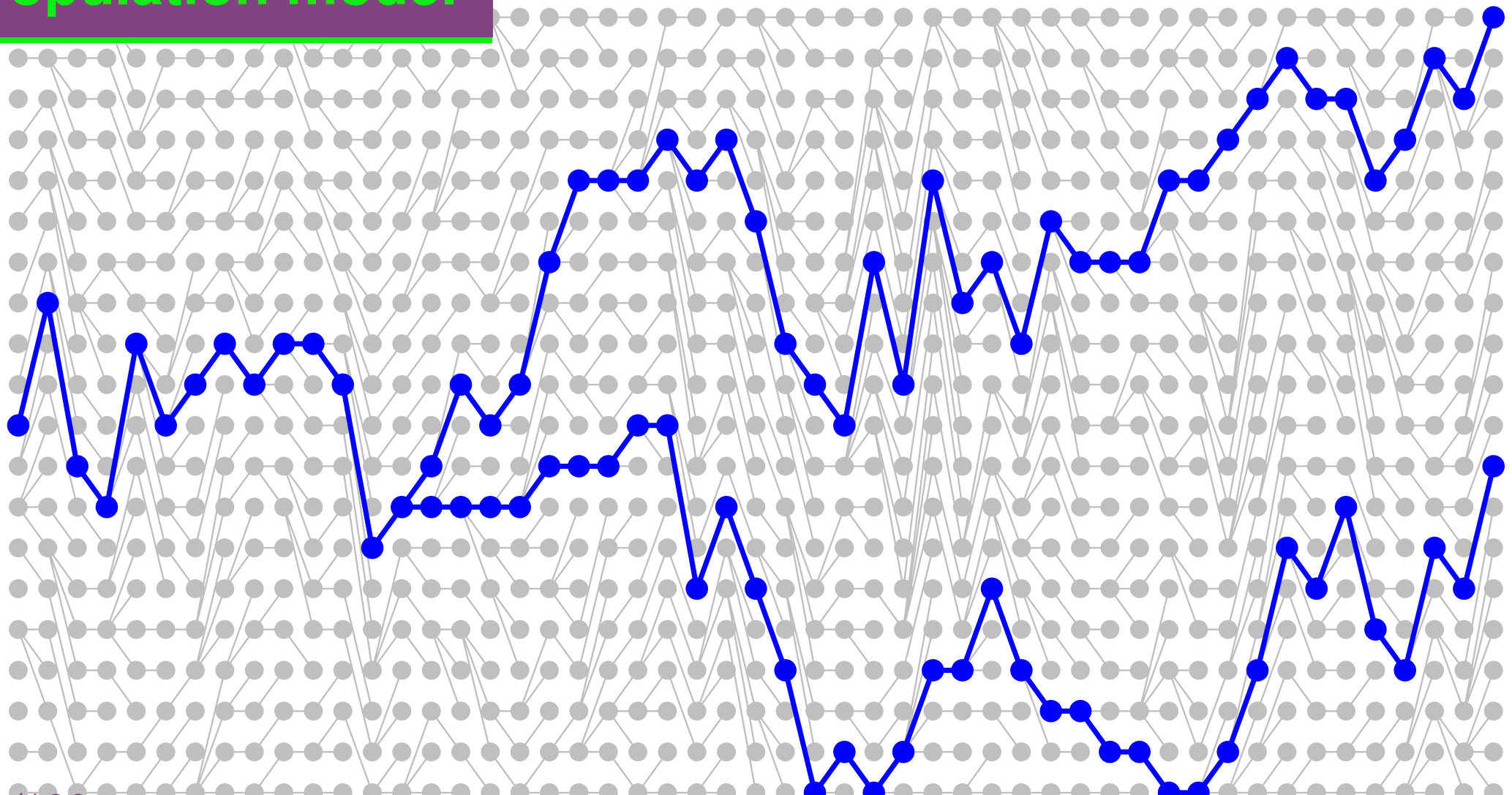
Population model



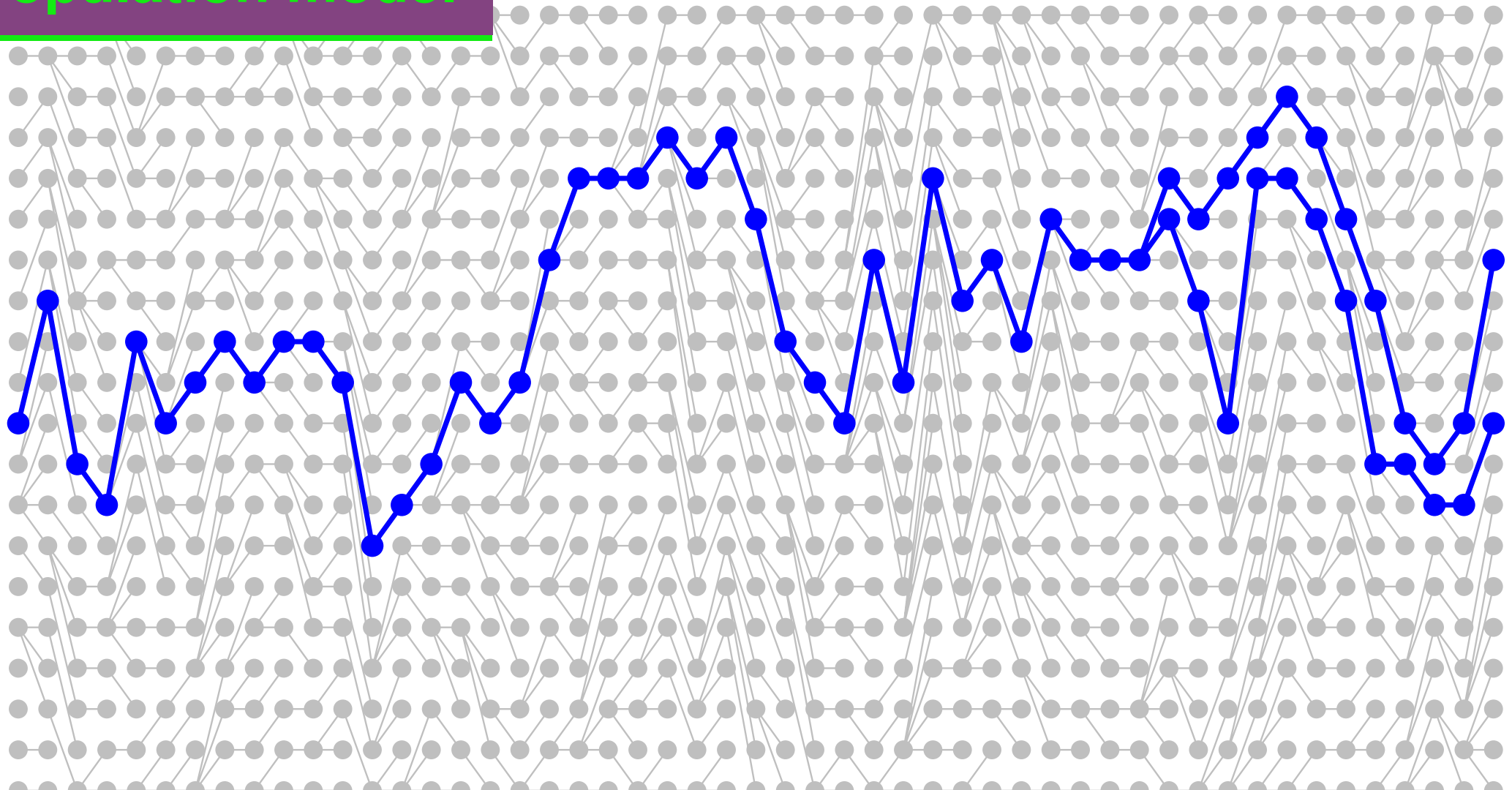
Past

Present

Population model



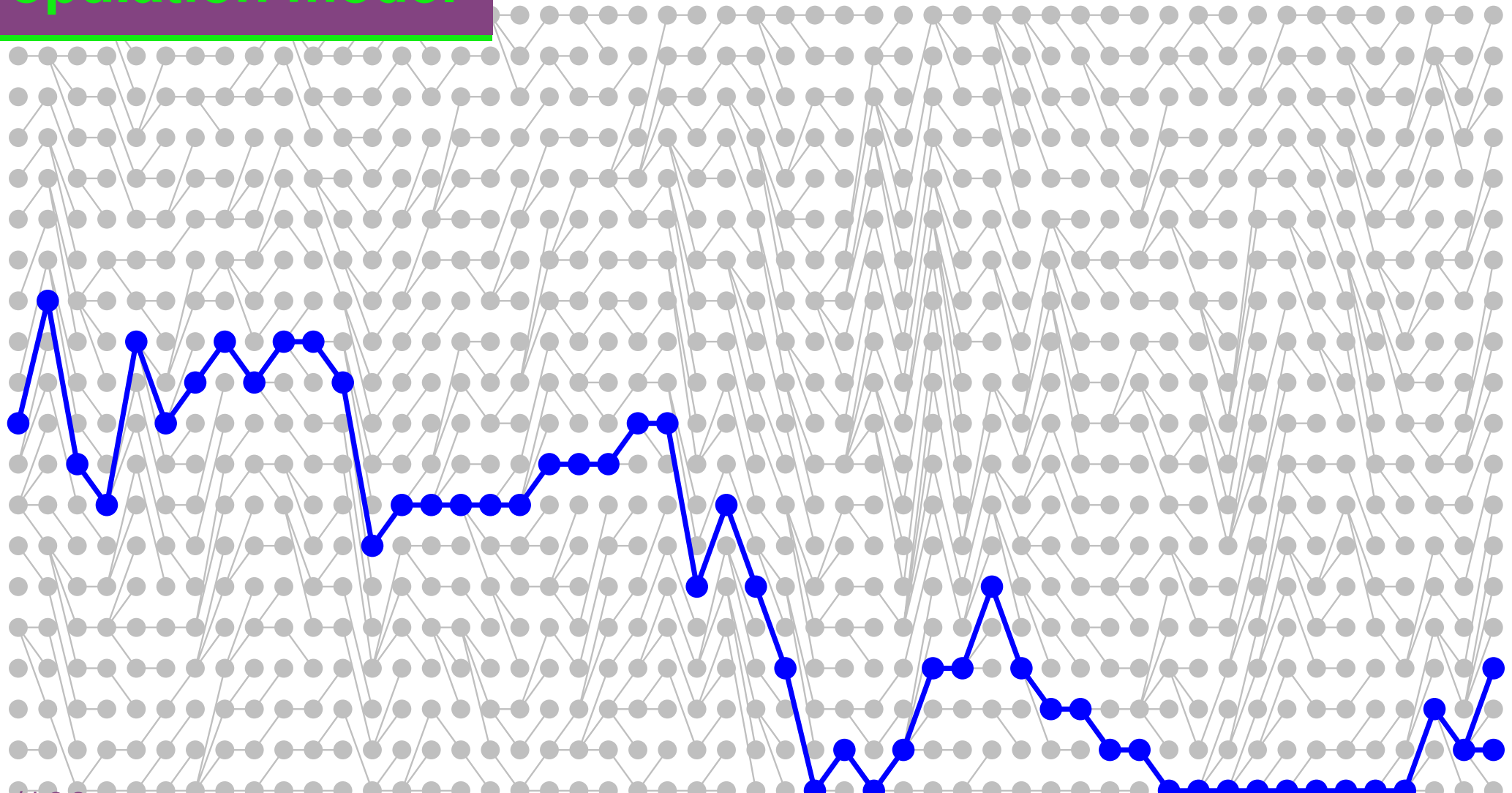
Population model



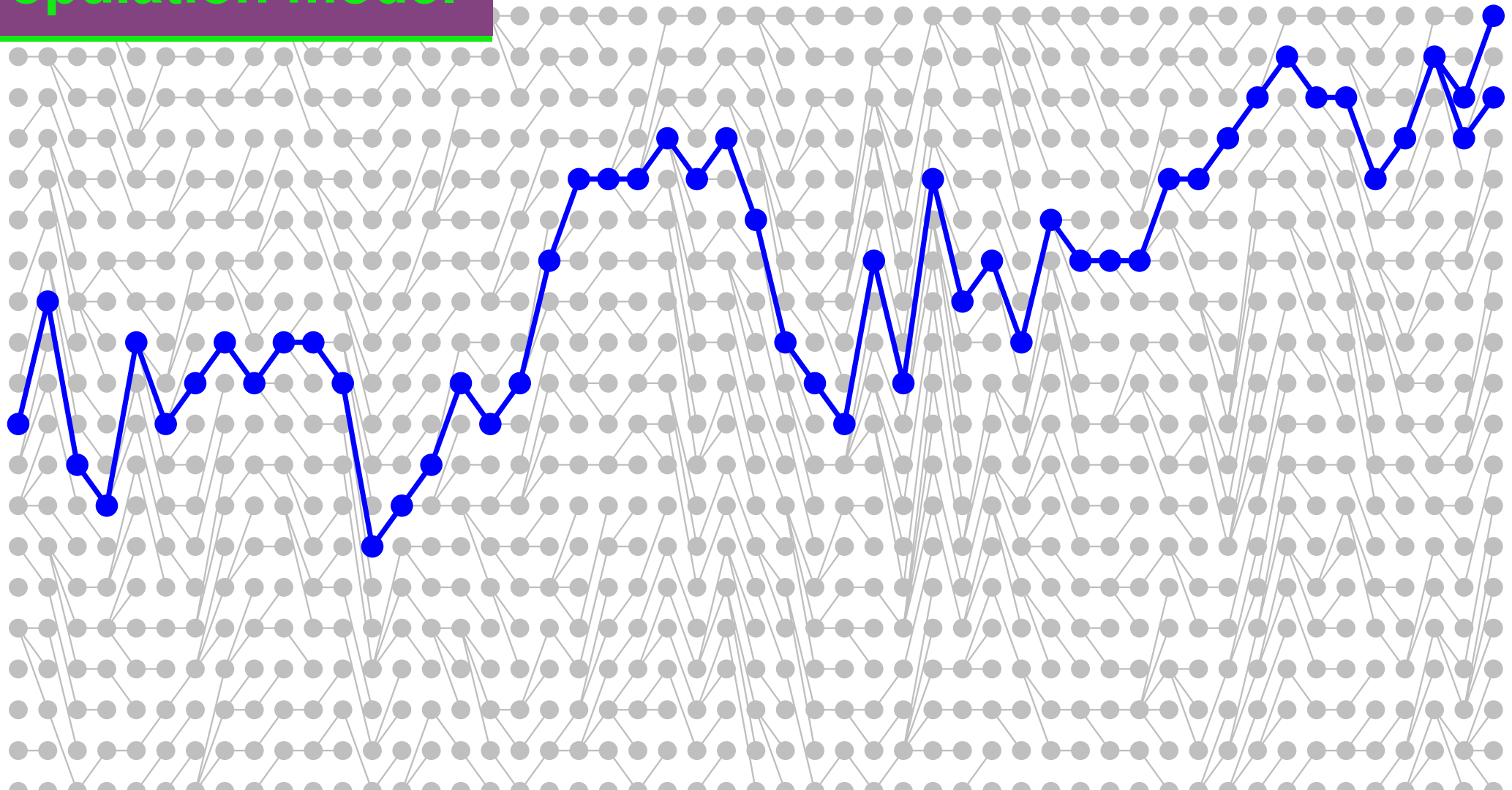
Past

Present

Population model



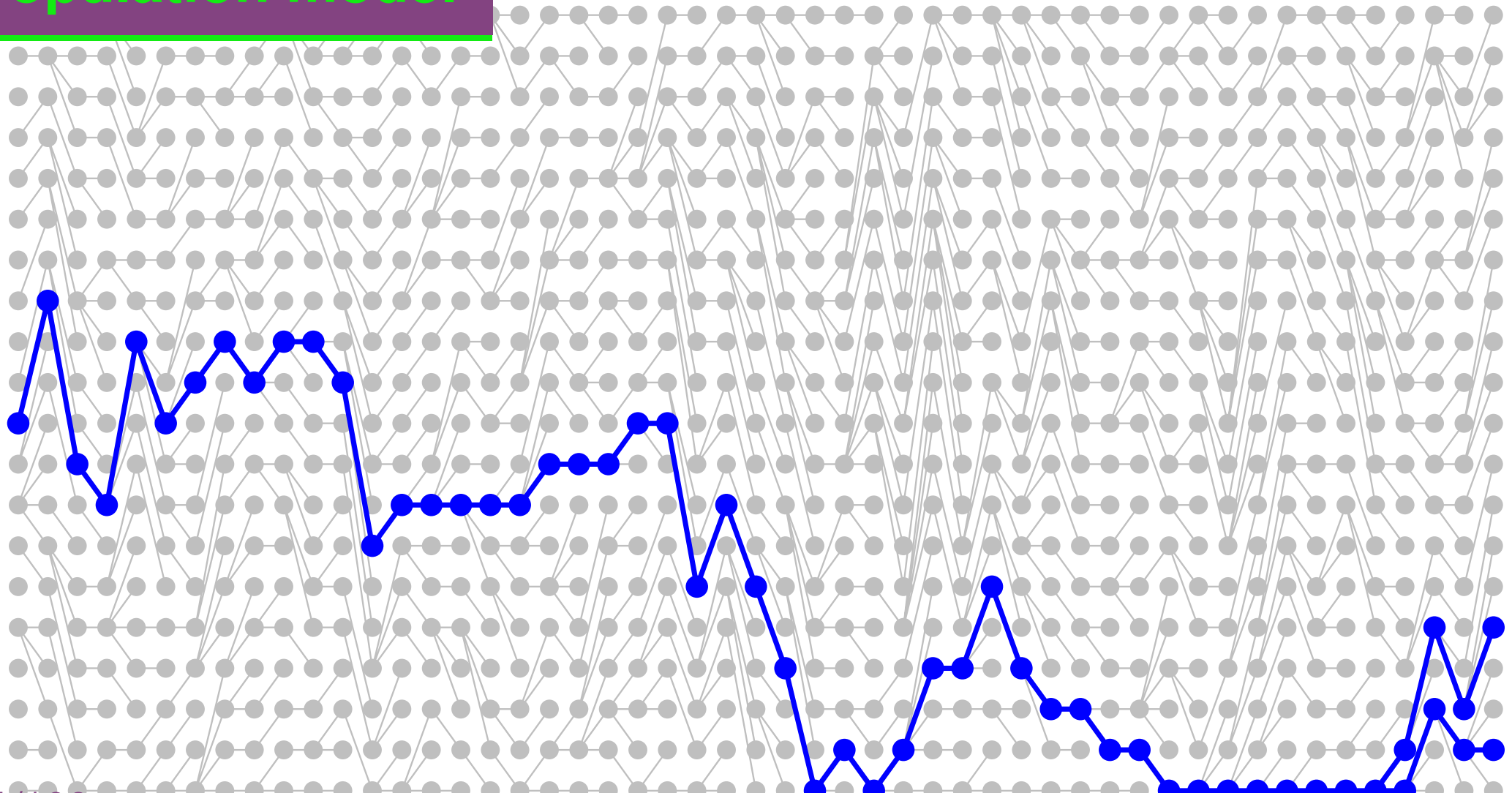
Population model



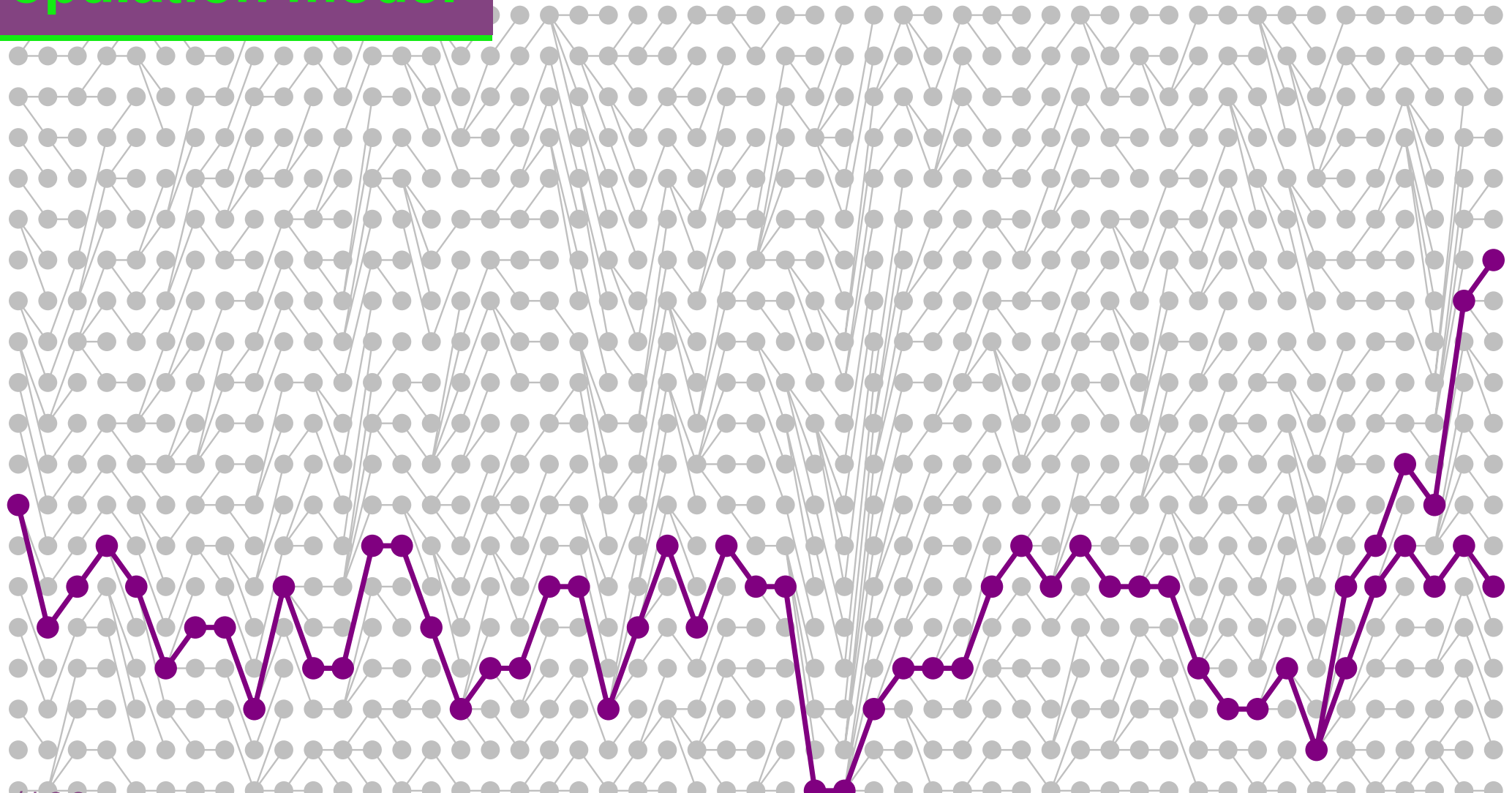
Past

Present

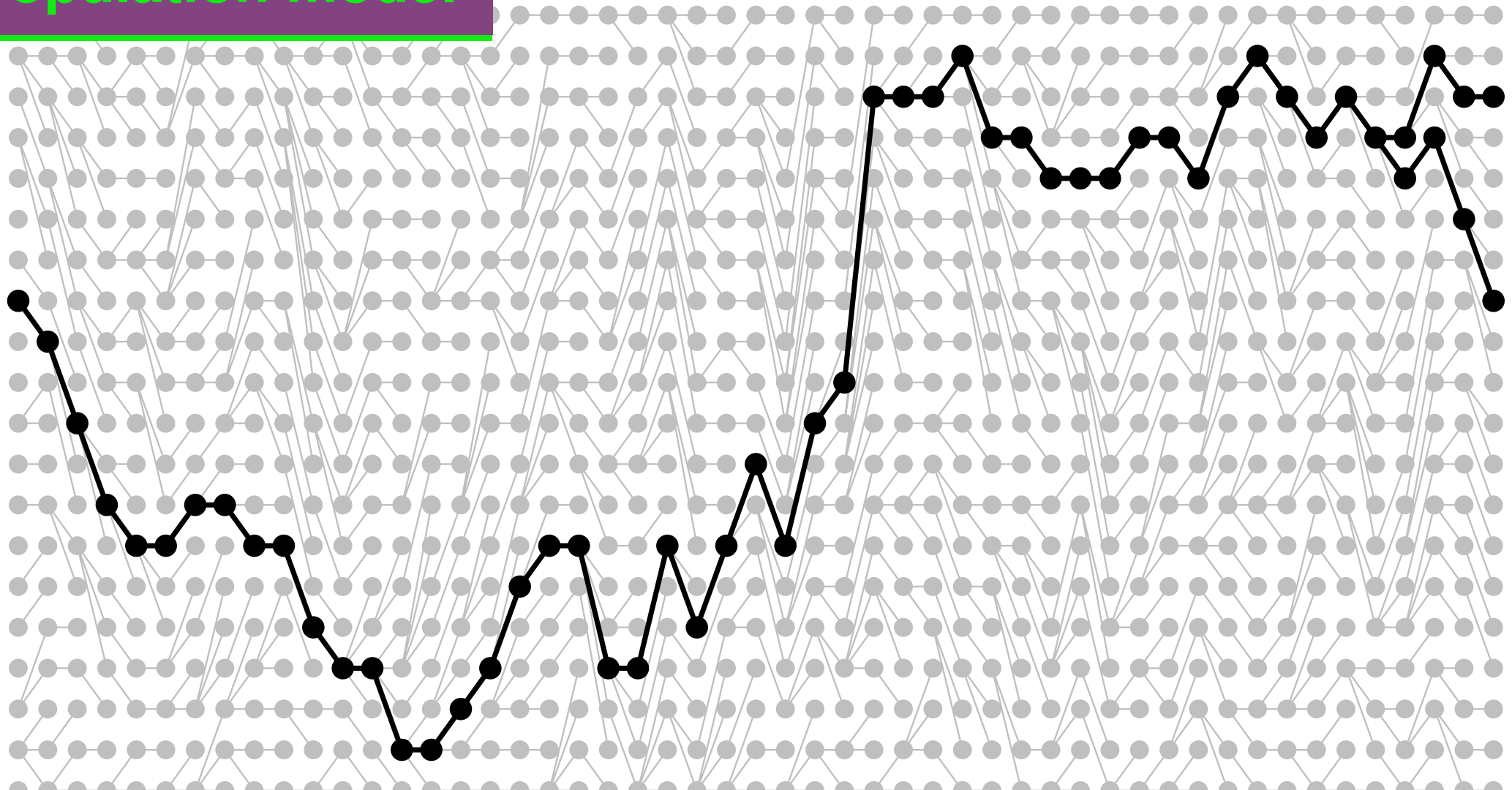
Population model



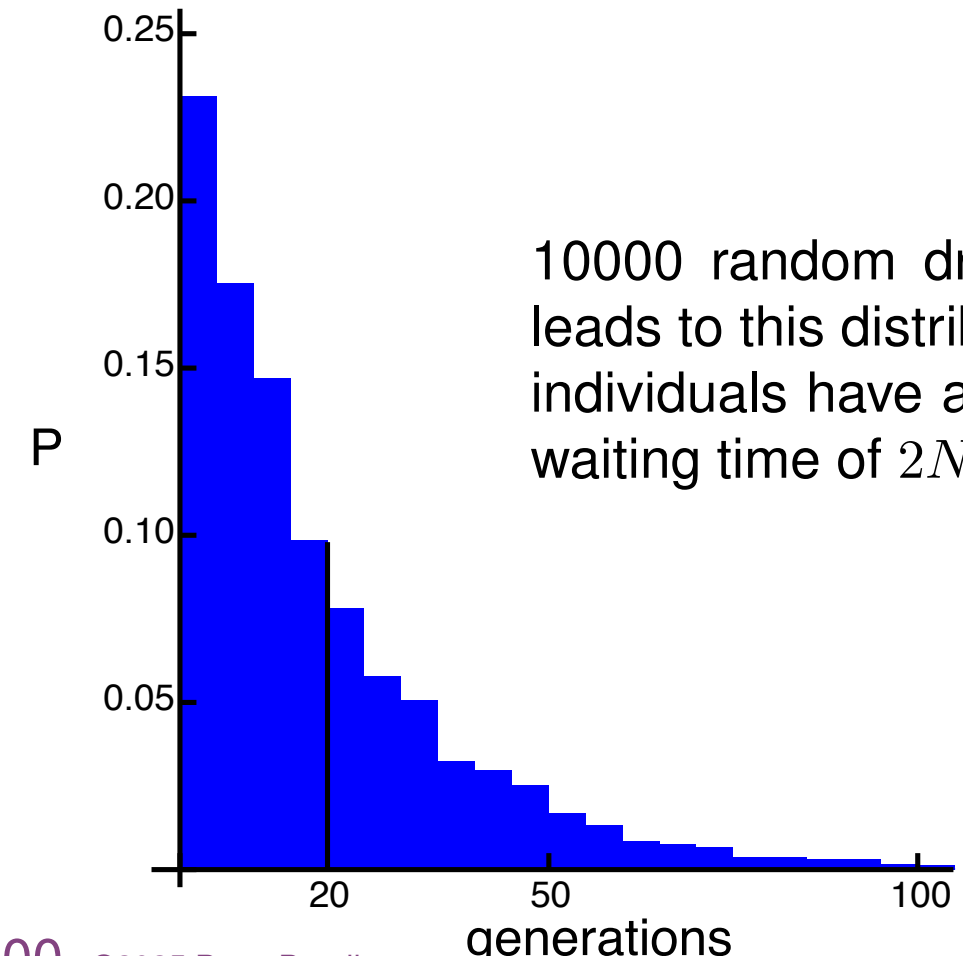
Population model



Population model



Probability Distribution



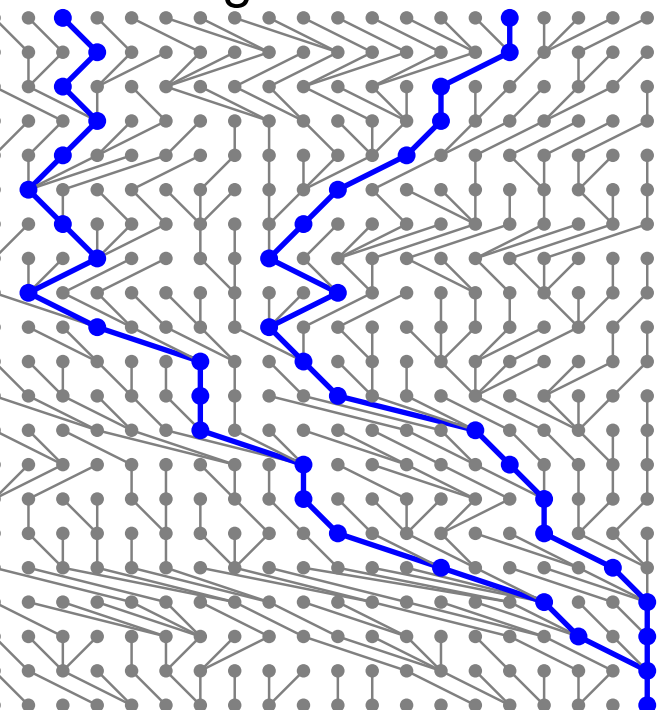
10000 random draw from a population with size $2N{=}20$ leads to this distribution of times until two randomly chosen individuals have a common ancestor. The observed mean waiting time of $2N{=}20.34$

Observations: Coalescence of two lineages

- For the time of coalescence in a sample of TWO , we will wait on average $2N$ generations assuming it is a Wright-Fisher population
- The model assumes that the generations are discrete and non-overlapping
- Real populations do not necessarily behave like a Wright-Fisher (the *'ideal' population)*
- *We assume that calculation using Wright-Fisher populations can be extrapolated to real populations.*

Other population models

Wright-Fisher



$\sigma^2_{\text{offspring}} \simeq 1$

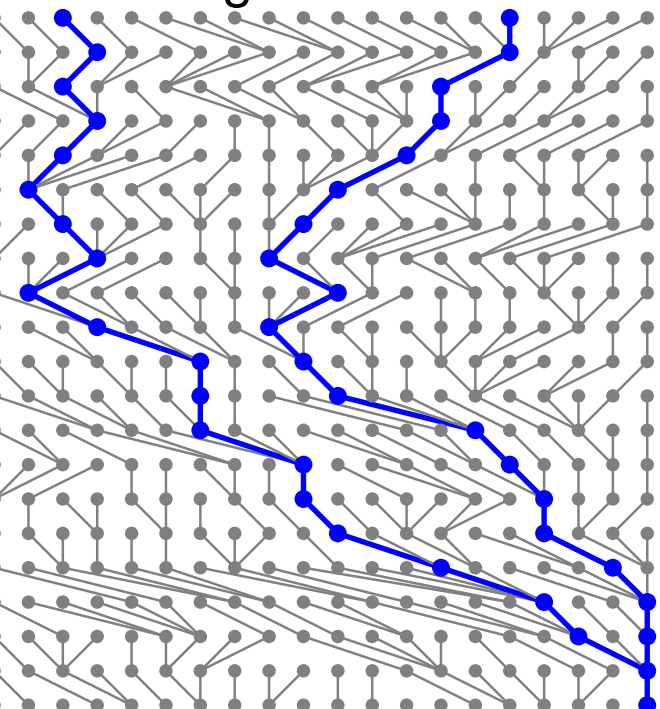
$\mathbb{E}(t) = 2N$

generation time $g = 1$

You can generate graphs like this using the python program *popsim* (check out my faculty page for the link)

Other population models

Wright-Fisher

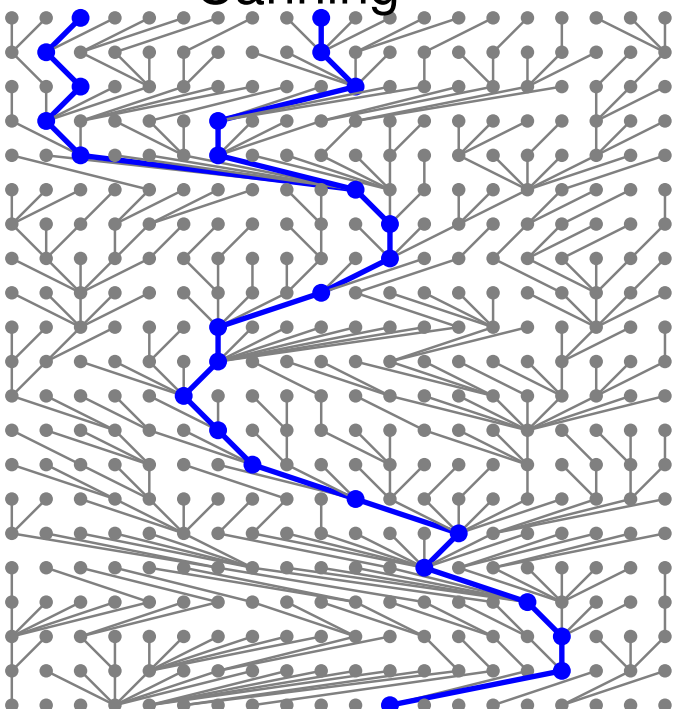


$\sigma^2_{\text{offspring}} \simeq 1$

$\mathbb{E}(t) = 2N$

generation time $g = 1$

Canning



$\sigma^2_{\text{offspring}} = x$

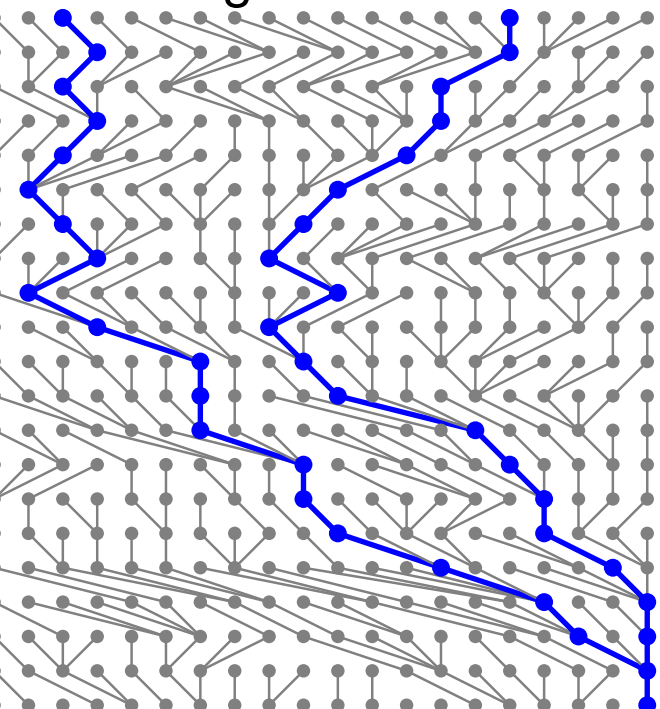
$\mathbb{E}(t) = 2N/x$

$g = 1$

  You can generate graphs like this using the python program *popsim* (check out my faculty page for the link)

Other population models

Wright-Fisher

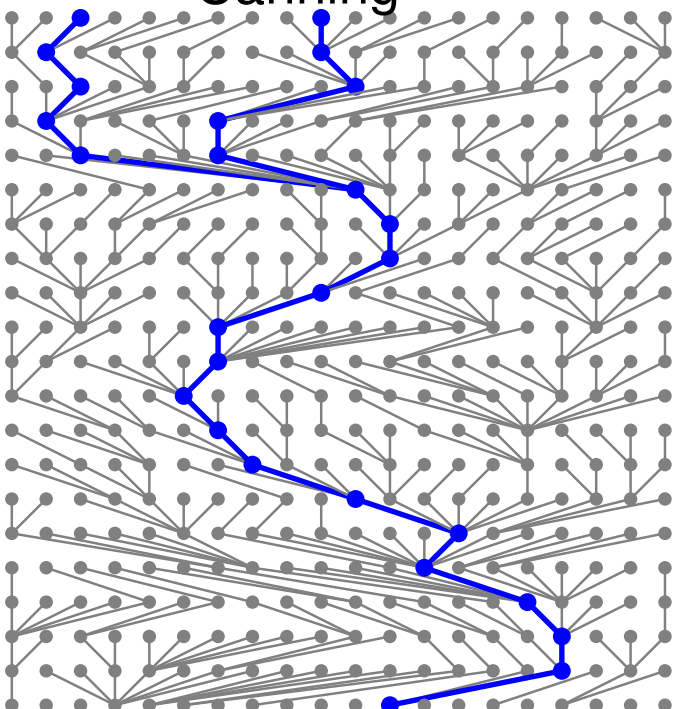


$\sigma^2_{\text{offspring}} \simeq 1$

$\mathbb{E}(t) = 2N$

generation time $g = 1$

Canning

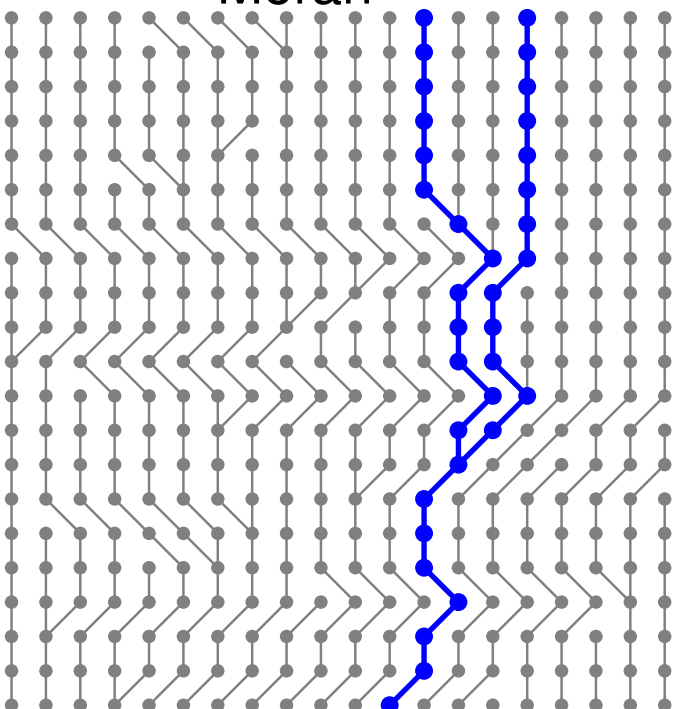


$\sigma^2_{\text{offspring}} = x$

$\mathbb{E}(t) = 2N/x$

$g = 1$

Moran



$\sigma^2_{\text{offspring}} = \frac{2}{2N}$

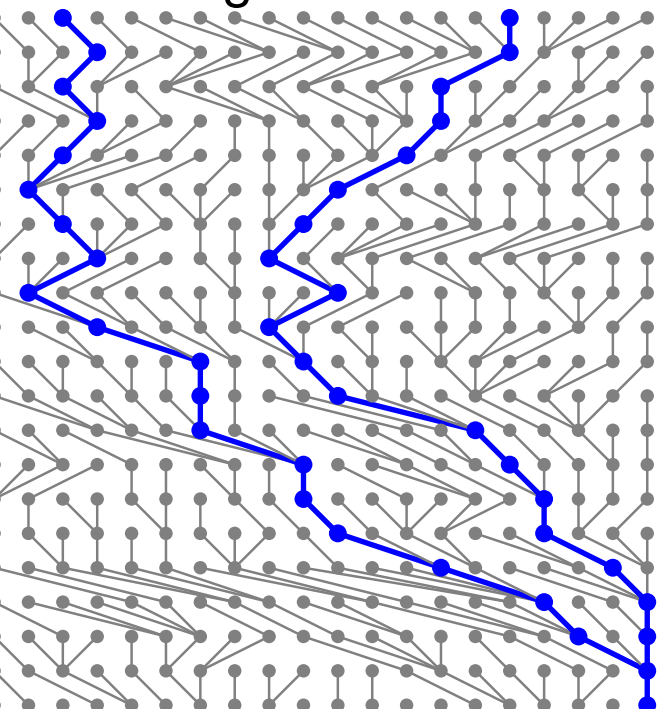
$\mathbb{E}(t) = \frac{1}{2}(2N)^2$

$g = 2N$

You can generate graphs like this using the python program *popvizard* (check out my faculty page for the link)

Other population models

Wright-Fisher

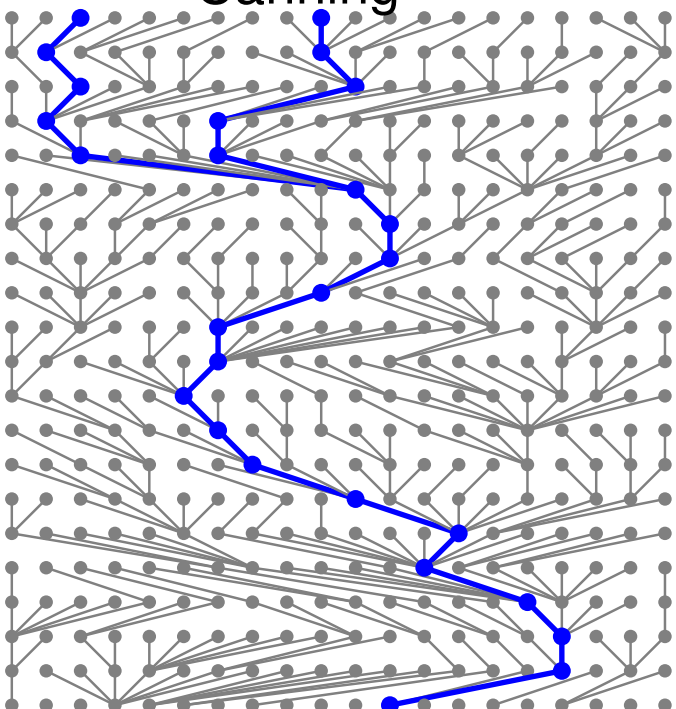


$\sigma^2_{\text{offspring}} \simeq 1$

$\mathbb{E}(t) = 2N$

generation time $g = 1$

Canning

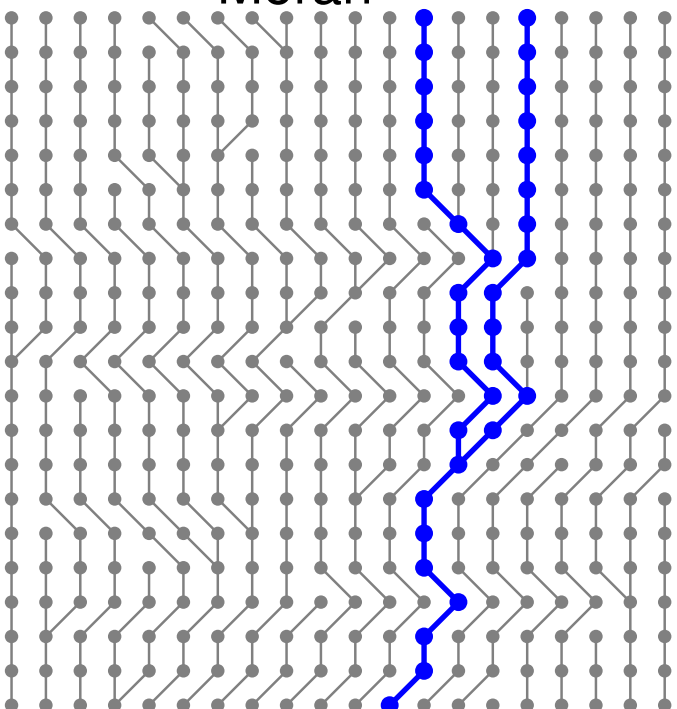


$\sigma^2_{\text{offspring}} = x$

$\mathbb{E}(t) = 2N/x$

$g = 1$

Moran



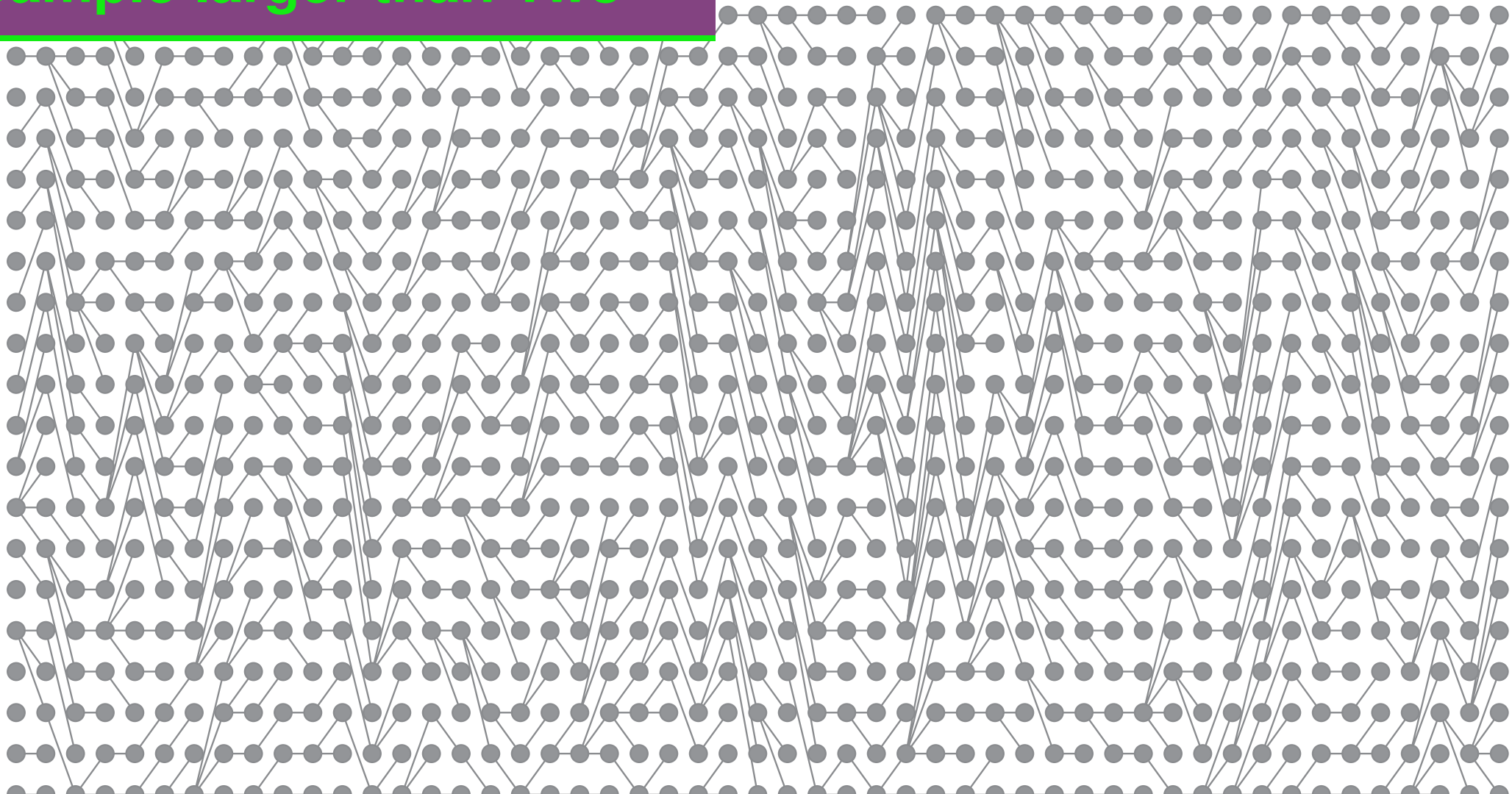
$\sigma^2_{\text{offspring}} = \frac{2}{2N}$

$\mathbb{E}(t) = \frac{1}{2}(2N)^2$

$g = 2N$

You can generate graphs like this using the python program *popsim* (check out my faculty page for the link)

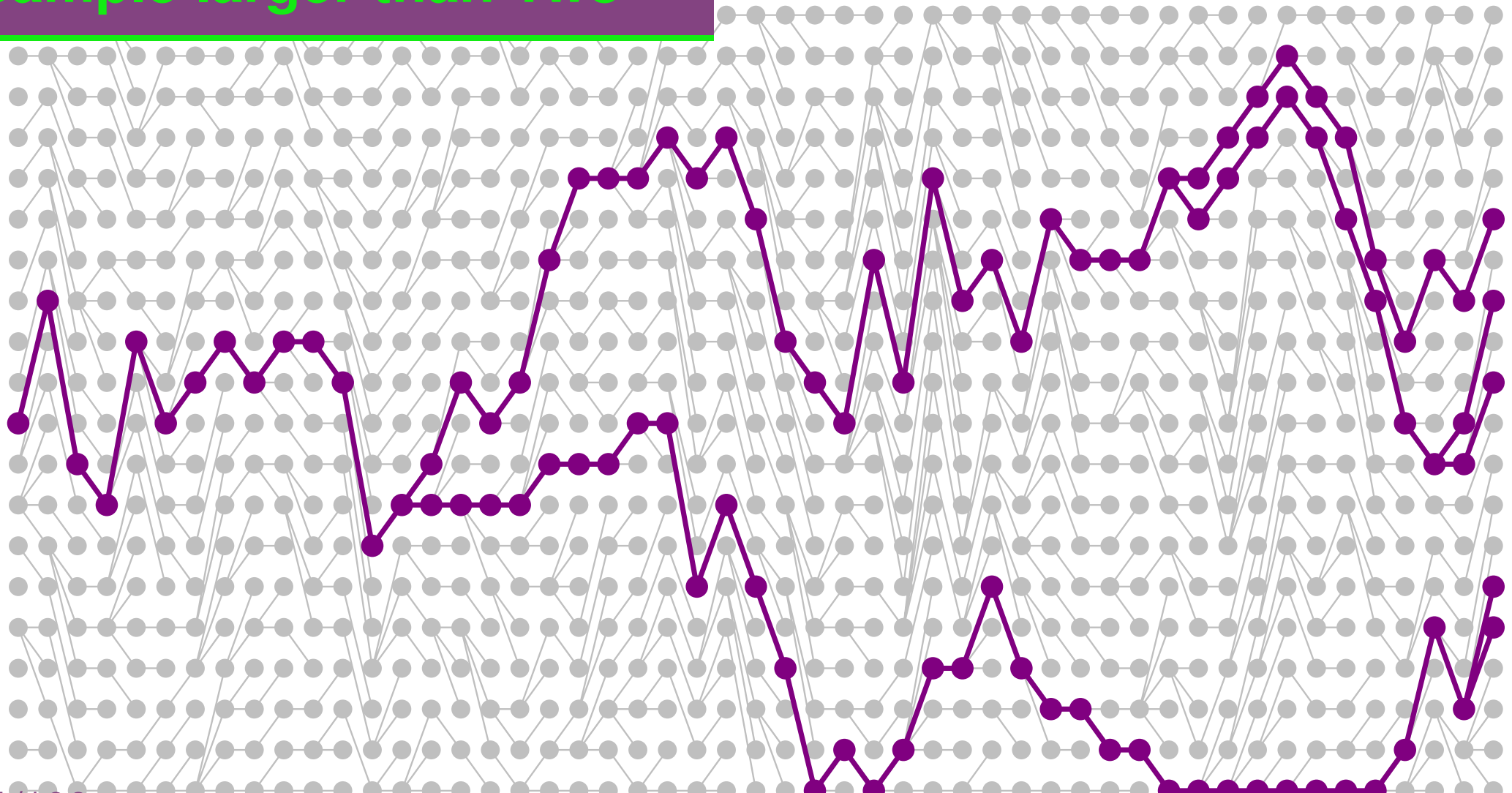
Sample larger than Two



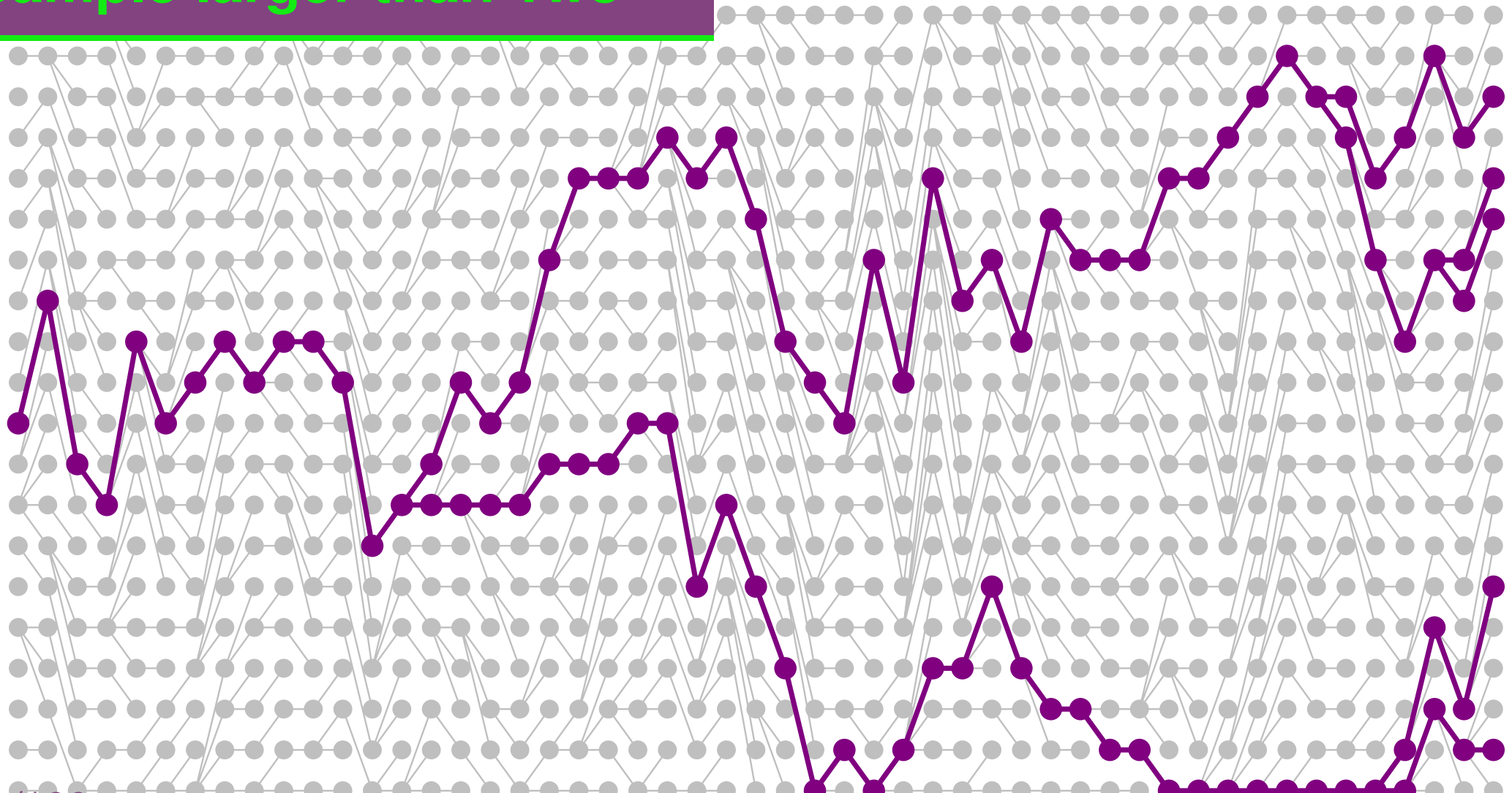
Past

Present

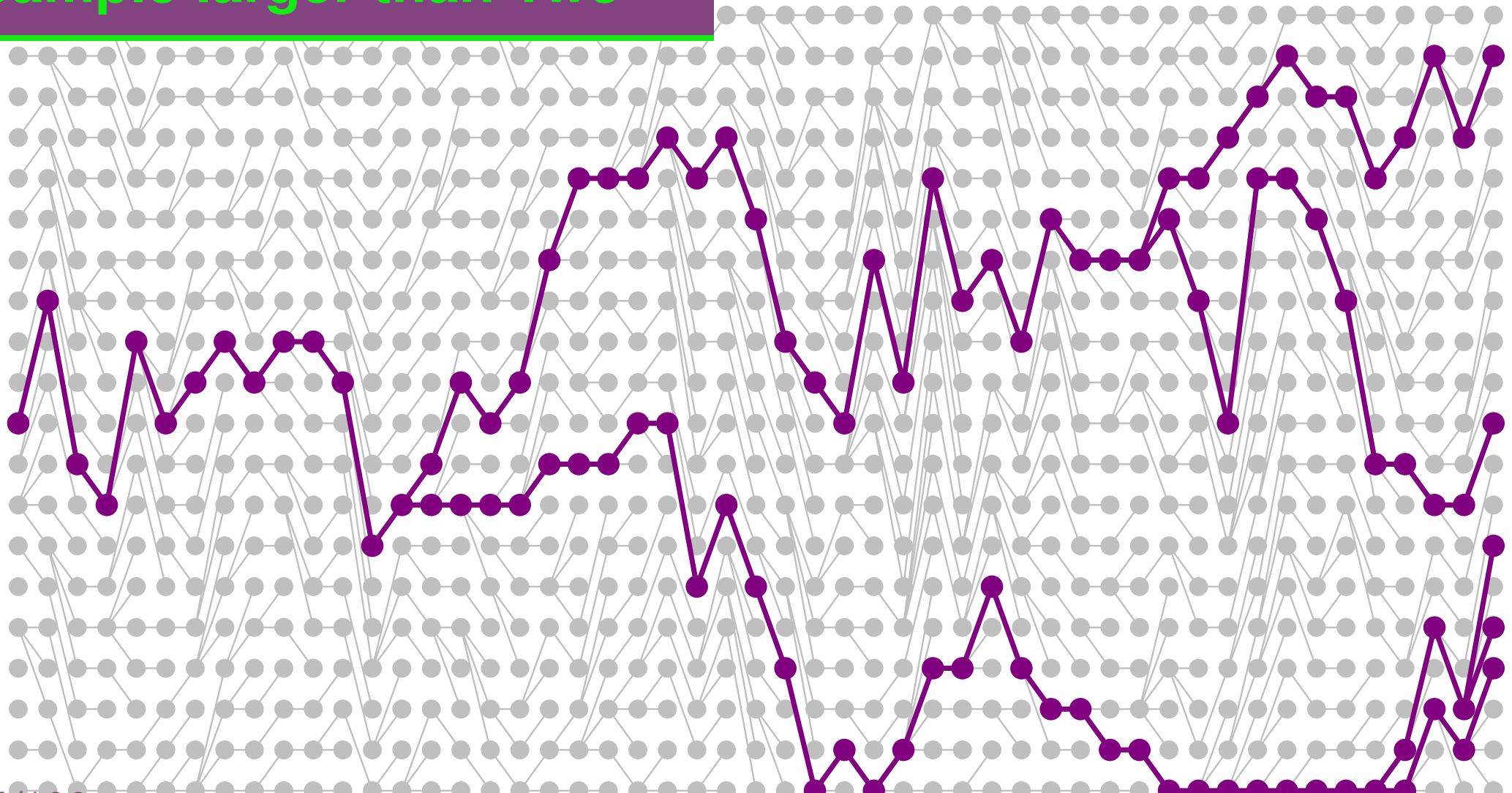
Sample larger than Two



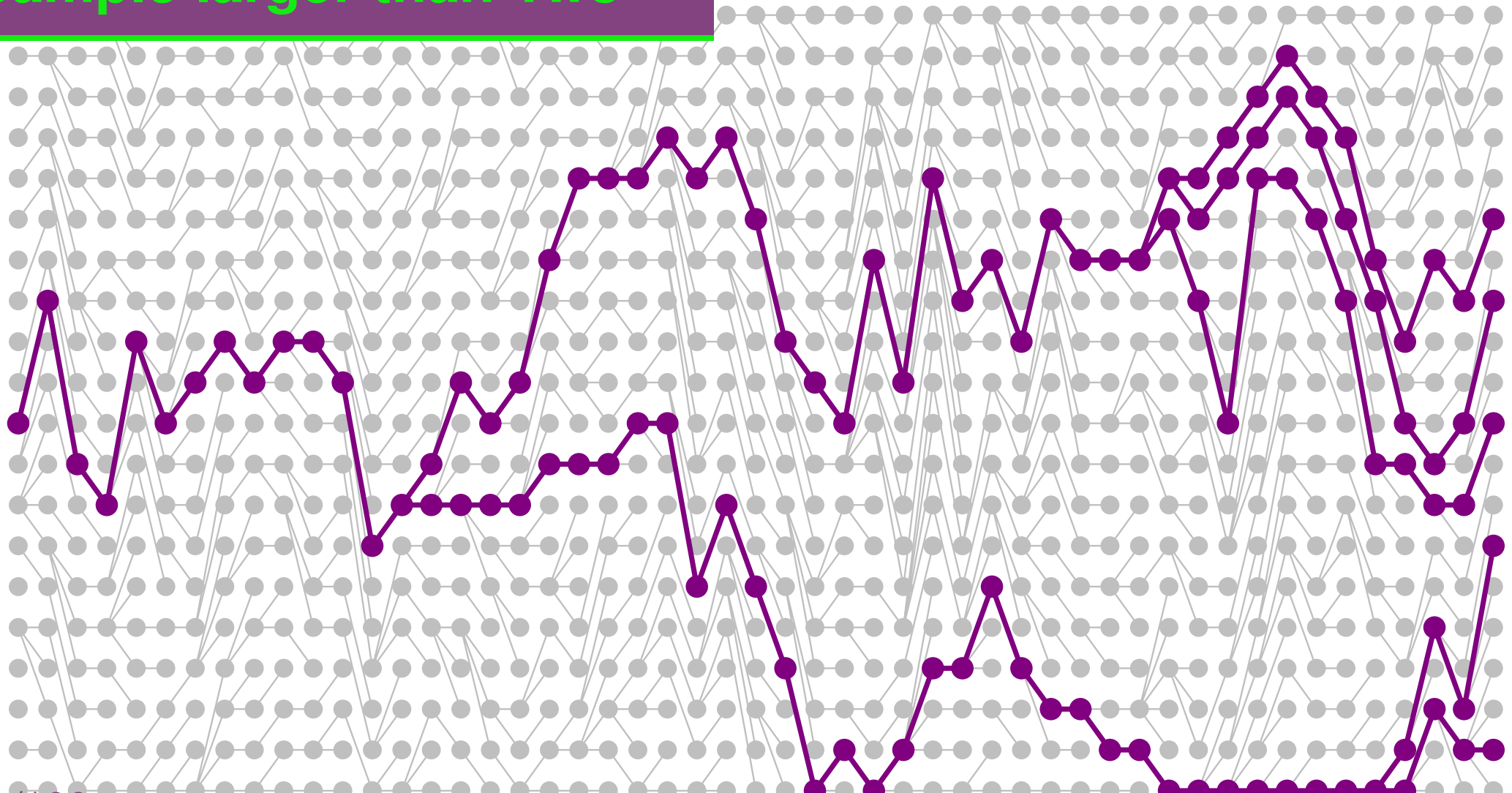
Sample larger than Two



Sample larger than Two



Sample larger than Two



Samples larger than two

Sir J. F. C. Kingman described in 1982 the n-coalecent. He showed the behavior of a sample of size n (instead of n I will use k in the following slides), and its probability structure looking backwards in time.

General findings:

$$\text{coalescence rate} = \binom{k}{2} = \frac{k(k-1)}{2}$$

Once a coalescence happened k is reduced to $k-1$ because two lineages merged into one. He then imposed a continuous approximation of the Canning's exchangeable model to get results.

Timescale

Sewall Wrights result on two lineages can be approximated:

In the discrete Wright-Fisher model we calculate the probability of non-coalescent during t generation; By using a suitable timescale τ such that one unit of scaled time corresponds to $2N$ generations, we can simplify to an continuous process

$$(1-\frac{1}{2N})^t = (1-\frac{1}{2N})^{(2N)\tau} \to e^{-\tau},$$

as N goes to infinity. For more than two lineages we use Kingman's result and use

$$e^{-\tau\binom{k}{2}}$$

for the probability of non-coalescence of k lineages during the time interval τ; we will elaborate on τ soon.

Timescale

Sewall Wrights result on two lineages can be approximated:

In the discrete Wright-Fisher model we calculate the probability of non-coalescent during t generation; By using a suitable timescale τ such that one unit of scaled time corresponds to $2N$ generations, we can simplify to an continuous process

$$(1-\frac{1}{2N})^t = (1-\frac{1}{2N})^{(2N)\tau} \to e^{-\tau},$$

For there curious: this is Poisson $\frac{\tau^k}{k!}e^{-\tau}$ with k events, here with $k{=}0 \to e^{-\tau}$

as N goes to infinity. For more than two lineages we use Kingman's result and use

$$e^{-\tau\binom{k}{2}}$$

for the probability of non-coalescence of k lineages during the time interval τ; we will elaborate on τ soon.

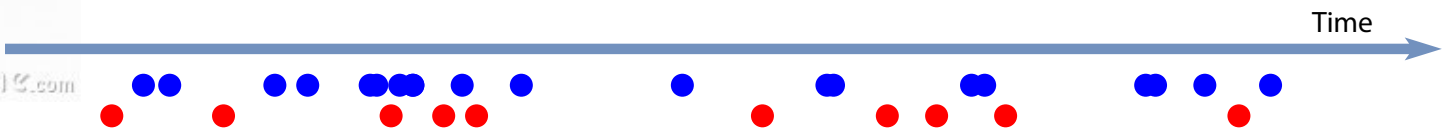
First analogy

First analogy

The time scale here is arbitrary, for example if the rate is 2 calls per 10 minutes; we then have a probability of getting no call for 10 minutes as

$$e^{-10\times 2/10} = 0.135$$

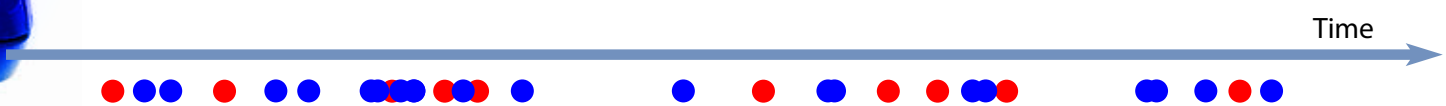
First analogy



If another type of call has a rate of 4 calls per 10 minutes; we then have a probability of getting no call for 10 minutes as I

$$e^{-10\times 4/10} = 0.018$$

First analogy

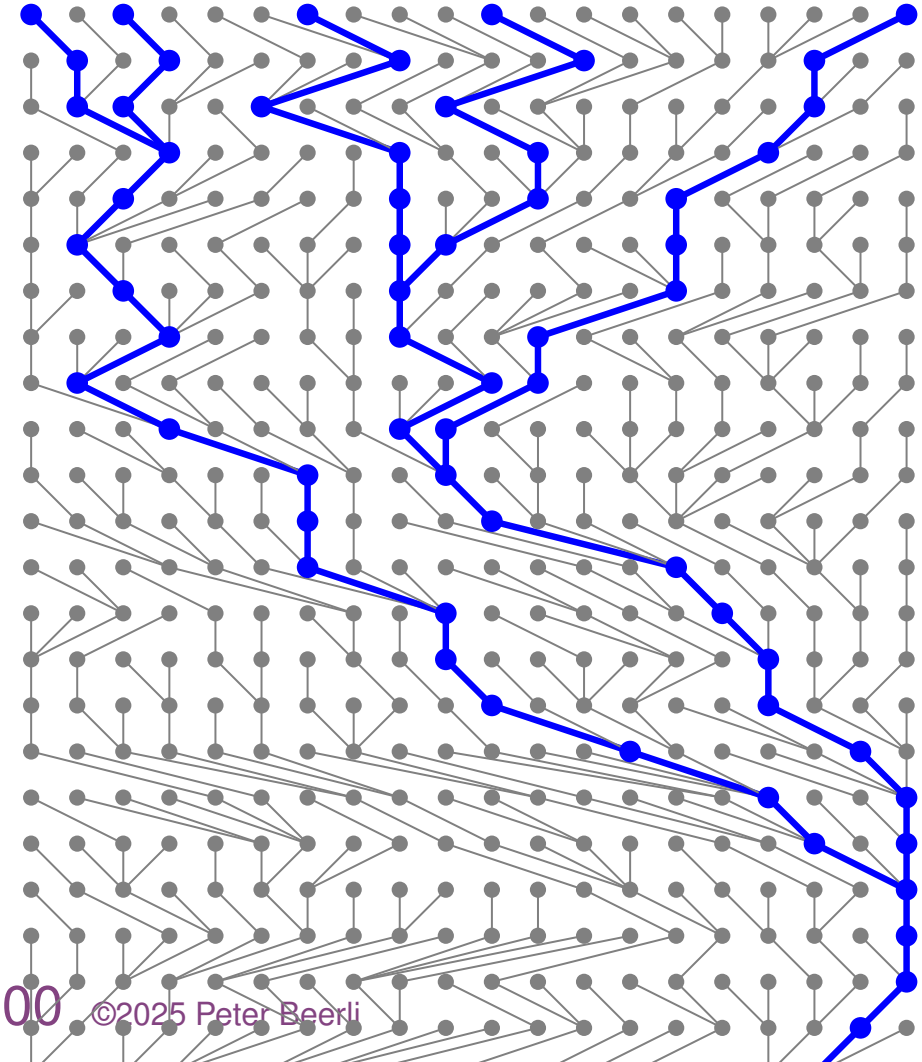


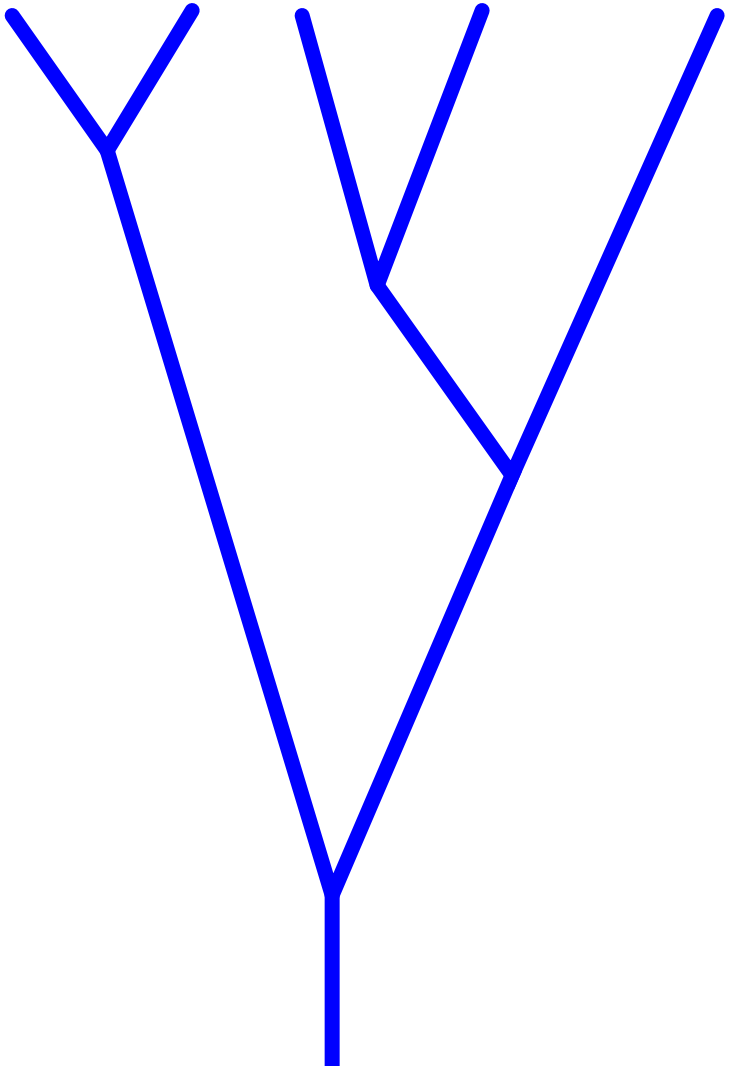
Having two type of calls with different rates $2/10$ and $4/10$; we then have a probability of getting no call for 10 minutes as

$$e^{-10\times(2/10)}\times e^{-10\times(4/10)} = e^{-10\times(2/10+4/10)} = 0.0024$$

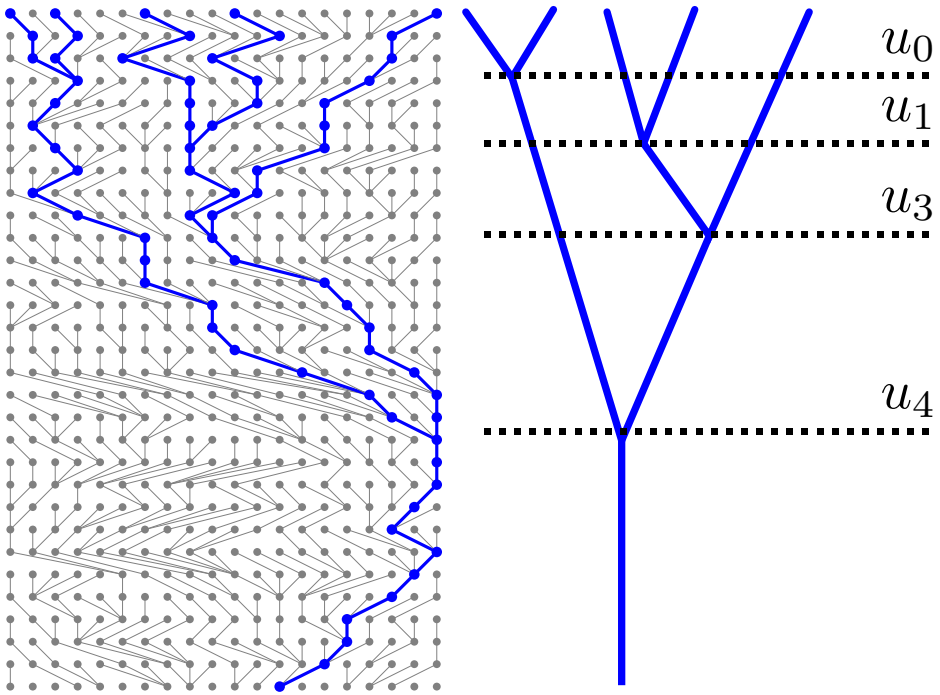
Second analogy

Samples larger than two



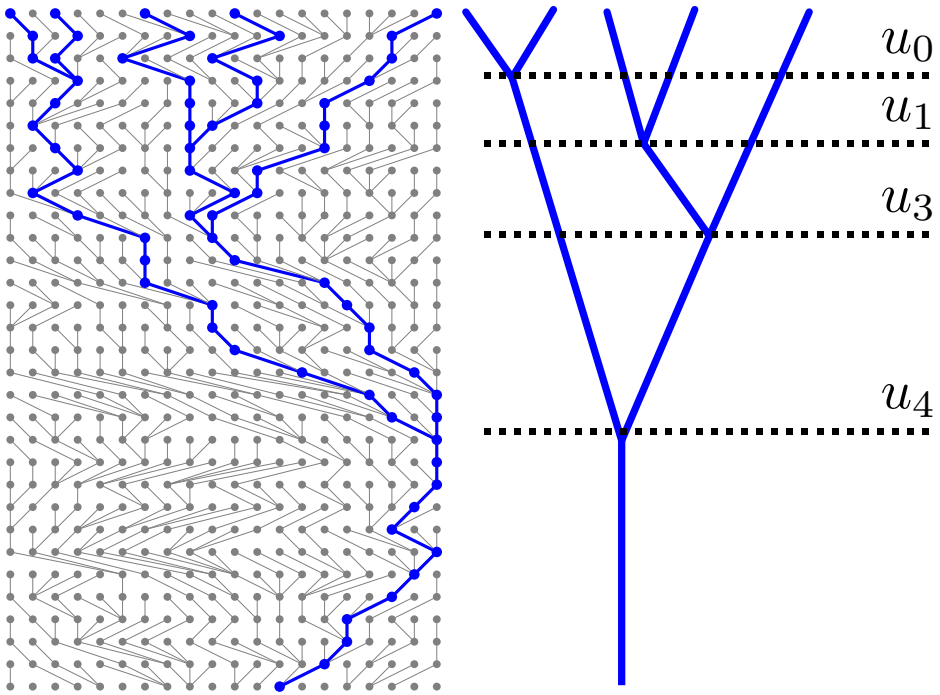


Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample of size k and the total population size N.

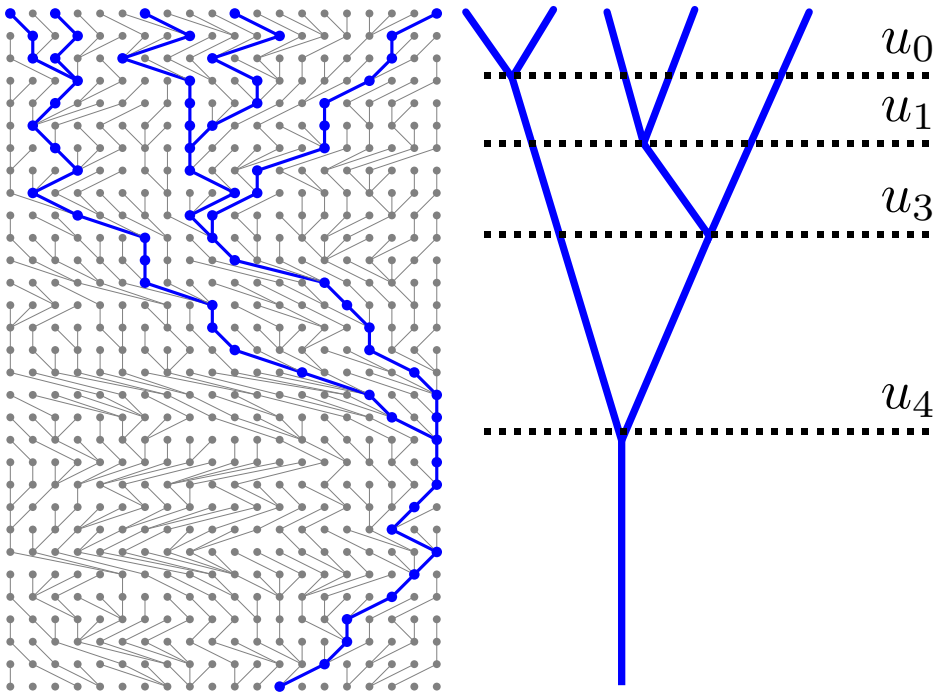
Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample of size k and the total population size N.

Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample of size k and the total population size N.

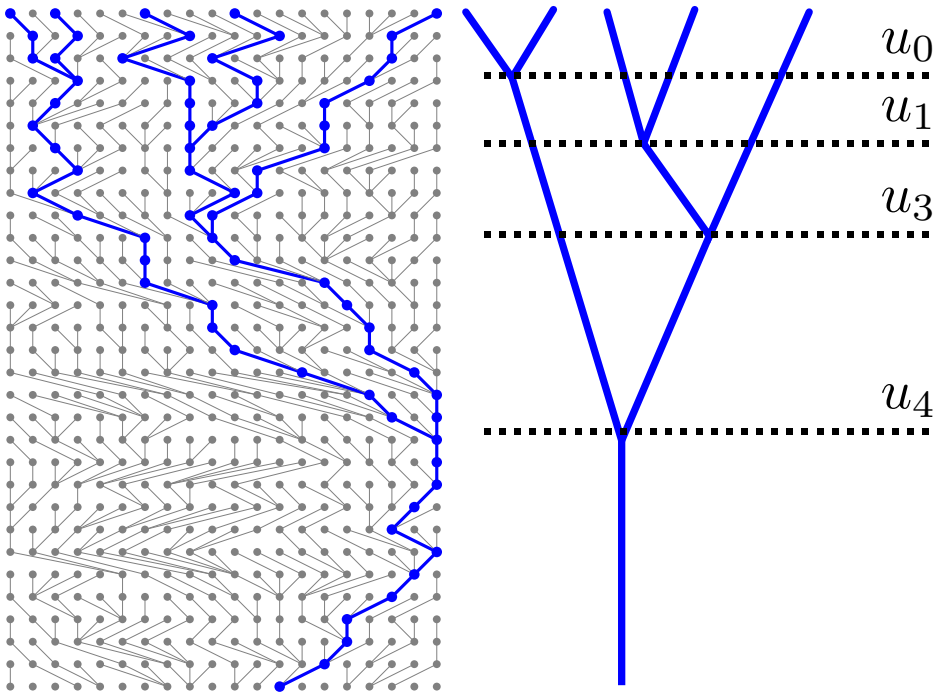
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with an exponential distribution:

$$\mathrm{P}(u_j|N) = e^{-u_j\lambda}\lambda$$

with the scaled coalescence rate

$$\lambda = \binom{k}{2}\frac{1}{2N} \times \mathrm{Prob}(\text{others do not coalesce})$$

Samples larger than two



Looking backward in time, the first coalescence between two random individuals is the result of a waiting process that depends on the sample of size k and the total population size N.

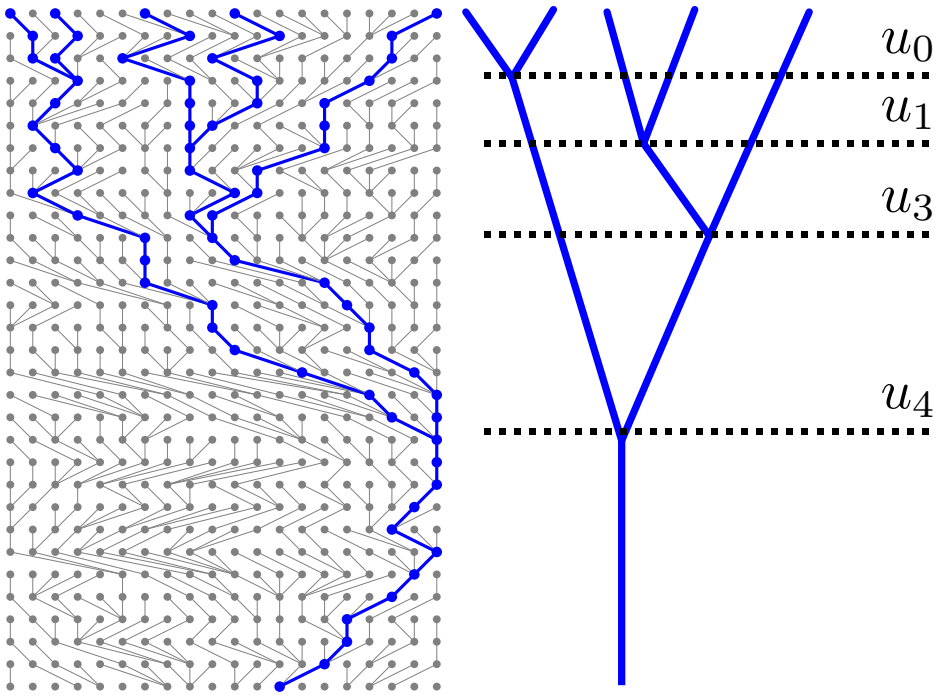
Using Kingman's coalescence rate and imposing a time scale we can approximate the process with a exponential distribution:

$$\mathrm{P}(u_j|N) = e^{-u_j\lambda}\lambda$$

with the scaled coalescence rate

$$\lambda = \binom{k}{2}\frac{1}{2N} = \frac{k(k-1)}{2(2N)} = \frac{k(k-1)}{4N}$$

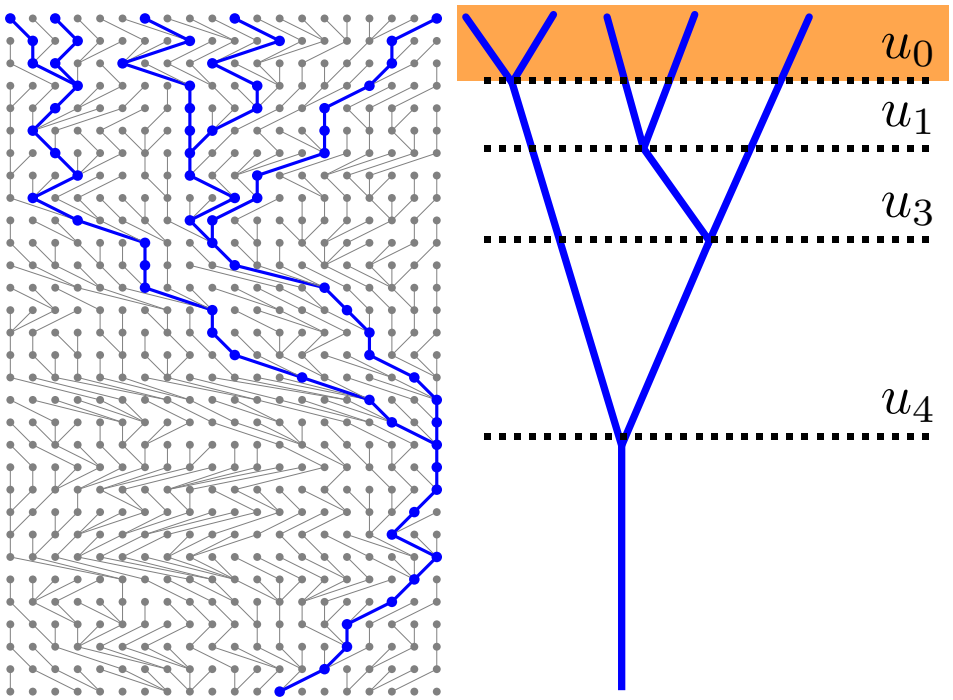
Samples larger than two



We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\mathrm{P}(G|N)$$

Samples larger than two

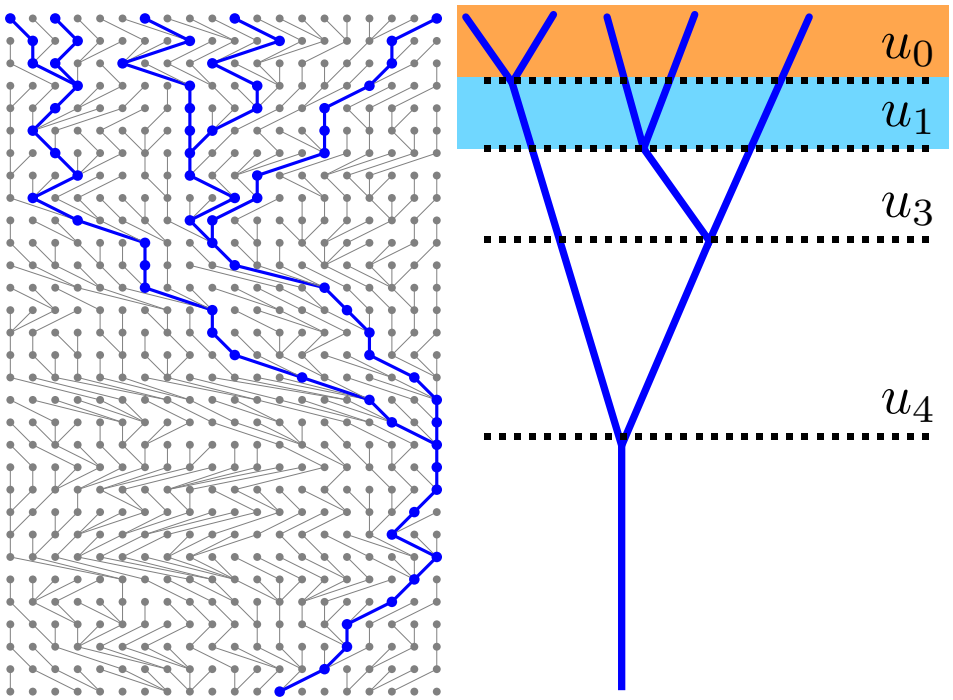


We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\mathrm{P}(G|N) = \quad \mathrm{P}(u_0|N, i_1, i_2)$$

$$\times$$

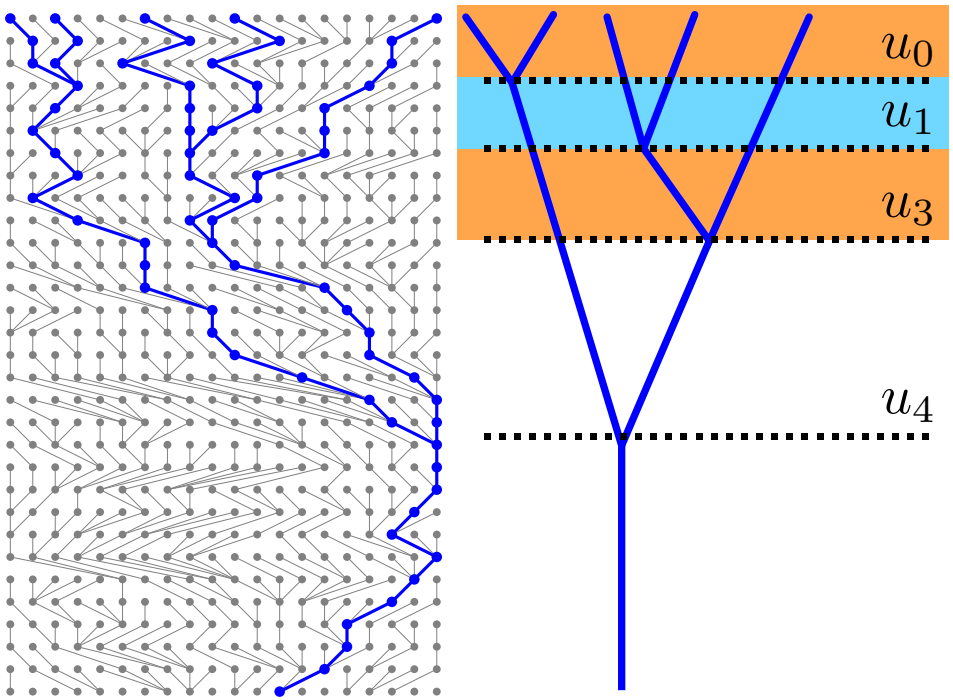
Samples larger than two



We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\mathrm{P}(G|N) = \mathrm{P}(u_0|N, i_1, i_2) \times \mathrm{P}(u_1|N, i_3, i_4)$$

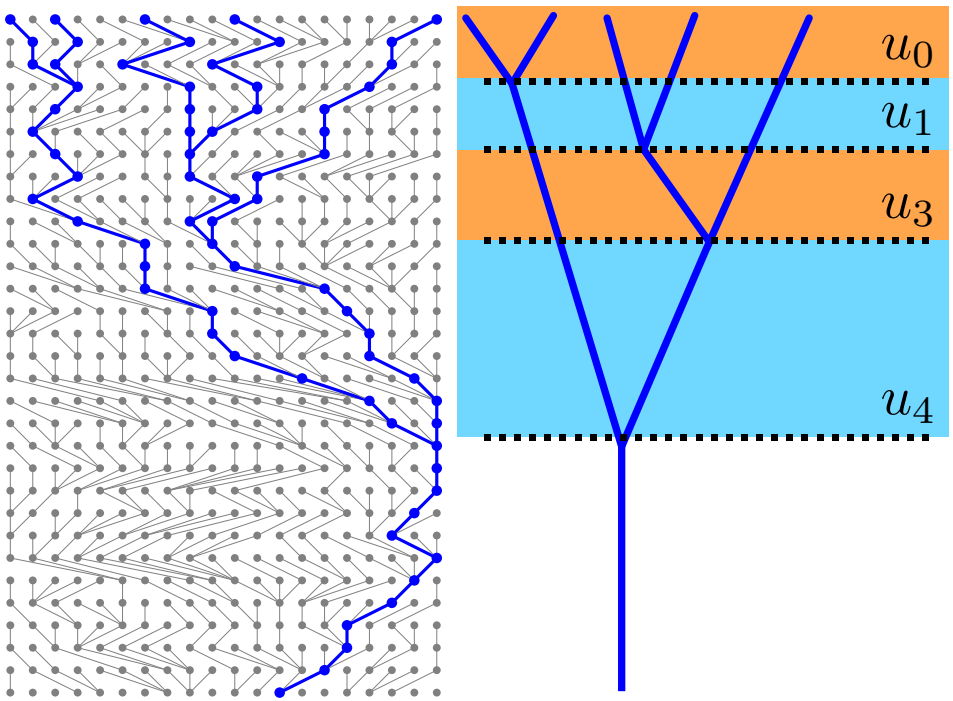
Samples larger than two



We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\begin{aligned}\mathrm{P}(G|N) = \quad & \mathrm{P}(u_0|N, i_1, i_2) \\ & \times \mathrm{P}(u_1|N, i_3, i_4) \\ & \times \mathrm{P}(u_3|N, i_{3,4}, i_5)\end{aligned}$$

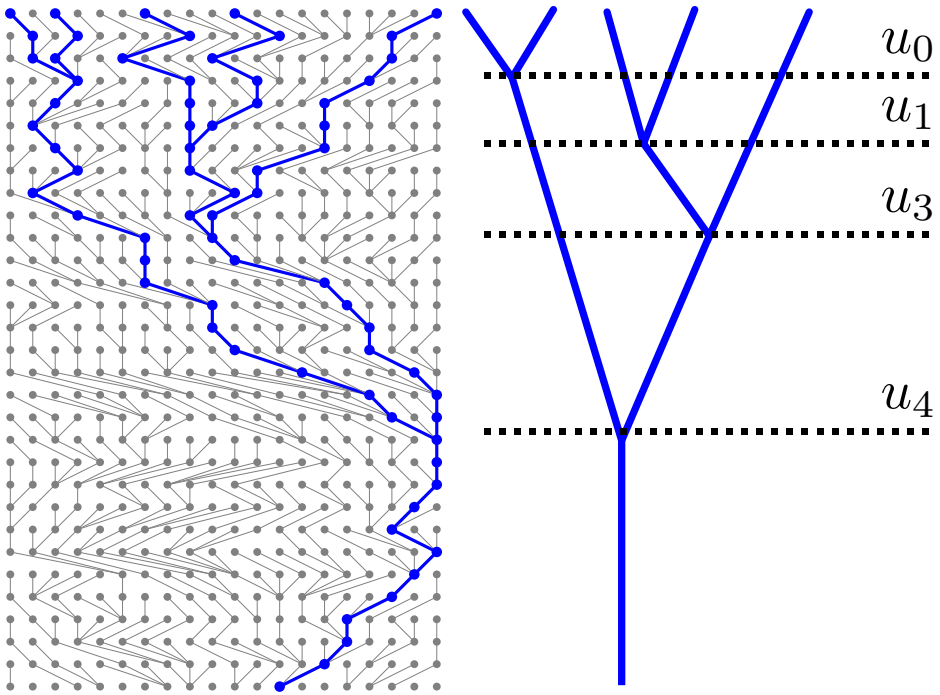
Samples larger than two



We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\begin{aligned}\mathrm{P}(G|N) = \quad & \mathrm{P}(u_0|N, i_1, i_2) \\ & \times \mathrm{P}(u_1|N, i_3, i_4) \\ & \times \mathrm{P}(u_3|N, i_{3,4}, i_5) \\ & \times \mathrm{P}(u_4|N, i_{1,2}, i_{3,4,5})\end{aligned}$$

Samples larger than two



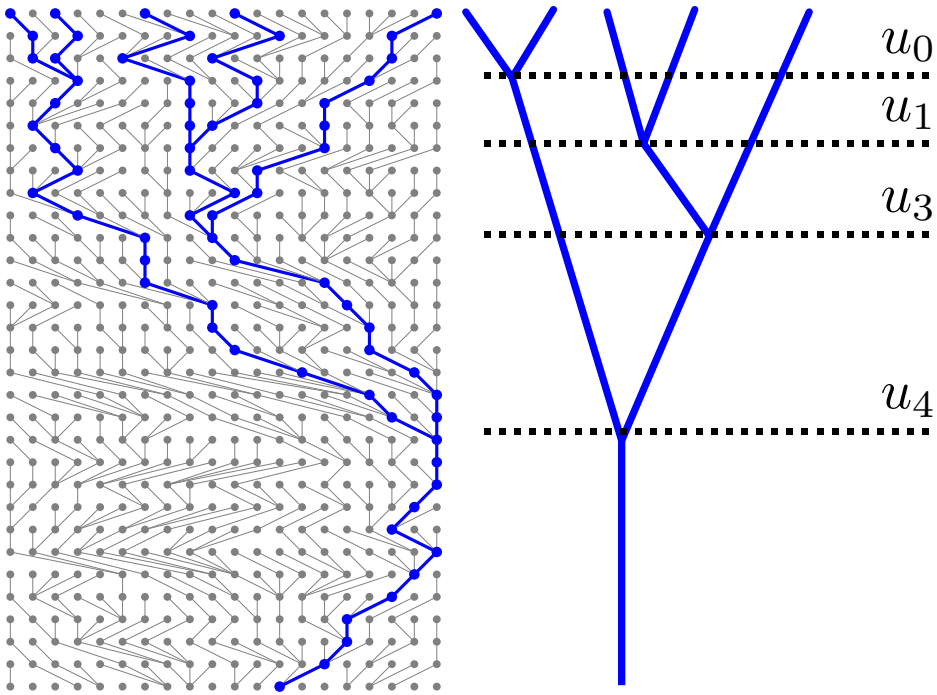
We are now able to calculate the probability of a whole relationship tree (Genealogy G). We assume that each coalescence is independent from any other:

$$\begin{aligned}\mathrm{P}(G|N) = \quad & \mathrm{P}(u_0|N, i_1, i_2) \\ & \times \mathrm{P}(u_1|N, i_3, i_4) \\ & \times \mathrm{P}(u_3|N, i_{3,4}, i_5) \\ & \times \mathrm{P}(u_4|N, i_{1,2}, i_{3,4,5})\end{aligned}$$

$$\mathrm{P}(G|N) = \prod_{j=0}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{k(k-1)}{4N} \frac{2}{k(k-1)} = \prod_{j=0}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{2}{4N}$$

Samples larger than two

$$\mathrm{P}(G|N) = \prod_{j=0}^{T} e^{-u_j \frac{k_j(k_j-1)}{4N}} \frac{2}{4N}$$



The expectations of the total time to coalescence is the sum of the expectations for each interval. Each interval has expectation

$$\mathbb{E}(u) = \frac{4N}{k(k-1)}$$

this leads to the expectation for the time of the most recent common ancestor

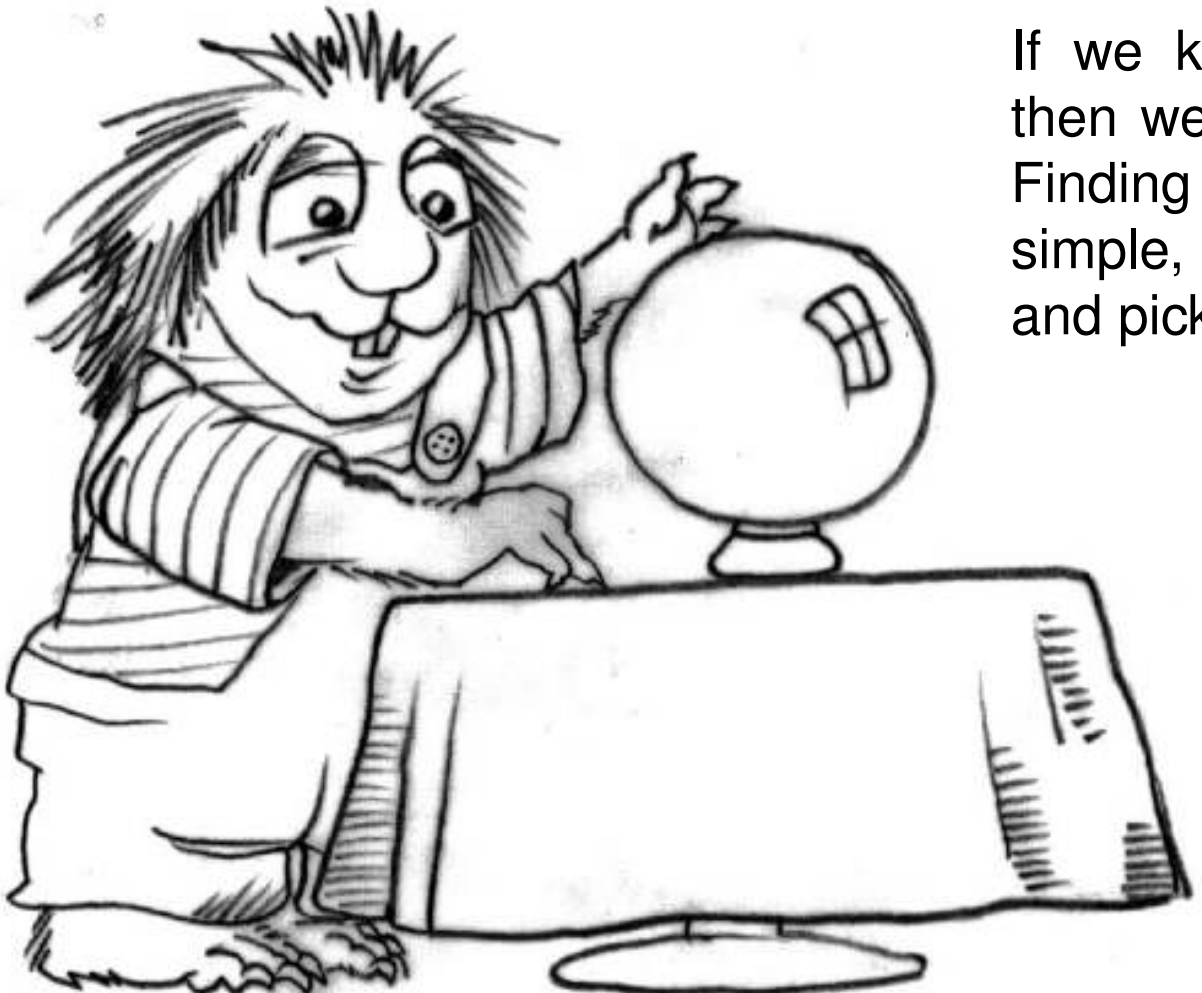
$$\mathbb{E}(\tau_{\mathrm{MRCA}}) = \text{Sum of the expectation of each time interval} = \sum_{j=0}^{J} \frac{4N}{k_j(k_j-1)}$$

$$\lim_{k\to\infty} \mathbb{E}(\tau_{\mathrm{MRCA}}) = 2N + \frac{2}{3}N + \frac{1}{3}N + \frac{1}{5}N + \frac{2}{15}N + \ldots = 4N \qquad \lim_{k\to\infty} \sigma(\tau_{\mathrm{MRCA}}) = 4N$$

What is it good for?

If we know the genealogy G with certainty then we can calculate the population size N. Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.

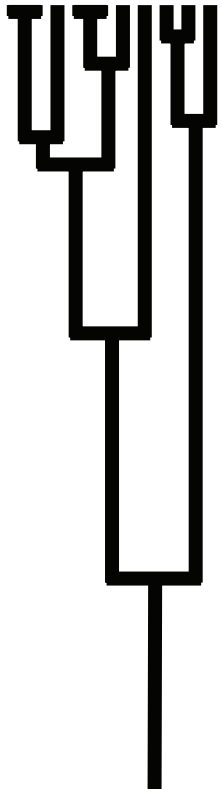
What is it good for?



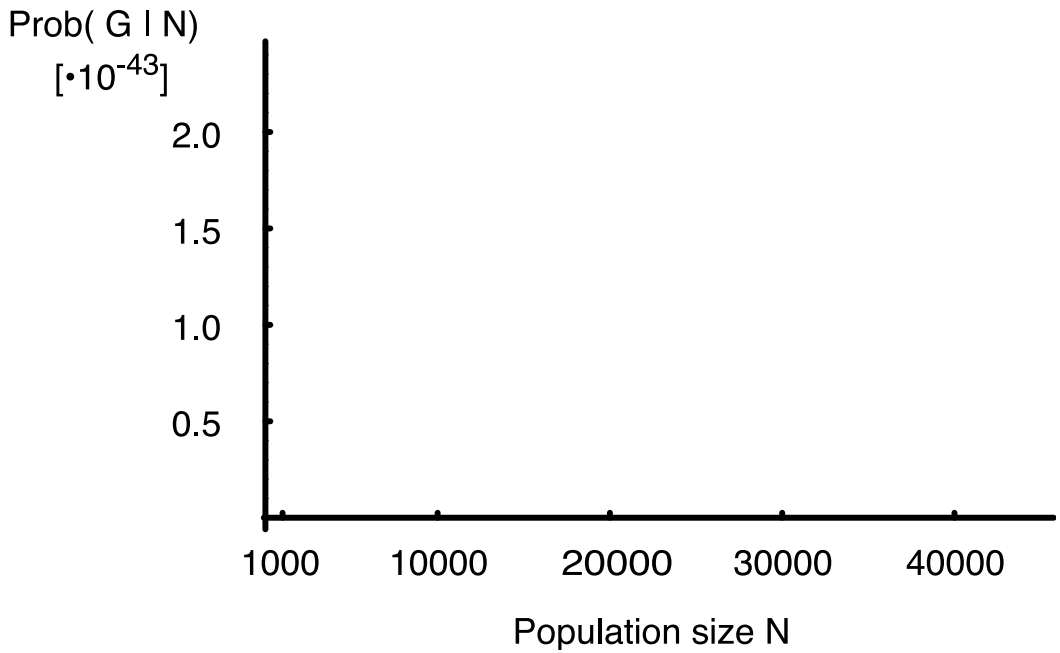
If we know the genealogy G with certainty then we can calculate the population size N. Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.

What is it good for?

If we know the genealogy G with certainty then we can calculate the population size N. Finding the maximum probability $\mathrm{P}(G|N,k)$ is simple, we evaluate all possible values for N and pick the value with the highest probability.



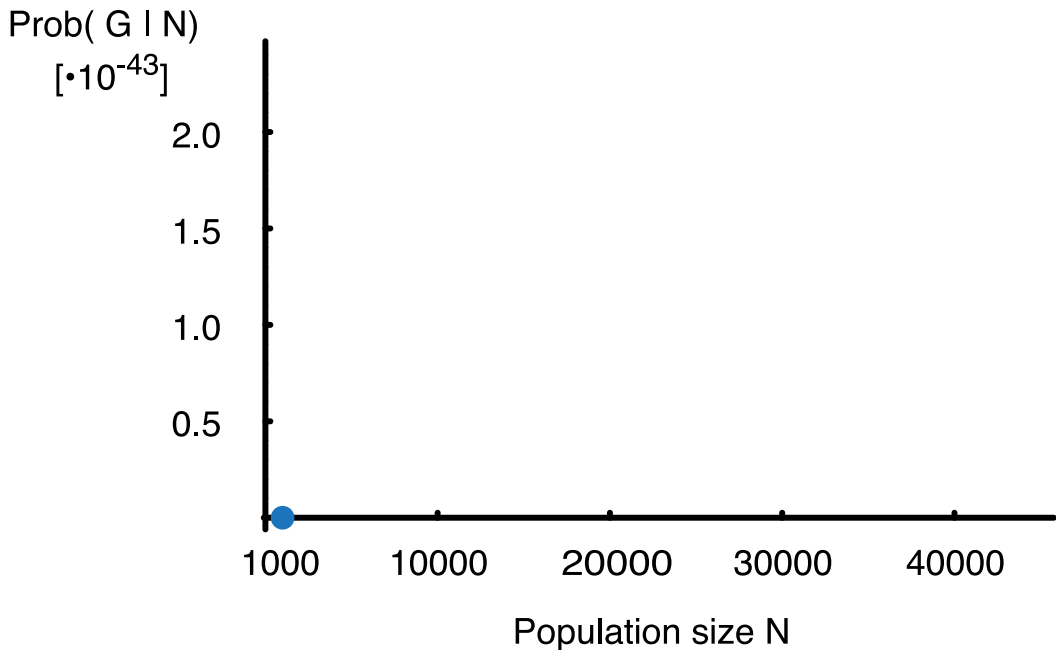
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

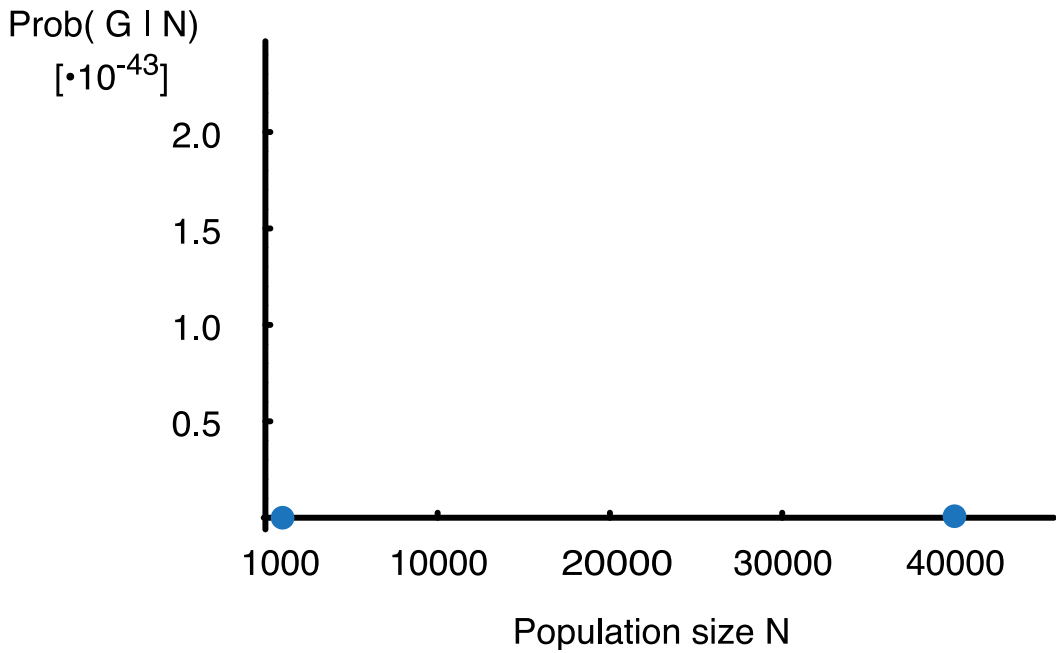
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

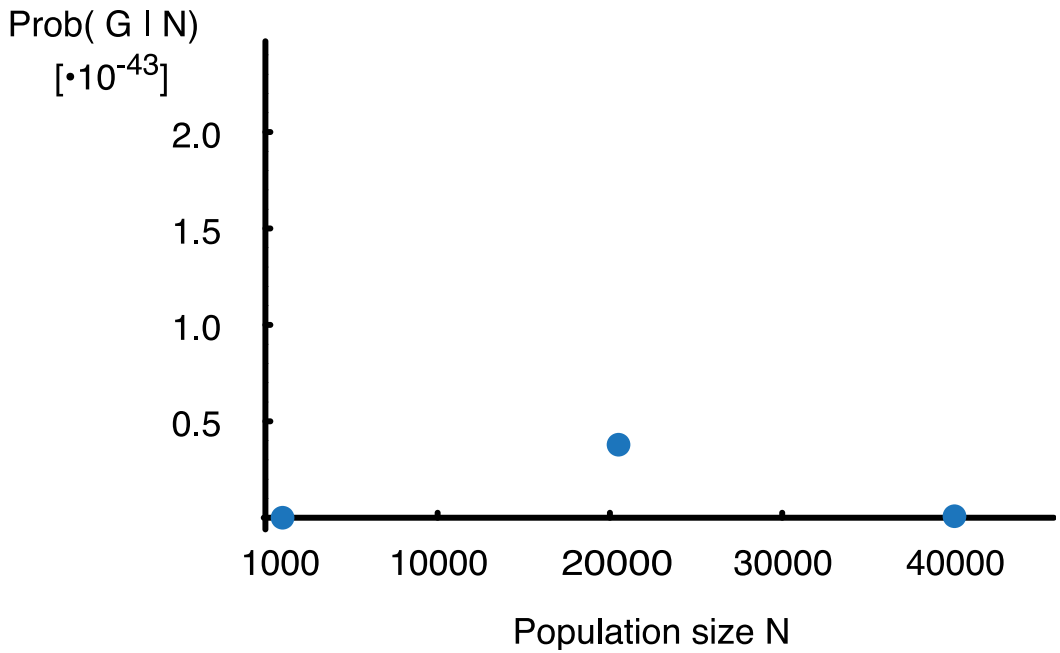
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

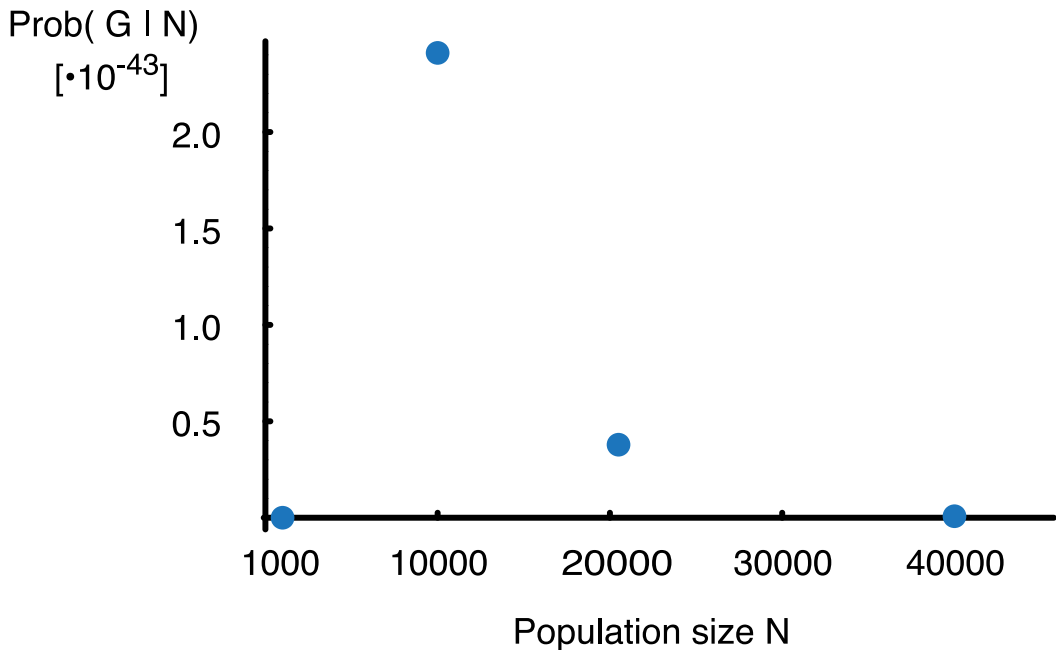
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

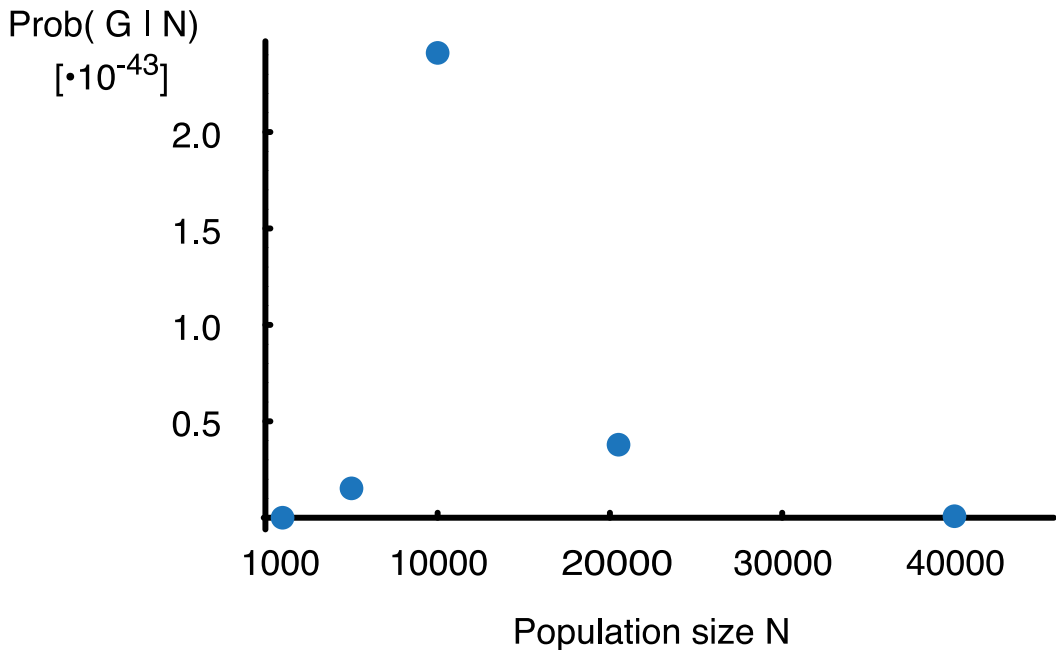
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

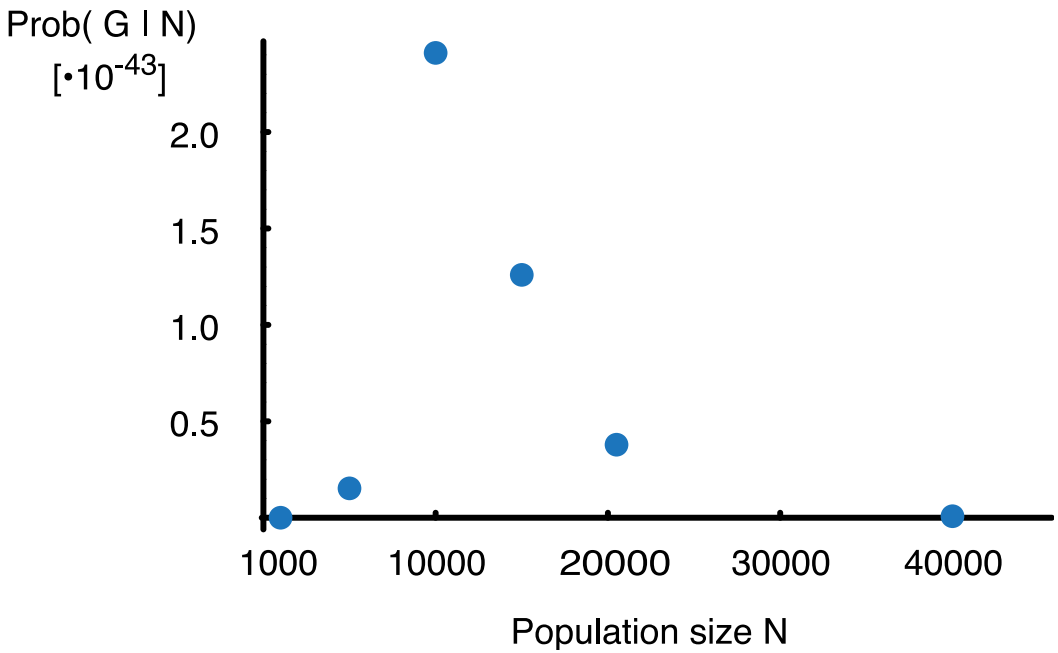
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

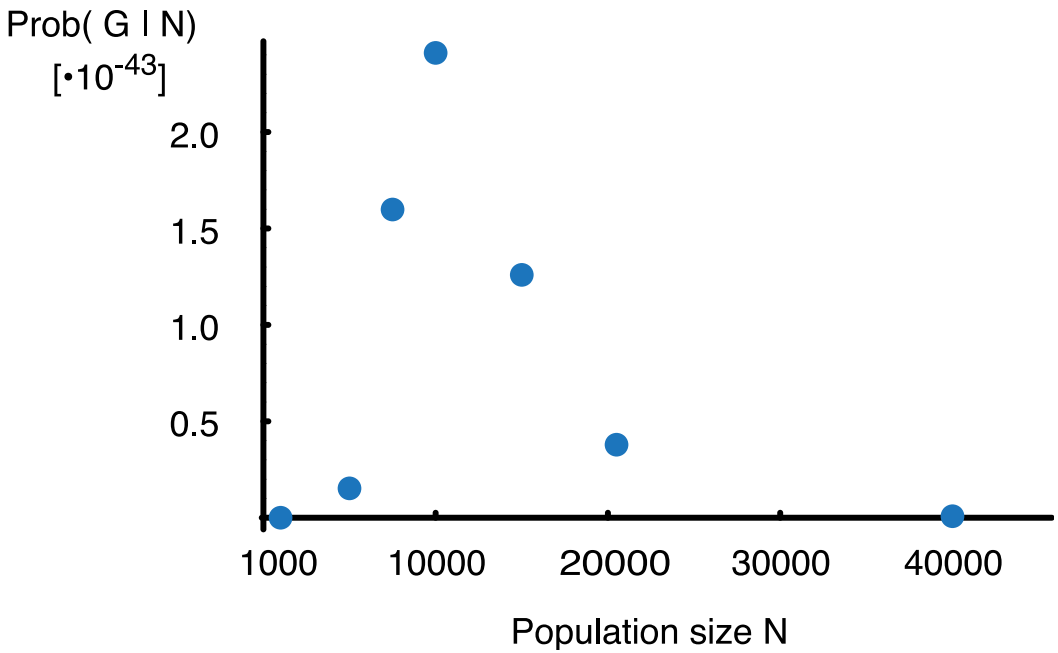
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

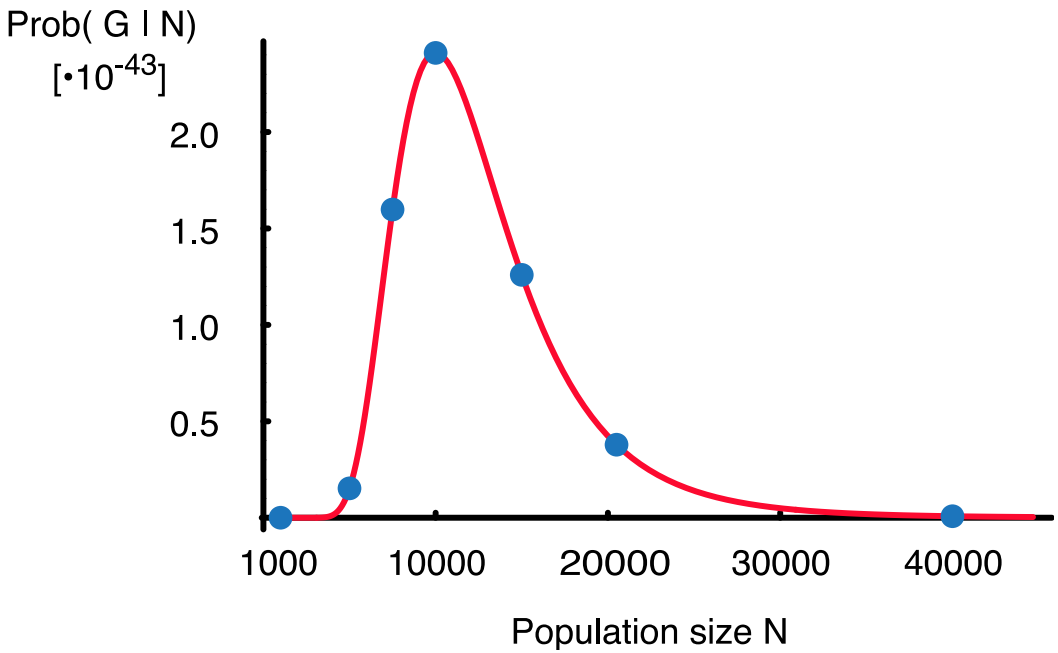
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

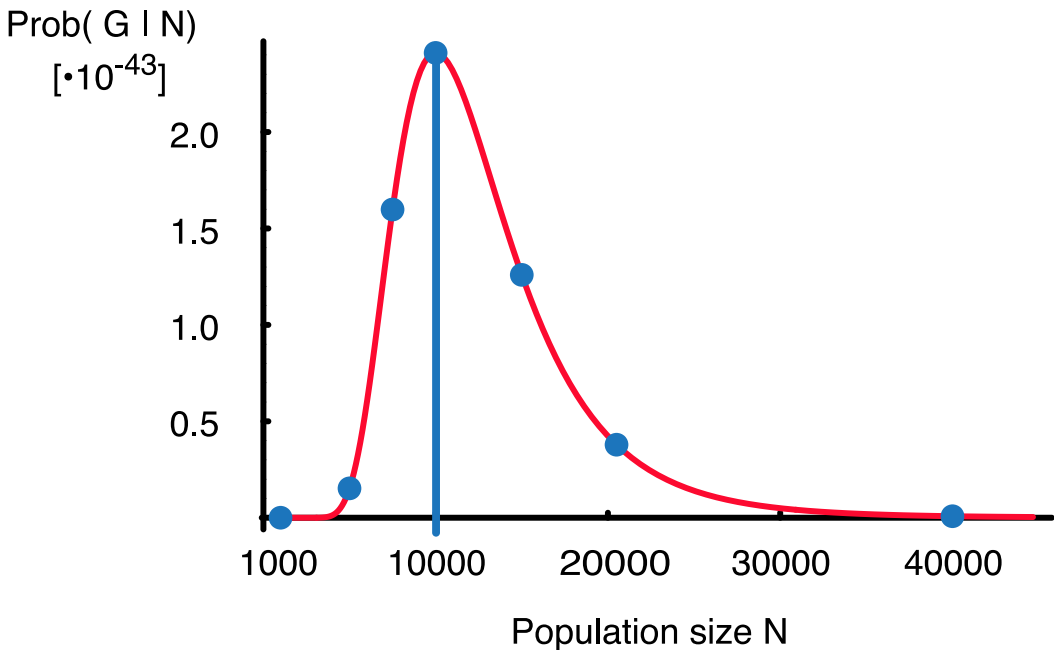
Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

Population size estimation



If an oracle gives us the true relationship tree G then we can calculate the population size N.

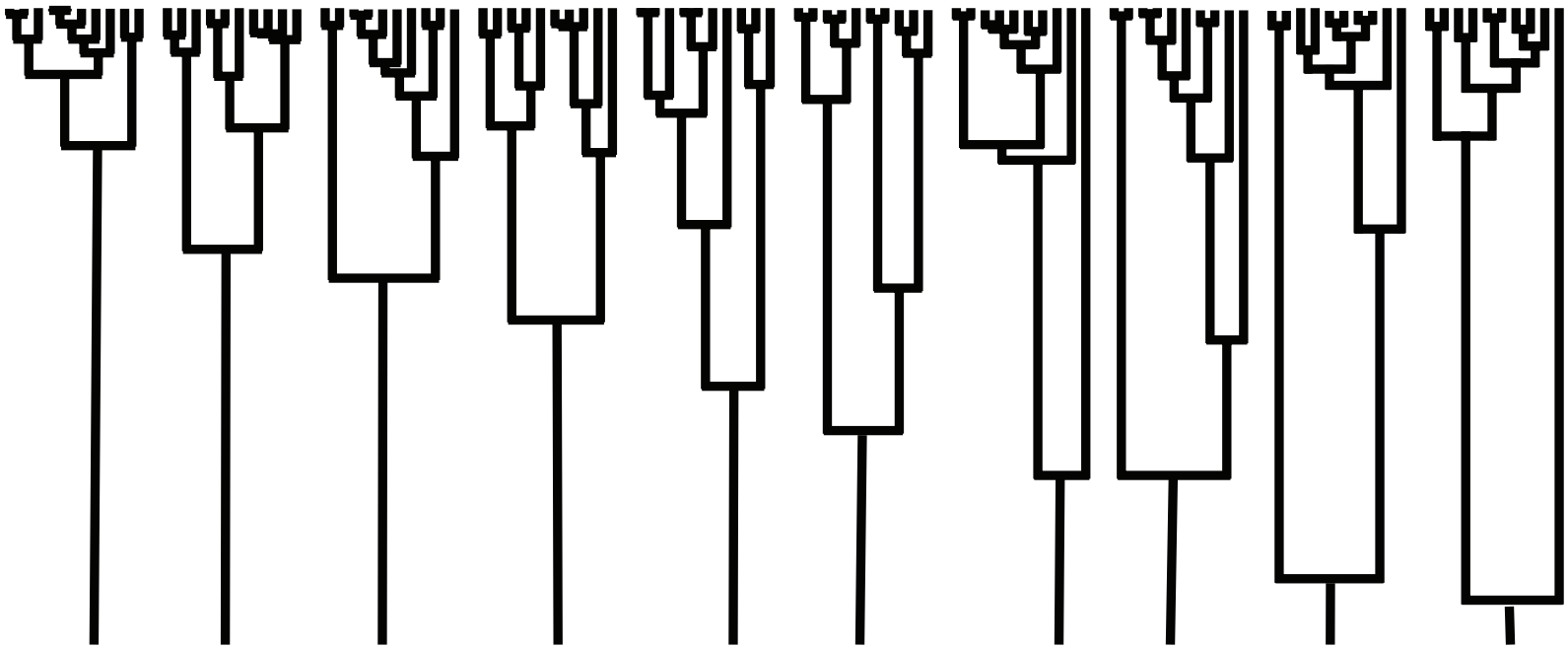
$$\mathrm{p(G|N,n)} = \prod_{\mathrm{k=2}}^{\mathrm{n}} \exp\left(-\mathrm{u_k}\frac{\mathrm{k(k-1)}}{\mathrm{4N}}\right)\frac{2}{\mathrm{4N}}$$

Population size estimation

There are at least two problems with the oracle-approach:

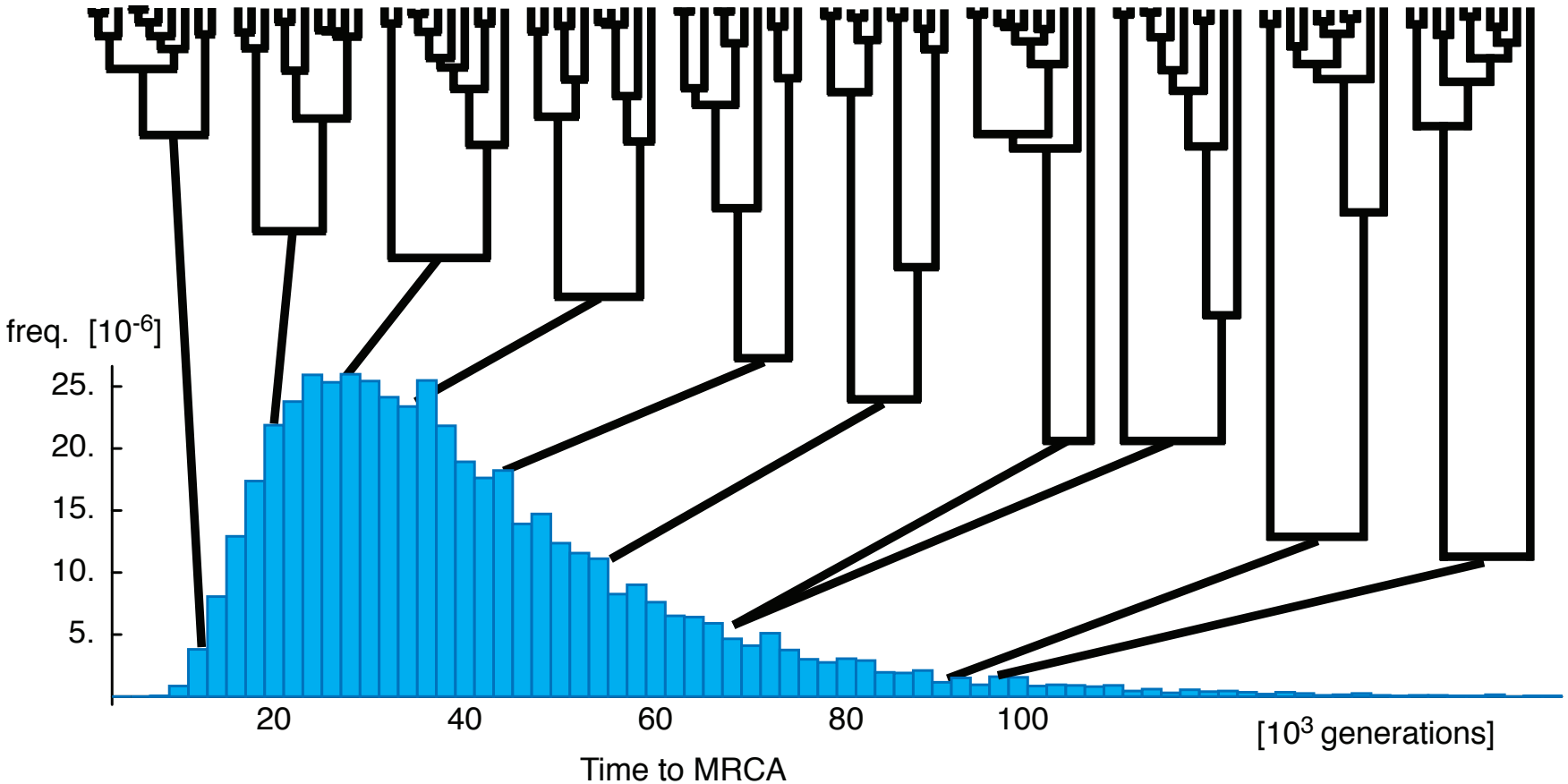
- There is no oracle to gives us clear information!
- We do not record genealogies, our data are sequences, microsatellite loci!
- What about the variability of the coalescent?

Variability of the coalescent process



All genealogies were simulated with the same population size $N_e = 10,000$

Variability of the coalescent process



MRCA = most recent common ancestor (last node in the genealogy)

Kingman's n-coalescent is an approximation

- All individuals have the same fitness (no selection).
- All individuals have the same chance to be in the sample (random sampling).
- The coalescent allows only merging two lineages per generation. This restricts us to to have a much smaller sample size than the population size.

$$n << N$$

- Yun-Xin Fu (2005) described the exact coalescent for the Wright-Fisher model and derived a maximal sample size $n<\sqrt{4N}$ for a diploid population. Although this may look like a severe restriction for the use of the coalescence in small populations, it turned out that the coalescence is rather robust and that even sample sizes close to the effective population size are not biasing immensely.

Ignoring multimerger coalescences

Here are the exact probabilities of 0, 1, or more coalescences with 10 lineages in diploid populations of different sizes:

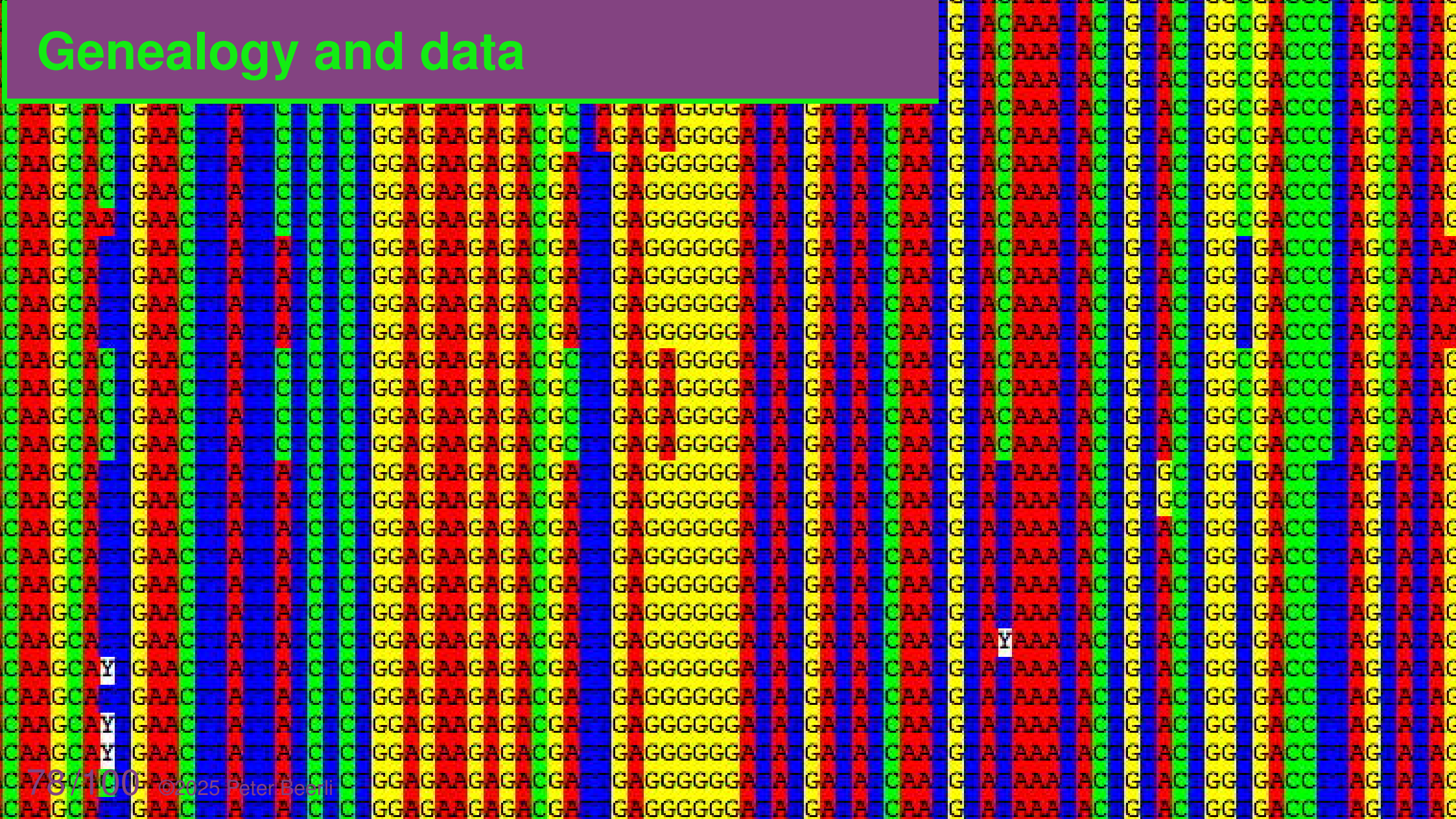
Population size	Coalescences within a single generation		
	0	1	>1
100	0.79560747	0.18744678	0.01694575
1000	0.97771632	0.02209806	0.00018562
10000	0.99775217	0.00224595	0.00000187

Note that increasing the population size by a factor of 10 reduces the coalescent rate for pairs by about 10-fold, but reduces the rate for triples (or more) by about 100-fold.

Observations

- Large samples coalesce on average in $4N$ generations.
- The time to the most recent common ancestor (TMRCA) has a large variance
- Even a sample with few individuals can most often recover the same TMRCA as a large sample.
- The sample size should be much smaller than the population size, although severe problems appear only with sample sizes of the same magnitude as the population size, or with non-random samples because Kingman's coalescence process assumes that maximally two sample lineages coalesce in any generation.
- With a known genealogy we can estimate the population size. Unfortunately, the true genealogy of a sample is rarely known.

Genealogy and data



Genealogy and data

Finding the best genealogy from such data is difficult

Genetic data and the coalescent

- Finite populations loose alleles due to genetic drift
- Mutation introduces new alleles into a population at rate μ
- With $2N$ chromosomes we can expect to see every generation $2N\mu$ new mutations. The population size N is positively correlated with the mutation rate μ.
- With genetic data sampled from several individuals we can use the mutational variability to estimate the population size.

Population size

The observed genetic variability

$$\mathcal{S} = f(N, \mu, n).$$

Different N and appropriate μ can give the same number of mutations. For example, for 100 loci sampled from 20 individuals with 1000bp each, we get :

N	μ	$4N\mu$	$\hat{S}$	σ_S^2
1250	10^{-5}	0.05	153.95	16.25
12500	10^{-6}	0.05	152.89	16.05

Using genetic variability alone therefore does not allow to disentangle N and μ. With multiple dated samples and known generation time we can estimate N and μ independently.

Mutation-scaled population size

By convention we express most results as the compound $N\mu$ and an inheritance scalar x, for simplicity we call this the mutation-scaled population size

$$\Theta = xN\mu,$$

where μ is the mutation rate per generation and per site. With a mutation rate per locus we use θ.

- for diploids: $\Theta = 4N\mu$.
- for haploids: $\Theta = 2N\mu$.

Time scale (second)

$$P(\text{no coalescence with } n \text{ lineages}) = \exp\left(-\tau\binom{k}{2}\right)$$

scaling time τ by the population size $2N$ and using t in generations we get $\tau = t\frac{1}{2N}$, this then leads to

$$P(\text{coalescence at } t_0 + t) = \exp(-t\lambda)\lambda \quad \text{with} \quad \lambda = \frac{1}{2N}\frac{k(k-1)}{2}$$

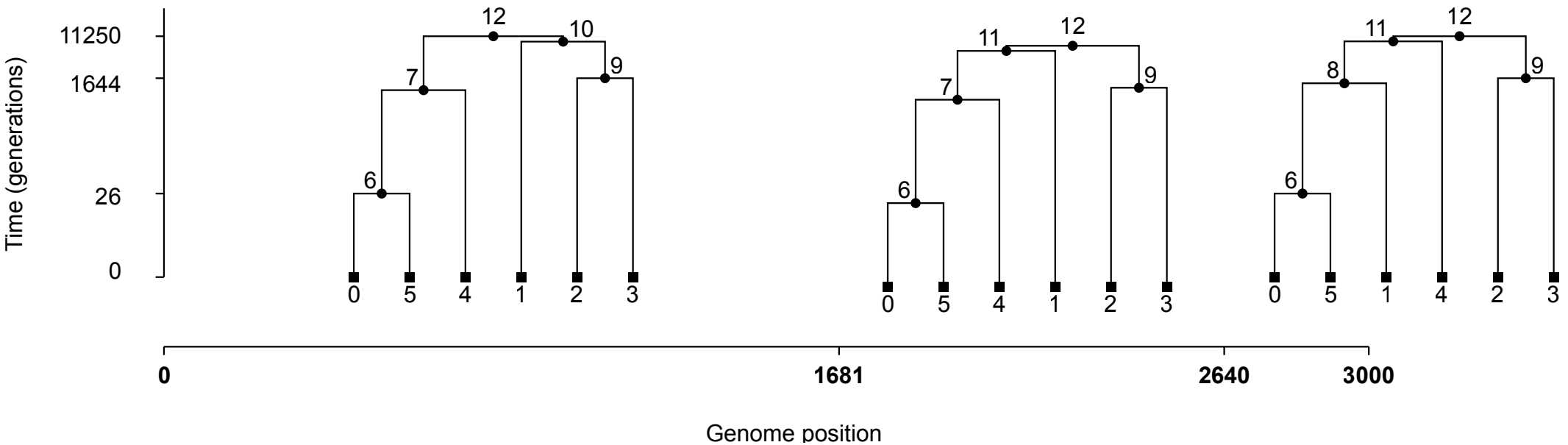
When we use DNA data, we assume there is a mutation model with a mutation rate μ; we include that in our scaling and use time t as scaled by expected mutation rate per generation and $\Theta = 4N\mu$:

$$P(\text{a coalescent at time } t_0 + t|t_0) = \frac{2}{\Theta}\exp(-t\frac{k(k-1)}{\Theta})$$

What is an individual for the coalescent

- Each site has a single coalescent tree
- Loci in close proximity share most likely the same coalescent tree, but dependent on the magnitude of the recombination rate, loci may be linked or unlinked.
- We assume that loci on different chromosomes are independent.
- Usually, we treat each locus as independent and combine the single locus estimates.

Each site can have different coalescents: Recombination



How to make inference using genetic data

Sequence data has some sites in some individuals that are different than others	$\Longrightarrow$	Mutation model (finite vs infinite)
Population model	$\Longrightarrow$	the Coalescent
Analysis method	$\Longrightarrow$	Summary statistics Maximum likelihood Bayesian Inference

How to make inference using genetic data

Sequence data has some sites in some individuals that are different than others	$\Longrightarrow$	Mutation model (finite vs infinite)
Population model	$\Longrightarrow$	the Coalescent
Analysis method	$\Longrightarrow$	Summary statistics Maximum likelihood Bayesian Inference

Genetic data and the coalescent

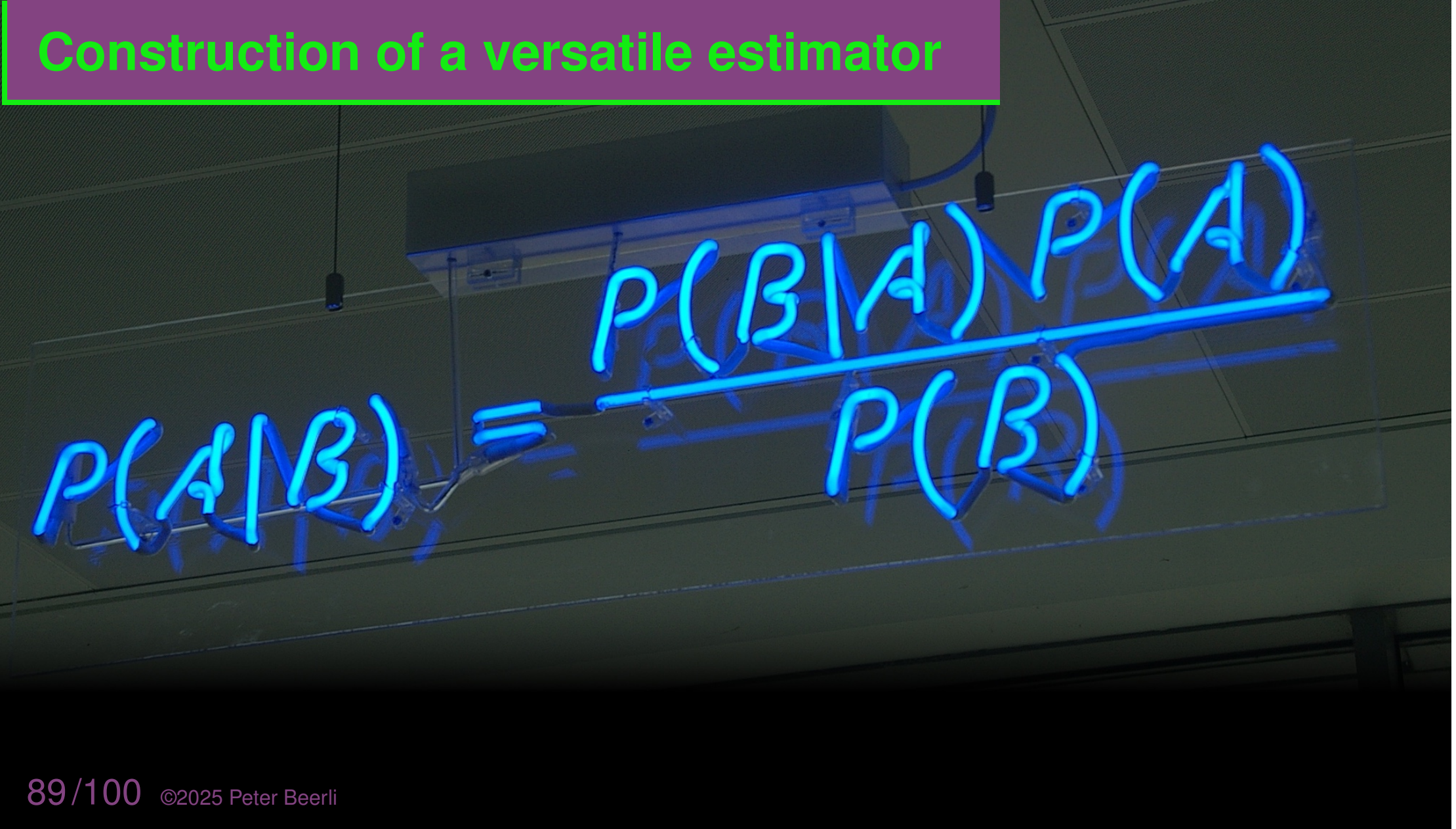
Using the infinite sites model we use the number of variable sites S per locus to calculate the mutation-scaled population size:

$$\theta_W = \frac{S}{\sum_{k=1}^{n-1} \frac{1}{k}}$$

from a sample of n individuals. For a single population the Watterson's estimator works marvelously well, but it is vulnerable to population structure.

Watterson's θ_W uses a mutation rate per locus! To compare with other work use mutation rate per site.

Construction of a versatile estimator



Construction of a versatile estimator

For Bayesian inference we want to calculate the probability of the model parameters given the data $\mathrm{p}(\mathrm{model}|\mathrm{D})$.

Coalescent	to describe the population genetic processes.
Mutation model	to describe the change of genetic material over time.

Construction of a versatile estimator

We calculate the Posterior distribution $\mathrm{p}(\Theta|\mathrm{D})$ using Bayes' rule

$$\mathrm{p}(\Theta|\mathrm{D}) = \frac{\mathrm{p}(\Theta)\mathrm{p}(\mathrm{D}|\Theta)}{\mathrm{p}(\mathrm{D})}$$

where $\mathrm{p}(\mathrm{D}|\Theta)$ is the likelihood of the parameters.

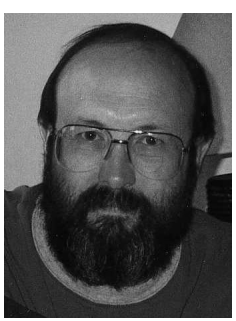
(almost) Felsenstein equation

$$p(D|\mathbf{\Theta}, G) = \mathrm{p(G|\mathbf{\Theta})p(D|G)}$$

$p(G|\mathbf{\Theta})$

The probability density of a genealogy given parameters.

$\mathrm{p(D|G)}$



The probability density of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.

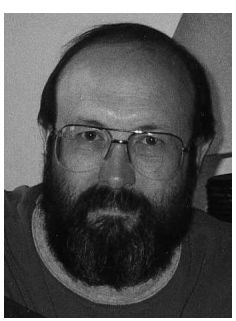
Felsenstein equation

$$p(D|\Theta) = \int_G \mathrm{p(G|\Theta)p(D|G)dG}$$

$p(G|\Theta)$

The probability density of a genealogy given parameters.

$\mathrm{p(D|G)}$



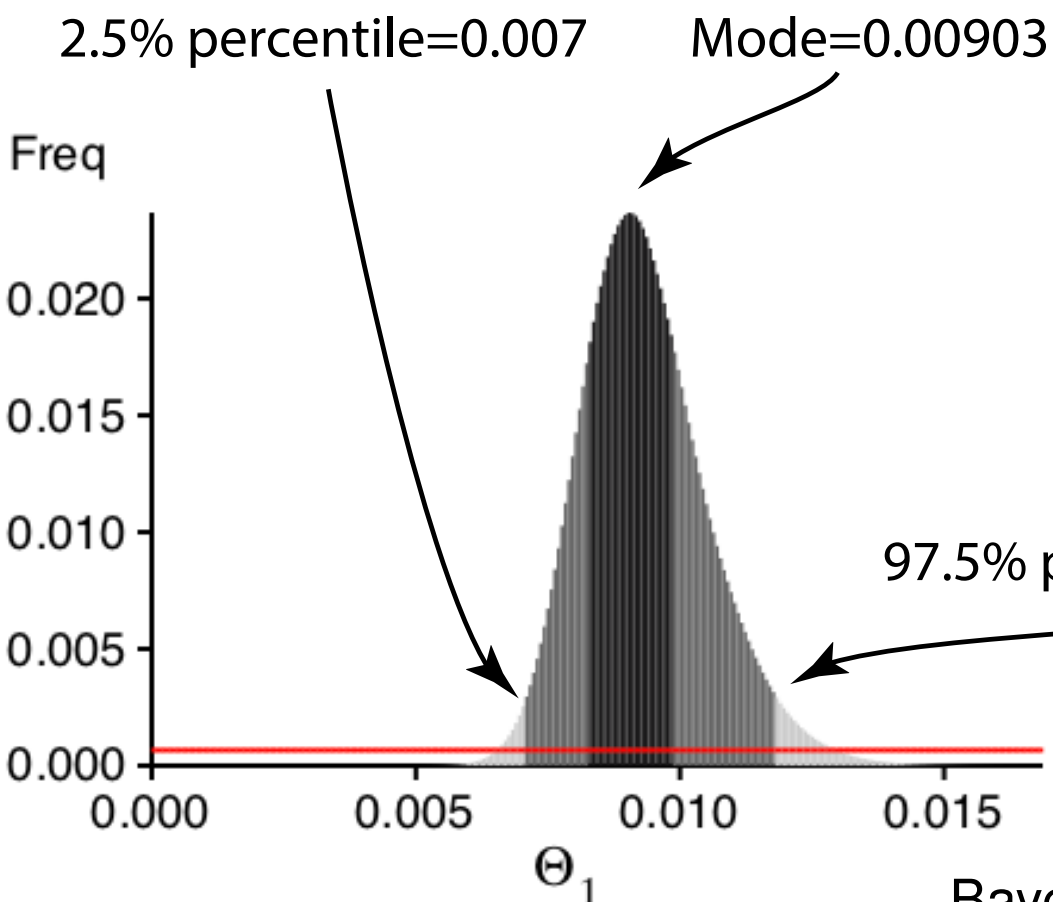
The probability density of the data for a given genealogy. Phylogeneticists know this as the tree-likelihood.

Problem with integration formula

$$p(D|\boldsymbol{\Theta}) = \int_G p(G|\boldsymbol{\Theta}) p(D|G) dG$$

The number of possible genealogies is very large and for realistic data sets, programs need to use Markov chain Monte Carlo methods.

Inference of population size



Median=0.00934
Mean=0.00934

97.5% percentile=0.0118

uniform prior

Bayesian inference: $\Theta = 0.00903$

Watterson Estimator $\Theta_W = 0.01003$

Mutation-scaled population size, revisited

By convention we express most results as the compound $N\mu$ and an inheritance scalar x, for simplicity we call this the mutation-scaled population size

$$\Theta = xN\mu,$$

where μ is the mutation rate per generation and per site. With a mutation rate per locus we use θ.

- for diploids: $\Theta = 4N\mu$.
- for haploids: $\Theta = 2N\mu$.
- For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta = 2N_f\mu$, and if the sex ratio is $1:1$ then $\Theta = N\mu$

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript N and call it the effective population size N_e, and consider Θ the mutation-scaled EFFECTIVE population size.

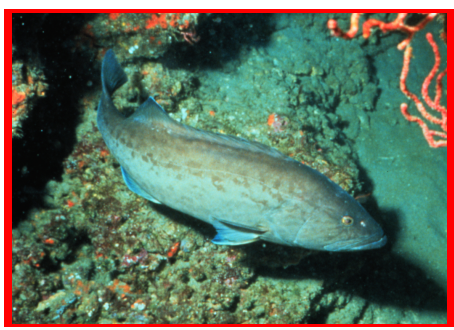
Mutation-scaled population size

By convention we express most results as the compound $N\mu$ and an inheritance scalar x, for simplicity we call this the mutation-scaled population size

$$\Theta = xN\mu,$$

where μ is the mutation rate per generation and per site. With a mutation rate per locus we use θ.

- for diploids: $\Theta = 4N\mu$.
- for haploids: $\Theta = 2N\mu$.



Gag Grouper starts out as a female and later in live becomes male.

- For mtDNA in diploids with strictly maternal inheritance this leads to $\Theta = 2N_f\mu$, and if the sex ratio is $1:1$ then $\Theta = N\mu$

Most real populations do not behave exactly like Wright-Fisher populations, therefore we subscript N and call it the effective population size N_e, and consider Θ the mutation-scaled EFFECTIVE population size.

Historical humpback whale population size

Humpback whales in the North Atlantic: Census population size around 12,000.

Historical humpback whale population size

using the data by Joe Roman and Stephen R. Palumbi (Science 2003 301: 508-510)

$\Theta = 2N_{♀}\mu$	0.01529	Population size of the North Atlantic population, estimated using migrate
$N_{♀} = \frac{\Theta}{2\mu}$	12,251	with $\mu = 5.2 \times 10^{-8}\mathrm{bp}^{-1}\mathrm{year}^{-1}$ and a generation time of 12 years
$N_e = N_{♀} + N_{♂}$	24503	Sex ratio is 1:1
$N_B = 2N_e$	49,006	ratio N_B/N_e assumed, using other data
$N_T = N_B \frac{N_{\text{juveniles}} + N_{\text{adults}}}{N_{\text{adults}}}$	78,410	from catch and survey data (used a ratio of 1.6)

using a mutation rate of Alter and Palumbi 2009; for nucDNA: $112,000 (45,000 - 235,000)$ (Ruegg et al. Conservation Genetics (2013) 14:103–114)