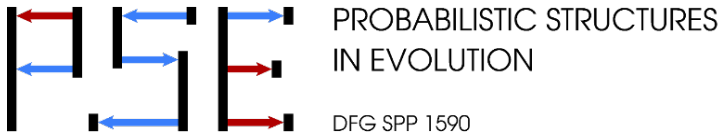
General coalescent models in population genetics

Jochen Blath

Joint work with

Bjarki Eldon, Adrián González Casanova, Noemi Kurt, Maite Wilke Berenguer, Eugenio Buzzoni...

Berlin-Paris Workshop Stochastic Analysis and Applications



Coalescent models and population genetics

Coalescent processes are powerful tools in the analysis of the probabilistic patterns exhibited by genetic samples drawn from a large biological population.

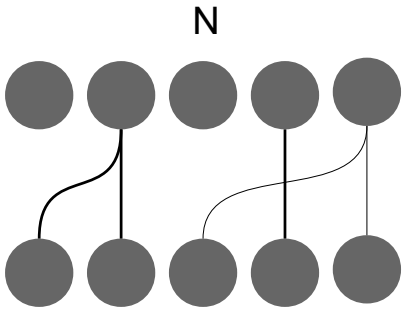
In this talk, we review some well-established coalescent models as well as some new coalescent-related objects that arise in non-classical scenarios and discuss potential applications.

Coalescent models and population genetics

Coalescent processes are powerful tools in the analysis of the probabilistic patterns exhibited by genetic samples drawn from a large biological population.

In this talk, we review some well-established coalescent models as well as some new coalescent-related objects that arise in non-classical scenarios and discuss potential applications.

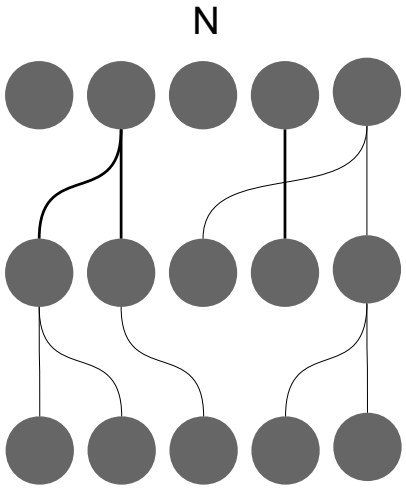
Background: The classical Wright-Fisher model



[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

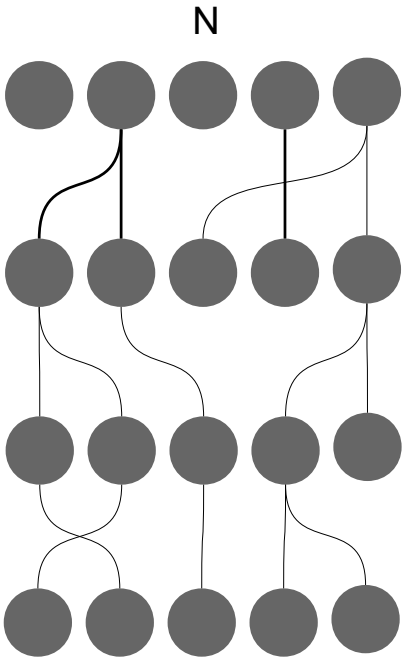
Background: The classical Wright-Fisher model



[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

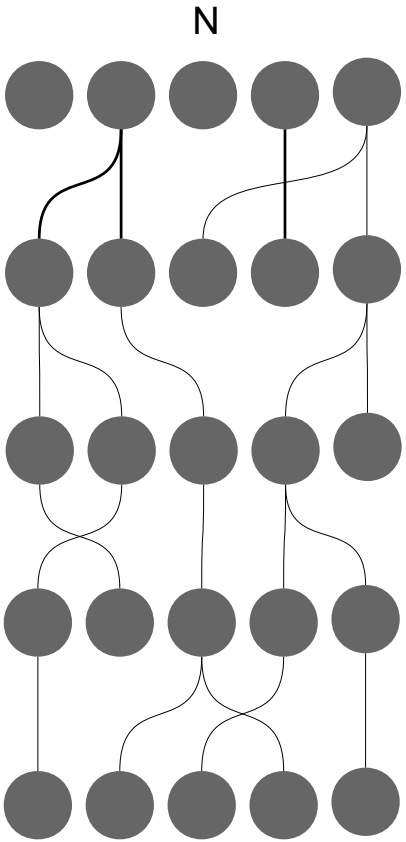
Background: The classical Wright-Fisher model



[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

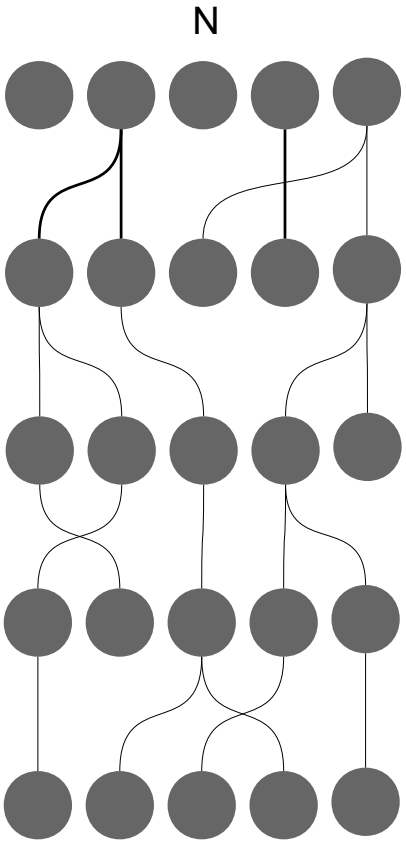
Background: The classical Wright-Fisher model



[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

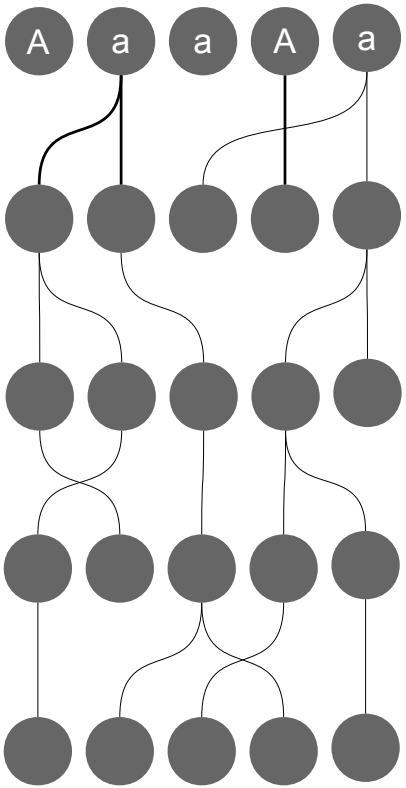
Background: The classical Wright-Fisher model



[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

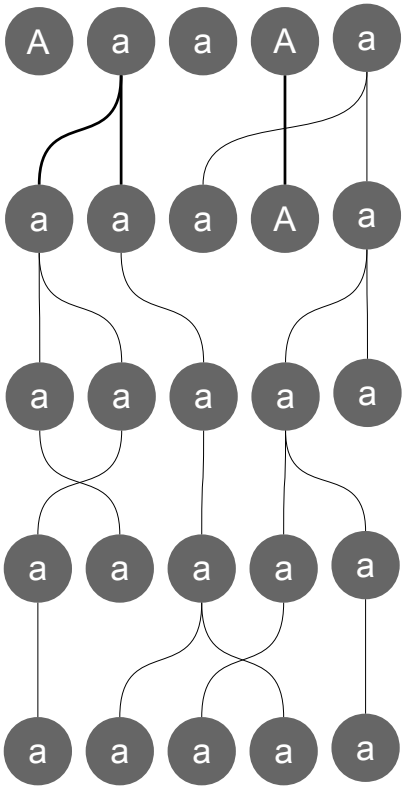
Background: The classical Wright-Fisher model



[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

Background: The classical Wright-Fisher model

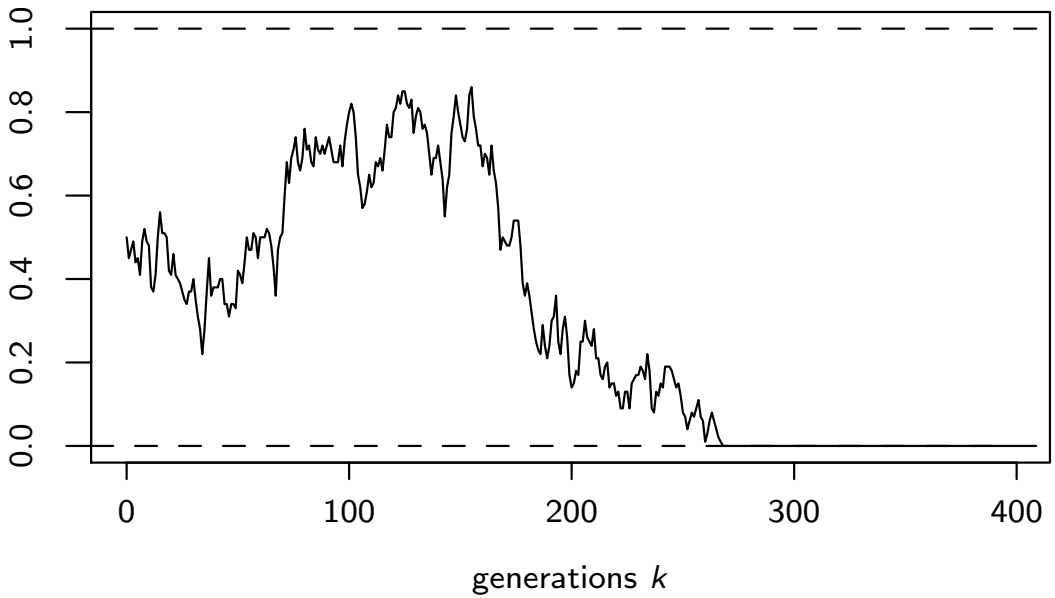


[Wright 1931, Fisher 1930]

- Haploid population of N individuals, reproducing in discrete generations, *fixed* population size.
- Selectively neutral: Each offspring chooses parent *uniformly* at random, copies its genetic type.
- Offspring distribution in each generation is thus symmetric multinomial.
- Simplest case: Two allelic types, $E = \{a, A\}$.

The Wright Fisher diffusion

- Let Z_k^N denote the fraction of type a individuals in generation k. It is a stochastic process; a simulation for $N = 100$:



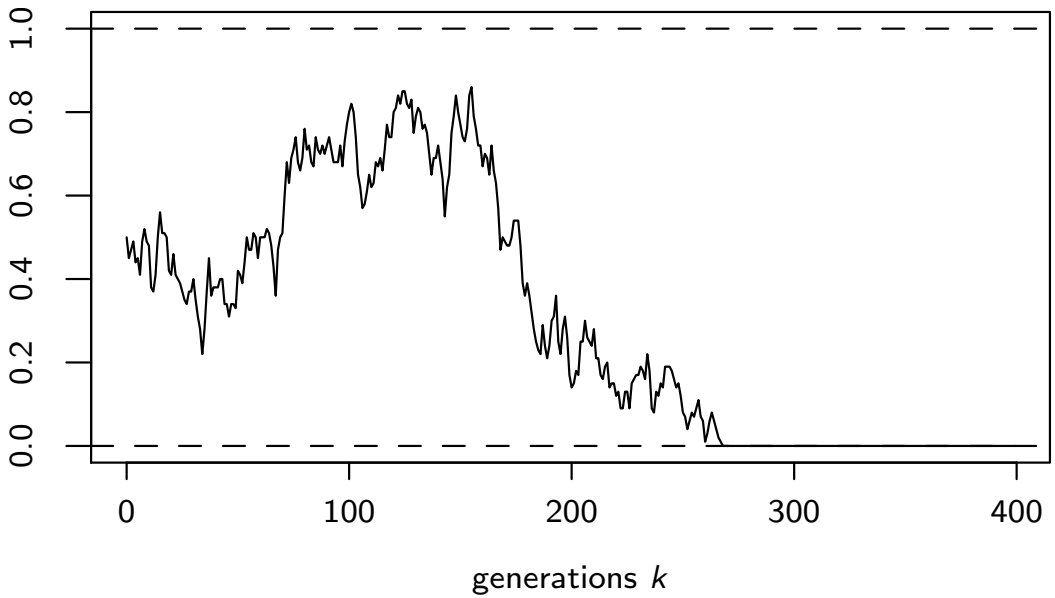
- Final absorption at 0 and 1 due to *random genetic drift* (fixation).
- Fact: Measuring time in units of N generations, $(Z^N_{\lfloor Nt \rfloor})_{t \geq 0}$ converges weakly to $(Z_t)_{t \geq 0}$ as $N \to \infty$, where

$$dZ_t = \sqrt{Z_t(1 - Z_t)}dB_t$$

is the *Wright-Fisher diffusion.*

The Wright Fisher diffusion

- Let Z_k^N denote the fraction of type a individuals in generation k. It is a stochastic process; a simulation for $N = 100$:

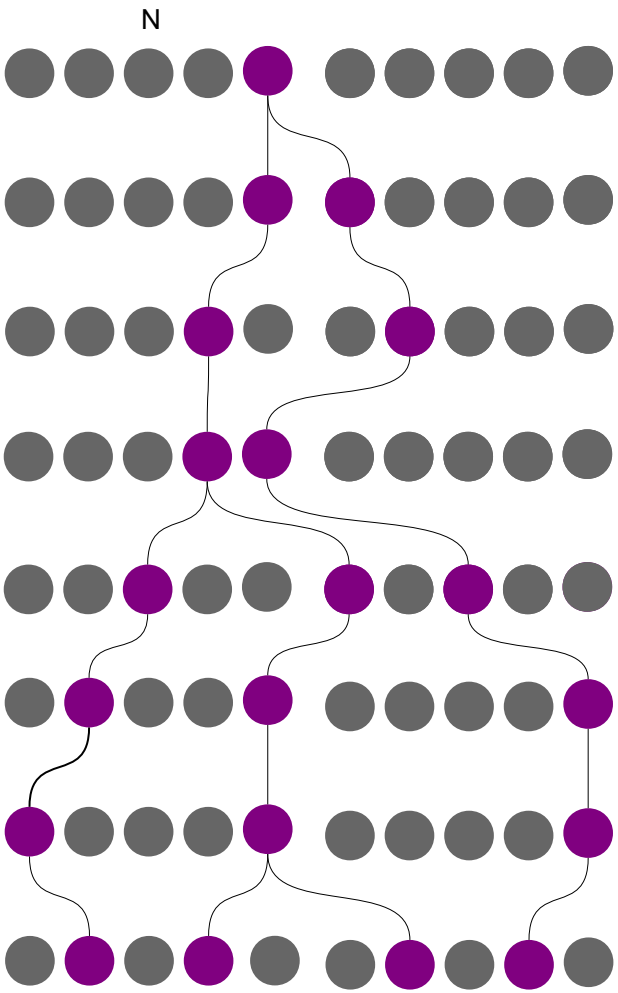


- Final absorption at 0 and 1 due to *random genetic drift* (fixation).
- Fact: Measuring time in units of N generations, $(Z^N_{\lfloor Nt \rfloor})_{t \geq 0}$ converges weakly to $(Z_t)_{t \geq 0}$ as $N \to \infty$, where

$$dZ_t = \sqrt{Z_t(1 - Z_t)}dB_t$$

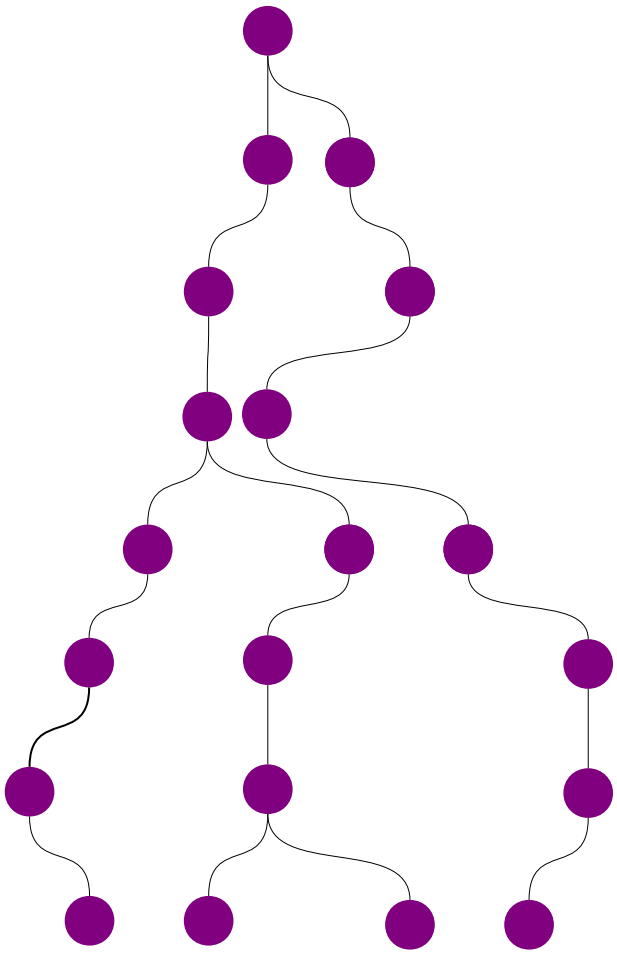
is the *Wright-Fisher diffusion.*

Ancestry of a sample in the Wright-Fisher model



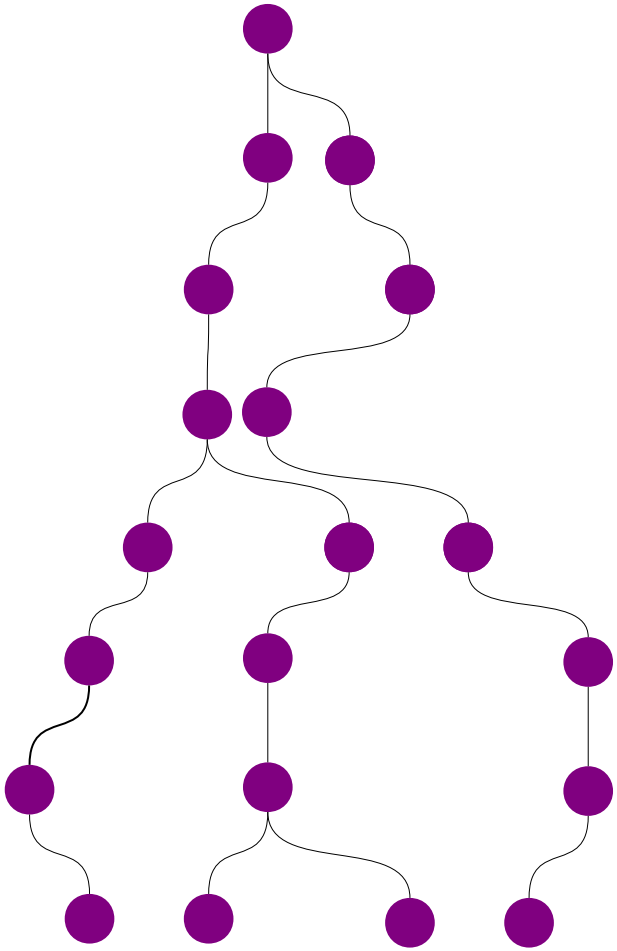
- Two ancestral lines merge once they reach a common parent. For each given pair of lines, this has probability $1/N$ per generation.
- For large populations, *if time is measured in units of N generations*, the time until two lines meet, converges to the exponential distribution with paramter 1.
- In fact, the whole the ancestral process of a sample, viewed as random tree, converges in distribution to the *Kingman coalescent* ([Kingman 1982])

Ancestry of a sample in the Wright-Fisher model



- Two ancestral lines merge once they reach a common parent. For each given pair of lines, this has probability $1/N$ per generation.
- For large populations, *if time is measured in units of N generations*, the time until two lines meet, converges to the exponential distribution with paramter 1.
- In fact, the whole the ancestral process of a sample, viewed as random tree, converges in distribution to the *Kingman coalescent* ([Kingman 1982])

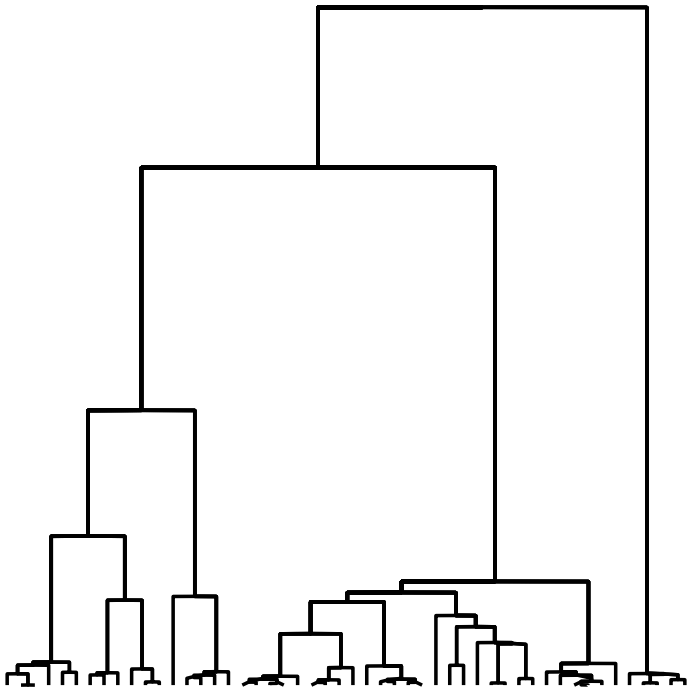
Ancestry of a sample in the Wright-Fisher model



- Two ancestral lines merge once they reach a common parent. For each given pair of lines, this has probability $1/N$ per generation.
- For large populations, *if time is measured in units of N generations*, the time until two lines meet, converges to the exponential distribution with paramter 1.
- In fact, the whole the ancestral process of a sample, viewed as random tree, converges in distribution to the *Kingman coalescent* ([Kingman 1982])

The coalescent [Kingman, 1982]

$N = 50$



- Each pair of lines merges at rate 1. Total rate if n lines present is $\binom{n}{2}$.
- Universal scaling limit for many population models (Moran model, etc.) essentially if offspring variance is finite.
- Standard null model in population genetics, sophisticated theory and inference methods available. See e.g. [Wakeley 2008].

Time to the most recent common ancestor T_{MRCA}

Many quantities of interest depend on the time to the most recent common ancestor of a sample (of size n, say). Denote this time by $T(n)$. Then, due to lack of memory,

$$T(n) \stackrel{d}{=} S_n + S_{n-1} + \cdots + S_2,$$

where the S_k are independent exponential random variables with rate $\binom{k}{2}$.

Time to the most recent common ancestor T_{MRCA}

Hence,

$$E[T(n)] = \sum_{k=2}^{n} E[S_k] = \sum_{k=2}^{n} \frac{2}{k(k-1)}$$
$$= 2\sum_{k=2}^{n} \Big(\frac{1}{k-1} - \frac{1}{k} \Big) = 2\Big(1 - \frac{1}{n}\Big).$$

Note that the whole population merges rather quickly! In particular,

$$1 = E[T(2)] \leq E[T(N)] < 2$$

(but keep in mind that time is measured in units of N).

The total length of the coalescent tree

In a similar fashion, we may compute the expected length $L(n)$ of the coalescent tree of a sample of size n. Indeed,

$$L(n) \stackrel{d}{=} nS_n + (n-1)S_{n-1} + \cdots + 2S_2,$$

so that

$$E[L(n)] = \sum_{j=2}^{n} j \frac{2}{j(j-1)} = 2\sum_{j=1}^{n-1} \frac{1}{j} \sim 2\log n.$$

Length grows slowly! This is very unfortunate from a statistical point of view. In fact, $E[L(n)]$ is proportional to the expected number of mutations seen in the sample.

It is not hard to show that $L_n - 2\log n$ is in the domain of attraction of the Gumbel extreme value distribution.

The total length of the coalescent tree

In a similar fashion, we may compute the expected length $L(n)$ of the coalescent tree of a sample of size n. Indeed,

$$L(n) \stackrel{d}{=} nS_n + (n-1)S_{n-1} + \cdots + 2S_2,$$

so that

$$E[L(n)] = \sum_{j=2}^{n} j \frac{2}{j(j-1)} = 2\sum_{j=1}^{n-1} \frac{1}{j} \sim 2\log n.$$

Length grows slowly! This is very unfortunate from a statistical point of view. In fact, $E[L(n)]$ is proportional to the expected number of mutations seen in the sample.

It is not hard to show that $L_n - 2\log n$ is in the domain of attraction of the GUMBEL extreme value distribution.

The total length of the coalescent tree

In a similar fashion, we may compute the expected length $L(n)$ of the coalescent tree of a sample of size n. Indeed,

$$L(n) \stackrel{d}{=} nS_n + (n-1)S_{n-1} + \cdots + 2S_2,$$

so that

$$E[L(n)] = \sum_{j=2}^{n} j \frac{2}{j(j-1)} = 2\sum_{j=1}^{n-1} \frac{1}{j} \sim 2\log n.$$

Length grows slowly! This is very unfortunate from a statistical point of view. In fact, $E[L(n)]$ is proportional to the expected number of mutations seen in the sample.

It is not hard to show that $L_n - 2\log n$ is in the domain of attraction of the GUMBEL extreme value distribution.

Kingman coalescent as random tree

Properties of the Kingman coalescent tree distribution are still an active area of research, beginning with simple properties observed by population geneticists as above (Watterson 1975, Hudson 1990,...).

See e.g. recent work of Kersting & Stanciu (2013) for a limit theorem for the joint law of the lengths of the internal and external branches...

... as well as related work by many others!

Kingman coalescent as random tree

Properties of the Kingman coalescent tree distribution are still an active area of research, beginning with simple properties observed by population geneticists as above (Watterson 1975, Hudson 1990,...).

See e.g. recent work of Kersting & Stanciu (2013) for a limit theorem for the joint law of the lengths of the internal and external branches...

... as well as related work by many others!

Background: More general coalescents

Donnelly & Kurtz (1999), Pitman (1999) and Sagitov (1999) independently introduce a general class of coalescent processes, which allow *multiple collisions.* These exchangeable coalescents with consistent rates are now known as **Λ-coalescents.**

Bertoin & Le Gall (2003) characterize these coalescents via flows of bridges.

The coalescent rates are easy to describe: If there are m blocks at present, each *given* k-tuple merges to a single block at rate

$$b_{m,k} = \int_{[0,1]} x^k (1-x)^{m-k} \frac{1}{x^2} \Lambda(dx), \tag{1}$$

$k \in \{2, \dots m\}$, where Λ is some finite measure on $[0, 1]$.

Background: More general coalescents

Donnelly & Kurtz (1999), Pitman (1999) and Sagitov (1999) independently introduce a general class of coalescent processes, which allow *multiple collisions.* These exchangeable coalescents with consistent rates are now known as Λ-**coalescents.**

Bertoin & Le Gall (2003) characterize these coalescents via flows of bridges.

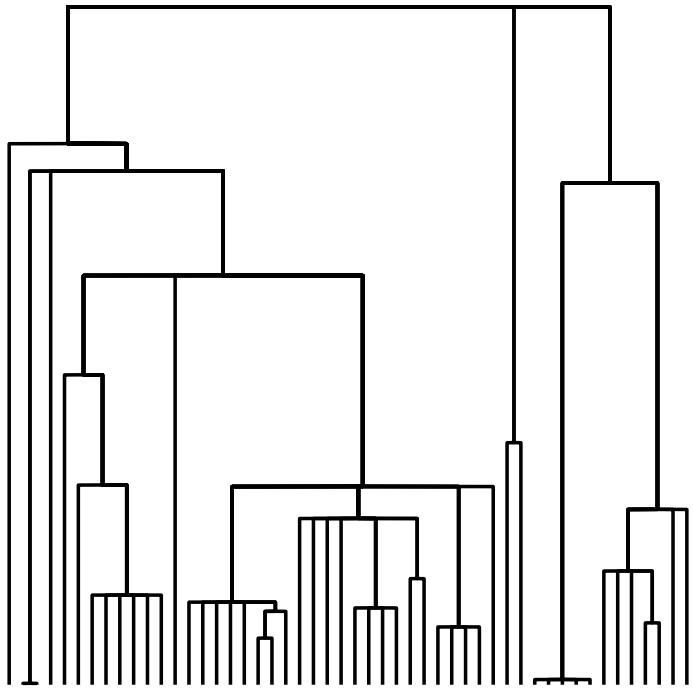
The coalescent rates are easy to describe: If there are m blocks at present, each *given* k-tuple merges to a single block at rate

$$b_{m,k} = \int_{[0,1]} x^k (1-x)^{m-k} \frac{1}{x^2} \Lambda(dx), \tag{1}$$

$k \in \{2, \dots m\}$, where Λ is some finite measure on $[0, 1]$.

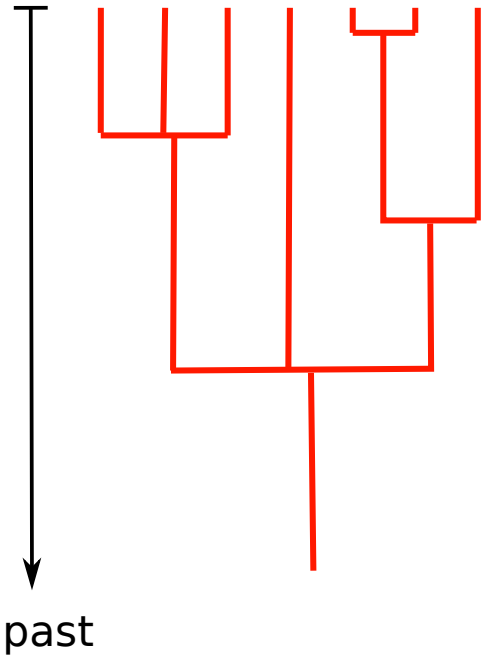
Example

The *Bolthausen-Sznitman* coalescent corresponds to $\Lambda = \text{Unif}_{[0,1]}$. A typical realisation may look like this ($N = 50$):



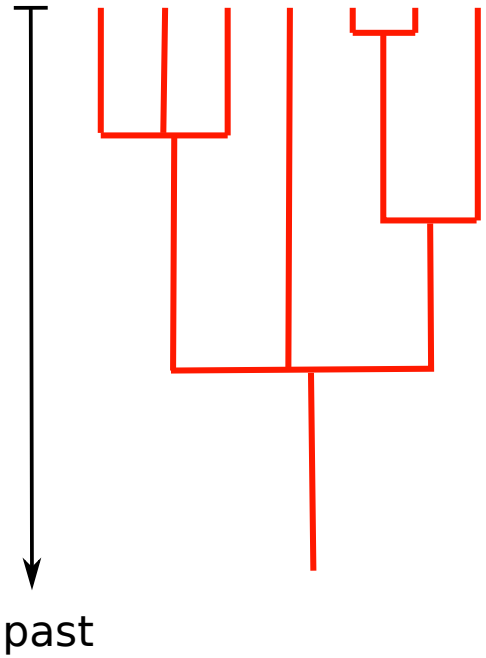
Of course, the tree properties of such a coalescent a very different from the standard Kingman coalescent.

Lambda- and Xi-coalescents



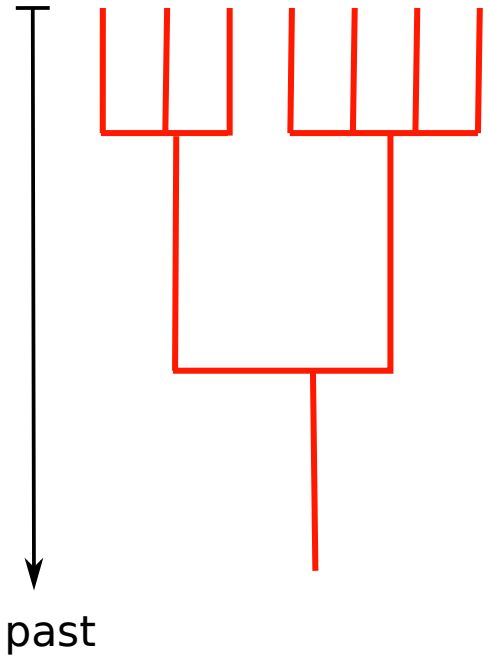
- Include Kingman's coalescent for $\Lambda = \delta_0$
- Limit model of genealogies for exchangeable population models with 'skewed' individual offspring distribution (esp. the case where Λ is a Beta$(2-\alpha, \alpha)$ measure, $\alpha \in (0, 2]$, or atomic $\delta_\psi, \psi \in [0, 1]$)
- Even more general: Ξ-coalescents Schweinsberg 2000; Möhle & Sagitov 2001. These allow *simultaneous* multiple mergers.

Lambda- and Xi-coalescents



- Include Kingman's coalescent for $\Lambda = \delta_0$
- Limit model of genealogies for exchangeable population models with 'skewed' individual offspring distribution (esp. the case where Λ is a Beta$(2-\alpha, \alpha)$ measure, $\alpha \in (0, 2]$, or atomic $\delta_\psi, \psi \in [0, 1]$)
- Even more general: Ξ-coalescents Schweinsberg 2000; Möhle & Sagitov 2001. These allow *simultaneous* multiple mergers.

Lambda- and Xi-coalescents



- Include Kingman's coalescent for $\Lambda = \delta_0$
- Limit model of genealogies for exchangeable population models with 'skewed' individual offspring distribution (esp. the case where Λ is a Beta$(2-\alpha, \alpha)$ measure, $\alpha \in (0, 2]$, or atomic $\delta_\psi, \psi \in [0, 1]$)
- Even more general: Ξ-coalescents Schweinsberg 2000; Möhle & Sagitov 2001. These allow *simultaneous* multiple mergers.

Lambda- and Xi-coalescents – are they really out there?

Mechanisms producing multiple or simultaneous multiple mergers in genealogy:

- Skewed offspring distribution ('sweepstakes reproduction' in Atlantic Cod, e.g. Hedgecock & Pudovkin's review in *Bull. Marine Sci.* 2011)
- Recurrent severe bottlenecks, leading to Ξ-coalescents (obtained as sub-ordinated Kingman coalescents), Birkner, Blath et al (2009)
- Selective sweeps (in combination with genetic hitchhiking) Durrett & Schweinsberg (2004, 2005); Coop & Ralph (2012)
- Selection in the guise of a fitness wave: Leading to Bolthausen-Sznitman coalescent, predictions of Desai, Walczak & Fisher (2013) and Neher & Hallatschek (2013), made rigorous by Schweinsberg (2015)
- Other mechanisms (e.g. host-parasite interaction), see Tellier & Lemaire *Mol Ecol* (2014) for an overview from the biological perspective

Lambda- and Xi-coalescents – are they really out there?

Mechanisms producing multiple or simultaneous multiple mergers in genealogy:

- Skewed offspring distribution ('sweepstakes reproduction' in Atlantic Cod, e.g. Hedgecock & Pudovkin's review in *Bull. Marine Sci.* 2011)
- Recurrent severe bottlenecks, leading to Ξ-coalescents (obtained as sub-ordinated Kingman coalescents), Birkner, Blath et al (2009)
- Selective sweeps (in combination with genetic hitchhiking) Durrett & Schweinsberg (2004, 2005); Coop & Ralph (2012)
- Selection in the guise of a fitness wave: Leading to Bolthausen-Sznitman coalescent, predictions of Desai, Walczak & Fisher (2013) and Neher & Hallatschek (2013), made rigorous by Schweinsberg (2015)
- Other mechanisms (e.g. host-parasite interaction), see Tellier & Lemaire *Mol Ecol* (2014) for an overview from the biological perspective

Lambda- and Xi-coalescents – are they really out there?

Mechanisms producing multiple or simultaneous multiple mergers in genealogy:

- Skewed offspring distribution ('sweepstakes reproduction' in Atlantic Cod, e.g. Hedgecock & Pudovkin's review in *Bull. Marine Sci.* 2011)
- Recurrent severe bottlenecks, leading to Ξ-coalescents (obtained as sub-ordinated Kingman coalescents), Birkner, Blath et al (2009)
- Selective sweeps (in combination with genetic hitchhiking) Durrett & Schweinsberg (2004, 2005); Coop & Ralph (2012)
- Selection in the guise of a fitness wave: Leading to Bolthausen-Sznitman coalescent, predictions of Desai, Walczak & Fisher (2013) and Neher & Hallatschek (2013), made rigorous by Schweinsberg (2015)
- Other mechanisms (e.g. host-parasite interaction), see Tellier & Lemaire *Mol Ecol* (2014) for an overview from the biological perspective

Lambda- and Xi-coalescents – are they really out there?

Mechanisms producing multiple or simultaneous multiple mergers in genealogy:

- Skewed offspring distribution ('sweepstakes reproduction' in Atlantic Cod, e.g. Hedgecock & Pudovkin's review in *Bull. Marine Sci.* 2011)
- Recurrent severe bottlenecks, leading to Ξ-coalescents (obtained as sub-ordinated Kingman coalescents), Birkner, Blath et al (2009)
- Selective sweeps (in combination with genetic hitchhiking) Durrett & Schweinsberg (2004, 2005); Coop & Ralph (2012)
- Selection in the guise of a fitness wave: Leading to Bolthausen-Sznitman coalescent, predictions of Desai, Walczak & Fisher (2013) and Neher & Hallatschek (2013), made rigorous by Schweinsberg (2015)
- Other mechanisms (e.g. host-parasite interaction), see Tellier & Lemaire *Mol Ecol* (2014) for an overview from the biological perspective

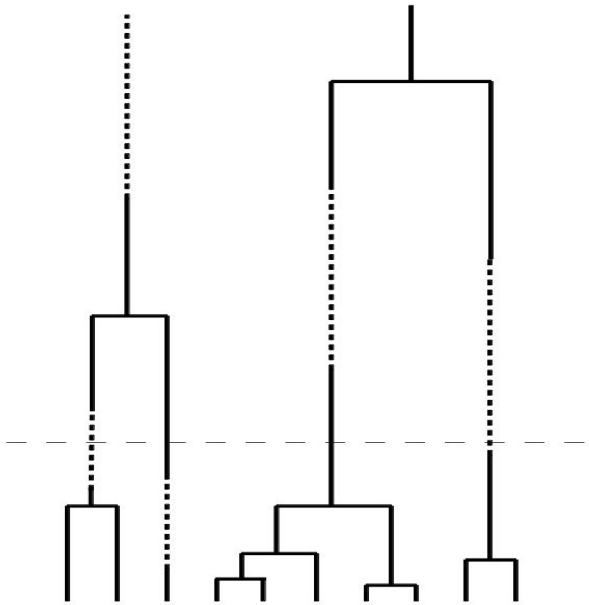
Lambda- and Xi-Fleming-Viot diffusions

Note that the above coalescent models are dual to Wright-Fisher-type "diffusions" with jumps.

Inferring the underlying coalescent distribution is in this way translated into inferring the jumps size distribution in the jump diffusion - related to inference questions in finance?

A new variant of coalescents

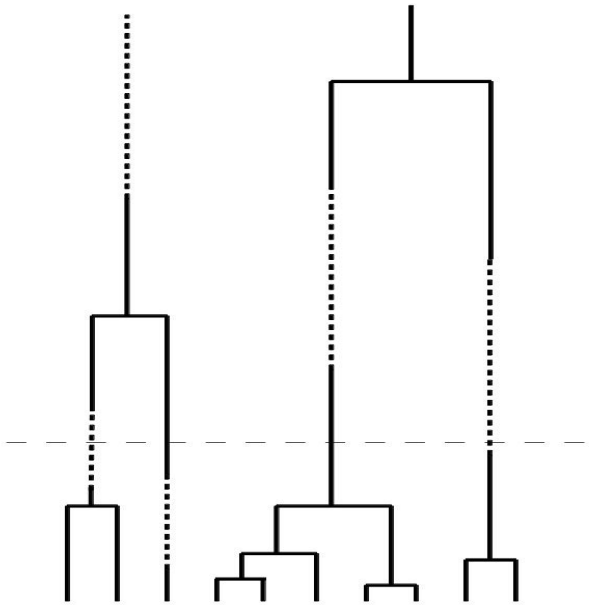
Consider (Kingman-, Lambda-, Xi-)coalescent processes in which lineages may, independently of each other, switch freely between the states turned "on" and "off".



Lineages that are currently turned "off" (dotted lines above) are not allowed to coalesce with other lineages (and do not contribute to the overall coalescence rate).

A new variant of coalescents

Consider (Kingman-, Lambda-, Xi-)coalescent processes in which lineages may, independently of each other, switch freely between the states turned "on" and "off".



Lineages that are currently turned "off" (dotted lines above) are not allowed to coalesce with other lineages (and do not contribute to the overall coalescence rate).

On/off-coalescents

We call the new class, after a discussion with A. Lambert, on/off-coalescents. Again, the tree properties change drastically in comparison with previous coalescent models.

An example of such colaescents arise e.g. in metapopulation/speciation models in the guise of the 'peripatric coalescent' (Lambert, Ma, 2015) or as 'seed bank coalescent', that we will describe below.

Indeed, in the seed bank scenario, lineages that are currently turned off correspond to 'seeds' or dormant individuals. Dormancy seems to be a widespread important evolutionary strategy, in particular among microbial communities.

On/off-coalescents

We call the new class, after a discussion with A. Lambert, on/off-coalescents. Again, the tree properties change drastically in comparison with previous coalescent models.

An example of such colaescents arise e.g. in metapopulation/speciation models in the guise of the 'peripatric coalescent' (Lambert, Ma, 2015) or as 'seed bank coalescent', that we will describe below.

Indeed, in the seed bank scenario, lineages that are currently turned off correspond to 'seeds' or dormant individuals. Dormancy seems to be a widespread important evolutionary strategy, in particular among microbial communities.

Dormancy in microbial populations

Seed banks play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial populations

Seed banks play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial populations

Seed banks play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial populations

Seed banks play an important role in microbial evolution [Lennon & Jones, *Nature reviews*, 2011]:

- Many microbial species exhibit *dormant forms*. That is, organisms can enter (and leave) a reversible state of low (resp. vanishing) metabolic activity. These forms can be short-lived but may also stay inactive for significant periods of time. A variety of bacteria can produce endospores or cysts that remain viable for many decades/centuries.
- Dormant microorganisms generate a *seed bank*, which comprises inactive individuals that are capable of being resuscitated.
- Initiation of dormancy may be triggered by environment, but may also happen spontaneously (*responsive* vs. *spontaneous switching*).
- A large fraction of the microorganisms in nature seem to be metabolically inactive.

Dormancy in microbial communities

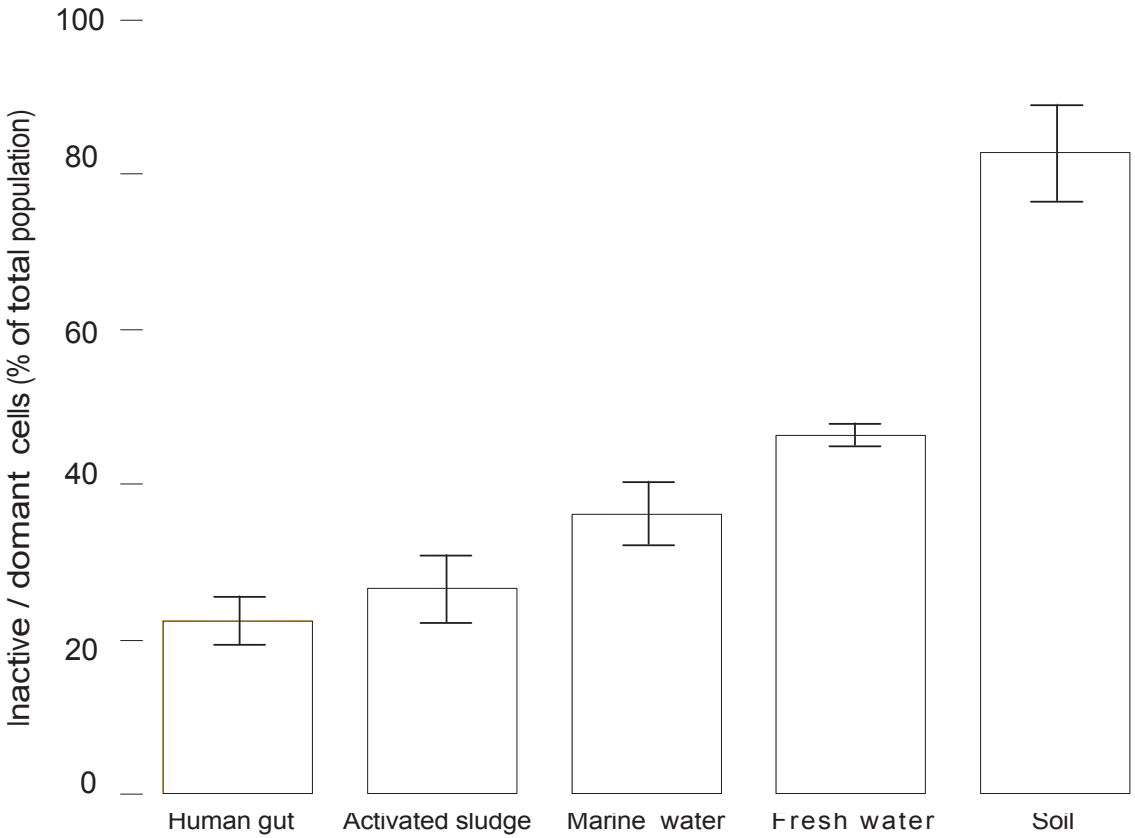


Figure: Percentage of inactive cells in microbial communities, data from [Lennon & Jones, 2011])

Modeling and investigating seed-banks

Despite many empirical studies and several theoretical works (including e.g. [Kaj, Krone & Lascoux 2001], [Vitalis, Glémin, Oliveiri 2004], [Tellier, Laurent, Lainer, Pavlidis, Stephan 2011], [Živković & Tellier 2011],...), the mathematical modeling of seed banks in population genetics appears to be still incomplete.

Seed banks in bacterial communities

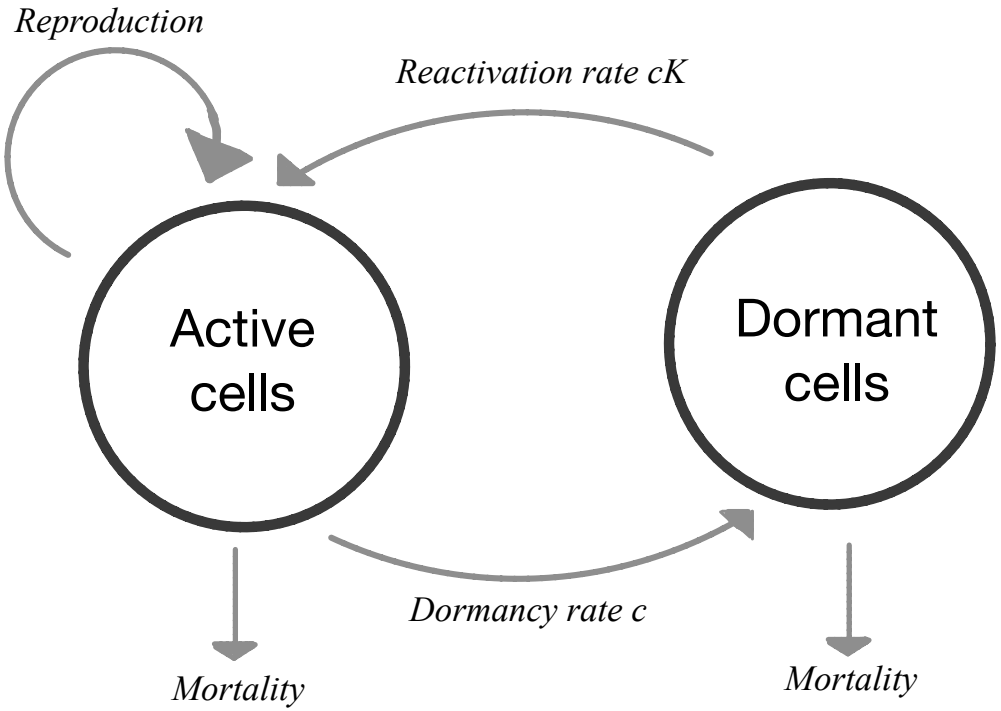
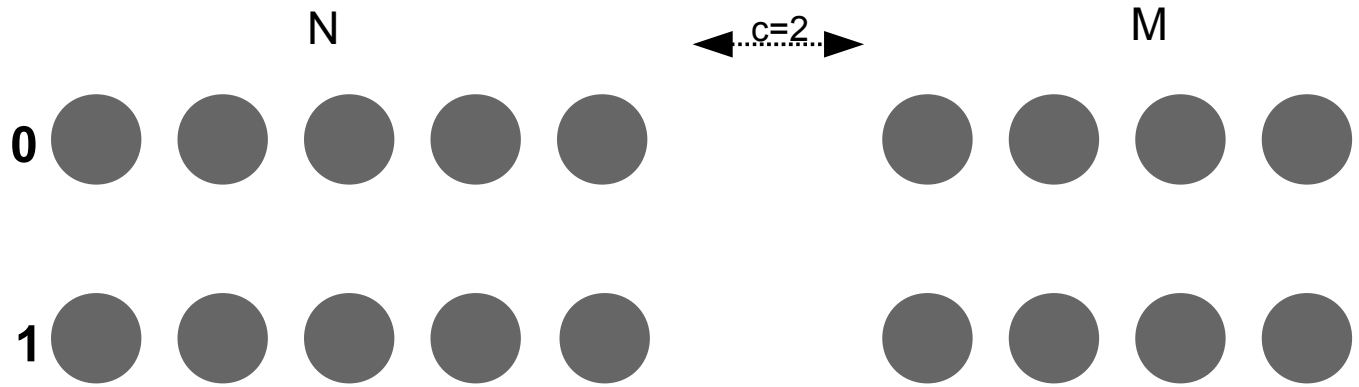


Figure: Dormancy: Initiaton and resuscitation, à la [Lennon & Jones 2011]

A Wright-Fisher model with seed bank

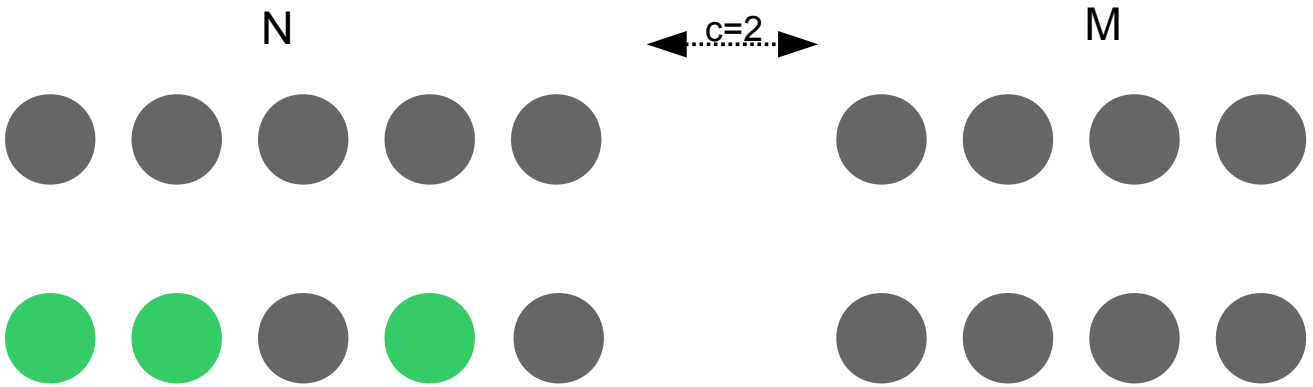
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

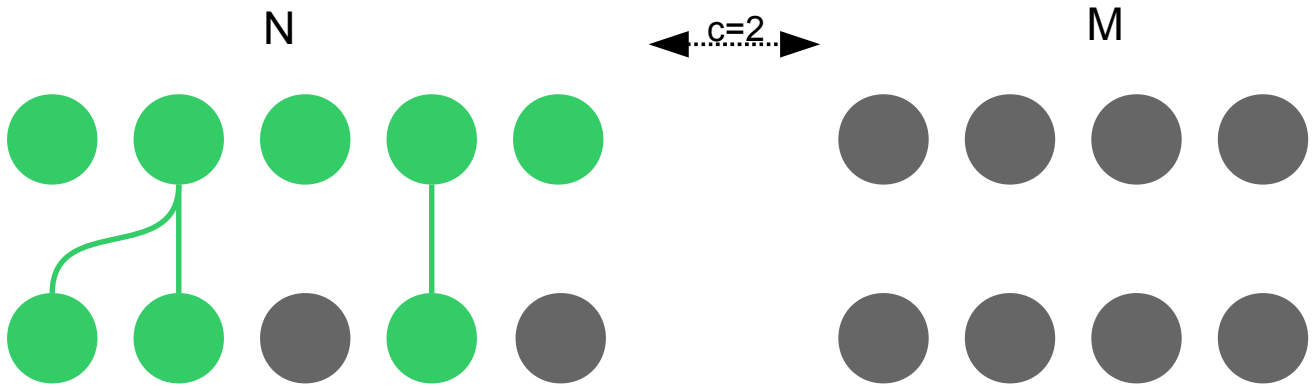
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

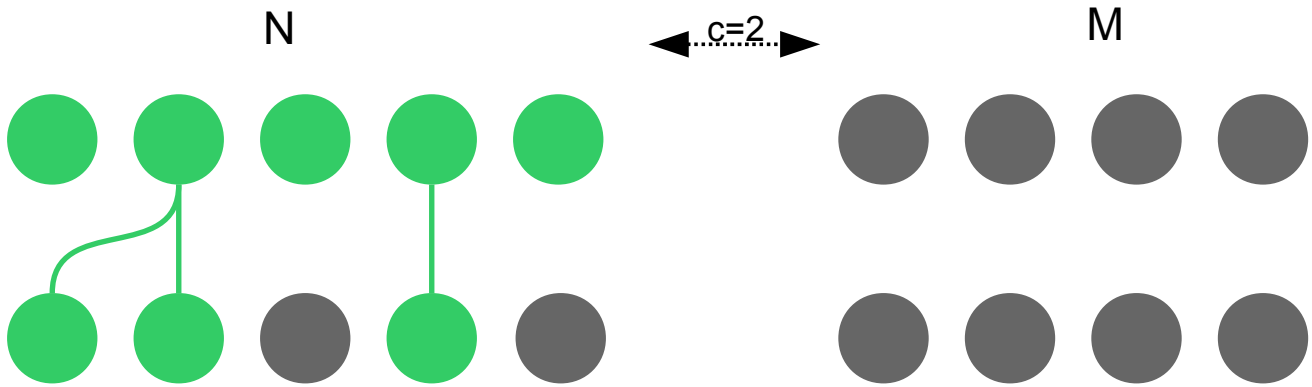
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

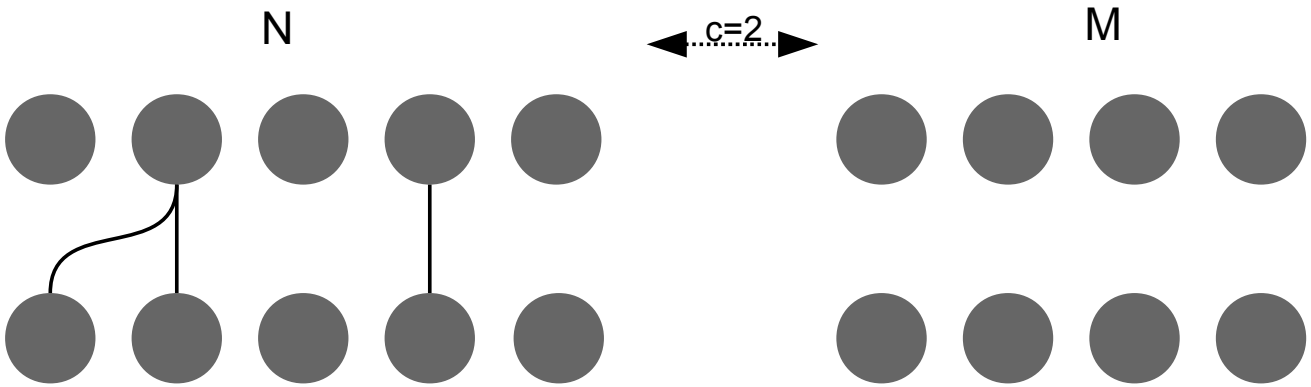
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

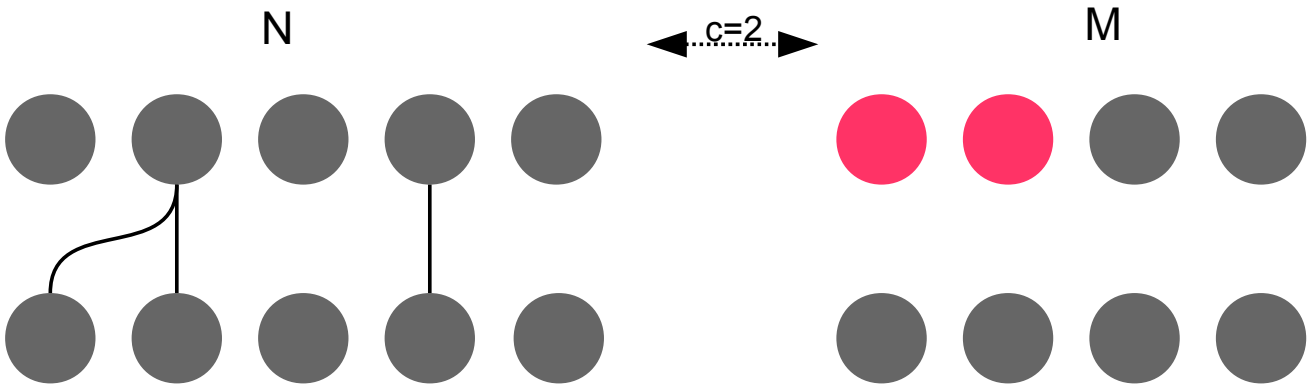
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

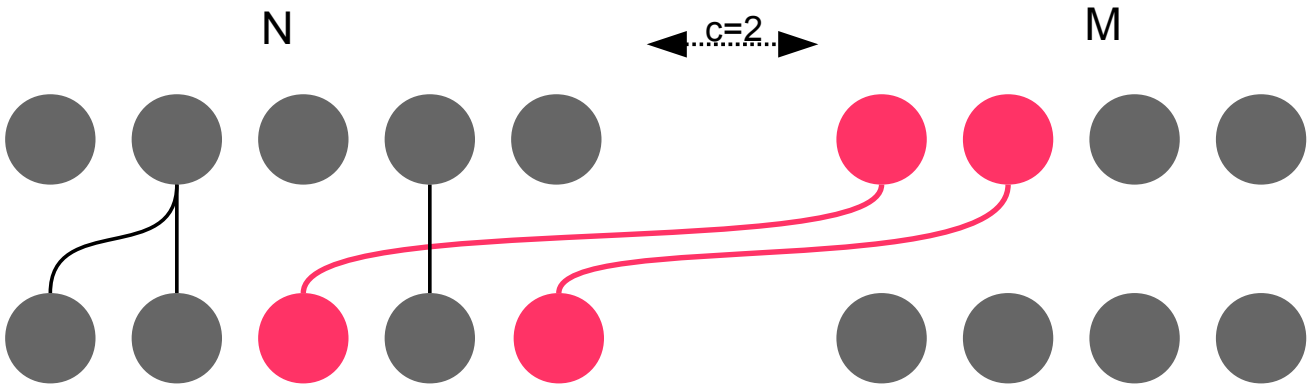
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

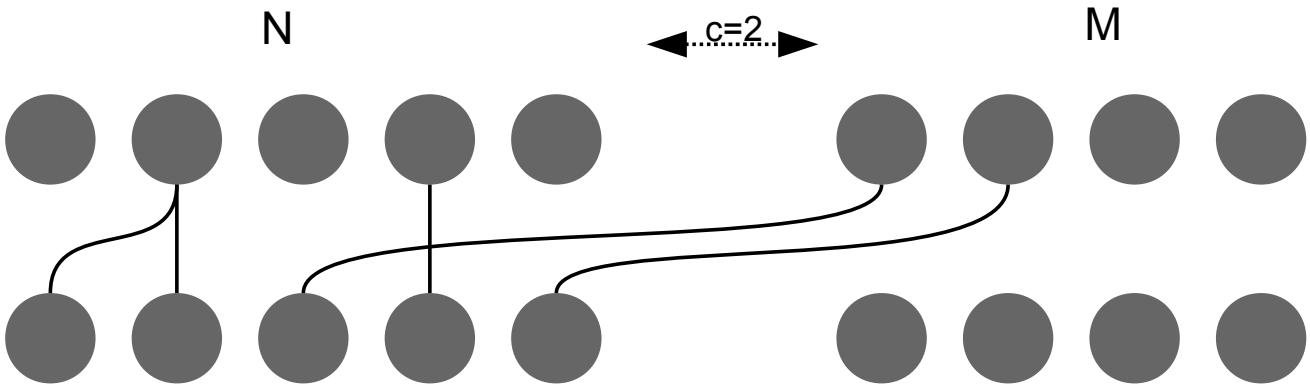
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

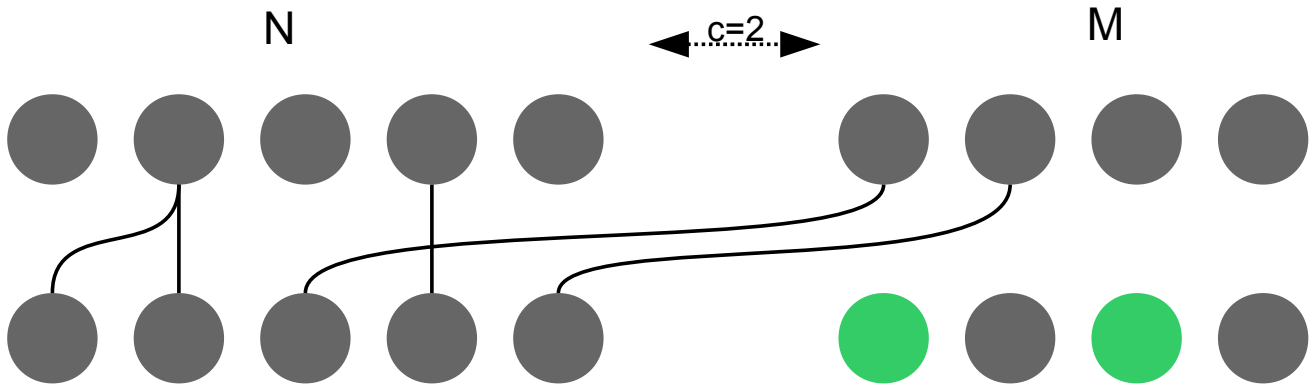
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

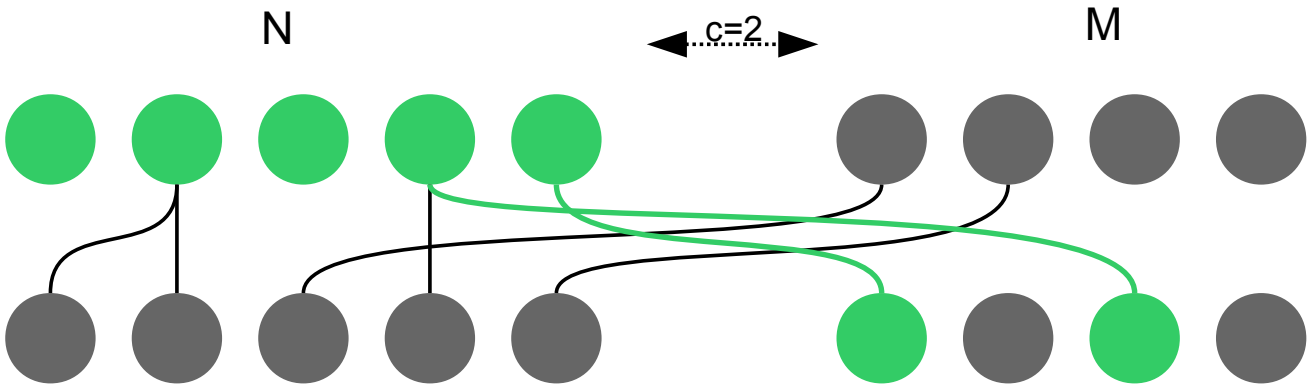
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

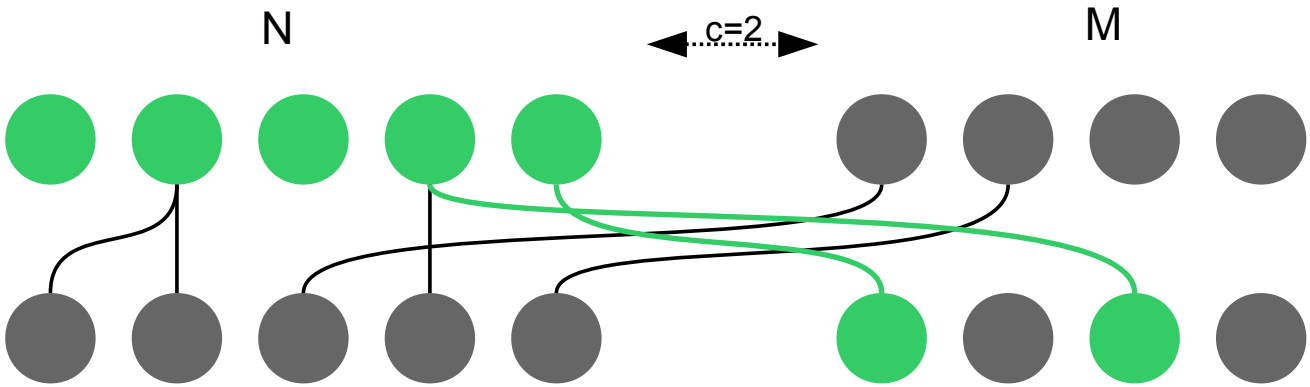
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

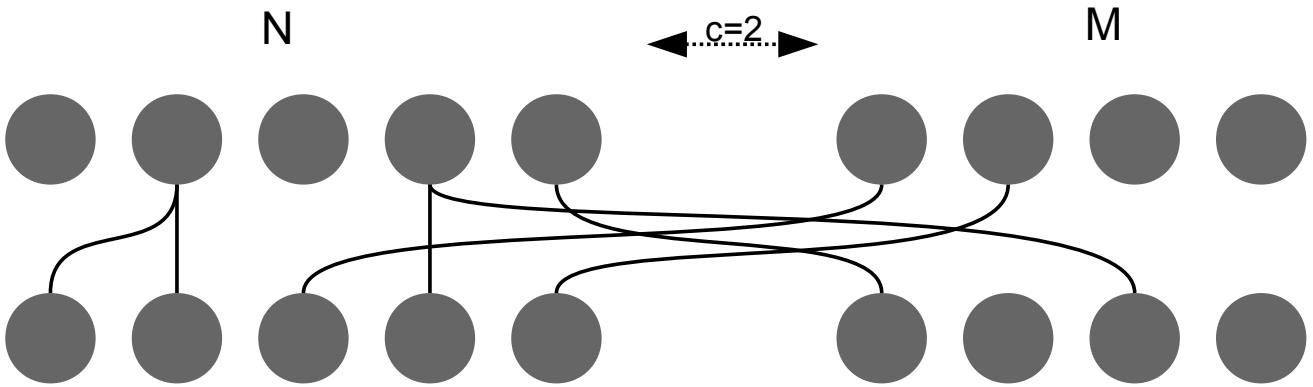
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

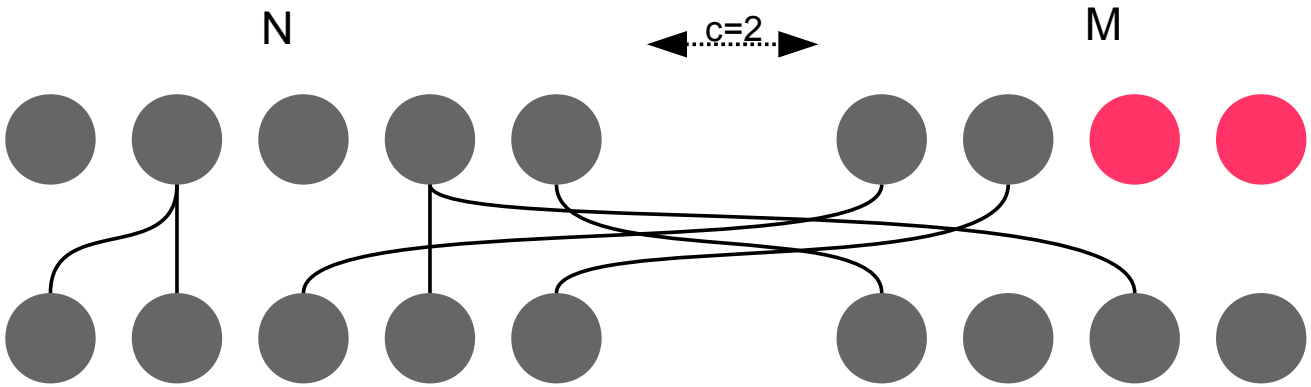
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

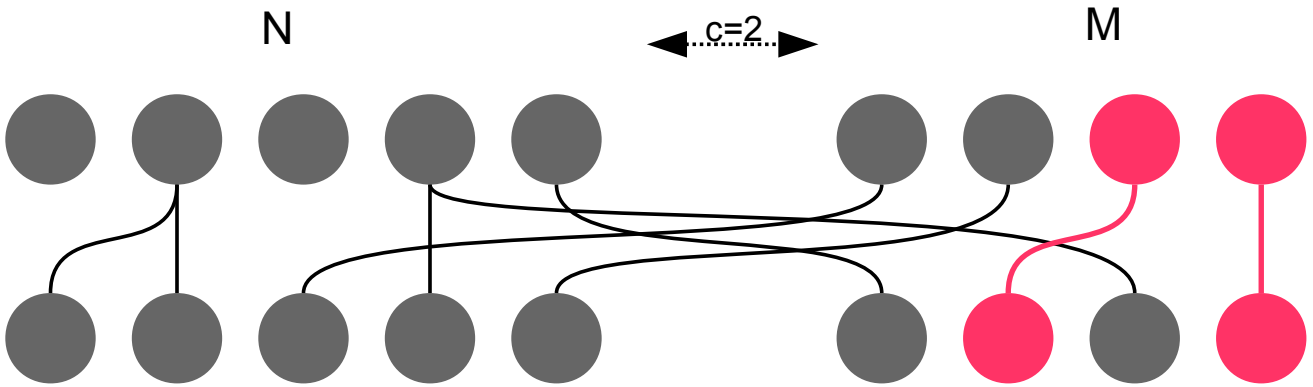
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

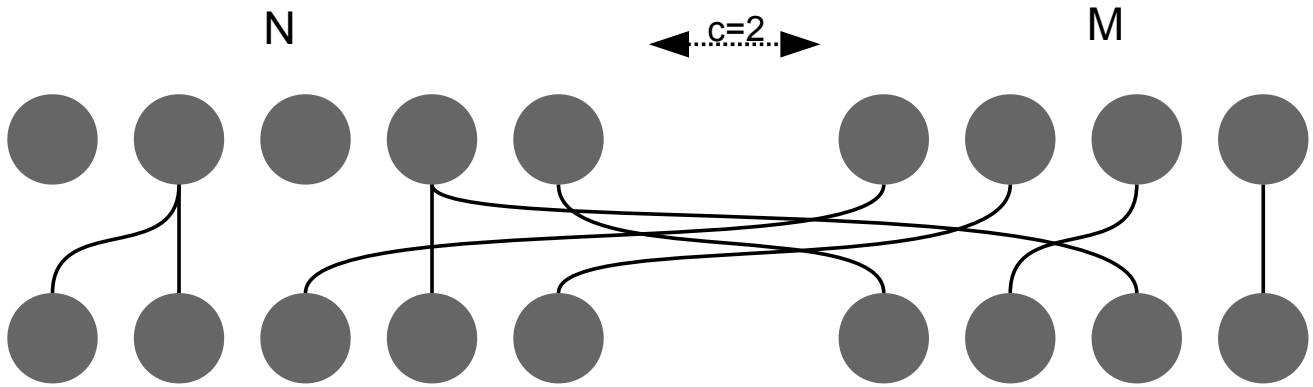
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

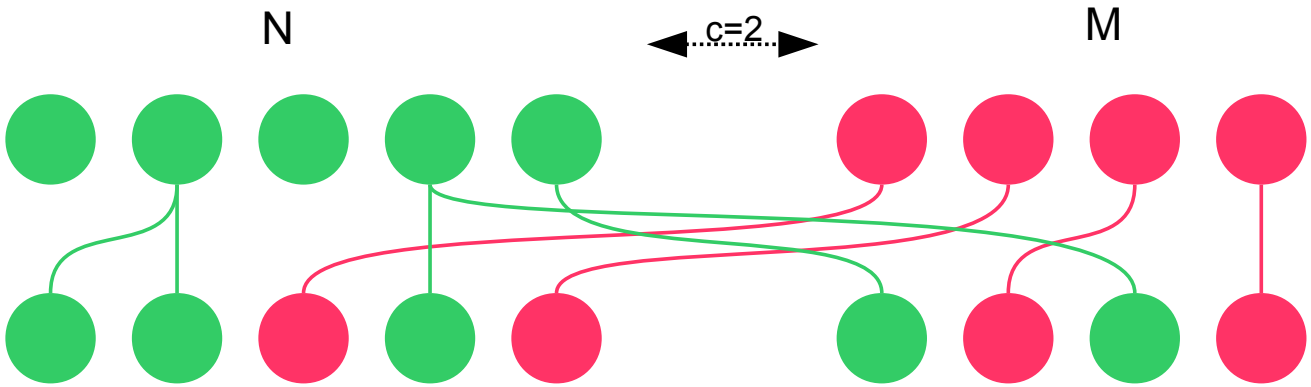
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

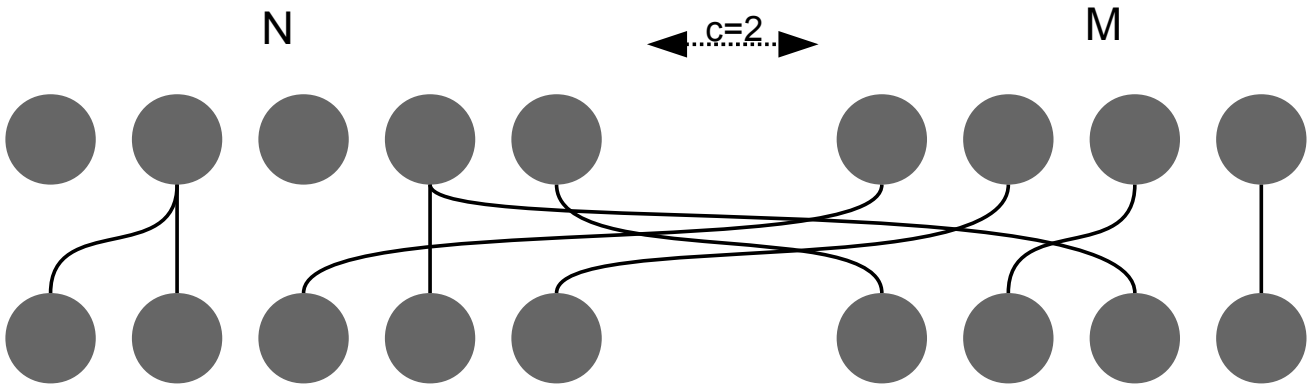
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

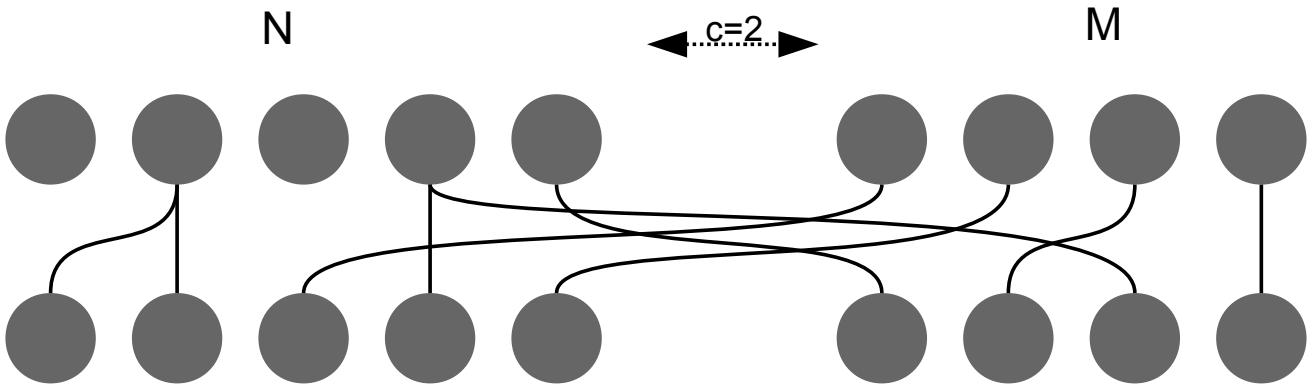
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

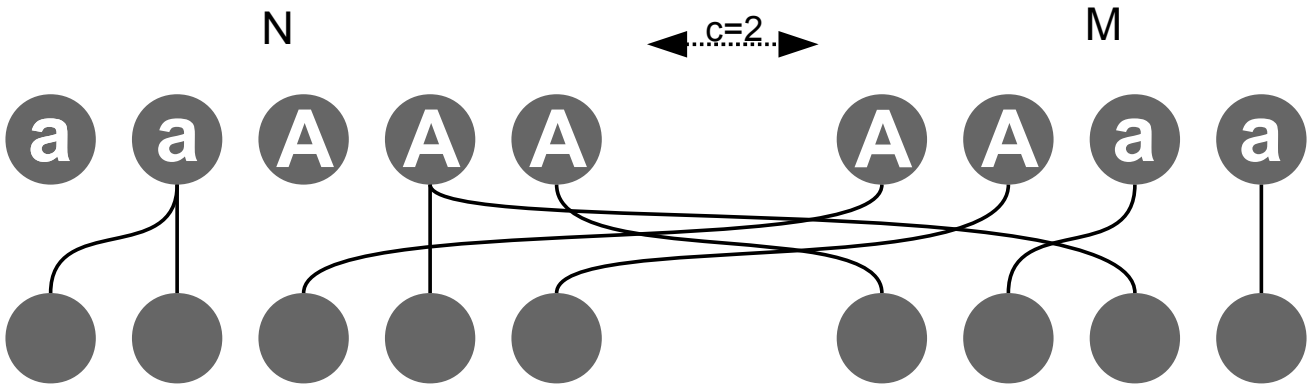
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$ *and type-space* $\{a, A\}$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

A Wright-Fisher model with seed bank

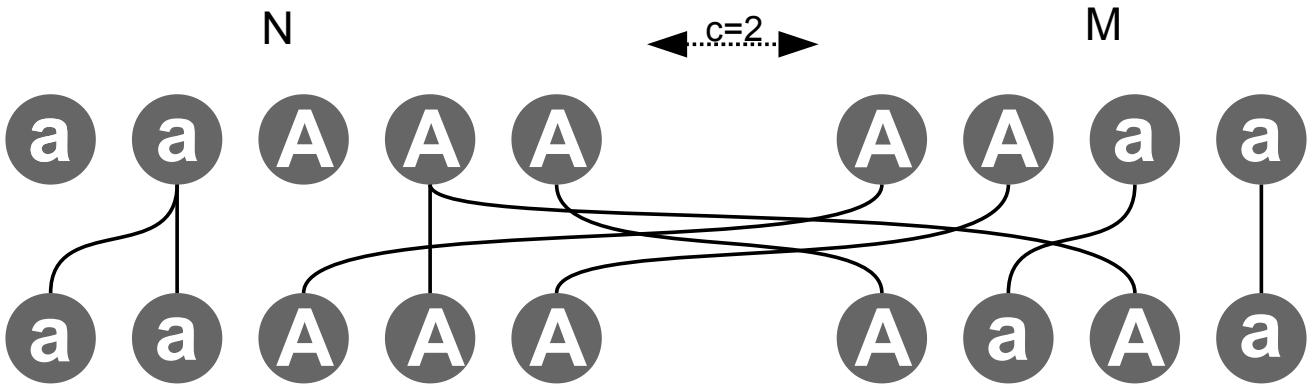
Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$ *and type-space* $\{a, A\}$



- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

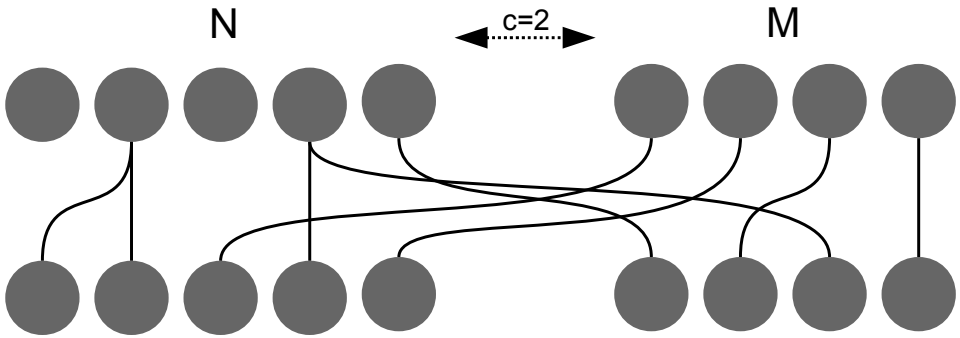
A Wright-Fisher model with seed bank

Again, consider a haploid population reproducing in discrete generations $k = 0, 1, 2...$ *and type-space* $\{a, A\}$

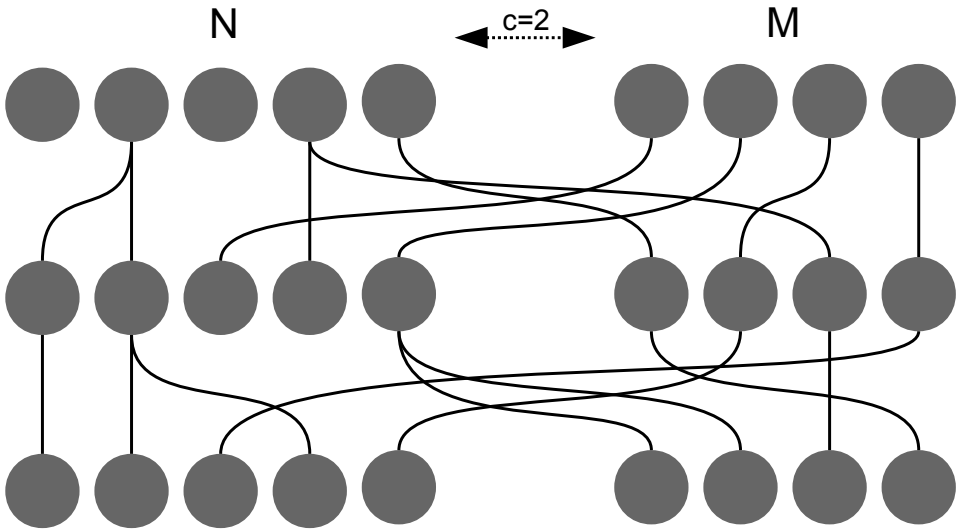


- N active individuals ("plants") produce $N - c$ active offspring by multinomial sampling
- c dormant individuals ("seeds") are selected uniformly to resuscitate: they produce exactly one active individual each
- N active individuals produce c dormant offspring by multinomial sampling
- $N - c$ dormant individuals stay in the seed-bank
- Offspring inherit the type of their parents

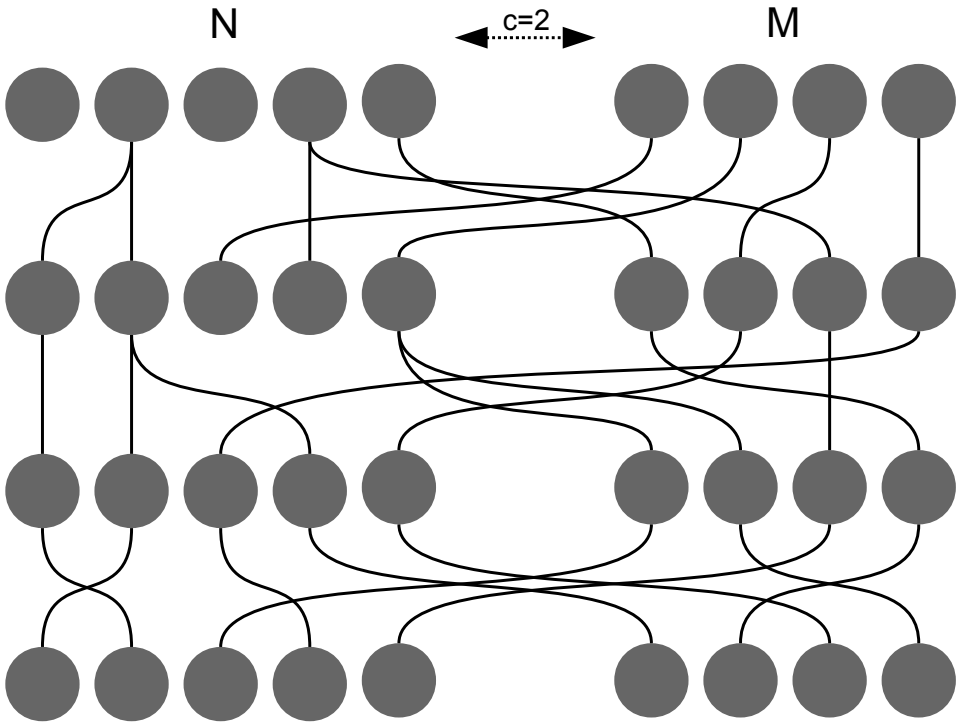
The Wright-Fisher model with seed bank



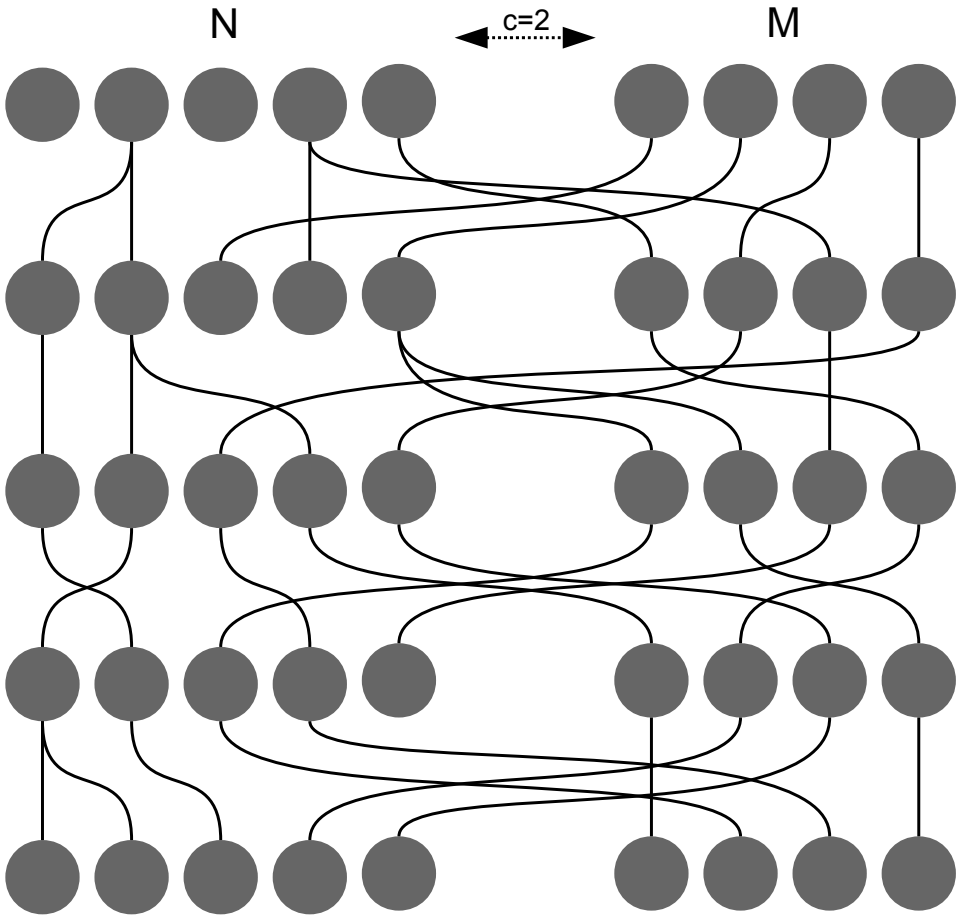
The Wright-Fisher model with seed bank



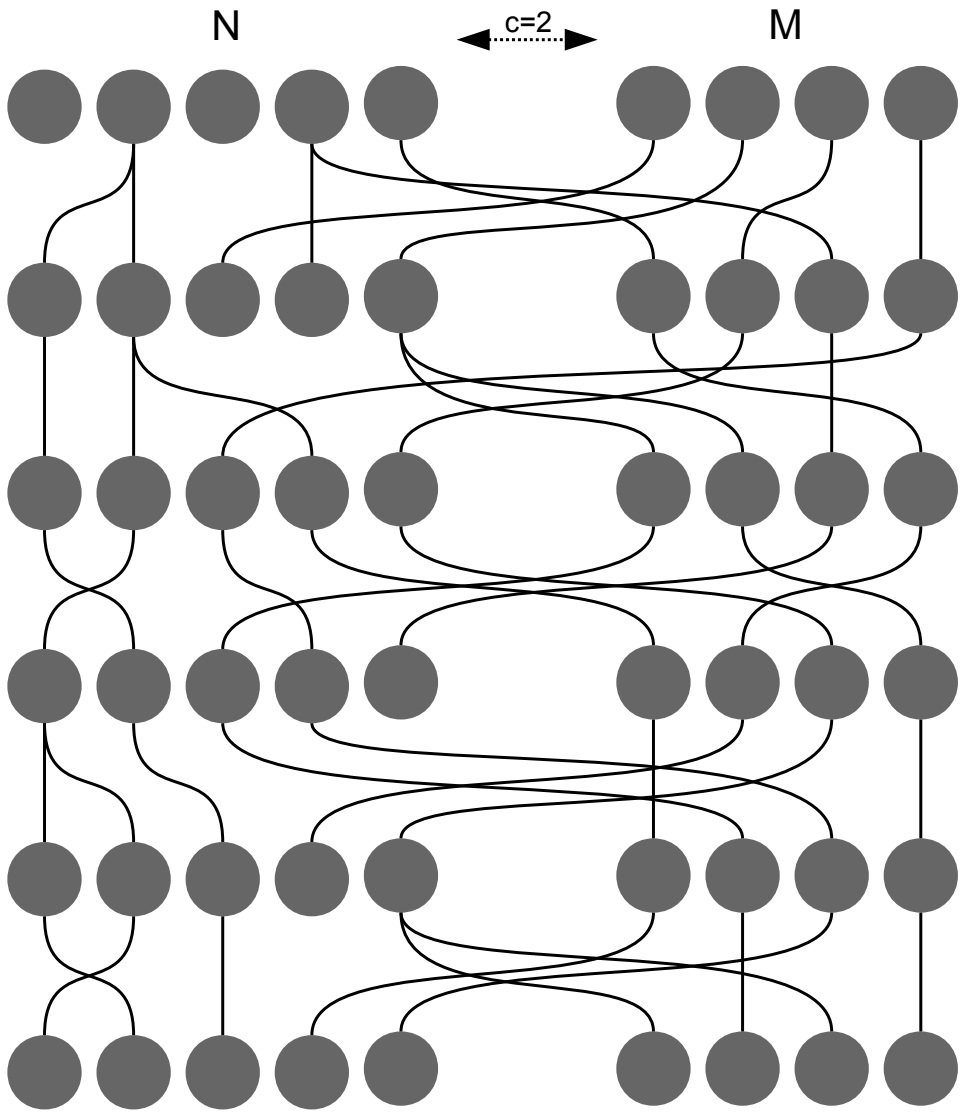
The Wright-Fisher model with seed bank



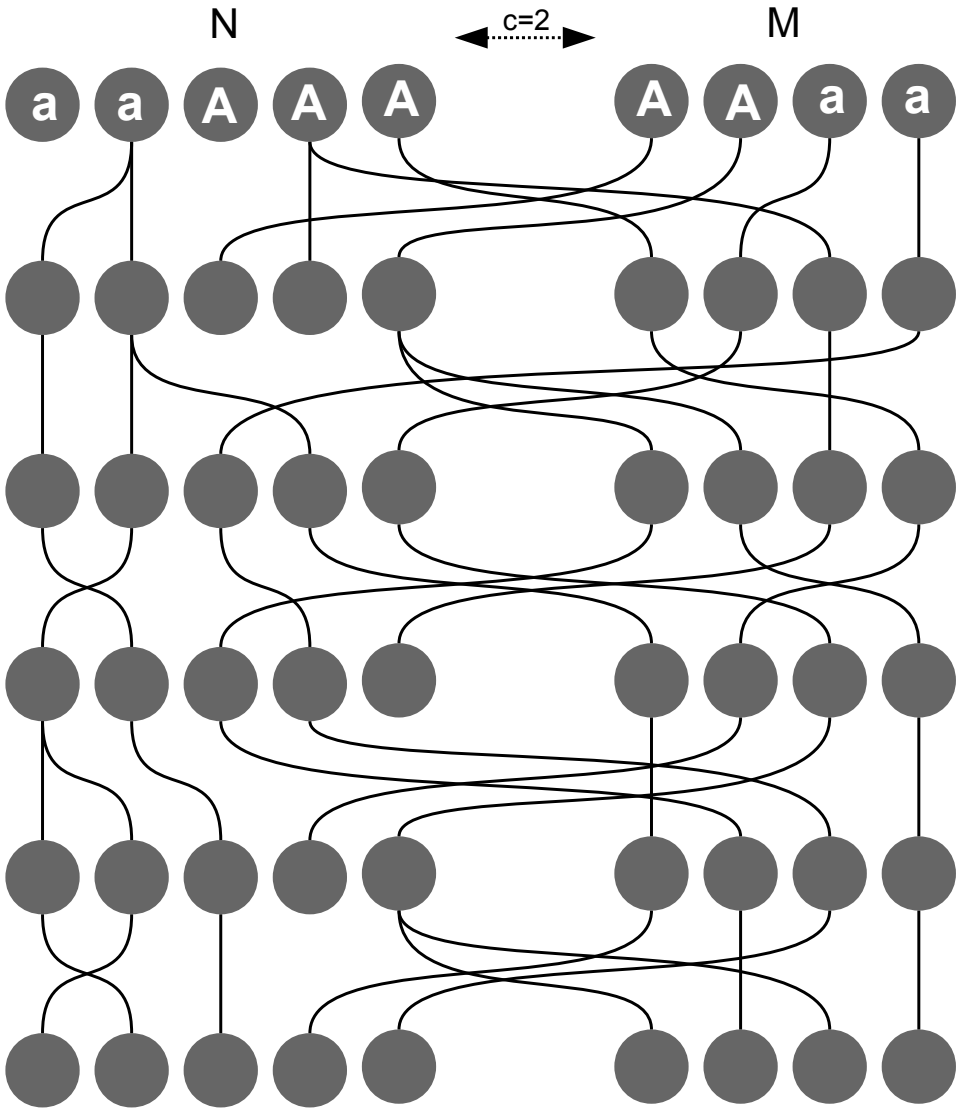
The Wright-Fisher model with seed bank



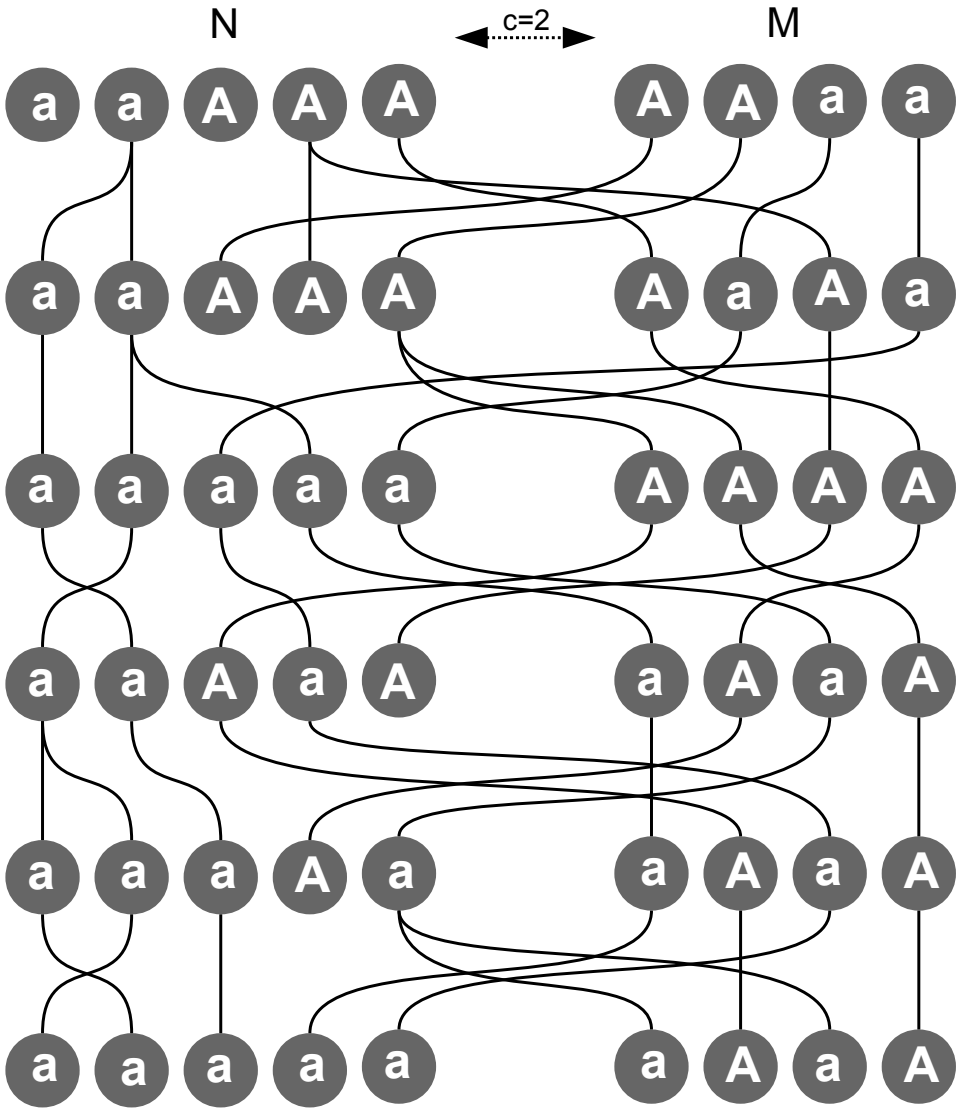
The Wright-Fisher model with seed bank



The Wright-Fisher model with seed bank



The Wright-Fisher model with seed bank



Wright-Fisher model with geometric seed-bank component

Fix pop.-size $N \in \mathbb{N}$, seed-bank size M, genetic type space E (e.g. $E = \{a, A\}$) and parameters $c > 0, K := N/M$ as before. Given initial type configurations $\xi_0 \in E^N$ and $\eta_0 \in E^M$, let

$$\xi_k := \big(\xi_k(i), i \in \{1, \ldots, N\}\big), \quad k \in \mathbb{N},$$

be the random genetic type configuration in E^N of the *plants* in generation k (obtained from the above mechanism), and

$$\eta_k := \big(\eta_k(j), j \in \{1, \ldots, M\}\big), \quad k \in \mathbb{N},$$

be the genetic type configuration of the *seeds* in E^M. We call the discrete-time Markov chain $(\xi_k, \eta_k)_{k \in \mathbb{N}_0}$ with values in $E^N \times E^M$ the *type configuration process* of the *Wright-Fisher model with geometric seed-bank component*.

Age structure in seed bank

We require that seed bank and active population are of comparable size: $N = K \cdot M(N)$ for some $K \in (0, \infty)$, independent of N.

Thus, the individual seed bank age distribution is *geometric* with parameter $c/M = cK/N$.

Age structure in seed bank

We require that seed bank and active population are of comparable size: $N = K \cdot M(N)$ for some $K \in (0, \infty)$, independent of N.

Thus, the individual seed bank age distribution is *geometric* with parameter $c/M = cK/N$.

Frequency chains of the geometric seed-bank model

We now specialise (again) to the bi-allelic case $E = \{a, A\}$.

Definition 1.1 (Frequency chains)

Let

$$X_k^N := \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{\xi_k(i)=a\}} \quad \text{and} \quad Y_k^M := \frac{1}{M} \sum_{j=1}^{M} \mathbf{1}_{\{\eta_k(j)=a\}}, \quad k \in \mathbb{N}_0. \tag{2}$$

The pair forms a discrete-time Markov chain taking values in $I^N \times I^M$, where

$$I^N = \left\{0, \frac{1}{N}, \frac{2}{N}, \ldots, 1\right\} \quad \text{and} \quad I^M = \left\{0, \frac{1}{M}, \frac{2}{M}, \ldots, 1\right\}.$$

Let $\mathbb{P}_{x,y}$ be the law of (X^N, Y^M) with initial frequencies x, y.

The limiting generator of the allele frequency processes

For $c, K \in (0, \infty)$ assume

$$M = M(N) = \frac{N}{K}$$

Proposition 1.2

For the above parameter choices, and suitable test functions f, consider the discrete generator A^N of $(X_k^N, Y_k^M)_{k\in\mathbb{N}}$:

$$A^N f(x,y) := N\,\mathbb{E}_{x,y}\Big[f\big(X_1^N, Y_1^M\big) - f(x,y)\Big].$$

Then, we have

$$\begin{aligned} Af(x,y) &:= \lim_{N\to\infty} A^N f(x,y) \\ &= c(y-x)\frac{\partial f}{\partial x}(x,y) + cK(x-y)\frac{\partial f}{\partial y}(x,y) + \frac{1}{2}x(1-x)\frac{\partial f^2}{\partial x^2}(x,y). \end{aligned}$$

The scaling-limit of the allele frequency process

The result follows from the usual Taylor expansion of $A^N f$ about (x, y) (lengthy details omitted, tightness is standard). We arrive at the following

Corollary 1.3 (Wright-Fisher diffusion with seed bank, [BGCKWB16])

Under the conditions of Proposition 1.2,

$$(X^N_{\lfloor Nt \rfloor}, Y^N_{\lfloor Nt \rfloor})_{t \geq 0} \Rightarrow (X_t, Y_t)_{t \geq 0}$$

on $D_{[0,\infty)}([0,1]^2)$*, where* $(X_t, Y_t)_{t \geq 0}$ *is a 2-dimensional diffusion solving*

$$\begin{aligned} dX_t &= c(Y_t - X_t)dt + \sqrt{X_t(1 - X_t)}dB_t, \\ dY_t &= cK(X_t - Y_t)dt, \end{aligned} \tag{3}$$

with $X_0 = x$, $Y_0 = y$.

The dual of the seed-bank frequency process

The classical Wright-Fisher diffusion $\{Z_t\}$ is known to be (moment-) *dual* to the *block counting process* $\{N_t\}$ of the Kingman-coalescent:

$$\mathbb{E}_z\left[Z_t^n\right] = \mathbb{E}^n\left[z^{N_t}\right],$$

where $\{N_t\}$ is a pure death chain which jumps from $n \to n-1$ at rate $\binom{n}{2}$. Here, $Z_0 = z \in [0,1], N_0 = n \geq 1$.

Such dual processes are often extremely useful in the analysis of the underlying system, and it is easy to see that our Wright-Fisher diffusion with geometric seed-bank component also has a nice dual.

The dual of the seed-bank frequency process

The classical Wright-Fisher diffusion $\{Z_t\}$ is known to be (moment-) *dual* to the *block counting process* $\{N_t\}$ of the Kingman-coalescent:

$$\mathbb{E}_z\left[Z_t^n\right] = \mathbb{E}^n\left[z^{N_t}\right],$$

where $\{N_t\}$ is a pure death chain which jumps from $n \to n-1$ at rate $\binom{n}{2}$. Here, $Z_0 = z \in [0,1], N_0 = n \geq 1$.

Such dual processes are often extremely useful in the analysis of the underlying system, and it is easy to see that our Wright-Fisher diffusion with geometric seed-bank component also has a nice dual.

The dual of the seed-bank frequency process

Definition 1.4

We define the *block counting process of the seed-bank coalescent* $(N_t, M_t)_{t\geq 0}$ to be the continuous time Markov chain started in $(N_0, M_0) \in \mathbb{N}^2$ with transitions

$$\begin{aligned} (n,m) &\mapsto (n-1, m+1) \quad \text{at rate} \quad cn \\ (n,m) &\mapsto (n+1, m-1) \quad \text{at rate} \quad cKm \\ (n,m) &\mapsto (n-1, m) \quad \text{at rate} \quad \binom{n}{2} \end{aligned}$$

Denote by $\mathbb{P}^{n,m}$ the distribution of $(N_t, M_t)_{t\geq 0}$ if started in $(N_0, M_0) = (n, m)$, and denote the corresponding expected value by $\mathbb{E}^{n,m}$.

Moment duality

It is easy to see that *eventually*, $N_t + M_t = 1$ (as $t \to \infty$), since the sum $M + N$ can be dominated by a pure death process.

Moreover, it is standard to show that $(N_t, M_t)_{t \geq 0}$ is the *moment dual* of $(X_t, Y_t)_{t \geq 0}$.

Theorem 1.5

For every $(x, y) \in [0, 1]^2$ *and every* $n, m \in \mathbb{N}$,

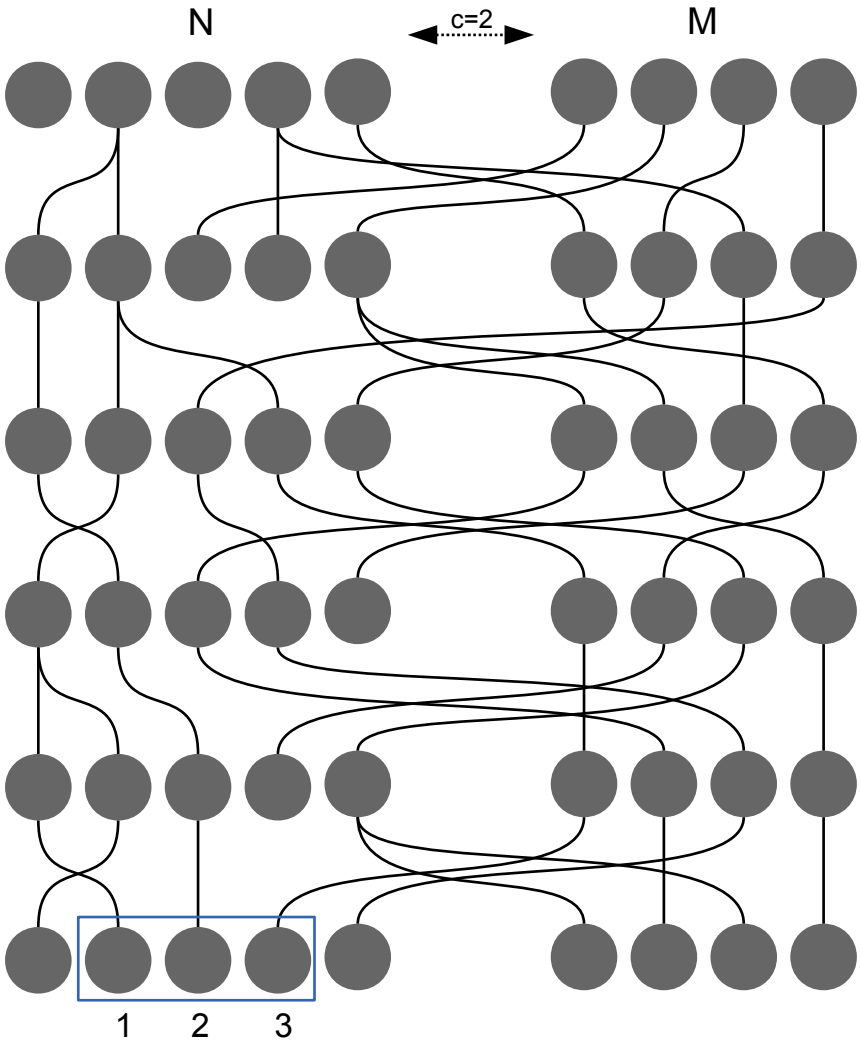
$$\mathbb{E}_{x,y}\left[X_t^n Y_t^m\right] = \mathbb{E}^{n,m}\left[x^{N_t} y^{M_t}\right].$$

The genealogy of a sample

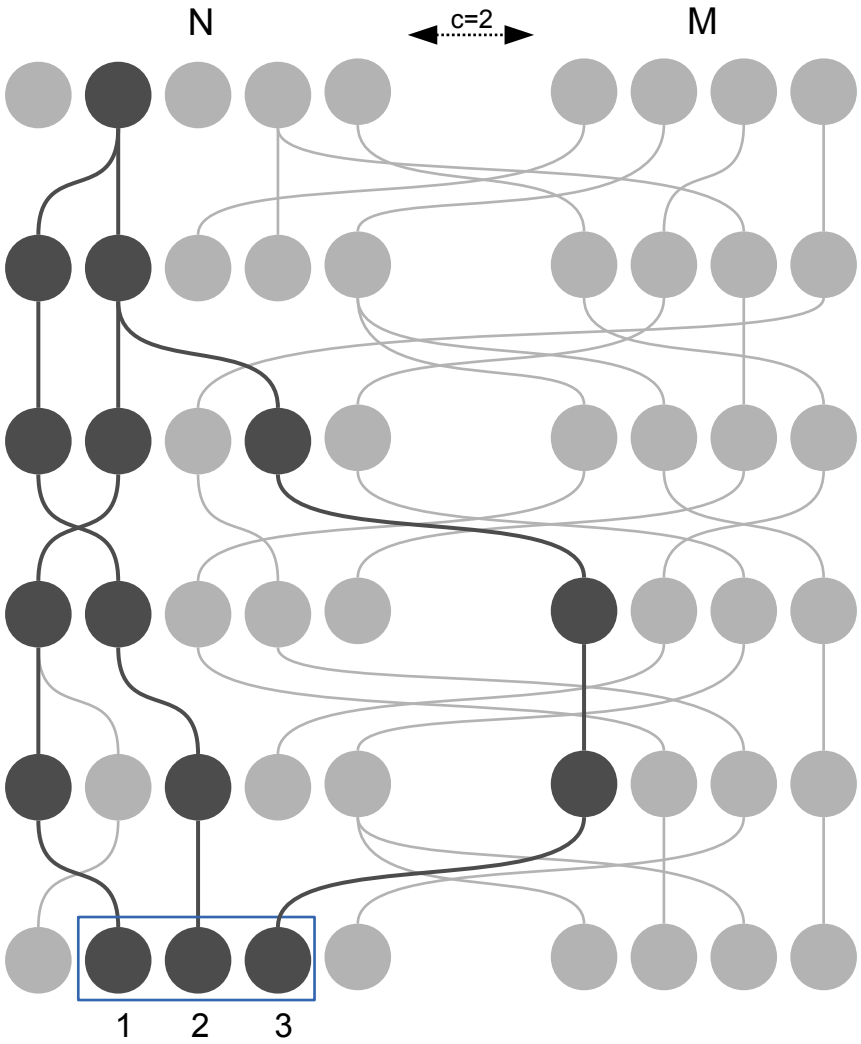
In view of the form of the block counting process, it is now easy to guess the stochastic process describing the limiting ancestral process of a sample taken from the Wright-Fisher model with geometric seed-bank component.

We now begin with its formal derivation.

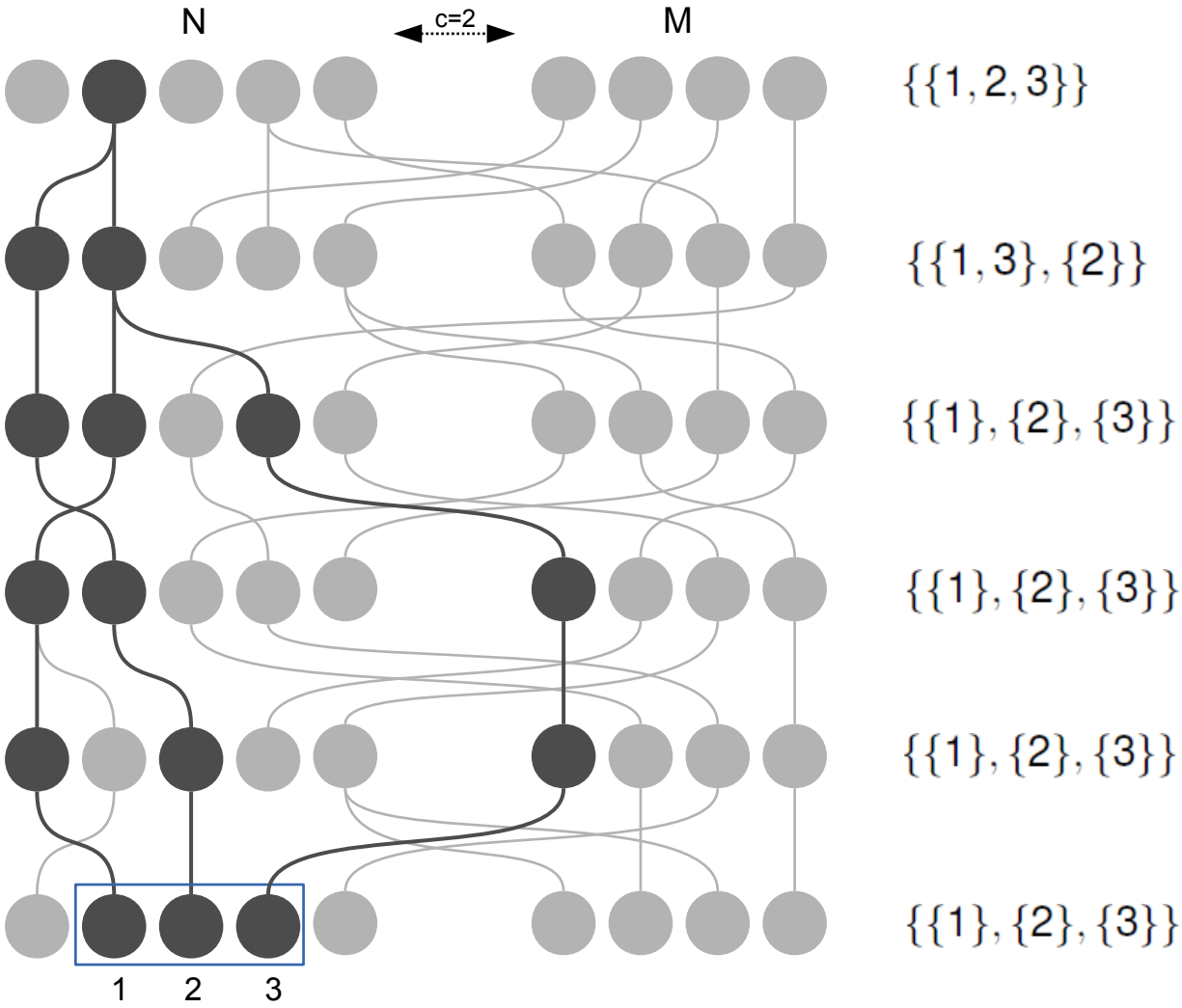
The genealogy of a sample



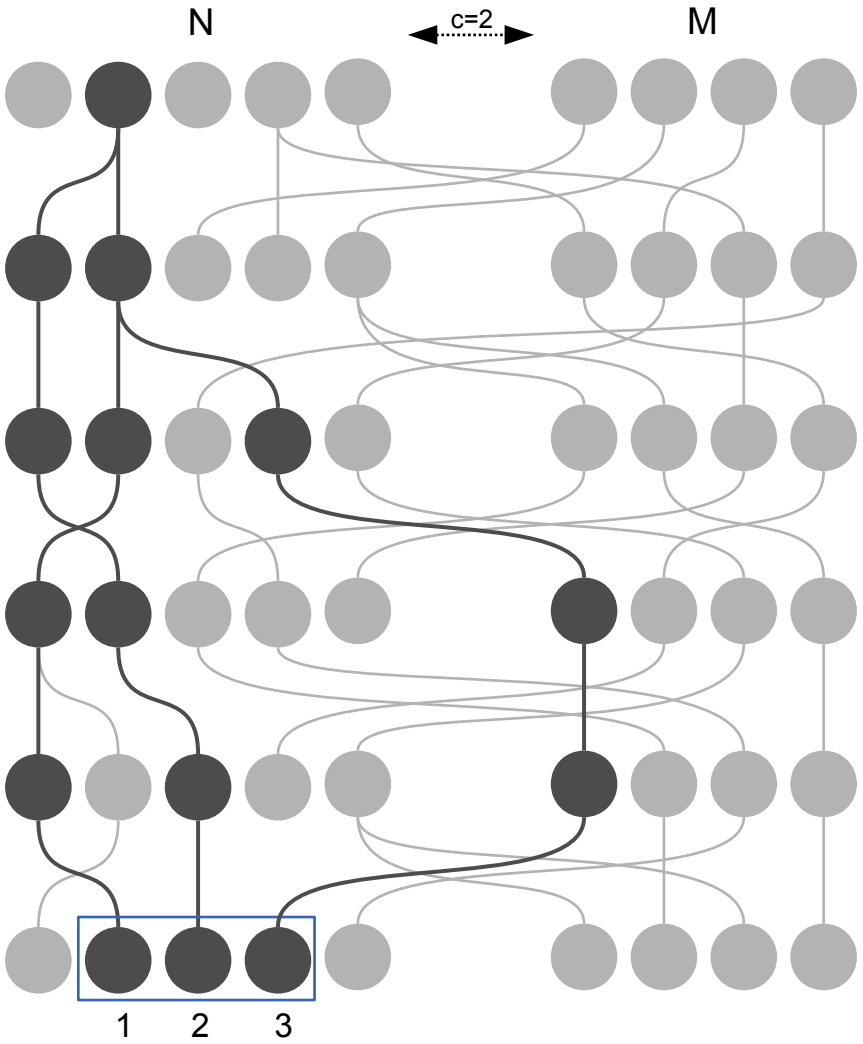
The genealogy of a sample



The genealogy of a sample



The genealogy of a sample



$\{\{1,2,3\}^p\}$

$\{\{1,3\}^p,\{2\}^p\}$

$\{\{1\}^p,\{2\}^p,\{3\}^p\}$

$\{\{1\}^p,\{2\}^p,\{3\}^s\}$

$\{\{1\}^p,\{2\}^p,\{3\}^s\}$

$\{\{1\}^p,\{2\}^p,\{3\}^p\}$

The genealogy of a sample and marked partitions

For $k \geq 1$, let $\mathcal{P}_k$ be the set of partitions of $[k]$. For $\pi \in \mathcal{P}_k$ let $|\pi|$ be the number of blocks of the partition π. We define a *marked partition* to be a partition in which each partition block carries a *flag* which can either be 'plant' or 'seed' (p or s).

For example, for $k = 5$, an element π of $\mathcal{P}_k^{\{p,s\}}$ is the marked partition

$$\pi = \big\{\{1,3\}^p\{2\}^s\{4,5\}^p\big\}.$$

The genealogy of a sample and partitions with flags, II

For two marked partitions $\pi, \pi' \in \mathcal{P}_k^{\{p,s\}}$ we write $\pi \succ \pi'$ if π' can be constructed by merging exactly two blocks of π both carrying the p-flag, and the resulting block in π' again carries a p-flag. For example

$$\big\{\{1,3\}^p\{2\}^s\{4,5\}^p\big\} \succ \big\{\{1,3,4,5\}^p\{2\}^s\big\}.$$

We use the notation $\pi \bowtie \pi'$ if π' can be constructed by switching the flag of precisely one block of π, for example

$$\big\{\{1,3\}^p\{2\}^s\{4,5\}^p\big\} \bowtie \big\{\{1,3\}^s\{2\}^s\{4,5\}^p\big\}.$$

The seed bank coalescent

Definition 1.6

For $k \geq 1$ and $c, K \in (0, \infty)$ we define the *seed bank coalescent* $(\Pi_t^{(k)})_{t \geq 0}$ with seed-bank intensity c and seed-bank size $1/K$ to be the continuous time pure-jump Markov process with values in $\mathcal{P}_k^{\{p,s\}}$, with transitions:

$$\begin{aligned}
\pi \to \pi' \quad & \text{at rate } 1 \text{ if } \pi \succ \pi', \\
\pi \to \pi' \quad & \text{at rate } c \text{ if } \pi \bowtie \pi' \text{ and one } p \text{ is replaced by one } s, \\
\pi \to \pi' \quad & \text{at rate } cK \text{ if } \pi \bowtie \pi' \text{ and one } s \text{ is replaced by one } p.
\end{aligned}$$

If $c = K = 1$, we speak of the *standard seed bank coalescent*.

The seed bank coalescent - Illustration

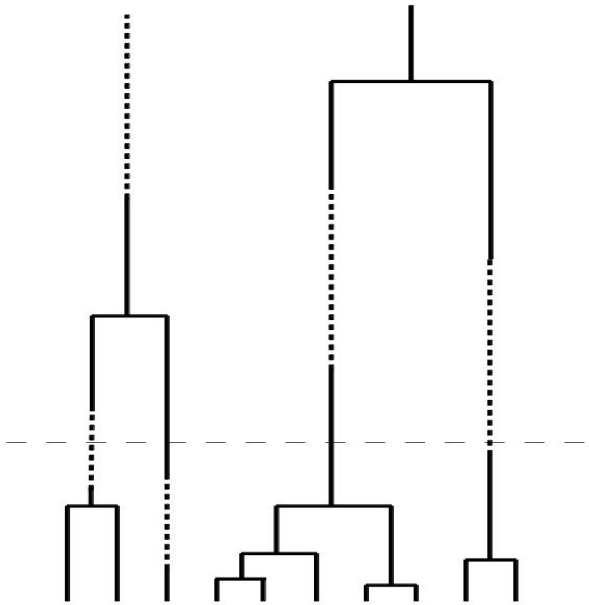


Figure: A possible realisation of the standard 10-seed bank coalescent. Dotted lines indicate 'inactive lineages' (carrying an s-flag, which are prohibited from merging). At the time marked with the dotted horizontal line the process is in state $\{\{1,2\}^s\{3\}^p\{4,5,6,7,8\}^p\{9,10\}^s\}$.

The seed bank coalescent as scaling limit

The seed bank coalescent appears as the limiting genealogy of a sample taken from the Wright-Fisher model with geometric seed-bank component in the same way as the Kingman coalescent in the classical Wright-Fisher model:

Corollary 1.7

Under the assumptions of Proposition 1.2, the ancestral process of an n-sample in the Wright Fisher model with geometric seed bank converges weakly as $N \to \infty$ to the seed-bank coalescent $(\Pi_t^{(n)})$ started with n plants.

Time to the most recent common ancestor

The seed bank causes a significant delay in the time to the most recent common ancestor.

Definition 1.8

Let $k \in \mathbb{N} \cup \{\infty\}$. We define the *time to the most recent common ancestor* of a sample of n plants to be

$$T(n) = \inf\{t > 0 : |\Pi_t^{(n)}| = 1 \text{ given that } \Pi_0 = \{\{1\}^p ... \{n\}^p\}\}.$$

Time to the most recent common ancestor

Theorem 1.9

For all $c, K \in (0, \infty)$, the seed bank coalescent satisfies

$$\mathbb{E}\big[T_{MRCA}[n]\big] \asymp \log\log n. \tag{4}$$

Recall the result for the Kingman coalescent, where

$$\mathbb{E}\big[T_{MRCA}[n]\big] = 2 - \frac{1}{n} \in [1, 2],$$

and a result for the *Bolthausen-Sznitman coalescent* in [Goldschmidt & Martin 2005], which exhibits the same time scale $\log\log n$ (but for a very different coalescent!).

Time to the most recent common ancestor

Theorem 1.9

For all $c, K \in (0, \infty)$, the seed bank coalescent satisfies

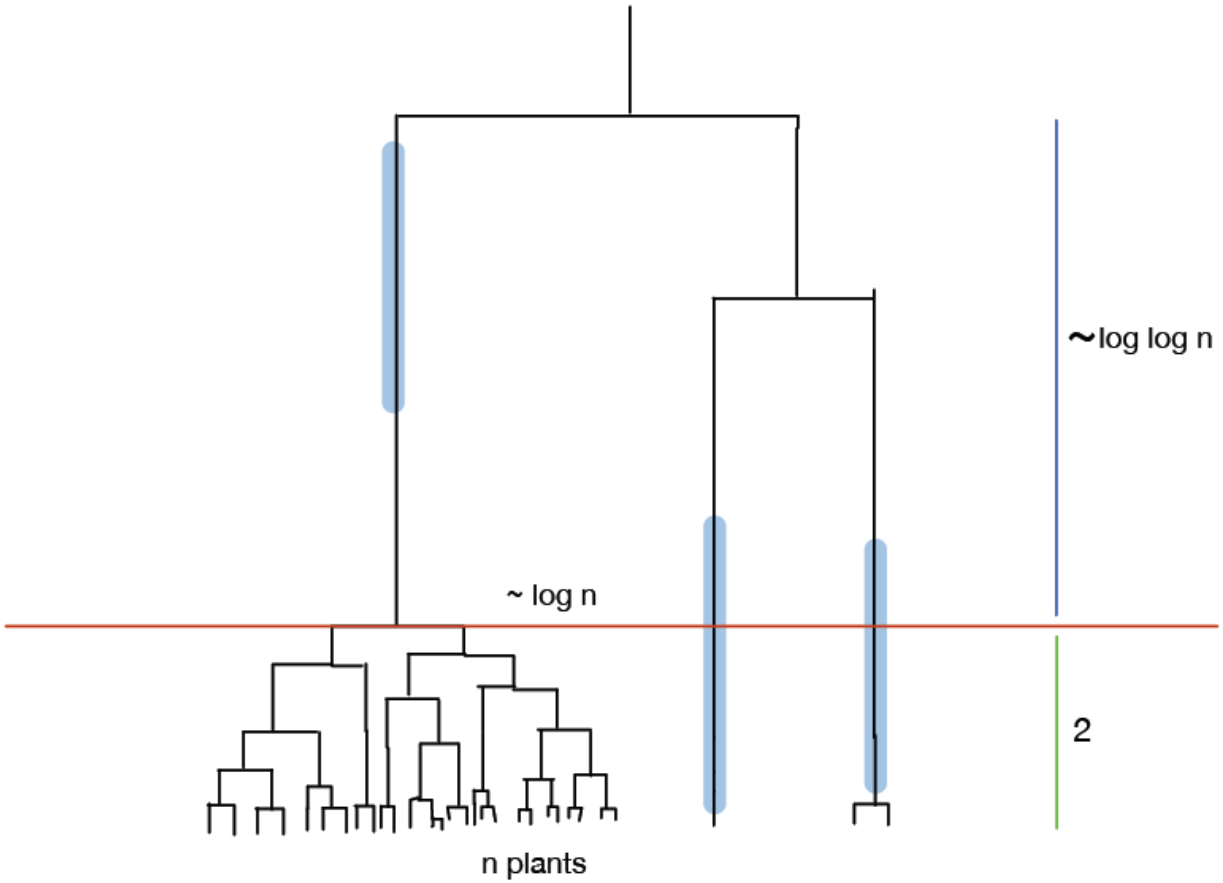
$$\mathbb{E}\big[T_{MRCA}[n]\big] \asymp \log\log n. \tag{4}$$

Recall the result for the Kingman coalescent, where

$$\mathbb{E}\big[T_{MRCA}[n]\big] = 2 - \frac{1}{n} \in [1, 2],$$

and a result for the *Bolthausen-Sznitman coalescent* in [Goldschmidt & Martin 2005], which exhibits the same time scale $\log\log n$ (but for a very different coalescent!).

Time to the most recent common ancestor



Relation to other models

The seed bank diffusion is related to two other interesting stochastic systems (work in progress with [E. Buzzoni, A. González)]:

- Seed bank diffusion scaling limit (with mutation) can be rewritten as a *stochastic delay differential equation*.
- Related to the two island model (where coalescences are completely blocked in one island) and the *structured coalescent* [Herbots 1997].

Characterization of stationary distribution (in presence of mutation)?

Relation to other models

The seed bank diffusion is related to two other interesting stochastic systems (work in progress with [E. Buzzoni, A. González)]:

- Seed bank diffusion scaling limit (with mutation) can be rewritten as a *stochastic delay differential equation*.
- Related to the two island model (where coalescences are completely blocked in one island) and the *structured coalescent* [Herbots 1997].

Characterization of stationary distribution (in presence of mutation)?

Relation to other models

The seed bank diffusion is related to two other interesting stochastic systems (work in progress with [E. Buzzoni, A. González)]:

- Seed bank diffusion scaling limit (with mutation) can be rewritten as a *stochastic delay differential equation*.
- Related to the two island model (where coalescences are completely blocked in one island) and the *structured coalescent* [Herbots 1997].

Characterization of stationary distribution (in presence of mutation)?

A seed bank stochastic delay differential equation

In the simplest case ($c = 1, K = 1$), started in $X_1(0) = x \in (0, 1)$, the seed bank diffusion is eqivalent to the one-dimensional sdde

$$\mathrm{d}X_1(t) = \left[\int_0^t e^{-(t-s)} X_1(s) \mathrm{d}s + x e^{-t} - X_1(t) \right] \mathrm{d}t + \sqrt{X_1(t)(1 - X_1(t))} \mathrm{d}B_t.$$

Interpretation: The type of any 'infinitesimal' currently resuscitated individual can be determined from the active population an exponentially distributed time ago.

Limiting behaviour

It is interesting to see if one can recover classical models if the seed bank switching rate c approaches 0 or infinity.

- If $c \to \infty$ (instantaneous switching), the genealogy converges to a classical Kingman coalescent, linearly delayed by a factor of $(K/(K+1))^2$.
- If $c \to 0$, then the active population is described by a classical Kingman coalescent while the seeds stay completely inactive.
- However, if one speeds up time by a factor of N and scales the switching rate by $c = 1/N$ (thus effectively moving to a 'fast' timescale), then a new object arises, describing the transmission of 'rare ancient material'.

Limiting behaviour

It is interesting to see if one can recover classical models if the seed bank switching rate c approaches 0 or infinity.

- If $c \to \infty$ (instantaneous switching), the genealogy converges to a classical Kingman coalescent, linearly delayed by a factor of $(K/(K+1))^2$.
- If $c \to 0$, then the active population is described by a classical Kingman coalescent while the seeds stay completely inactive.
- However, if one speeds up time by a factor of N and scales the switching rate by $c = 1/N$ (thus effectively moving to a 'fast' timescale), then a new object arises, describing the transmission of 'rare ancient material'.

Limiting behaviour

It is interesting to see if one can recover classical models if the seed bank switching rate c approaches 0 or infinity.

- If $c \to \infty$ (instantaneous switching), the genealogy converges to a classical Kingman coalescent, linearly delayed by a factor of $(K/(K+1))^2$.
- If $c \to 0$, then the active population is described by a classical Kingman coalescent while the seeds stay completely inactive.
- However, if one speeds up time by a factor of N and scales the switching rate by $c = 1/N$ (thus effectively moving to a 'fast' timescale), then a new object arises, describing the transmission of 'rare ancient material'.

The ancient material process

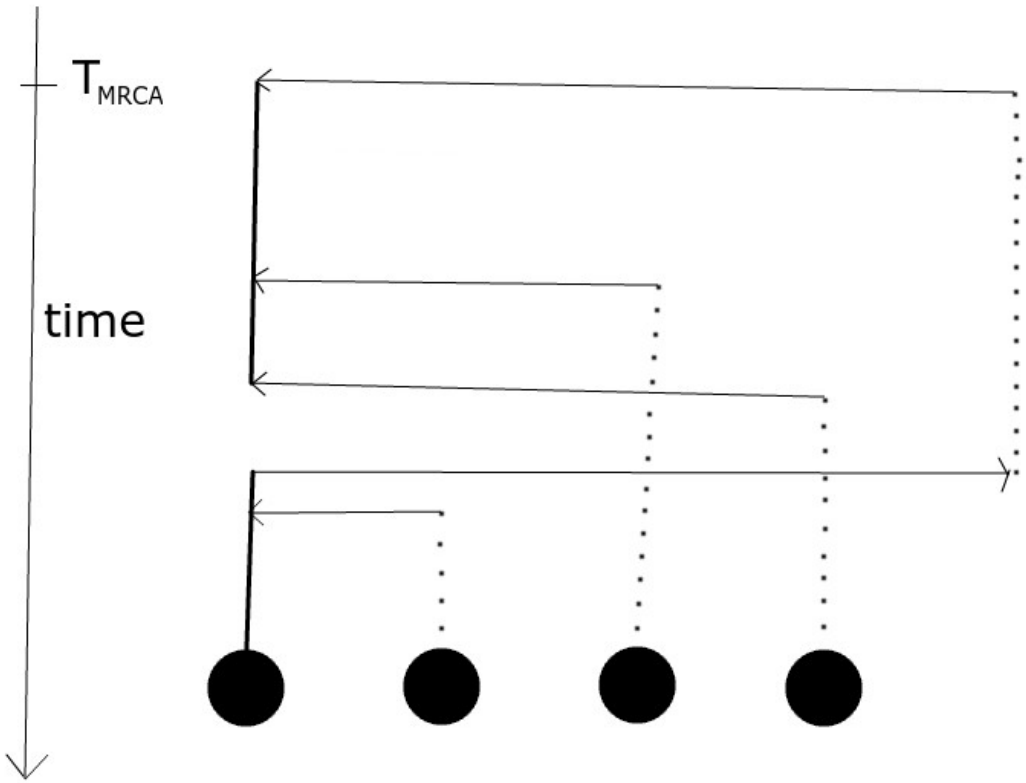
- The 'ancient material process' has a block counting process that can be defined as the time-continuous Markov chain $(N(t), M(t))$ with state space $\{0,1\} \times \mathbb{N}$.

  For $n_2 \in \mathbb{N}$ it jumps from the state $(1, n_2)$ to $(1, n_2 - 1)$ at rate Kn_2, from $(1, n_2)$ to the state $(0, n_2 - 1)$ at rate 1, and from the state $(0, n_2)$ to the state $(1, n_2 - 1)$ at rate Kn_2, all other transition rates being zero.
- This process can be extende to the state space $\mathbb{N} \times \mathbb{N}$ so that for every initial condition $(N(0), M(0)) = (n_1, n_2) \in \mathbb{N} \times \mathbb{N}$,

$$(N(0+), M(0+)) = (1_{n_1>0}, n_2)$$

  holds almost surely.
- That is, in this new coalescent-related object all possible plant-coalescences happen instantaneously, while switches appear as usual.

The ancient material process



... thank you for your attention!

Talk mostly based on:

- Blath, González Casanova, Kurt, Wilke Berenguer: A new coalescent for seed bank models, *Annals of Applied Probability*, 2016
- Blath, Eldon, González Casanova, Kurt, Wilke Berenguer: Genetic variability under the seed bank coalescent, *Genetics*, 2015
- Blath, Buzzoni, González Casanova, *Manuscript in progress*, 2016

The same coalescent structure has been derived independently in a metapopulation context in ecology:

- Lambert, Ma: The coalescent in peripatric metapopulations, *Journal of Applied Probability*, 2015