# Some mathematical models from population genetics

Alison Etheridge
University of Oxford

with thanks to numerous collaborators, especially Nick Barton, IST Austria

CMAP5, May 2023

# What we have so far: Wright-Fisher/Moran models

In time units of $N_e$ generations, $p =$ proportion $a$-alleles

- (Forwards time) The Wright-Fisher diffusion (with and without selection)

$$dp_t = -sp_t(1-p_t)dt + \sqrt{p_t(1-p_t)}dW_t;$$

- (Backwards time) The Kingman coalescent/ ASG

$$n_t \mapsto n_t - 1 \text{ at rate } \binom{n_t}{2}, \quad n_t \mapsto n_t + 1 \text{ at rate } sn_t;$$

- Sampling probabilities

$$\mathbb{E}[p(t)^{n(0)}] = \mathbb{E}[p(0)^{n(t)}]$$

Stronger result holds. Kingman coalescent really describes genealogy of random sample from (neutral) population.

# Adding spatial structure: subdivided populations

Population subdivided into demes = islands = colonies

- Vertices of graph, $i \in I$;
- $i \sim j$ if $i$, $j$ neighbours
- $N_i =$ population size in deme $i$

**Structured Wright-Fisher model**

Reproduction in discrete generations

- neutral Wright-Fisher within each deme
- proportion $m_{ij}$ of individuals in deme $i$ migrate to deme $j$

$$N_i \sum_{j \sim i} m_{ij} = \sum_{j \sim i} N_j m_{ji}$$

# Genealogy of structured Wright-Fisher model

1. Two lineages sampled from deme $i$

$$\mathbb{P}\big[\text{coalesce in } j \neq i \text{ in previous generation}\big] = \frac{\binom{m_{ji}N_j}{2}}{\binom{N_i}{2}} \frac{1}{N_j}$$

$$\mathbb{P}\big[\text{coalesce in } i \text{ in previous generation}\big] = \frac{\binom{N_i - \sum_{j\sim i} m_{ji}N_j}{2}}{\binom{N_i}{2}} \frac{1}{N_i}$$

# Genealogy of structured Wright-Fisher model

1. Two lineages sampled from deme $i$

$$\mathbb{P}\big[\text{coalesce in } j \neq i \text{ in previous generation}\big] = \frac{\binom{m_{ji}N_j}{2}}{\binom{N_i}{2}} \frac{1}{N_j}$$

$$\mathbb{P}\big[\text{coalesce in } i \text{ in previous generation}\big] = \frac{\binom{N_i - \sum_{j\sim i} m_{ji}N_j}{2}}{\binom{N_i}{2}} \frac{1}{N_i}$$

2. Two lineages sampled from demes $i \neq j$

$$\mathbb{P}\big[\text{coalesce in } k \notin \{i,j\} \text{ in previous generation}\big] = \frac{m_{ki}N_k}{N_i} \frac{m_{kj}N_k}{N_j} \frac{1}{N_k}$$

$$\mathbb{P}\big[\text{coalesce in } j \text{ in previous generation}\big] = \frac{m_{ji}N_j}{N_i} \frac{\left(N_j - \sum_{l\sim j} m_{lj}N_l\right)}{N_j} \frac{1}{N_j}$$

# Scaling limit: the structured coalescent

- $N_i = O(N)$ (large)
- $m_{ij} = O(1/N)$

# Scaling limit: the structured coalescent

- $N_i = O(N)$ (large)
- $m_{ij} = O(1/N)$

$$\mathbb{P}\big[\text{simultaneous migration and coalescence}\big] = O(1/N^2)$$

$$\mathbb{P}\big[\text{simultaneous or multiple mergers}\big] = O(1/N^2)$$

$$\mathbb{P}\big[\text{single lineage at } i \text{ migrates}\big] = \sum_{j \sim i} \frac{m_{ji} N_j}{N_i} = O(1/N)$$

# Scaling limit: the structured coalescent

- $N_i = O(N)$ (large)
- $m_{ij} = O(1/N)$

$$\mathbb{P}\big[\text{simultaneous migration and coalescence}\big] = O(1/N^2)$$

$$\mathbb{P}\big[\text{simultaneous or multiple mergers}\big] = O(1/N^2)$$

$$\mathbb{P}\big[\text{single lineage at } i \text{ migrates}\big] = \sum_{j\sim i} \frac{m_{ji}N_j}{N_i} = O(1/N)$$

**The structured coalescent** $\underline{n} = (n_i)_{i\in I}$:

- $\begin{cases} n_i \mapsto n_i - 1 \\ n_j \mapsto n_j + 1 \end{cases}$ at rate $n_i \frac{N_e(j)}{N_e(i)} m_{ji}$
- $n_i \mapsto n_i - 1$ at rate $\frac{1}{2N_e(i)} n_i\,(n_i - 1)$

# Scaling limit: the structured coalescent

- $N_i = O(N)$ (large)
- $m_{ij} = O(1/N)$

$$\mathbb{P}\big[\text{simultaneous migration and coalescence}\big] = O(1/N^2)$$

$$\mathbb{P}\big[\text{simultaneous or multiple mergers}\big] = O(1/N^2)$$

$$\mathbb{P}\big[\text{single lineage at } i \text{ migrates}\big] = \sum_{j\sim i} \frac{m_{ji}N_j}{N_i} = O(1/N)$$

**The structured coalescent** $\underline{n} = (n_i)_{i\in I}$:

- $\begin{cases} n_i \mapsto n_i - 1 \\ n_j \mapsto n_j + 1 \end{cases}$ at rate $n_i \frac{N_e(j)}{N_e(i)} m_{ji}$
- $n_i \mapsto n_i - 1$ at rate $\frac{1}{2N_e(i)} n_i\,(n_i - 1)$

Ancestral lineages drawn into more populous demes

# Forwards in time?

$N_i \sum_{j\sim i} m_{ij} = \sum_{j\sim i} m_{ji} N_j$, $m_{ij} = O(1/N)$

Alleles $a$, $A$. $p_i(t) =$ proportion of type $a$ in deme $i$ at time $t$

$\Delta p_i$ change across single generation

# Forwards in time?

$N_i \sum_{j\sim i} m_{ij} = \sum_{j\sim i} m_{ji} N_j$, $m_{ij} = O(1/N)$

Alleles $a$, $A$. $p_i(t) =$ proportion of type $a$ in deme $i$ at time $t$

$\Delta p_i$ change across single generation

$$\begin{aligned}
\mathbb{E}\big[\Delta p_i\big] &= \frac{1}{N_i}\Big(\big(1-\sum_{j\sim i} m_{ij}\big)N_i p_i + \sum_{j\sim i} m_{ji} N_j p_j\Big) - p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i} p_j - \frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i
\end{aligned}$$

# Forwards in time?

$N_i \sum_{j\sim i} m_{ij} = \sum_{j\sim i} m_{ji} N_j$, $m_{ij} = O(1/N)$

Alleles $a$, $A$. $p_i(t) =$ proportion of type $a$ in deme $i$ at time $t$

$\Delta p_i$ change across single generation

$$\begin{aligned}
\mathbb{E}\big[\Delta p_i\big] &= \frac{1}{N_i}\Big(\big(1-\sum_{j\sim i} m_{ij}\big)N_i p_i + \sum_{j\sim i} m_{ji} N_j p_j\Big) - p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i} p_j - \frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i
\end{aligned}$$

$$\frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i = \frac{1}{N_i}\sum_{j\sim i} N_j m_{ji} p_i$$

# Forwards in time?

$N_i \sum_{j\sim i} m_{ij} = \sum_{j\sim i} m_{ji} N_j$, $m_{ij} = O(1/N)$

Alleles $a$, $A$. $p_i(t) =$ proportion of type $a$ in deme $i$ at time $t$

$\Delta p_i$ change across single generation

$$\begin{aligned}
\mathbb{E}\big[\Delta p_i\big] &= \frac{1}{N_i}\Big(\big(1-\sum_{j\sim i} m_{ij}\big)N_i p_i + \sum_{j\sim i} m_{ji} N_j p_j\Big) - p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i} p_j - \frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i}\big(p_j - p_i\big)
\end{aligned}$$

$$\frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i = \frac{1}{N_i}\sum_{j\sim i} N_j m_{ji} p_i$$

# Forwards in time?

$N_i \sum_{j\sim i} m_{ij} = \sum_{j\sim i} m_{ji} N_j$, $m_{ij} = O(1/N)$

Alleles $a$, $A$. $p_i(t) =$ proportion of type $a$ in deme $i$ at time $t$

$\Delta p_i$ change across single generation

$$\begin{aligned}
\mathbb{E}\big[\Delta p_i\big] &= \frac{1}{N_i}\Big(\big(1-\sum_{j\sim i} m_{ij}\big)N_i p_i + \sum_{j\sim i} m_{ji} N_j p_j\Big) - p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i} p_j - \frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i}\big(p_j - p_i\big)
\end{aligned}$$

$\mathbb{E}\big[(\Delta p_i)^2\big] = \frac{1}{N_i}\big(p_i(1-p_i)+O(1/N)\big)$ $\qquad$ $\mathtt{Cov}\big(\Delta p_i, \Delta p_j\big) = O(1/N^2)$

# Forwards in time?

$N_i \sum_{j\sim i} m_{ij} = \sum_{j\sim i} m_{ji} N_j$, $m_{ij} = O(1/N)$

Alleles $a$, $A$. $p_i(t) =$ proportion of type $a$ in deme $i$ at time $t$

$\Delta p_i$ change across single generation

$$\begin{aligned}
\mathbb{E}\big[\Delta p_i\big] &= \frac{1}{N_i}\Big(\big(1-\sum_{j\sim i} m_{ij}\big)N_i p_i + \sum_{j\sim i} m_{ji} N_j p_j\Big) - p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i} p_j - \frac{1}{N_i}\sum_{j\sim i} m_{ij} N_i p_i \\
&= \sum_{j\sim i} m_{ji}\frac{N_j}{N_i}\big(p_j - p_i\big)
\end{aligned}$$

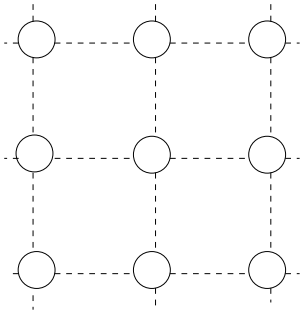
$\mathbb{E}\big[(\Delta p_i)^2\big] = \frac{1}{N_i}\big(p_i(1-p_i)+O(1/N)\big)$ $\qquad$ $\mathtt{Cov}\big(\Delta p_i, \Delta p_j\big) = O(1/N^2)$

As $N \to \infty$ recover a system of diffusions coupled through migration

# Kimura's stepping stone model

$\sum_j N_e(i)m_{ij} = \sum_j N_e(j)m_{ji}$

$$dp_i = \sum_j \frac{N_e(j)}{N_e(i)} m_{ji}(p_j - p_i)dt + \sqrt{\frac{1}{N_e(i)} p_i(1-p_i)}\, dW_i$$



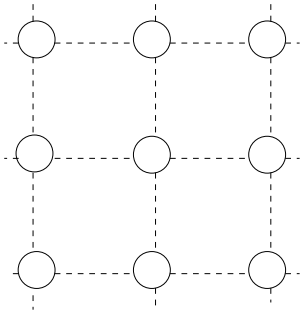
$\{W_i\}_{i \in I}$ independent Brownian motions

System of W-F diffusions coupled through migration

# Kimura's stepping stone model

$\sum_j N_e(i)m_{ij} = \sum_j N_e(j)m_{ji}$

$$dp_i = \sum_j \frac{N_e(j)}{N_e(i)} m_{ji}(p_j - p_i)dt + \sqrt{\frac{1}{N_e(i)} p_i(1-p_i)}\, dW_i$$



$\{W_i\}_{i \in I}$ independent Brownian motions

System of W-F diffusions coupled through migration

The structured coalescent $\underline{n}$:

- $\begin{cases} n_i \mapsto n_i - 1 \\ n_j \mapsto n_j + 1 \end{cases}$ at rate $n_i \frac{N_e(j)}{N_e(i)} m_{ji}$
- $n_i \mapsto n_i - 1$ at rate $\frac{1}{2N_e(i)} n_i \left(n_i - 1\right)$

# Duality

*for simplicity* $N_i \equiv N_e$

$$dp_i = \sum_j m_{ji}(p_j - p_i)dt + \sqrt{\tfrac{1}{N_e} p_i(1-p_i)}\, dW_i \qquad \underline{p^n} := \prod_{i \in I} p_i^{n_i}.$$

# Duality

*for simplicity* $N_i \equiv N_e$

$$dp_i = \sum_j m_{ji}(p_j - p_i)dt + \sqrt{\tfrac{1}{N_e} p_i(1-p_i)}\, dW_i \qquad \underline{p}^{\underline{n}} := \prod_{i \in I} p_i^{n_i}.$$

$$\begin{aligned} d\underline{p}^{\underline{n}} &= \sum_i n_i \underline{p}^{\underline{n}-\underline{e}_i} \sum_j m_{ji}(p_j - p_i)dt \\ &\quad + \sum_i \frac{1}{N_e} \frac{n_i(n_i-1)}{2} \underline{p}^{\underline{n}-2\underline{e}_i} p_i(1-p_i)dt + \text{ martingale term} \end{aligned}$$

# Duality

*for simplicity* $N_i \equiv N_e$

$dp_i = \sum_j m_{ji}(p_j - p_i)dt + \sqrt{\frac{1}{N_e} p_i(1-p_i)}\, dW_i \qquad \underline{p}^{\underline{n}} := \prod_{i \in I} p_i^{n_i}.$

$$
\begin{aligned}
d\underline{p}^{\underline{n}} &= \sum_i n_i \underline{p}^{\underline{n}-\underline{e}_i} \sum_j m_{ji}(p_j - p_i)dt \\
&\quad + \sum_i \frac{1}{N_e} \frac{n_i(n_i-1)}{2} \underline{p}^{\underline{n}-2\underline{e}_i} p_i(1-p_i)dt + \text{ martingale term} \\
&= \sum n_i \sum_j m_{ji} \left( \underline{p}^{\underline{n}+\underline{e}_j-\underline{e}_i} - \underline{p}^{\underline{n}} \right) dt \\
&\quad + \sum_i \frac{1}{N_e} \binom{n_i}{2} \left( \underline{p}^{\underline{n}-\underline{e}_i} - \underline{p}^{\underline{n}} \right) dt + \text{ martingale term}
\end{aligned}
$$

# Duality

*for simplicity* $N_i \equiv N_e$

$dp_i = \sum_j m_{ji}(p_j - p_i)dt + \sqrt{\frac{1}{N_e} p_i(1-p_i)}dW_i \qquad \underline{p}^{\underline{n}} := \prod_{i \in I} p_i^{n_i}.$

$$
\begin{aligned}
d\underline{p}^{\underline{n}} &= \sum_i n_i \underline{p}^{\underline{n}-\underline{e}_i} \sum_j m_{ji}(p_j - p_i)dt \\
&\quad + \sum_i \frac{1}{N_e} \frac{n_i(n_i-1)}{2} \underline{p}^{\underline{n}-2\underline{e}_i} p_i(1-p_i)dt + \text{ martingale term} \\
&= \sum n_i \sum_j m_{ji} \left(\underline{p}^{\underline{n}+\underline{e}_j-\underline{e}_i} - \underline{p}^{\underline{n}}\right)dt \\
&\quad + \sum_i \frac{1}{N_e} \binom{n_i}{2} \left(\underline{p}^{\underline{n}-\underline{e}_i} - \underline{p}^{\underline{n}}\right)dt + \text{ martingale term}
\end{aligned}
$$

$\underline{n} \mapsto \underline{n} + \underline{e}_j - \underline{e}_i$ at rate $n_i m_{ji}$

$\underline{n} \mapsto \underline{n} - \underline{e}_i$ at rate $\frac{1}{N_e}\binom{n_i}{2}$

$$\frac{d}{du}\mathbb{E}\left[\underline{p}_{\underline{u}}^{\underline{n}_{t-u}}\right] = 0$$

$$\mathbb{E}\left[\underline{p}_{\underline{t}}^{\underline{n}_0}\right] = \mathbb{E}\left[\underline{p}_{\underline{0}}^{\underline{n}_t}\right].$$

# Interpretation

$$\mathbb{E}\left[\underline{p}_t^{n_0}\right] = \mathbb{E}\left[\underline{p}_0^{n_t}\right].$$

- Sample $n_i(0)$ individuals from deme $i$, $\sum_i n_i(0) < \infty$,
- Probability all type $a$ is $\mathbb{E}\left[\underline{p}_0^{n_t}\right]$

# Interpretation

$$\mathbb{E}\left[\underline{p}_t^{n_0}\right] = \mathbb{E}\left[\underline{p}_0^{n_t}\right].$$

- Sample $n_i(0)$ individuals from deme $i$, $\sum_i n_i(0) < \infty$,
- Probability all type $a$ is $\mathbb{E}\left[\underline{p}_0^{n_t}\right]$

**Example** Suppose $I = \mathbb{Z}^2$
For any finite sample, eventually $\underline{n}_t$ is a singleton, so all individuals in the sample are of the same type.

# Interpretation

$$\mathbb{E}\left[\underline{p}_t^{n_0}\right] = \mathbb{E}\left[\underline{p}_0^{n_t}\right].$$

- Sample $n_i(0)$ individuals from deme $i$, $\sum_i n_i(0) < \infty$,
- Probability all type $a$ is $\mathbb{E}\left[\underline{p}_0^{n_t}\right]$

**Example** Suppose $I = \mathbb{Z}^2$
For any finite sample, eventually $\underline{n}_t$ is a singleton, so all individuals in the sample are of the same type.

Need to account for mutation in our model

# Adding mutation

Simplest example:

- Infinitely many alleles model of mutation: each individual in each generation, independently, with small probability $\mu$ mutates to a type never before seen in the population
- *Probability of identity by descent* of two individuals, $F$, $=$ probability no mutation since most recent common ancestor (MRCA)
- Equivalently $F = (1-2\mu)^T \approx \exp(-2\mu T)$ is the Laplace transform of the distribution of the time to the MRCA.

The neutral mutation rate dictates the timescales over which we can reconstruct information about genealogies.

# Isolation by distance

In a population in which individuals typically migrate to geographically close subpopulations, and new mutations continuously accumulate, $\mathbb{P}[\text{two individuals in same allelic state}]$ declines with increasing separation.

Isolation by distance (Wright 1943)

# Isolation by distance

In a population in which individuals typically migrate to geographically close subpopulations, and new mutations continuously accumulate, $\mathbb{P}[\text{two individuals in same allelic state}]$ declines with increasing separation.

Isolation by distance (Wright 1943)

In $\mathbb{Z}$ with nearest neighbour migration there is an explicit expression for the probability of identity. It declines exponentially with distance. But the exact formula is very special.

# Probability of identity in subdivided population

Population on $\mathbb{Z}^2$, $N$ individuals per deme, discrete generations

- Reproduction according to Wright-Fisher model in each deme;
- Proportion $g_1(x-y)$ of offspring in deme $x$ migrate to deme $y$.

$T=$ time to MRCA of sample of size two

$$F(x) = \mathbb{E}_x[(1-2\mu)^T]$$

($x$ vector in $\mathbb{Z}^2$)

$$\psi_t(x) = \mathbb{P}_x[T=t],$$

$$\psi_1(x) = \frac{G_1(x)}{N}, \quad G_1(x) = \int g_1(x,z)g_1(0,z)dz.$$

# Calculating $F(x)$

If $t > 1$, partition over location immediate ancestors

$$\psi_t(x) = \sum_y G_1(x-y)\psi_{t-1}(y) - \frac{1}{N} G_1(x)\psi_{t-1}(0).$$

Then

$$(*) \qquad \psi_t(x) = \frac{1}{N}\left( G_t(x) - \sum_{\tau=1}^{t-1} G_{t-\tau}(x)\psi_\tau(0) \right)$$

# Calculating $F(x)$

If $t > 1$, partition over location immediate ancestors

$$\psi_t(x) = \sum_y G_1(x-y)\psi_{t-1}(y) - \frac{1}{N} G_1(x)\psi_{t-1}(0).$$

Then

$$(*) \qquad \psi_t(x) = \frac{1}{N}\left(G_t(x) - \sum_{\tau=1}^{t-1} G_{t-\tau}(x)\psi_\tau(0)\right)$$

$\widetilde{G}(z,x) = \sum_{t=1}^{\infty} G_t(x) z^t$, discrete Laplace transform of $G$

# Calculating $F(x)$

If $t > 1$, partition over location immediate ancestors

$$\psi_t(x) = \sum_y G_1(x-y)\psi_{t-1}(y) - \frac{1}{N} G_1(x)\psi_{t-1}(0).$$

Then

$$(*) \qquad \psi_t(x) = \frac{1}{N}\left( G_t(x) - \sum_{\tau=1}^{t-1} G_{t-\tau}(x)\psi_\tau(0) \right)$$

$\widetilde{G}(z,x) = \sum_{t=1}^{\infty} G_t(x) z^t$, discrete Laplace transform of $G$

Write $\phi(z,x) = \mathbb{E}_x[z^T]$,

$$(\dagger) \qquad \phi(z,x) = \frac{1}{N}\widetilde{G}(z,x)\Big(1 - \phi(z,0)\Big)$$

(convolution $\rightarrow$ product under LT)

# Calculating $F(x)$

$\phi(z,x) = \mathbb{E}_x[z^T]$,

$$(\dagger) \qquad \phi(z,x) = \frac{1}{N}\widetilde{G}(z,x)\Big(1-\phi(z,0)\Big)$$

Set $x = 0$ in $(\dagger)$,

$$\phi(z,0) = \frac{1}{N}\widetilde{G}(z,0)\big(1-\phi(z,0)\big) \rightsquigarrow \qquad \phi(z,0) = \frac{\widetilde{G}(z,0)}{N+\widetilde{G}(z,0)}$$

# Calculating $F(x)$

$\phi(z,x) = \mathbb{E}_x[z^T]$,

$$(\dagger) \qquad \phi(z,x) = \frac{1}{N}\widetilde{G}(z,x)\Big(1-\phi(z,0)\Big)$$

Set $x = 0$ in $(\dagger)$,

$$\phi(z,0) = \frac{1}{N}\widetilde{G}(z,0)\big(1-\phi(z,0)\big) \rightsquigarrow \qquad \phi(z,0) = \frac{\widetilde{G}(z,0)}{N+\widetilde{G}(z,0)}$$

Substitute back in $(\dagger)$

$$\phi(z,x) = \frac{\widetilde{G}(z,x)}{N+\widetilde{G}(z,0)}$$

# Calculating $F(x)$

$\phi(z,x) = \mathbb{E}_x[z^T]$,

$$(\dagger) \qquad \phi(z,x) = \frac{1}{N}\widetilde{G}(z,x)\Big(1-\phi(z,0)\Big)$$

Set $x = 0$ in $(\dagger)$,

$$\phi(z,0) = \frac{1}{N}\widetilde{G}(z,0)\big(1-\phi(z,0)\big) \rightsquigarrow \quad \phi(z,0) = \frac{\widetilde{G}(z,0)}{N+\widetilde{G}(z,0)}$$

Substitute back in $(\dagger)$

$$\phi(z,x) = \frac{\widetilde{G}(z,x)}{N+\widetilde{G}(z,0)}$$

If $g_1$ approximately Gaussian

$$\frac{1}{N}\widetilde{G}(z,0) = \frac{1}{2\mathcal{N}}\log\Big(\frac{1}{\sqrt{1-z}}\Big); \quad \frac{1}{N}\widetilde{G}(z,x) = \frac{1}{\mathcal{N}}K_0\Big(\frac{|x|}{\sigma}\sqrt{1-z}\Big)$$

$\mathcal{N} = 2N\pi\sigma^2$ is Wright's neighbourhood size, $K_0$ modified Bessel function of second kind of degree zero.

# Calculating $F(x)$

Have shown

$$\phi(z,x) \approx \frac{K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right)}{\mathcal{N} - \log(\sqrt{1-z})}$$

# Calculating $F(x)$

Have shown

$$\phi(z,x) \approx \frac{K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right)}{\mathcal{N} - \log(\sqrt{1-z})}$$

DIVERGES as $x \to 0$

# Calculating $F(x)$

Have shown

$$\phi(z,x) \approx \frac{K_0\left(\frac{|x|}{\sigma}\sqrt{1-z}\right)}{\mathcal{N} - \log(\sqrt{1-z})} \quad \text{DIVERGES as } x \to 0$$

Assume solution constant over small scale $\kappa$, use $K_0(y) \approx -\log y$ as $y \downarrow 0$, set $z = 1 - 2\mu \approx \exp(-2\mu)$ and substitute:

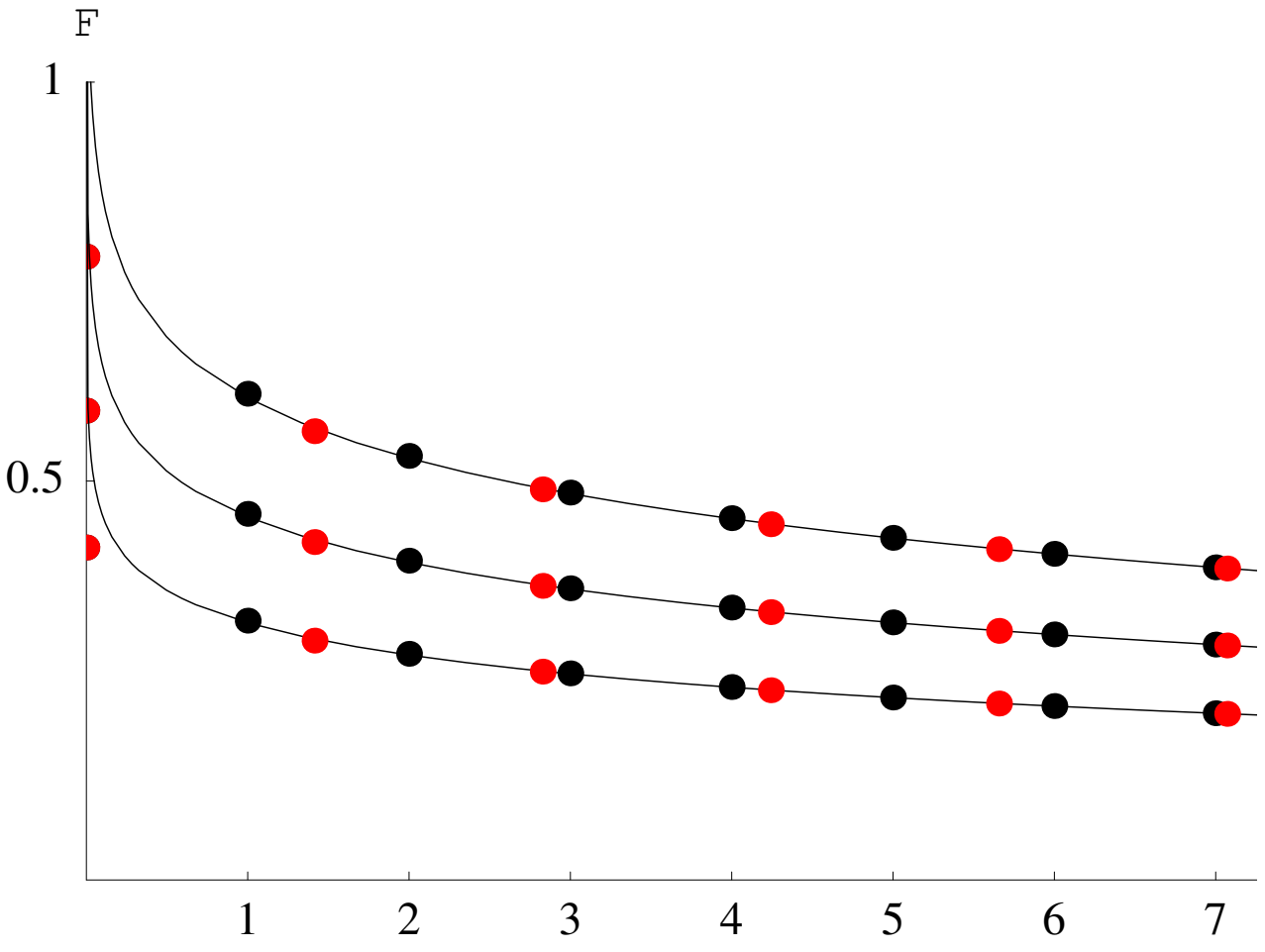
$$(*) \qquad F(x) = \mathbb{E}_x[e^{-2\mu T}] \approx \frac{K_0\big(|x|/l_\mu\big)}{\mathcal{N} + \log(l_\mu/\kappa)} \qquad |x| > \kappa$$

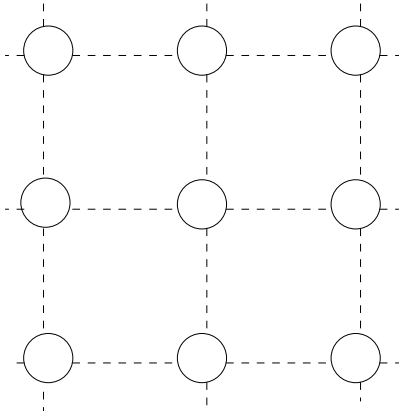
where $l_\mu = \sigma/2\mu$,

$$\mathbb{E}_0[e^{-2\mu T}] \approx \frac{\log(l_\mu/\kappa)}{\mathcal{N} + \log(l_\mu/\kappa)}.$$

$(*)$ is known as the Wright-Malécot formula.

# Malécot-Wright approximation for the stepping stone model



$F = \mathbb{P}[\text{identity}]$

# The unreasonable effectiveness of the Kingman coalescent

Common to use Kingman coalescent even for natural populations
Replace census population size by an effective population size

$N_e =$ number of individuals needed in an idealised population for specified quantity of interest (eg rate of change of genetic diversity) to be the same as in the real population.

# The unreasonable effectiveness of the Kingman coalescent

Common to use Kingman coalescent even for natural populations
Replace census population size by an effective population size

$N_e =$ number of individuals needed in an idealised population for specified quantity of interest (eg rate of change of genetic diversity) to be the same as in the real population.

For Buri's data we saw $N_e = N/\sigma^2$ where $\sigma^2$ was variance in number of offspring of a single fly.

Typically, $N_e < N$, possibly $\ll N$.

# The unreasonable effectiveness of the Kingman coalescent

Common to use Kingman coalescent even for natural populations
Replace census population size by an effective population size

$N_e =$ number of individuals needed in an idealised population for specified quantity of interest (eg rate of change of genetic diversity) to be the same as in the real population.

For Buri's data we saw $N_e = N/\sigma^2$ where $\sigma^2$ was variance in number of offspring of a single fly.

Typically, $N_e < N$, possibly $\ll N$.

Why does it work?

# Sampling uniformly from the torus $\mathbb{T}(L) \subset \mathbb{Z}^2$

$T =$ time to MRCA two individuals sampled uniformly from $\mathbb{T}(L)$

- $T_0 =$ time to first come into same deme
- $t_0 =$ time to coalesce started from same deme

$$T = T_0 + t_0$$

# Sampling uniformly from the torus $\mathbb{T}(L) \subset \mathbb{Z}^2$

$T =$ time to MRCA two individuals sampled uniformly from $\mathbb{T}(L)$

- $T_0 =$ time to first come into same deme
- $t_0 =$ time to coalesce started from same deme

$$T = T_0 + t_0$$

$X_t =$ distance between two lineages (for convenience *continuous time* r.w.) Uniform stationary distribution $\mathbb{P}_\pi[X_t = 0] = 1/L^2$

# Sampling uniformly from the torus $\mathbb{T}(L) \subset \mathbb{Z}^2$

$T =$ time to MRCA two individuals sampled uniformly from $\mathbb{T}(L)$

- $T_0 =$ time to first come into same deme
- $t_0 =$ time to coalesce started from same deme

$$T = T_0 + t_0$$

$X_t =$ distance between two lineages (for convenience *continuous time* r.w.) Uniform stationary distribution $\mathbb{P}_\pi[X_t = 0] = 1/L^2$

$$\mathbb{E}_\pi\Big[\text{time up to } L^2 \text{ lineages in same colony}\Big] = \int_0^{L^2} \mathbb{P}_\pi[X_t = 0]dt = 1$$

# Sampling uniformly from the torus $\mathbb{T}(L) \subset \mathbb{Z}^2$

$T =$ time to MRCA two individuals sampled uniformly from $\mathbb{T}(L)$

- $T_0 =$ time to first come into same deme
- $t_0 =$ time to coalesce started from same deme

$$T = T_0 + t_0$$

$X_t =$ distance between two lineages (for convenience *continuous time* r.w.) Uniform stationary distribution $\mathbb{P}_\pi[X_t = 0] = 1/L^2$

$$\mathbb{E}_\pi\Big[\text{time up to } L^2 \text{ lineages in same colony}\Big] = \int_0^{L^2} \mathbb{P}_\pi[X_t = 0]dt = 1$$

If $X_0 = 0$, local CLT $\implies$ $\mathbb{P}_0[X_t = 0] \approx 1/(4\pi\sigma^2 t)$

$$\mathbb{E}_0\Big[\text{time up to } L^2 \text{ lineages in same colony}\Big] = \int_0^{L^2} \mathbb{P}_0[X_t = 0]dt$$

$$\approx \frac{\log(L^2)}{4\pi\sigma^2}$$

# The distribution of $T_0$

$$1 = \int_0^{L^2} \mathbb{P}_\pi[X_t = 0]dt = \int_0^{L^2} \mathbb{P}_\pi[T_0 = s] \int_0^{L^2-s} \mathbb{P}_0[X_t = 0]dtds$$

$$\approx \mathbb{P}_\pi[T_0 \leq L^2] \frac{\log(L^2)}{4\pi\sigma^2}$$

So $\mathbb{P}_\pi[T_0 \leq L^2] \approx \frac{2\pi\sigma^2}{\log L}$. $\rightsquigarrow T_0 = O(L^2 \log L)$

# The distribution of $T_0$

$$1 = \int_0^{L^2} \mathbb{P}_\pi[X_t = 0]dt = \int_0^{L^2} \mathbb{P}_\pi[T_0 = s] \int_0^{L^2-s} \mathbb{P}_0[X_t = 0]dtds$$

$$\approx \mathbb{P}_\pi[T_0 \le L^2] \frac{\log(L^2)}{4\pi\sigma^2}$$

So $\mathbb{P}_\pi[T_0 \le L^2] \approx \frac{2\pi\sigma^2}{\log L}$. $\rightsquigarrow T_0 = O(L^2 \log L)$

$\tau := T_0/(L^2 \log L)$,

# The distribution of $T_0$

$$1 = \int_0^{L^2} \mathbb{P}_\pi[X_t = 0]dt = \int_0^{L^2} \mathbb{P}_\pi[T_0 = s] \int_0^{L^2-s} \mathbb{P}_0[X_t = 0]dtds$$

$$\approx \mathbb{P}_\pi[T_0 \le L^2] \frac{\log(L^2)}{4\pi\sigma^2}$$

So $\mathbb{P}_\pi[T_0 \le L^2] \approx \frac{2\pi\sigma^2}{\log L}$. $\rightsquigarrow T_0 = O(L^2 \log L)$

$\tau := T_0/(L^2 \log L)$,

Random walk to equilibriates over $\mathbb{T}(L)$ in $o(L^2 \log L)$ Cox & Durrett (2002)

# The distribution of $T_0$

$$1 = \int_0^{L^2} \mathbb{P}_\pi[X_t = 0]dt = \int_0^{L^2} \mathbb{P}_\pi[T_0 = s] \int_0^{L^2-s} \mathbb{P}_0[X_t = 0]dtds$$

$$\approx \mathbb{P}_\pi[T_0 \le L^2] \frac{\log(L^2)}{4\pi\sigma^2}$$

So $\mathbb{P}_\pi[T_0 \le L^2] \approx \frac{2\pi\sigma^2}{\log L}$. $\rightsquigarrow T_0 = O(L^2 \log L)$

$\tau := T_0/(L^2 \log L)$,

Random walk to equilibriates over $\mathbb{T}(L)$ in $o(L^2 \log L)$ Cox & Durrett (2002)

$$\mathbb{P}[\tau > s + t | \tau > s] = \mathbb{P}[\tau > t] \quad \text{as } L \to \infty$$

i.e (asymptotically) $\tau$ has exponential distribution

$$\mathbb{P}_\pi\left[T_0 > \frac{L^2 \log L}{2\pi\sigma^2} t\right] \to e^{-t}$$

# The distribution of $t_0$

$$\mathbb{P}_0\big[\text{lineages coalesce before jump apart}\big] = \frac{\frac{1}{N}}{\frac{1}{N} + 2m}$$

# The distribution of $t_0$

$$\mathbb{P}_0\big[\text{lineages coalesce before jump apart}\big] = \frac{\frac{1}{N}}{\frac{1}{N} + 2m}$$

$R_0 =$ return time

$$\mathbb{E}_0[t_0] = N + \Big(\frac{\frac{1}{N} + 2m}{\frac{1}{N}} - 1\Big)\mathbb{E}[R_0]$$

# The distribution of $t_0$

$$\mathbb{P}_0\big[\text{lineages coalesce before jump apart}\big] = \frac{\frac{1}{N}}{\frac{1}{N} + 2m}$$

$R_0 =$ return time

$$\mathbb{E}_0[t_0] = N + \Big(\frac{\frac{1}{N} + 2m}{\frac{1}{N}} - 1\Big)\mathbb{E}[R_0]$$

Kac's Lemma: $\mathbb{E}[R_0] = 1/(2m\pi(0)) = L^2/(2m)$

$$\mathbb{E}[t_0] = N(L^2 + 1)$$

# The distribution of $t_0$

$$\mathbb{P}_0\big[\text{lineages coalesce before jump apart}\big] = \frac{\frac{1}{N}}{\frac{1}{N} + 2m}$$

$R_0 =$ return time

$$\mathbb{E}_0[t_0] = N + \Big(\frac{\frac{1}{N} + 2m}{\frac{1}{N}} - 1\Big)\mathbb{E}[R_0]$$

Kac's Lemma: $\mathbb{E}[R_0] = 1/(2m\pi(0)) = L^2/(2m)$

$$\mathbb{E}[t_0] = N(L^2 + 1)$$

- Unless $N$ grows with $L$, $T_0$ dominates

# Genealogy

- Sample of size $k$: when first pair of lineages coalesces, positions remaining lineages uncorrelated with their starting points.

# Genealogy

- Sample of size $k$: when first pair of lineages coalesces, positions remaining lineages uncorrelated with their starting points.

- On timescale $L^2 \log L$ genealogy uniform sample from $\mathbb{T}(L) \to$ Kingman coalescent as $L \to \infty$ Zähle, Cox, Durrett (2005)

# Genealogy

- Sample of size $k$: when first pair of lineages coalesces, positions remaining lineages uncorrelated with their starting points.

- On timescale $L^2 \log L$ genealogy uniform sample from $\mathbb{T}(L) \to$ Kingman coalescent as $L \to \infty$ Zähle, Cox, Durrett (2005)

Census population size grows with $L^2$ so this does not explain the timescale seen in real populations