Bayesian inference for structured population stochastic epidemic models given final outcome data

P. D. O'Neill

University of Nottingham, UK

Outline

Structured populations

- Motivation: real-life populations rarely mix homogeneously.
- Instead there is frequently some inherent structure.
- Examples include:
  - Animals kept in pens
  - A town with homes, schools, workplaces
  - Animals inhabiting specific habitats

Models

- We may reasonably suppose that population structure influences disease transmission.
- In particular it is sensible to allow **rates** of transmission between individuals to depend on the common structures (if any) that they occupy.

Two-level mixing household model

Ball, Mollison and Scalia-Tomba, Annals of Applied Probability, 1997

Population of N individuals partitioned into households.

Households may be different sizes.

Initially, all individuals susceptible except for a few infectives.

At any time point, each individual is either

- susceptible
- infective
- recovered and immune

Epidemic ends when there are no more infectives.

Two-level mixing household model: infectious periods

An infectious individual remains so for a time T_I.

T_I is a specified non-negative random variable.

The T_I's for different individuals are usually i.i.d.

At the end of infectious period, individual becomes immune.

Two-level mixing household model: transmission

Whilst infectious, an individual has per-individual contacts....

- ....with household members at rate λ_L;
- ....with all individuals at rate λ_G/N.

i.e. contacts occur at times given by points of a Poisson process

All contact processes are independent.

Each such contact with a susceptible causes the susceptible to become infective immediately.

Two-level mixing household model: comments

- No latent periods -

  but including them does not affect final outcome.
- Can make the model multi-type; 'type' may refer to age, vaccination status, ...

Two-level mixing household model: threshold

Consider a branching process in which

'individuals = households'

Each individual has mean offspring $E[T]\lambda_G$, where

$T=$ number infected in a typical household.

Then $R_* = E[T]\lambda_G$ is a threshold parameter for the branching process as

$$\text{number of households} \to \infty,$$

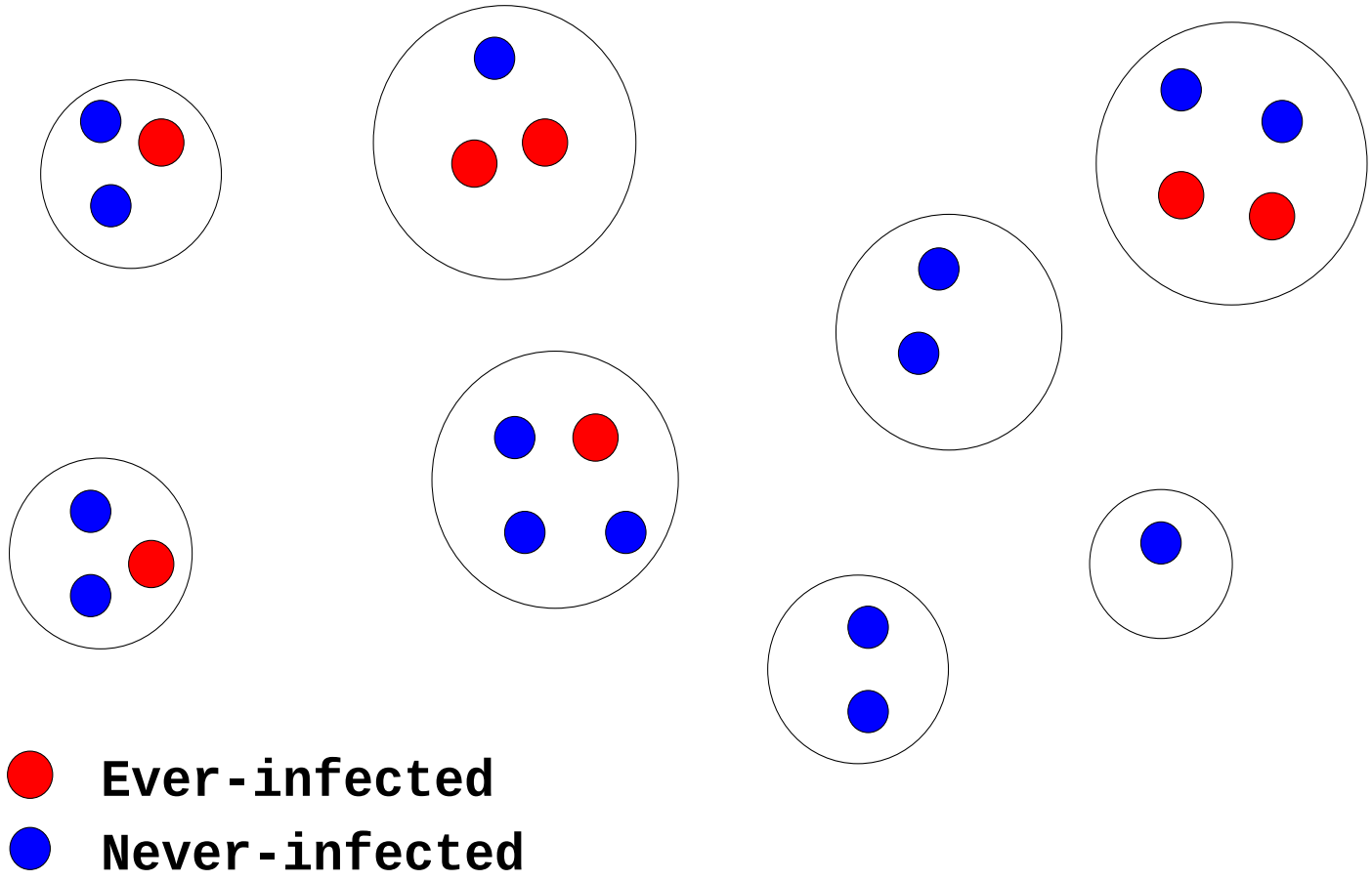
i.e. process may explode if and only if $R_* > 1$.

Data: overview

We consider data consisting of

- knowledge of the population structure
- final outcome (ever infected, or not) for each individual
- any 'type' information (for multi-type model)

2-level mixing model: data



Data: single-type example

Data on influenza outbreak in Tecumseh, Michigan, winter 1980-81.

	Susceptibles per household						
No. infected	1	2	3	4	5	6	7
0	44	62	47	38	9	3	2
1	10	13	8	11	5	3	0
2		9	2	7	3	0	0
3			3	5	1	0	0
4				1	0	0	0
5					1	0	0
6						0	0
7							0
Total	54	84	60	62	19	6	2

Inference: overview

We wish to infer information about the infection rate parameters λ_L and λ_G given the data (x, say).

In the Bayesian framework this means we focus on

$$\pi(\lambda_L, \lambda_G|x) \propto \pi(x|\lambda_L, \lambda_G)\pi(\lambda_L, \lambda_G)$$

Inference: the problem

Unfortunately, the likelihood

$$\pi(x|\lambda_L, \lambda_G)$$

is analytically and numerically intractable in all but very simple cases.

Inference: a solution

One way around the problem is to impute an (unobserved) description of the process of infectious contacts, i.e. who each person would infect if they themselves were infected.

Note that the contact processes of different individuals are independent.

Note also that it is only necessary to consider the contact processes of individuals who ever become infected (which the data tell us).

Inference: a solution

Call the imputed contact process information G. Then

$$\pi(\lambda_L, \lambda_G, G|x) \propto \pi(x|G)\pi(G|\lambda_L, \lambda_G)\pi(\lambda_L, \lambda_G)$$

and in particular

- $\pi(x|G)$ is just 1 or 0 (if G does/does not agree with data)
- $\pi(G|\lambda_L, \lambda_G)$ has a simple product form

MCMC algorithm

Sample-based inference can then be performed using an MCMC algorithm.

The parameters of interest in the algorithm are

- local infection rate λ_L
- global infection rate λ_G
- representation of the contact process G

Random graph representation

Joint work with Nikos Demiris, MRC, Cambridge

Demiris and O'Neill, J. Roy. Stat. Soc. Series B, 2005

Consider a single infective (i) and a single susceptible (j) within a household.

Recall that i contacts j at the points of a Poisson process of rate λ_L whilst infectious, i.e. for a time $T_I^{(i)}$, say.

Thus,

$$P(i \text{ infects } j | T_I^{(i)}) = 1 - \exp(-\lambda_L T_I^{(i)}).$$

Random graph representation

Conditional upon the value of $T_I^{(i)}$, all of i's household members have the same probability

$$1 - \exp(-\lambda_L T_I^{(i)})$$

of being contacted by i, and all such contacts are independent.

Furthermore, the contact processes of different individuals are independent.

Random graph representation

Thus the (local) contact process can be represented by a random graph in which an edge from one individual=vertex (i) to another (j) has probability

$$p_L^{(i)} = 1 - \exp(-\lambda_L T_I^{(i)})$$

Suppose i has $m^{(i)}$ ever-infected household members. Then the probability that i contacts a specified set of $n_L^{(i)}$ of them is

$$(p_L^{(i)})^{n_L^{(i)}} (1 - p_L^{(i)})^{m^{(i)} - n_L^{(i)}}$$

Random graph representation

The global contact process is similar, yielding between-individual edge probabilities

$$p_G^{(i)} = 1 - \exp(-\lambda_G T_I^{(i)}/N)$$

Note that global contacts may occur between any two members of the population.

Random graph representation

Call the ever-infected population A (n individuals), and the never-infected population C ($N-n$ individuals). Then

$$\pi(G|\lambda_L,\lambda_G)=\prod_{i=1}^{n}(p_L^{(i)})^{n_L^{(i)}}(1-p_L^{(i)})^{m^{(i)}-n_L^{(i)}}(p_G^{(i)})^{n_G^{(i)}}(1-p_G^{(i)})^{N-n_G^{(i)}}$$

$$\times P(A \text{ does not infect } C)$$

where

$$p_L^{(i)}=1-\mathrm{e}^{-T_I^{(i)}\lambda_L} \text{ etc,}$$

$$n_L^{(i)} = \text{number of edges that } i \text{ has in } G,$$

$$m^{(i)} = \text{number of infected individuals in } i\text{'s household.}$$

Random graph representation

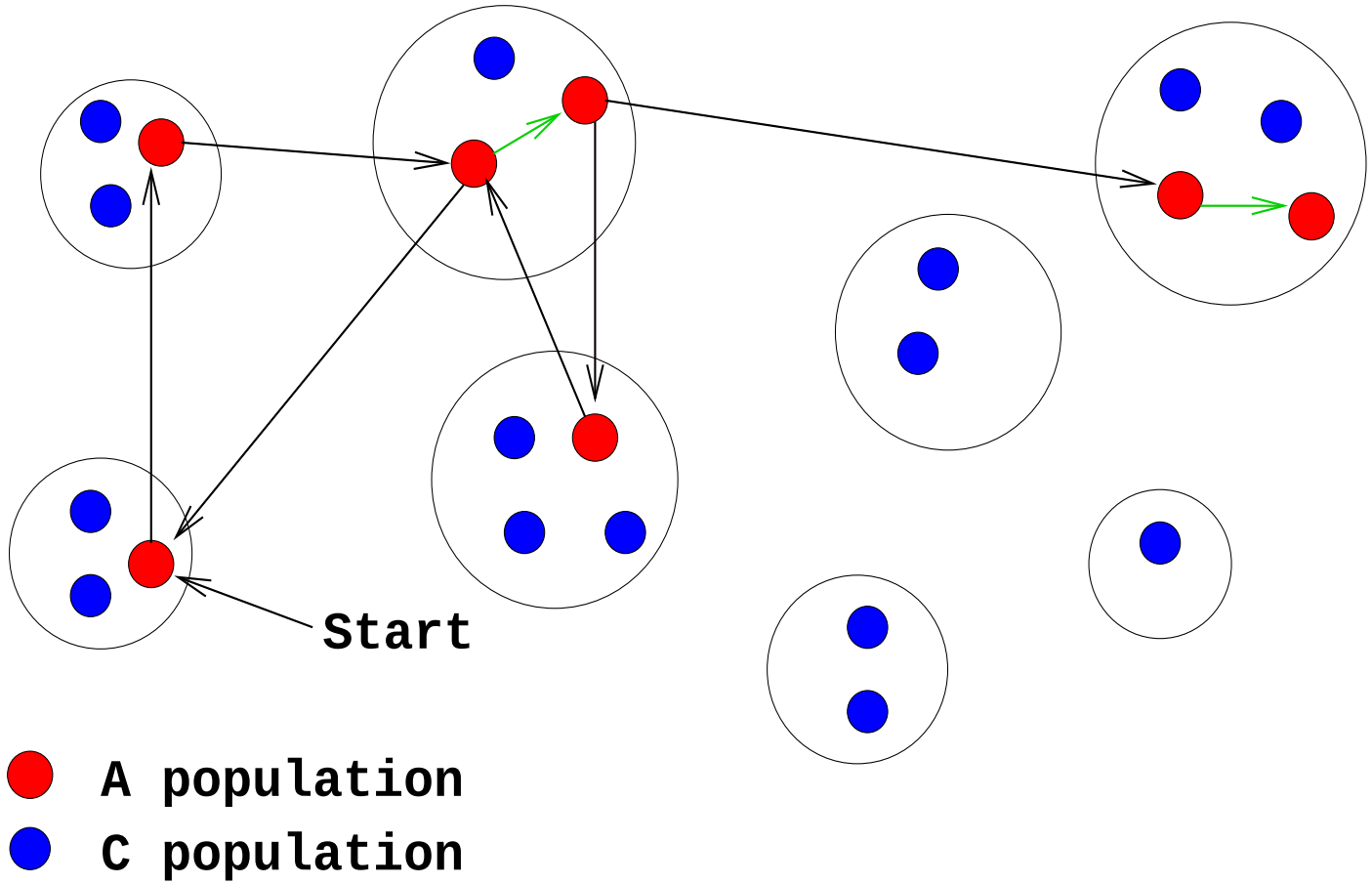
Furthermore,

$$P(A \text{ does not infect } C) = \mathrm{e}^{-\lambda_G T_A (N-n)/N},$$

where

$$T_A = \sum_{i=1}^{n} T_I^{(i)}.$$

Random graph representation



MCMC implementation

The parameters λ_L and λ_G can be updated by (e.g.) a Metropolis random walk with Gaussian proposals.

The graph G can be updated by adding and deleting edges at random.

The acceptance probabilities are easily evaluated.

Checking that G agrees with the data, i.e. G is appropriately connected, is computationally costly.

Poisson representation

Recall that individual i contacts each of their $m^{(i)}$ ever-infected household members at rate λ_L.

Equivalently, i has local contacts at rate $m^{(i)}\lambda_L$, and each contact is chosen uniformly at random from the $m^{(i)}$ individuals. Note that the contacts may be repeated.

Poisson representation

Thus if an individual i has $x_L^{(i)}$ specified local contacts, the contribution to the likelihood is

$$\left(\frac{\mathrm{e}^{-\lambda}\lambda^{x_L^{(i)}}}{x_L^{(i)}!}\right)\left(\frac{1}{m^{(i)}}\right)^{x_L^{(i)}}$$

where

$$\lambda = \lambda_L T_I^{(i)} m^{(i)}.$$

Global contacts are similar.

Poisson representation

Thus

$$\pi(G|\lambda_L,\lambda_G) = \prod_{i=1}^{n}\left(\frac{\mathrm{e}^{-\lambda_L(i)}\lambda_L(i)^{x_L^{(i)}}}{x_L^{(i)}!}\right)\left(\frac{1}{m^{(i)}}\right)^{x_L^{(i)}}$$

$$\times\left(\frac{\mathrm{e}^{-n\lambda_G/N}(n\lambda_G/N)^{x_G^{(i)}}}{x_G^{(i)}!}\right)\left(\frac{1}{n}\right)^{x_G^{(i)}}$$

$$\times P(A \text{ does not infect } C)$$

where

$$\lambda_L(i) = \lambda_L T_I^{(i)} m^{(i)}.$$

MCMC implementation

Under the Poisson representation it is possible to update all parameters according to Gibbs steps - other than checking connectivity.

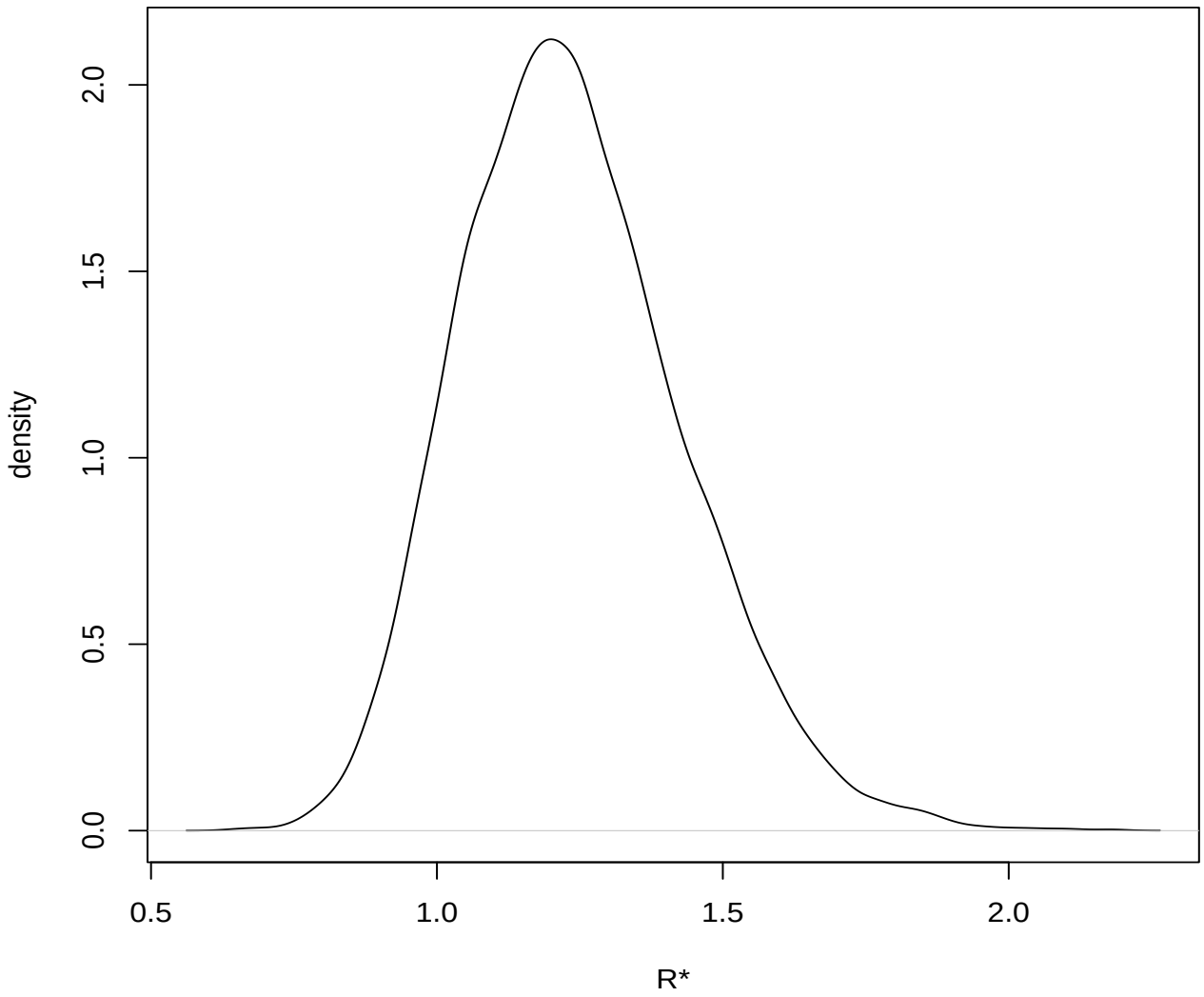
Number of contacts of an individual is Poisson.

The list of contacts is just i.i.d. uniform draws.

The parameters λ_L and λ_G can be updated via their full conditional Gamma distributions (assuming conjugate priors).

As before, the main computational cost is checking that the list of contacts agrees with the data.

Tecumseh data: R_* posterior density



Tecumseh data: summaries

	Parameter		
	λ_L	λ_G	R_*
Mean	0.048	0.190	1.24
Median	0.047	0.190	1.22
S. dev.	0.010	0.024	0.20
95% CI	(0.030,0.070)	(0.15,0.24)	(0.90,1.66)

	Parameter	
	local links	global links
Mean	48.8	98.9
Median	49	99
S. dev.	5.81	4.73
95% CI	(37,60)	(91,109)

Multitype models

Suppose the population contains individuals of k different observable types.

These typically correspond to covariates such as age, vaccination status, etc.

Infectious periods for different types may be different.

Multitype models

Infection-rates between types i, j are denoted

- λ_{ij}^{L} (Local)
- λ_{ij}^{G} (Global).

Define matrices

- $\Lambda^{L} = (\lambda_{ij}^{L})$
- $\Lambda^{G} = (\lambda_{ij}^{G})$

Model has $2k^2$ infection-rate parameters.

MCMC for multitype case

- Random graph representation - edges from individual i to type j individuals;
- Poisson representation - type i individual has Poisson $(T_I^{(i)} \lambda_{ij}^L)$ local contacts with each type j individual in household, etc.
- Likelihoods generalise easily.
- Updates are similar to single-type case.

Example - Tecumseh age data

Haber, Longini and Cotsonis, Biometrics, 1988

Adults and children (2 types).

289 households of sizes 1 - 5.

62 out of 491 adults infected.

63 out of 180 children infected.

So 125 infected individuals out of 671.

Observing a sample of the population

In reality we rarely observe the entire population.

In order to accomodate this, suppose that

- A is the observed infected population
- B is the unobserved population
- C is the observed uninfected population

Note that the data tell us about A and C.

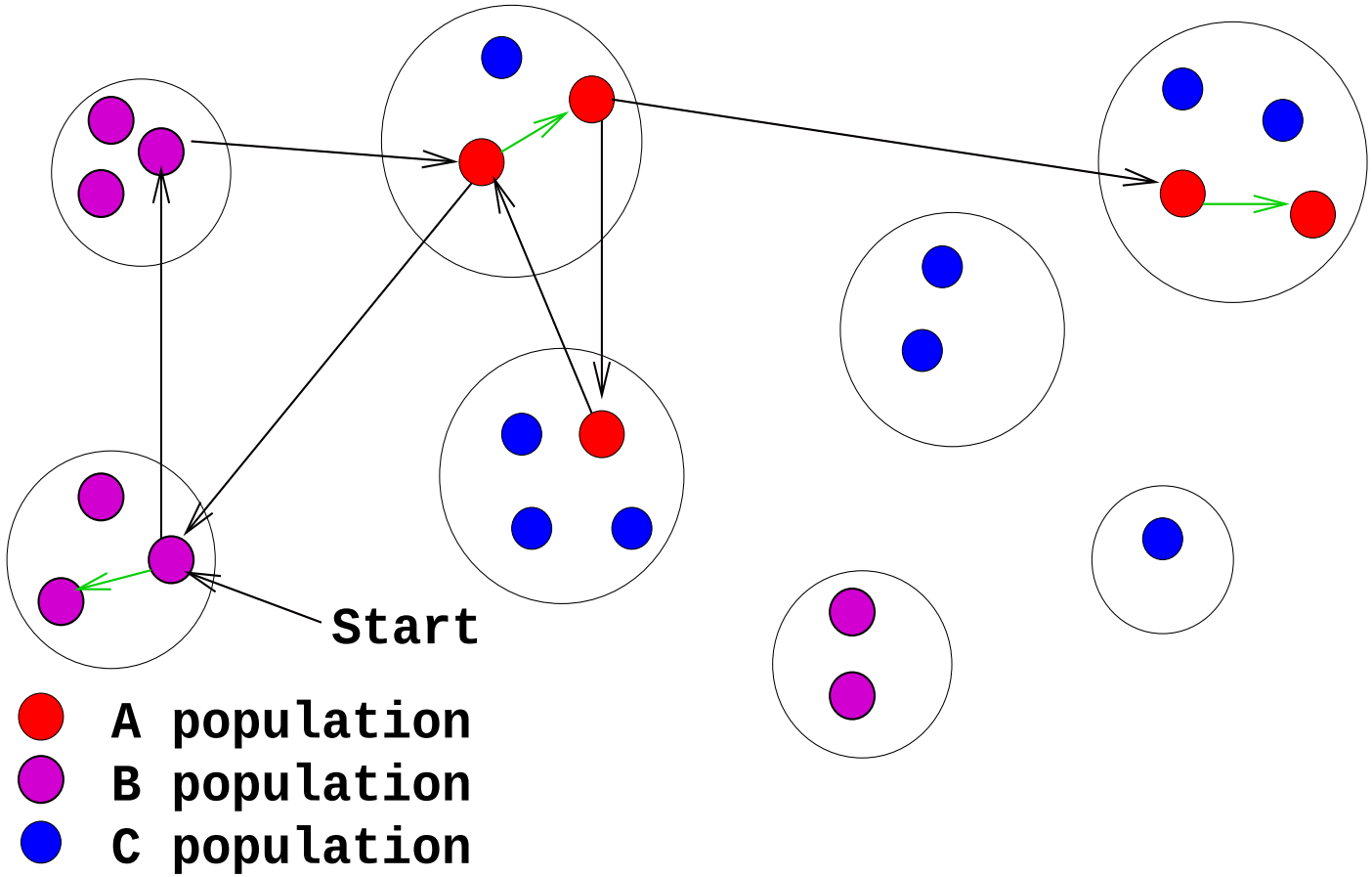
We must assume something about population structure of B.

Observing a sample of the population

Methods can proceed as before:

- Construct graph/contact list on A and B only
- Explicitly include all individuals and contacts in B
- Checking connectivity is slower

2-level mixing model: sample observed



Likelihood

As before, $\pi(G|\Lambda^L, \Lambda^G)$ is a product, over either

- 1. All individuals in A and B
- 2. All of A and those (currently) ever-infected in B

Although 2 is neater, updating graph G can alter likelihood dramatically and so acceptance probabilities are less straightforward to compute.

Effect of observing a sample

Let $\alpha \in (0, 1]$ denote fraction of population observed.

As $\alpha \downarrow 0$

- estimation becomes more precise
- $P(R_* > 1)$ increases....

Roughly: **if** we observe infected cases in a sample with α small, **then** it is unlikely that the epidemic died out early on....

....and this effect becomes more pronounced as α decreases.

Effect of observing a sample

Results from a small artificial dataset

7 households, each with 2 individuals.

Model assumes $\lambda_{ij}^L = \lambda_{kj}^L$, $\lambda_{ij}^G = \lambda_{kj}^G$.

(This means: type just affects susceptibility).

	α			
	1	0.5	0.1	0.05
$E(R_*\|x)$	2.7	2.6	2.3	2.3
$S.Dev.(R_*\|x)$	1.03	0.82	0.60	0.57
$P(R_* < 1\|x)$	0.016	0.0046	0.0002	$< 10^{-4}$

Effect of observing a sample

Similar results are seen for the Tecumseh age data set:

$\alpha = 1$ gives 95%CI for $R_* = (0.99, 1.56)$

$\alpha = 0.5$ gives 95%CI for $R_* = (1.04, 1.44)$

Work in progress

1. Improving MCMC mixing

Algorithms struggle to mix well with large (unobserved) populations.

How to move around more efficiently?

Also inherent problem with checking connectivity....

2. Perfect simulation

Set of all graphs is finite and partially ordered

Certain monotonicity properties hold (e.g. number of contacts stochastically increases with infection rate)

This suggests that coupling-from-the-past is feasible....

2-level mixing model