

Inferring Transmission Structure from HIV Sequence Data via Latent Spatial Poisson Processes

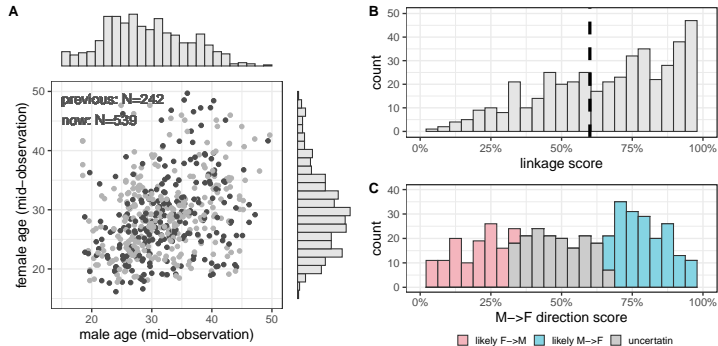
Fan Bu, Oliver Ratmann and Jason Xu
on behalf of the Rakai Community Cohort Study and PANGEA-HIV

Updated in June, 2022

Major objectives

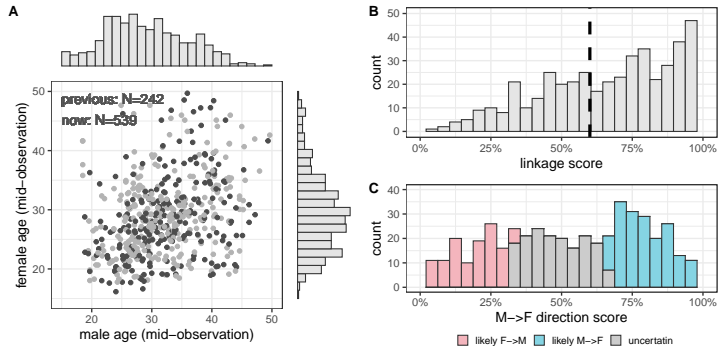
- ▶ Infer age structure in HIV transmissions based on phylogenetic analysis outcomes
- ▶ **Without** any pre-classification on phylogenetic summary statistics while making use of more data
- ▶ Exploit a **continuous** spatial process (with marks) and thus avoid manual discretization

The data



- A** male and female ages (mid-observation) for each potential transmission pair (s_{i1}, s_{i2})
- B** “posterior” linkage score $\ell_i \in [0, 1]$
- C** “posterior” direction score $d_i \in [0, 1]$

The data



A male and female ages (mid-observation) for each potential transmission pair (s_{i1}, s_{i2})

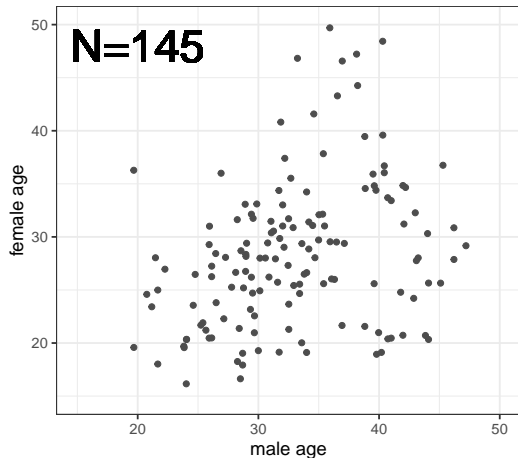
B “posterior” linkage score $\ell_i \in [0, 1]$

C “posterior” direction score $d_i \in [0, 1]$

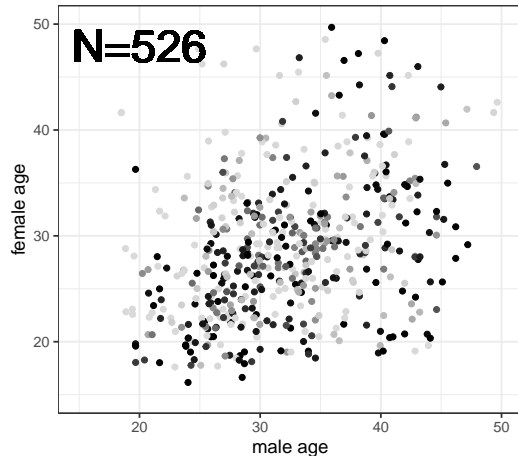
Goal: leverage the phylogenetic scores (ℓ_i and d_i) to infer transmission links and directions

The model

Existing approach: pre-classify points on ℓ_i and d_i ; discretize on age bands
(below: pre-classified MF transmissions)

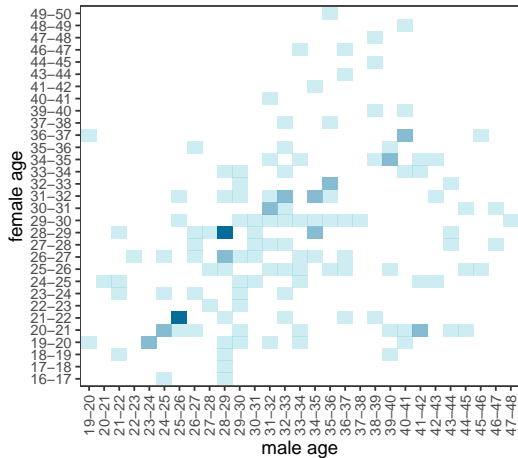


Our approach: **no need** for any pre-classification or discretization, directly accounting for uncertainty



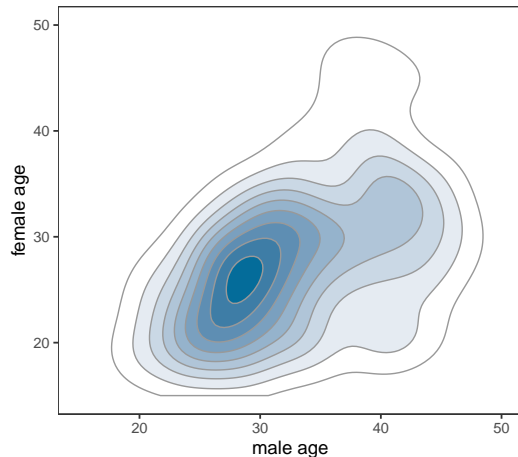
The model

Existing approach:



End result: transmission rate π_{ab} between (discrete) age groups a and b

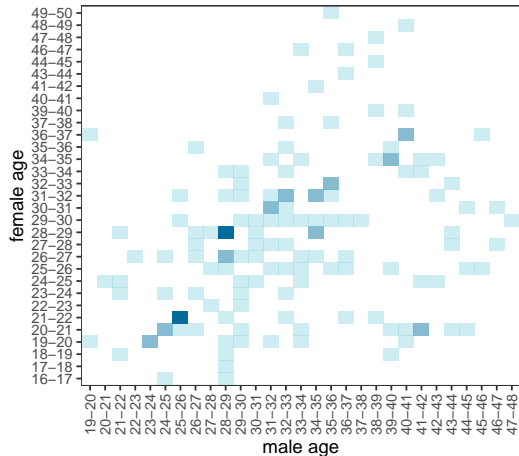
Our approach:



End result: a **continuous** spatial intensity function $\lambda(\cdot, \cdot)$

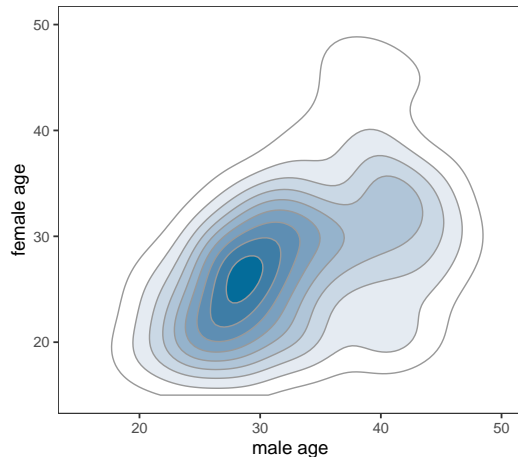
The model

Existing approach:



π_{ab} specific to age discretization;
hard to convert to different resolutions

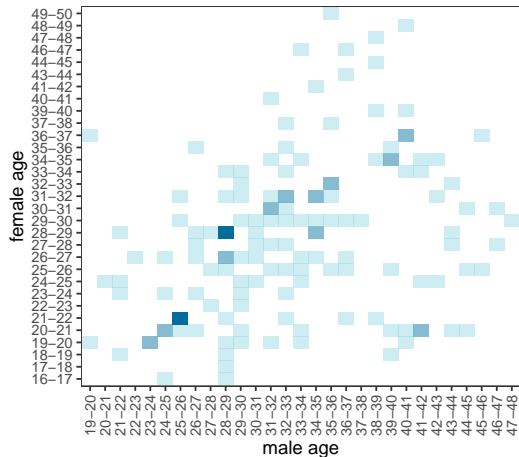
Our approach:



Easy to discretize to any age group
resolution by integrating $\lambda(\cdot, \cdot)$

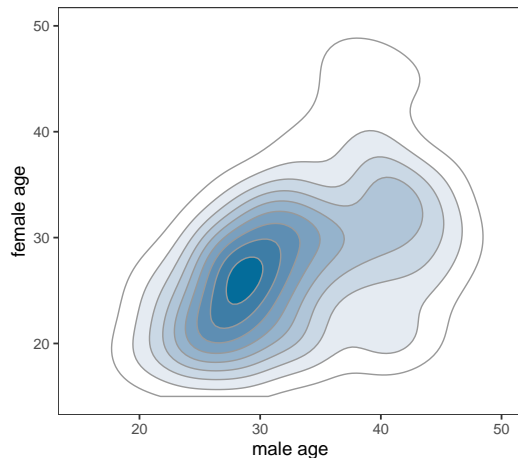
The model

Existing approach:



π_{ab} specific to age discretization;
hard to convert to different resolutions

Our approach:



E.g., Source age (marginal) for recipient age $b = \int_{a \in \mathcal{A}} \lambda(a, b) da$

The model

- ▶ Assume age pairs $\mathbb{E} = \{(s_{i1}, s_{i2})\}_{i=1}^N$ are point patterns from a 2D spatial Poisson process with intensity function λ .
- ▶ Introduce (latent) variable c_i as (unknown) type label for point i :
 - ▶ $c_i = 0$: no transmission event
 - ▶ $c_i = 1$: M \rightarrow F transmission
 - ▶ $c_i = 2$: F \rightarrow M transmission

The model

- ▶ Assume age pairs $\mathbb{E} = \{(s_{i1}, s_{i2})\}_{i=1}^N$ are point patterns from a 2D spatial Poisson process with intensity function λ .
- ▶ Introduce (latent) variable c_i as (unknown) type label for point i :
 - ▶ $c_i = 0$: no transmission event
 - ▶ $c_i = 1$: M \rightarrow F transmission
 - ▶ $c_i = 2$: F \rightarrow M transmission
- ▶ Two parts of model:
 1. Spatial process (mixture model on spatial density function)
 2. “Marks” distribution: logit-normal model on the linkage and direction scores

More on the spatial process

Decompose the intensity function $\lambda(\cdot)$ into scale and density functions, and then model the density function as a mixture.

$$\lambda(\cdot) = \gamma\pi(\cdot)$$

$$\pi(\cdot) = \sum_{k \in \{0,1,2\}} p_k \pi_k(\cdot)$$

$$\pi_k(\cdot) = \sum_{h=1}^{H_k} w_{kh} \text{BVN}(\cdot; \theta_{kh}, \Sigma_{kh}).$$

More on the spatial process

Decompose the intensity function $\lambda(\cdot)$ into scale and density functions, and then model the density function as a mixture.

$$\lambda(\cdot) = \gamma \pi(\cdot)$$

$$\pi(\cdot) = \sum_{k \in \{0,1,2\}} p_k \pi_k(\cdot)$$

$$\pi_k(\cdot) = \sum_{h=1}^{H_k} w_{kh} \text{BVN}(\cdot; \theta_{kh}, \Sigma_{kh}).$$

Spatial density function π_k :

- ▶ “continuous” version of discrete rate π_{ab} ’s
- ▶ integrates to 1 over the 2D age space
- ▶ bivariate normal mixture model with Dirichlet process prior (flexible number of components)

Likelihood function

Construct **complete data** likelihood given type labels c_i 's for parameters

$$\Theta = \{\gamma, \mathbf{p}, \boldsymbol{\mu}, \sigma_\ell^2, \sigma_d^2, \{(\theta_{kh}, \Sigma_{kh})\}, \{\alpha_k\}\}.$$

$$\begin{aligned} L(\Theta; \{\mathbf{x}_i\}, \{c_i\}, \{\mathbf{s}_i\}) &= \gamma^N \frac{e^{-\gamma}}{N!} \prod_{k \in \{0,1,2\}} \prod_{i: c_i=k} p_k f_k(\mathbf{s}_i) \phi_k(\mathbf{x}_i) \\ &= \prod_{i=1:N} \phi(\mathbf{x}_{i1} \mid \tilde{\mu}_{\ell,i}, \sigma_\ell^2) \phi(\mathbf{x}_{i2} \mid \tilde{\mu}_{d,i}, \sigma_d^2) \\ &\quad \times \gamma^N \frac{e^{-\gamma}}{N!} \prod_{k \in \{0,1,2\}} \prod_{i: c_i=k} \left(p_k \sum_{h=1}^{H_k} w_{kh} \text{BVN}((s_{i1}, s_{i2}); \theta_h, \Sigma_h) \right). \end{aligned}$$

blue terms: spatial process for 2D paired ages

red terms: “marks” distribution for phylogenetic scores

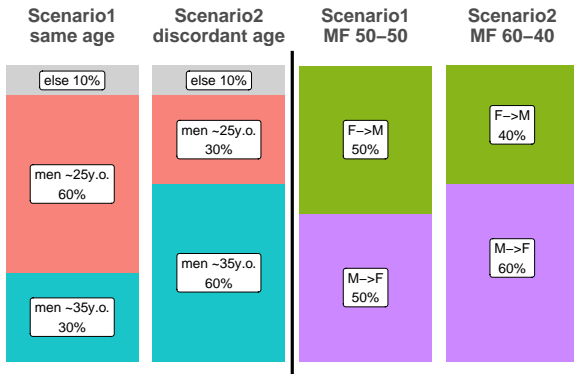
Inference

- ▶ Data augmentation approach in a Bayesian inference framework to make it tractable
- ▶ Treat unknown type label c_i as latent variables; in each iteration:
 1. sample c_i conditional on everything else
 2. sample parameters Θ conditional on configurations of c_i

Simulation study - setup

Two key problems:

- ▶ male source age for young women (~ 15 -24 y.o.)
- ▶ proportions of each transmission direction (more MF or more FM?)



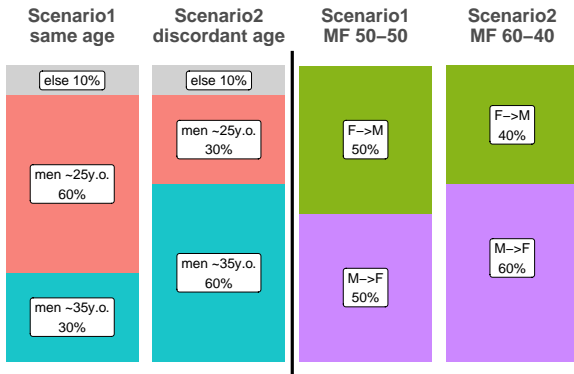
Simulation study - setup

Two key problems:

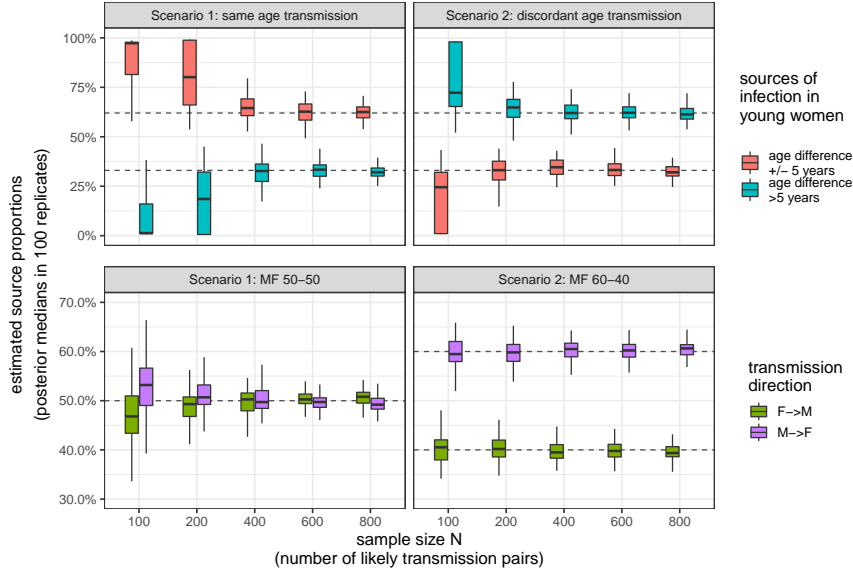
- ▶ male source age for young women (~ 15 -24 y.o.)
- ▶ proportions of each transmission direction (more MF or more FM?)

Sample size

$N = 100, 200, 400, 600, 800$;
100 runs for each setup.



Simulation study - results



Simulation study - results commentary

- ▶ As sample size N increases, better accuracy at estimating the quantities of interest
- ▶ Satisfactory performance with $N = 400$ or 600 already; this is the sample size range of the real data
- ▶ For the age source problem, difference between younger/older men proportions is over-estimated when N is small – to be expected with the parsimony nature of the Dirichlet process prior

Real data analysis

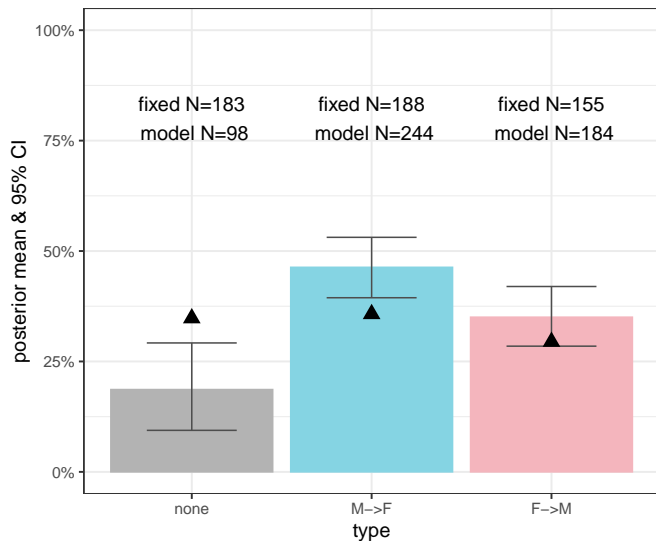
Data overview

- ▶ ~ 540 potential transmission pairs, with male and female ages, and linkage and direction scores obtained from phylogenetic analysis
- ▶ pre-processing: keep pairs with $\ell_i > 0.2 \rightarrow 526$ pairs in total

Comparison of our approach with existing approach

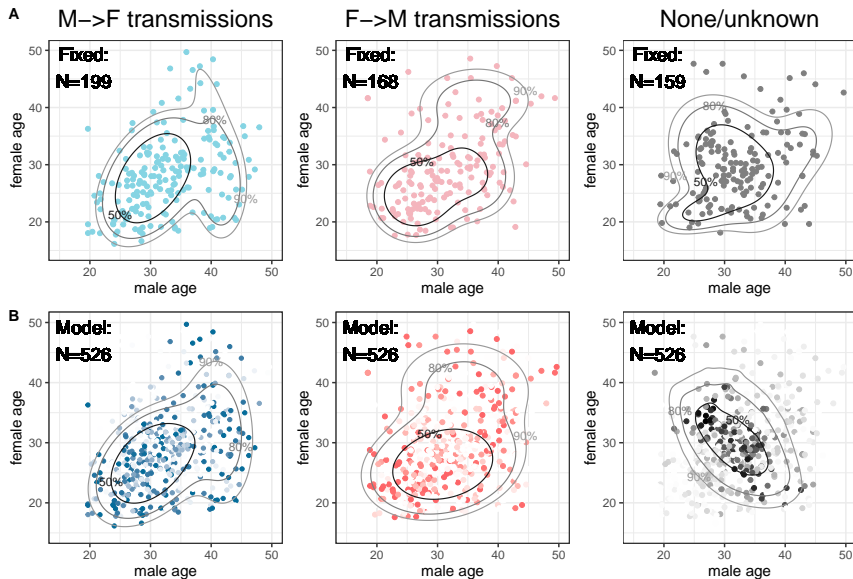
- ▶ **“fixed”**: pre-classify with fixed type labels. $\ell_i > 0.6 \rightarrow$ real events; $d_i > 0.5 \rightarrow$ MF transmission, otherwise FM transmissions.
- ▶ **“model”**: full model with flexible type labels learned as latent variables.

Real data analysis - transmission type proportions



- ▶ ▲ = proportions using **fixed** thresholds ($\ell_i > 0.6$, $d_i > 0.5 \rightarrow M \rightarrow F$, $d_i < 0.5 \rightarrow F \rightarrow M$)
- ▶ our model **includes more** data points as transmission events
- ▶ yet, ratio between MF and FM directions stays similar

Real data analysis - transmission age structure with uncertainty



Real data analysis - transmission age structure with uncertainty

- ▶ Row A: “**fixed**” analysis with pre-classification; Row B: “**model**”-learned flexible type labels, with point color shades representing the posterior probabilities of each data point belonging to each type.
- ▶ Contour lines: 50%, 80% and 90% highest-density regions for point patterns (MAP estimates of spatial process).
- ▶ Similar point patterns are identified by proposed model compared to a fixed-label ad-hoc approach.

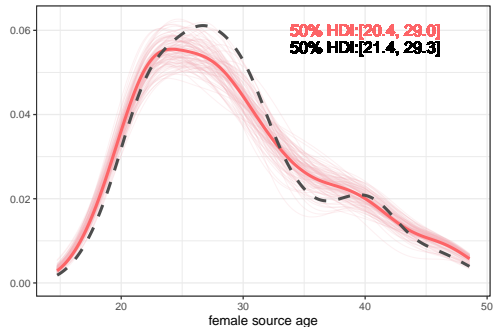
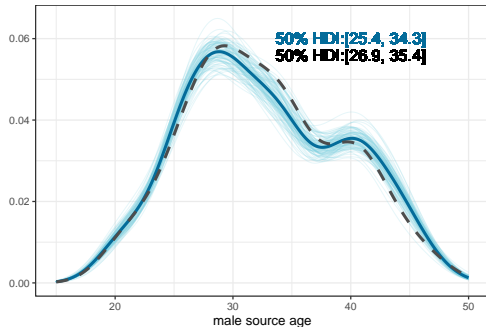
Real data analysis - transmission age structure with uncertainty

- ▶ Row A: “**fixed**” analysis with pre-classification; Row B: “**model**”-learned flexible type labels, with point color shades representing the posterior probabilities of each data point belonging to each type.
- ▶ Contour lines: 50%, 80% and 90% highest-density regions for point patterns (MAP estimates of spatial process).
- ▶ Similar point patterns are identified by proposed model compared to a fixed-label ad-hoc approach.
 - ▶ Overall age assortative transmissions, but...
 - ▶ Older men transmit to a wide range of female age groups (both young and old) and get infected by slightly older women (still wide range).
 - ▶ Much more old-men-to-young-women transmissions than old-women-to-young-men.
 - ▶ No clear patterns for the “none/unknown” type – can be considered as background noise.

Real data analysis - transmission age structure with uncertainty

- ▶ Row A: “**fixed**” analysis with pre-classification; Row B: “**model**”-learned flexible type labels, with point color shades representing the posterior probabilities of each data point belonging to each type.
- ▶ Contour lines: 50%, 80% and 90% highest-density regions for point patterns (MAP estimates of spatial process).
- ▶ Similar point patterns are identified by proposed model compared to a fixed-label ad-hoc approach.
- ▶ However, our model is able to differentiate the strengths of evidence among different data points and thus leverage and respect data uncertainty.

Real data analysis - source age distribution (overall)

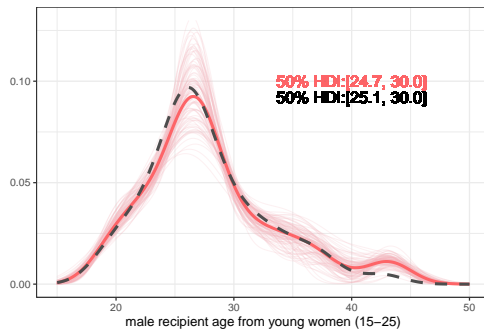
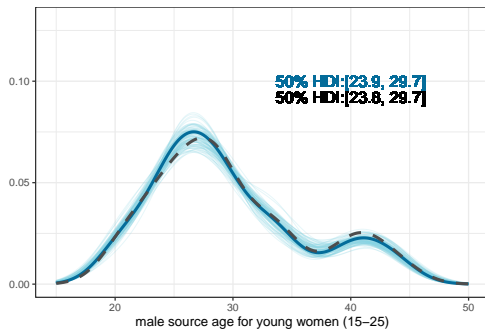


Dark dashed lines: “**fixed**” analysis; solid colored lines: “**model**” with 100 posterior samples, overlaid with posterior mean.

There are more older male sources than older female sources.

Real data analysis - the story of young women (15-25 y.o.)

Whom do they get infected by, and whom do they transmit to?

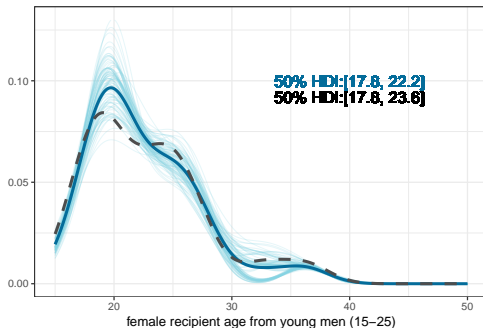
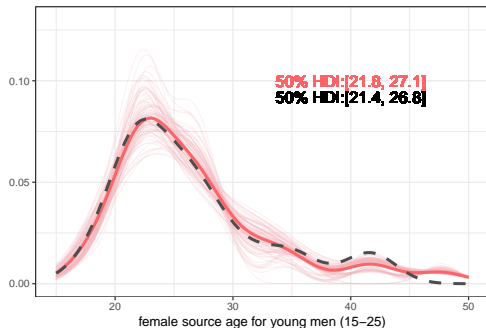


Notable peak at male source around 40y.o., but younger women transmit more to younger men.

There are “sugar-daddy” transmissions that result in HIV infections in young women (a deviation from typical age assortative behavior).

Real data analysis - the story of young men (15-25 y.o.)

Whom do they get infected by, and whom do they transmit to?



Very different story for young men: mainly get infected by young women, and mainly transmit back to young women. Age assortative behavior.

Summary

- ▶ Proposed a novel model for inferring HIV transmission flow between age groups
- ▶ Included more data points in the model while considering uncertainty in identifying transmission links/directions
- ▶ Utilized spatial process with marks to leverage both phylogenetic information and age structure
- ▶ Avoided manual discretization with a continuous construction

Thank you!

Supplement slides

More on the “marks” distributions

Specify a logit-normal model for the phylogenetic scores.

$$\begin{aligned}\text{logit}(\ell_i) \mid c_i &\sim N(\tilde{\mu}_{\ell,i}, \sigma_{\ell}^2), \\ \text{logit}(d_i) \mid c_i &\sim N(\tilde{\mu}_{d,i}, \sigma_d^2),\end{aligned}$$

where

$$\begin{aligned}\tilde{\mu}_{\ell,i} &= \mu_{\ell} \mathbb{1}[c_i \neq 0], \\ \tilde{\mu}_{d,i} &= \mu_d \mathbb{1}[c_i = 1] + \mu_{-d} \mathbb{1}[c_i = -1].\end{aligned}$$

More on the “marks” distributions

Specify a logit-normal model for the phylogenetic scores.

$$\begin{aligned}\text{logit}(\ell_i) \mid c_i &\sim N(\tilde{\mu}_{\ell,i}, \sigma_{\ell}^2), \\ \text{logit}(d_i) \mid c_i &\sim N(\tilde{\mu}_{d,i}, \sigma_d^2),\end{aligned}$$

where

$$\begin{aligned}\tilde{\mu}_{\ell,i} &= \mu_{\ell} \mathbb{1}[c_i \neq 0], \\ \tilde{\mu}_{d,i} &= \mu_d \mathbb{1}[c_i = 1] + \mu_{-d} \mathbb{1}[c_i = -1].\end{aligned}$$

$\ell_i > 0.5 \rightarrow$ more likely to be real transmission

$d_i > 0.5 \rightarrow$ more likely to be M \rightarrow F transmission

$d_i < 0.5 \rightarrow$ more likely to be F \rightarrow M transmission