Supporting Information for

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

Web Appendix A: Additional Details for Model and Inference

Here we provide a summary of our MCMC sampling algorithm for Bayesian inference, as detailed in Section 3 of the main text.

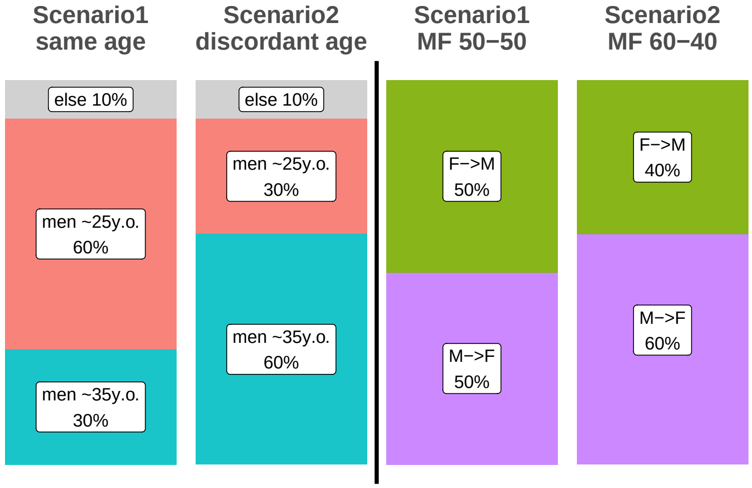
Web Algorithm 1 MCMC inference with data augmentation

```
procedure INFERENCE
    Directly draw samples for $\gamma$ with
        $\gamma \mid \{\mathbf{x_i}\}, \{\mathbf{s_i}\} \sim Ga(\alpha_0 + N, \beta_0 + 1).$
    Randomly initialize parameter values $\Theta^{(0)}$ (except for $\gamma$).
    Randomly assign initial type labels $\{c_i^{(0)}\}$.
    for $t = 1 : T$ do
        (1) Sample $\boldsymbol{\mu}^{(t)}$, $\sigma_\ell^{2(t)}$, and $\sigma_d^{2(t)}$ conditional on $\{c_i^{(t-1)}\}$ and logit-transformed signals $\mathbf{x}_i$:
            $\mu_\ell \mid \sigma_\ell^{2(t-1)}, \{\ell_i\}, \{c_i^{(t-1)}\} \sim N_{(0,\infty)}\left(\sum_{i:c_i \neq 0} \mathrm{logit}(\ell_i), \sigma_\ell^{2(t-1)}/N_+\right);$
            $\mu_d \mid \sigma_d^{2(t-1)}, \{d_i\}, \{c_i^{(t-1)}\} \sim N_{(0,\infty)}\left(\sum_{i:c_i = 1} \mathrm{logit}(d_i), \sigma_d^{2(t-1)}/N_1\right);$
            $\mu_{-d} \mid \sigma_d^{2(t-1)}, \{d_i\}, \{c_i^{(t-1)}\} \sim N_{(-\infty,0)}\left(\sum_{i:c_i = -1} \mathrm{logit}(d_i), \sigma_d^{2(t-1)}/N_1\right);$
            $\sigma_\ell^2 \mid \mu_\ell^{(t)}, \{\ell_i\}, \{c_i^{(t-1)}\} \sim \text{inv-Gamma}\left(\frac{\nu_0 + N_+}{2}, \frac{\nu_0\sigma_0^2 + \sum_{i:c_i \neq 0}(\mathrm{logit}(\ell_i) - \mu_\ell^{(t)})^2}{2}\right);$
            $\sigma_d^2 \mid \mu_d^{(t)}, \mu_{-d}^{(t)}, \{d_i\}, \{c_i^{(t-1)}\} \sim \text{inv-Gamma}\left(\frac{\nu_0 + N_+}{2}, \frac{\nu_0\sigma_0^2 + \sum_{i:c_i = 1}(\mathrm{logit}(d_i) - \mu_d^{(t)})^2 + \sum_{i:c_i = -1}(\mathrm{logit}(d_i) - \mu_{-d}^{(t)})^2}{2}\right).$
        (2) For each data point $i$, sample $c_i^{(t)}$ from
            $Pr(c_i = k \mid \Theta^{(t-1)}) \propto p_k f_k(\mathbf{s}_i) \phi_k(\mathbf{x}_i).$
        (3) sample each $\mathbf{p}^{(t)}$ conditional on $N_k$, the total number of points with $c_i^{(t)} = k$:
            $\mathbf{p} \mid \{c_i^{(t)}\} \sim Dir\left(q_{-1} + N_{-1}, q_0 + N_0, q_1 + N_1\right).$
        for each type $k$ do
            (4.a) Conditionally sample DP precision parameter $\alpha_k^{(t)}$ and component weights $w_{kh}^{(t)}$, using updating steps described in ?.
            (4.b) For each data point $i$, sample a component latent indicator $z_i^{(t)}$ conditional on $w_{kh}^{(t)}$ and $\theta_{kh}^{(t-1)}, \Sigma_{kh}^{(t-1)}$ by
                $Pr(z_i = h \mid w_{kh}^{(t)}, \theta_{kh}^{(t-1)}, \Sigma_{kh}^{(t-1)}, \mathbf{s}_i) \propto w_{kh}^{(t)} \varphi(\mathbf{s}_i \mid \theta_{kh}^{(t-1)}, \Sigma_{kh}^{(t-1)}).$
            (4.c) For each component $h$, sample $\theta_{kh}^{(t)}, \Sigma_{kh}^{(t)}$ conditional on all data points $\mathbf{s}_i$ with $z_i^{(t)} = h$:
                $\theta_{kh} \mid \Sigma_{kh}^{(t-1)}, \{z_i^{(t)}\}, \{\mathbf{s_i}\} \sim \text{BVN}\left((m_h(\Sigma_{kh}^{(t-1)})^{-1} + \Sigma_0^{-1})^{-1}\left(\sum_{i:z_i^{(t)}=h} (\Sigma_{kh}^{(t-1)})^{-1}\mathbf{s_i} + \theta_0\Sigma_0^{-1}\theta_0\right), (m_h(\Sigma_{kh}^{(t-1)})^{-1} + \Sigma_0^{-1})^{-1}\right);$
                $\Sigma_{kh} \mid \theta_{kh}^{(t)}, \{z_i^{(t)}\}, \{\mathbf{s_i}\} \sim \text{inv-Wishart}\left(\nu + m_h, \left(S_0^{-1} + \sum_{i:z_i^{(t)}=h} (\mathbf{s_i} - \theta_{kh}^{(t)})(\mathbf{s_i} - \theta_{kh}^{(t)})^T\right)^{-1}\right).$
    Return MCMC samples for parameters $\Theta$ and type labels $\{c_i\}$
```

Web Appendix B: Supplemental Materials for Simulation Study

Parameters and settings in the simulations

Below we detail the parameter choices in the simulation study described in Section 3 of the main text. We also include a diagram that illustrates the different scenarios compared in our simulations in Web Figure 1.



Web Figure 1: Graphic illustration of the simulation setup. We consider two different scenarios for each epidemiological question of interest. (1) Age of male sources for young women between 15 and 24 (left panel); scenario "**same age**" has most young women infections attributable to young men around 25 who are of similar age to those infected, whereas in scenario "**discordant age**" most such transmissions are attributable to older men around 35. (2) Proportions of male-to-female and female-to-male events; for scenario "**MF 50-50**" the two transmission directions have equal incidents, while in scenario "**MF 60-40**" there are slightly more MF transmissions which contribute to 60% of total infections.

1: age structure of male sources for infections in young women We consider two scenarios with different setup in the BVN mixture model of the spatial density function $f_1(\cdot)$ for MF transmissions:

- **same age**: For women aged between 15-24, infections from younger men take the majority; here, for density function $f_1(\cdot)$, we set the component centered at $(35, 20)^T$ (i.e., older men sourced transmission) to have mixture weight 0.3, and the component centered at $(25, 20)^T$ (i.e., younger men sourced transmission) to have mixture weight 0.6.

- **discordant age**: For women aged between 15-24, infections from older men take the majority; here, for function $f_1(\cdot)$, we set the component centered at $(35, 20)^T$ to have mixture weight 0.6, and the component centered at $(25, 20)^T$ to have mixture weight 0.3.

For each scenario, the other 10% probability mass for MF transmissions is spread across all other bivariate normal components of the density function f_1.

2. proportions of MF and FM transmission events We consider the following two scenarios with different value choices for the type probability $\mathbf{p}$:

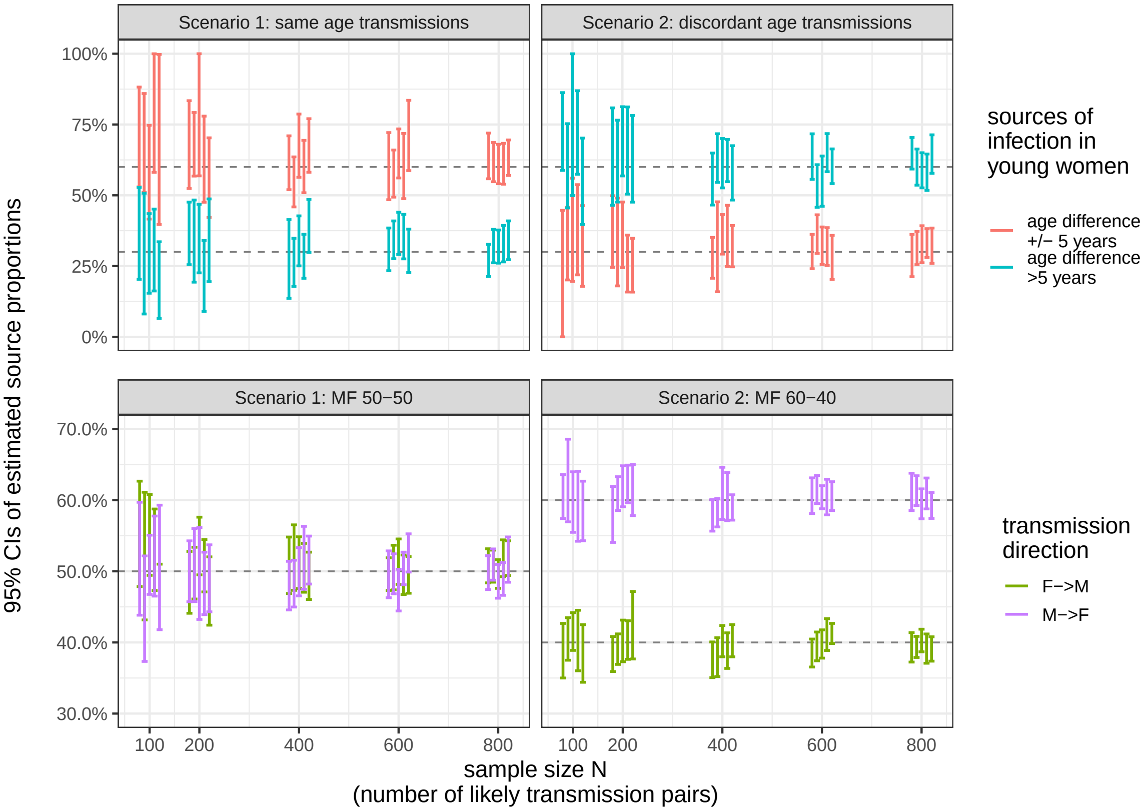
- **MF 50-50**: There are equal proportions of MF and FM events, and each contribute to about 50% of all real transmission events; we set 25% of events to be non-transmission events, and thus we have $p_0 = 0.25, p_{-1} = p_1 = 0.375$.
- **MF 60-40**: There are more MF events than FM events, and MF events contribute to about 60% of all transmissions; again, we set 25% of events to be non-transmissions, and thus we have $p_0 = 0.25, p_{-1} = 0.3, p_1 = 0.45$.

To simulate the spatial patterns, we use 6 different BVN mixture components to construct the three density functions f_0, f_{-1}, f_1; for the signal distributions, we set $\mu_\ell = 2$, $\mu_d = 1.5, \mu_{-d} = -1.5$, and $\sigma_\ell^2 = \sigma_d^2 = 1$.

For each simulation, we run the Bayesian MCMC sampler for 3000 iterations with a 1000-iteration burn-in period, and adopt the following hyper-parameters for the priors: $a_0 = 1, b_0 = 0.02$ (prior for γ), $\nu_0 = 2, \sigma_0^2 = 1$ (priors for σ_ℓ^2 and σ_d^2), $q_i = 1, i = 1, 2, 3$ (prior for $\mathbf{p}$), $\theta_0 = (0, 0)^T, \Sigma_0 = \left(\begin{smallmatrix} 10^4 & 0 \\ 0 & 10^4 \end{smallmatrix}\right)$ (priors for θ_{kh}'s), $\nu = 2, S_0 = I_2$ (priors for Σ_{kh}'s), and $a = 2, b = 3$ (priors for α_k's). We have experimented with various hyper-parameter values and have found that the results are not sensitive to value changes within reasonable ranges.

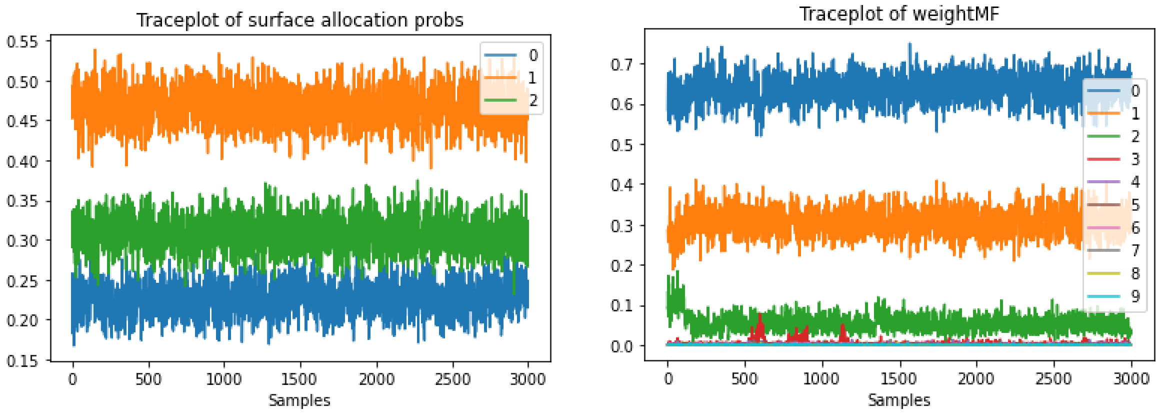
Simulation study results

In addition to the posterior medians results shown in the main text, within each simulation run, we can also look at the posterior credible intervals. In Web Figure 2, we plot the 95% posterior credible intervals for the proportions of male source age groups in each scenario in the top row, and for proportions of each transmission directions as well in the bottom row. For each sample size N and scenario setting, the plotted credible intervals are acquired from **five** (5) randomly selected simulations among 100 total runs. We can see that the majority of these credible intervals do cover the truth (marked by dashed lines in each sub-plot), and when N gets larger, the credible intervals get narrower, indicating less uncertainty when more data are available.



Web Figure 2: 95% posterior credible intervals of the simulation study. Results are shown for **five** (5) randomly selected simulation run for each N and each scenario. Each error bar shows the 95% posterior credible interval. The true parameter values are marked in dashed lines in each subplot.

Moreover, within each run of the simulation study, we can also inspect the MCMC traceplots (such as the ones shown in Web Figure 3) to check for convergence and inference quality, which is also similarly done for the real data case study. For example, in subplot (a), we plot the values of entries in $\mathbf{p}$ (the probability/probability vector for the three transmission types) sampled across all 3000 iterations of the MCMC sampler, where the three lines represent MF, FM and non-event transmission surfaces from top to bottom. Inspecting this traceplot, we can see that the sampled probability values stabilize after the initial 500 iterations or so (used as the "burn-in" period) and, in later iterations, fluctuate



(a) Traceplot of the surface probabilities (entries of $\mathbf{p}$). 0 = non-event, 1 = MF transmission surface, 2 = FM transmission surface.

(b) Traceplot of the BVN component mixture weights (with label switching accounted for) for the male-to-female (MF) transmission surface. Each number/color represents a unique BVN component.

Web Figure 3: Traceplots for transmission surface type probabilities (subplot (a)) and BVN mixture component weights for FM surface (subplot (b)) throughout the 3000 iterations of the MCMC sampler. Example from a randomly picked simulation run with $N = 600$.

around a value close to ground truth (in this case, MF and FM probabilities have an approximate ratio of 60% versus 40%), which is an indicator of convergence. Also, in subplot (b), we show the traceplot of the sampled mixture weights (of the BVN components) for the MF surface. Again, we see that the weights have stabilized in later iterations after some exploration in the early steps, and there are three dominating components where the two biggest ones have weights around 0.6 and 0.3, respectively, which is very close to the true mixture weight values. Note that in real data analysis, we can also use the traceplots as a graphical diagnostic tool to check MCMC convergence.

Web Appendix C: Supplemental Materials for the Case Study

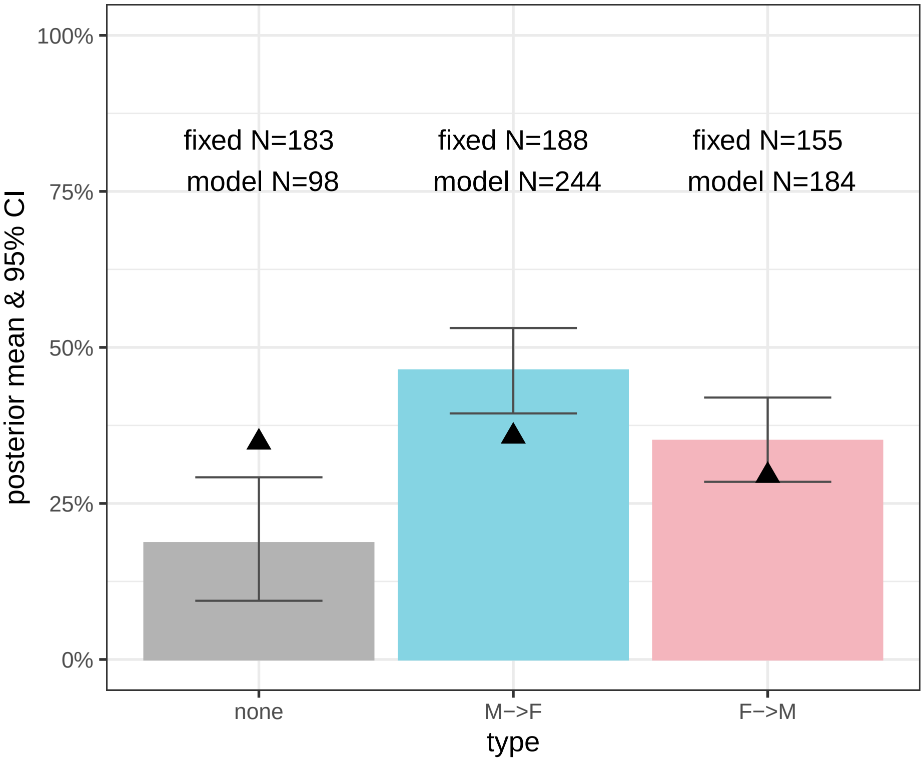
Full analysis with flexible point types

In the full analysis, we pre-specify$\mu_d = 1.5$ and $\mu_{-d} = -1.5$ to imply that the d_i's with $i = 1$ are centered around 0.817 and the d_i's with $i = -1$ are centered around 0.182. We note that analysis results are not considerably sensitive to the choices of these parameters, and through experiments, we've found that altering these values within a reasonable range produces consistent results.

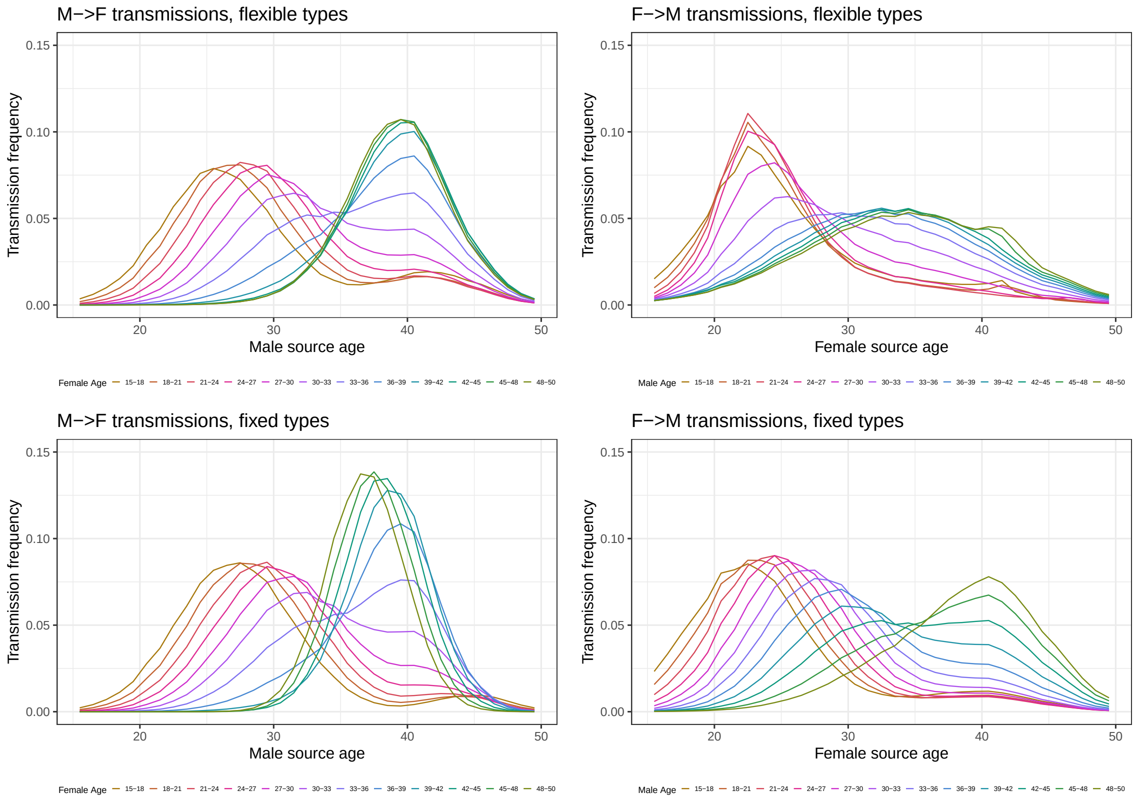
Moreover, for a pair with $d_i = 1$ or 0 (extremely large or small direction scores), in the inference process, we conditionally assign it to the MF or FM surface depending on its ℓ_i score - if c_i is sampled to be nonzero in the inference algorithm, then we would set $c_i = 1$ if $d_i = 1$ and otherwise set $c_i = -1$.

Additional analysis results

Here we provide a graphical illustration as supplement to Table 1 in the main text. As described in Section 5, our model ("**model**") recognized more data points as real transmission events compared to the fixed-type analysis ("**fixed**"), thus effectively leveraging more information from the data.



Web Figure 4: Proportions of data points assigned to different type labels (MF, FM or none) identified by the full model (colored bars), compared to partial analysis with fixed point types (marked by ▲'s). The colored bars show the posterior median proportions of type labels, with black errorbars showing the 95% credible intervals; the number of points assigned to each type (using posterior median numbers for the full model) is also annotated on the plot. The full model consistently identifies significantly more data points as MF or FM transmission events than the fixed model, while the relative ratios between MF and FM event counts are similar between the full and partial analyses (fixed: MF/FM ≈ 1.22, full model: MF/FM ≈ 1.32).



Web Figure 5: Source age distribution for recipients in each 3-year age band. **Top row**: results learned from full analysis without pre-fixing point type labels. **Bottom row**: results learned from the partial analysis with pre-fixed point types. Left column: male sources and female recipients; right column: female sources and male recipients.

Source age distribution by specific recipient age groups

In Web Figure 5, for each 3-year age band of recipients, we plot the relative frequency curve for the age of heterosexual sources. On the left column, we show inferred results for male to female (MF) transmissions – each curve corresponds to each 3-year age group of female recipients (e.g., the leftmost curve represents the age distribution of male sources for women between 15 to 18); on the right column, we show results in the same manner for female to male (FM) transmissions. The full analysis (flexible point types) results are presented in the top row, while partial analysis (fixed point types) results are shown on the bottom.

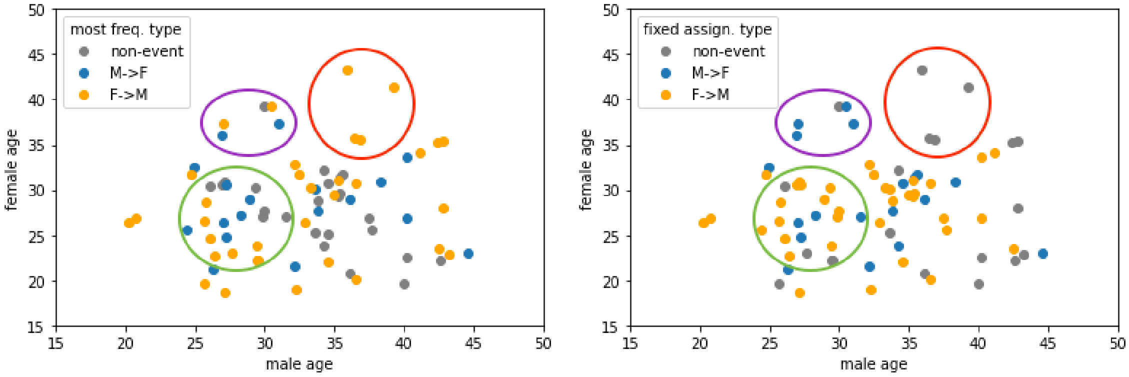
We can see that for different age groups of recipients, the age distributions of their sources could be very different. On the left column (male-to-female transmissions), we can see that younger women may get infected by both younger and older men, but older women main get infected by similarly aged men. On the right side (female-to-male transmissions), we see a more age assortative behavior (young men mainly get infected by young women and older men get infected by older women); however, we also see a bump of female sources at around age 40 for younger men, which also agrees with the general findings discussed in the main paper.

How our model leverages uncertain data points

A central feature of our proposed model is to probabilistically and flexibly classify point types, particularly for the ones with medium direction scores and/or low linkage scores (the uncertain points). Those uncertain points play an important role in inferring the spatial patterns, borrowing information across different types, and helping leverage spatial information for point type allocation.

For a data point i, the posterior mean probability vector $\hat{p}_i = (\hat{p}_{i,0}, \hat{p}_{i,1}, \hat{p}_{i,-1})^T$ for its type indicator c_i represents the level of uncertainty we have about its type. (Here, for example, $\hat{p}_{i,1}$ is the relative frequency of i being assigned to the MF surface across all posterior samples.) *If the entropy of $\hat{p}_i$ is high, then we have high uncertainty about i's type.*

In Web Figure 6, we plot the 77 data points with classification entropy > 0.8 (as a reference, $(0.1, 0.2, 0.7)^T$ would have entropy $= 0.802$). On the left, the points are colored according to the types they are *most frequently* assigned to in the posterior samples (as in the sampling iterations of the full model). On the right, the color represents the type a point would get assigned to given fixed pre-classification (as in the flexbile-type analysis).



Web Figure 6: Flexible-type data points with classification entropy > 0.8 (77 points in total). Left: point color represents the most frequent type assignment; right: point color represents the fixed threshold classification.

One of the most distinct differences between the two plots is the classification of the three points in the upper right corner (inside the red circle). In the full model, they tend to be assigned to the FM surface as their linkage scores are relatively high (close to the 0.6 fixed threshold) and their locations are close to the FM surface mass near $(40, 40)$. Therefore, these points contribute (at least probabilistically) to characterizing the transmission flow patterns from older women to older men.

Another notable difference is within the green circles of the two plots. The full model would, in fact, switch the type allocations for most of those points across the iterations of the inference algorithm, since they are mostly flexible-type points with close-to-threshold direction and linkage scores. (Note that this region is a high-density region for both MF and FM surfaces.) By exploring the possible configurations of data point classification, the model effectively allows spatial information to be shared between different surfaces (especially between MF and FM transmission surfaces) while accounting for the uncertainty in the point types. This is exactly what we hope to achieve through a hierarchical model.

Finally, within the purple circles, the two points at the upper left corner tend to be more frequently classified as FM transmission events, as opposed to the MF classification using fixed thresholds. This is actually in part due to the slightly different spatial patterns of the two transmission directions learned from all other data points: in the FM surface (bottom-middle panel in Figure 4 of the main text), there is a slightly denser mass for older-women to younger-men transmissions, and hence these two points tend to be assigned to the FM surface slightly more frequently. This is an example of how the learned spatial patterns help inform data point classification throughout the iterations of the inference procedure, thanks to joint modeling of the spatial process and the signal distributions as described in Section 2 of the main text.