1 Introduction

The class of models for aggregate relational data that we consider all involve modeling responses from a negative binomial distribution, with the mean μ_{ik} equal to the number of expected response for individual i about knowing a number of people in a group of interest $\mathcal{G}_k$.

$$y_{ik} \sim \text{NegBin}(\omega_k \mu_{ik}, \omega_k) \qquad E(y_{ik}) = \mu_{ik} \qquad Var(y_{ik}) = \mu_{ik} + \frac{\mu_{ik}}{\omega_k}$$

1.1 Random Mixing

The most basic model treats this expectation as simply d_i, the degree of individual i, multiplied by the proportion of the population that is in $\mathcal{G}_k$. Letting $\delta_{jk} = \mathbb{I}\{j \in \mathcal{G}_k\}$, we can derive this expression as follows:

$$\begin{aligned}
\mu_{ik} &= \sum_{j=1}^{d_i} \mathbb{E}[\delta_{jk} | i \to j] \qquad (1)\\
&= \sum_{j=1}^{d_i} \mathbb{P}(j \in \mathcal{G}_k | i \to j)\\
&= d_i \mathbb{P}(j \in \mathcal{G}_k | i \to j)\\
&= d_i \mathbb{P}(j \in \mathcal{G}_k)\\
&= d_i \left(\frac{N_k}{N}\right)
\end{aligned}$$

1.2 Non-Random Gender Mixing

If we do not want to simply assume that the way individual i mixes with alters depends on gender then we can also model non-random gender mixing.

$$\begin{aligned}
\mu_{ik} &= \sum_{j=1}^{d_i} \mathbb{P}(j \in \mathcal{G}_k | g_i, i \to j) \qquad (2)\\
&= d_i \mathbb{P}(j \in \mathcal{G}_k | g_i, i \to j)\\
&= d_i \sum_{g_j} \mathbb{P}(j \in \mathcal{G}_k, g_j | g_i, i \to j)\\
&= d_i \sum_{g_j} \mathbb{P}(j \in \mathcal{G}_k | g_j, g_i, i \to j) p(g_j | g_i, i \to j)\\
&= d_i \sum_{g_j} \mathbb{P}(j \in \mathcal{G}_k | g_j) \rho_{g_i g_j}\\
&= d_i \sum_{g_j} \rho_{g_i g_j} \mathbb{P}(j \in \mathcal{G}_k | g_j)\\
&= d_i \sum_{g_j} \rho_{g_i g_j} \left(\frac{N_{k,g_j}}{N_{g_j}}\right)
\end{aligned}$$

Here $\rho_{g_i g_j}$ is then a latent variable that can be inferred as the mixing rate between egos of gender g_i and alters of gender g_j.

1.3 Non-Random Age and Gender Mixing

If we also believe that egos of certain ages might mix differently with alters of other ages, then, we can also model non-random age and gender mixing. Suppose the each individual i belongs to some age category a_i (e.g. 0-17, 18-24, etc.).

$$
\begin{aligned}
\mu_{ik} &= \sum_{j=1}^{d_i} \mathbb{P}(j \in \mathcal{G}_k | a_i, g_i, i \to j) \qquad (3)\\
&= d_i \mathbb{P}(j \in \mathcal{G}_k | a_i, g_i, i \to j)\\
&= d_i \sum_{a_j, g_j} \mathbb{P}(j \in \mathcal{G}_k, a_j, g_j | a_i, g_i, i \to j)\\
&= d_i \sum_{a_j, g_j} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j, a_i, g_i, i \to j) p(a_j, g_j | a_i, g_i, i \to j)\\
&= d_i \sum_{a_j, g_j} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j) \rho_{(a_i, g_i)(a_j, g_j)}\\
&= d_i \sum_{a_j, g_j} \rho_{(a_i, g_i)(a_j, g_j)} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j)\\
&= d_i \sum_{a_j, g_j} \rho_{(a_i, g_i)(a_j, g_j)} \left(\frac{N_{k, a_j, g_j}}{N_{a_j, g_j}} \right)
\end{aligned}
$$

Here $\rho_{(a_i, g_i)(a_j, g_j)}$ is then a latent variable that can be inferred as the mixing rate between egos of age category a_i, gender g_i and alters of age category a_j, gender g_j.

1.4 Issues

In our experiments, the non-random age-gender mixing model suffered from bias/variance and identifiability issues because the mixing rate parameters lacked constraints (other than summing to 1). In an effort to resolve this issue, we now propose a model with more structure and fewer parameters.

2 Kernel Model

In this new approach, we first assume that age is continuous ($a_i \in (-\infty, \infty)$) rather than just binning age into categories. This allows us to model the mixing rate of an ego with age a_i with an alter of age a_j as a Gaussian kernel defined smoothly over all possible a_j.

Additionally, we also model the alter degree d_j for the first time. Interestingly, this modeling yields different results depending on whether we set up the model from the perspective of the alter or from the perspective of the ego (as we've done in all previous models above).
Modifying the derivation from 1.3, the mean can be derived as follows:

$$\begin{aligned}
\mu_{ik} &= d_i \mathbb{P}(j \in \mathcal{G}_k | a_i, g_i, i \to j) \qquad (4)\\
&= d_i \sum_{g_j} \int_{a_j} \mathbb{P}(j \in \mathcal{G}_k, a_j, g_j | a_i, g_i, i \to j) da_j \\
&= d_i \sum_{g_j} \int_{a_j} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j, a_i, g_i, i \to j) p(a_j, g_j | a_i, g_i, i \to j) da_j \\
&= d_i \sum_{g_j} \int_{a_j} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j) p(a_j | g_j, a_i, g_i, i \to j) p(g_j | a_i, g_i, i \to j) da_j \\
&= d_i \sum_{g_j} \int_{a_j} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j) p(a_j | g_j, a_i, g_i, i \to j) p(g_j | g_i, i \to j) da_j \\
&= d_i \sum_{g_j} p(g_j | g_i, i \to j) \int_{a_j} \mathbb{P}(j \in \mathcal{G}_k | a_j, g_j) p(a_j | g_j, a_i, g_i, i \to j) da_j \\
&= d_i \sum_{g_j} \rho_{g_i g_j} \int_{a_j} \mathcal{N}(a_j | \mu_{g_j,k}, \sigma^2_{g_j,k}) \mathcal{N}(a_j | a_i, \lambda_{g_i g_j}) da_j \\
&= d_i \sum_{g_j} \rho_{g_i g_j} \frac{e^{-\frac{(a_i - \mu_{g_j,k})^2}{2(\lambda_{g_i g_j} + \sigma^2_{g_j,k})}}}{\sqrt{2\pi(\lambda_{g_i g_j} + \sigma^2_{g_j,k})}}
\end{aligned}$$

Here $\rho_{g_i g_j}$ is the same latent variable as in 1.2 that can be inferred as the mixing rate between egos of gender g_i and alters of gender g_j. Additionally, $\lambda_{g_i g_j}$ is a latent variable that can be inferred as the kernel bandwidth of the age mixing kernel. Essentially, small values of $\lambda_{g_i g_j}$ indicate that egos of gender g_i only know alters of gender g_j that are close to the ego in age (whereas larger values indicate the egos know alters of a wide range of ages, not necessarily just those close in age).

Additionally, $\mu_{g_j,k}$ and $\sigma^2_{g_j,k}$ are just estimated from the population data about group G_k. These values are analogous (though not exactly equal) to the mean and standard deviation of the ages of alters in group G_k with gender g_j.

3 Simulation

3.1 Data

We simulate responses to questions about 12 names using estimated age means/variances (for each name) and simulated respondent degrees, name overdispersions, kernel lengthscales, and gender mixing rates in an attempt to determine whether our model can recover the parameters.

$$d_i \sim \log \mathcal{N}(6.2, 0.5) \qquad \frac{1}{\frac{1}{\omega_k} + 1} \sim Beta(10, 2)$$

Var	Linda	Jen.	Karen	Kim.	Emily	Steph.	Mark	Jacob	Kevin	Kyle	Adam	Bruce
μ_k	63.3	37.4	56.1	39.8	28.0	35.6	49.2	22.5	38.8	25.6	31.0	62.4
σ_k	10.4	10.6	13.9	13.0	22.9	14.3	14.9	18.0	16.2	10.8	16.2	16.7
ω_k	4.23	8.42	7.65	3.83	9.79	10.91	5.01	2.80	2.22	2.60	12.69	4.95

$$\lambda = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 225 & 144 \\ 100 & 256 \end{pmatrix}$$

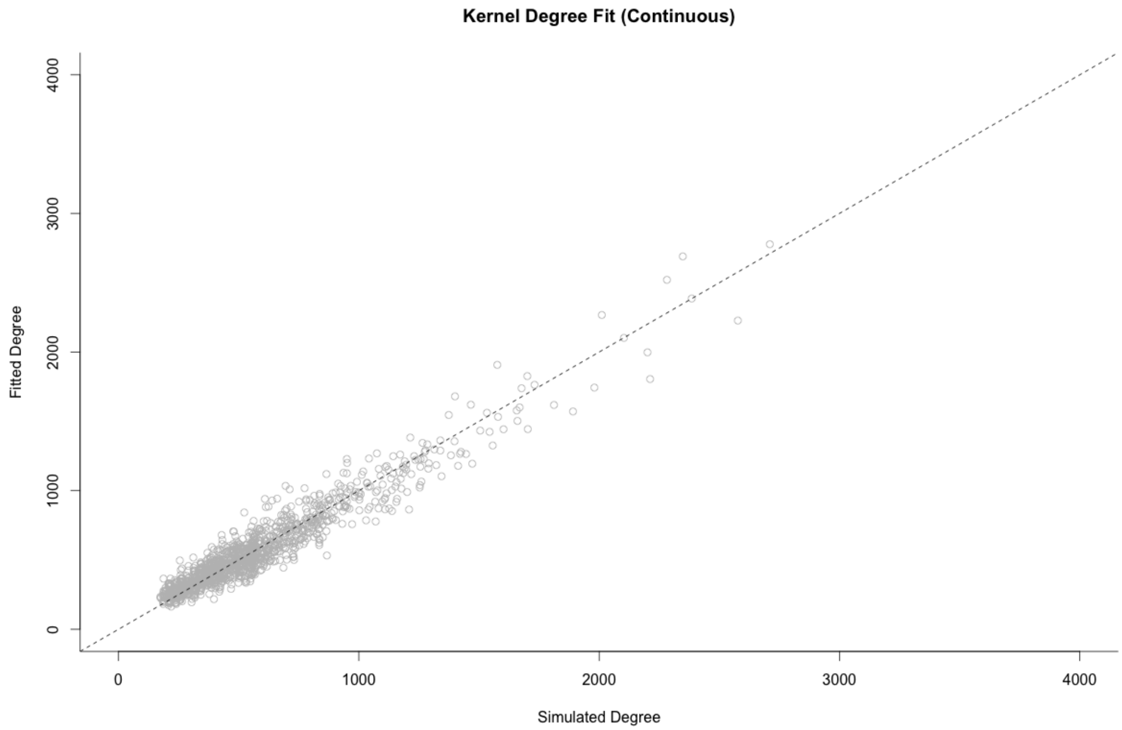
$$\rho = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.6 & 0.4 \\ 0.45 & 0.55 \end{pmatrix}$$

With these "true" parameter values, we then simulate two data sets: one from the continuous model and one from the discrete model. Our results outlining how well the continuous model recovers the parameters from each data set are then discussed in the next sections. In the final section, we discuss and interpret the results of fitting the continuous model on real data.

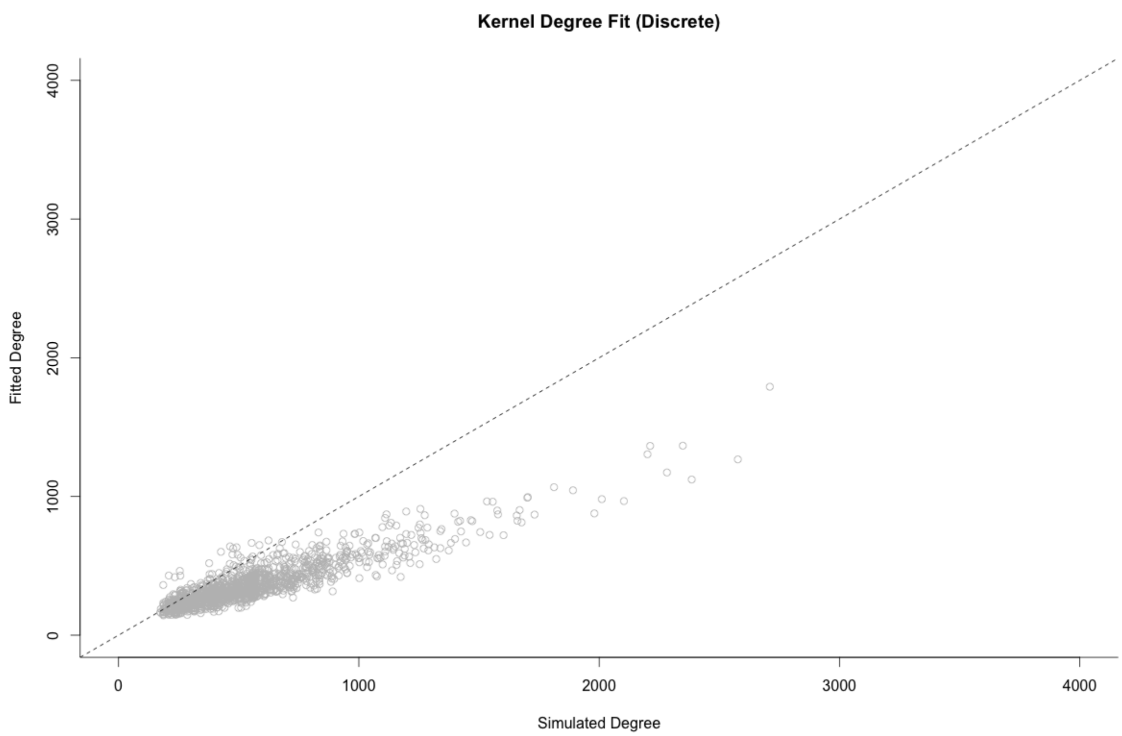
3.2 Results

Respondent Degrees (d_i)

The degrees are recovered well from the continuous model data, with a correlation of 0.97 between the simulated values and their posterior means.

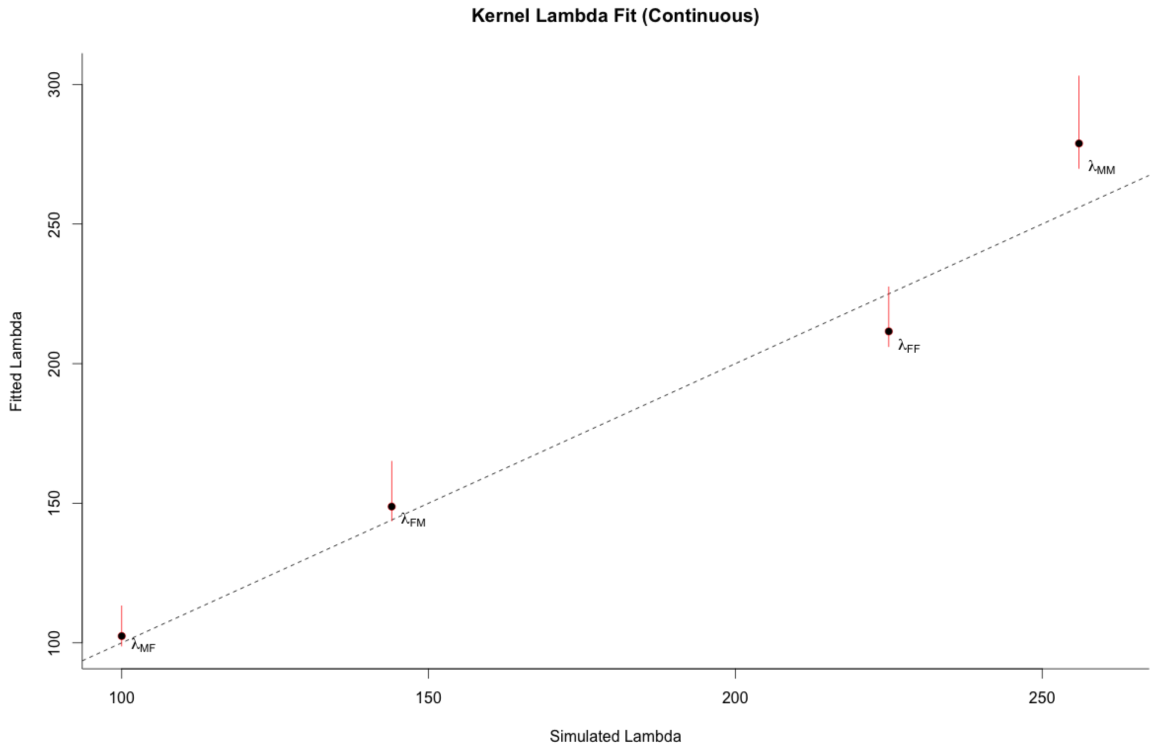


The degrees are largely underestimated for the discrete model data with a strangely linear offset.

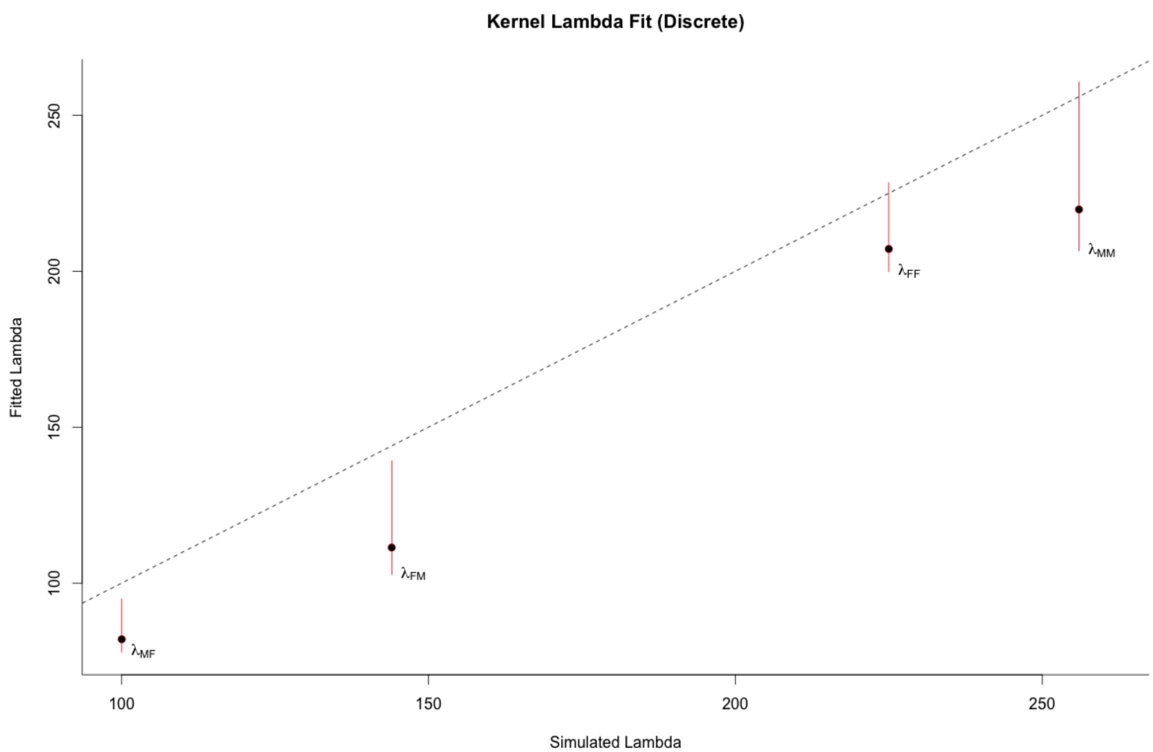


Kernel Lengthscale ($\lambda_{g_i g_j}$)

The continuous model does a great job of recovering the lambdas from the continuous model data.

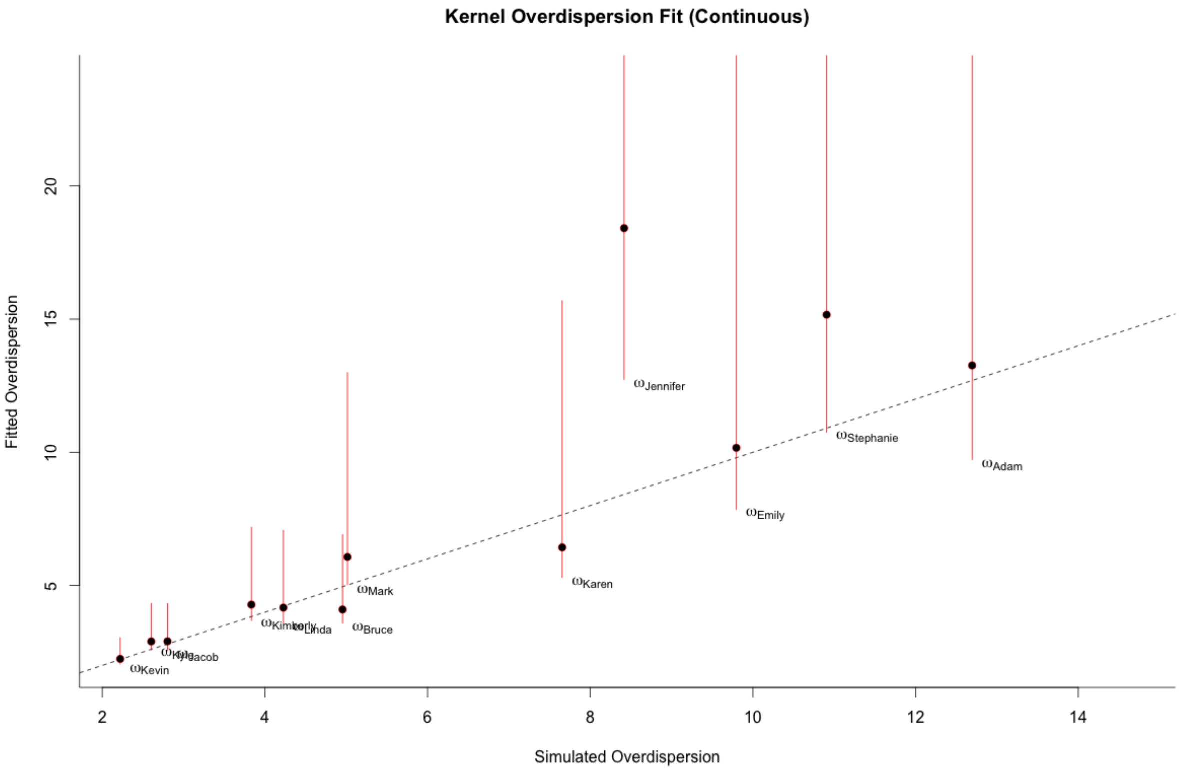


The continuous model underestimates the lambdas from the discrete model data.

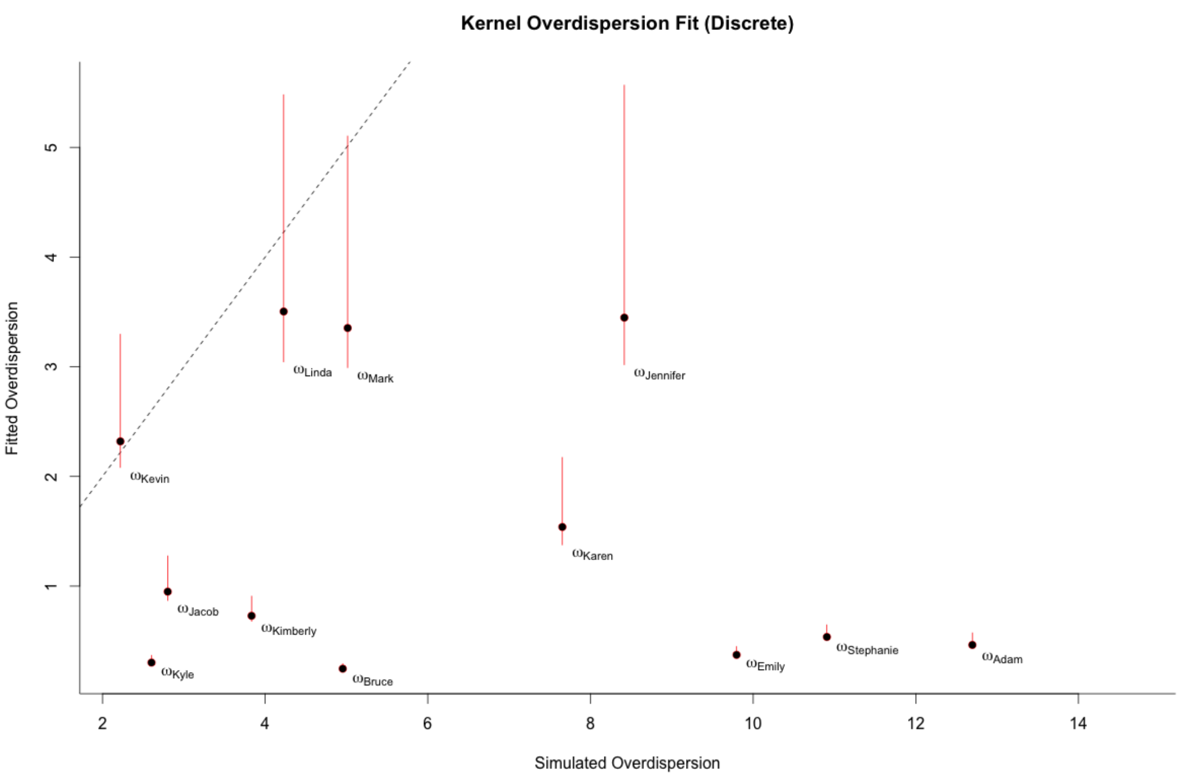


Name Overdispersions (ω_k)

The overdispersions are also mostly contained within the central 95% of their posterior distributions, but posterior uncertainty is quite large. Here the posterior median is preferred to the posterior mean.

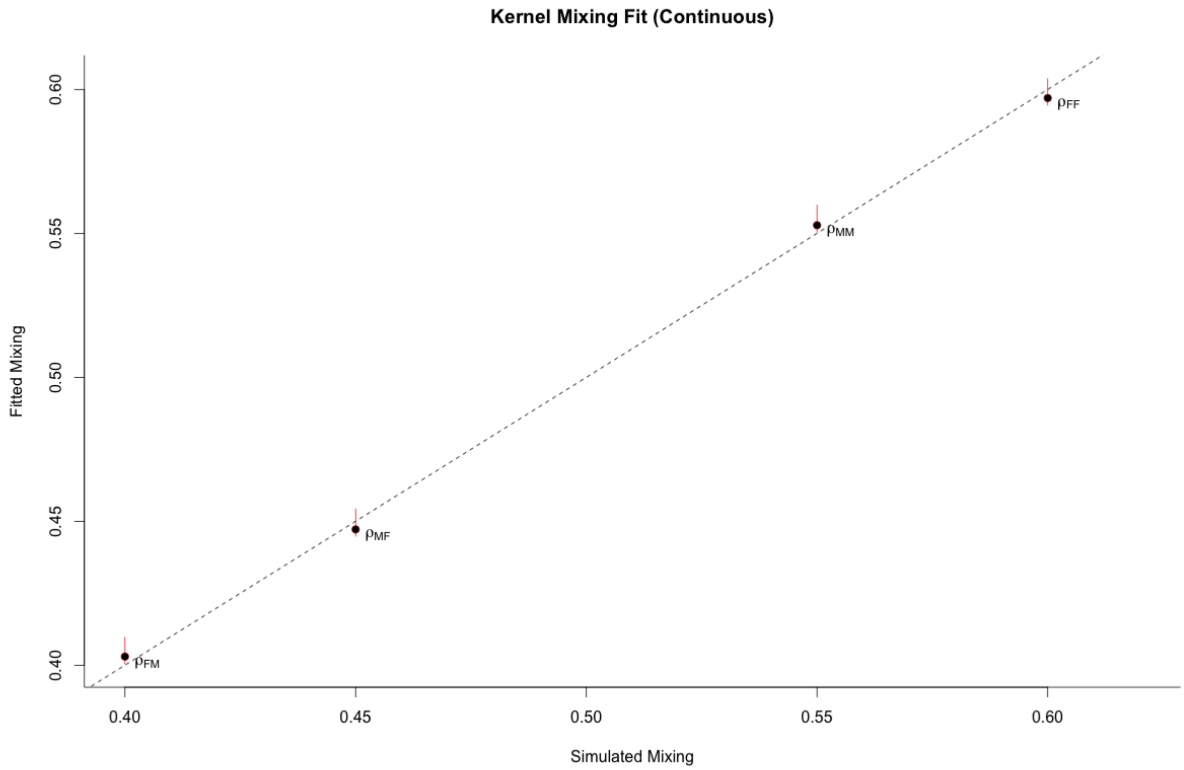


The overdispersions are poorly recovered from the discrete model data.

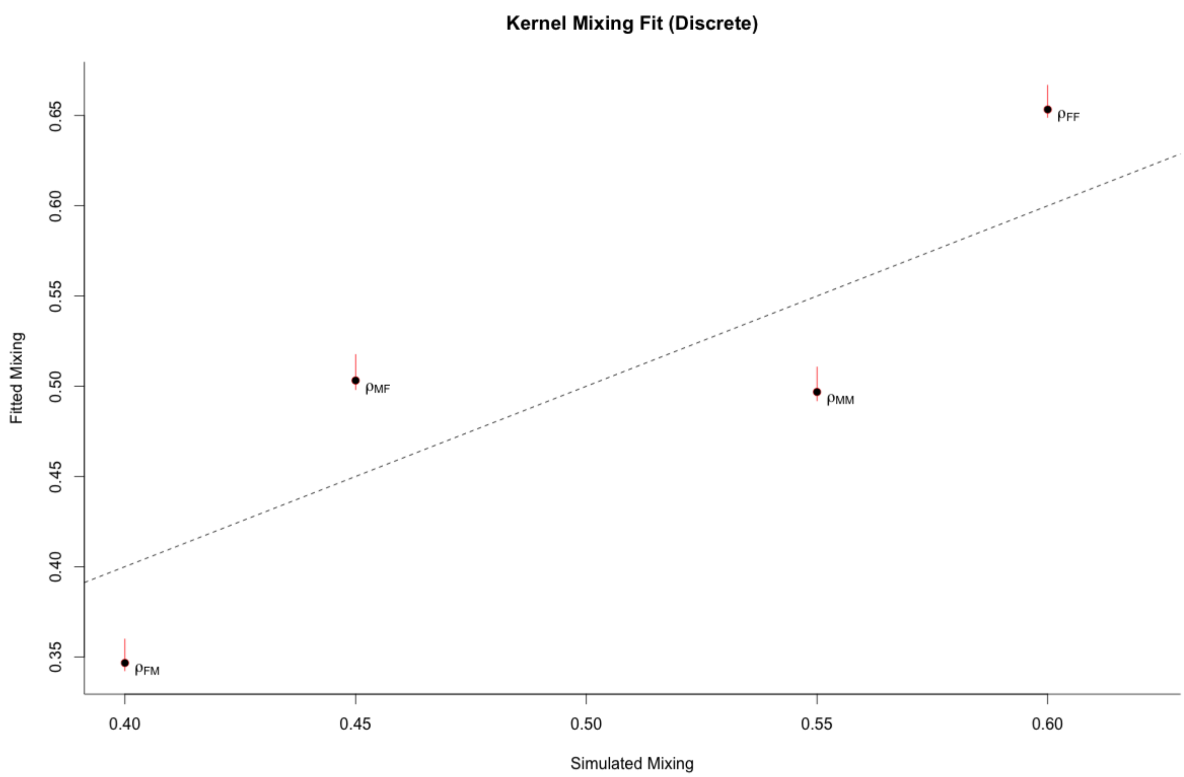


Gender Mixing Rates ($\rho_{g_i g_k}$)

Lastly, the gender mixing rates are recovered very well from the continuous model data.



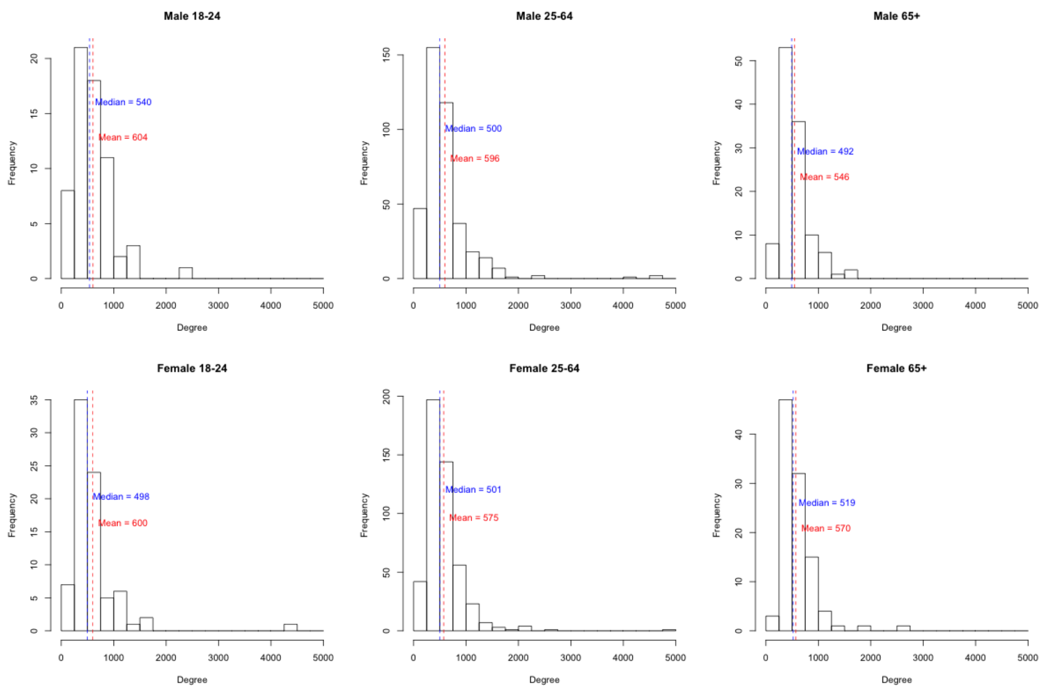
However, the mixing is poorly recovered from the discrete model data.



4 Omni Data Results

Respondent Degrees

The continuous model estimates imply decreasing network size by age for men, but possibly increasing by age for women.



Kernel Lengthscales

The length scales estimated from the actual data are quite large (compared to the ones we choose for the simulated data), implying very flat kernels. The important distinction here, then, is perhaps the relative size of the lengthscales.

$$\lambda_{BAYES} = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 5092 & 8334 \\ 4545 & 13710 \end{pmatrix}$$

Indeed, it seems that the female to female kernel is much tighter than the male to male kernel, implying that women tend to know a narrow age range of other women but men tend to know a wide age range of other men. Alternatively, the female to male kernel is much wider than the male to female kernel, implying that women know a wider age range of men while men known a very narrow age range of women.

Gender Mixing Rates

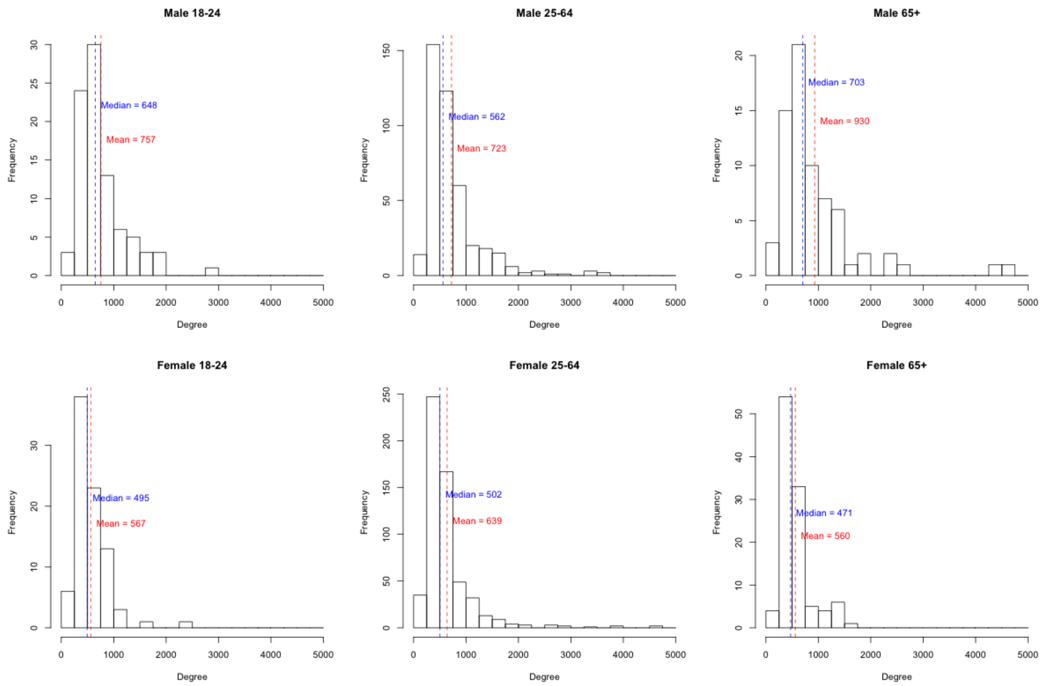
The gender mixing rates seem to correlate with the kernel lengthscales. Here we see that about 54% of a female's network is female, while 61% of a male's network is male. This also implies that females mix more with males than the other way around.

$$\rho_{BAYES} = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.54 & 0.46 \\ 0.39 & 0.61 \end{pmatrix}$$

5 McCarty Data Results

Respondent Degrees

The McCarty data imply network size is a minimum for middle aged men, while it is a maximum for middle aged women.



Kernel Lengthscales

The relative lengthscales for the McCarty data are somewhat similar to the relative lengthscales estimated from the Omni data.

$$\lambda_{BAYES} = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 1150 & 767 \\ 1827 & 1740 \end{pmatrix}$$

Namely, we see once again that the female to female kernel is much tighter than the male to male kernel. Additionally, we see once again that the female to male kernel is wider than the male to female kernel.

Gender Mixing Rates

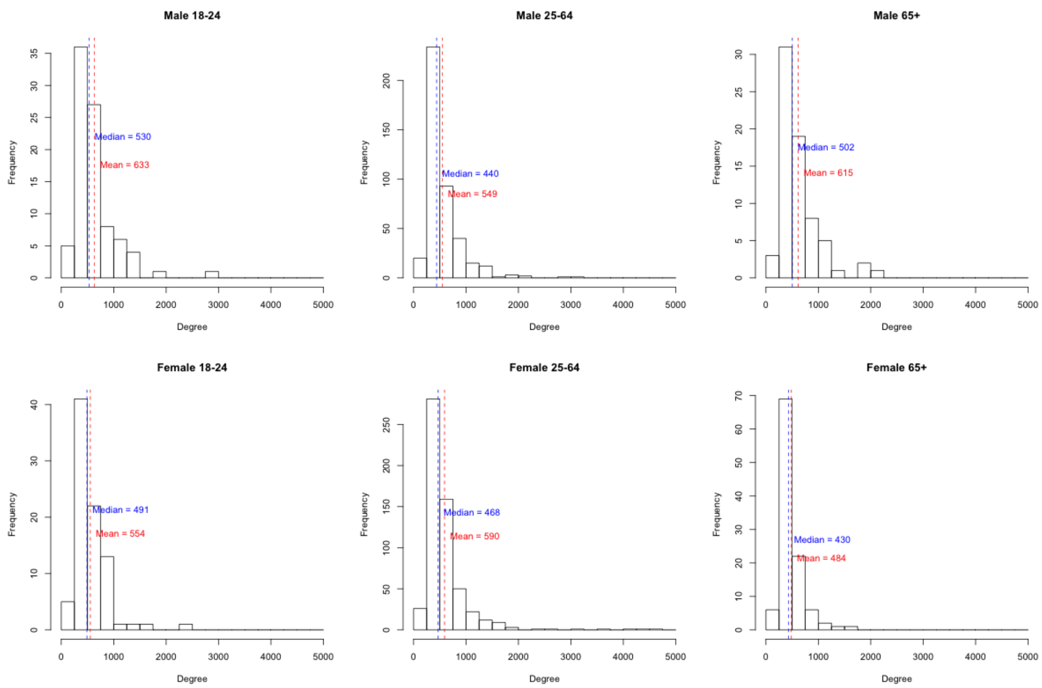
The gender mixing rates, however, are quite surprising, and don't correlate well with the kernel lengthscales.

$$\rho_{BAYES} = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.27 & 0.73 \\ 0.31 & 0.69 \end{pmatrix}$$

6 McCarty Data Results (Without Michael, Robert, David)

Respondent Degrees

The McCarty data imply network size is a minimum for middle aged men, while it is a maximum for middle aged women.



Kernel Lengthscales

The relative lengthscales for the McCarty data are somewhat similar to the relative lengthscales estimated from the Omni data.

$$\lambda_{BAYES} = \begin{pmatrix} \lambda_{FF} & \lambda_{FM} \\ \lambda_{MF} & \lambda_{MM} \end{pmatrix} = \begin{pmatrix} 2514 & 1483 \\ 5064 & 1956 \end{pmatrix}$$

Namely, we see once again that the female to female kernel is much tighter than the male to male kernel. Additionally, we see once again that the female to male kernel is wider than the male to female kernel.

Gender Mixing Rates

The gender mixing rates, however, are quite surprising, and don't correlate well with the kernel lengthscales.

$$\rho_{BAYES} = \begin{pmatrix} \rho_{FF} & \rho_{FM} \\ \rho_{MF} & \rho_{MM} \end{pmatrix} = \begin{pmatrix} 0.30 & 0.70 \\ 0.31 & 0.69 \end{pmatrix}$$