Modelling TB transmission in clusters of maximum size two

Shannon Gallagher

December 13, 2019

Abstract. We examine the transmission of TB of clusters of individuals of maximum size two.
Keywords: tuberculosis; difficult problems;

1 Modelling the spread of TB

There are many elements to consider in disease transmission of pulmonary tuberculosis (TB). First of all, we have to consider time elements vs. generations. For example consider the case where the primary case, person 0 infects person 1 at time 0. Person 1 then goes on to infect person 2 at time 1. Finally, person 0 infects person 3 at time 2. Person 0 is the primary case and persons 1 and 3 are in the first generation, as they are one contact removed from person 0. Person 2 is in generation 2 yet was infected before person 3. This situation is illustrated in Figure 1.

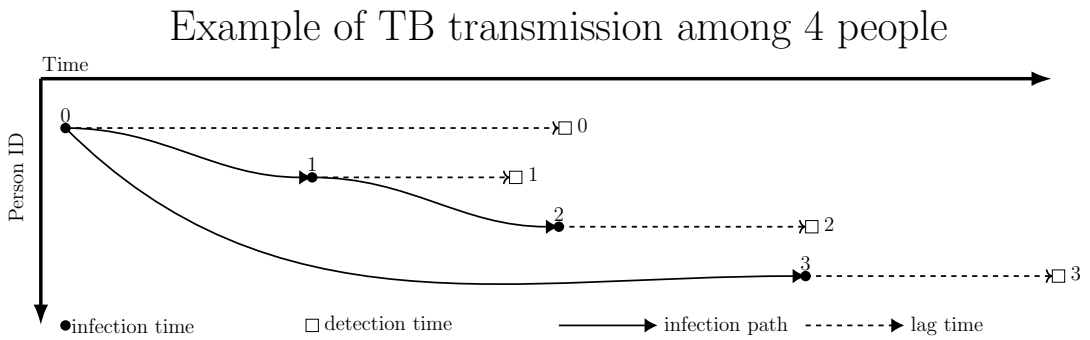

Figure 1: Graphical depiction of spread of TB among four hypothetical people. The x-axis represents time and each row in the y-axis represents the status of a person. A circle represents the infection time of TB and the square the detection time. The solid lines represent a transmission path between people, and the dashed lines denote the lag between detection and infection time.

In TB transmission, the latent period is especially important as the detection time of a secondary case can occur before the detection time of the primary case. This is illustrated in Fig. 1 where although person 0 is infected before person 1, we detect person 1's infection first.

Another covariate of importance in our analysis is the smear status (+/-) of the individual. The smear status is the result of a test given to determine whether the individual currently has TB. We assume that the date of the smear result is the same as the date of detection.

We would like to estimate:

- the effect of smear in transmitting TB
- the transmission sequence of individuals within a cluster

An ideal, complete set of TB transmission data would include infection time, detection time, and infector ID, as shown in Table 1. This ideal data set would allow us to infer information both about the time and generation of the transmission of TB. For example, we could assess the importance of smear status by modelling the probability that the secondary case is infected as a function of the primary case's smear status and gap in infection time between the primary and secondary case. The second item would be trivial, as the data contains the transmission sequence.

Table 1: Ideal TB transmission data.

Person	**Smear**	**Inf. Time**	**Det. Time**	**Infector ID**
0	+	-1	2	–
1	+	0	1	0
2	-	1	3	1
3	+	2	4	0

Unfortunately, within each cluster, we only examine the smear status and detection time of each individual, which makes both items considerably more difficult to answer. Due to this increased difficulty, we first begin to answer these questions in a special case where the clusters are of maximum size two. That is, in each cluster there is only one or two individuals who are infected with TB. Moreover, we assume that given the cluster has two individuals, the primary case transmitted TB to the secondary case, and the primary case received the disease from an exogenous source. That is for indiviudals i and j in cluster k, we assume that one of either $i \to j$ or $j \to i$ occurred.

1.1 Models for clusters of maximum size two

We examine the following models. Let G_{ij} be the graph where i transmits the disease to j ($i \to j$). Let $n_k \in \{1, 2\}$ be the size of the cluster k for $k = 1, \ldots, K$.

For clusters of size $n_k = 1$, the probability of *not* transmitting the disease onward is

$$T_k \sim \text{Bernoulli}(\pi_{k,1})$$
$$\pi_{k,i} = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{k,i})}}$$

where $T_k \sim \text{Bernoulli}(\pi_{k,i})$ is a Bernoulli draw indicating whether the individual transmitted the disease onward, $\pi_{k,i}$ is the probability of individual i transmitting the disease to another, and $X_{k,i}$ the smear status of individual i.

For clusters of size two ($n_k = 2$), we know either person 1 transmitted it to 2 or *vice versa*. Then the probability that the infection order is i, j is

$$P(\text{inf } 1 = i, \text{inf } 2 = \text{j}) = \pi_{k,1} w_{k,1} + \pi_{k,2} w_{k,2},$$

where w_1 and w_2 are the weights of person 1 and person 2 transmitting the disease.

The likelihood then for clusters of size 1 and clusters of size two is

$$\mathcal{L}(w_1, w_2, \beta_0, \beta_1) = \prod_{k \in \{k:n_k=1\}} (1 - \pi_{k,1}) \prod_{k \in \{k:n_k=2\}} (\pi_{k,1} w_{k,1} + \pi_{k,2} w_{k,2}) \tag{1}$$

We examine different weighting structures as functions of the gap time, which we define as difference in detection time between the individuals, $\delta_{k,1} = d_{k,2} - d_{k,1}$ and $\delta_{k,2} = -d_{k,1}$.

Weights 1 – No gap: Our first weighting structure places equal weight on both individuals being the infector. As such, the gap time is not relevant,

$$w_{k,i} = \frac{1}{2}. \tag{2}$$

Weights 2 – G_{ij} known: If we were somehow given the infection order, we could maximize the likelihood to estimate whether the smear status is important in modelling the transmission. Let $\mathcal{I}\{\cdot\}$ be the indicator function of its arguments. The weights are

$$w_{k,i} = \mathcal{I}\{G_{ij} = g_{12}\}, \tag{3}$$

which puts all the weight on the individual actually transmitting the disease. This reduces the problem to a regular logistic regression problem of predicting transmission probability given smear status. This situation is mostly theoretical as we are unlikely to actually know the transmission sequence but allows us to better test our model under controlled simulations.

Weights 3 – Normally distributed gap time: We let the weights be the *pdf* of a univariate normal distribution with mean gap timeμ and variance σ^2,

$$w_{k,i} = \phi(\delta_{k,1}). \tag{4}$$

When $\delta_{k,1} > 0$, then the detection time for individual 1 occurred before individual 2 and, intuitively, would expect that it is more likely that $1 \to 2$ than $2 \to 1$. If $\mu > 0$, then $w_{k,1} > w_{k,2}$ and so this weighting structure reflects this intuition.

Weights 4 – Standardized normals: We let the weights be the standardized version of the weights in Eq. (4)

$$\begin{aligned} w_{k,i} &= \frac{\phi(\delta_{k,1})}{\phi(\delta_{k,1}) + \phi(\delta_{k,2})} \\ w_{k,1} &= \frac{1}{1 + e^{-2\delta_{k,1}\nu}}. \\ w_{k,2} &= \frac{e^{-2\delta_{k,1}}}{1 + e^{-2\delta_{k,1}\nu}}, \end{aligned} \tag{5}$$

where $\nu = \mu/\sigma^2$. The reason we standardize the normal weights is insure $w_{k,1} + w_{k,2} = 1$ and so that we can more easily compare models using a statistical test (as explained later).

In summary, we look at four different weight structures. The first is equally weighting transmission sequences of $i \to j$ and $j \to i$. The parameters are $\theta_1 = (\beta_0, \beta_1)$ The second set of weights looks only at the transmission probability as a function of smear status and assumes we know the transmission order. The parameters are $\theta_2 = (\beta_0, \beta_1)$ The third set weights the probability of transmission by the gap time, where we assume the gap time is normally distributed. The parameters are $\theta_3 = (\beta_0, \beta_1, \mu, \sigma^2)$. The fourth set modifies the third set so that $w_{k,1} + w_{k,2} = 1$. The parameters are $\theta_4 = (\beta_0, \beta_1, \nu)$ where $\nu = \mu/\sigma^2$. Note that in the fourth set of weights, μ and σ^2 are not identifiable because we combine the two into a ratio estimate.

Simulations indicate that for all four of these weighting structures, the MLEs for β_0 and β_1 are consistent (see an example in Fig. 2).

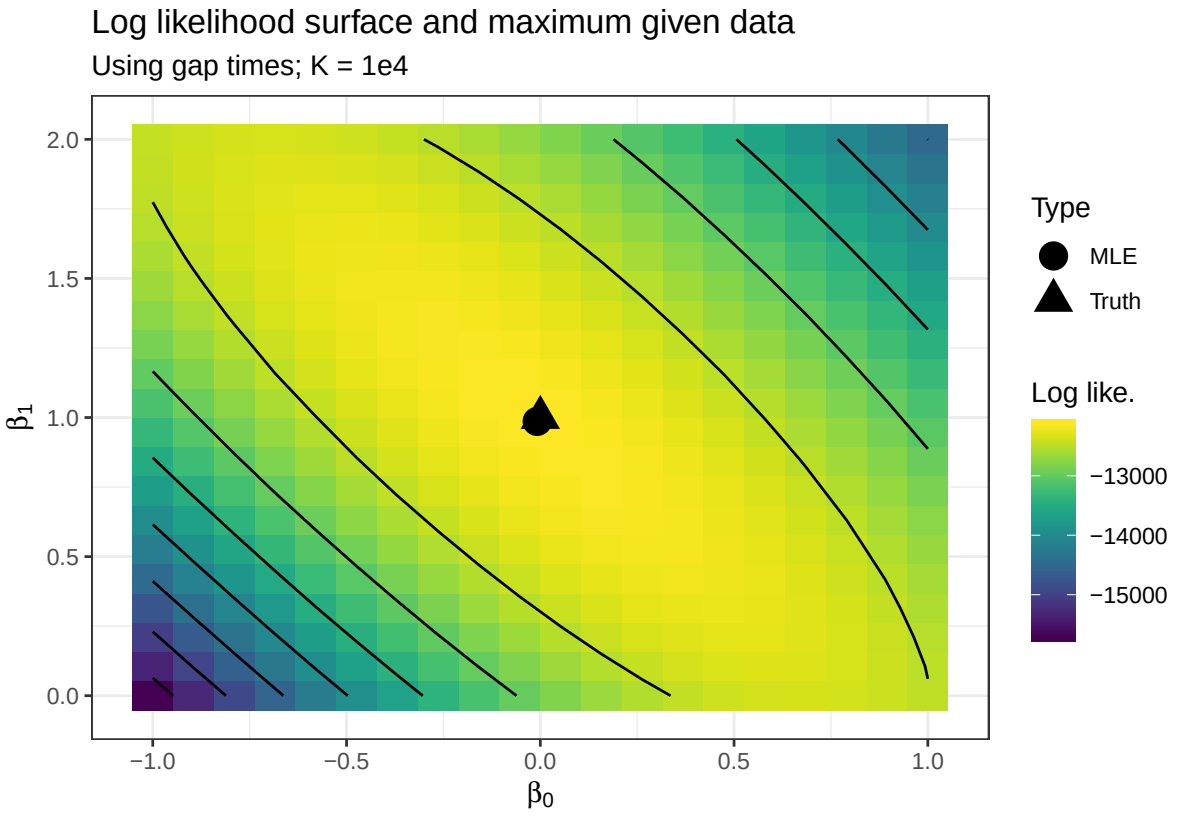

Figure 2: MLE surface from simulations for β_0 and β_1 from weighting structure 3.

1.2 Model selection: gap vs. no gap

Of the weighting schemes we presented, only the weights in Eqs. (2), (4), and (5) are relevant to our problem, as we will not observe the true transmission sequence (e.g. weights in Eq. (3)). Of these weighting schemes, only the likelihoods in Eqs. (2) and (5) are comparable to one another. The likelihood of weighting structure 4 with parameters $\theta_4 = (\beta_0, \beta_1, \nu)$ is equal to the likelihood of weighting structure 1 with parameters $\theta_1 = (\beta_0, \beta_1)$ when $\nu = 0$. The parameter $\nu = \mu/\sigma^2$ will be zero when either $\mu = 0$ or $\sigma^2 \to \infty$. That means, intuitively, if the gap time has no relevance to the problem then $\nu = 0$ and we can use the simpler model without gap time.

We can compare the models with weighting structure 1 and 4 using a likelihood ratio

test. Since $\Theta_1 \subseteq \Theta_4$ we can test

$$\begin{aligned} \mathrm{H}_0 &: \nu = 0 \\ \mathrm{H}_A &: \nu \neq 0 \end{aligned}$$

Since the MLEs of β_1, β_2 and ν are in the interior, then we can apply Wilke's theorem, which says that -2($\log \frac{L_0}{L_A}$) $\rightsquigarrow \chi^2(1)$, wehre L_0 is the maximum likelihood under the null and L_A is the maximum likelihood under the alternative. Thus, we are able to test whether the gap time and smear status matters in the spread of TB as opposed to smear status alone.

2 Simulating data

In order to asess the validity of our models, we find it helpful to first simulate data. We generate data through the following process, which is summarized algorithmically in Alg. 1. We first fix the number of clusters K. Then, without loss of generality, we set the initial *infection* time for each individual $A_{k,1}$ to time $t_{k,1} = 0$. We then determine the *detection* time for individual 1 with a draw from a distribution F_1, $d_{k,1} \sim F_1$. Following that, we determine the individual smear status $X_{k,1}$ for individual 1 in cluster k, which is an i.i.d. Bernoulli draw with probability p,

$$X_{k,i} \sim \text{Bernoulli}(p).$$

We then determine the probability $\pi_{k,1}$ that the agent will transmit the disease to another. Here, $\pi_{k,1}$ is a function of the smear status, namely,

$$\begin{aligned} \pi_{k,i} &= f(\beta_0, \beta_1, X_{k,i}) \\ &= \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_{k,i})}}. \end{aligned}$$

We then determine if individual $A_{k,i}$ transmits the disease based on an i.i.d. Beroulli draw T_k,

$$T_k \sim \text{Bernoulli}(\pi_{k,1})$$

If $T_k = 0$, then the individual does not transmit the disease. If $T_k = 1$, then we initialize individual $A_{k,2}$, the secondary case of individual $A_{k,1}$. Individual $A_{k,2}$'s *infection* time is determined by a draw from a distribution F_2,

$$t_{k,2} \sim F_2. \tag{6}$$

The detection time is the infection time plus some noise drawn from distribution F_1.

$$d_{2,k} | t_{k,2} \sim t_{k,2} + F_1.$$

Finally, the gap time δ_k for cluster k is determined by the difference in detection times between the two individuals, $\delta_k = d_{k,2} - d_{k,1} = d_{k,2}$ or is NA if there is only one individual

in the cluster.

for $k = 1, \ldots, K$ **do**
- Initialize agent $A_{k,1}$
- Initialize **infection** time to zero: $t_{k,1} = 0$
- Determine the **detection** time $d_{k,1}|t_{k,1} \sim F_1$
- Determine smear status:
- $X_{k,1} \sim \text{Bernoulli}\,(p)$
- Determine if agent transmits TB:
- $\pi_{k,1} = f(\beta_0, \beta_1, X_{k,1})$
- $T_k \sim \text{Bernoulli}(\pi_{k,1})$
- **if** $T_k == 1$ **then**
  - Initialize agent $A_{k,2}$
  - Determine **infection** time: $t_{k,2} \sim F_2$
  - Determine **detection** time: $d_{k,1}|t_{k,2} \sim t_{k,2} + F_1$
  - Determine smear status:
  - $X_{k,2} \sim \text{Bernoulli}\,(p)$
  - Determine gap time:
  - $\delta_k = d_{k,2} - d_{k,1}$
- **end**

end

Algorithm 1: Process for simulating TB transmission data

We illustrate the data generating process in Alg. 1 with the following example. We set the number of clusters to $K = 20$, and let $p = 0.7$, $\beta_0 = 0$, $\beta_1 = 1$, $F_1 \overset{d}{=} N(\mu_L = 1, \sigma_L = 0.5)$ and $F_2 \overset{d}{=} N(\mu_W = 1, 0.2)$. Note that because we assume draws from F_1 and F_2 are independent from one another, then the gap time $\delta_k \overset{d}{=} F_1$. We show the clusters visually in Figure 3. Each row in the plot shows a cluster of one or two individuals. The first individual is colored blue and the second individual is colored in orange. The triangles represent infection time and the circles are detection time. The lines are the gap times between infection times (dotted) and detection times (solid). Note that for cluster 4, the second individual infected was detected before the first individual.

Note that when we use normal distributions for F_1 and F_2, that it is possible (but perhaps unlikely depending on our parameter choices) to simulate "non-sensical" data. For example, we could possibly sample an infection time for the the secondary case that is before the primary case, which is physical impossibility. For the moment, we ignore this feature.

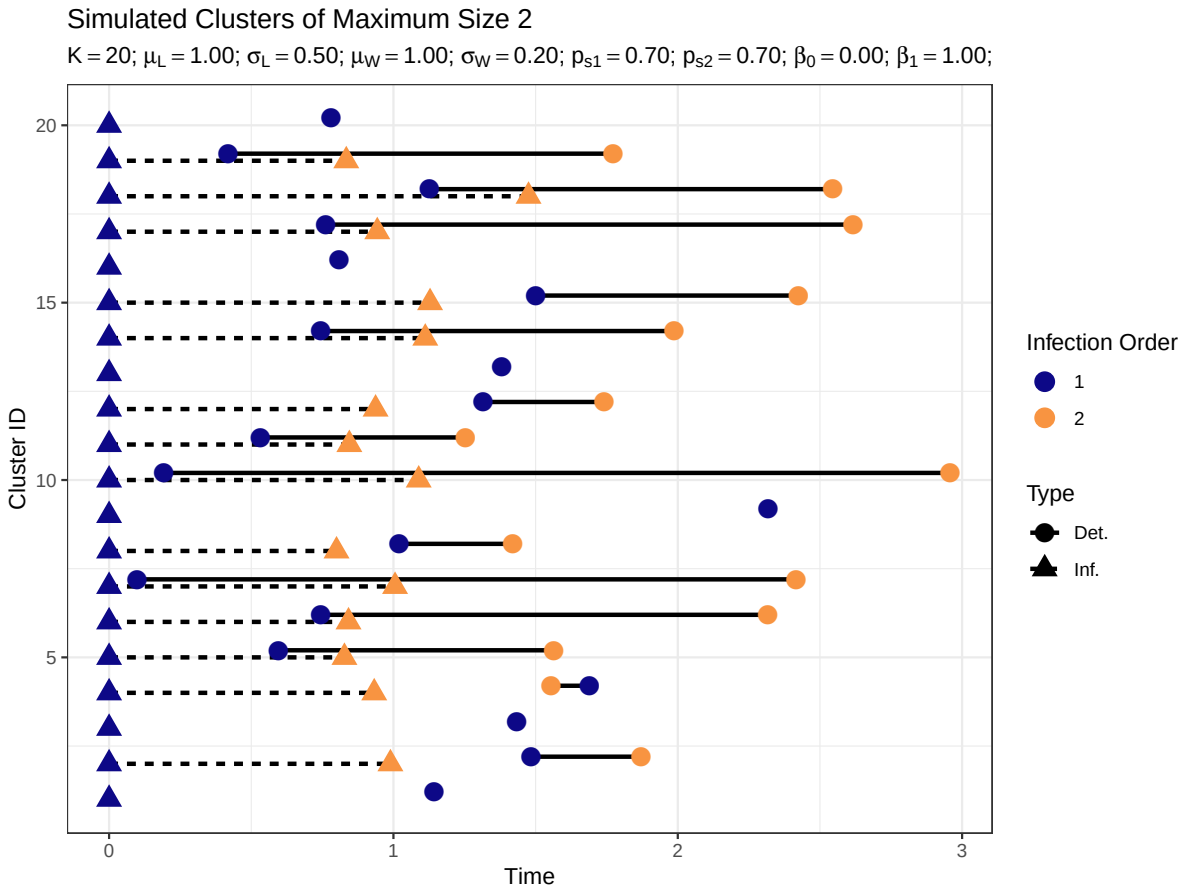

Figure 3: Visual depiction of simulated clusters.

3 Application: TB in MD

The data consists of recorded suspected cases of TB in Maryland from 2003-2009. There are 1137 incidence cases with 70 covariates. Each row in the data set consists of a suspected TB patient. For demographic information, there is age, sex, zipcode, ethnicity, HIV status, homelessness status, and more. In addition, there is a variety of information about the two different tests (smear and NAAT/MTD) along with the genotype PCR clustering results, which are present for 65.79% of the cases. The incidence map of infection is shown in Figure 4.

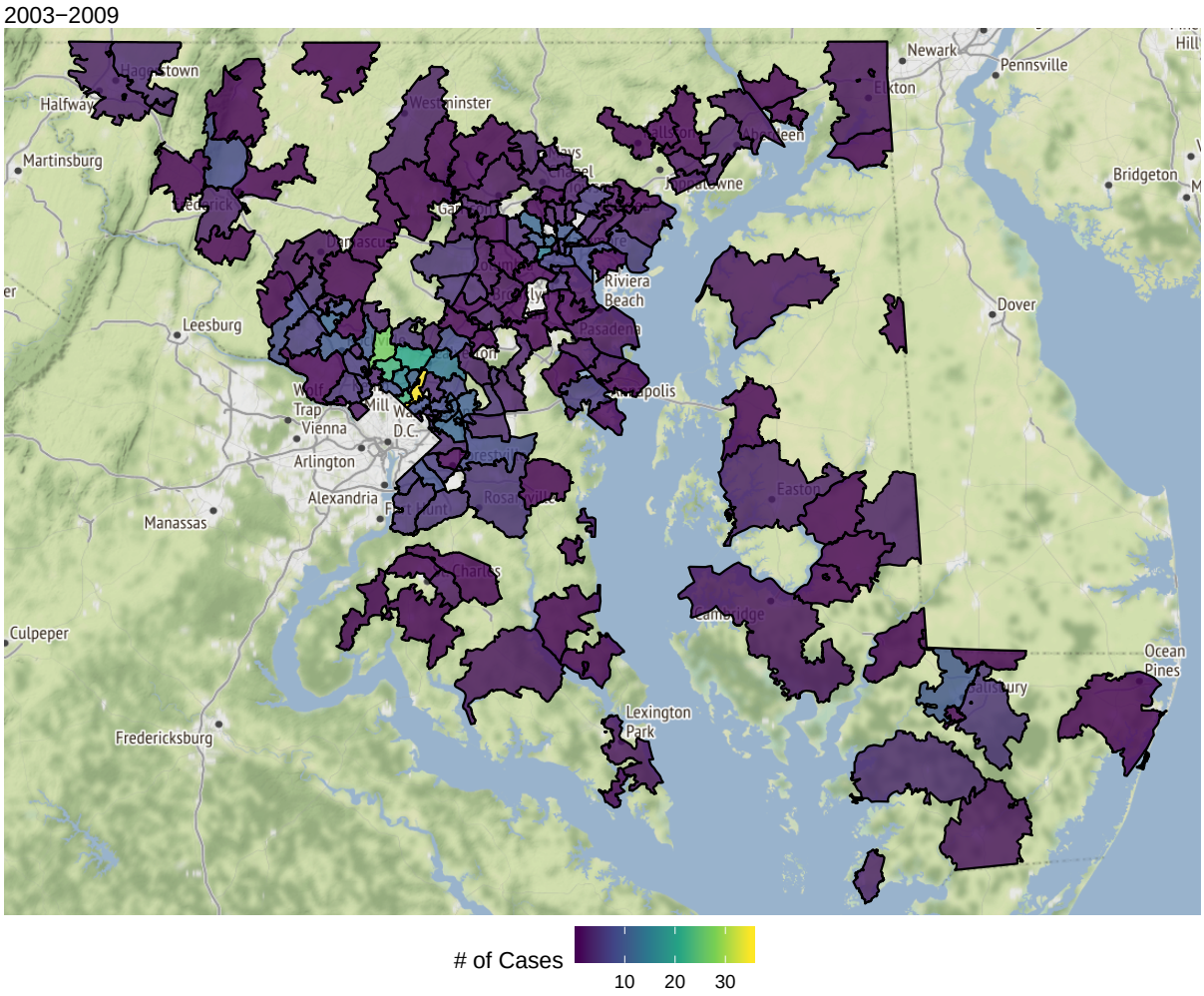

Figure 4: Incidence of TB in Maryland zipcodes from 2003-2009.

There is genotype information for 158 total clusters, 75 of which contain a single person and 41 of which contain exactly two people. We will limit our analysis to these clusters of one or two people. We will also assume that for the clusters of two people, one of the individuals transmitted TB to the other. These clusters are visualized in Fig. 5. The clusters are partitioned by their size and show the detection time of each case where the y-axis is the cluster ID. The smear negatives are colored blue and the smear positives are colored orange. For visual purposes, individuals in the same cluster are connected by a black line. Note that for the clusters of size two, it is not uncommon to have detection gap times of over one year!

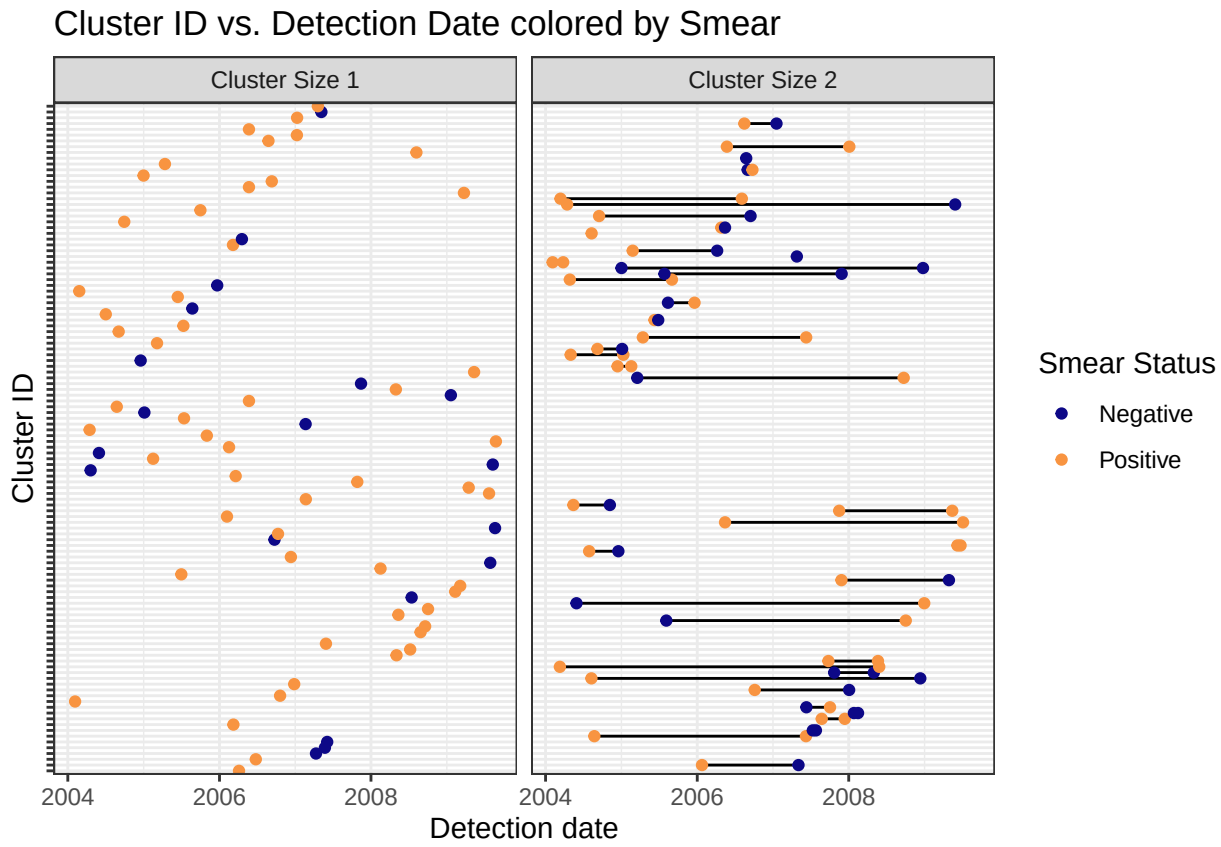

Figure 5: Detection date and smear status of individuals in a cluster. The black lines connects individuals in the same cluster.

We first fit the likelihood model where the gap is not relevant (Eqs. (1) and (2)). The MLE is $\hat{\beta}_0$ = -0.22 (95% CI: [-0.95, 0.49]), $\hat{\beta}_1$ = -0.74 (95% CI: [-1.73, 0.25]). We also fit the second likelihood model where the probability of transmission is weighted by the standardized normal *pdfs* of the gap times (Eqs. (1) and (5)). The MLE for this model is $\hat{\beta}_0 = -0.15$, $\hat{\beta}_1 = -0.88$, and $\nu = \mu/\sigma^2$ = -0.88 but we could not estimate the SEs because the Fisher information matrix was singular.

We could, however, compute the log likelihood for both models and hence test whether gap time is relevant, i.e. $\nu \neq 0$. Our test statistic is $-2 \log L_0/L_A = 2.37$ where L_0 is the maximum likelihood under the null (no gap time $\nu = 0$) and L_A is the maximum likelihood under the alternative ($\nu \neq 0$). The p-value is 0.12 and so we would conclude that gap times are not relevant in this case. This result is not too surprising as the standard error of the gap times is 561 days and the sample mean is 555 days and so the plug-in estimate of ν is 0.002.

It is interesting to note that the log-likelihood for a cluster k where the individuals have the same smear status have the same log-likelihood contribution in both weighting schemes because ($\pi_{k,1} = \pi_{k,2}$ and so $w_{k,1}\pi_{k,1} + w_{k,2}\pi_{k,2} = \pi_{k,1}$). Since the estimates of the βs from both weighting structures are similar to one another, this means that much of the difference in the likelihood can be attributed to the clusters with discordant signs.

Another aspect this result brings in to question is our assumption that one of the individuals had to have transmitted it to the other individuals. We will next consider models that permit an exogenous source to infect both individuals in a cluster of size 2.

3.1 Mean and SE of weighting structures

We explore the difference in estimates of β_1 under different data generation schemes. Specifically, for each data generation scheme, we specify initial parameters and generate 100 different data sets from the initial parameters and generation scheme. For each data set, we estimate β_1 for each of the four different weighting schemes as the MLEs for the given data set. Said another way, for each different data set, we have four estimates of β_1, one from each of our four weighting structures. The four different weighting structures are 1) DAG known; 2) No gap; 3) Standardized pdf gap times; and 4) Normal pdf gap times. The four weighting structures are described above in Equations (2)-(5). The four different data generation schemes are: 1) Uniform$(-M,\ M)$ gap times; 2) Randomly permuted gap times; 3) Weakly informative gap times; and 4) Strongly informative gap times.

The four different data generation schemes specify different distributions for the gap timeδ_k for cluster k (given $n_k = 2$) in the following four ways:

$$\begin{aligned}
&\delta_k \sim \text{Uniform}\,(-10, 10) \quad \text{(Uniform gap)} \\
&\delta_k \sim (2 \cdot \text{Bernoulli}(0.5) - 1) \times N(\mu = 10, \sigma^2 = 10) \quad \text{(Permuted gap)} \\
&\delta_k \sim N(\mu = 10, \sigma^2 = 10) \quad \text{(Weakly informative)} \\
&\delta_k \sim N(\mu = 10, \sigma^2 = 1) \quad \text{(Strongly informative).}
\end{aligned} \tag{7}$$

The data is generated via the algorithm in Section 1 with $p = 0.7$, $\beta_0 = 0$, $\beta_1 = 1$, $K = 1000$, and F_1 as specified in Eq. (7).

The simulation results are displayed in Figure 6. The x-axis shows the weighting structure and the y-axis shows the estimate (point) and 95% CI (bars) of β_1, the effect of smear status on probability of disease transmission. The plots are partitioned by the different data generation schemes. All but one of the estimates (Normal pdf gap times for the uniform gap time generation data scheme) have mean estimates of β_1 as approximately 1. Why is that one seemingly biased?? Possible explanation is that I start the optimizer at the average positive value of the gap time, which can be quite far from zero so I may not be getting the true MLE.

With regards to the standard error (SE) of β_1 for each of the four weighting schemes, note that "DAG known" (far left in each graph) has the smallest SE, which makes sense as this is a "best-case" scenario. The SE for β_1 for the "No gap" model is always visibly larger than that of "DAG known" but not substantially so. This means, that we can recover the effect of smear status on the probability of transmission even if we do not know the true DAG or know anything about the gap times. When the gap times are informative (bottom row), both 'Std. pdf gap times' and 'Normal pdf gap times' do approximately well as the 'DAG known' model in terms of the size of the SE. When the gap time is irrelevant (top row), the SE is visibly larger but not much larger than not using gap time at all.

Since the bar lengths are not easily comparable in the top row of Fig. 6, we include Table 2 which shows the ratio of the sample variance of the different estimates:

$$\left(\frac{SE\left(\hat{\beta}_1^{(i)}\right)}{SE\left(\hat{\beta}_1^{(4)}\right)} \right)^2$$

where $i = 1, 2, 3$ and the superscript (i) refers to the ith weighting structure, respectively. For example, for the strongly informative gap time data generation scheme, the ratio of the sample variance of the 'No gap' model to the 'Std. pdf gap times' model is 1.29. This means that we need approximately 29% more data for the former model to have standard errors of the same magnitude as the latter model. We see that 'Std. pdf gap times' is just as good as 'No gap' model when the gap times are informative but also do not hurt much when the gap time is irrelevant. In comparison to the 'Normal pdf gap times' model, the 'Std. pdf gap times' model needs more data to have the same magnitude SE under the uniform data generation, but we note that the average estimate of β_1 for uniform data generation scheme and the 'Normal pdf gap times' is not as accurate as that of the 'Std. pdf gap times' model.

In summary, having gap time in the model always helps when the gap time is relevant to the probability of transmission and does not hurt much when the gap time is irrelevant. Moreover, when gap time is even weakly informative, we do about as well as having known the actual DAG.

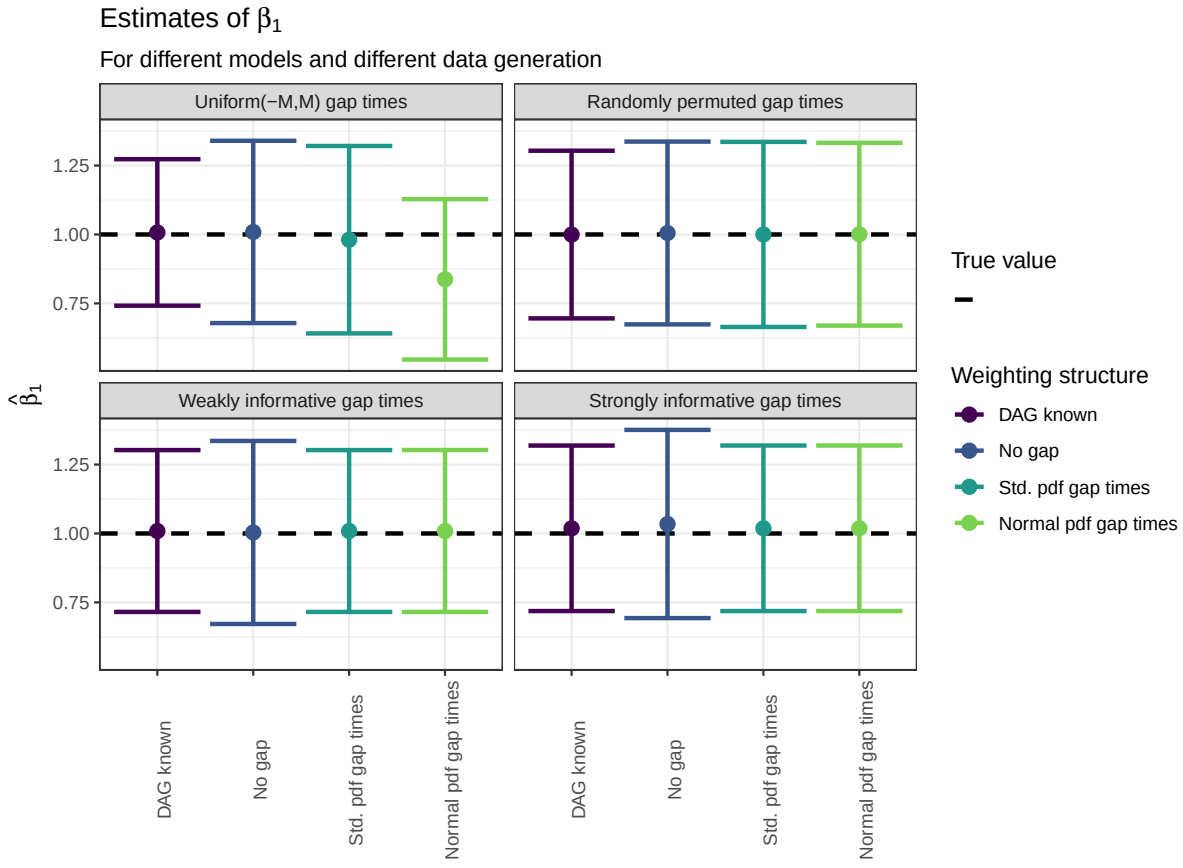

Figure 6: Result of estimates of β_1 from 100 different data sets for each of the four different data generation schemes. The average and 95% CIs are plotted.

Table 2: Ratio of sample variances of $\hat{\beta}_1$ for different weighting structures.

Data Type	(DAG known: Std. pdf)	(No gap: Std. pdf)	(Normal pdf: Std. pdf)
Uniform(-M,M) gap times	0.611	0.946	0.733
Randomly permuted gap times	0.821	0.975	0.975
Weakly informative gap times	1.000	1.280	1.001
Strongly informative gap times	1.000	1.293	1.000

4 Incorporating exogeneous sources of infection

In the previous section, we examined clusters of sizes one and two. When the cluster was of size 2, we assumed that person 1 infected person 2 or *vice versa*, which is a reasonable assumption due to both the genetic data linking the people and the biological plausibility in that TB can have a latent period of years.

In our previous analyses, we assumed that there was primary infection within the cluster whose infection originated from an unobserved, exogenous source. In those analyses, we assumed that the primary infection transmitted TB to the secondary person. However, it is possible that the "secondary" person was also infected from an exogenous source. If the secondary person was also infected exogenously, then either a transmission path can be traced back directly to the primary infection (possibly through unobserved individuals) or no such path can be found. That is, in the first case, the primary person *indirectly* infected person 2 and in the second case, the infections of person 1 and 2 are conditionally independent of one another given some previous infector. Thus given the cluster size $n_k = 2$ the following items show the possible infection paths where O_1 and O_2 represent outside unobserved infections and 1 and 2 represent our observed infections inside the cluster:

1. $O_1 \to 1 \to 2$
2. $O_1 \to 2 \to 1$
3. $O_1 \to 1 \to O_2 \to 2$
4. $O_1 \to 2 \to O_2 \to 1$
5. $O_1 \to 1, O_1 \to 2$

If we care only about whether person 2 was in person 1's direct transmission path, then senarios 3 and 4 can be reduced to scenario 5. If, on the other hand, we only care whether person 1 infected person 2 (or 2 to 1) directly or indirectly, scenarios 3 and 4 would reduce to 1 and 2, respectively.

We let $T_k \in \{0,1\}^3$ represent a transmission event for cluster k where $(1,0,0)$ represents the scenario where both people receive the disease exogenously; $(0,1,0)$ represents the scenario where person 1 directly infects person 2; and $(0,0,1)$ represents the scenario where person 2 directly infects person 1.

Let O_k be a random variable such that $O_k = 1$ if both persons receive TB exogenously and is 0 otherwise. As a consequence, if $O_k = 1$ then $T_k = (1,0,0)$. Let Z_k be a conditional Bernoulli random variable, given $O_k = 1$. If $Z_k = 0$ then person 1 infects person 2 and if $Z_k = 1$ then person 2 infects person 1.

Let the distributions of O_k and Z_k be given by

$$O_k \sim \text{Bernoulli}\left(f(d(k,1,2))\right)$$
$$Z_k \sim \text{Bernoulli}\left(\frac{\pi_{k,2}}{\pi_{k,1}+\pi_{k,2}}\right)$$

Thus T_k is given by

$$T_k = \begin{cases} (1,0,0) & \text{if } O_k = 1 \\ (0,1,0) & \text{if } O_k = 0 \text{ and } Z_k = 0 \\ (0,0,1) & \text{else} \end{cases}$$

where $d(k,i,j)$ measures the distance/similarity between persons i and j in cluster k, and f is a function that takes in a similarity/distance d and outputs a number between 0 and 1, i.e. $f : [0,\infty) \to [0,1]$ and $\pi_{k,i}$ for $i = 1,2$ is given by

$$\pi_{k,i} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_{k,i})}}$$

where $X_{k,i}$ is the smear status of person i in cluster k and β_0 and β_1 are parameters.

As such, T_k can be written as a Multinomial draw of size 1 where the probabilities are obtained from the above equations,

$$\begin{aligned} T_k &\sim \text{Multinomial}\left(1,(p_0,p_1,p_2)\right) \\ p_0 &= f(d(k,1,2)) \\ p_1 &= (1-p_0)\frac{\pi_{k,1}}{\pi_{k,1}+\pi_{k,2}} \\ p_2 &= (1-p_0)\frac{\pi_{k,2}}{\pi_{k,1}+\pi_{k,2}}. \end{aligned}$$

Note that $p_0 + p_1 + p_2 = 1$ and are non-negative so we have specified a valid probability distribution.

The likelihood is thus

$$\mathcal{L}(\beta_0,\beta_1;T_k,n_k,X_k) = \prod_{k:n_k=1}(1-\pi_{k,1}) \prod_{k:n_k=2} p_0^{T_{k,0}} p_1^{T_{k,1}} p_2^{T_{k,2}}.$$

The log likelihood is then

$$\ell(\beta_0,\beta_1) = \sum_{k:n_k=1}\log(1-\pi_{k,1}) + \sum_{k:n_k=2}\left(T_{k,0}\log p_0 + T_{k,1}\log p_1 + T_{k,2}\log p_2\right)$$

We would like to maximize the above quantity. Unfortunately, we do not know T_k. We can use the EM algorithm where the E step finds the expected value of T_k and the M step maximizes the log likelihood given $E[T_k]$.

4.1 Distance/Similarity EDA

We assume that direct transmission of TB can only occur between two people who are close to one another. Of course, this is true in the most literal sense as TB is transferred

through an airborne vector or infected surfaces. We assume that agents who are more similar to one another in demographic characteristics are more likely to directly transmit the disease. Demographic information in our data includes the zipcode, HIV status, homelessness status, age, race, ethnicity, employment status, age and sex. One question to consider is whether we should include detection date or not when calculating the similarity between people.

4.1.1 Physical distance

Perhaps the most natural aspect of similarity to consider is physical distance between people. In the data, we have the variable *zipcode*, which corresponds to the area of residence of the person. In Figure 7, we plot the cross-zipcode transmission of TB where an edge between zipcodes represents the existence of a pair of people within the same infection cluster, one from each zipcode. The edges are colored by the total number of times a unique pair of people within the same cluster exists in zipcode A and zipcode B. The edge thickness is displayed in such a way that closer distances are thicker than further distances. The distance between zipcodes is calculated by taking the Haussdorf distance between zipcodes, which is the minimum distance between the boundaries of two zipcodes. As such, if two boundaries touch, the distance is zero.

We were surprised to see the large number of pairs in the same cluster that were physically distant from one another. This can mean many things, however. First of all, the pairs of people could have been visiting each other or at a common location such as work or school. Alternatively, the people may have transmitted the disease through closer intermediaries (possibly within the cluster or possibly unobserved). It is difficult to see the effect size of near-zipcode transmission network in Fig. 7 due to the large number of cross-zip code interaction crowding the figure.

In Figure 8, we show the same graph as Figure 7 but group the edges based on their distance between zipcodes. From this, we see a number of pairs in a cluster are phsyically close to one another. The top graph in Figure 8 shows the the empirical 0-50% quantiles of the zipcodes, i.e. the close zipcodes. The lower graph shows the 50-100% quantiles, i.e. the far zipcodes. From the top graph, we see many close interactions are within the DC suburbs (bottom left) and Baltimore (center) areas, which also correspond to areas that have higher population density.

Finally, Figure 9 shows the number of pairs within an infection cluster that are also within the same zipcode. This figure shows that many individuals within a infection cluster reside in the same zipcode. In fact, we find that about 25% of the 83 clusters of more than one person are contained in the same county. As such, physical distance between individuals does seem to be important in transmission of TB.

4.1.2 Other similarities

It is possible for individuals to be similar to one another in features besides physical distance. For example, in our data set we also have demographic features including race, age, ethnicity, sex, homelessness, HIV status, unemployment status, and more. We assume there is some degree of homophily (i.e. like likes like) in our population. In terms of disease transmission, we assume this indicates that less similar individuals are less likely to infect one another. We frame this in terms of the negative (i.e. 'less') as opposed to the positive (i.e. 'more') because while it is not necessarily intuitive to assume that individuals are more likely to infect one another due to some similarities, it does

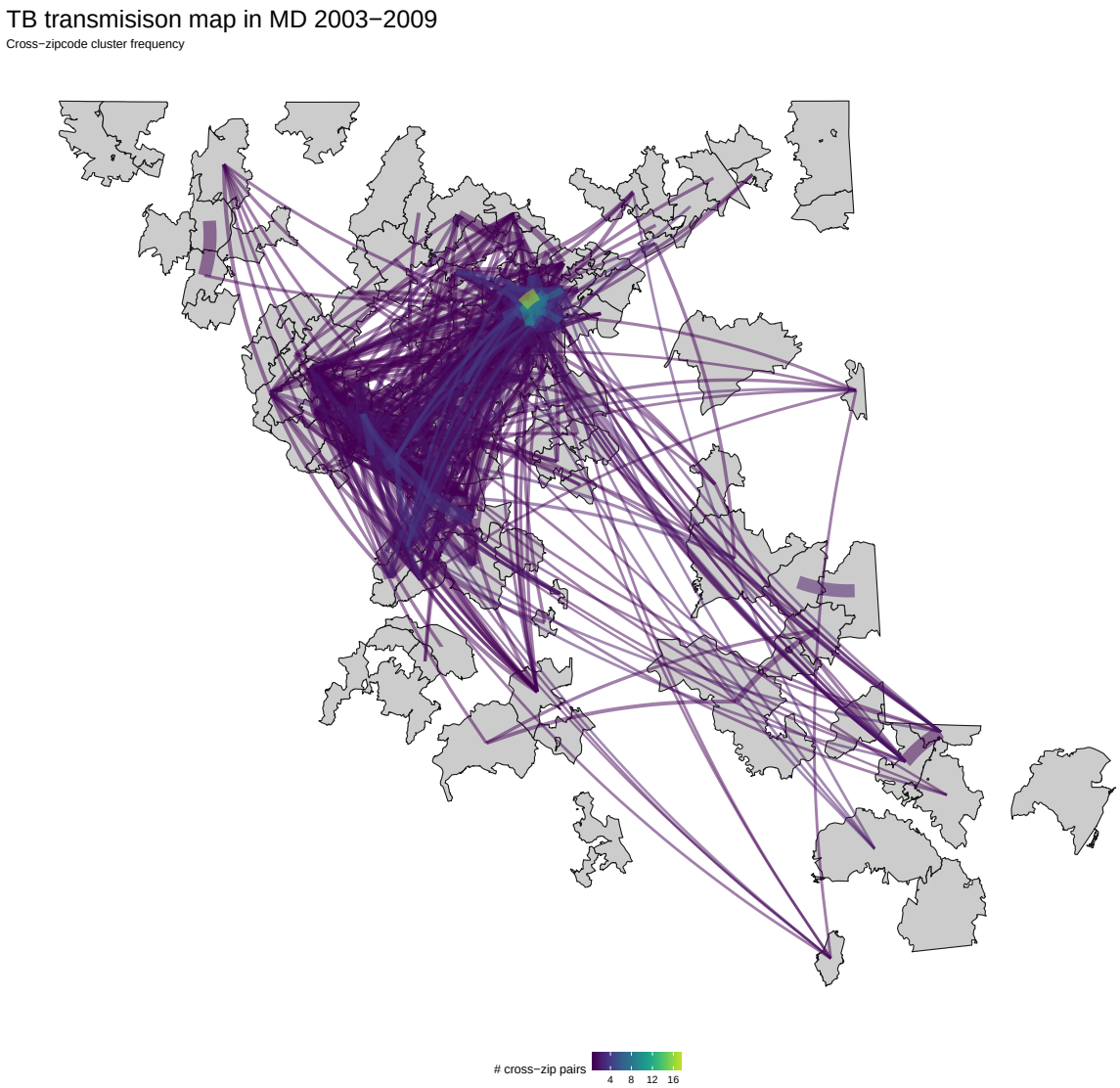

Figure 7: Transmission network of TB. The edge color is the number of unique pairs within an infected cluster for zipcode A and B. The edge thickness is such that closer distances between zipcodes are thicker than those further away.

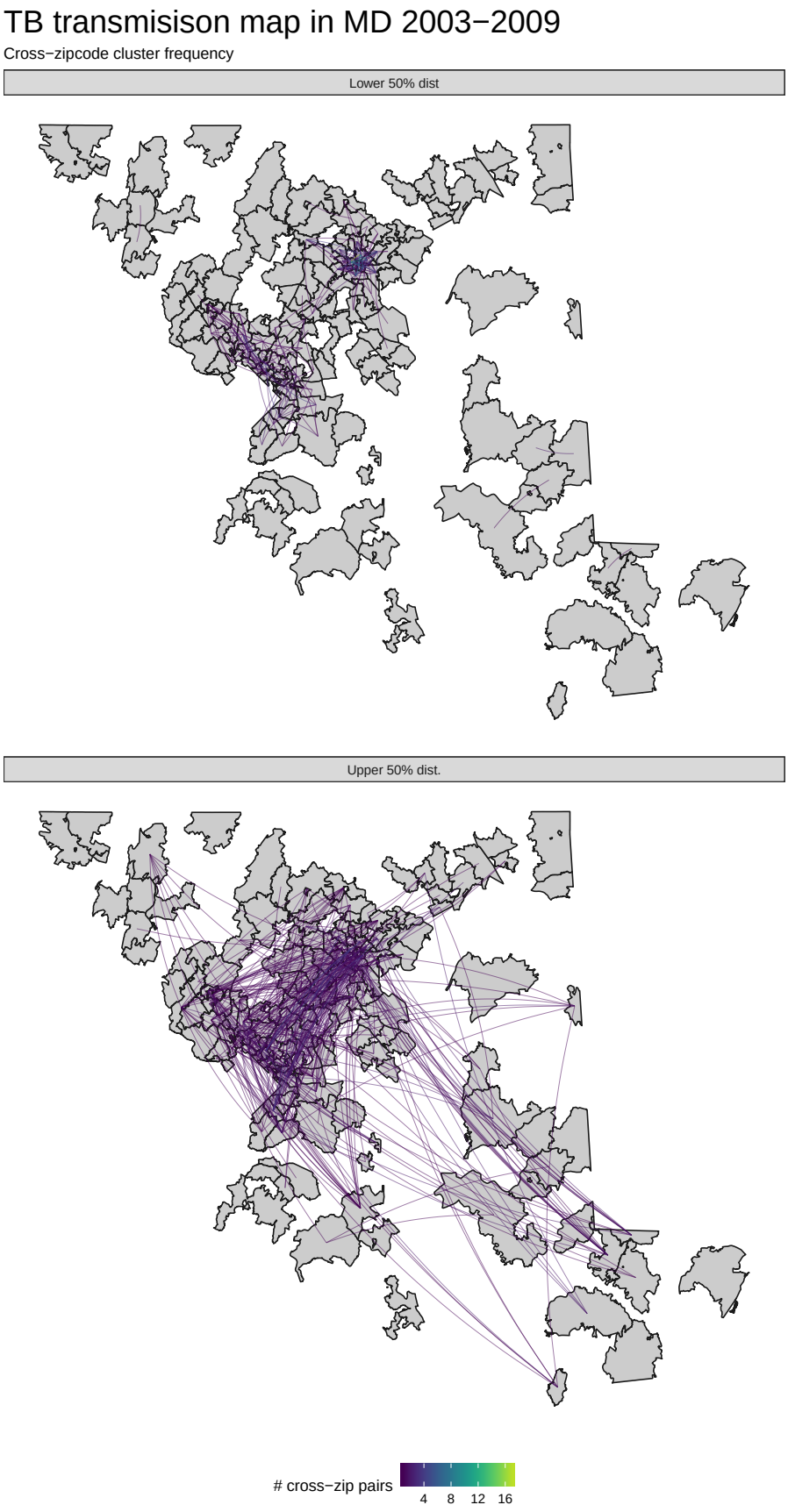

Figure 8: Transmission network of TB. The edge color is the number of unique pairs within an infected cluster for zipcode A and B. The edge thickness is such that closer distances between zipcodes are thicker than those further away.

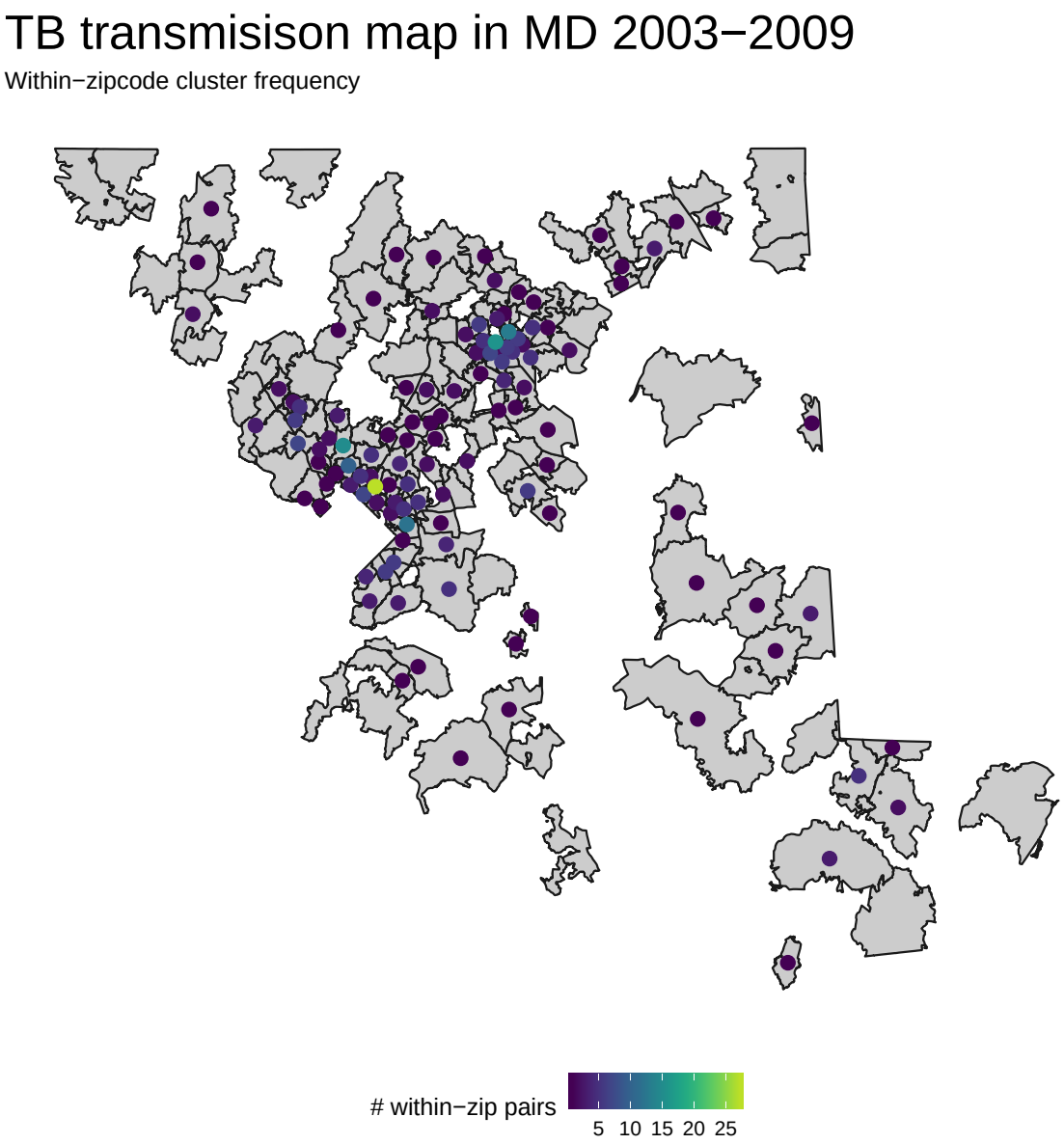

Figure 9: Transmission of TB within a zipcode. The point color is the number of unique pairs within an infected cluster for zipcode A.

seem intuitive that individuals with almost no similarities are less likely to infect one another.

Statistically, we define the similarity between two individuals i and j as

$$s(i,j) = e^{-D(i,j)^2} \tag{8}$$

where $D(i,j)$ is the distance between individuals i and j. We define $D(i,j)$ as the aggregate distance of categorical, continuous, and location-based features of the individuals, respectively d_f, d_g, d_h,

$$D(i,j) = \sum_{f \text{ cts}} w_f d_g(i,j) + \sum_{g \text{ cat}} w_g d_g(i,j) + \sum_{h \text{ loc.}} w_h d_h(i,j), \tag{9}$$

where the subscripts f, g, h refer to features within the sets of continuous, categorical, and location features, respectively and w_f, w_g, w_h are weights for those features.

In this application, we use distances d_f, d_g, d_h that are scaled between 0 and 1 based on the data:

$$d_f(i,j) = \frac{d_f^*(i,j)}{\max_{i,j} d_f^*(i,j)}.$$

Specifically, we take $d_f(i,j)$ to be the normal Euclidean distance between features, $d_g(i,j)$ to be the 0/1 distance and $d_h(i,j)$ to be the Haussdorf distances between the zipcode of i and the zipcode of j. We let all the weights be equal, $w_f = w_g = w_h = 1$ for all f, g, h. The reason we scale the distances between 0 and 1 is so that the different categories are comparable to one another. We note that $\max_{i,j} d_f^*(i,j)$ is a fixed quantity that is pre-determined before any subsetting of the data is done. This way any subset of the data is comparable to one another.

Our function D is a true distance in the sense that it is 1) $D(i,j) \geq 0$ non-negative, 2) $D(i,i) = 0$ identity is zero, 3) $D(i,j) = D(j,i)$ symmetric, and 4) $D(i,k) \leq D(i,j) + D(j,k)$ sub-additive because it is a linear combination of distance metrics. As such the similarity score $s(i,j)$ lies between 0 and 1. The similarity scores between every pair of individuals in clusters of at least size two is plotted in Figure 10. The darker colors represent a lower similarity score (close to zero) and represent individuals who are "far away" from one another. Higher similarity scores (close to one) are represented by lighter colors. Notice how the off-diagonal is very bright, which makes sense as individuals are exactly similar with themselves. The individuals are ordered by their cluster identity so the first n_1 individuals belong to cluster 1, the next n_2 individuals belong to cluster 2 and so on. In the figure, there are lighter blocks of color on the off diagonal, which is a good indication that individuals within the same cluster are more similar to one another than those not in the cluster.

In Figure 11, we plot similarity scores of individuals within clusters. Specifically, these are the block diagonals from Figure 10 reordered so that the most similar individuals within clusters are next to one another. For example, compare Cluster 100 (top left) to Cluster 96 (bottom right). The two individuals within cluster 100 are very different from one another whereas the two within 96 are more similar to one another and almost indistinguishable. We may think that it was less likely that the two individuals in Cluster 100 transmitted the disease to one another than those in cluster 96. In other words, we may suspect that *both* infections in cluster 100 came from exogenous sources.

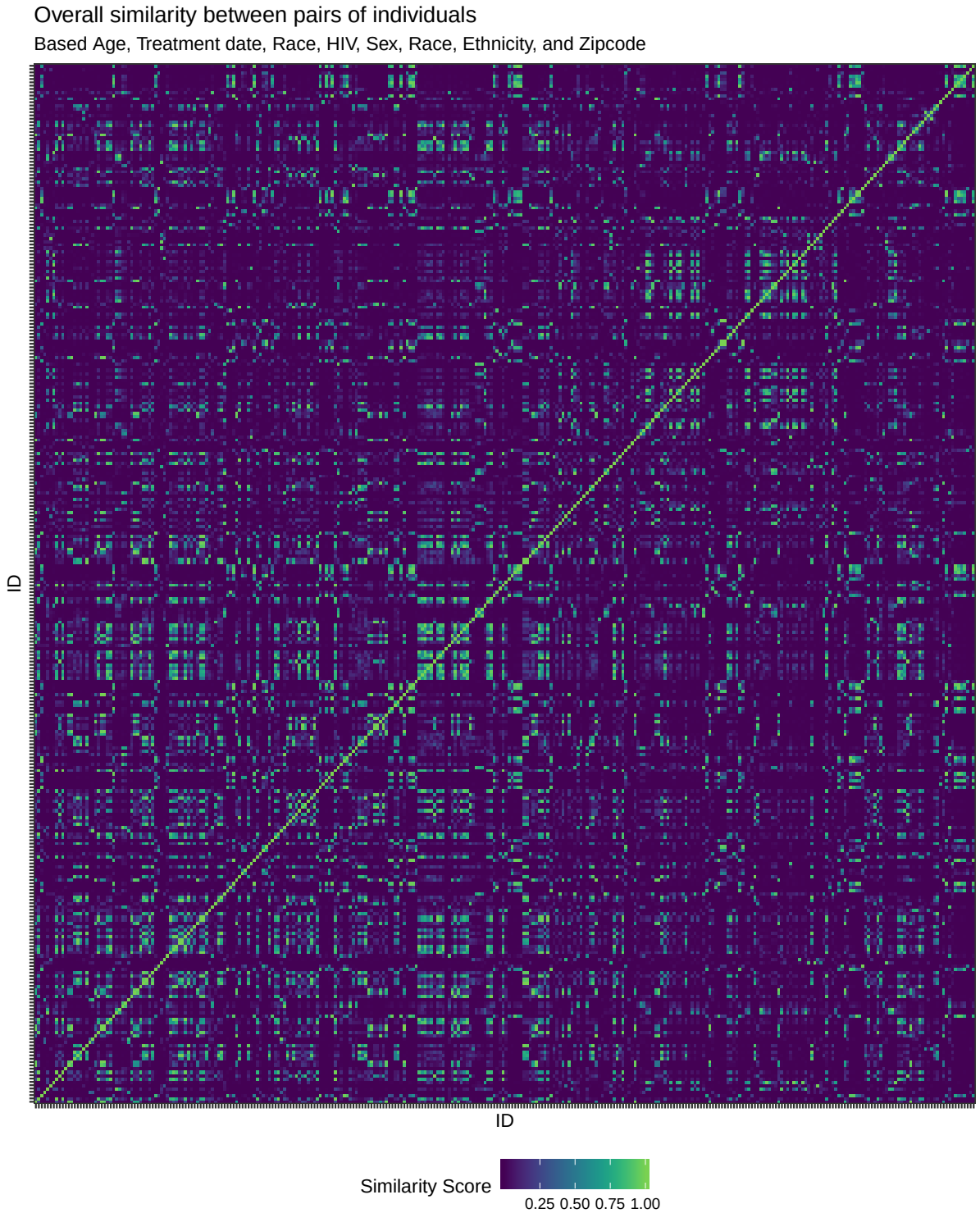

Figure 10: Similarity scores for pairs of individuals based on age of case at treatment, treatment date, race, HIV status, Sex, Race, Ethnicity, and zip code distance. The x and y axis refer to identities of the individuals and are plotted in increasing order by cluster. We see that the off-diagonal then is lighter than the others, which indicates that, in general, similarity scores between individuals within-clusters is higher than similarity scores of individuals within different clusters.

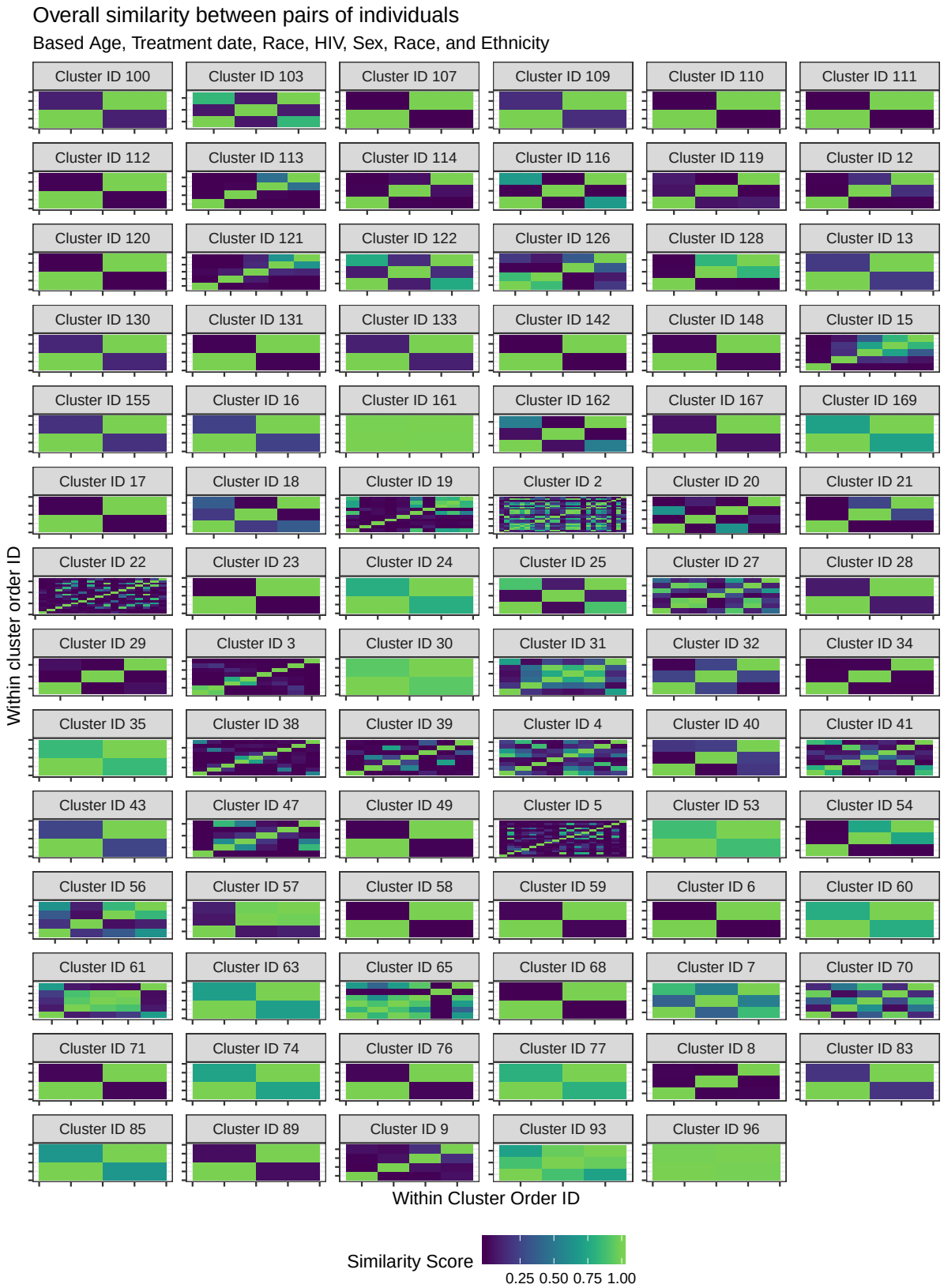

Figure 11: Similarity scores for pairs of individuals within a cluster based on age of case at treatment, treatment date, race, HIV status, Sex, Race, Ethnicity, and zip code distance. The x and y axis refer to identities of the individuals and are ordered by similarity score.

Another interesting example in Figure 10 is that of Cluster 2 (row 6, column 4) and Cluster 22 (row 7, column 1), which are two of the largest clusters in the data set. We see that the individuals in Cluster 2 are highly similar but those in in Cluster 22 are mostly dissimilar.

Finally, we examine the empirical distribution of similarity scores between those within the same cluster and those within different clusters. The results are plotted in Figure 12. There are 972 unique pairs that are in the same cluster compared to 48169 pairs that are in different clusters. The empirical distributions are similar but we do note some separation in larger similarity scores between within-cluster and out-of cluster pairs. We note that there are few "moderate" similarity scores between values of .3 and .7 for either group, which indicates that individuals are typically very similar or not similar at all.

5 Coin flipping simulation for two people

We assume every cluster has at least one infectious individual in it. This primary individual (uniquely defined as the person in the cluster with the smallest infection time) receives her infection from an exogenous or outside (O) source. The second person is then generated (or not) by flipping two coins. The first coin implies an infection from the first source and the second coin from the primary individual. If neither coin has a successful result, then person 2 is not generated. In the case where person 2 is not generated, the cluster is of size 1.

Let $P(O) = \alpha$ be the probability that an outsider infects person 2. Let $f(\beta, X_1)$ be the probability that primary person infects the secondary person where X_1 is some vector of covariates such as smear status, age, or shoe size.

Our observed data is then of the form $D = (C_k, \mathbf{X}_{k,1}, \mathbf{X}_{k,2})_{k=1\ldots K}$ where C_k is the cluster ID of cluster k, $\mathbf{X}_{k,1}$ is the vector of covariates of one of the individuals and $\mathbf{X}_{k,2}$ is the vector of covariates for the other individual. We emphasize that $X_{k,1}$ does not necessarily have to be the primary infector as defined above. If a second person does not exist, we say $\mathbf{X}_{k,2} = \text{NA}$. Let n_k be the size of cluster k (so either 1 or 2). Then the likelihood of observing data D given parameters α_0 and β is

$$\begin{aligned}\mathcal{L}(D; \alpha_0, \beta) = &\prod_{k:n_k=1} (1-\alpha_0)(1-f(\beta, \mathbf{X}_1)) \\ &\times \prod_{k:n_k=2} \sum_{i=1}^{2} [\alpha_0 \times f(\beta, \mathbf{X}_i) + (1-\alpha_0) \times f(\beta, \mathbf{X}_i) + \alpha_0 \times (1-f(\beta, \mathbf{X}_i))]\end{aligned}$$

5.1 Now with one coin

Let α be the probability of person 2 being infected from the outside. Let $p(X_{ki})$ be the probability of i infecting person j in cluter k. We assume that the probability of an outside and inside infection are independent events. Then the probability of no onward transmission within a cluster is $P(n_k = 1) = (1-\alpha)(1-p(X_{k,i}))$.

The probability then that the observed data in cluster k is $C_k = \{x_{ka}, x_{kb}\}$ and $C_k =$

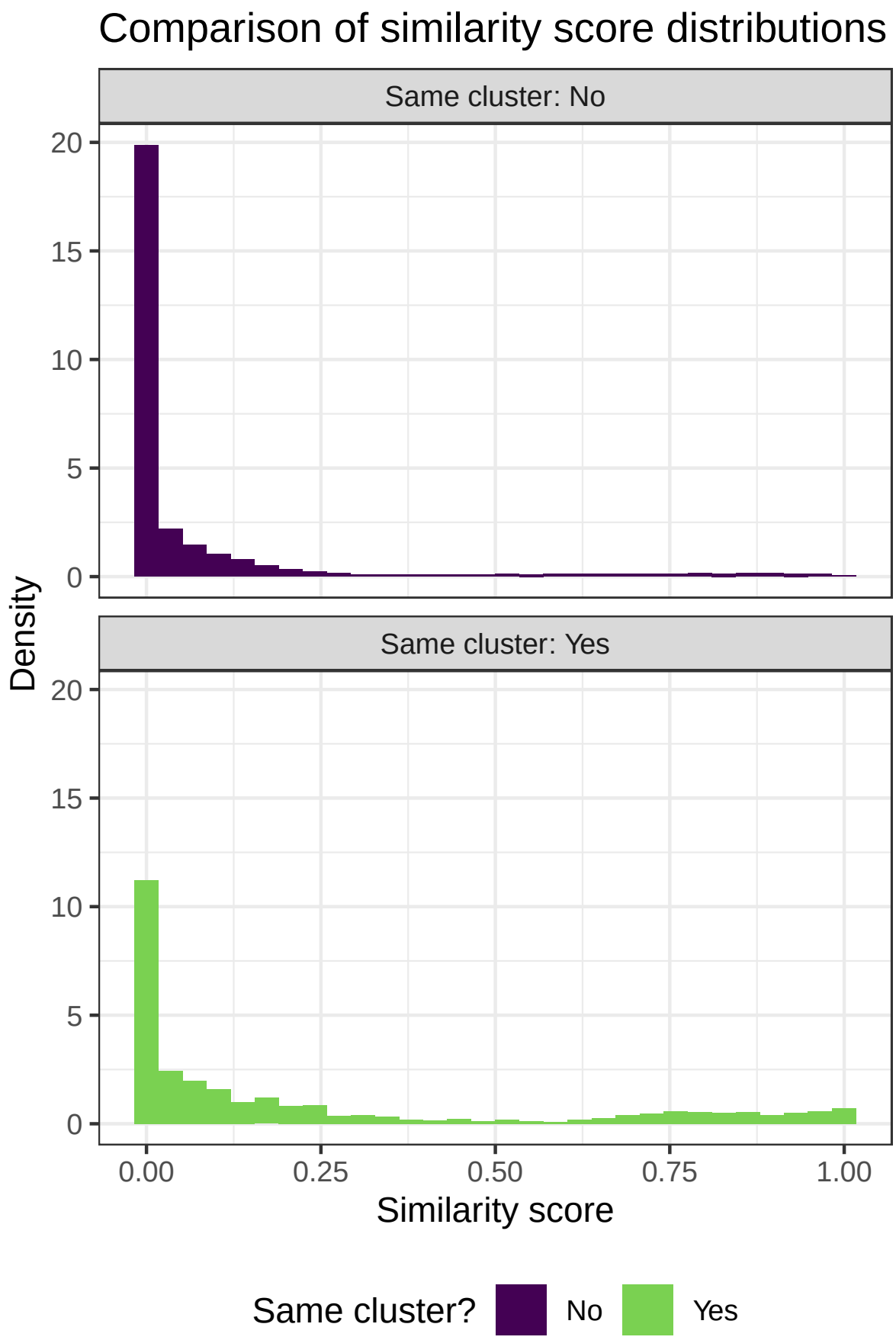

Figure 12: Empirical distributions of similarity scores for pairs without (top) and within (bottom) clusters of TB transmission.

$\{x_{ka}\}$ is

$$P(C_k = \{x_{ka}, x_{kb}\}) = [1 - (1-\alpha)(1-p(X_{ka})]\, P(a=1) + [1 - (1-\alpha)(1-p(X_{kb})]\, P(a=2)$$
$$P(C_k = \{x_{ka}\}) = (1-\alpha)(1-p(X_{ka}))$$

The likelihood of observing clusters $1, \ldots, K$ $C_k = c_k$ is then

$$\begin{aligned}
\mathcal{L}(\alpha, \beta; c_1, \ldots, c_K) &= \prod_{k:n_k=1} P(C_k = \{x_{k1}\}) \times \prod_{k:n_k=2} P(C_k = \{x_{ka}, x_{kb}\}) \\
&= \prod_{k:n_k=1} (1-\alpha)(1-p(X_{k1}))\times \\
&\quad \prod_{k:n_k=2} [1 - (1-\alpha)(1-p(X_{ka})]\, P(a=1) + [1 - (1-\alpha)(1-p(X_{kb})]\, P(a=2)
\end{aligned}$$

If we assume $P(a=1) = P(a=2)$, meaning that either individual has the same probability of being the primary infection the likelihood reduces to

$$\begin{aligned}
\mathcal{L}(\alpha, \beta; c_1, \ldots, c_K) \propto &\prod_{k:n_k=1} (1-\alpha)(1-p(X_{k1}))\times \\
&\prod_{k:n_k=2} 2 - (1-\alpha)(2 - p(X_{ka}) - p(X_{kb})).
\end{aligned}$$

The log likelihood is then

$$\begin{aligned}
\ell(\alpha, \beta; c_1, \ldots, c_K) = &\sum_{k:n_k=1} \log(1-\alpha) + \log(1-p(X_{k1}))+ \\
&\sum_{k:n_k=2} \log\left[2 - (1-\alpha)(2 - p(X_{ka}) - p(X_{kb}))\right]
\end{aligned}$$

Let the data generation process be the following:

1. Simulate a pool of $M > 2K$ individuals with demographic characteristics such as age, sex, smear status, and shoe size.
2. Randomly choose K individuals to be the primary infection in cluster k.
3. Flip a weighted coin $Z_k \sim \text{Bernoulli}\left((1-\alpha)(1-p(X_{k1}))\right)$. If $Z_k = 1$, we randomly choose another individual from the pool to be the secondary infection in cluster k.
4. If there is a secondary infection, we randomly scramble the order of the two individuals.

6 Next steps

For clusters of size two.

- New simulation function. two steps to put in simulate_tb
- simulate agent pool -¿ just generate a bunch of people

- simulate transmission -¿ instead of infection. because more and less confusing
- simulate_tb_clusts2o -¿ the latest and greatest way of simulation

$$\mathcal{L}(\alpha, \beta; c_1, \ldots, c_K) \propto \prod_{k:n_k=1} (1-\alpha)(1-f(X_{k1}))\times \prod_{k:n_k=2} (f(X_{k1})+f(X_{k2})) + \alpha(2-f(X_{k1})-f(X_{k2})).$$

7 Clusters of maximum size 3

Having shown the likelihood for the 2 cluster case was correctly specified, we can move onto the case of where we allow for a maximum of three infections, where all may have been infected from an exogenous source.

We now add an explicit variable infection time denoted t_{ki}. Within the cluster, we assume an ordering of infection $t_{k1} < t_{k2} < t_{k3}$. We impose the constraint that individual i cannot infect individual $i > j$. One can think of t_{ki} as the time of contact with their potential infector.

With this explicit constraint, we adapt our generation process to the following.

We select K primary infections. For each cluster k, we do the following transmission process. The process is sequential. Person 1 first infects individuals. He has a probability of p_1 each of infecting person 2 and person 3. These are independent transmissions.

Person 1

If person 1 infects person 2 and person 3, we are done as we have our maximum number of infections.

If person 1 infects person 2 and not person 3, we move to person 2 movements.

If person 1 infects person 3 and not person 2, we move to person 3 movements.

If person 1 infects neither person 2 nor person 3, we move to Outside (O(2)) movements.

Person 2

If person 2 infects person 3, we are done.

If person 2 does not infect person 3, we move to person O(3) movements.

O(2)

If the outside infects person 2, we move to person 2 movements.

If the outside does not infect person 2, we are done.

O(3)

If the outside infects person 3, we are done.

If the outside does not infect person 3, we are done.

The likelihood is described the following equations. The important part to note is the sequence. We assume a population of 3 people. The f

$$\mathcal{L}(\beta, \alpha) = \prod_{k:n_k=1} \mathcal{L}_1(\beta, \alpha) \times \prod_{k:n_k=2} \mathcal{L}_2(\beta, \alpha) \times \prod_{k:n_k=3} \mathcal{L}_2(\beta, \alpha)$$

$$\mathcal{L}_1(\beta,\alpha) = \prod_{k:n_k=1} (1-p_1)^2(1-\alpha)^2$$

$$\begin{aligned}\mathcal{L}_2(\beta,\alpha) = \prod_{k:n_k=2} \sum_{A\neq B\in\{1,2\}} & [p_A(1-p_A)(1-p_B)(1-\alpha) \\ & + (1-p_A)(1-\alpha)p_A \\ & + (1-p_A)^2\alpha(1-p_B)(1-\alpha) \\ & + (1-p_A)^2(1-\alpha)\alpha]\end{aligned}$$

$$\begin{aligned}\mathcal{L}_3(\beta,\alpha) = \prod_{k:n_k=3} \sum_{A\neq B\in\{1,2,3\}} & [p_A^2 \\ & + p_A(1-p_A)p_B \\ & + p_A(1-p_A)\alpha \\ & + p_A(1-p_A)(1-p_B)\alpha \\ & + (1-p_A)^2(p_B)\alpha \\ & + (1-p_A)^2(1-p_B)\alpha^2]\end{aligned}$$

Enumerating the paths

- Size 1
  1. $1 \not\to 2, 1 \not\to 3, O \not\to 2, O \not\to 3$
- Size 2 (2 permutations)
  1. $1 \to 2, 1 \not\to 3, 2 \not\to 3, O \not\to 3,$
  2. $1 \not\to 2, 1 \to 3, O \not\to 2$
  3. $1 \not\to 2, 1 \not\to 3, O \to 2, 2 \not\to 3, O \to 3$
  4. $1 \not\to 2, 1 \not\to 3, O \not\to 2, O \to 3$
- Size 3 (6 permutations)
  1. $1 \to 2, 1 \to 3$
  2. $1 \to 2, 1 \not\to 3, 2 \to 3$
  3. $1 \not\to 2, 1 \to 3, O \to 2$
  4. $1 \to 2, 1 \not\to 3, 2 \not\to 3, O \to 3$
  5. $1 \not\to 2, 1 \not\to 3, O \to 2, 2 \to 3$
  6. $1 \not\to 2, 1 \not\to 3, O \to 2, 2 \not\to 3, O \to 3$

So there are a total of $1 + 8 + 36 = 45$ paths if we don't know the order of infection (according to time). If somehow we do know the ordering, this reduces to 11.

- Toss til failure model. Outside infects then keep going.
- Enumerate likelihood for up to clusters of 5
- Simulate some clusters max size 5
- Random thinning of graph, alpha a function of betas. Work it out when max 3 pick 2.

8 Enumerating all likelihoods for clusters up to size 5

Let α be the probability of infection from the outside and p_k be the probability of successful infection *from* individual k to another individual. We assume all transmissions are independent of one another.

The below show the unique trees up to isomorphism, meaning to get the full likelihood, we must sum over all permutations of the labels.

8.1 $K = 1$

Path via adjacency matrix

Adjacency Matrix	Likelihood
1. $\begin{pmatrix} 1 \end{pmatrix}$	1. α

8.2 $K = 2$

Adjacency Matrix	Likelihood
1. $\begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$	1. α
2. $\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	2. $\alpha(1-\alpha)p_1$

8.3 $K = 3$

Adjacency Matrix	Likelihood
1. $\begin{pmatrix} 1 & 1 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}$	1. α^3
	2. $\alpha^2(1-\alpha)p_1$
2. $\begin{pmatrix} 1 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix}$	3. $\alpha(1-\alpha)^2 p_1^2(1-p_1)p_2$
	4. $\alpha(1-\alpha)^2 p_1(1-p_1)p_2$

3. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 0 \end{pmatrix}$

4. $\begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$

8.4 $K = 4$

Adjacency Matrix

1. $\begin{pmatrix} 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

2. $\begin{pmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

3. $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

4. $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

5. $\begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

6. $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

7. $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

8. $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 \end{pmatrix}$

Likelihood

1. α^4
2. $\alpha^3(1-\alpha)p_1$
3. $\alpha^2(1-\alpha)^2 p_1(1-p_1)(1-p_2)p_3$
4. $\alpha^2(1-\alpha)^2 p_1^2$
5. $\alpha^2(1-\alpha)^2 p_1(1-p_1)p_2$
6. $\alpha(1-\alpha)^3 p_1^3$
7. $\alpha(1-\alpha)^3 p_1^2(1-p_1)p_2$
8. $\alpha(1-\alpha)^3 p_1(1-p_1)^2 p_2(1-p_2)p_3$

9. $\begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$

8.5 $K = 3$

Adjacency Matrix

1. $\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
2. $\begin{pmatrix} 1 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
3. $\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
4. $\begin{pmatrix} 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
5. $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
6. $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$
7. $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

Likelihood

1. α^5
2. $\alpha^4(1-\alpha)p_1$
3. $\alpha^3(1-\alpha)^2p_1^2$
4. $\alpha^3(1-\alpha)^2p_1(1-p_1)p_2$
5. $\alpha^2(1-\alpha)^3p_1^3$
6. $\alpha^2(1-\alpha)^3p_1^2(1-p_1)(1-p_2)p_3$
7. $\alpha^2(1-\alpha)^3p_1^2(1-p_1)p_2$
8. $\alpha^2(1-\alpha)^3p_1(1-p_1)^2p_2(1-p_2)p_3$
9. $\alpha^2(1-\alpha)^3p_1(1-p_1)^2(1-p_2)^2p_3(1-p_3)p_4$
10. $\alpha(1-\alpha)^4p_1(1-p_1)^3p_2(1-p_2)^2p_3(1-p_3)p_4$
11. $\alpha(1-\alpha)^4p_1^2(1-p_1)^2p_2(1-p_2)^2(1-p_3)p_4$
12. $\alpha(1-\alpha)^4p_1^2(1-p_1)^2p_2^2$
13. $\alpha(1-\alpha)^4p_1^3(1-p_1)p_2$
14. $\alpha(1-\alpha)^4p_1^4$
15. $\alpha(1-\alpha)^4p_1(1-p_1)^3p_2(1-p_2)^2p_3^2$
16. $\alpha(1-\alpha)^4p_1(1-p_1)^3p_2^2(1-p_2)p_3$

8. $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

9. $\begin{pmatrix} 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$

10. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$

11. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$

12. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

13. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

14. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

15. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

16. $\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}$

8.6 My current problem 12/11

The data D can be summarized in terms of each cluster where n_k is the size of the cluster and X_k is the matrix of covariates (order does not matter).

$$D = \{(n_k, X_k)\}_{1:K}$$

Then the likelihood is equal to the product of the likelihood of the independent clusters,

$$\begin{aligned}
\mathcal{L}(\alpha, \beta_0, \beta_1; D) &= \prod_{k=1}^{K} \mathcal{L}_k(\alpha, \beta_0, \beta_1, D_k) \\
\mathcal{L}_k(\alpha, \beta_0, \beta_1, D_k) &= P(\text{Cluster size} = n_k) \\
&= \sum_{\text{ordering}} \sum_{p} P(\text{path } p | \text{ordering}, n_k)
\end{aligned}$$