My project title

"No man is an island, Entire of itself,

Every man is a piece of the continent, A part of the main."

John DONNE, *Meditation XVII*

Author:

John SMITH

Supervisor:

Dr. Mark BROWN

March 11, 2015

Statement of Originality

I certify that this thesis, and the research to which it refers, are the product of my own work, conducted during the current year of the MRes in Biomedical Research at Imperial College London. Any ideas or quotations from the work of other people, published or otherwise, or from my own previous work are fully acknowledged in accordance with the standard referencing practices of the discipline.

The Outbreaker model referenced in this thesis is the work of Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser and Neil Ferguson at the Department of Infectious Disease Epidemiology at Imperial College London. For this thesis I conceived the new spatial likelihood, programmed this likelihood into the existing code, created all simulations and performed all resulting analysis myself.

Abstract

This is where my abstract goes

Acknowledgements

This is where my acknowledgements go

Contents

Abbreviations

This is where my abbreviations go MCMC - Markov chain Monte Carlo

Chapter 1

Introduction

In recent years new technological advances have made the collection of DNA
sequence data fast and cost effective [source - find from other paper claiming
same thing]. At the same time new statistical techniques have risen to
popularity which take advantage of these new gains in computational power,
an introduction to these techniques can be found by O'Neill (2002) and
Streftaris and Gibson (2004). For researchers modelling infectious diseases
such as Teunis et al. (2013) and Wallinga and Teunis (2004) this has meant
a move away from modelling the incidence of a disease in a population to
reconstructing specific outbreak scenarios. The aim of this work has been
to infer who infected whom from data on cases during an outbreak, this
can then tell us other information about the transmission methods of the
pathogen. To reconstruct an outbreak, past authors such as Haydon et al.
(2003) and Jombart et al. (2011) have built transmission trees using the
locations and infection dates of cases, others such as Cottam et al. (2008)
have used phylogenetic approaches. Further work by Ypma et al. (2013) and
Jombart et al. (2014) has combined the use of both of these types of data
in the same model. When reconstructing outbreaks, one area which has not

been investigated is the role in which different groups within a population may effect the transmission dynamics of the outbreak.

In certain outbreak scenarios it may be possible to separate the cases of infection into two or more distinct groups. For instance, Cauchemez et al. (2011) investigated how the transmission dynamics of an influenza outbreak were shaped by school-aged children. An interesting question in this scenario might be whether individuals in the same group are more likely to infect each other and less likely to infect individuals in the other group. Previous work which has sought to estimate transmission rates between different groups within a population has involved using large epidemiological datasets to fit a different compartmental model for each group as in O'Neill (2002) or Cauchemez et al. (2004). Or it has required the creation of bespoke models for specific outbreak scenarios by Cottam et al. (2008), Ypma et al. (2013) and others. The earliest attempt to infer who infected whom began with Haydon et al. (2003) who devised a technique to reconstruct a foot-and-mouth outbreak in the UK. The cases in this model were animals on farms which meant that it was easy to track which animals had come into contact with one another and when each animal had fallen ill. Later work by Cottam et al. (2008) and Ypma et al. (2013) provided new models which took into account genetic data and were designed for use on outbreaks between human hosts. Around the same time Morelli et al. (2012) used a Markov Chain Monte Carlo approach to attempt outbreak reconstruction, this method is heavily computational and for larger outbreaks with many parameters this method has only become feasible as computing power has increased. Crucially these approaches seek to estimate transmission rates between groups in isolation from other epidemiological properties; this approach relegates the study of the group structure of an outbreak to an af-

terthought, to be performed after more pragmatic epidemiological analyses have been undertaken.

In this paper I will develop a group framework for an existing method for outbreak reconstruction by Jombart et al. (2014). In this approach, epidemiological and genetic data collected from outbreaks are used to infer who infected who using a computationally intensive Bayesian model. I will show how this method can be extended to include further data about the group structure of the population and see how well I can estimate parameters representing transmission probabilities between different groups. Additionally, this group data may also serve to improve the quality of the model output in situations where genetic data is not present; in the past genetic data has been shown to play an important role in placing constraints on potential transmission trees, which speeds up the search for likely transmission trees in previous research by Jombart et al. (2014). By introducing this group framework to an effective existing model it is hoped that this will allow the study of the group structure of the population during an outbreak to begin sooner, as well as improving the model's ability to infer who infected whom in certain situations.

I will measure my success by studying analysing the results of two different simulations. In the first simulation I will test how well my extended model infers the group transmission probability parameters. In the second simulation I will consider whether adding the group framework to the model has helped with the model's original task of outbreak reconstruction. In the Methods section I will discuss in more detail the methods used in previous outbreak construction models. I will then explain the techniques used in the Outbreaker method and show how I have added to them.

Chapter 2

Methods

2.1 Past Models

The Outbreaker package by Jombart et al. (2014) uses a Markov chain Monte Carlo process to try and sample from the posterior probability distribution of various parameters and pieces of augmented data. Specifically, it uses DNA sequence data and dates of disease onset or dates of DNA sequence collection obtained from an outbreak along with a generation time distribution and a time-to-collection distribution to infer the immediate ancestor of each case. These likely ancestors can be combined into a transmission tree, and if we are fairly confident in the truth of our assembled transmission tree then we can infer further properties about the outbreak from it. These properties include the rate of mutation of nucleotides in the DNA sequence of the pathogen and the effective reproduction numbers, the average number of secondary cases caused by each case, of individuals through time (which has important properties concerning the potential of an outbreak to become an epidemic, see Grassly and Fraser (2008)).

The Outbreaker model builds upon previous methods by Haydon et al.

(2003), Cottam et al. (2008),Morelli et al. (2012) and Ypma et al. (2013) that use a similar process of assigning a likelihood value to transmission trees (out of potentially millions of tree configurations) and then searching for the tree with the maximum likelihood value or using the likelihood of a tree to sample from the posterior distribution over all possible trees given the data we have collected. The earliest implementation of this approach was Haydon et al. (2003), who proposed a likelihood function for transmission trees which might define the spread of foot-and-mouth disease between farms in the UK. Their likelihood function for each transmission event is a product of two independent terms. The first term gives a likelihood value based upon how well the period during which farm A was infectious overlaps with the predicted time period during which farm B was infected. The better these time periods overlap, the more likely this transmission event was. The second term gives a likelihood value based on how far apart the two farms are, seeing as animals on different farms do not freely mix (especially during an outbreak) then there is only the possibility of an aerial infection, this is more plausible the closer the two farms are. For each tree they considered the likelihood of all of the separate events in the tree and came to an overall likelihood for the tree. Finally, Haydon et al. (2003) proposed an algorithm which would work towards finding the transmission tree with the highest overall likelihood.

As genetic sequencing became faster and more affordable, Cottam et al. (2008) could expand upon this previous model by including a genetic likelihood term. Now that most case data also included a DNA sequence of the pathogen then the DNA sequences could be compared to see if they could help infer the immediate ancestor of each case. Many infectious diseases mutate quickly, therefore mutations in the DNA sequences can occur between

each generation of cases in an outbreak. We can compare the DNA sequences and produce a likelihood that one case is the ancestor of another case. This likelihood depends on how similar the two sampled DNA sequences are, the genetic likelihood of one case being the ancestor of another is higher when their DNA sequences are more similar. Cottam et al. (2008) proceeded by selecting the transmission tree configurations which had the highest genetic likelihoods and then using the previous epidemiological likelihood (based on infection and collection times by Haydon et al. (2003)) to choose 4 final trees which accounted for 95% of the sum total of the likelihood for every possible tree. A new model was formulated by Ypma et al. (2013), who combined the genetic and epidemiological likelihoods into a single term, therefore removing the assumption that the two likelihoods are independent. This is an important assumption to consider because more mutations will occur in a DNA sequence over longer periods of time so we expect that there will be some correlation between the generation times and the number of mutations that we find.

Simultaneously, Morelli et al. (2012) used a Markov Chain Monte Carlo (MCMC) approach to sample from the posterior distribution of transmission trees given the collected data. This technique begins with a tree and then moves the suspected ancestors of each case around according to certain probability rules to form a new tree. The likelihood of both of these trees are calculated and the new tree is accepted as a sample with a probability calculated from the ratio of the two likelihoods. The chain and the movement rules are constructed so that the probability of accepting a tree as a sample is equal to the posterior probability of the tree given the data. We can then look at the trees with the highest posterior probabilities or consider the posterior probability that one case is an ancestor of another.

The Outbreaker model is a combination of these approaches, it uses an MCMC approach with independent genetic and epidemiological likelihoods. It also allows for unsampled cases to occur between two cases and a more complex account of the DNA sequence mutations. Unlike the previous approaches it has also been written as a package for the programming language R (R Core Team (2014)) which means it can be run on personal computers by people with less technical computing skills within a reasonable amount of time. To understand the Outbreaker model we must first look at MCMC methods in general and understand how we can use an MCMC method to sample from a specified distribution.

2.2 Bayesian Statistics

Before covering MCMC methods it is necessary to cover some basic concepts of Bayesian statistics. Bayesian statistics differs from classical frequentist statistics in that when we are trying to infer the value of a parameter for a distribution we define our existing knowledge about the value of the parameter in a prior distribution. In frequentist statistics we assume that the parameter value is an unknown constant that we will try to estimate, the result of our Bayesian inferences is a posterior probability distribution. More formally we have a prior distribution $q(\theta)$ which is (in most cases) a standard probability distribution and a likelihood function $p(D|\theta)$ which tells us the likelihood of our data, D, given θ. The result of Bayesian inference is the posterior distribution $p(\theta|D)$ which tells us the probability of values of θ given our data D. These probability distributions are connected by the fundamental Bayesian formula, taken from Robert (2007):

$$p(\theta|D) = \frac{p(D|\theta)q(\theta)}{p(D)} \qquad (2.1)$$

13

2.3 Markov Chain Monte Carlo Processes

Markov Chain Monte Carlo processes are a combination of two statistical tools, the easiest of the two is Monte Carlo methods. Monte Carlo methods use a large amount of samples from specific probability distributions, these samples may then be used for other purposes such as aproximating parameters of a probability distribution. To work out the expected value of a probability distribution analytically we would integrate over every possible value in the distribution multiplied by the probability of it occurring:

$$E(X) = \int_{-\infty}^{\infty} x \cdot p(x) dx \tag{2.2}$$

Using Monte Carlo methods we would approximate this integration by sampling from the probability distribution thousands of times and taking the mean average of the results:

$$E(X) \approx \frac{1}{N} \sum_{i=1}^{N} X_i \tag{2.3}$$

Put simply, Monte Carlo methods are a way of approximating values of interest given a large amount of samples from a specific probability distribution.

Voss (2014) defines a Markov chain as follows: "a stochastic process $X = (X_j)_{j \in \mathbb{N}_0}$ with values in a set S is a *Markov chain*, if

$$P(X_j \in A_j | X_{j-1} \in A_{j-1}, X_{j-2} \in A_{j-2}, \cdots, X_0 \in A_0) = P(X_j \in A_j | X_{j-1} \in A_{j-1}) \tag{2.4}$$

for all $A_0, A_1, \cdots, A_j \subset S$ and all $j \in \mathbb{N}$".

The output of a Markov chain is a chain of states in the state space S

where the next state in the chain is chosen by probabilities which are only dependent on the current state. If I am at the current state $X_{j-1} \in A_{j-1}$ then my probability of moving to A_j is only dependent on A_{j-1} and not any other previous states such as A_{j-2} and so on.

We can then define a *transition density* as "a map $p : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ such that:

(a) $p(x, y) \geq 0$ for all $x, y \in \mathbb{R}^d$; and

(b) $\int_{\mathbb{R}^d} p(x, y) dy = 1$ for all $x \in \mathbb{R}^d$

If the Markov chain X can be described by a transition density, then we have

$$P(X_j \in A | X_{j-1} = x) = \int_A p(x, y) dy \tag{2.5}$$

from Voss (2014). This gives us a probability distribution which tells us the probability of moving to another state given our current state. If we run our Markov chain for a long time we might be interested in know what the probability of the chain being at a certain state is. Again, using the notation from Voss (2014), we can define a *stationary distribution*: "A probability density $\tau : \mathbb{R}^d \to [0, \infty)$ is a *stationary density* for a Markov chain on the state space $\mathbb{R}^d$ with transition density p, if it satisfies

$$\int_S \tau(x) p(x, y) dx = \tau(y) \tag{2.6}$$

for all $y \in \mathbb{R}^d$." This means that the probabilities of moving between states become fixed and characterised by a probability density τ. Therefore if we run the Markov chain for a long time we can consider the states which it outputs as samples from the probability density τ. If we are able to sample from a stationary distribution which is useful to us, we can combine the samples from a Markov chain with Monte Carlo algorithms together, these

are Markov chain Monte Carlo (MCMC) methods.

We can use the Metropolis-Hastings algorithm to find Markov Chains that have a specified stationary distribution, as described in Gilks et al. (1996). Therefore our stationary distribution could be very complex and we could still use a fairly simple Markov Chain to sample from it. We can then use these stationary distribution samples in our Monte Carlo methods to make approximations about the distribution. Metropolis-Hastings can also be used in a Bayesian setting by specifying the stationary distribution as a posterior distribution of interest. This means that instead of having to find a posterior distribution analytically we can instead use an MCMC process to sample from it and then make inferences about the distribution from our samples. This is what the Outbreaker model does in the specific context of finding the posterior distribution of transmission trees with given outbreak data. The posterior distribution that the Outbreaker model tries to sample from is complex, yet we can use the relatively straightforward Metropolis-Hastings algorithm to construct a Markov Chain with a stationary distribution equal to our posterior distribution.

2.4 The Metropolis-Hastings Algorithm

Voss (2014) describes how we can use the Metropolis-Hastings algorithm to sample from our target density τ as follows:

- Start with a value X_0 that is from the target density, thus $\tau(X) > 0$.

- We then need a transition density $p(x|y)$ where $p(x|\cdot)$ is the probability density of the next possible states of the Markov chain given that the previous state was x. We then sample a value X_1 from the distribution $p(X_0|\cdot)$.

- We then calculate

$$\alpha(X_0, X_1) = min\left(\frac{\tau(X_1)p(X_1|X_0)}{\tau(X_0)p(X_0|X_1)}, 1\right) \tag{2.7}$$

- We then generate a random variable $U_1 \sim U[0,1]$, if $\alpha(X_0, X_1) > U_1$ then we accept X_1 as a sample from τ. If not then we set $X_1 \leftarrow X_0$ and accept this as a sample.

- We repeat this process for thousands of iterations, saving all of the accepted values in a chain. These values are samples from our target density τ.

The need to find the values $\tau(X)$ presents a problem because it requires that we know (or can at least find probability values from) the target distribution which we are trying to sample from and this might not be the case. We can get around this when looking for posterior densities by substituting in the likelihood function which is usually easier to calculate. If our target density is a posterior density of the form $\tau(\theta|D)$ with parameter θ and observed data D then we can write this as

$$\tau(\theta|D) = \frac{\tau(D|\theta) \times \tau(\theta)}{\tau(D)} \propto \tau(D|\theta) \times \tau(\theta) \tag{2.8}$$

Since D represents fixed data, $\tau(D)$ is a constant and therefore cancels out in the equation for $\alpha(x, y)$ so we are left with

$$\alpha(X_0, X_1) = min\left(\frac{\tau(D|X_0)\tau(X_0)p(X_1|X_0)}{\tau(D|X_1)\tau(X_1)p(X_0|X_1)}\right) \tag{2.9}$$

Therefore to use Metropolis Hastings to sample from a posterior distribution we only need to be able to construct the likelihood function and calculate values from the prior densities of our parameters.

2.5 The Outbreaker MCMC Process

Outbreaker uses the Metropolis-Hastings algorithm to sample from the posterior distribution of transmission trees when we have data on an outbreak. There is a transition density that moves around parameters such as the rate of DNA mutation and then accepts or reject the candidate parameter based on the genetic likelihood defined in the Outbreaker model. Additionally Outbreaker uses augmented data which are pieces of data that are moved around as if they were parameters and accepted or rejected. In the context of Outbreaker each case i has an ancestor α_i; a transition density is used to suggest a new candidate ancestor, the likelihood of this potential ancestry is calculated depending on how well the infection time, group, and DNA sequence data fit together between the cases. We can now go on to to discuss the new group structure data and group likelihood.

2.6 Group Data and Parameters

As previously mentioned, certain outbreak scenarios lend themselves to a model whereby the population is separated into distinct groups, people in these groups could have different levels of contact between members of their own group and members of other groups. This could potentially lead to different rates of transmission within and between different groups. One example of this could be groups of patients on different wards of a hospital, if someone on one ward falls ill it seems plausible that they are more likely to transmit this infection to another patient on their own ward rather than a patient on a different ward. If the outbreak spreads through several wards we could use our knowledge of what ward cases are on to assess the probability that one case infected another and help us construct a transmission tree.

Another example could be to separate cases into groups based on their age, Cauchemez et al. (2011) found that influenza transmission in a community was characterised by "back-and-forth waves of transmission between the school, the community, and the household".

We represent these differing rates of transmission between l groups as parameters in an $l \times l$ transmission probability matrix where the element m_{ij} is the probability that a case's infector is in group j given that the case is in group i.

$$M = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1l} \\ m_{21} & m_{22} & \cdots & m_{2l} \\ \vdots & \vdots & \vdots & \vdots \\ m_{l1} & m_{l2} & \cdots & m_{ll} \end{pmatrix}$$

To introduce these parameters into the MCMC process we need to define a likelihood function that will be used to accept or reject candidate probabilities depending on how well they fit the data and a transition density (also known as a "move") that will produce candidate transmission probabilities.

2.7 Group Likelihood

The existing likelihood function of the outbreaker model by Jombart et al. (2014) is composed of the product of the genetic and epidemiological likelihoods. The genetic likelihood between a case i and a proposed ancestor α_i is given by:

$$\Omega_i^1 = p(s_i | \alpha_i, s_{\alpha_i}, \kappa_i, \mu) \tag{2.10}$$

Where s_i is the DNA sequence of case i, s_{α_i} is the DNA sequence of case α_i, κ_i is the number of unsampled cases between i and α_i, and μ is the rate of mutation of the DNA sequences of the pathogen.

The epidemiological likelihood between a case i and a proposed ancestor α_i is given by:

$$\Omega_i^2 = p(t_i|T_i^{inf})p(T_i^{inf}|\alpha_i, T_{\alpha_i}^{inf}, \kappa_i)p(\kappa_i|\pi) \qquad (2.11)$$

Where t_i is the collection date of s_i, T_i^{inf} is the collection date of s_i, $T_{\alpha_i}^{inf}$ is the collection date of s_{α_i}, and π is the proportion of sampled cases from the outbreak.

Jombart et al. (2014) then define their full likelihood as:

$$\Omega_i^1 \times \Omega_i^2 \times p(\alpha_i) \qquad (2.12)$$

To write the group likelihood, Ω_i^3, we can think of the transmission event between case i with group g_i and its immediate ancestor α_i with group g_{α_i} as a Bernoulli trial with $P_{g_i g_{\alpha_i}}$ chance of succeeding. We know the group membership of each case and the current candidate ancestor α_i, therefore the likelihood of a transmission event between i and α_i for the candidate transmission rates matrix, M, is given by:

$$\Omega_i^3 = p(g_i|\alpha_i, \kappa_i, g_{\alpha_i}, M)$$

and the likelihood of a Bernoulli trial is the probability that the event takes place, which can be found by calculating the transmission probability matrix, so for this particular case the likelihood is:

$$\Omega_i^3 = P_{xy}$$

In the Outbreaker model each transmission event is assumed to be independent of other events, therefore the group likelihood for a whole transmission

tree is the product of all of the individual likelihoods for each transmission event (or the sum of the group log likelihoods).

$$\Omega^3 = \prod_i \Omega_i^3 = \prod_i P_{g_i g_{\alpha_i}}$$

This likelihood term is multiplied onto the existing likelihood term to give an overall likelihood for a case: $\Omega_i^1 \times \Omega_i^2 \times \Omega_i^3$. This assumes that the group, epidemiological and genetic likelihoods are all independent. This assumption simplifies the likelihood term but in real outbreak data we would expect to see some correlation between the likelihood terms. For example, if transmission rates really were higher within a group than between other groups we might expect that observed DNA sequences are generally more similar between cases in the same group because mutations that occur between two cases in the same group are more likely to stay within that group, therefore distinguishing the DNA sequences from these cases from those belonging to other groups. Having defined our group likelihood term we must now decide upon the way in which the Metropolis-Hastings algorithm will move the parameters in the transmission rate matrix to produce new candidate rates.

2.8 Transmission Rate Matrix Move

The proposal distribution for moving elements of the transmission probabilities matrix is not straightforward because the values in the matrix must satisfy $0 \geq m_{ij} \leq 1$ and the values of each row sum to 1. We implemented a move for the transmission probabilities matrix in the MCMC algorithm which proposes new probabilities for each row of the matrix at a time, l candidate probabilities are sampled from the Dirichlet distribution

with concentration parameters equal to the old probability values in the chain multiplied by a constant value. Because the Dirichlet distribution is not symmetrical we need to introduce a correction factor. The probability distribution function for the Dirichlet distribution for K probabilities with concentration parameters $(\alpha_1, \cdots, \alpha_K)$ is given by:

$$Dir(\alpha, x) = \frac{1}{B(\alpha)} \prod_{i=1}^{K} x_i^{\alpha_i - 1} \qquad (2.13)$$

The prior distribution for each row is a symmetric Dirichlet distribution where all concentration parameters are equal and multiplied by a constant provided by the user. The prior multiplication constant reflects how confident the user is that the probabilities in the matrix are equal, larger values reflect a belief that the transmission probabilities are not equal between groups. This gives the following process for updating an element in the transmission rates matrix, it is the standard Metropolis-Hastings algorithm except we have taken the logarithm of values to preserve accuracy during computation:

- For row i we take the current probabilities, $\mathbf{p}_n = \{p_{i1}, p_{i2}, \cdots, p_{il}\}$ and sample candidate probabilities: $\mathbf{p}_{n+1} \sim Dir(x; m\mathbf{p})$ where m is the multiplying constant.

- Calculate the log ratio:

$$\log(\Omega_3(\mathbf{p}_{n+1})) - \log(\Omega_3(\mathbf{p}_n))$$
$$+ \log(Dir(\mathbf{p}_{n+1}; \mathbf{p}_n)) - \log(Dir(\mathbf{p}_n; \mathbf{p}_{n+1})) \qquad (2.14)$$
$$+ \log(Dir(\alpha; \mathbf{p}_{n+1})) - \log(Dir(\alpha; \mathbf{p}_n))$$

Where $\log(\Omega_3(\mathbf{p}_{n+1})) - \log(\Omega_3(\mathbf{p}_n))$ is the ratio of the group likelihoods

of the old and new parameters and

$$\log(Dir(\mathbf{p}_{n+1}; \mathbf{p}_n)) - \log(Dir(\mathbf{p}_n; \mathbf{p}_{n+1}))$$

is the correction factor for the proposal distribution, and

$$\log(Dir(\alpha; \mathbf{p}_{n+1})) - \log(Dir(\alpha; \mathbf{p}_n))$$

are the values of the prior distributions with the user-specified multiplication constant c and concentration parameter α with α_i all equal.

- If the log ratio is greater than 0, we accept $\mathbf{p}_{n+1}$ as a sample from the posterior distribution

- If the log ratio is less than 0 then we generate a random uniform number, U, and if $\log(U)$ is less than or equal to the log ratio then we accept $\mathbf{p}_{n+1}$ as a sample from the posterior distribution. If the log ratio is less than 0 and $\log(U)$ then we reject $\mathbf{p}_{n+1}$ and draw a new candidate $\mathbf{p}_{n+1}$.

I implemented the move within the existing model using the programming languages C and R, I also implemented a tuning feature which increases or decreases the multiplication constant of the Dirichlet proposal distribution to keep the acceptance probability of the move between 25% and 50%. The result of this process is a number of samples of the group transmission rate parameters from the posterior distribution, we can now go on to discuss how we can analyse this output and how we can produce data to test the extended model.

2.9 Simulating Outbreaks With Group Structure

To test the new group framework in Outbreaker, we need to be able to fit the model to data which was generated using a population that has groups which have varying tranmission rates between them. The Outbreaker package has its own outbreak simulation procedure, `simOutbreak`, which we can extend to generate outbreak data that has the desired group structure. We can modify the formula which chooses the infector of an infected individual (in the program code this is implemented by first choosing the ancestors and then choosing who has been infected by them), currently the individual is chosen by sampling from a multinomial distribution with probabilities:

$$\frac{w(t - t_i)}{\sum_i w(t - t_i)}$$

where w is the generation time distribution. We can incorporate the probability of transmission within and between members of different groups into these probabilities by supplying a transmission probability matrix. If we continue with the notation g_i as the group of case i and g_{α_i} as the group of the immediate ancestor of case i then we can introduce the group transmission probabilities into the multinomial distribution as follows:

$$\frac{P_{g_i g_{\alpha_i}} \cdot w(t - t_i)}{\sum_i P_{g_i g_{\alpha_i}} \cdot w(t - t_i)}$$

At the start of the outbreak we would have n individuals, n_1 of whom are in group 1, n_2 of whom are in group 2 and so on up to group l. We can then specify the exact numbers in each group ,perhaps indirectly through proportions of the population in each group, when the simulation begins. We would also specify the $l \times l$ matrix of transmission probabilities which would be

the true, unknown parameters in our model tests. Then the procedure takes place just as before but now new cases have different probabilities of being infected by the existing cases depending on the within and between group transmission probabilities. We can then modify the output of `simOutbreak` to colour the nodes of the transmission tree depending on group so it is easy to see how the outbreak has moved around the group structure. Imported cases are assigned to the existing groups with a probability proportional to the relative sizes of the groups, here we are assuming a scenario where imported cases inherit the group transmission probabilities of a group once they join it.

Having implemented this method in `simOutbreak`, the user can now pass a matrix to the `simOutbreak` function giving the true values of the transmission probabilities within and between groups, the user must also specify the number of individuals within each group. The number of individuals in each of the groups must sum to the overall number of individuals in the simulation, individuals are then assigned to groups randomly. Imported cases are assigned a group based on the relative sizes of the groups defined by the user. Users can now also colour nodes on the plotted transmission trees by their group membership, this allows us to see how the transmission tree is affected by different group sizes and transmission probabilities. For example, the tree in figure 2.1 was created and coloured using three groups with the transmission probability matrix

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

We can now go on to generate some data to see how the method performs
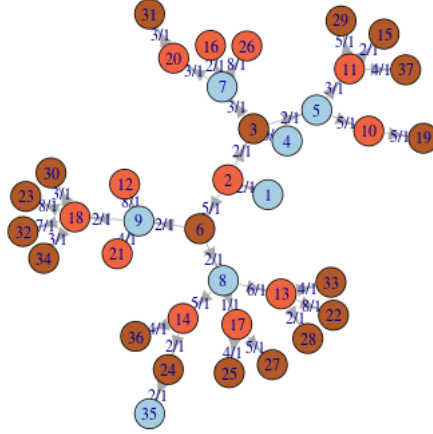
Transmission tree from simulated outbreak

Figure 2.1: A transmission tree constructed from a randomly generated outbreak using the group transmission probability matrix defined above. The direction of arrows between nodes determines the direction of infection. The nodes are coloured by the group of the individual

on generated data and data from a real outbreak with group structure.

2.9.1 Simulation 1

For the first set of simulations we required datasets where outbreaker could already infer the majority of correct ancestries, we could then give outbreaker the group data and an uninformative prior and see how well the posterior distributions of the transmission probability parameters capture the group dynamics of the data.

We used an imagined scenario of collecting outbreak data from districts within a city in West Africa during the 2014 Ebola Virus Disease (EVD) outbreak. The study of the outbreak led the WHO Ebola Response Team (2014) to hypothesise that the geographical spread of EVD was in part due

to a large amount of population movement between cities in bordering countries. We aimed to recreate this migration led transmission on a smaller scale between districts in a city. We simulated datasets where infected individuals are divided into groups based upon their district of residence and the probabilities of transmission between groups is dependent on the amount people travelling between the two districts. If many people commute between districts A and B regularly then it is more likely that an individual unknowingly infected with EVD will travel from district A to district B (or vice versa) and transmit EVD to the people who live there during their trip. To make the scenario as realistic as possible we used the estimated epidemiological properties of EVD estimated during the recent analysis by WHO Ebola Response Team (2014) and others, a full description of the parameter values used can be found in Table 2.1.

Table of pathogen parameters for simulation 1

Table 2.1: The table shows the values of the epidemiological properties of the pathogen used in the `simOutbreak` function, the right hand column shows the source from which the parameter value was taken.

Parameter	Value	Source
Serial Interval	Gamma distribution with mean = 13.5, s.d. = 9.2	WHO Ebola Response Team (2014)
$R0$	2.1	WHO Ebola Response Team (2014)
Mutation Rates	Substitution rate per site per day = 5.479452e-06	Gire et al. (2014)
Sequence Length	19000 bases	Volchkov et al. (1999)

I also had to provide `simOutbreak` with further simulation parameters that would characterise the outbreak which I would then analyse with outbreaker, a full description of these parameters can be found in Table 2.2.

Table of `simOutbreak` parameters used in simulation 1

Table 2.2: The table shows the values of the parameters passed to the `simOutbreak` function to simulate outbreak datasets

Parameter	Value
Number of groups	4
Group sizes (number of hosts)	75,75,25,25 (=200)
Group transition matrix	$\begin{pmatrix} 0.65 & 0.1 & 0.15 & 0.1 \\ 0.1 & 0.6 & 0.1 & 0.2 \\ 0.05 & 0.15 & 0.4 & 0.4 \\ 0.15 & 0.05 & 0.4 & 0.4 \end{pmatrix}$
Duration	50
Spatial	FALSE
Number of Iterations and Burn-in	1e5,2e4

Finally, outbreaker was run 4 times on each dataset. One run included both group and DNA sequence data, one run included only group data, one run included only DNA sequence data, and one run included neither types of data (leaving just onset times). All runs were computed for 100000 iterations with a 20000 iteration burn-in period. We simulated 440 datasets using these parameters and collected the results from each run of outbreaker on each dataset using the high-performance cluster. I discarded 14 datasets where there was not at least one case from each group in the outbreak leaving 426 datasets and their corresponding results. I then analysed the proportion of correct ancestries inferred on each run for each dataset to check that the transmission tree was adequately inferred using DNA sequence data alone and that the outbreaker model including the group framework was behaving correctly.

Once I was satisfied that the extended outbreaker model was running properly I could then begin to scrutinise the posterior density samples for the transmission probability parameters. Each simulation was generated using the same transmission probability matrix but the number of cases varied so the simulations as a whole should give a good idea of how well the param-

28

eters are inferred in a variety of situations. Our first step in evaluating the posterior samples for each run was to see in the true parameter value (used to generate the dataset) fell inside the 95% equal-tails interval (a credible interval with 2.5% of the posterior density in each tail), I then counted how many times this occurred for each parameter over all of the simulations to give myself a rough idea of how often outbreaker infers a posterior distribution with a reasonable probability of giving the true parameter value.

2.9.2 Simulation 2

I also hoped to show that the group likelihood could help to infer correct ancestries in situations where the genetic and epidemiological likelihoods were not so effective. These situations are characterised by an outbreak where there are low rates mutation in the pathogen DNA sequences and a fairly long pathogen generation time. In these situations the original outbreaker model without group data struggles to infer the correct ancestor for two reasons. The genetic likelihood cannot narrow down the ancestor because there are few mutations between cases so most previous cases will have very similar genetic likelihoods. Secondly, the epidemiological likelihood struggles because the fairly long generation time means that for a newly infected case outbreaker has to look quite far back into the past for potential ancestors, this will bring up many candidate ancestors and outbreaker have no other way to determine who the correct ancestor is likely to be. In Figure 2.2, if there are not many mutations between cases, outbreaker would have trouble inferring an ancestor out of cases 1 to 5.

The group likelihood can help in this situation if the cases are divided into groups and we have are confident of what the group transmission probabilities in the situation are. If we are sure that most transmission takes

Outbreaker inference without groups

Figure 2.2: The figure shows a transmission tree, each node is an infected individual and each full line shows a previously inferred transmission event. The node with a "?" is a newly infected case, the dotted lines represent potential infectors of "?". The figure aims to show that in this scenario outbreaker would have difficulty inferring the correct infector of "?"
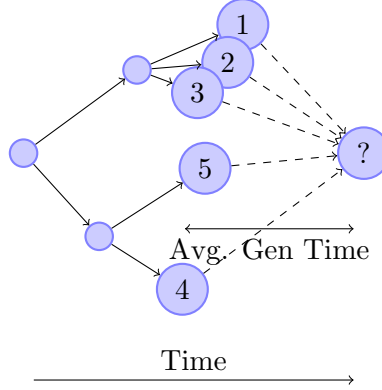

place within groups we could provide a prior that heavily promotes unequal transmission probabilities. If the data has a true group structure where transmission happens overwhelmingly within groups this will be inferred quickly by the model because it is encouraged by the prior. Therefore when we go to assess the group likelihood of a particular ancestry, it will give a much higher likelihood value to two cases within the same group.

Returning to the scenario in Figure 2.3, if the newly infected case belongs to group A and there are 5 candidate ancestors, one of whom belongs to group A, then the likelihood of the connection between the two cases from group A will be much higher and therefore outbreaker will infer this ancestry. Therefore if our prior knowledge that the transmission rates are very unequal is true then we will have biased outbreaker towards the correct ancestries based on their group membership. This is how the group structure of the data and a strong prior can help outbreaker infer correct ancestries in certain situations where the other data is not as useful. In the

transmission tree below the nodes are coloured by group membership. If we are trying to guess an ancestor for the new case and we suspect that most transmission occurs within groups then we would guess node 5. If our prior knowledge is accurate then we are making a sensible guess because it would be most likely to have been node 5 that infected our new case. Adding in a group structure and prior knowledge has helped us infer the correct ancestor. To test this we created datasets that had small groups with 3-4 people in

Outbreaker inference with groups

Figure 2.3: The figure shows a transmission tree, each node is an infected individual and each full line shows a previous inferred transmission event. Nodes are coloured by their group membership. The node with a "?" is a newly infected case, the dotted lines represent potential infectors of "?". The figure aims to show that with group data, outbreaker would make the sensible inference that case 5 is the infector of "?".
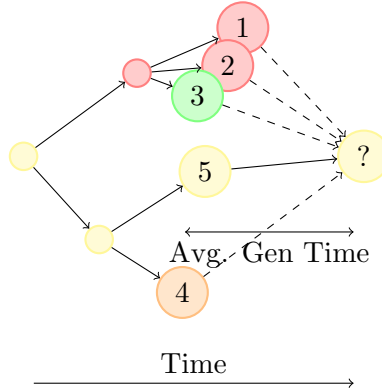
(households) and one larger group (the community), the true transmission probabilities were such that cases with a household had very high probabilities of infecting individuals in the same household, a much lower probability of infecting individuals in the community and an extremely low probability of infecting individuals in other households. Cases in the community had a reasonable probability of infecting others in the community and the rest of the probability was assigned equally to infected a member from each household. Table 2.3 gives the exact parameters given to the `simOutbreak`

function to produce each dataset.

Table of parameters for simulation 2

Table 2.3: The table gives the values of parameters used to create simulated datasets for simulation 2

Parameter	Value
Generation time	Gamma($\alpha = 1.72, \beta = 0.16$)
Community group size	15
Number of households	Pois($\lambda = 12$)
Household size	Pois($\lambda = 4$) and always ≥ 2
$R0$	1.6
Within house transmission probability	0.9999
Within community transmission probability	0.1
DNA sequence mutations	Left to `simOutbreak` default

For this simulation I generated several datasets and analysed them using the extended outbreaker model with various values for the Dirichlet prior multiplication constant. It is not obvious or easy to find a good value for the constant that works well across many datasets so I did not attempt to run many instances of this simulation on the high performance clusters. Further work may be able to determine the properties of a dataset which could suggest a sensible prior value.

Chapter 3

Results

3.1 Testing Procedures

I tested the extended method on simulated data with a group structure generated by the new `simOutbreak` function. As explained in the Introduction, I ran two different simulations which tested different aspects of the new group framework. The first simulation aimed to see how well outbreaker estimated the transmission probabilities parameters. To do this I ran the extended outbreaker model on datasets where the original outbreaker model already does a good job of inferring correct ancestries. From this I can see if outbreaker does a good job at inferring transmission probability parameters when it infers correct ancestries. The second simulation aimed to see whether the group framework can improve the inference of correct ancestries on datasets where the original outbreaker model does not perform very well. If I give outbreaker strong prior parameters with regards to the group structure of the data, does this help outbreaker infer correct ancestries?
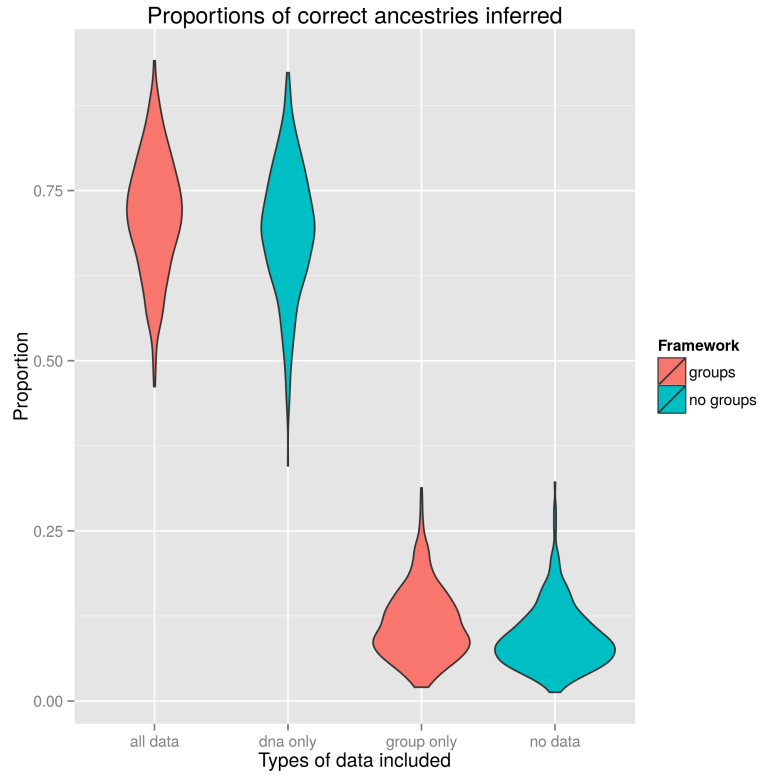
Impact of group framework on proportion of correct ancestries
inferred

Figure 3.1: This violin plot shows the distribution of the proportion of correctly inferred ancestries in the consensus ancestries of outbreaker runs with difference parameters.

3.2 Simulation 1: Estimating Transmission Probabilities

Below are the results of my analysis on the data collected from the first simulation based on the method described in section 2.9.1. In line with previous results from Jombart et al. (2014), outbreaker infers a much higher proportion of correct ancestries when DNA sequence data is included. Comparing runs with and without group data also shows that the outbreaker model including the group framework performs as well as, if not very slightly bet-

Log-posterior values of chain
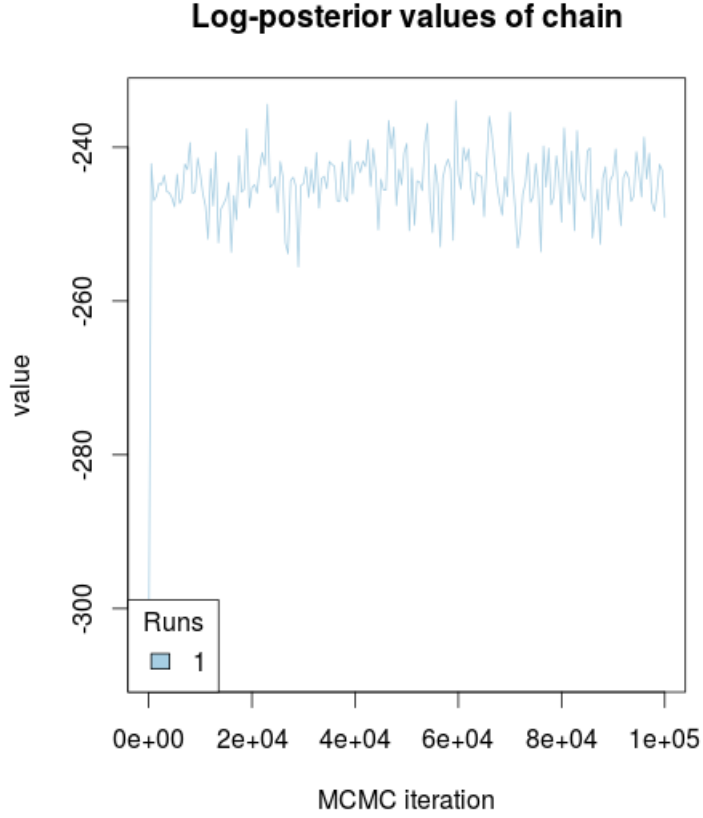
**Mixing quality of an extended outbreaker run**

Figure 3.2: This plot shows the values of the log-posterior density at every 500 steps of the chain for an example run of the extended outbreaker model on a simulated dataset

ter, than the original outbreaker model. We also checked the convergence of the model likelihood including the group likelihood for several datasets to ascertain that the model with the group data was mixing well and converged upon a posterior distribution suitably, an example of one of these MCMC traces is shown below. The outbreaker runs including group data had uninformative priors that suggested that all group transmission probabilities are equal. The matrix below shows the corresponding proportion of times that
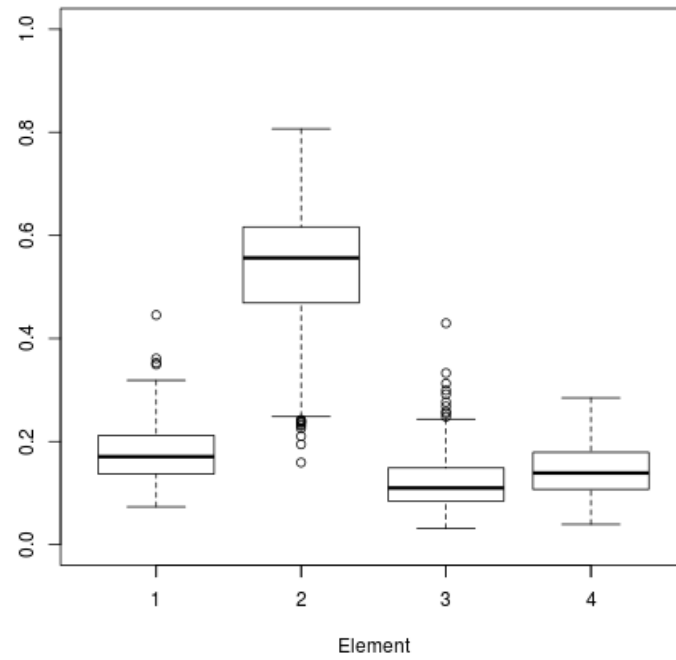
the true parameter was within the 95% equal-tails interval of each run.

$$\begin{pmatrix} 0.8 & 0.87 & 0.97 & 0.99 \\ 0.86 & 0.89 & 0.99 & 0.89 \\ 0.54 & 0.88 & 0.88 & 0.91 \\ 0.89 & 0.48 & 0.91 & 0.87 \end{pmatrix}$$
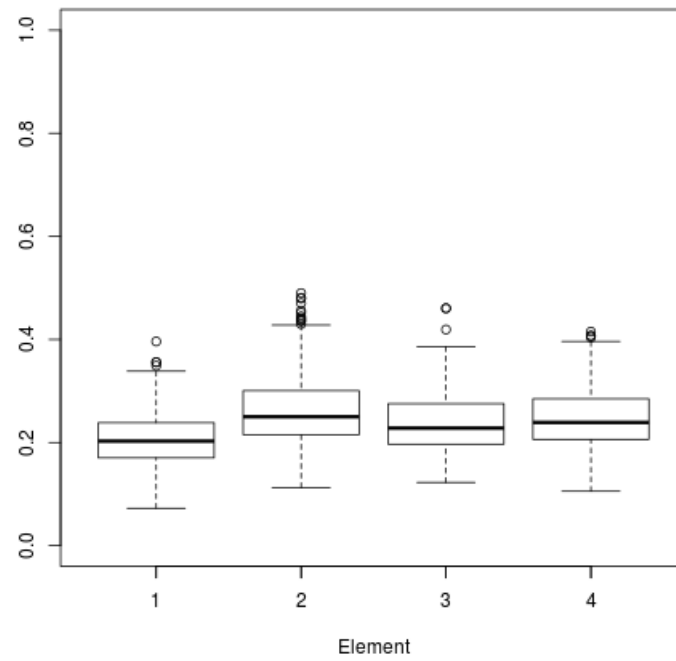
We also calculated the posterior sample median for each transmission prob-ability parameter and produced a boxplot to describe where the sample medians fell over all of the simulations:
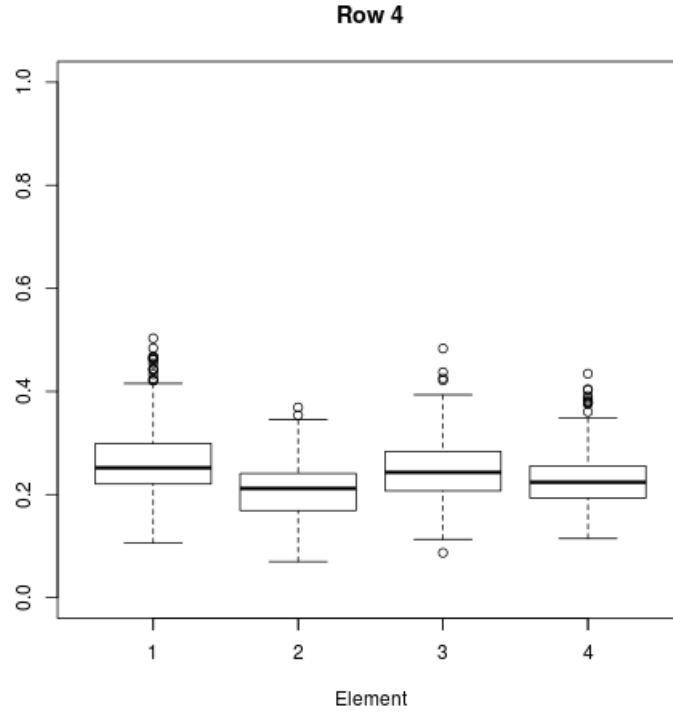
Row 1

36

Row 2

Row 3

Row 4
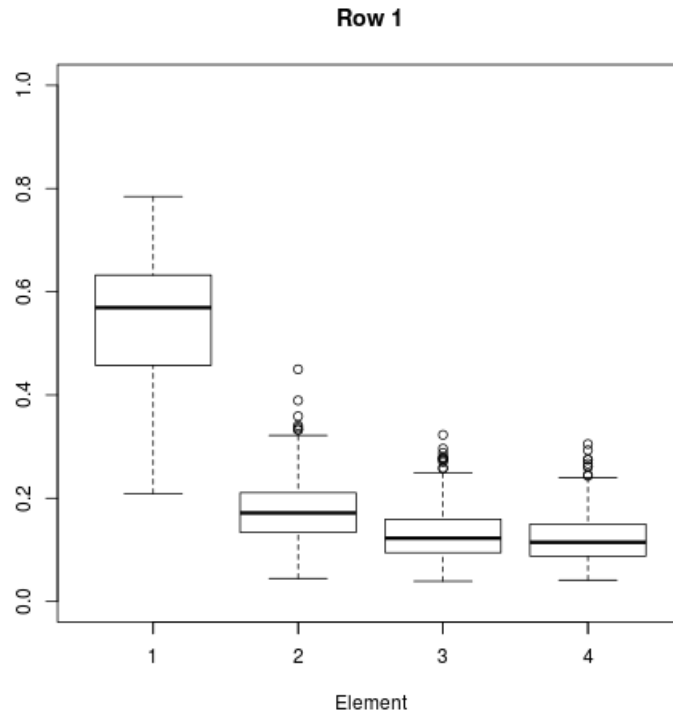
3.3   Simulation 2: Inferring Correct Ancestries

The aim of this was to try and create datasets where most transmission happened within households and infection spread from household to household via the community where the equal probability of moving to a new household modelled random mixing in a community setting. In these datasets there would be several new cases occurring in different households at the same time, if we take into account the household structure then we would assume that a new case in household A was infected by a previous case in household A. If we do not have any group framework in outbreaker then we have a collection of cases that happened around the same time with barely any differences in the DNA sequences between them.

The transmission tree below is from such a dataset, the node colours rep-

resent the group that the case belongs to. The light blue nodes are members of the community, other colours are members of different households.
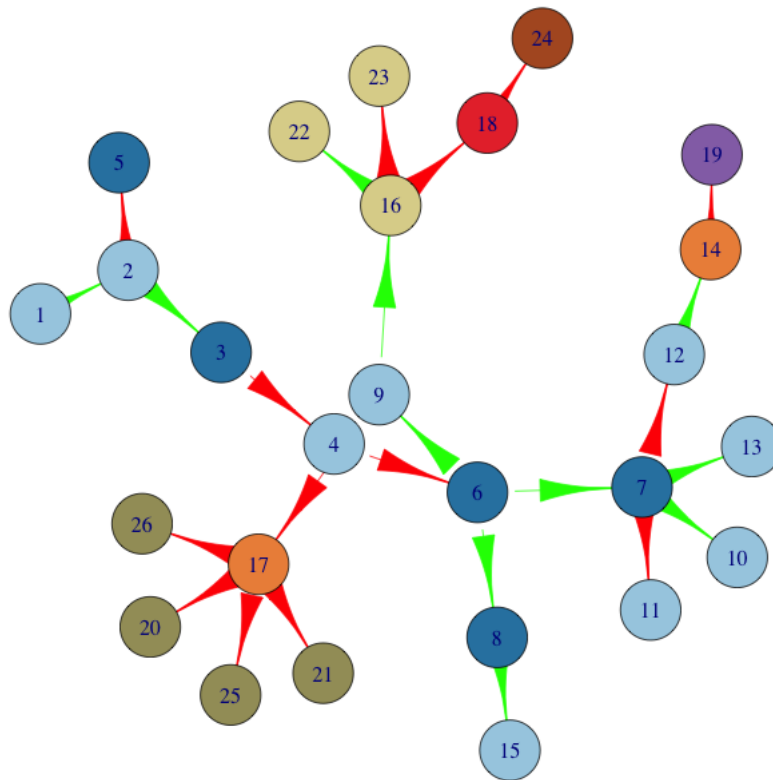
True ancestries
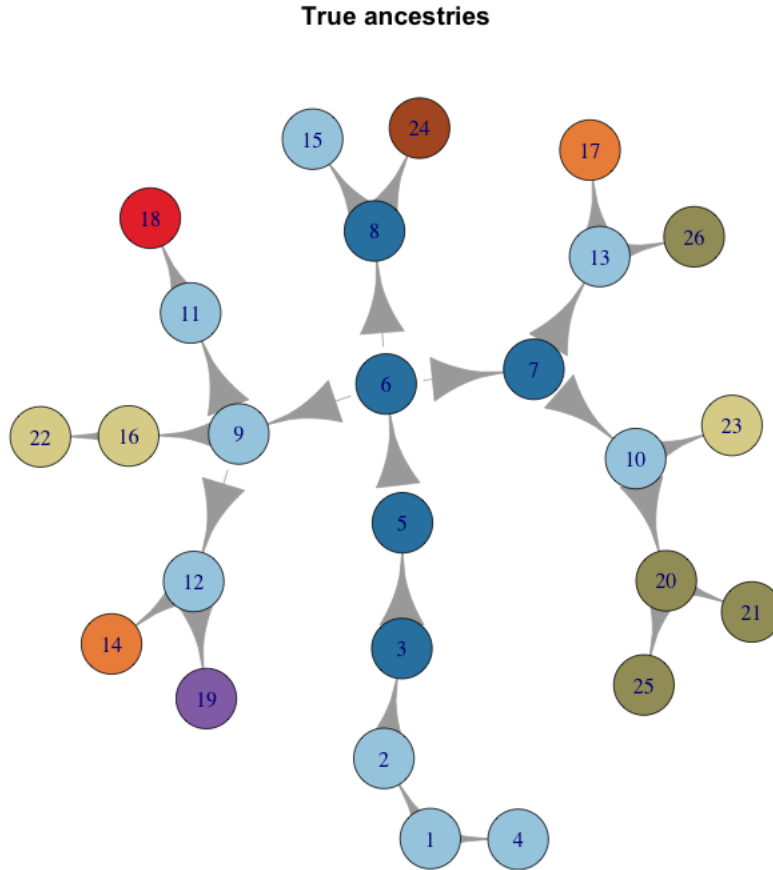
As described, most transmission takes place between members of the same household and the infection mostly moves from household to household via the community. We ran two different instances of outbreaker on this dataset to see which one could infer the correct ancestries better, one run had no group data and the other had group data and priors that strongly favoured unequal transmission probabilities. For this ancestry the outbreaker run with group data inferred 42% of correct ancestries (11 out of 25) whereas the run without group data inferred 23% of correct ancestries (6 out of 25). The plots below show the consensus transmission trees obtained from
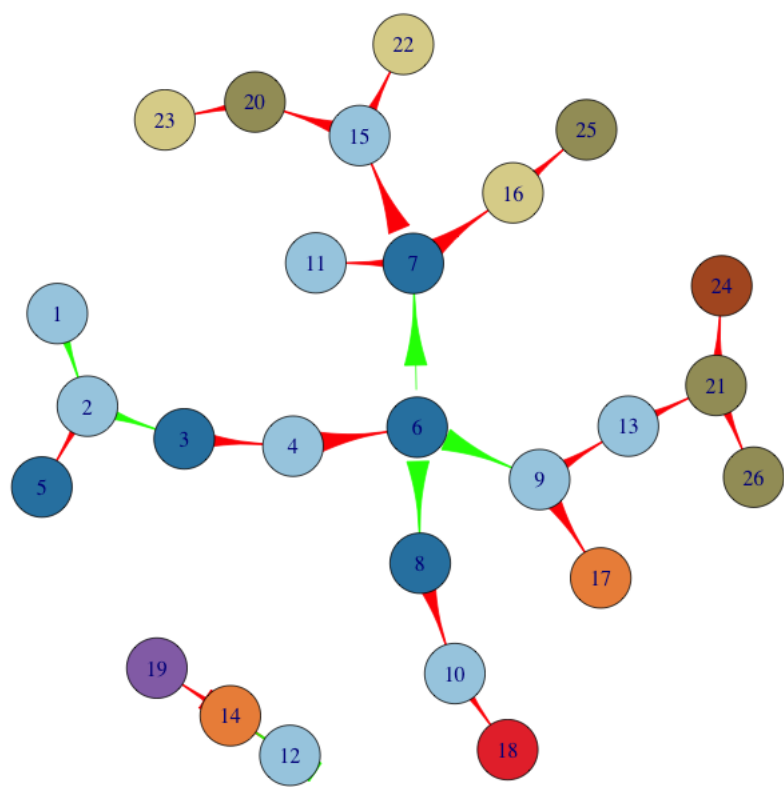
outbreaker's `get.tTree()` function. The nodes are again coloured based on group membership and the edges are coloured green if they are correct inferences of the true ancestry above.

Consensus ancestries with spatial data

Consensus ancestries with no spatial data

Chapter 4

Discussion

4.1 Results

4.1.1 Estimating Transmission Parameters

The results of the EVD group simulations show that with uninformative priors the model is able to infer transmission probability parameters well where the true parameter values are greater than 0.05. For the case where the true parameter value was 0.05 the true parameter value only fell inside the 95% equal-tails interval of the posterior samples around half of the time. One explanation for this could be that the priors used in the test were uninformative which means that there was a bias in the model towards equal transmission probabilities within a row. We can see from the boxplots of the posterior medians for each simulation that the posterior samples for elements (3,2) and (4,1) are slightly lower than the other elements in their respective rows, perhaps a prior which favoured more unequal probabilities across rows would allow for more accurate inference of very small transmission probabilities.

One area the model did well in was inferring the larger transmission

probabilities of elements (1,1) and (2,2). The true parameter was within the equal-tails interval for each run in the majority of cases, although if we study the boxplots of the posterior medians we can see that the posterior medians vary a lot between runs. This suggests that the equal-tails intervals for these elements were quite wide for many runs because there must be cases where the posterior median was not close to the true parameter value but the true parameter value was still in the equal-tails interval of the posterior. The posterior median boxplots for rows 1 and 2 do not show the same features for other parameters, this means that the output of the model is still suggesting that the elements (1,1) and (2,2) could be larger than the others even if the posterior median is not as large as it should be. This could suggest to people using the model that transmission probabilities are not equal even if it does not deliver an accurate inference of the true parameter values.

4.1.2   Inferring Correct Ancestries

For the particular outbreak shown in the results section, adding the group framework to outbreaker greatly improved the proportion of correct ancestries inferred by outbreaker. The benefits of including group data can be seen from the consensus ancestries in the results section, given what we know about the rates of transmission between groups the outbreaker model with the group likelihood produces more sensible results in several cases. The group data has caused outbreaker to infer the correct ancestry between cases 16 and 22, it has also caused outbreaker to guess an ancestry from 16 to 23 which is a sensible inference since it was unlikely that the true ancestry would look as it does with two separate infections of the same household (this is unlikely because households are small relative to the overall population so if two members of a household are infected it is very unlikely that

43

the individual who infected the third member would no be one of the other two members). However, the majority of the ancestries correctly inferred by outbreaker with group data which were not also inferred by outbreaker without group data are transmission events between members of the community and members of households. One explanation for this could be that correct (or more sensible) inferences involving household members elsewhere in the tree made the the correct inferences between individuals 7 and 10 or 7 and 13 more probable.

4.2 Modelling Assumption

Our model makes some simplifying assumptions which are important to recognise. We must also consider how they may effect the interpretation and usefulness of the model output. The first and main assumption is that the likelihood terms for the genetic, epidemiological and group data are independent. This means that if we choose a candidate ancestry then we expect the values of each of the separate likelihoods not to be correlated. If the two cases that we have chosen occur within a short period of time then they would have a high epidemiological likelihood. Due to the short space of time there would not have been many mutations in the DNA sequences of the pathogen, therefore the cases would have a high genetic likelihood. This means that it is unlikely that the values of these two likelihoods would be entirely independent. However, combining these likelihood functions to take account of the correlation would be a very tricky process and would inevitably produce a more complex likelihood function - this has consequences in terms of computational effort when we attempt to compute the likelihood function for thousands of moves during an MCMC run. It is also hard to gauge what sort of effect this assumption will have on the output of the

model without computing a combined likelihood and comparing the results.

A second, assumption is that transmission probabilities between groups are modelled as remaining constant over time. However, the fact that the transmission probabilities are not modelling as changing over time might not stop us from noticing if this is the case during outbreak analysis. This way of modelling the transmission probabilities serves to constrain the situations in which the extended model would be appropriate. In the two simulation scenarios envisioned it is probably safe to assume that the transmission rate between groups stays fairly constant over the period of the outbreak. However this does rule out the model being used to assess the effectiveness of an intervention on an outbreak because we would expect that interventions might change the probability of transmissions between different groups. Although data from two outbreaks that are similar in every other respect apart from the presence of an intervention strategy during one of the outbreaks could be used to compare the effectiveness of the intervention strategy at preventing between group transmissions.

4.3   Further Work

The new parameters in the model cannot be considered too helpful if they unduly restrict other parts of the model. In this case the significant restraint placed on the rest of the model by the group transmission parameters is that we must assume that there are no unsampled cases between infected individuals, thus limiting the situations in which we can use the extended model to those where we can be sure that we have collected data on every case in an outbreak. These sorts of outbreak would be relatively small and in places where data can easily be collected, thankfully the types of scenario where we have imagined that this method will be useful match this

description.

When we are dealing with transmission between two cases where there are unsampled cases between them we do not know the group membership of the unsampled cases which causes problems when we come to assess the likelihood of this transmission event. The solution to this problem is to use a modified transmission probability matrix M^{κ} where κ is the number of unsampled cases. The values of this modified matrix would then be used in likelihood calculations, this could be implemented in future updates of the outbreaker model.

Bibliography

S. Cauchemez, F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle. A bayesian mcmc approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23:3469–3487, 2004.

S. Cauchemez, A. Bhattarai, T. L. Marchbanks, R. P. Fagan, S. Ostroff, N. Ferguson, and D. Swerdlow. Role of social networks in shaping disease transmission during a communtiy outbreak of 2009 h1n1 pandemic influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 108(7):2825–2830, 2011.

E. Cottam, Gal Thbaud, Jemma Wadsworth, John Gloster, Leonard Mansley, David J. Paton, Donald P. King, and Daniel T. Haydon. Integrating genetic and epidemiological data to determine transmission pathways of foot-and-mouth disease virus. *Proceedings. Biological sciences / The Royal Society*, 275(1637):887–95, 2008.

W. R Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice.* Chapman and Hall/CRC, 1996.

Stephen K. Gire, Augustine Goba, Kristian G. Andersen, and Rachel S. G. Sealfon. Genomic surveillance elucidates ebola virus origin and transmission during the 2014 outbreak. *Science*, 345(6202):1369–1372, 2014.

Nicholas C. Grassly and Christophe Fraser. Mathematical models of infectious diseases. *Nature Reviews Microbiology*, 6:477–487, 2008.

D. T Haydon, M Chase-Topping, D. J Shaw, L Matthews, J. K Friar, J Wilesmith, and Woolhouse M. E. J. The construction and analysis of epidemic trees with reference to the 2001 foot-and-mouth outbreak. *Proceedings of the Royal Society B: Biological Sciences*, 270(1511):121–127, 2003.

Thibaut Jombart, Rosalind M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–90, 2011.

Thibaut Jombart, Anne Cori, Xavier Didelot, Simon Cauchemez, Christophe Fraser, and Neil Ferguson. Bayesian reconstruction of disease outbreaks by combining epidemiologic and genomic data. *PLoS Computational Biology*, 10(1):e1003457, 2014.

Marco J. Morelli, Gal Thbaud, Jol Chaduf, Donald P. King, Daniel T. Haydon, and Samuel. Soubeyrand. A bayesian inference framework to reconstruct transmission trees using epidemiological and genetic data. *PLoS Computational Biology*, 8(11):e1002768, 2012.

Philip D. O'Neill. A tutorial introduction to bayesian inference for stochastic epidemic models using markov chain monte carlo methods. *Mathematical Biosciences*, 180:103–114, 2002.

The R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2014. URL `http://www.R-project.org`.

C. P. Robert. *The Bayesian Choice*. Springer, second edition, 2007.

G Streftaris and G. J Gibson. Bayesian inference for stochastic epidemics in closed populations. *Statistical Modelling 4*, 4(1):63–75, 2004.

P. Teunis, J. C. M. Heijne, F. Sukhrie, J. van Eijkeren, M. Koopmans, and M. Kretzschmar. Infectious disease transmission as a forensic problem: who infected who? *J. R. Soc. Interface*, 10, 2013.

V. E. Volchkov, V. A. Volchkova, A. A. Chepurnov, V.M. Blinov, O. Dolnik, Netesov S.V., and H. Feldmann. Characterization of the l gene and 5' trailer region of ebola virus. *J. Gen. Biology*, 80:355–62, 1999.

Jochen Voss. *An Introduction to Statistical Computing - A Simulation-Based Approach*. WIley Series in Computational Statistics. John Wiley and Sons Ltd, 2014.

Jacco Wallinga and Peter Teunis. Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American journal of epidemiology*, 160(6):509–16, 2004.

The WHO Ebola Response Team. Ebola virus disease in west africa - the first 9 months of the epidemic and forward projections. *The New England Journal of Medicine*, 371(16):1481–1495, 2014.

Rolf J. F Ypma, W. Marijn van Ballegooijen, and J. Wallinga. Relating pylogenetic trees to transmission trees of infectious disease outbreaks. *Genetics*, 195(3):1055–62, 2013.