Adding a spatial element to Outbreaker

December 24, 2014

"No man is an island entire of itself; every man is a piece of the continent, a part of the main; "
John Donne, Devotion Upon Emergent Occasions.

1 Previous work

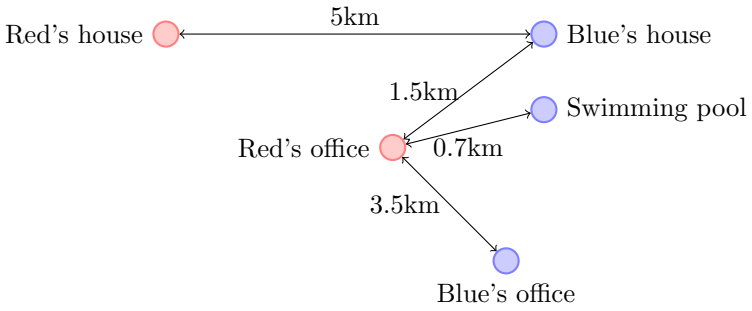
Outbreaker is an R package which fits a Bayesian model to genetic and epidemiological data collected from an outbreak. Outbreaker produces a final transmission tree, infers parameter estimates such as R_0 and the mutation rate μ and also infers augmented data such as infection times and unsampled cases using the data provided. The model behind Outbreaker is the most recent improvement in a line of work which aims to reconstruct the transmission trees of epidemics using epidemiological, geographical and subsequently genetic data. The general approach is to construct a likelihood function that gives a likelihood value for each particular tree, we can then use these likelihood values to sample from the posterior distribution of transmission trees or more simply we can scrutinise the trees which have the highest likelihoods. Before DNA sequencing became cheap and efficient it was a lot more difficult to sample DNA sequence data for the pathogen for each case of an outbreak. The earliest implementations of the likelihood approach use more traditional epidemiological case data such as times of infection and recovery as well as data about the geographical distance between two cases. Haydon et al 2002 engineer this approach to use on the 2001 UK FMDV outbreak, they construct potential tranmission trees by adding nodes to an initial tree (constructed using contact tracing data) based on the probability that one farm infected another. The probability that farm A infected farm B is based of how well the period that farm A was infectious overlaps with the predicted time duration in which farm B was infected. Additionally, the researchers consider how far apart the two farms are, farms which are a long distance away from each other are much less likely to infect each other.
Cottam el al expand on this work by expanding the analysis to include the genomic data of samples collected during the 2001 FMDV outbreak, they construct a genomic likelihood which depends on the expected number of nucleotide changes in the sequence in the time between the two cases. They select the trees which have the highest genomic likelihood and then assess these trees based on

their epidemiological likelihood. This reduced the pool of genomically plausible trees to 4 final trees which encompassed 95The genomic and epidemiological likelihoods were combined into a single likelihood (along with geographical likelihood involving distances between farms) by Ypma et al in 2012. This process makes use of all data available and allowed the researchers to infer likely tranmissions of avian flu on Dutch farms. Morelli et al 2012 made further improvements to the model by moving to a Bayesian framework with a combined genetic and epidemiological likelihood and removed the assumption that the two likelihoods are independent. The farms which were infectious at a similar point in time we expect will have similar pathogen DNA sequences so there is no reason to expect that the two likelihoods should be independent. The researchers then used an MCMC approach to sample from the posterior distribution on the tree and model parameters to reconstruct a transmission tree.
The model behind Outbreaker takes this process another step further by using the generation time distribution as part of the epidemiological likelihood (an approach first used by Wallinga and Teunis for inferring transmission networks), introducing the parameter κ to account for unsampled generations between cases and generalising the ideas used in the previous models to account for the various types of pathogens that infect humans rather than farms such as requiring that the epidemic has one source. The aim of this project will be to incorporate a spatial element to the Outbreaker model which takes into account geographical and contact tracing data, hopefully this will further improve the reconstruction of transmission trees or allow us to provide estimates for other parameters of interest.

2 Geographical data and Outbreaker

Previous work on farms (citations) incorporated geographical data into the likelihood of transmission between farms by assigning a value based on the Euclidean distance between the two farms. For example Haydon et. al fitted an exponential function to some data for known infections between farms and this was then used to assign geographical likelihoods. This approach, although it works well for static farms, does not translate across well to outbreaks in humans and the Outbreaker model for several reasons. Some of these reasons are theoretical problems and some are incompatibilities between the Outbreaker model and the idea of spatial likelihoods.

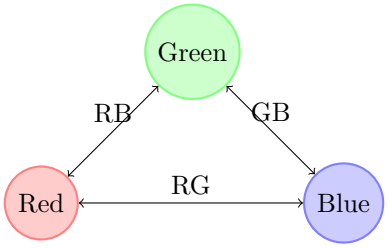
2.1 Theoretical problems



The theoretical problem with a spatial likelihood for transmission between two human cases is that humans are not a fixed entity like farms and move around for reasons of work and leisure. We would need to decide upon a distance measure between two cases that represents how far apart the two cases when the Euclidean distance between the two cases varies at different times. For example two people, who I will call Red and Blue, could live 5km apart which might lead us to conclude that the distance between Red and Blue is 5km. However Red might work in an office which is 3.5km from the office in which Blue works, or the office in which Red works could be 1.5km from Blue's house. Even more confusingly Blue might go swimming at a pool which is only 0.7km away from Red's office. How would we go about deciding how far apart Red and Blue are? We could take the minimum distance of 0.7km between the swimming pool and Red's office, but this might not be a good representation of how far apart Red and Blue are - Blue could only go swimming once a month or might only go when Red is at home. We could again be dogmatic and declare that we will take this distances between where Red and Blue live, but this ignores the obviously plausible scenario of Red or Blue being infected at their workplace. To get a good idea of what would be a good distance between Red and Blue we would have to know a lot of information about where they go and when, collecting this data for anything other than a small outbreak would be extremely difficult. There simply is no catch-all measurement of distance between Red and Blue that would be applicable in even most circumstances, we would need to know a vast amount of information about the lives of each case and also about the transmission mechanism of the pathogen (which we might even be undertaking the analysis to try and understand!). When the level of information on the whereabouts of each case is so detailed we may as well be dealing with contact tracing data, which would recommend a different approach.
It is worth noting that this is not a hard problem and can be solved by making decisions about what constitutes a distance between two cases, but these decisions can be made under certain conditions when we already have knowledge of the pathogen and the case number is small enough to track the travels and habits of each case.

2.2 Model incompatabilities

The reason that the Outbreaker model does not sit well with geographical likelihoods is that it allows for the existence of unsampled generations between cases. This lack of knowledge about the intermediate generations is usually dealt with using convolutions of probability distributions, if Outbreaker wants to calculate the epidemiological likelihood betwen Red and Blue whilst assuming that there is a missing generation, Green say, between them then we would look at the estimated time that Red was infected and the estimated time that Blue was infected and assess the plausibility that the interval between these two times is long or short enough for two generations of infection to have occured within it. Speaking more rigorously, we would calculate the probabilitiy that two random variables drawn from the generation time distribution of the pathogen sum to our observed interval, this gives us a likelihood of the parameters of the generation time distributuion given our observed data.
The problem with using this technique for Euclidean (or other) distances is that the individual Euclidean distances between Red, Blue and Green do not all sum together to give the distance between Red and Blue. In the example below, $RB + GB \neq RG$.



The number of generations inbetween Red and Blue does not have any bearing on how far apart Red and Blue are, there could be 10 unsampled generations between Red and Blue yet Red and Blue could only be a small distance apart, what would be important in assessing the likelihood of this transmission chain would be the distances between each unsampled generation but because they are unsampled we don't know what these distances are. We cannot try and sum the likelihoods of all possible distances because there are infinitely many combinations of possible locations for Green which leave Red and Blue the same distance apart. In densely sampled outbreaks the technique of using convolutions to assess likelihood would not be fruitful. If we know that the outbreak has been fully sampled we can use some exponential function as a geographical likelihood, but this is rarely the case (and it is also hard to know for certain whether it is the case or not).

3 Contact tracing data

4 Transmission between groups

In certain outbreak scenarios cases may belong to two or more different groups which have different levels of contact within and between groups and therefore potentially different rates of transmission events within and between groups. One example of this could be groups of cases on different wards of a hospital, a case on one ward is more likely to transmit infection to cases on the same ward than cases on a different ward because they have more contact with people in the same ward as them. Another example could be classes of school children, there may be a higher level of transmission between children in the same classroom and a lower level of transmission between children from all classes out in the playground. We can incorporate this feature into the Outbreaker model using a matrix which contains the parameters for the probability of transmission within and between each group. For two groups A and B the matrix would be as follows:

$$\begin{pmatrix} P_{aa} & P_{ab} \\ P_{ba} & P_{bb} \end{pmatrix}$$

Where P_{xy} is the probability that someone from group x infects someone from group y, note that for each group j, $\sum_i P_{ji} = 1$.

Using this we can construct a *group likelihood* which is the likelihood that the proposed transmissions between and within groups took place given the proposed tree and the proposed group transmission parameters. We can model a single transmission event between a person in group x and a person in group y as a Bernoulli trial with probability P_{xy} of succeeding. Successive transmissions are therefore successive Bernoulli trials with differing probabilities of success depending on the group membership of the cases involved. To work out the group likelihood of a transmission event we would need to consider the relevant element in the transmission matrix, to work out the group likelihood of the whole tree we would need to take the product of the likelihood of all transmission events in the tree.

The existing pseudo-likelihood function of the outbreaker model is composed of the product of the genetic and epidemiological likelihoods. We need to construct a group likelihood function to multiply onto the existing likelihood function, the group likelihood function will give us a likelihood value for the probabilities in our transmission matrix given the observed and augmented data. Some of the observed and augmented data will not play a role in determining how likely the transmission probabilities are, such as the genetic sequences and timings of each case. The data that will play a role are α_i, the index of the ancestor of case i (and their corresponding group) and κ_i, the number of generations between i and κ_i.
Gap

Let g_i be the group of case i and g_{α_i} be the group of the most recent ancestor of case i. The probability of a case from g_i infecting a case from g_{α_i} is given by the corresponding entry in the transmission matrix, $P_{g_i g_{\alpha_i}}$. Denote the transmission matrix M. We can first split the probability into two terms because the outbreaker model considers tranmissions to be independent events, so the probability of the group of α_i is only dependent on the ancestor of α_i and not case i.

$$Prob(g_{\alpha_i} = y, g_i = x|M) = Prob(g_{\alpha_i} = y|M)Prob(g_i = x|g_{\alpha_i} = y, M)$$

Let us denote α_i as case j, then we can rewrite the first term and we know the value of the second term:

$$Prob(g_j = y|M) \cdot P_{xy}$$

The first term can be expanded by considering α_j, since the probability of g_j being y depends on what group α_j is in:

$$Prob(g_j = y|M) = Prob(g_j = y, g_{\alpha_j} = z|M)$$

which can again be manipulated

$$Prob(g_{\alpha_j} = z|M)Prob(g_j = y|g_{\alpha_j} = z, M) = Prob(g_{\alpha_j} = z|M) \cdot P_{zy}$$

This chain will continue backwards up the connections in the proposed transmission tree until it reaches the root of the tree, at this point $Prob(g_{root}|M) = 1$ since the group of the root is given and not contingent on a further ancestor.

Therefore the likelihood of transmission between case i and α_i is the product of all of the probabilities of transmissions between all of the nodes from the root of the tree down to the one that we are currently considering.

Or:

Alternatively the group structure of the model could affect the term $p(\alpha_i)$ in the general pseudo-likelihood. Since the probability that α_i is the ancestor of i is affected by their respective group membership. Thus $p(\alpha_i)$ is now

$$p(\alpha_i|g_i, g_{\alpha_i}, M) = P_{g_i g_{\alpha_i}}$$

for individual i and

$$\prod_i P_{g_i g_{\alpha_i}}$$

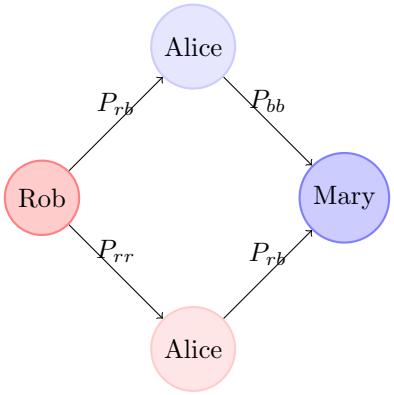
for the entire tree.

4.1 What to do if κ is greater than 1

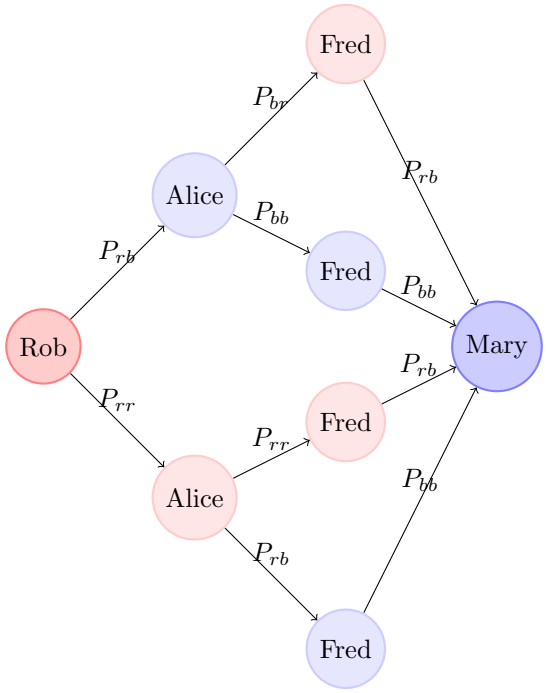
One problem arises when we are dealing with transmission between two cases where there are unsampled generations inbetween because we do not know the

group membership of the unsampled cases. We can get around this problem by summing over all of the possible group membership combinations of the missing cases to arrive a likelihood of transmission between the two known cases. For example if Rob is in red ward and Mary is in blue ward and we are considering the group likelihood of a transmission event between them within one unsampled generation, Alice say, then we would sum the likelihoods that Rob infected Alice who then infected Mary for the case that Alice is a member of any possible group - in this case red or blue ward.
For small values of κ we can easily calculate the sum of these likelihoods, in the image below the group likelihood that Rob infected Mary would be

$$(P_{rb} \cdot P_{bb}) + (P_{rr} \cdot P_{rb})$$



I have called the unsampled case Alice but in reality we would know nothing about them apart from the fact that we must assume that they belong to one of the groups that we are looking at. For a given value of κ we have to consider 2^κ different potential group combinations so for cases where κ is large we will have to perform a lot of computations. Below is an example for when $\kappa = 2$ by adding Fred to the transmission path.



If we have n groups then calculating the group likelihood for a transmission event with κ unsampled cases between them will require n^{κ} individual calculations for each possible permutation of the different groups that the unsampled cases might belong to.

4.2 Why bother doing this?

An analysis undertaken where cases have been separated into groups depending on their contact patterns could give us interesting information about the transmission dynamics of the pathogen. Differences in the probabilities of transmissions between and within groups could help narrow down the processes which are driving transmission. We could perhaps find that P_{ab} is much larger than P_{ba} and we could then look into why this might be the case, perhaps each day group A takes part in a buffet lunch before group B and this is why infection only passes in one direction. There could also be differences in the values of P_{aa} and P_{bb} and we might try to determine why one group has less chance of infecting each other, this might be that one group has better hygiene practices or not as sociable.

Additionally if we are fairly confident in the transmission tree that we construct then we can look at where in the tree the infection moves between groups. We can then investigate these cases to try and figure out how the infection moves between groups. Perhaps Fred from red ward used to go for a sneaky smoke with Alice from blue ward before Alice and then Fred subsequently became

symptomatic. There might be a common feature of all of the cases of infection between groups, we might find that the infection was passed from scout leader to scout leader who then each passed on the infection to the scouts in their troop. We could narrow down the transmission event to the monthly scout leader poker night.

4.3 Potential implementations in the current MCMC algorithm

Teunis et al 2014 describe a way of updating a matrix where the sums of the rows must be one. First the entries P_{ab} are logit transformed:

$$z_{ab} = \log \frac{P_{ab}}{1 - P_{ab}}$$

Then we update z_{ab} with $\epsilon \sim N(0, 1)$:

$$z'_{ab} = z_{ab} + \epsilon$$

Finally we reverse logit transform all of the z'_{ab} and scale the entries in the row so that they sum to 1:

$$v'_{ab} = \frac{\exp z'_{ab}}{1 + \exp z'_{ab}}$$

gives the reverse logit transform and we rescale by simply dividing by the sum of all entries in the row:

$$v_{ab} = \frac{v'_{ab}}{\sum_{a=1}^{l} v'_{ab}}$$

4.4 Assumptions made

Adding in a separate group likelihood assumes that the likelihood functions involved are independent of each other, this is not obviously the case with the data. We might expect correlation between someone's group membership, when they were infected and the DNA sequence of their pathogen seeing as for the most part you are much more likely to be infected by someone within your own group.

4.5 Further simplifying assumptions that could be made

You could make the simplifying assumption that the probability of within group transmission is the same for all groups and only estimate one within group transmission probability, this might be a sensible assumption for the case where the groups are all identical but separate such as identical wards on different floors of a hospital (as in same number of patients and same hygiene practices etc).

4.6 MRSA dataset

A good data set to experiment with is the MRSA outbreak data from the special care baby unit (SCBU) of a hospital in Cambridge. The researchers used whole genome sequencing to identify cases that came from the same lineage of MRSA to rule out that there had been several concurrent infections with different strains of MRSA at the same place. The researchers found that there was transmission between infants in the SCBU, their mothers and the wider hospital community. One staff member was identified to have unknowingly reintroduce MRSA to the SCBU after the infection had died out there and a deep clean had taken place.

It would be interesting to see if we could come to the same conclusions by fitting the Outbreaker model with different groups.

4.7 Recreating simulated outbreaks

Existing outbreak simulations could be adapted to replicate the dynamics of transmission between two different groups. Outbreaker could then be run on the simulated outbreak data to see how well the transmission tree and other parameters are inferred.

An outbreak simulator is already included in the Outbreaker function `simOutbreak`, this could be modified to incorporate the new group dynamics. Specifically we could modify the function which chooses the infector of an infected individual (in the code this is implemented by first choosing the ancestors and then choosing who has been infected by them), currently this is sampled 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 could try to incorporate the probability of transmission within and between groups into the probability that one case in an infector of another. If we denote P_{ic} as the probability that a person from the group to which the newly infected case c belongs to is infected by someone from the group to which case i belongs to, where i is the same index used in the generation time distribution:

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

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 and perhaps we would choose a

group which the infection will begin from. Then the procedure takes place just as before but now new cases have different probabilities of being infected by the exisiting cases depending on the within and between group transmission probabilities. Finally 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. We would also need to come up with a rule for the group membership of imported cases, they could either be assigned to the existing groups at some specified frequency or marked as group unknown which would bring up questions about how to treat these cases in terms of transmission probabilities. In some situations such as hospital wards it might be sensible to take the first approach, our groups might be in one large ward and one small ward so we would expect imported cases to have joined one of the wards and although their ancestors are not present in our transmission tree we expect them to inherit the transmission dynamics of their group once they join one. For the second scenario we would have to decide what would be a sensible transmission probability between each group and a case where the group is unknown, we might assign a "baseline" transmission probability that we use for any cases where one or both of the groups are unknown during a transmission event. This is a better approach if our groups are set within a larger community that has some probability of transmission between the members (this is equivalent to assigning all imported cases to an extra group with equal within and between group transmission). Another benefit of this second approach is that it also provides a way of dealing with cases where group membership data is missing.