Tracing Infectious Diseases using Genetic and Spatial Data

Bryan Hooi
Advisor: Susan Holmes

Stanford University

May 15, 2014

Abstract

The analysis of viral genetic sequence data collected during disease outbreaks has emerged as a promising new tool for understanding infectious disease dynamics and designing control measures against infectious diseases. Hence, there is a need for statistical methodologies that effectively integrate genetic data with other epidemiological data to perform inference on the underlying disease dynamics.

In this project, we develop a Markov chain Monte Carlo framework which incorporates genetic, temporal and spatial data into a single framework that infers who infected whom among a group of patients, as well as various disease-related parameters, while allowing for missing data. Using simulations, we show that our algorithm determines who infected whom more accurately than existing methods, and has the additional benefit of jointly inferring unknown parameters such as the disease mutation rate, transmission rate, and contact network related parameters. We apply our approach to analyze 433 H1N1 viral genetic sequences drawn from the early stages of the 2009 H1N1 influenza pandemic, and to estimate the basic reproductive number of the pandemic.

We then extend our framework: firstly, we investigate how best to incorporate spatial data by comparing several ways of measuring the distance between two locations: namely, geographical distance, travel time and route length. Secondly, we extend the disease transmission model to allow for mixtures of different transmission routes, for example, allowing diseases to travel along an air traffic network as well as over land, and extend our inference algorithm to this setting.

Contents

1 Introduction

1.1 Compartment Models vs. Network Models

In the study of infectious diseases, the most common model has long been the compartment or "mean field" model, in which each infected individual is equally likely to spread the disease to any susceptible member of the population (Kermack & McKendrick, 1932). However, this approach ignores the fact that individuals are embedded in contact networks along which epidemics generally spread. The heterogeneity between individuals in contact networks has been shown to have important implications on epidemic dynamics as well as on the recommended strategies for controlling the spread of these epidemics (Anderson, 1991).

1.2 Inferring Epidemic Contact Networks

Due to the importance of the contact networks over which diseases spread, understanding these networks using past epidemic data are have important implications on the design of control measures for future outbreaks. While many studies have studied contact networks by simulating from a variety of random graph models, relatively few have tackled the associated inference problem of inferring the parameters of the contact network (such as the p in Erdos-Renyi $G(n,p)$ models) from epidemiological data. Among those that do are Britton and O'Neill (2002), which models the contact network as an Erdos-Renyi random graph, and uses a Markov chain Monte Carlo approach based on epidemic data to infer its parameters. This line of work has been advanced by Groendyke et al. (2010), who use the SEIR stochastic epidemic model, and further by Groendyke et al. (2011) to use exponential-family random graph models to model the transmission process over the contact network.

1.3 Genetic Sequence Data

As genetic sequence data collected during disease outbreaks becomes increasingly commonplace, such data has emerged as a promising way to better understand disease dynamics. So far, genetic data has primarily been used in the form of phylogenetic tree analysis (Grenfell et al., 2004). Others such as Jombart et al. (2011) use a maximum parsimony and likelihood based approach to derive the most likely ancestor of each patient's viral genetic sequence. Ypma et al. (2010) construct a likelihood function combining genetic, spatial and temporal data, assuming these components are independent of one another to construct a Bayesian inference scheme to obtain posterior intervals for the viral mutation parameters and the transmission history of the disease. Morelli et al. (2012) uses a more complex Bayesian inference scheme that allows for dependence between the likelihood components.

2 Problem Description

2.1 Data

The data consists of 433 viral H1N1 genetic sequences sequenced at the hemagglutinin and neuraminidase genes, as well as the geographical (latitude/longitude) coordinates of the associated patients. The sequences are based on freely available data on GenBank (Benson et al., 2010), aligned, and with disease detection time and date annotated by Jombart et al. (2011).

We computed genetic distances between each pair of sequences using the Kimura (1980) distance metric, which uses separate probabilities for transitions (mutations between A and G, or C and T) versus transversions (all other single-base mutations). These genetic distances were computed using the `ape` package (Paradis et al., 2004).

2.2 Overview of our project

Our project has the following main sections, all related to the overall theme of incorporating genetic and spatial data to understand the dynamics of infectious diseases.

- **Descriptive Statistics**: we summarize the data and describe the distribution of H1N1 patients geographically and temporally.
- **Genetic Clustering**: to better understand how genetic sequences relate to disease transmission, we form clusters of patients with similar genetic sequences and examine how these clusters relate to epidemic progression.
- **Correlation between Distance Metrics**: we formally validate the relationships we have found between genetic similarity, geography and time by performing hypothesis tests.
- **Bayesian Inference using Markov Chain Monte Carlo**: we develop a statistical model describing the generative process for epidemic and genetic data. We then construct a Markov Chain Monte Carlo algorithm that takes as input the detection times, locations, and genetic sequences of each patient (some of which may be missing), and makes inference on 1) parameters of the epidemic model (disease transmission rate and genetic mutation rate), 2) parameters of the contact network, and 3) who infected whom in the epidemic. We then test the algorithm on real and simulated data.

We then extend the framework in two ways:

- **Comparing spatial distance metrics**: we assess how best to incorporate spatial distance into the model by comparing several distance metrics: geographical distance, travel time and route length distance, where the latter two are computed using Google Maps API, to determine which distance metric is most informative about disease transmission.

- **Mixture models for disease transmission:** we extend the disease transmission model in our inference framework to allow for a mixture of several different transmissions routes, motivated by the importance of air travel as a contributor to transmission routes. This allows us to both estimate the importance of air traffic for disease transmission, as well as to more accurately reconstruct disease outbreaks by taking air traffic into account.

2.3 Relation to existing work

Our project builds on the work of Jombart et al. (2011), who developed the `seqTrack` algorithm, which uses a maximum parsimony and likelihood approach to infer genealogies from genetic and epidemiological data. We also build on Britton and O'Neill (2002), who also perform inference on the parameters of random graph models, but only use the times of detection and infection of each patient as input, not genetic sequence data, and also do not model contact networks.

Our work differs from these in that we set up a comprehensive inference framework that explicitly models the contact network as a random graph, combining this with models for epidemic transmission and genetic mutation. In this way we combine spatial, temporal and genetic data to jointly infer parameters related to disease transmission, mutation and the contact network, as well as who infected whom.

Moreover, while most existing inference approaches use relatively simple SI, SIR or SEIR based transmission models, we consider mixture models for disease transmission which allow diseases to transmit along multiple routes (for example, air and land based transmission). Finally, to our knowledge, the problem of comparing spatial distance metrics (geographic distance, travel time and route length) to determine which is most informative about disease transmission has not been addressed in the literature.

3 Descriptive Statistics

We first show the distribution of H1N1 patients geographically and over time. Geographically, patients are generally clustered in a relatively small number of locations, particularly around New York, Mexico City and California.

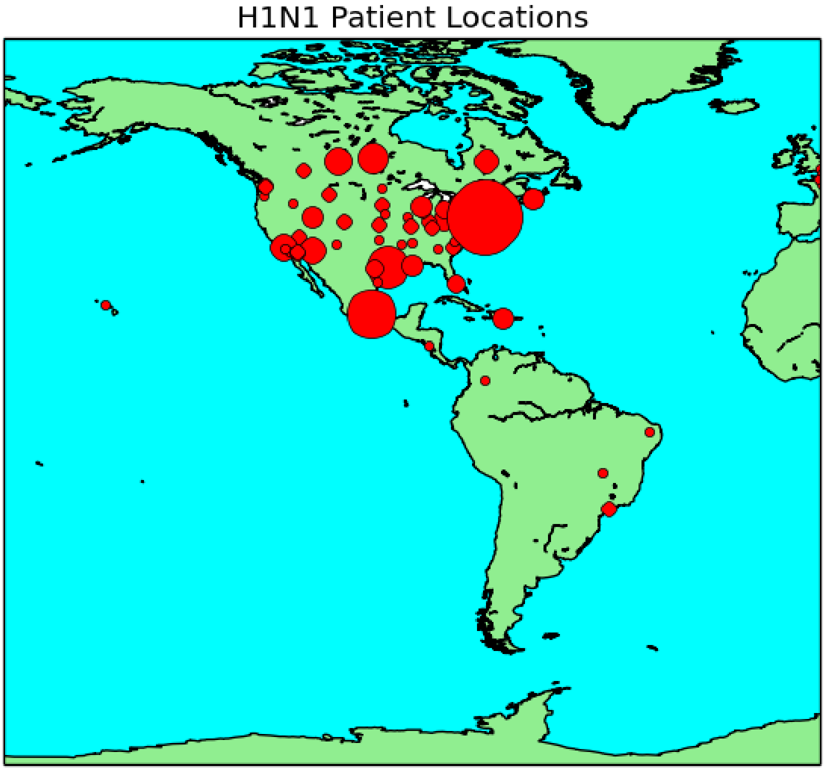


Figure 1: Distribution of patients in North and South America. The smallest circles represent a single patient; larger circles are locations with more than one patient, where the area of the circle is proportional to the number of patients. For ease of view, this only displays the 290 patients in North and South America (out of the full dataset of 433 patients).

As for their distribution over time, the number of cases detected reaches a peak around day 30, after which it slows considerably.

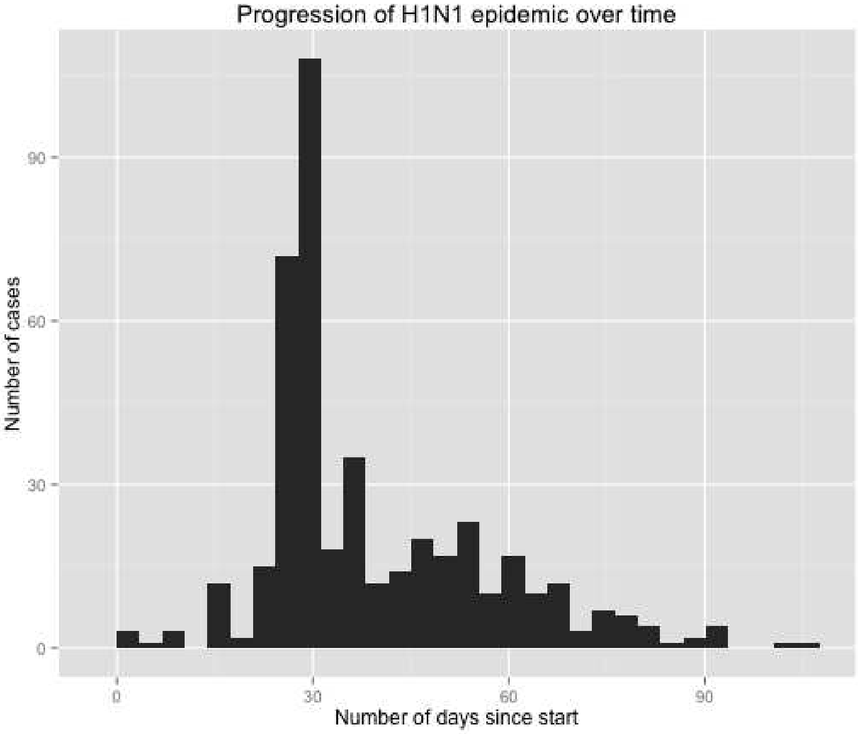


Figure 2: Distribution of patients' detection times over time since the start of the epidemic.

4 Genetic Clustering

We use genetic distance to separate the patients' viral genetic sequences into clusters of similar sequences. The clustering is done by k-medoids using the Partitioning Around Medoids (PAM) algorithm (Kaufman et al., 1987), with the Kimura (1980) distance metric.

The k-medoids algorithm is similar to the k-means algorithm, but works with an arbitrary matrix of distances between datapoints; rather than using points in Euclidean space as cluster centroids like k-means, k-medoids designates chosen data points as cluster centers. We used the implementation of K-medoids in the `cluster` package (Maechler et al, 2014).

Figure 3 and 4 show the reasonably strong separation of genetic clusters geographically: the outbreaks in California and Texas are mainly composed of cluster 2, while the outbreak in Mexico City is a mix of clusters 2 and 3. The largest outbreak in New York is dominated by cluster 1: the dataset contains 123 patients from New York, 104 of which are cluster 1. On the whole, we see strong evidence that genetic

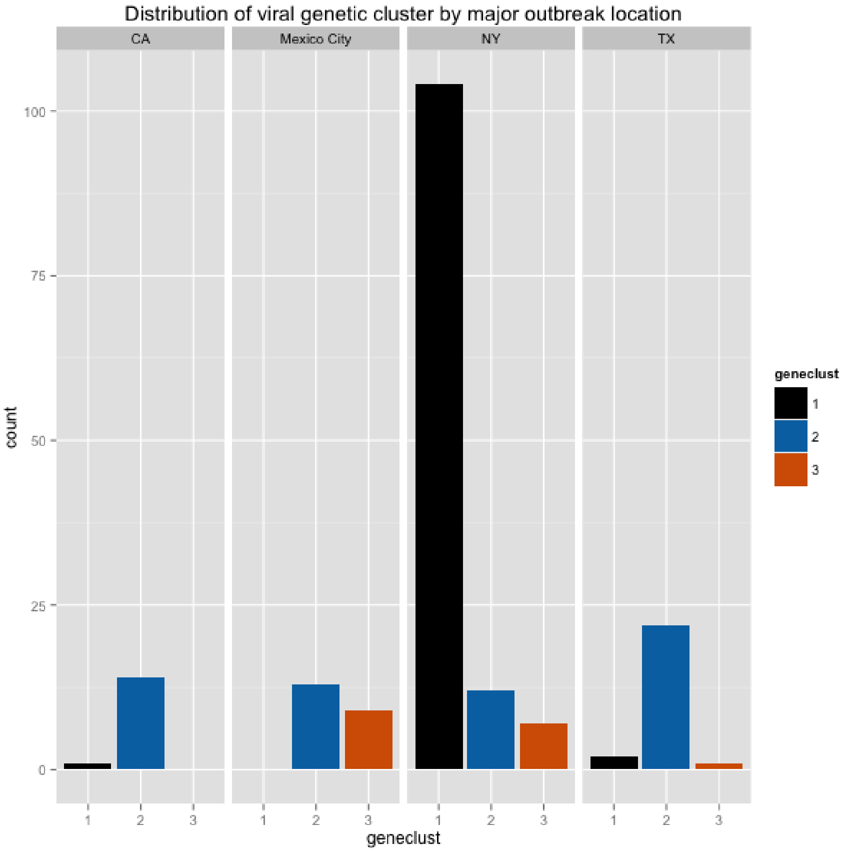


Figure 3: Distribution over gene clusters within each of the largest four outbreak locations.

clusters also separate geographically.

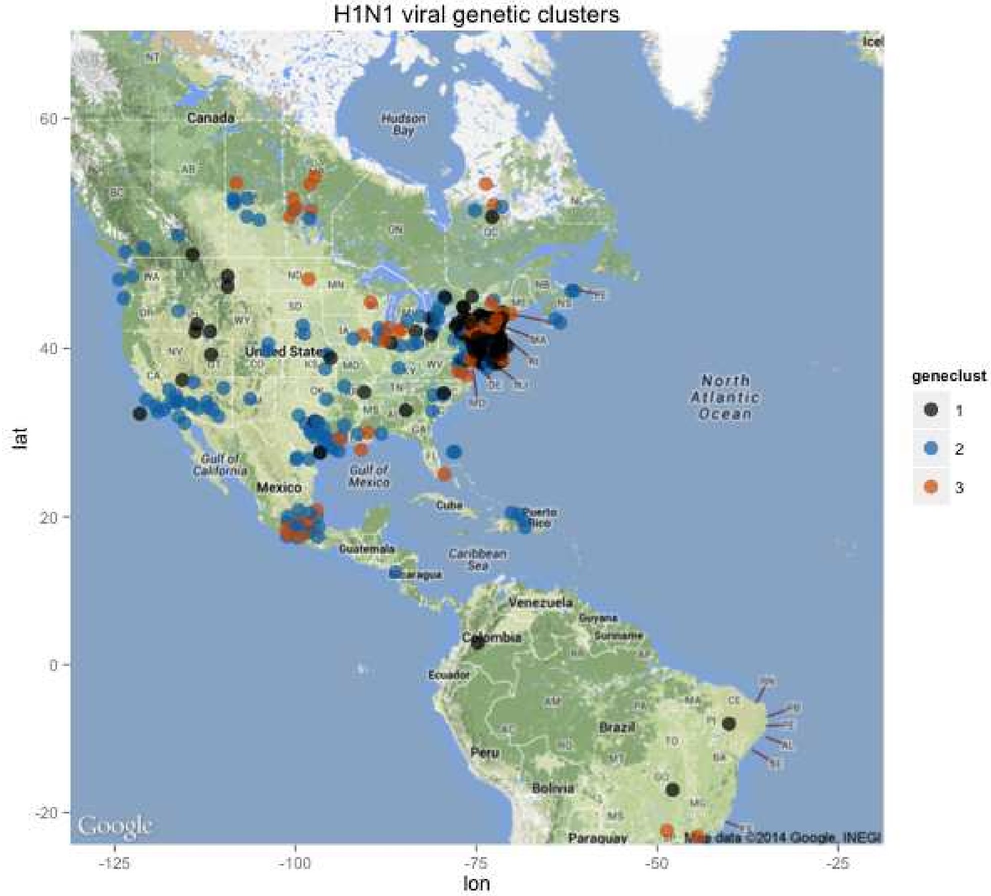


Figure 4: H1N1 patients, labelled according to their viral genetic cluster.

Plotting the clusters against time, we also see a distinction in the initial infection times of each cluster, showing that genetic clusters have a strong dependence with time as well.

On the whole, cluster 1 (black) primarily consists of a large outbreak in New York, and emerges the latest. Cluster 2 emerges the earliest but dies out around day 40, and is geographically evenly dispersed, comprising most of the cases in California and Texas. Cluster 3 is a relatively small cluster, mainly appearing in Mexico City and New York between days 15 and 40.

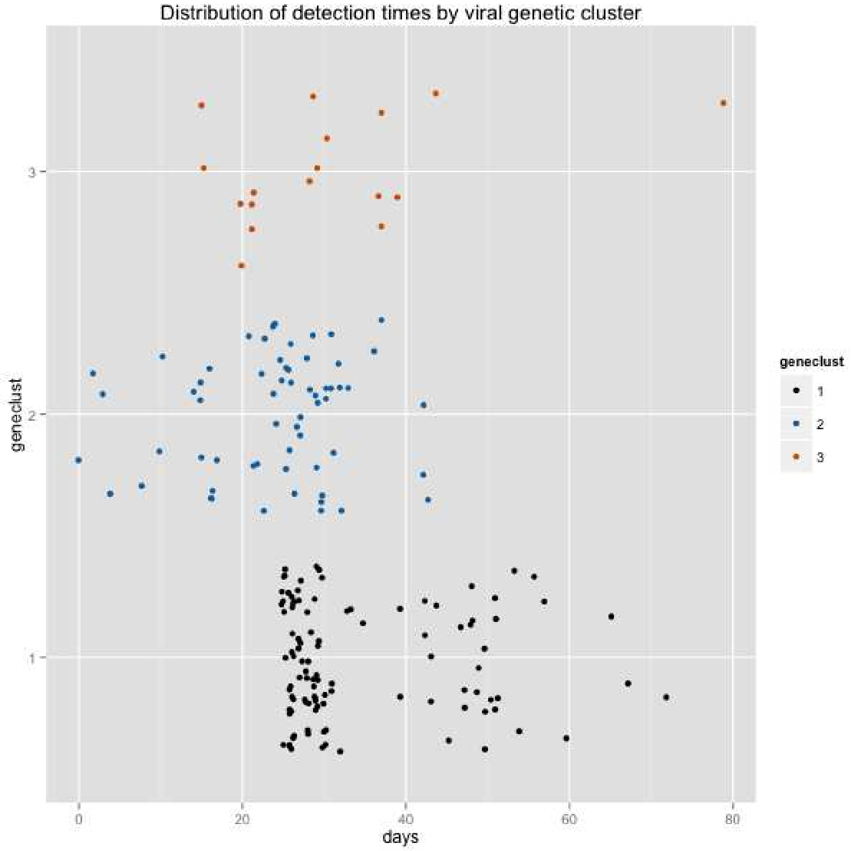


Figure 5: Distribution of detection times (in days since the initial case), separated into 3 genetic clusters.

5 Correlation between Distance Metrics

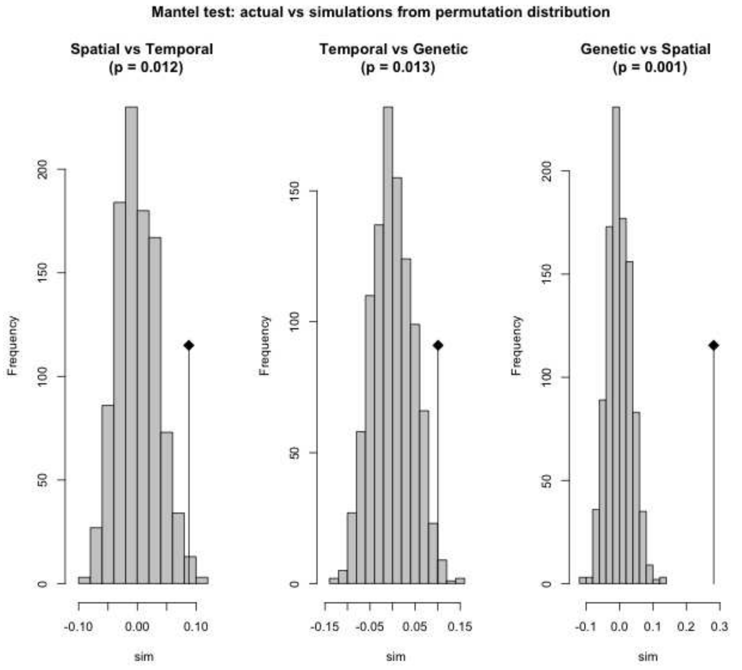
Having observed relationships between genetic distance, time, and spatial distances, we now test for these relationships more formally.

Observe that we can treat time as a distance metric, where the temporal distance between any two cases is the absolute difference in their detection times: in this way, spatial, temporal and genetic distance form 3 distance metrics on the set of patients. In line with our analysis in Figures 3 to 5, we expect that these three distance metrics are correlated: clusters of related cases are likely to be closely packed geographically, should occur around the same time, and the virus sequences involved should have high genetic similarity.

We test for this correlation between distance metrics using the Mantel test (Mantel, 1967). This is a test for correlation between two matrices, which in our case are the distance matrices computed between our patients. The null hypothesis is that there is no relationship between the distance matrices, which we test by computing the Pearson product-moment correlation coefficient between the matrices. We then compare this statistic to its null distribution, obtained by applying a randomly chosen permutation to both the rows and columns of one of the matrices, and computing the correlation coefficient using the permuted matrix.

We compute three Mantel statistics: 1) spatial and temporal distance, 2) temporal and genetic distance, and 3) genetic and spatial distance.

The results show an especially clear correlation between genetic and spatial distance, and a less strongly significant correlation between temporal distance and the other distance metrics:


Mantel test: actual vs simulations from permutation distribution

Spatial vs Temporal (p = 0.012)

Temporal vs Genetic (p = 0.013)

Genetic vs Spatial (p = 0.001)

Figure 6: Mantel test results for comparing pairs of distance matrices. The diamond indicates the observed test statistic (correlation between the actual distance matrices), and the histogram shows the null distribution of this statistic computed from 1000 permutations of one of the matrices.

6 Bayesian Inference using Markov Chain Monte Carlo

6.1 Model

In this section we describe the model we use for contact networks, epidemics, and genetic mutations.

The population is assumed to have a fixed population size n. We model the underlying, unobserved contact network structure by a random graph G over these nodes: the edges in G represent contact between the two patients. In this section, we model G using a $G(n,p)$ Erdos-Renyi model.

Next, we model epidemics over this contact network. Denote by H the directed graph corresponding to actual infective contacts: i.e. an edge $e_{ij} \in H$ iff one of the patients i and j infected the other.

We use a simplified version of the standard SIR (Susceptible, Infectious, Recovered) model in epidemiology: in our case, at any time individuals can only be in one of two states: Susceptible or Infectious. Infectious individuals can transmit infections to neighboring susceptible individuals by making infectious contacts. The time that passes between when individual i is infected and when i makes an infectious contact with a susceptible neighbor j follows an exponential distribution with an unknown parameter β, which we call the transmission rate. Letting t_{ij} be the time taken for i to infect j (assuming j is not infected by another node), we can write this as:

$$t_{ij} = \begin{cases} Exponential(\beta) & \text{if } e_{ij} \in G \\ \infty & \text{if } e_{ij} \notin G \end{cases}$$

Since in general a susceptible node may have multiple infectious neighbors, we model it as being infected at the first time it receives an infectious contact from any of its neighbors. The index case, or the first infected patient, is given an infection time of 0. Thus, letting I_i be the infection time of node i, we have the following recurrence:

$$I_i = \begin{cases} 0 & \text{if } i \text{ is the index case} \\ \min_{j:e_{ij} \in G}(I_j + t_{ji}) & \text{otherwise} \end{cases} \tag{1}$$

Finally, we define the process of viral genetic evolution over an epidemic. To prevent the creation of a large number of latent variables, we will not model the sequences directly, but instead model the number of mutations, i.e. differences between sequences.

We use the simple Jukes-Cantor (1969) model of evolution, which assumes equal base frequencies of the bases A, T, C and G, as well as equal mutation rates between any two of these bases. We use a Poisson process model which models genetic mutations as independent rare events, with the result that the number of mutations occurring during the transmission from i to j follows a Poisson distribution with parameter α. We refer to α as the genetic mutation rate of the disease. Since the amount of time passing in the intervening period between i and j's times of infection is $|I_j - I_i|$, thus letting D_{ij} be the genetic distance between i and j, we have:

$$D_{ij} \sim Poisson(\alpha |I_j - I_i|) \quad \text{if } e_{ij} \in H$$

The distance between two nodes which are not directly connected in H is the sum of distances D_{ij} along the shortest path between i and j in the graph H.

Figure 7 shows the output of a simulation run involving 15 nodes.

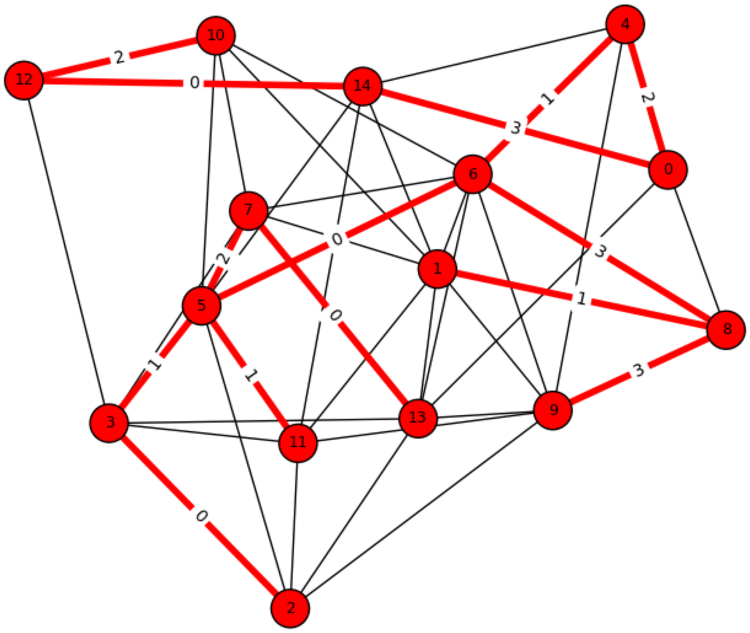


Figure 7: Output of a simulated contact network (edges in black), an epidemic run over this contact network (edges in red), and the numbers of mutations that occurred as the virus spread during the epidemic (edge labels). The first node affected by the epidemic is node 0 (upper-right). This simulation run has 15 nodes, $p = 0.4$, $\beta = 1, \alpha = 5.0$.

6.2 Dijkstra-based Simulation Algorithm

In this section, we describe a simple approach based on Dijkstra's algorithm used to speed up the process of simulating epidemics using the model described in the previous subsection. This is especially useful in simulating large training sets involving many such simulated epidemics to use in training a prediction algorithm, as we will need in Section 7.

Algorithm 1 Algorithm for simulating epidemics

```
• Input: a contact network $G$
• Output: a simulated epidemic along the contact network

1. For each edge $(i, j) \in G$, sample $t_{ij} \sim$ Exponential($\beta$)
2. Choose a random starting node $k$ to be the index case (i.e. the first infected patient)
3. Run Dijkstra's algorithm starting from $k$.
4. For each node $i$, the infection time of $i$ is the shortest path length from $k$ to $i$, and the node that infected $i$ is the second-last node in the shortest path from $k$ to $i$.
```

To show that this algorithm works, we note that the recurrence (1) defining infection times I_i is exactly the recurrence used by Dijkstra's algorithm to compute shortest paths from k to each node. Furthermore, the infector of node i is $\text{argmin}_{j:e_{ij}\in G}(I_j + t_{ji})$, which is the second-last node in the shortest path from k to i.

6.3 Markov chain Monte Carlo algorithm

We design a Markov chain Monte Carlo (MCMC) algorithm to infer the parameters of a random contact graph based on observed genetic and epidemiological data. The following summarizes the notion we use:

- $G = (V, E)$: random contact network
- H: subgraph consisting of the edges along which disease transmission occur
- I_j : infection time for patient j
- D_{ij} : genetic distance between patient i and j
- p: Erdos-Renyi edge generation probability
- α : disease genetic mutation rate
- β : disease transmission rate

The following Bayesian directed acyclic graph defines the relationships and conditional independencies between the variables in our model:

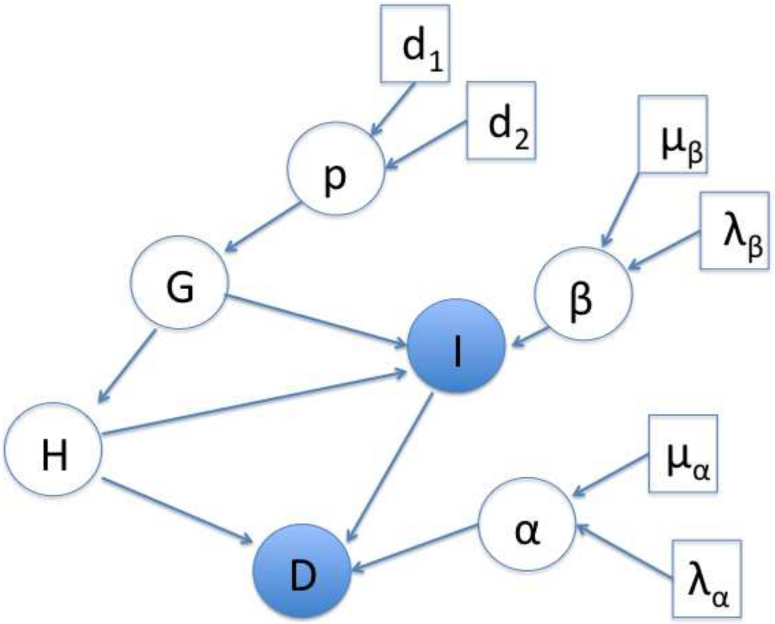


Figure 8: Bayesian DAG depicting variables in MCMC model. Circles are variables, squares are hyperparameters. Shaded circles represent observed data, while unshaded circles are unknown (latent variables). G is the contact network, H is the subgraph of epidemic edges, I the vector of infection times, and D the matrix of genetic distances. α is the genetic mutation rate and β is the epidemic transmission rate.

We use a Gibbs sampling scheme. To make the inference process more efficient, we use conjugate priors for our parameters p, β, α: since each edge of G is drawn randomly based on a $Bernoulli(p)$ distribution, we set a $Beta(d_1, d_2)$ prior on p. This allows us to straightforwardly draw p from the conditional distribution given G, by simply sampling p from a $Beta(d_1 + |G|, d_2 + \frac{n(n-1)}{2} - |G|)$ distribution, where $|G|$ is the number of edges in G. For β and α, which are rate parameters for an exponential and a Poisson distribution respectively, we place Gamma conjugate priors with hyperparameters as shown in the diagram. This will be described in more detail in the following section.

6.3.1 Gibbs sampling equations

The following equations give the conditional distributions we use to sample each variable given the others. Let A_i be the ancestor of i, i.e. the node that infected node i. Let L be the length of each patient's genetic sequence. Let G_{ij} be 1 if edge $(i, j) \in E$, and 0 otherwise.

Network, transmission and mutation parameters (p, β, α)

We resample p based on the usual update for a Bernoulli variable with a Beta conjugate prior based on observed data. The graph has $|G|$ edges present and $\binom{n}{2} - |G|$ edges absent, so the conditional distribution of p is a $Beta(d_1 + |G|, d_2 + \binom{n}{2} - |G|)$ distribution.

The update for β comes from the fact that $n - 1$ is the number of transmissions which actually occurred, and $\sum_{i,j \in G} |I_j - I_i|$ is the sum over edges of the time that elapsed during which one node was infected and the other was not. In other words, this is the time during which transmissions could have occurred; thus, the conditional distribution is $Gamma(\mu_\beta + n - 1, \lambda_\beta + \sum_{i,j \in G} |I_j - I_i|)$. Note that we derive similar Gibbs sampling updates more formally in the mixture models section (Section 10), which is a generalization of this case.

Analogously, the update for α comes from the fact that $\sum_{i,j \in H} D_{ij}$ is the sum of genetic distances along edges by which disease transmission occurred, and hence is the number of genetic mutations which actually occurred. $\sum_{i,j \in H} |I_j - I_i|$ is the total time during which genetic mutations could have occurred, so the conditional distribution is $Gamma(\mu_\alpha + \sum_{i,j \in H} D_{ij}, \lambda_\beta + \sum_{i,j \in H} |I_j - I_i|)$.

Graph edges (G_{ij})

We resample each edge G_{ij} independently as follows: if the edge is currently used as one of the epidemic transmission edges, then its conditional probability given the other variables is clearly 0. Otherwise, the probability that the edge is present depends on its prior (p) multiplied by the likelihood arising from the lack of a transmission that occurred along this edge, with transmission probability β and an elapsed time of $|I_j - I_i|$ in which a transmission could potentially have occured. As such, the probability of G_{ij} being present is proportional to $p \exp(-\beta |I_j - I_i|)$, while its probability of being absent is proportional to $1 - p$. We normalize these to get the exact conditional distribution for G_{ij}.

Ancestors (A_{ij})

We resample A_i (the ancestor of node i, that is, the patient that transmitted the disease to patient i) as follows: if node j has a later infection time than i, the probability that j is the ancestor of i is 0. Otherwise, since all nodes have the same transmission rate β, the only differences in likelihood between candidates for the ancestor of i come from the likelihood of the observed genetic sequences: intuitively, genetic sequences which are more similar to i's are more likely to come from an ancestor of i.

Thus, we will evaluate the likelihoods of each potential ancestor j of i based on their genetic sequences. Since genetic mutation follows a Poisson process, the number of genetic differences between i and j follows a Poisson distribution with mean $\alpha|I_j - I_i|$, so:

$$\begin{aligned} & P(\text{exactly } D_{ij} \text{ genetic differences between } i \text{ and } j|A_i = j) \\ = \; & \frac{(\alpha|I_j - I_i|)^{D_{ij}} \exp(-\alpha|I_j - I_i|)}{(D_{ij})!} \end{aligned} \tag{2}$$

However, note that this gives the combined likelihood of all possible genetic sequences with D_{ij} differences; in reality, we have observed a particular one of these. As such, we have to divide the above probability by the number of sequences having D_{ij} genetic differences from i's genetic sequence, which we will now compute.

Letting L be the length of the genetic sequence in base pairs, the D_{ij} mutations occur at a subset of these L bases, so the number of ways to choose the mutation positions is $\binom{L}{D_{ij}} = \frac{L^{D_{ij}}}{(D_{ij})!} + o(L^{D_{ij}})$, where $\frac{o(L^{D_{ij}})}{L^{D_{ij}}} \to 0$ as $L \to \infty$, and we ignore this term as we assume that L is large compared to D_{ij}. Note that in both the influenza case and in our simulations, $L \approx 3000$, since this is the approximate combined length in base pairs of the hemagglutinin and neuraminidase sequences for H1N1 influenza, whereas the genetic distances D_{ij} are generally at most 20. Once the mutation positions are chosen, each base pair that is different from the original sequence can be chosen from among 3 possibilities (the 4 base pairs other than the one in the original sequence). Thus the number of genetic sequences that differ from the original in exactly D_{ij} positions is close to $\frac{(3L)^{D_{ij}}}{(D_{ij})!}$. Combining this with equation (1) we get:

$$\begin{aligned} & P(\text{observed genetic sequence of } i|A_i = j) \\ = \; & \frac{(\alpha|I_j - I_i|)^{D_{ij}} \exp(-\alpha|I_j - I_i|)}{(D_{ij})!} \Big/ \frac{(3L)^{D_{ij}}}{(D_{ij})!} \\ = \; & (\frac{\alpha|I_j - I_i|}{3L})^{D_{ij}} \exp(-\alpha|I_j - I_i|) \end{aligned}$$

Finally, the probability that $A_i = j$ can be obtained from the above using Bayes rule:

$$P(A_i = j|\text{others}) = \frac{P(\text{observed genetic sequence of } i|A_i = j)}{\sum_k P(\text{observed genetic sequence of } i|A_i = k)}$$

Gibbs update equations

Finally, combining the discussion so far leads to the following system of update equations:

$$
\begin{aligned}
P(p|\text{others}) &= Beta(d_1 + |G|, d_2 + \binom{n}{2} - |G|) \\
P(\beta|\text{others}) &= Gamma(\mu_\beta + n - 1, \lambda_\beta + \sum_{i,j \in G} |I_j - I_i|) \\
P(\alpha|\text{others}) &= Gamma(\mu_\alpha + \sum_{i,j \in H} D_{ij}, \lambda_\beta + \sum_{i,j \in H} |I_j - I_i|) \\
P(G_{ij} = 1|\text{others}) &\propto \begin{cases} 1 & \text{if } A_i = j \text{ or } A_j = i \\ \frac{p\exp(-\beta|I_j - I_i|)}{(1-p)+p\exp(-\beta|I_j - I_i|)} & \text{otherwise} \end{cases} \\
P(A_i = j|\text{others}) &\propto \begin{cases} (\frac{\alpha|I_j - I_i|}{3L})^{D_{ij}} \exp(-\alpha|I_j - I_i|) & \text{if } I_j < I_i \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$

6.3.2 Rao-Blackwellized estimation

The simplest way to estimate p based on our MCMC samples would be to simply average the sampled values of p. Letting θ_i^1,θ_i^2,...,θ_i^N, be the Gibbs-sampled values for a particular parameter θ_i, this involves predicting the empirical mean:

$$\hat{\theta}_i = \sum_{t=1}^{n} \theta_i^t$$

The Rao-Blackwellized estimator decreases the predictive variance of p. Rather than predicting the empirical mean, we predict the conditional mean of θ_i given the current values of the rest of the variables:

$$\hat{\theta}_i = \sum_{t=1}^{n} E(\theta_i^{t+1}|\theta_{-i}^t)$$

By predicting the conditional mean, our prediction no longer incorporates the additional sampling variance arising from the random sampling of p given the other variables. This estimator has been shown to have lower asymptotic variance than the empirical mean (McKeague et al., 2000).

Furthermore, in many cases the Rao-Blackwellized estimator can be computed naturally as part of the Gibbs sampling process, and hence does not affect the computational complexity of the inference algorithm. As part of Gibbs sampling, we keep track of the posterior distribution of p given the other variables at the previous step; often, the mean of this posterior distribution is available in closed form. In our case, the posterior of p is a Beta distribution. As such, letting $|G_t|$ be the number of edges in the Gibbs-sampled contact network at time t:

$$\begin{aligned}\hat{p} = & \textstyle\sum_{t=1}^{n} E(p|\text{others}) \quad \text{where } p \sim Beta(d_1 + |G_t|, d_2 + \frac{n(n-1)}{2} - |G_t|) \\ = & \textstyle\sum_{t=1}^{n} \frac{d_1+|G_t|}{d_1+d_2+\frac{n(n-1)}{2}}\end{aligned}$$

6.3.3 Implementation

We implemented the MCMC inference framework using Python. This implementation, as well as accompanying code used to generate the analyses and figures in this paper, can be found on the author's website. The README file for the software package summarizes the functionality and inputs and outputs of the various programs in the package, and is included as an Appendix.

7 Evaluation

We evaluate our method using simulation - we simulate random contact networks from a random graph model with known parameter, a random epidemic using this contact network, and a genetic distance matrix arising from viral evolution along this epidemic, then test our algorithms on the simulated data.

The key metrics we are interested in are whether the disease parameters (mutation rate α and transmission rate β) have been inferred accurately, and whether the algorithm has accurately inferred who infected whom among our groups of patients. Simulation parameters were chosen to reflect the statistics of the real H1N1 dataset as closely as possible.

As for the real H1N1 data, since we do not have ground truth in terms of disease parameters or who infected whom, we cannot directly evaluate the model in the same way. We indirectly evaluate the model by using our learned model to infer epidemic parameters such as the basic reproduction number R_0. For H1N1 influenza, R_0 is fairly well measured in the literature, estimated to be between 1.7 and 1.8 for the early stage of the US epidemic (White et al., 2009), which is the period our data is drawn from. We can thus evaluate our method by using it to compute posterior predictions for R_0 and evaluating how well this agrees with the literature.

7.1 Who infected whom

For assessing the algorithm's inferences for who infected whom, we compare the algorithm to the `seqTrack` package of Jombart et al. (2011), which combines a parsimony-based approach with maximum likelihood to infer who infected whom from a set of patients and their genetic sequences. The following figure shows the result of comparing our MCMC algorithm to `seqTrack` over a range of mutation rate values.

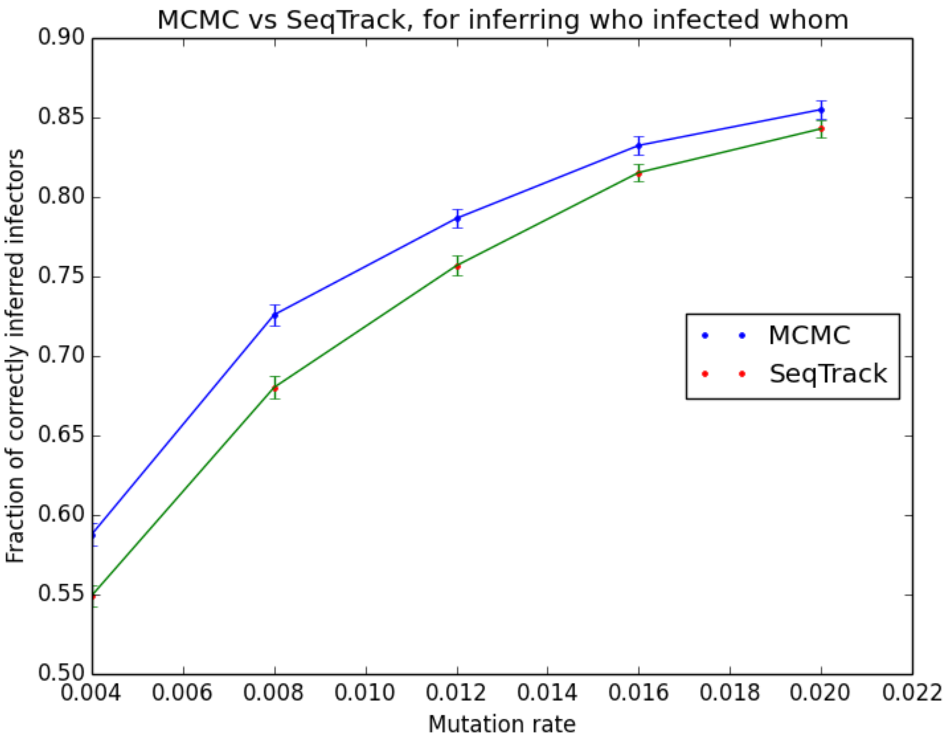


Figure 9: Comparison between MCMC and SeqTrack for inferring who infected whom. Each data point is an average of 1500 trials run with the corresponding mutation rate. The fraction of correctly inferred infectors is computed as follows: for each patient, we use MCMC (or SeqTrack) to infer their most likely infector (that is, the patient that infected this patient). We then plot the fraction of correct inferences made, averaging the results over 1500 trials.

We observe that MCMC performs significantly better than `seqTrack` over a range of mutation rates. Note that as the mutation rate increases, both algorithms do better because the average number of genetic mutations between any two patients tends to increase. Since patients tend to be genetically close to their true infector, a higher number of genetic mutations on average means that we have a higher signal-to-noise ratio with which to infer the most likely ancestor of each patient.

7.2 Disease parameters

As we have mentioned, a key goal of the MCMC approach was to jointly infer disease-related parameters, particularly the disease transmission rate and the genetic mutation rate. In contrast, `seqTrack` only allows us to predict who infected whom. The following figures shows the MCMC-inferred values of these parameters, plotted against the true values.

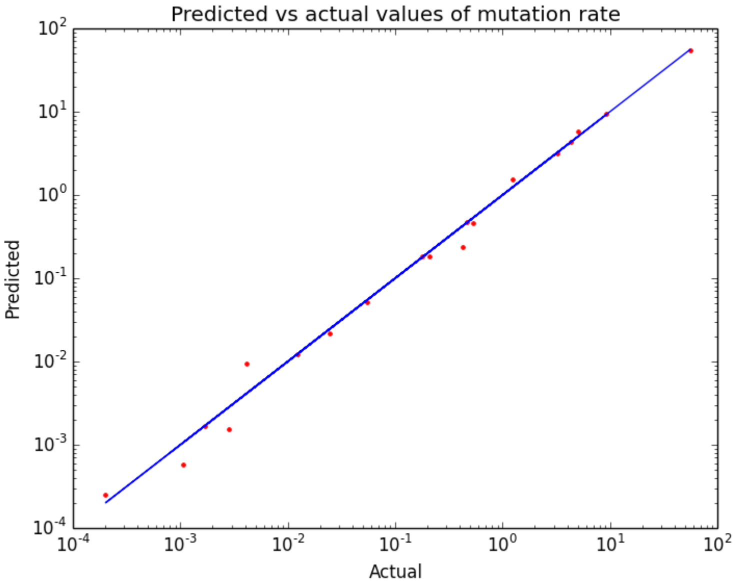


Figure 10: Here we use MCMC to infer the disease mutation rate parameter α. Each of the red points comes from a single simulated epidemic, with the true value of α plotted on the x-axis, and the value inferred by MCMC on the y-axis. The line $y = x$ is plotted in blue. The closer a point is to the line, the more accurate the inferred value.

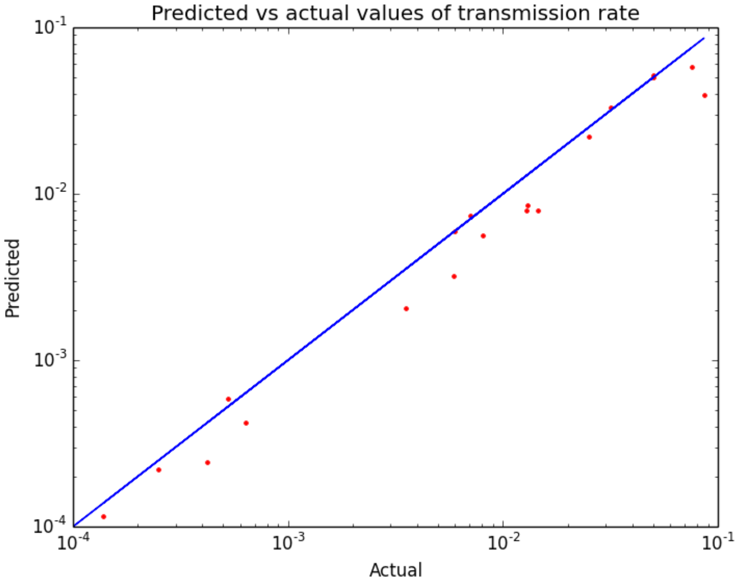


Figure 11: Disease transmission rate parameter β, as inferred by MCMC algorithm

We observe from both plots that the points are close to the main diagonal, showing that the parameters have been inferred accurately.

7.3 Contact network

While the MCMC algorithm is fairly accurate in inferring the disease parameters and who infected whom, it is fairly poor at inferring the Erdos-Renyi contact network parameter p, as we will see. Intuitively, the large sampling variance in MCMC when we resample the presence or absence of each of the n^2 possible edges (where n is the number of patients) results in high variance for p. Moreover, with the contact network being unobserved, the observed data provides a relatively weak signal for p because the edges which are not used as part of the epidemic have minimal influence on the observed data, so this weak signal is swamped by the noise arising from MCMC sampling.

To fix this problem, our idea is therefore to summarize the observed data into a relatively small number of features which predict p more effectively. We do this by hand-engineering features, and using prediction algorithms, namely Random Forests and Gradient Boosting Machines (GBMs), to predict the contact network parameter p from these features. Assuming we can predict p accurately, we can use this value of p as a fixed constant in MCMC inference, circumventing the poor performance of MCMC in estimating p and thereby making the rest of the MCMC inference more accurate as well.

The features and prediction algorithms will be described in more detail in the rest of this section.

To evaluate our approaches in an unbiased manner, we compute the test error of the Random Forests and Gradient Boosting Machines by using them to predict p on a test set of simulated epidemics that is independent from the training set on which they were trained. The following table shows the mean squared error (MSE) and mean absolute error (MAE) of predicting p on the test simulations, compared with that of MCMC.

Model	**MSE**	**MAE**
Random Forest	0.011	0.074
Gradient Boosting	0.010	0.071
MCMC	0.053	0.18

Table 1: Mean squared error (MSE) and mean absolute error (MAE) of predicting p using random forests, gradient boosting machines, and Markov Chain Monte Carlo

As the table, as well as Figure 12 indicates, the prediction error of Random Forests and Gradient Boosting Machines are much lower than that of MCMC.

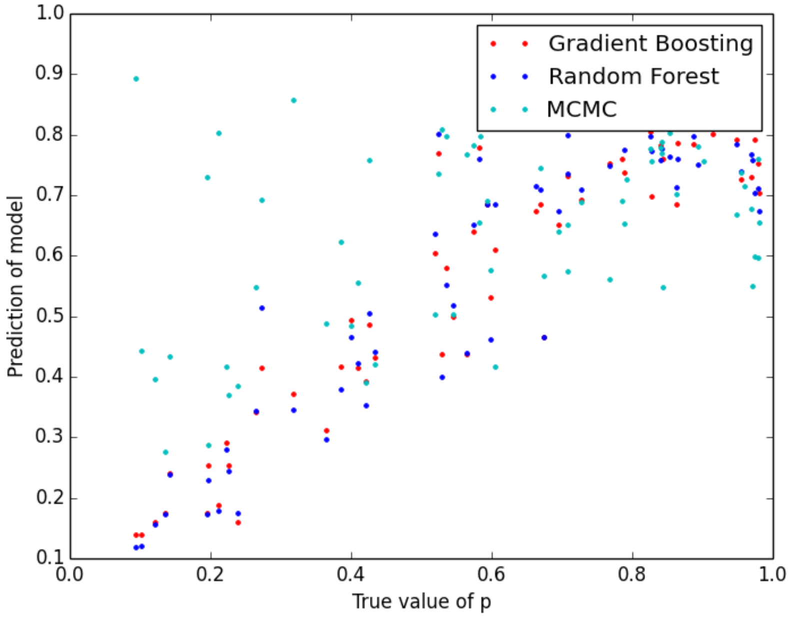


Figure 12: Predictions of random forest and gradient boosting machine models against true values when inferring the unknown parameter p in $G(n, p)$ random contact networks. Each point represents one simulation of 100 patients.

7.3.1 Feature Engineering

We use the following features of the observed data to predict the unknown parameter p of Erdos-Renyi graphs.

- F1: Latest infection time among all patients
- F2: Sum of infection times
- F3: Largest number of genetic differences between any two sequences
- F4: Sum of genetic differences between each two sequences

These features are used as input to two prediction models, Random Forests (RFs) and Gradient Boosting Machines (GBMs) - chosen for their ability to fit nonlinear functions as well as interactions between predictors. The prediction models were implemented and trained using Python's scikit-learn library (Pedregosa et al., 2011). We tuned the RF and GBM model parameters slightly to ensure that the number of estimators in each case was sufficient - we used 500 estimators for the Random Forest model and 300 for the Gradient Boosting model.

We first trained the prediction models, random forests and gradient boosting machines, on 400 independent simulations (with different values of p), and tested on 80 more simulations, with the results plotted in Figure 12.

7.3.2 Variable Importance

Both the Random Forests and Gradient Boosting models have standard methods for evaluating the importance of each feature. In both cases, this works by measuring the importance of a feature based on the average decrease in squared error when we split the feature space according to that variable as part of the training process. A more detailed description of variable importance can be found in Hastie et al. (2005).

Model	**F1**	**F2**	**F3**	**F4**
Random Forest	0.12	0.27	0.28	0.32
Gradient Boosting	0.11	0.25	0.32	0.32

Table 2: Variable importances based on the two models for features. F1 and F2 are the infection time based features, while F3 and F4 are the genetic distance based features.

The max genetic distance between any two genetic sequences (F4) is the most important feature, while the sum of genetic distances (F3) is not far behind. The genetic distance-based features seem to be more predictive than the infection time-based features. This could be because the genetic distance-based features are able to capture the fact that two patients who were infected at around the same time may not be closely related (e.g. if the branches leading from the index case to these two patients separated very early). The infection time-based features, however, would only see that these two patients were infected at around the same time, and thus treat them as similar.

7.3.3 Summary

In summary, we showed that MCMC predicts who infected whom more accurately than similar methods, in particular `seqTrack`., and additionally infers the disease mutation rate and transmission rate accurately. It is, however, poor at predicting the contact network parameter p; we showed that this can be mitigated by using an alternate feature engineering approach to predict p, which we can then use as a fixed constant within the MCMC process.

In the following section, we apply the MCMC framework to the H1N1 dataset.

8 H1N1 Influenza Analysis

The following figure plots the inferred disease transmissions from the H1N1 dataset, obtained using MCMC inference:

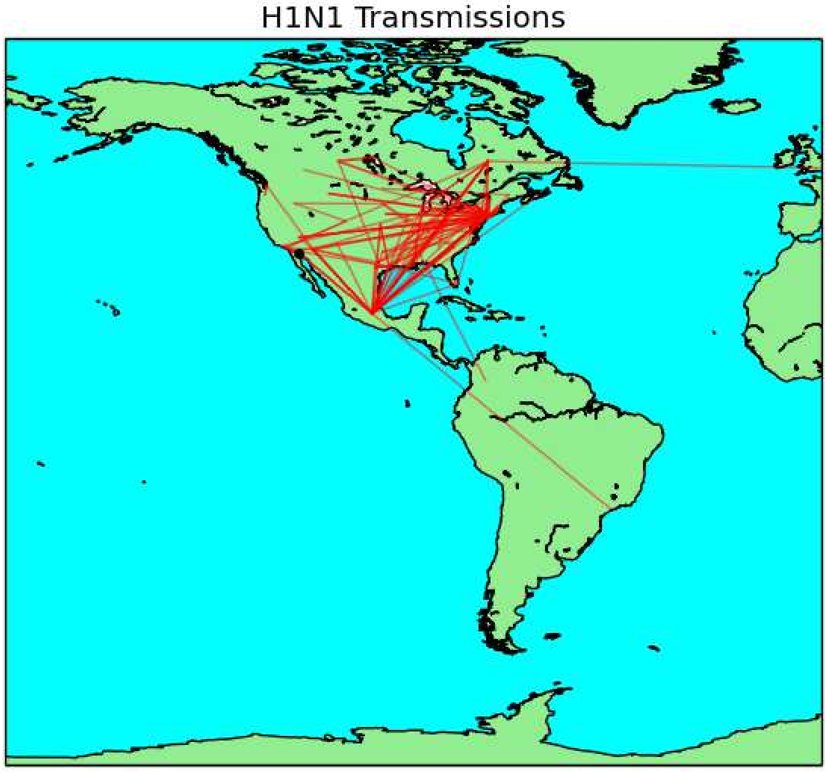


Figure 13: Transmissions inferred by the Markov chain Monte Carlo algorithm. The index case is shown as a black circle. Following the red lines, we see that the disease was transmitted from the index case first to Mexico City, then to a larger outbreak in New York.

The most likely ancestors for each patient define a series of inferred transmissions. They suggest that the disease started out in El Centro, California, then transmitted toward Mexico City, in which a sizable outbreak seems to have occurred. From there, it spread to a number of locations across the United States (as well as Brazil). We also observe a second large outbreak in New York, during which the disease transmitted heavily within New York and also to other locations within the United States.

These findings agree broadly, but not perfectly, with accounts in the literature of the H1N1 viral progression. The disease is believed to have first spread in Mexico, but since many patients from Mexico are not in our dataset, our algorithm instead identifies a patient in El Centro, CA as the first case. The Centers for Disease

Control and Prevention (CDC; 2009) report that southern California was one of the first places in the U.S. affected by the outbreak. Lessler (2009) describes a later outbreak in a high school in Queens, New York, identifying a linkage with travel of several of the patients to Mexico, which agrees with our findings.

8.1 Estimation of Basic Reproductive Number R_0

Based on the output of the MCMC algorithm, we now show how to estimate the basic reproductive number R_0 of the epidemic, which is the number of secondary infections that result from each infection in a completely susceptible population. R_0 is of important interest as it defines a threshold determining whether the infection will spread in a population, and is therefore important to estimating the risk of an epidemic, as well as for disease elimination and vaccination programs to determine the level of control needed to eliminate the disease.

Wallinga (2007) suggest estimating R_0 based on the early stage of the epidemic, based on the idea that epidemic transmission in this phase approximates that of a completely susceptible population. Following this approach, we estimate R_0 as the average number of patients infected by one of the patients who was infected in the early phase of the epidemic, where the 'early epidemic' is defined by a variable threshold. That is, letting I_j be the infection time of patient j and n_j be the number of patients infected by patient j, we define the set of patients infected before time t :

$$S(t) = \{j : I_j < t\}$$

Then our estimate for R_0 is:

$$\hat{R_0}(t) = \frac{\sum_{j \in S(t)} n_j}{|S(t)|}$$

Since there is no fixed method for choosing the threshold defining the early epidemic, we plot a curve displaying the estimated R_0 for each possible threshold.

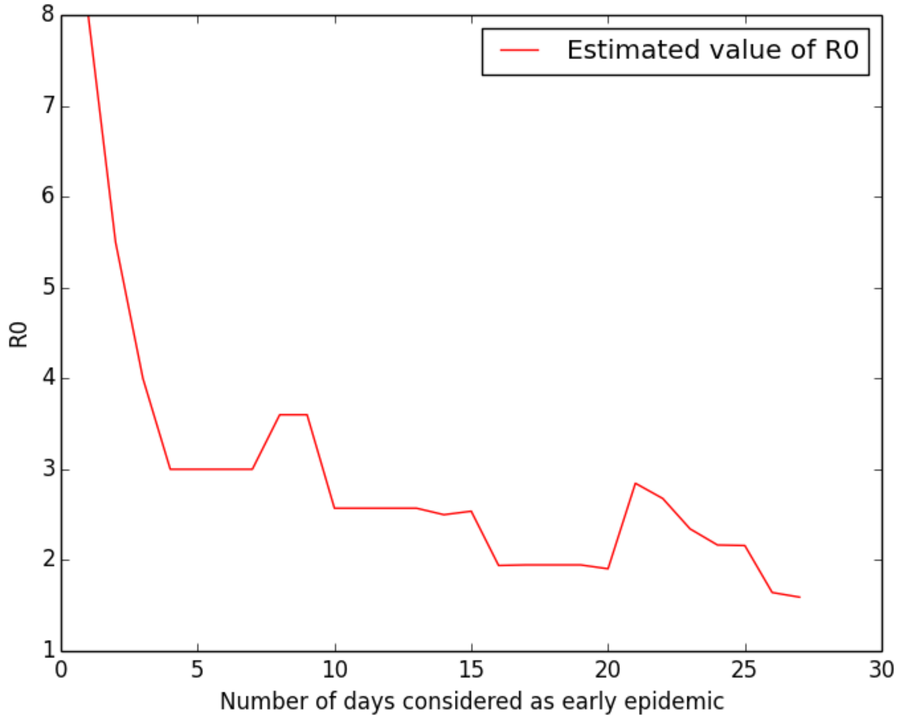


Figure 14: Estimated value of basic reproductive number R_0 against the choice of threshold defining the early epidemic.

The figure shows that the estimated values generally range from 1.6 to 3.0, which agrees with the values in the literature, but has large variance. Here we use only a single epidemic, which constitutes a small sample size with highly dependent patients. In contrast, the 1.7 to 1.8 value in the literature is based on multiple separate outbreaks. We expect that extending our method to larger datasets might be a promising approach to estimating R_0 based on genetic data, without requiring time-consuming contact tracing to determine the secondary infections of each patient.

In the next section, we extend the MCMC framework by considering several possible choices of ways for measuring the spatial distance between two locations: geographic distance, travel time and route length distance.

9 Comparing Spatial Distance Metrics

Spatial distance metrics are an important part of infectious disease epidemiology, and multiple studies have shown spatial clustering in the dynamics of infectious diseases (Ruiz-Moreno, 2010). However, it is not known whether disease transmissions are most consistent with the usual geographic distance 'as the crow flies', or the amount of time or distance taken to travel over land from one location to another, which can be obtained by querying Google Maps API. Intuitively, the former would be more consistent with transmissions occurring mainly as a result of air travel, and the latter with travel over land.

In this section, we use the H1N1 genetic sequences described previously to compare 3 spatial distance metrics: geographical distance, travel time, and route length distance, to determine which is most informative for disease transmission, and hence, which we would use in an inference algorithm such as those in the next section, where we take spatial distance metrics into account.

Specifically, we expect shorter distances along each metric to correlate with higher likelihood of disease passing between two patients; thus, we want to find which distance metric is most informative about transmission. We compute geographical distance from patients' latitude and longitude coordinates, which were annotated by Jombart et al. (2011). Travel time and route length refer to the time and distance taken along the shortest path returned when querying Google Maps API for a driving route from one point to the other. Since Google Maps considers distances along roads, we subsetted the dataset to the 285 patients living in the United States and Mexico, for which valid distances could be obtained.

9.1 Comparing Likelihoods of Distance Metrics

We first use Mantel tests to check for significant correlations between each of the three spatial distance metrics (geographic distance, route length and travel time distance), and genetic distance. The results show significant correlations in all three cases ($p = 0.001$).

Our next goal is to determine which among the 3 geographical distance metrics is most strongly informative of disease transmission.

Our approach is to first compute a likelihood for each distance metric, assuming the probability density of disease transmission across a distance of d decreases with d according to an exponential distribution with parameter λ, i.e. $e^{-\lambda d}$. In this way we can evaluate the likelihood of any particular sequence of transmissions. However, since we do not know the true transmission history of the disease (i.e. which patient transmitted to which patient), we estimate the transmission history based on genetic data: we use a maximum likelihood based approach based only on genetic data and temporal data to estimate the transmission history. We then compute the likelihood of this transmission history using each of the spatial distance metrics. Since λ is unknown, we use the maximum likelihood value of λ under each spatial distance metric.

For each estimated transmission, say from patient A to patient B, let I_{AB} be the event in which patient A infects patient B. Since we know that exactly one patient infected B, the likelihood under distance metric D is:

$$P(I_{AB}|D,\lambda) = \frac{e^{-\lambda D(A,B)}}{\sum_{C \in Patients-B} e^{-\lambda D(C,B)}}$$

By multiplying together such likelihood terms over each such pair of patients in which one infected the other, we obtain the combined likelihood of D and λ. We then maximize the combined likelihood over λ by performing a grid search over λ, inspecting plots of likelihood against λ to ensure that the range and interval size is appropriate. This results in a likelihood for each distance metric.

Next, for each pair of spatial distance metrics, we obtain a likelihood ratio by comparing the likelihoods computed on the original dataset. We estimate the sampling variance of these likelihood ratios by the bootstrap (Efron, 1993).

Note that in the following, we choose to compute and display likelihood ratios (rather than likelihoods themselves) because of our use of pairing: for each bootstrap sample, we evaluate its likelihood with respective to each of the distance metrics (say, duration and geographic distances); we then take the ratio of these likelihoods. Note that the likelihoods evaluated on a particular bootstrap replicate are related; hence using the pairing approach reduces variance and increases statistical power as is the usual case for paired tests. As a result, the bootstrapping process allows us to estimate sampling variances for the likelihood ratios, but not the likelihoods themselves.

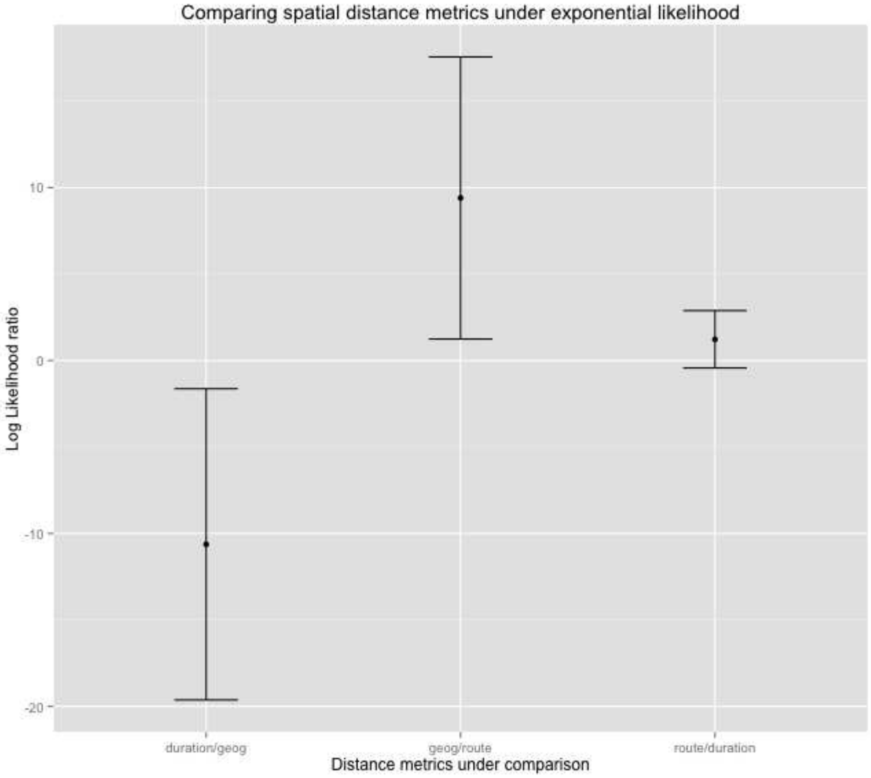


Figure 15: Log likelihood ratios, comparing two spatial distance metrics at a time. For example, the first column shows a significant negative log likelihood ratio, meaning geographic distance has significantly higher likelihood than duration time. Confidence intervals are bootstrap 95% confidence intervals (with Bonferroni correction).

Figure 15 shows that geographic distance has significantly higher likelihood than both of the other two metrics (hence the significant negative likelihood ratio for duration/geographic, and the significant positive ratio for geographic/route). In contrast, route and duration differ in likelihood to a much smaller (and not statistically significant) degree.

We also perform the same analysis using a power law density instead, where the probability density of transmission over distance x is proportional to $x^{-\alpha}$:

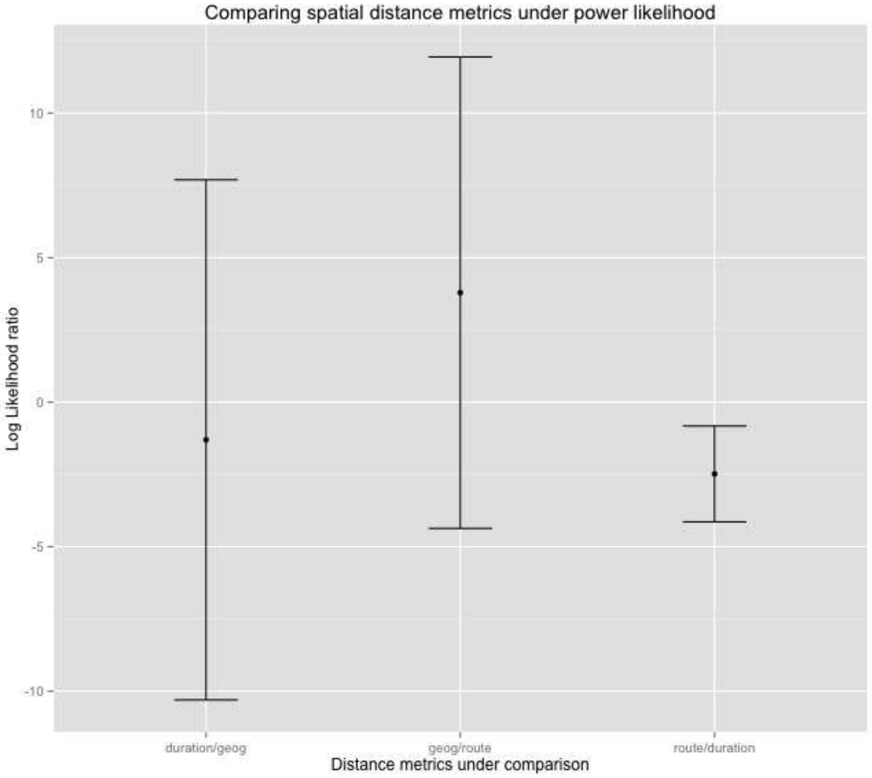


Figure 16: Log likelihood ratios, comparing two distances at a time, under power law ($x^{-\alpha}$) transmission density. Confidence intervals are bootstrap 95% intervals (with Bonferroni correction).

Here route distance has a significantly lower likelihood than duration distance, but the other comparisons are not statistically significant.

In summary, in the exponential case, geographic distance has significantly higher likelihood than the other two distances, and also seems to perform at least as well in the power law case. The better performance of geographic distance is consistent with the fact that air travel - not just land travel - is an important mode of transmission.

10 Mixture Models for Disease Transmission

Empirical studies of influenza such as Brownstein (2006) attest to the important influence of international air travel on influenza spread, and mathematical models of influenza such as those by Chong (2012) have recommended air travel restrictions as an important control measure for influenza pandemics.

In this section we extend the disease transmission model in our MCMC inference framework to allow for a mixture of several different transmissions routes. The program takes as input a directed or undirected graph (for example, an air traffic network). It allows for weighted graphs, since the number of passengers travelling along a particular air traffic route is of key importance in modeling disease spread.

The flight network data comes from the Bureau of Transportation Statistics domestic flight database, which contains the total number of passengers travelling between any two cities within the year of 2009. Note that this does not consider each segment of a passenger's flight as a separate journey; instead, the journey is considered as a whole, from its initial start city to its final ending city.

10.1 Model

We modify our earlier model from Section 6 as follows: rather than a single contact network G, we now have K graphs, which are the contact networks corresponding to the transmission routes. In both simulations and real data, $K = 2$, representing air and land based transmissions.

Denote these contact network graphs as $G_k = (V_k, E_k)$, where $k = 1, 2, \ldots, K$. These graphs are allowed to be directed and weighted. Higher weight indicates higher infectiousness: for example, in the air traffic network, an edge (i, j) represents the presence of a flight route from patient i to patient j's location, and the weight of this edge is the number of passengers travelling along this flight route.

To control the overall level of infectiousness of each transmission route, define infectiousness parameters $\lambda_1, \ldots, \lambda_K$. The transmission model is then as follows: the time taken for the disease to transmit along an edge with weight w with corresponding infectiousness parameter λ_k follows an exponential distribution with rate parameter $w\lambda_k$. As before, an uninfected node with multiple edges joining it with infected nodes is infected the first time it receives the disease transmission along any of these edges.

These parameters λ_k are unobserved and will be inferred as part of MCMC inference, along with the mutation rate α, and the unknown vector A of ancestors, where A_i indicates the ancestor (or infector) of patient i.

10.2 MCMC Inference

The Gibbs sampling update for the mutation rate α, which samples from the conditional distribution $P(\alpha|\text{ others})$, is unchanged from the non-mixture case described in Section 6. This is because we have changed the disease transmission model but not the mutation model; moreover, the inference process for the mutation rate α depends only on the ancestor vector A, and the observed data. Since Gibbs sampling involves conditioning on A, we conclude that inference from α is unchanged.

Since we have modified the transmission model, we need to compute the Gibbs sampling equations for the infectiousness parameters λ_k, and the ancestors vector A.

Infectiousness Parameters λ_k

Let $w_{ij}^{(k)}$ represent the weight of the edge (i, j) in graph G_k. Introduce auxiliary variables $z_{ij} \in \{1, \ldots, K\}$ for each i, j such that $(i, j) \in E = \bigcup_{k=1}^{K} E_k$. Intuitively, z_{ij} is a 'switch variable' that controls which of the K mixture components caused an infection along the edge from node i to node j. We introduce these variables because directly computing the conditional distribution $P(\lambda_1, \ldots, \lambda_K|\text{others})$ leads to conditional distributions for λ which cannot be expressed in a simple form, as well as leading to dependence between the λ_k. Therefore, we introduce z_{ij} into the Gibbs sampling process, which involves first sampling from the conditional distribution $P(\lambda_1, \ldots, \lambda_K|z, \text{ others})$, followed by sampling from $P(z|\lambda, \text{ others})$. As we will now show, the conditional distribution $P(\lambda_1, \ldots, \lambda_K|z, \text{ others})$ factorizes over the λ_k, and can be computed efficiently and in a simple form when we use conjugate priors.

We start by assigning uniform priors to z_{ij}:

$$P(z_{ij} = k) = \frac{1}{K} \text{ for all } (i, j) \in E$$

To λ_k we assign a $Gamma(\alpha_k, \beta_k)$ prior:

$$P(\lambda_k) = \frac{\beta_k^{\alpha_k}}{\Gamma(\alpha_k)} \lambda_k^{\alpha_k - 1} e^{-\beta_k \lambda_k} \propto \lambda_k^{\alpha_k - 1} e^{-\beta_k \lambda_k}$$

The conditional distribution for λ_k can be obtained by multiplying the prior with the likelihood of the observed series of disease transmissions. Each of the numbered equations are explained in the text below:

$$\begin{align}
P(\lambda_1, \ldots, \lambda_K| \text{ others}) &\propto \prod_{k=1}^{K} (\lambda_k^{\alpha_k - 1} e^{-\beta_k \lambda_k} \tag{3} \\
&\quad \times \prod_{i,j:z_{ij}=k} \lambda_k w_{ij}^{(k)} e^{-\lambda_k w_{ij}^{(k)} |t_i - t_j|} \tag{4} \\
&\quad \times \prod_{\substack{i,j \in E_k \setminus H \\ \text{or } i,j \in H, z_{ij} \neq k}} (e^{-\lambda_k w_{ij}^{(k)} |t_i - t_j|})) \tag{5} \\
&\propto \prod_{k=1}^{K} \lambda_k^{\alpha_k - 1} e^{-\beta_k \lambda_k} \prod_{(i,j) \in E_k} e^{-\lambda_k w_{ij}^{(k)} |t_i - t_j|} \lambda_k^{n_k} \tag{6} \\
&\quad \text{where } n_k = |\{i, j : z_{ij} = k\}| \nonumber
\end{align}$$

In the above, Equation 3 comes from the priors set on λ_k.

Equation 4 comes from the fact that for each i, j such that $z_{ij} = k$, we know that i transmitted the disease to j along transmission route k. Therefore, we multiply in the density for an $Exponential(\lambda_k w_{ij}^{(k)})$ distribution evaluated at $|t_i - t_j|$, since $\lambda_k w_{ij}^{(k)}$ is the rate parameter of the exponential distribution of the time taken for this transmission to occur, and $|t_i - t_j|$ is the actual time taken.

Equation 5 corresponds to edges for which either a transmission did not occur $(i, j \in E_k \backslash H)$, and edges where a transmission occurred but was from a different mixture component $(i, j \in H, z_{ij} \neq k)$. For each of these, mixture component k did not transmit the disease, so we need to multiply in a likelihood term for the probability that the corresponding $Exponential(\lambda_k w_{ij}^{(k)})$ variable had a value of greater than $|t_i - t_j|$, which has probability $e^{-\lambda_k w_{ij}^{(k)} |t_i - t_j|}$.

Note that Equation 6 implies that the conditional probability $P(\lambda_1, \ldots, \lambda_K |$ others$)$ factorizes as a product:

$$P(\lambda_1, \ldots, \lambda_K | \text{ others}) = \prod_{k=1}^{K} P(\lambda_k | \text{ others})$$

Which implies that $\lambda_1, \ldots, \lambda_K$ are independent in the conditional distribution. Moreover, from Equation 6 we have:

$$\begin{aligned} P(\lambda_k | \text{ others}) &= \lambda_k^{\alpha_k - 1} e^{-\beta_k \lambda_k} \prod_{(i,j) \in E_k} e^{-\lambda_k w_{ij}^{(k)} |t_i - t_j|} \lambda_k^{n_k} \\ &= \lambda_k^{\alpha_k - 1 + n_k} e^{-\lambda_k (\beta_k + \sum_{(i,j) \in E_k} w_{ij}^{(k)} |t_i - t_j|)} \end{aligned}$$

Thus the conditional distribution involves a simple conjugate prior update:

$$\lambda_k | \text{ others} \sim Gamma(\alpha_k - 1 + n_k, \beta_k + \sum_{(i,j) \in E_k} w_{ij}^{(k)} |t_i - t_j|)$$

Next, we show how to derive the conditional distribution for our auxiliary variables z_{ij} given the rest of the variables. We use the fact that given independent exponentially distributed random variables $X_1, X_2, \ldots, X_n$ with rate parameters $\lambda_1, \lambda_2, \ldots, \lambda_n$, the distribution of the index of the variable which achieves the minimum is distributed according to:

$$P(k = \text{argmin}_i(X_i)) = \frac{\lambda_k}{\lambda_1 + \cdots + \lambda_n}$$

We apply this fact, noting that $\{z_{ij} = k\}$ is the event in which the kth mixture component was the first to transmit the disease along edge (i, j). Since the time taken for the kth mixture component to transmit the disease according this edge follows an exponential distribution with mean $\lambda_k w_{ij}^{(k)}$, we get that:

$$P(z_{ij} = k | \text{ others}) = \frac{\lambda_k w_{ij}^{(k)}}{\sum_{k'=1}^{K} \lambda_{k'} w_{ij}^{(k')}}$$

Ancestors vector A

Here, we can use the fact that the minimum of independent exponentially distributed random variables has itself an exponential distribution, with rate parameter equal to the sum of the rate parameters of the original variables. In this way, all the edges between any two nodes i and j effectively collapse into a single edge, with an associated exponential distribution.

Note that when sampling A as part of Gibbs sampling, we condition on the infectiousness parameters λ_k, so the combined exponential distributions governing transmission along each edge are all known. As a result, performing inference on A is now the same problem as it was in the non-mixture case in Section 6, resulting in the same update equation as before.

10.3 Results

The following figures show that as in our original MCMC inference, we are able to accurately infer the disease mutation rate, as well as both transmission rates, as shown by most of the points lying close to the diagonal.

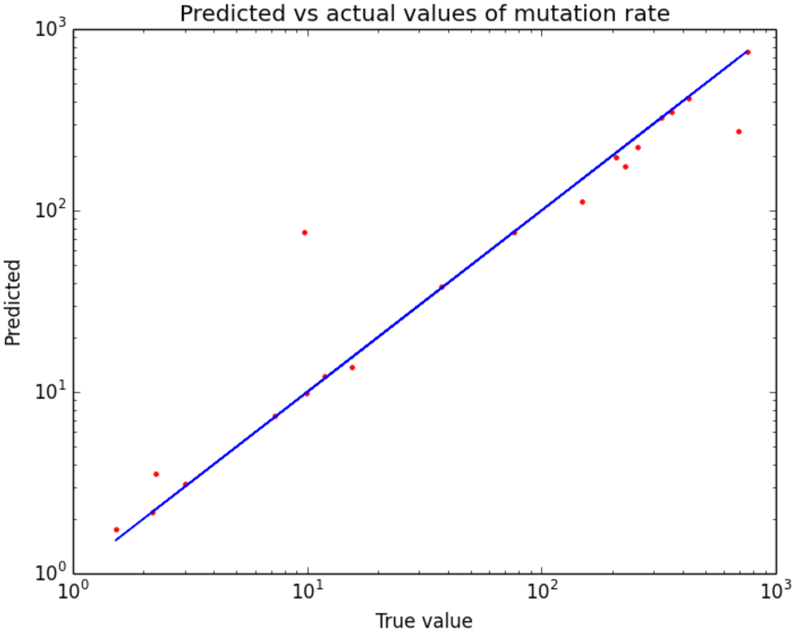
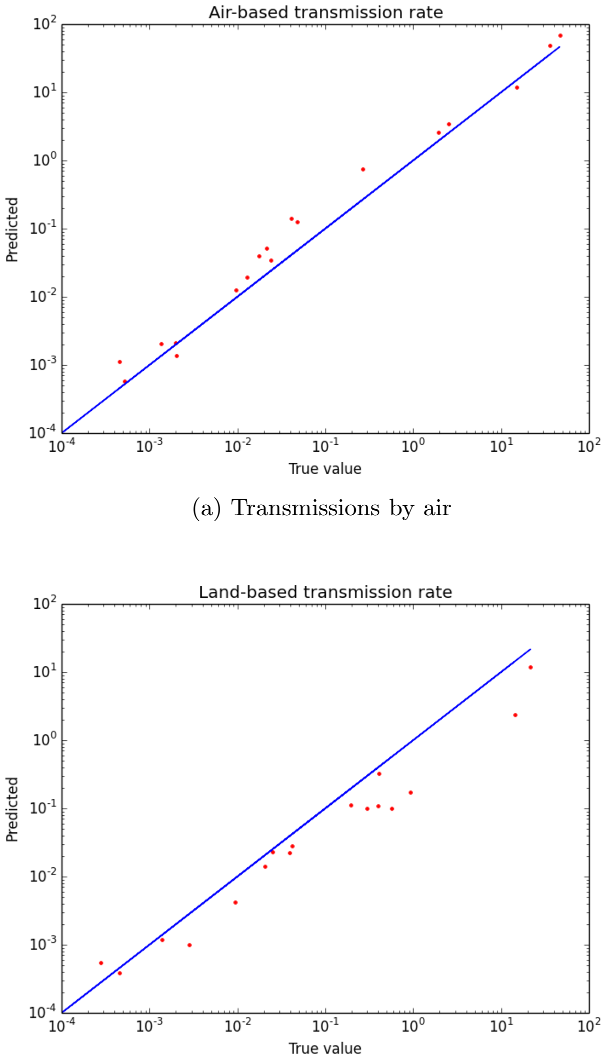


Figure 17: MCMC inference for mutation rate in simulations. Each point represents a simulation, which is plotted according to the true and predicted values of the mutation rate.



(a) Transmissions by air

(b) Transmissions by land

Figure 18: MCMC inference for land and air-based transmission rate in simulations

As such, our simulations suggest that even with mixtures of transmission routes, the algorithm is able to recover each of the transmission rates fairly accurately.

The following shows the result of applying MCMC with mixture models for transmission to our original H1N1 dataset:

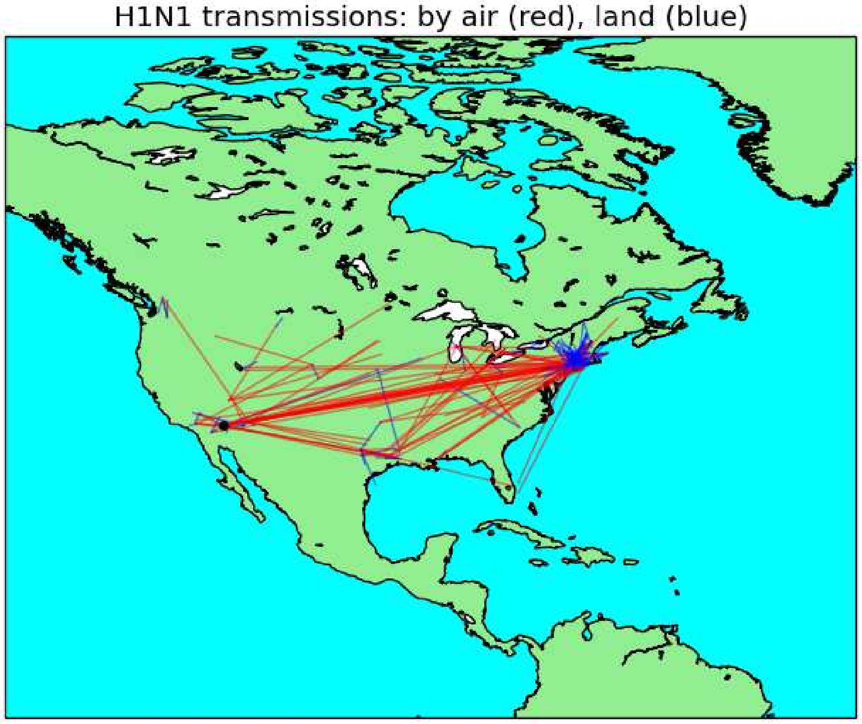


Figure 19: Inferred disease transmissions, colored by route: land-based transmissions are blue, air-based transmissions are red.

The above figure shows a large number of long-distance transmissions into and out of New York City inferred as air-based transmissions. In accordance with intuition, these are generally relatively busy flight routes. The large outbreak in New York is primarily inferred as spreading over land, which is consistent with its highly localized nature. On the whole, the conclusion of the importance of air travel in spreading the disease agrees with analysis of the outbreak by Khan (2009), who finds clear correlations between air travel patterns and H1N1 infection.

Appendix A: README of MCMC Software Package

The README file for the software package summarizes the functionality and inputs and outputs of the program.

```
Infectious Disease Inference using Genetic Data
```

```
==============================================================================

Overview
------------------------------------------------------------------------------
This software package implements Markov Chain Monte Carlo framework for using
genetic, spatial and temporal data to infer various disease-related parameters,
contact network parameters, and who infected whom among a group of patients.


Data Processing
(data.py)
------------------------------------------------------------------------------
Contains functionality to load in several data files. The data files used below
are all included in this package.

- Patient data table, containing locations, infection time and (optional) city,
  latitude and longitude for each of a group of patients.

- Genetic data file, which contains the viral genetic sequences for each
  patient, in the form of .fasta files.

Optional additional data files:

- Additional spatial distance matrices: route length and travel time distance
  matrices are included, computed using Google Maps API.

- For using tramission mixture models: a table containing number of passengers
  travelling between any two U.S. airports along domestic flights for the year
  2009.


MCMC Inference
(mcmc.py)
------------------------------------------------------------------------------
- The runGibbs function implements the main MCMC inference, taking in a set of
  observed data, number of iterations to run and burn in time. It returns a
  tuple containing the inferred values of the contact network parameter p, the
  disease transmission rate beta, the disease mutation rate alpha, and the
  index of the inferred ancestor of each patient (-1 indicating the inferred
  index case)

Example usage:
```

```

times = [1, 2, 3] # disease detection times
geneDist = [
[0, 1, 2],
[1, 2, 3],
[2, 3, 4] ] # genetic distance matrix
runGibbs((times, geneDist), 70, 20)

- Functions to plot the inferred transmissions from MCMC inference on a world
  map

- Infer the basic reproductive number R0 based on the results from MCMC
  inference


Simulation functions
(sim.py)
------------------------------------------------------------------------------
Contains functionality to:

- Simulate sample epidemics

- Simulate sample genetic distance matrices

- Test the MCMC inference algorithm on a set of simulated epidemics and output
  metrics and plots comparing its inferences with the true values


Mixture models
(mixture.py)
------------------------------------------------------------------------------
- Implements MCMC inference with mixture models for disease transmission. In
  addition to the parameters taken by the original MCMC algorithm, this takes a
  list of weighted directed graphs, along which disease transmissions are
  assumed to occur. Returns most of the same disease related parameters as the
  original MCMC inference algorithm, with a list of transmission parameters
  corresponding to each of the contact networks.
```

References

[1] Swine Influenza A (H1N1) infection in two children — Southern California, March–April 2009.

http://www.cdc.gov/mmwr/preview/mmwrhtml/mm5815a5.htm. Accessed: 2014-05-14.

[2] RM Anderson and RM May. Infectious disease of humans. *Dynamics and control*, 1991.

[3] Dennis A Benson, Ilene Karsch-Mizrachi, David J Lipman, James Ostell, and Eric W Sayers. Genbank. *Nucleic acids research*, 38(suppl 1):D46–D51, 2010.

[4] Tom Britton and Philip D. O'neill. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29(3):375–390, 2002.

[5] John S Brownstein, Cecily J Wolfe, and Kenneth D Mandl. Empirical evidence for the effect of airline travel on inter-regional influenza spread in the united states. *PLoS medicine*, 3(10):e401, 2006.

[6] Ka Chun Chong and Benny Chung Ying Zee. Modeling the impact of air, sea, and land travel restrictions supplemented by other interventions on the emergence of a new influenza pandemic virus. *BMC infectious diseases*, 12(1):309, 2012.

[7] Leon Danon, Ashley P. Ford, Thomas House, Chris P. Jewell, Matt J. Keeling, Gareth O. Roberts, Joshua V. Ross, and Matthew C. Vernon. Networks and the epidemiology of infectious disease. *Interdisciplinary Perspectives on Infectious Diseases*, 2011, March 2011.

[8] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*, volume 57. CRC press, 1994.

[9] Bryan T Grenfell, Oliver G Pybus, Julia R Gog, James LN Wood, Janet M Daly, Jenny A Mumford, and Edward C Holmes. Unifying the epidemiological and evolutionary dynamics of pathogens. *Science*, 303(5656):327–332, 2004.

[10] Chris Groendyke, David Welch, and David R Hunter. Bayesian inference for contact networks given epidemic data. *Scandinavian Journal of Statistics*, 38(3):600–616, 2011.

[11] Chris Groendyke, David Welch, and David R Hunter. A network-based analysis of the 1861 hagelloch measles data. *Biometrics*, 68(3):755–765, 2012.

[12] T. Jombart, R. M. Eggo, P. J. Dodd, and F. Balloux. Reconstructing disease outbreaks from genetic data: a graph approach. *Heredity*, 106(2):383–390, February 2011.

[13] Leonard Kaufman and Peter Rousseeuw. *Clustering by means of medoids.* North-Holland, 1987.

[14] William O Kermack and Anderson G McKendrick. Contributions to the mathematical theory of epidemics. ii. the problem of endemicity. *Proceedings of the Royal society of London. Series A*, 138(834):55–83, 1932.

[15] Kamran Khan, Julien Arino, Wei Hu, Paulo Raposo, Jennifer Sears, Felipe Calderon, Christine Heidebrecht, Michael Macdonald, Jessica Liauw, Angie Chan, et al. Spread of a novel influenza a (h1n1) virus via global airline transportation. *New England journal of medicine*, 361(2):212–214, 2009.

[16] Justin Lessler, Nicholas G Reich, and Derek AT Cummings. Outbreak of 2009 pandemic influenza a (h1n1) at a new york city school. *New England Journal of Medicine*, 361(27):2628–2636, 2009.

[17] Martin Maechler, Peter Rousseeuw, Anja Struyf, Mia Hubert, and Kurt Hornik. *cluster: Cluster Analysis Basics and Extensions*, 2014. R package version 1.15.1 — For new features, see the 'Changelog' file (in the package source).

[18] Ian W McKeague and Wolfgang Wefelmeyer. Markov chain monte carlo and rao–blackwellization. *Journal of statistical planning and inference*, 85(1):171–182, 2000.

[19] Marco J Morelli, Gaël Thébaud, Joël Chadœuf, 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.

[20] Diego Ruiz-Moreno, Mercedes Pascual, Michael Emch, and Mohammad Yunus. Spatial clustering in the spatio-temporal dynamics of endemic cholera. *BMC infectious diseases*, 10(1):51, 2010.

[21] Laura Forsberg White, Jacco Wallinga, Lyn Finelli, Carrie Reed, Steven Riley, Marc Lipsitch, and Marcello Pagano. Estimation of the reproductive number and the serial interval in early phase of the 2009 influenza a/h1n1 pandemic in the usa. *Influenza and Other Respiratory Viruses*, 3(6):267–276, 2009.

[22] RJF Ypma, AMA Bataille, A Stegeman, G Koch, J Wallinga, and WM Van Ballegooijen. Unravelling transmission trees of infectious diseases by combining genetic and epidemiological data. *Proceedings of the Royal Society B: Biological Sciences*, 279(1728):444–450, 2012.