Bayesian inference for contact networks given epidemic data *

Penn State Department of Statistics Technical Report #10-02

Chris Groendyke
Department of Statistics
Pennsylvania State University
cxg928@stat.psu.edu

David Welch
Department of Statistics
Pennsylvania State University
jdw21@stat.psu.edu

David R. Hunter
Department of Statistics
Pennsylvania State University
dhunter@stat.psu.edu

February 25, 2010

Abstract

In this article, we estimate the parameters of a simple random network and a stochastic epidemic on that network using data consisting of recovery times of infected hosts. The SEIR epidemic model we fit has exponentially distributed transmission times with gamma distributed latent (exposed) and infective periods on a network where every tie exists with the same probability, independent of other ties. We employ a Bayesian framework and MCMC integration to make estimates of the joint posterior distribution of the model parameters. We discuss the accuracy of the estimates of different parameters under various prior assumptions and, in particular, show that it is possible in many scientifically interesting cases to accurately recover the network parameter p. We demonstrate some of the important aspects of our approach by studying a measles outbreak in Hagelloch, Germany in 1861 consisting of 188 affected individuals. We provide an R package to carry out these analyses, which is available publicly on the Comprehensive R Archive Network (CRAN).

Keywords: Erdős-Rényi, exponential random graph model (ERGM), stochastic SEIR epidemic, MCMC, measles

*This work was funded by NIH grant R01-GM083603-01.

1 Introduction

In studying the dynamics of epidemics, the dominant model has long been the "mean field" or "random mixing" model that assumes that an infected individual may spread the disease to any susceptible member of the population (Bailey, 1950; Kermack and McKendrick, 1927). An alternate assumption under which the epidemic spreads only across the edges of a contact network within a population may result in much different epidemic dynamics (Ferrari, 2006; Keeling and Eames, 2005; Meyers, Pourbohloul, Newman, Skowronski, and Brunham, 2005). Much of the work based on this alternate assumption relies heavily on simulations: Some network is taken as given or simulated to have certain properties, then a disease outbreak is simulated on the network and the properties of the epidemic studied; see, for example, Volz (2008) and Barthelemy, Barrat, Pastor-Satorras, and Vespignani (2005).

This article takes a different approach, extending work of Britton and O'Neill (2002), hereafter referred to as BO2002, to consider the central question of statistical inference: Given epidemic data assumed to have arisen from the spread of some disease across a network, what can we say about the properties of the disease spread and the network on which it spread? Ascertaining these properties will allow us to learn about the contact networks associated with certain diseases, thereby enabling researchers to test competing theories about transmission of disease and to devise better containment strategies. In particular, we address some of the practical issues of implementing the framework described by BO2002 such as determining the areas of the parameter space in which parameter estimation might be expected to be fruitful and implementing the software necessary to perform the type of inference described; we also suggest generalizations of the network model used in BO2002. The primary purpose for presenting these extensions is to move toward the goal of developing an inferential methodology of practical use.

The remainder of this paper is organized as follows: in Section 2 we review the models used in this study, including the model of the population network structure and the model governing the dynamics of the spread of an epidemic through the population. In Section 3 we discuss Bayesian inference for the model parameters. In Section 4 we discuss the MCMC algorithm used to obtain samples from the desired posterior distributions. Section 5 tests our methodology on multiple simulated data sets, and goes on to apply it to data from a measles outbreak in Hagelloch, Germany in 1861. We then offer some conclusions and a discussion of possible extensions and future work in Section 6.

2 Network and Epidemic Models

2.1 Network Structure

We consider a finite population of fixed size N in which the contact structure between individuals is modeled as an Erdős-Rényi random graph. That is, let $\mathcal{G} = (V, D)$ be a graph with vertices $V = \{1, \ldots, N\}$ corresponding to the N individuals in the population. For $i, j \in V$, $i \neq j$, the (undirected) edge $\{i, j\}$ is in $\mathcal{G}$ with probability p, independently of all other pairs of vertices. For convenience, we let D_{ij} denote the event that $\{i, j\} \in \mathcal{G}$. Here, a "contact" is interpreted as the occurrence of a physical association of two individuals that could be sufficient for disease transmission, though not all contacts between infective and susceptible individuals are guaranteed to result in transmission. In particular, "contact" will have different interpretations in different disease contexts. The Erdős-Rényi model used here is one particular type of a more general class of exponential-family random graph models (ERGMs). We discuss the possibility of extending this type of analysis to more general ERGMs in Section 6.1.

2.2 SEIR Epidemic Model

We describe the spread of a disease through the population by an SEIR model that divides the population into four groups: Susceptible, Exposed, Infective, and Removed; see Keeling and Rohani (2008) for details of this model. Individuals are in the exposed state for a period of time modeled by a gamma random variable with parameters k_E and θ_E, after which time they move to the infective state; the length of time spent in this state is given by a gamma random variable with parameters k_I and θ_I. The disease spreads across the edges in the network from infective individuals to susceptible ones, where the time until transmission across a given edge is modeled by an exponential random variable with mean $1/\beta$. Using gamma random variables to model the lengths of time spent in the exposed and infected states (as opposed to the exponential random variables used by BO2002) increases the flexiblity of the model, but also increases the number of parameters that we must estimate. Indeed, Ray and Marzouk (2008), who also used gamma random variables to model these periods, note that they were unable to perform meaningful inference on their full set of parameters, though this may be due at least in part to the paucity of their data (the data set they consider had a total of only 32 infected individuals).

Finally, when an infective individual can no longer transmit the disease (e.g., due to recovery or death), he or she belongs to the removed group and plays no further part in the spread of the epidemic. The epidemic continues until there are no remaining exposed or infective individuals in the population. Clearly, the dynamics of the epidemic and the proportion of the population that becomes infected depend heavily on the parameters in the network model (p) and in the epidemic model ($\beta, k_E, \theta_E, k_I, \theta_I$).

3 Inference

3.1 Data and Notation

The data we consider are the removal times for each infected node. The data may also include the exposure and/or infective times for each node, but in general, these will be unknown. The exposure, infective, and removal times for node j are denoted by $E_j, I_j,$ and R_j, respectively. The sets of all exposure, infective, and recovery times are $\mathbf{E} = (E_1, E_2, \ldots, E_N), \mathbf{I} = (I_1, I_2, \ldots, I_N),$ and $\mathbf{R} = (R_1, R_2, \ldots, R_N)$; we will denote the entire set of times $(\mathbf{E}, \mathbf{I}, \mathbf{R})$ by $\mathbf{T}$. We assign a value of ∞ to $E_b, I_b,$ and R_b for any node b that was not infected during the course of the epidemic. Denote the identity of the initial exposed (which is, in general, unknown) by κ. Because we will sometimes need to treat the exposure time of the initial exposed separately from that of the other infecteds, we denote by $\mathbf{E}_{-\kappa}$ the set of the exposure times except for the initial infected, i.e., $\mathbf{E} \setminus E_\kappa$. For convenience, we label the nodes so that the ones who were infected during the epidemic are $1, \ldots m$, where m is the number of nodes that were ultimately infected, and $1 \leq m \leq N$.

For this analysis, we perform inference on the model parameters using a Bayesian approach. We will denote the prior distribution for the parameter δ by $\pi_\delta(\cdot)$, the likelihood function by $L(\mathbf{T}|\delta_1, \delta_2, \ldots)$, and the posterior distribution of δ by $\pi_\delta(\cdot|\mathbf{T})$.

Denote the fixed but unknown contact network in the population by $\mathcal{G}$ and the associated transmission tree (pathway along which the epidemic spreads) by $\mathcal{P}$. $\mathcal{P}$ is a directed subgraph of the undirected graph $\mathcal{G}$. We will say that the edge $(a, b) \in \mathcal{P}$ iff a infects b. Note that if $(a, b) \in \mathcal{P}$, we must have

$$I_a < E_b < R_a. \tag{1}$$

We also have the following relationships: $m - 1 = |\mathcal{P}| \leq |\mathcal{G}| \leq \binom{N}{2}$, where $|\mathcal{P}|$ and $|\mathcal{G}|$ denote the number of (directed) edges in $\mathcal{P}$ and the number of (undirected) edges in $\mathcal{G}$, respectively. This notation borrows from and extends that used in BO2002.

Figure 1 shows a example of a contact network $\mathcal{G}$ within a population of $N = 25$ individuals. This network was simulated using an Erdös-Rényi model with $p = 0.15$. Superimposed on this contact network is the transmission tree $\mathcal{P}$ generated from a simulated SEIR stochastic epidemic, with $\beta = 0.15$ and $k_I = k_E = \theta_I = \theta_E = 3$. Figure 2 illustrates the spread of this epidemic over time.

3.2 Likelihood Calculation

In order to calculate the likelihood function for this model, it would be necessary to sum over all possible values of $\mathcal{G}$ and $\mathcal{P}$:

$$L(\mathbf{T}|\beta, k_E, \theta_E, k_I, \theta_I, p) = \sum_{\mathcal{G},\mathcal{P}} L(\mathbf{T}|\beta, k_E, \theta_E, k_I, \theta_I, p, \mathcal{G}, \mathcal{P}) f(\mathcal{G}, \mathcal{P}|p)$$

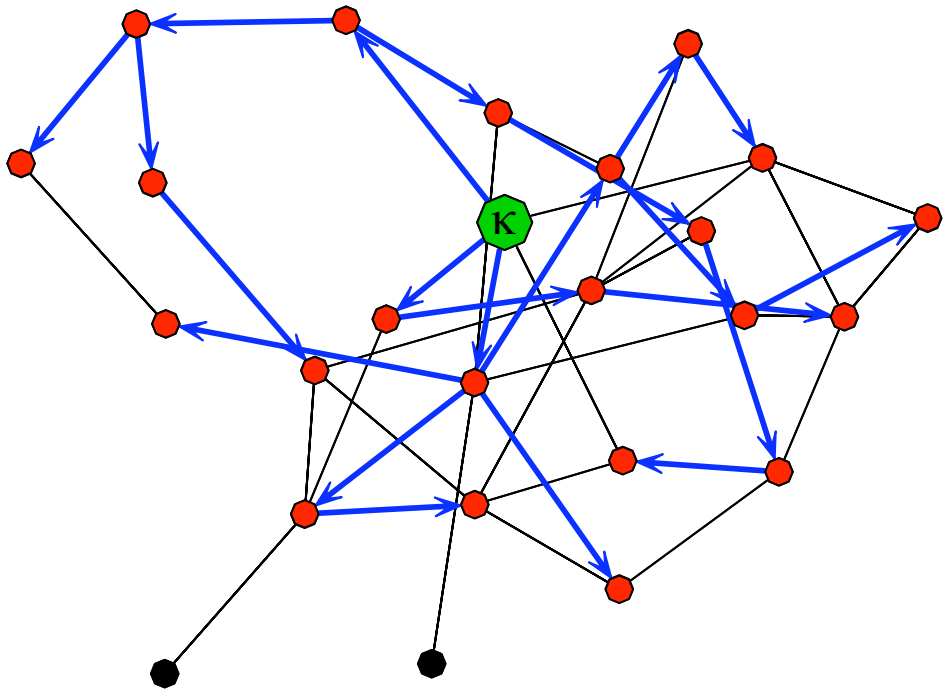

Figure 1: A realization of an Erdös-Rényi contact network ($\mathcal{G}$) with a simulated SEIR epidemic. The large green node represents the initial exposed individual (κ). Red nodes represent individuals who were infected during the course of the epidemic, while the black nodes remained susceptible throughout. The blue arrows show the path of the epidemic across the network (i.e, the transmission tree $\mathcal{P}$). The black lines indicate edges in the contact network across which the epidemic did *not* travel (i.e., $\mathcal{G} \setminus \mathcal{P}$).

$$= \sum_{\mathcal{G}} \sum_{\mathcal{P}} L(\mathbf{T}|\beta, k_E, \theta_E, k_I, \theta_I, p, \mathcal{G}, \mathcal{P}) f(\mathcal{P}|\mathcal{G}) f(\mathcal{G}|p).$$

For all but the smallest problems, this summation contains too many terms to practically compute, so we treat $\mathcal{G}$ and $\mathcal{P}$ as extra parameters; given the values of $\mathcal{G}$ and $\mathcal{P}$, the likelihood is relatively simple to compute. We therefore estimate $\mathcal{G}$ and $\mathcal{P}$ along with the other parameters of interest in the model. For similar reasons, we also condition on the initial exposure time, E_κ. The likelihood only depends on p through $\mathcal{G}$, allowing us to write

$$L(\mathbf{E}_{-\kappa}, \mathbf{I}, \mathbf{R}|\beta, k_E, \theta_E, k_I, \theta_I, \mathcal{G}, \mathcal{P}, E_\kappa).$$

We can calculate the likelihood as $L_1 L_2 L_3 L_4$, where L_1 is the contribution to the

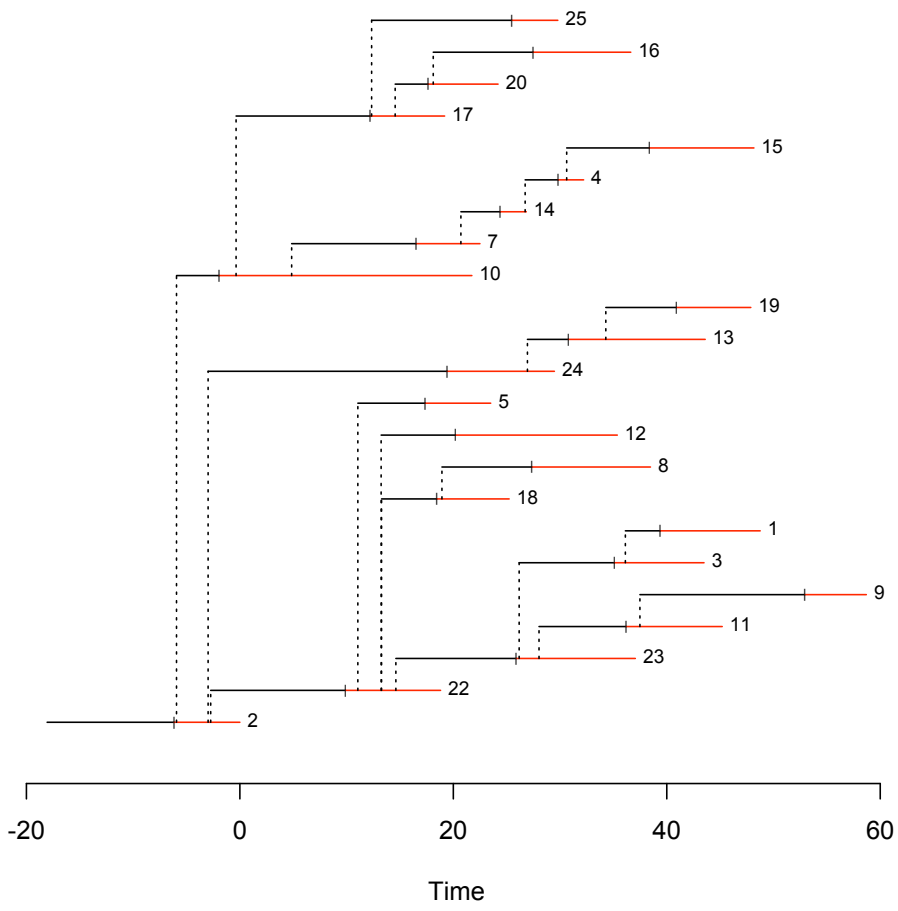

Figure 2: Progression of the SEIR epidemic through time, as produced by function **plotepitree()** in **R** package **epinet**. Vertical dashed line segments show the infection pathway. Horizontal solid line segments show the time periods that the individuals were in the exposed (black) and infective (red) stages of the epidemic. The identities (node numbers) of the individuals are given to the right of their respective epidemic periods.

likelihood from the edges over which the epidemic was transmitted (i.e., $\mathcal{P}$), L_2 is the contribution to the likelihood from the edges over which the epidemic did not pass ($\mathcal{G} \setminus \mathcal{P}$), and L_3 and L_4 are the contributions due to the transition (from exposed to infective) and removal processes, respectively. The likelihood function is defined to be 0 for any values of $\mathbf{T}$ that violate Inequality (1).

L_1 and L_2 are given by

$$L_1 = \beta^{m-1} \exp\left[-\beta \sum_{(a,b)\in\mathcal{P}} (E_b - I_a)\right]$$

and

$$L_2 = \exp\left[-\beta \sum_{(a,b)\in\mathcal{G}\setminus\mathcal{P}} \left[\{(E_b \wedge R_a) - I_a\} \vee 0\right]\right],$$

so that

$$L_1 L_2 = \beta^{m-1} \exp\left[-\beta A\right],$$

where

$$\begin{aligned} A &= \sum_{(a,b)\in\mathcal{P}} (E_b - I_a) + \sum_{(a,b)\in\mathcal{G}\setminus\mathcal{P}} \left[\{(E_b \wedge R_a) - I_a\} \vee 0\right] \\ &= \sum_{a=1}^{m}\sum_{b=1}^{m} \mathbf{1}\left(D_{ab}\right) \cdot \mathbf{1}\left(I_a < E_b\right) \cdot \left[\{(E_b \wedge R_a) - I_a\} \vee 0\right] + \sum_{a=1}^{m} S(a) \cdot (R_a - I_a). \quad (2) \end{aligned}$$

$\mathbf{1}(\cdot)$ is the indicator function and $S(a)$ denotes the number of susceptible (never infected) nodes that node a shares an edge with. A is therefore the total amount of "infectious pressure" applied over the course of the epidemic. The expression (2), which is analagous to that derived in Neal and Roberts (2005), is the form we use in our algorithm.

L_3 and L_4 are given by

$$L_3 = \frac{\left[\prod_{i=1}^{m}(I_i - E_i)\right]^{k_E - 1} \theta^{-mk_E} e^{-B/\theta_E}}{\left[\Gamma(k_E)\right]^m}$$

and

$$L_4 = \frac{\left[\prod_{i=1}^{m}(R_i - I_i)\right]^{k_I - 1} \theta^{-mk_I} e^{-C/\theta_I}}{\left[\Gamma(k_I)\right]^m},$$

where $B = \sum_{i=1}^{m}(I_i - E_i)$ and $C = \sum_{i=1}^{m}(R_i - I_i)$ may be regarded respectively as the total "transition pressure" and "removal pressure" applied over the course of the epidemic.

3.3 Prior Distributions

For some model parameters, we typically use conjugate prior distributions; these distributions are often preferable when they are available, as they can simplify and / or accelerate the process of updating these parameters. In particular, the beta distribution is conjugate for the network parameter p; the inverse gamma distribution is conjugate for the epidemic parameters θ_E and θ_I; and the gamma distribution is conjugate for the parameter β. When it is necessary to infer the exposure and / or infective times, we assign them uninformative (flat) prior distributions; when necessary, we assign a uniform prior for κ. For the k_E and k_I parameters, we use gamma or uniform prior distributions.

In choosing parameters for the prior distributions, we can obtain guidance from independent information known about the disease and / or population, as well as from the scientific literature in some cases. For example, much work has been done to study the lengths of times that individuals infected with the measles virus spend in the exposed and infective states; we can use this information to constuct prior distributions for the parameters governing these periods. Regarding the parameter p, if we have reason to believe that the network under consideration is likely to be sparse, we might then choose a beta distribution that places greater mass on the smaller values of p.

4 MCMC Algorithm

Here we describe the MCMC algorithm used to produce samples from the posterior distributions of the parameters. At each iteration, we update each parameter in turn: $\{\mathcal{P}, \mathcal{G}, p, \beta, k_E, \theta_E, k_I, \theta_I, \mathbf{I}, \mathbf{E}, \kappa\}$, using the methods described below. Experimentation indicates that updating the parameters in a fixed order results in better mixing of the Markov chain than choosing a random update order for each cycle. Note that in the case where the exposure times are assumed to be known, we do not update $\mathbf{E}$; similarly, when the infective times are known, we need not update $\mathbf{I}$. Only in the case in which both $\mathbf{E}$ and $\mathbf{I}$ are unknown do we need to infer κ. This algorithm is based in part on the algorithm described in BO2002. However, we did not find that the "mixing step" described by those authors significantly improved the performance of the algorithm, and hence did not include it in our algorithm. Neal and Roberts (2005) give an algorithm based on a different representation of the network model; we discuss the relative merits of the two parameterizations in Section 4.6.

4.1 Updating $k_E, k_I, p, \beta, \theta_E, \theta_I$

These parameters can be updated via a standard Hastings step. We propose updated values from a uniform distribution centered at the current value of the parameter. Alternatively, the p, β, θ_E, and θ_I parameters can be updated using Gibbs samplers from their conditional distributions, if appropriate prior distributions are used. Let $X \sim$ gamma(a, b) indicate that X has a gamma distribution with density $x^{a-1}b^{-a}e^{-x/b}/\Gamma(a)$ for $x > 0$; let $W \sim \text{IG}(c, d)$ indicate that W has an inverse gamma distribution, i.e., $1/W \sim$ gamma$(c, 1/d)$; let $Y \sim$ beta(q, z) indicate that Y has a beta distribution with parameters q and z on $(0, 1)$; and let $U \sim \mathcal{U}(a, b)$ indicate that U has a uniform distribution on (a, b).

If we assign the following prior distributions as described in Section 3.3: $\pi_\beta(\beta) \sim$ gamma (a_β, b_β), $\pi_{\theta_I}(\theta_I) \sim$ IG (a_I, b_I), $\pi_{\theta_E}(\theta_E) \sim$ IG (a_E, b_E), and $\pi_p(p) \sim$ beta(c, d), then the corresponding full conditional distributions of these parameters are:

$$\pi_\beta(\beta|\mathbf{T}) \sim \text{gamma}\left(m + a_\beta - 1, \frac{1}{A + 1/b_\beta}\right),$$

$$\pi_{\theta_E}(\theta_E|\mathbf{T}) \sim \text{IG}\left(mk_E + a_E, \frac{1}{B + 1/b_E}\right),$$
$$\pi_{\theta_I}(\theta_I|\mathbf{T}) \sim \text{IG}\left(mk_I + a_I, \frac{1}{C + 1/b_I}\right) \text{ and}$$
$$\pi_p(p|\mathbf{T}) \sim \text{beta}\left(|\mathcal{G}| + c, \binom{N}{2} - |\mathcal{G}| + d\right),$$

where $A, B,$ and C are as defined in Section 3.2.

4.2 Updating $\mathcal{G}$

Since we are assuming that the existence of each edge is independent of all other edges, we can generate each edge individually in order to sample from the full conditional distribution of $\mathcal{G}$. We calculate the full conditional probability of D_{ij} for each (i, j), assuming without loss of generality that $E_i < E_j$:

$$P\left(D_{ij}|\mathbf{T}, \mathcal{P}, \beta, p\right) = \frac{P(\mathbf{T}|D_{ij}, \mathcal{P}, \beta, p)P(D_{ij}|\mathcal{P}, p)}{P(\mathbf{T}|D_{ij}, \mathcal{P}, \beta, p)P(D_{ij}|\mathcal{P}, p) + P(\mathbf{T}|D_{ij}^c, \mathcal{P}, \beta, p)P(D_{ij}^c|\mathcal{P}, p)},$$

where $P(D_{ij}|\mathcal{P}, p) = p$, unless the edge appears in $\mathcal{P}$, in which case $P(D_{ij}|\mathcal{P}, p) = 1$, and the values of $P(\mathbf{T}|D_{ij}, \mathcal{P}, \beta, p)$ and $P(\mathbf{T}|D_{ij}^c, \mathcal{P}, \beta, p)$ vary depending on the status (ultimately infected or never infected) of nodes i and j. Note that we only need to consider the data associated with these two nodes, rather than the entirety of $\mathbf{T}$. If $(i, j) \in \mathcal{P}$ then $P\left(D_{ij}|\mathbf{T}, \mathcal{P}, \beta, p\right) = 1$, since $(i, j) \in \mathcal{P} \Rightarrow \{i, j\} \in \mathcal{G}$. Otherwise, if $(i, j) \notin \mathcal{P}$, then

$$P\left(D_{ij}|\mathbf{T}, \mathcal{P}, \beta, p\right) = \frac{\exp\left(-\beta\left[\{(R_i \wedge E_j) - I_i\} \vee 0\right]\right) \cdot p}{1 - p + \exp\left(-\beta\left[\{(R_i \wedge E_j) - I_i\} \vee 0\right]\right) \cdot p}.$$

Recall that $E_k = I_k = R_k = \infty$ for $k > m$; we also use the convention that $\infty - \infty = 0$ for the purpose of evaluating the above probabilities.

4.3 Updating $\mathcal{P}$

Updating the transmission tree consists of determining, for each infected node except the initial exposed, which node infected it. Let $\mathcal{P}_j$ denote the parent of node j and $\pi_{\mathcal{P}_j}(r)$ denote the prior probability that node r is the parent of j. The candidate nodes for the parent of node j (i.e., the node that infected j) are exactly those nodes i for which $\{i, j\} \in \mathcal{G}$ and $I_i \leq E_j \leq R_i$. Denote these candidate nodes by $i_1, \ldots, i_k$. Then the probability that i_t is the parent of j, given that one of the candidates is known to

have infected j, is

$$\frac{\beta \exp\left(-\beta\left[\sum_{i\in\{i_1,\ldots,i_k\}} E_j - I_i\right]\right)\cdot \pi_{\mathcal{P}_j}(i_t)}{\sum_{a=1}^{k} \beta \exp\left(-\beta\left[\sum_{i\in\{i_1,\ldots,i_k\}} E_j - I_i\right]\right)\cdot \pi_{\mathcal{P}_j}(i_a)} = \frac{\pi_{\mathcal{P}_j}(i_t)}{\sum_{a=1}^{k}\pi_{\mathcal{P}_j}(i_a)}$$

a function of only the prior assumptions. If we assume that $\pi_{\mathcal{P}_j}(r)$ is the same for all j and r (we will often make this uniform assumption in the absence of other information, though we consider other possibilities in Section 5.2.2), then each of the candidates is equally likely to be the parent. To find the parent of node j, we simply find the parent candidates and sample from among them according to their respective probabilities. We repeat this for each infected node (except the initial exposed) in order to produce a sample from the full conditional distribution of $\mathcal{P}$.

4.4 Updating κ

Note that we only need to update κ in the cases in which both $\mathbf{E}$ and $\mathbf{I}$ are not fully known. To perform this update, we use a method similar to that described in BO2002. The prior assumption is that each of the m infected nodes is equally likely to be the initial infected, i.e., $\pi_\kappa(i) = 1/m$ for all i. Given the current value of κ, we choose a proposed value for κ^* by sampling uniformly from the set $\{j : (\kappa, j) \in \mathcal{P}\}$. We propose new values for $\mathcal{P}, \mathbf{E},$ and $\mathbf{I}$ that are consistent with κ^* in the following manner. First, we swap the values of E_κ and I_κ with those of E_{κ^*} and I_{κ^*}, respectively. Then, we replace the edge (κ, κ^*) in $\mathcal{P}$ with (κ^*, κ). We determine whether or not to accept the proposed values according to the appropriate Hastings ratio.

4.5 Updating $\mathbf{E}, \mathbf{I}$

We update each element of $\mathbf{E}_{-\kappa}$ in a uniformly random order, and finally update E_κ. We use a Hastings step to update each element of $\mathbf{E}$. For each $j \neq \kappa$, we first identify the parent of j in $\mathcal{P}$, i.e., the node that infected j. Since i must have been infected (and not yet recovered) when j became exposed, and since j enters the exposed phase before the infected phase, we must have $I_i < E_j < \min\{I_j, R_i\}$. The proposed updated value for E_j is generated from a uniform distribution on the interval of its possible values.

Our method for updating $\mathbf{I}$ is similar to the method for $\mathbf{E}$, updating each I_j in a uniformly random order. We find a proposed value for each I_j by sampling uniformly from the interval of its possible values and accept each proposal according to the appropriate Hastings ratio.

4.6 Implementation

We have built a package for **R** (R Development Core Team, 2009), named **epinet**, containing software which implements the algorithm described above. This software is publicly available on the Comprehensive R Archive Network (cran.r-project.org), and will in the future be maintained to reflect future extensions and/or generalizations made to the model, such as the ERGM extensions discussed in Section 6.1.

The internal representation of the graph structure, which is based on the binary tree representation used in the **ergm** package (Handcock, Hunter, Butts, Goodreau, Morris, and Krivitsky, 2009), allows for efficient storage of the graph, especially for sparse graphs. There are several reasons that we chose to extend the MCMC algorithm described in BO2002, rather than the algorithm detailed in Neal and Roberts (2005). The first reason is scalability. As noted in Neal and Roberts (2005), it is not necessary to know the entirety of the graph $\mathcal{G}$ in order to calculate the likelihood for our model. Neal and Roberts (2005) proposes using a subgraph $\mathcal{F}$ that does not consider the edges between never-infected individuals; this subgraph consists of $N \cdot m$ dyads. Our algorithm, using the expression for the likelihood given in (2), only explicitly considers the edges between two infected individuals. We also must keep track of the number of never-infected individuals that each of the m infecteds is connected to. Thus, the algorithm we describe requires storage and updating of $m^2 + m$ rather than $N \cdot m$ elements. In cases where $N \gg m$, i.e., when only a small portion of a population is infected, this savings in computing resources may be significant.

In our experience, the algorithm described in Neal and Roberts (2005) ran significantly more slowly than did the algorithm described in 4. When the parameter p is updated in the Neal and Roberts algorithm, each edge in $\mathcal{F}$ is also updated, causing this update to be very slow. This phenomenon will only be exacerbated as we move to more complicated ERGM models (see Section 6.1) which have more parameters, as $\mathcal{F}$ would need to be updated with each of them. The two methods gave roughly similar results in terms of mixing, as measured by number of effective samples produced, though the relative performances of the algorithms varied by parameter and dataset.

5 Applications

5.1 Simulated Epidemics

We simulate epidemics over Erdős-Rényi networks with varying values of p and β and attempt to recover these parametersusing the algorithm described above. Our primary interest here lies in the parameter describing the network model, p.

One difficulty in performing inference for this model lies in distinguishing the effects of the network parameter p from the epidemic parameter β; as discussed in BO2002, the spread of an epidemic throughout a population could be due to either a fast transmission

rate (high value of β) or a more fully connected network (large value of p). This ambiguity can lead to difficulties in estimating (or distinguishing the effects of) these two parameters. Hence, there is often a significant negative correlation between the samples produced for these two parameters—as the values of p increase, the values of β decrease, and vice-versa, leading in some cases to a narrow estimated posterior distribution for the quantity $p \cdot \beta$ but much more dispersed posteriors for these two parameters individually.

To explore this issue, we consider ten different simulated Erdős-Rényi networks of $N = 40$ individuals with $p = 0.1, 0.2, \ldots, 1$. Over each of these ten networks, we simulate epidemics with five different values of β: 0.01, 0.05, 0.1, 0.5, and 1. The values of the other epidemic parameters are set in each case at $k_I = k_E = \theta_I = \theta_E = 5$. We ran the MCMC algorithm described above with full data, assuming that the exposure, infective, and recovery times were all known. We chose uniform priors for each parameter, specifically $\pi_\beta \sim \mathcal{U}(0, 1.5), \pi_p \sim \text{beta}(1, 1), \pi_{k_I} \sim \mathcal{U}(3, 7), \pi_{\theta_I} \sim \mathcal{U}(3, 7), \pi_{k_E} \sim \mathcal{U}(3, 7)$, and $\pi_{\theta_E} \sim \mathcal{U}(3, 7)$. In each case, we ran the algorithm for 100 million iterations, thinning every 100 iterations to obtain at least 50,000 approximately independent samples from each of the parameters.

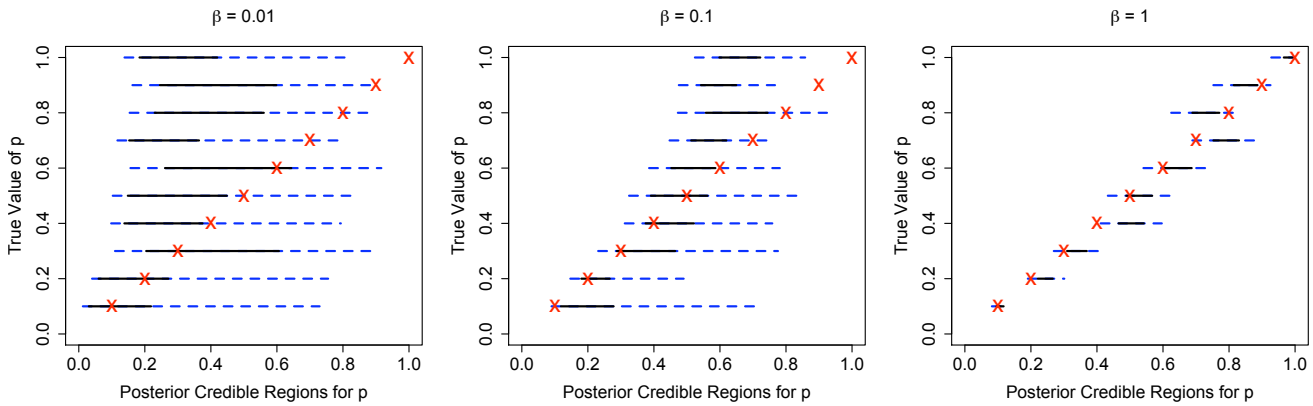

Figure 3: 50% (black, solid) and 90% (blue, dashed) posterior credible regions for p, for ten different simulated Erdős-Rényi networks with $p = 0.1, 0.2, \ldots, 1$, and simulated SEIR epidemics with $\beta = 0.01$ (left), $\beta = 0.1$ (center), and $\beta = 1$ (right).

Figure 3 gives 50% and 90% posterior credible intervals for the parameter p for $\beta = 0.01, 0.1,$ and 1, and for each of the 10 values of p. We can see from this that some areas of the p and β parameter space lend themselves to meaningful inference for the network parameter p, while in other regions, the resulting posterior distributions for p are not very informative. In particular, when $\beta = 0.01$, the posterior credible intervals for p are relatively wide and tend not to be able to distinguish the different values of

p, whereas for $\beta = 1$, the data allow us to better recover the parameter p.

Figure 4 shows scatterplots of the posterior distribution of $\log(p)$ and $\log(\beta)$ for the SEIR epidemic simulations where $\beta = 0.01, 0.1,$ and 1; p was set to 0.2 in each case. (We chose a relatively low value of p as a basis for comparison because we are more likely to encounter smaller values of this parameter in the actual networks of interest.) While the correlations between these two parameters are negative in each case as expected, the lower values of β show a much more substantial negative correlation.

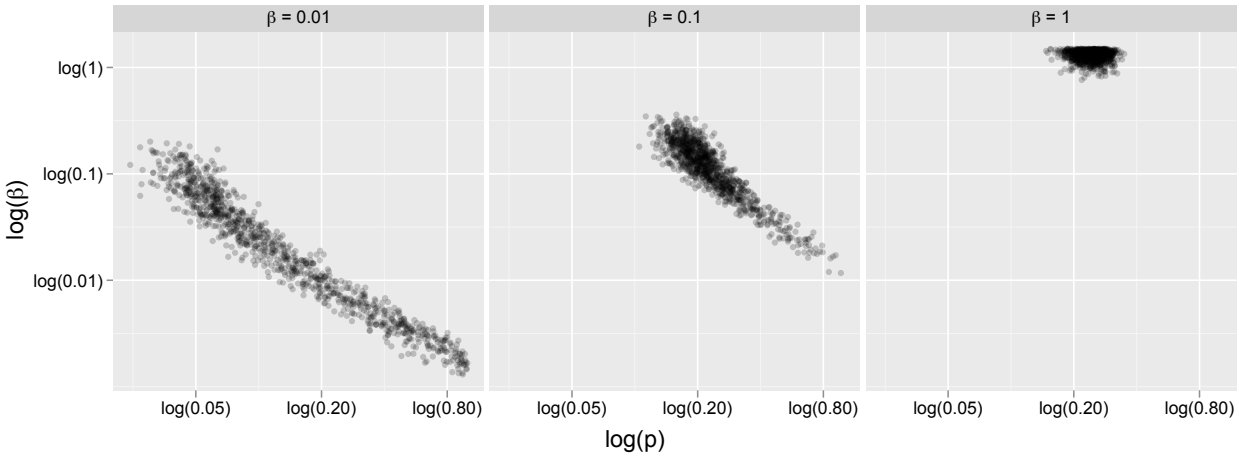

Figure 4: Scatterplots of $\log(p)$ vs $\log(\beta)$ for a simulated Erdős-Rényi network with $p = 0.2$ and three different simulated SEIR epidemics with $\beta = 0.01, 0.1,$ and 1. Each plot shows 1,000 points sampled from the respective posterior samples for these parameters. Posterior correlations between $\log(p)$ and $\log(\beta)$ were estimated at $-0.97, -0.90,$ and -0.05 for $\beta = 0.01, 0.1,$ and 1, respectively.

These results indicate that, while distinguishing the effects of p and β is indeed difficult for some combinations of these parameters, it is nonetheless possible to perform meaningful inference for p elsewhere in large, qualitatively meaningful portions of the parameter space, particularly for the smaller values of p and the relatively large values of β. In real populations, networks tend to be fairly sparse, especially as the number of nodes increases, which means that we would expect small values of p. Therefore, our simulation results are encouraging from the standpoint of fitting these models to real data.

5.2 Hagelloch Measles Data

We consider data arising from a measles epidemic that spread through the small town of Hagelloch, Germany in 1861. This data set, which contains data on 188 infected in-

dividuals, is notable for its completeness and depth of data (Pfeilsticker, 1863). Following Neal and Roberts (2004), we assume that the entirety of the susceptible population (which consists of the children born after the previous measles outbreak in Hagelloch) was ultimately infected in this epidemic. We assume that each individual's infective period begins one day prior to the onset of prodromes and ends three days after the appearance of rash. As the data do not contain any information about the exposure times of the individuals ($\mathbf{E}$), we will treat these times as unknown and infer them.

We initially performed inference on this data set under two different sets of prior assumptions for the parameters. In the first case, we used uniform priors for all parameters, so $\pi_\beta \sim \mathcal{U}(0,2), \pi_p \sim \mathcal{U}(0,1), \pi_{k_I} \sim \mathcal{U}(15,25), \pi_{\theta_I} \sim \mathcal{U}(0.25,0.75), \pi_{k_E} \sim \mathcal{U}(8,20)$, and $\pi_{\theta_E} \sim \mathcal{U}(0.25,1)$. In the second case, we used the conjugate prior distributions described in Section 4.1, so that $\pi_\beta \sim \text{gamma}(2,0.5), \pi_p \sim \text{beta}(1,1), \pi_{k_I} \sim \text{gamma}(20,1), \pi_{\theta_I} \sim \text{IG}(3.5,1), \pi_{k_E} \sim \text{gamma}(20,1)$, and $\pi_{\theta_E} \sim \text{IG}(3.5,1)$. In both cases, we used uniform priors for the transmission tree. There is one individual in the data set who is recorded as showing symptoms of the disease approximately one month after the epidemic had otherwise subsided, making inclusion of this individual in the epidemic questionable. We ran our analyses both including and excluding this individual. In each case, we ran the algorithm for 20 million iterations, thinning every 200 iterations to obtain at least 5,000 approximately independent samples from each of the parameters.

Figure 5 gives histograms of the samples from the posterior distributions of p and β for the case with the conjugate prior distributions and excluding the outlier. We can see that the data indicates a strong signal for p, with an estimated posterior mean of 0.046 and median of 0.042. This would correspond to an average degree (within the susceptible population) of approximately 8. β has an estimated posterior mean of 0.96 and median of 0.81. The estimates for these parameters remained largely unchanged over the various sets of assumptions used. We also note that the correlation between $\log(p)$ and $\log(\beta)$ is roughly -0.5, which is far enough from -1 to allow us to extract separate information concerning p and β.

The basic reproduction number, R_0, is defined as the expected number of infectious contacts that a single infective has in a totally susceptible population (Keeling and Rohani, 2008). Analogous to the expression derived in BO2002, we can express this quantity as $R_0 = N \cdot p \cdot \text{P}(X < Y)$, where X is the time that it takes an individual to cause an infection along a given edge after entering the infective state and Y is the time to the individual's removal after becoming infective. For the SEIR model we consider here, X is exponentially distributed with mean $1/\beta$ and $Y \sim \text{gamma}(k_I, \theta_I)$, so that

$$R_0 = N \cdot p \cdot \left(1 - \left(\frac{1}{1+\beta\theta_I}\right)^{k_I}\right). \tag{3}$$

Note that this expression reduces to the formula given in BO2002 in the case that the length of the infective periods are modeled by an exponential random variable, i.e.,

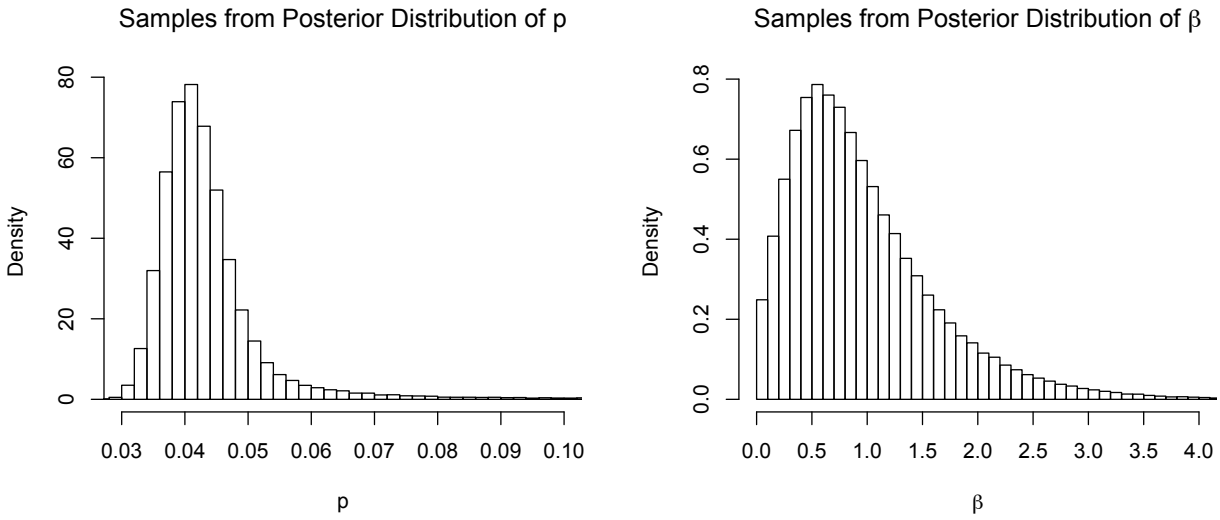

Figure 5: Histogram of values sampled from the posterior density for parameters p and β, excluding the outlying data point and using conjugate prior distributions.

$k_I = 1$. For the values of the $\beta, \theta_I,$ and k_I that we consider in this example, Equation (3) will typically be slightly less than $N \cdot p$, an individual's mean number of contacts.

Figure 6 gives a histogram of the posterior samples for R_0, as calculated by Equation (3) using the posterior parameter samples for $\beta, \theta_I,$ and k_I produced by the Hagelloch data analysis. As is evident from the figure, the posterior mean and mode are in the 7 to 8 range. This is comparable with figures reported in the literature on measles. For instance, Huang (2008) gives a range of 5.8 to 14.3; and Edmunds, Gay, Kretzschmar, Pebody, and Wachmann (2001) give a range of 6.1 to 10.2 for different outbreaks, using data that while much later than the Hagelloch data are still from pre-vaccination Europe. We caution that our results should not be extrapolated to measles outbreaks in general, since they come from only a single outbreak, but nonetheless our R_0 values are consistent with existing results.

The two sets of priors did not result in dramatically different posterior distributions for any of the parameters (note that the prior assumption for p was actually the same in both cases). We did, however, notice a $10\% - 20\%$ decrease in runtimes for the cases where the conjugate prior distributions were used; this was due to the ability to sample directly from the conditional distributions of many of the parameters rather than relying on computationally expensive Metropolis-Hastings steps.

Because we were required to infer the exposure times in this epidemic, it is also interesting to examine the posterior estimates for the parameters governing the exposed

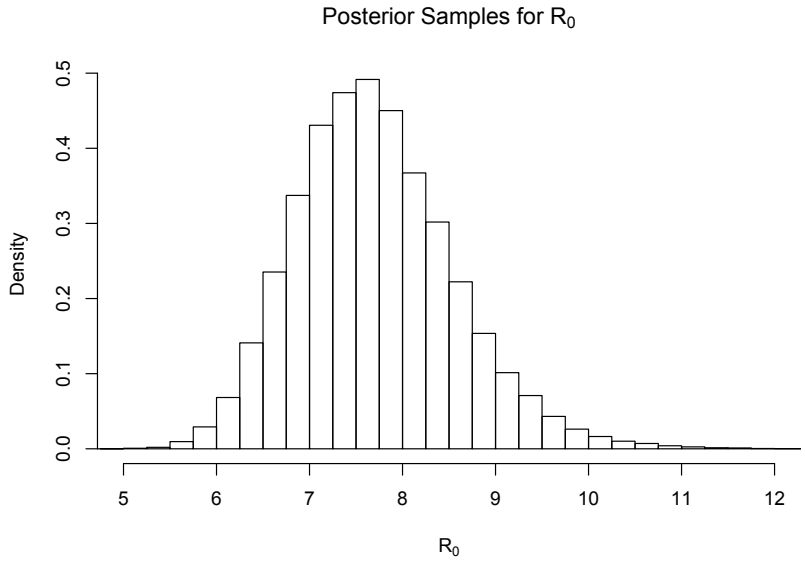

Figure 6: Posterior samples of the basic reproduction number R_0 for the Hagelloch measles data, calculated using Equation (3) using conjugate prior distributions and excluding the outlier.

periods of the individuals. As we are using a gamma(k_E, θ_E) random variable to model the lengths of these periods, their estimated mean and variance are given by $k_E\theta_E$ and $k_E\theta_E^2$, respectively. Figure 7 shows plots of the estimated posterior distributions of these quantities, both including and excluding the outlying data point.

We can see that removing the outlier from this dataset caused a modest decrease in the the estimate of the mean exposed period. Much more noticeable, though, is the effect that removing this outlier had on the corresponding estimated variance – a decrease of over 40%. These results indicate that this outlier was indeed having a large effect on the estimates of the exposed length parameters. None of the other parameters in the model, however, was significantly affected by the inclusion of this outlier. Our posterior mean estimate of 10.3 days for the mean length of the exposed period seems quite reasonable; other authors have estimated the length of this period to be between 6 and 10.3 days for measles (Anderson and May, 1982; Gough, 1977; Schenzle, 1984).

5.2.1 Posterior Predictive Modeling

We assess the quality of fit of the models used for this analysis by simulating SEIR epidemics over Erdős-Rényi networks, using epidemic and network parameter values sampled from the joint posterior distribution produced by the Hagelloch data analysis.

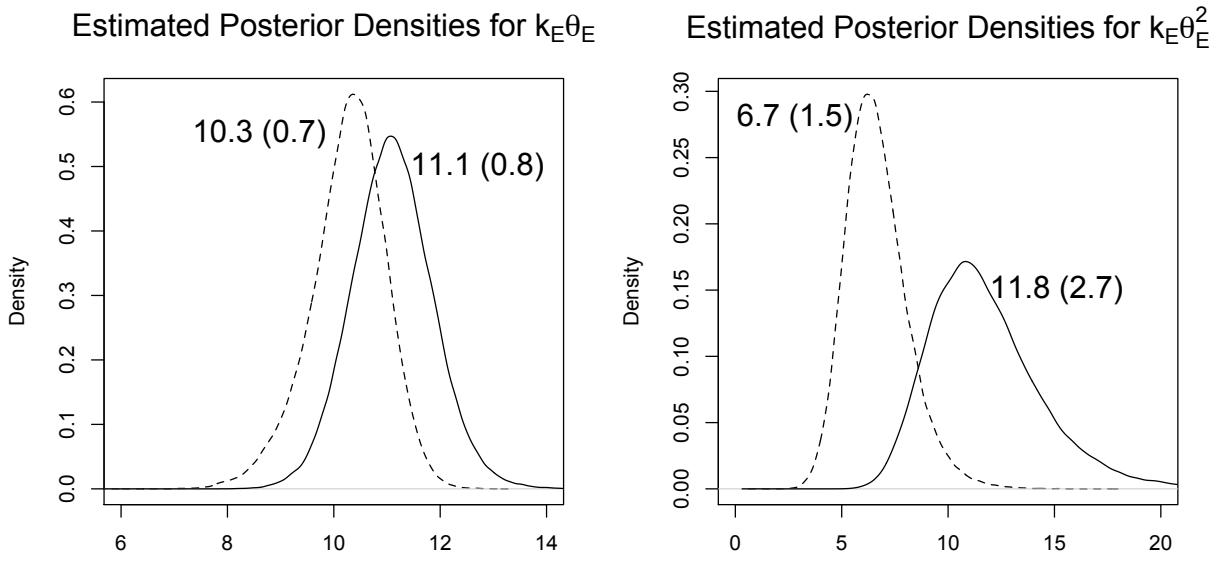

Figure 7: Estimated posterior densities for the mean ($k_E\theta_E$, left panel) and variance ($k_E\theta_E^2$, right panel) of the length of the inferred exposed periods. The solid lines represent estimates based on all data points, while the dashed lines indicate estimates excluding the outlier. The estimated mean (standard deviation) is also given for each density. The prior assumption for $k_E\theta_E$ had a mean of 8 and a standard deviation of 6.9; the prior assumption for $k_E\theta_E^2$ had a mean of 5.3 and infinite variance.

We then compared these simulated epidemics with the original Hagelloch measles epidemic data. One statistic of interest is the number of individuals in the infective stage of the disease, as a function of time. Figure 8 gives a comparison of these epidemic curves for the simulations as compared to the actual data. We see that the epidemic increases more rapidly, and dies out earlier, than the model predicts. However, this is unsurprising given the simplistic contact network model used here: An Erdős-Rényi model may capture the correct mean degree of a network, but the degree distribution itself is fully determined (and approximately normally distributed) once this mean degree is specified. A more realistic network model might more accurately capture the tendency for some nodes to have large degrees—for instance, it should be possible to modify the model in order to capture effects due to spatial location of residence, household, and classroom, as Neal and Roberts (2004) identify as important factors. Despite the simplicity of the Erdős-Rényi model used here, however, the model appears to be a useful first approximation to reality.

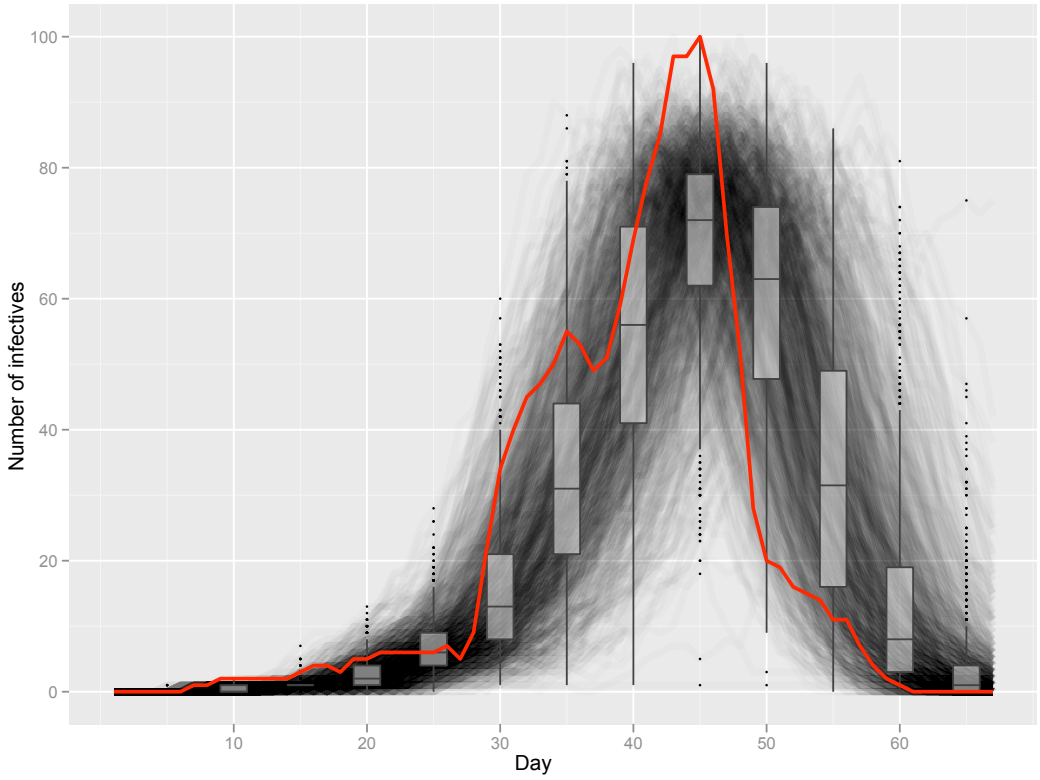

Figure 8: Comparison of the number of individuals in the infective stage of the epidemic, as a function of time. The individual epidemics simulated from samples taken from the joint posterior distribution of the parameters are traced in grey, while the original Hagelloch measles data is plotted in red. Summaries (in the form of boxplots) of the number of infectives from across the 1,000 simulations are given at intervals of five days.

5.2.2 Incorporating Additional Information

We next consider a situation in which the data provide additional information beyond the infective and recovery times considered above. In particular, the Hagelloch data set contains information about the actual transmission tree for this epidemic. For each individual, a "putative parent" is given, i.e., the data contains an indication of who is the most likely to have infected each individual. We use this information in order to construct a more informative prior distribution for the transmission tree $\mathcal{P}$. Rather than assuming a uniform prior for each node j, i.e., $\pi_{\mathcal{P}_j}(r)$ the same for all r, we will incorporate the additional information by placing more prior weight on the putative parent than on the other individuals.

To study the effect of this additional information, we used the algorithm above to perform inference under different sets of prior assumptions for the transmission tree, holding all other prior assumptions constant. (The same conjugate distributions and hyperparameters used in the previous analysis were used in both cases here.) Figure 9 shows the posterior distribution of the parent for a representative node under the uniform transmission tree prior as well as a prior assumption that puts 8 times as much weight on the putative parent node for each individual. As expected, using a more informative prior assumption for the transmission tree leads to a more concentrated posterior distribution for this parameter. Note, however, that in some cases even this prior did not result in a large posterior probability assigned to the putative parent. (In fact, for some nodes, zero posterior probability was assigned to the putative parent.) In other words, we did not, in all cases, find evidence to support each individual's assignment of putative parent.

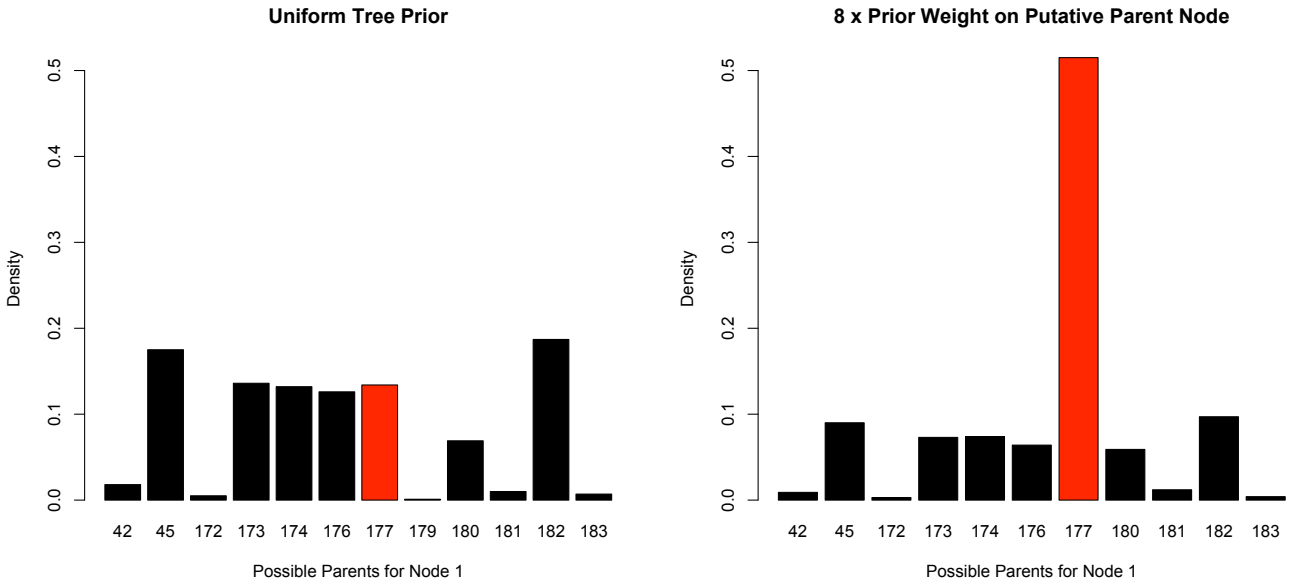

Figure 9: Barplots of the samples for the posterior distribution of the parent of Node 1 under a uniform prior assumption (left panel) and an assumption that places 8 times as much prior mass on the putative parent node (right panel) as the other nodes. In the Hagelloch data set, node 177 (highlighted in red) is identified as the putative parent for node 1. Only nodes with positive posterior probabilities are shown in the charts.

We also consider the impact that this additional information will have on the inference for the others parameters in the model. Incorporating this additional information

has almost no impact on the inferential results for the network parameter p, but does have an effect on some of the other parameters. Most notably, increasing the prior weight placed on the putative parent node caused a slight change in the estimated exposure lengths; in particular the estimated mean exposure length decreased, whereas the estimated variance of the exposure periods increased. This is not surprising, given the data. As mentioned above, many of the assignments of the putative parent node are questionable on the grounds that they would necessitate unreasonably short exposure periods, some as short as 1 or 2 days. As we give more prior weight to these assignments, the estimates exposure lengths must necessarily decrease in mean and increase in variance to accomodate.

6 Discussion

Performing inference for the parameters of a network model, given only data about the infective and removal times of individuals, can be very difficult. The examples in the previous section, however, show that it is indeed possible in many cases to use this type of data to extract meaningful information about the structure of the underlying population. Since this structure is known to play an important role in the dynamics of epidemics, developing novel methods of inferring this structure—a subject which until recently has received relatively little attention—is potentially very valuable.

In this paper, we have expanded the framework provided by BO2002 to perform an analysis of the areas of the parameter space for which estimation is most likely to provide meaningful results. We have also performed inference for the network and epidemic parameters for an actual data set under various sets of prior assumptions; the results obtained were shown to be in concordance with the relevant known scientific information. These developments, demonstrated using efficient new software we have implemented, show that the original framework of BO2002 is viable for datasets much larger than those considered previously in the literature (Britton and O'Neill, 2002; Ray and Marzouk, 2008). It suggests that extensions of the model should be explored (see Section 6.1 below); after all, the Erdős-Rényi network model used here is certainly overly simplistic. In particular, it will be important to consider how to incorporate more data in the network models used. For example, for certain viral diseases we may have genetic data about viruses sampled from infected individuals, and due to the relatively rapid rate of mutations that occur in the viral genomes, these data can inform the structure of the transmission tree $\mathcal{P}$. We can in turn use this additional information to help improve the quality of the inference for the parameters of interest. Or we may have information about the physical locations of various nodes at various times, which could be employed in a more realistic model for the contact network $\mathcal{G}$.

6.1 Extensions of the Network Model

One of the extensions of the model above that we might consider consists of using a more general ERGM to model the interactions in population, as opposed to the Erdős-Rényi model. The ERGM model is very flexible; by specifying various types of graph statistics, we can achieve a wide variety of possible models, and hence provide a more general framework for performing the type of inference described above.

A more complicated ERGM network structure will of course necessitate some modifications to our inference and MCMC algorithm. The likelihood function would need to be modified to include the entire vector of ERGM parameters ($\boldsymbol{\eta}$), rather than just p, so that we would have

$$\begin{aligned} L(\mathbf{E}, \mathbf{I}, \mathbf{R}|\beta, k_E, \theta_E, k_I, \theta_I, \boldsymbol{\eta}) &= \sum_{\mathcal{G},\mathcal{P}} L(\mathbf{E}, \mathbf{R}|\beta, k_E, \theta_E, k_I, \theta_I, \boldsymbol{\eta}, \mathcal{G}, \mathcal{P}) f(\mathcal{G}, \mathcal{P}|\boldsymbol{\eta}) \\ &= \sum_{\mathcal{G}} \sum_{\mathcal{P}} L(\mathbf{E}, \mathbf{I}, \mathbf{R}|\beta, k_E, \theta_E, k_I, \theta_I, \boldsymbol{\eta}, \mathcal{G}, \mathcal{P}) f(\mathcal{P}|\mathcal{G}) f(\mathcal{G}|\boldsymbol{\eta}). \end{aligned}$$

We will have to modify the MCMC algorithm described in Section 4 to reflect the more general ERGM case. For instance, updating the η parameter using a Metropolis-Hastings algorithm would be more difficult, since the Hastings ratio involves the ratio of ERGM normalizing constants for the current parameter value $\eta^{(0)}$ and the new proposal η^*. While this ratio of normalizing constants is trivially easy to calculate in the simplistic case of the Erdős-Rényi model, it is computationally intractable for some other ERGMs (Hunter, Goodreau, and Handcock, 2008; Snijders, 2002). Hence, more complicated updating and estimation schemes such as those described in Snijders (2002) or Hunter, Handcock, Butts, Goodreau, and Morris (2008) may be necessary. Nonetheless, certain types of ERGMs, called dyadic independence models (Hunter et al., 2008), avoid the difficulties of estimating the ratio of normalizing constants while still incorporating useful statistics, such as geographic data on the individual nodes.

Acknowledgments

The authors are grateful to Peter Neal for supplying the Hagelloch dataset and to Shweta Bansal for insightful comments on an early draft of the manuscript. This work is supported by NIH grant R01-GM083603-01.

References

Anderson, R. and R. May (1982). Directly transmitted infections diseases: control by vaccination. *Science 215*(4536), 1053.

Bailey, N. (1950). A simple stochastic epidemic. *Biometrika 37*(3-4), 193–202.

Barthelemy, M., A. Barrat, R. Pastor-Satorras, and A. Vespignani (2005). Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *Journal of Theoretical Biology 235*(2), 275–288.

Britton, T. and P. O'Neill (2002). Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure. *Scandinavian Journal of Statistics 29*(3), 375–390.

Edmunds, W., N. Gay, M. Kretzschmar, R. Pebody, and H. Wachmann (2001). The pre-vaccination epidemiology of measles, mumps and rubella in Europe: implications for modelling studies. *Epidemiology and infection 125*(03), 635–650.

Ferrari, M. (2006). *Mixing models and the geometry of epidemics*. Ph. D. thesis, Pennsylvania State University.

Gough, K. (1977). The estimation of latent and infectious periods. *Biometrika 64*(3), 559.

Handcock, M. S., D. R. Hunter, C. T. Butts, S. M. Goodreau, M. Morris, and P. Krivitsky (2009). *ergm: A Package to Fit, Simulate and Diagnose Exponential-Family Models for Networks*. Seattle, WA. Version 2.2-2. Project home page at url-http://statnetproject.org.

Huang, S. (2008). A new SEIR epidemic model with applications to the theory of eradication and control of diseases, and to the calculation of R0. *Mathematical Biosciences 215*(1), 84–104.

Hunter, D. R., S. M. Goodreau, and M. S. Handcock (2008). Goodness of fit for social network models. *Journal of the American Statistical Association 103*, 248–258.

Hunter, D. R., M. S. Handcock, C. T. Butts, S. M. Goodreau, and M. Morris (2008). ergm: A package to fit, simulate and diagnose exponential-family models for networks. *Journal of Statistical Software 24*(3).

Keeling, M. and K. Eames (2005). Networks and epidemic models. *Journal of the Royal Society Interface 2*(4), 295.

Keeling, M. and P. Rohani (2008). Modeling infectious diseases in humans and animals. *Clinical Infectious Diseases 47*, 864–6.

Kermack, W. and A. McKendrick (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London 115*(772), 700–721.

Meyers, L., B. Pourbohloul, M. Newman, D. Skowronski, and R. Brunham (2005). Network theory and SARS: predicting outbreak diversity. *Journal of theoretical biology 232*(1), 71–81.

Neal, P. and G. Roberts (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics 5*(2), 249.

Neal, P. and G. Roberts (2005). A case study in non-centering for data augmentation: Stochastic epidemics. *Statistics and Computing 15*(4), 315–327.

Pfeilsticker, A. (1863). Beiträge zur Pathologie der Masern mit besonderer Berücksichtigung der statistischen Verhältnisse, M.D. thesis, Eberhard-Karls Universität, Tübingen.

R Development Core Team (2009). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Ray, J. and Y. Marzouk (2008). A bayesian method for inferring transmission chains in a partially observed epidemic. In *Proceedings of the Joint Statistical Meetings: Conference Held in Denver, Colorado, August 3-7, 2008*. American Statistical Association.

Schenzle, D. (1984). An age-structured model of pre-and post-vaccination measles transmission. *Mathematical Medicine and Biology 1*(2), 169.

Snijders, T. A. B. (2002). Markov Chain Monte Carlo estimation of exponential random graph models. *Journal of Social Structure 3*(2).

Volz, E. (2008). SIR dynamics in random networks with heterogeneous connectivity. *Journal of Mathematical Biology 56*(3), 293–310.