Scandinavian Journal of Statistics Theory and Applications

Scandinavian Journal of Statistics
doi: 10.1111/j.1467-9469.2010.00721.x

Bayesian Inference for Contact Networks Given Epidemic Data

CHRIS GROENDYKE

Department of Statistics, Pennsylvania State University

DAVID WELCH and DAVID R. HUNTER

Department of Statistics and Center for Infectious Disease Dynamics, Pennsylvania State University

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 exposed and infectious periods on a network where every edge exists with the same probability, independent of other edges. We employ a Bayesian framework and Markov chain Monte Carlo (MCMC) integration to make estimates of the joint posterior distribution of the model parameters. We discuss the accuracy of the parameter estimates under various prior assumptions and show that it is possible in many scientifically interesting cases to accurately recover the parameters. We demonstrate 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.

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

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 infectious individual may spread the disease to any susceptible member of the population (Kermack & McKendrick, 1927; Bailey, 1950). 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 (Keeling & Eames, 2005; Meyers *et al.*, 2005; Ferrari, 2006). 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 *et al.* (2005).

This article takes a different approach, extending work of Britton & O’Neill (2002) 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 Britton & O’Neill (2002) 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 Britton & O’Neill (2002). The primary purpose for presenting these extensions is to move towards the goal of developing an inferential methodology of practical use.

The remainder of this article 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 datasets, 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 modelled as an Erdős-Rényi random graph, which we denote by $\mathcal{G}$. That is, we define the vertex set $V=\{1,\ldots,N\}$ corresponding to the N individuals in the population, and for two distinct vertices $i,j\in V$, we let $\{i,j\}$ denote the undirected edge between them; we will write $\{i,j\}\in\mathcal{G}$ if there exists a contact between vertices i and j and presume that any such contact exists with probability p, independently of the existence of any other contact. 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 infectious 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, infectious and removed; see Keeling & Rohani (2008) for details of this model. Individuals are in the exposed state for a period of time modelled by a Gamma random variable with mean $k_E\theta_E$ and variance $k_E\theta_E^2$, after which time they move to the infectious state; the length of time spent in this state is given by a Gamma random variable with mean $k_I\theta_I$ and variance $k_I\theta_I^2$. The disease spreads across the edges in the network from infectious individuals to susceptible ones, where the time until transmission across a given edge is modelled by an exponential random variable with mean $1/\beta$. Using Gamma random variables to model the lengths of time spent in the exposed and infectious states (as opposed to Britton & O'Neill, 2002, who used an exponential distribution for the length of time spent in the infectious state of their SIR model) increases the flexiblity of the model, but also increases the number of parameters that we must estimate. Indeed, Ray & 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 dataset they consider had a total of only 32 infected individuals).

Finally, when an infectious individual can no longer transmit the disease (e.g. because of 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 infectious 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 ($\beta, k_E, \theta_E, k_I, \theta_I$). While we are not aware of any statistical or probabilistic analyses of this model in the literature, there have been several studies that simulate SEIR outbreaks on Erdős-Rényi networks (Rahmandad & Sterman, 2008; Kenah, 2009).

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 infectious times for each node, but in general, these will be unknown. The exposure, infectious and removal times for node j are denoted by E_j, I_j and R_j, respectively. The sets of all exposure, infectious and removal 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 by κ; since the identity of this individual will not in general be known, κ may be considered a parameter to estimate. 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 exposed, that is, $\mathbf{E}\backslash E_\kappa$. 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 a generic parameter (say, δ) by $\pi_\delta(\cdot)$, the likelihood function by $L(\mathbf{T} \mid \delta, \ldots)$, and the posterior distribution of δ by $\pi_\delta(\cdot \mid \mathbf{T})$.

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 $\mathcal{P}$. $\mathcal{P}$, whose root node is κ, 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 Britton & O'Neill (2002).

Figure 1A shows an example of a contact network $\mathcal{G}$ within a population of $N=25$ individuals. This network was simulated using an Erdős-Rényi model with $p=0.15$. Superimposed on this contact network in Fig. 1B 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

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

$$\begin{aligned} L(\mathbf{T} \mid \beta, k_E, \theta_E, k_I, \theta_I, p) &= \sum_{\mathcal{G}, \mathcal{P}} L(\mathbf{T} \mid \beta, k_E, \theta_E, k_I, \theta_I, p, \mathcal{G}, \mathcal{P}) f(\mathcal{G}, \mathcal{P} \mid p) \\ &= \sum_{\mathcal{G}} \sum_{\mathcal{P}} L(\mathbf{T} \mid \beta, k_E, \theta_E, k_I, \theta_I, p, \mathcal{G}, \mathcal{P}) f(\mathcal{P} \mid \mathcal{G}) f(\mathcal{G} \mid p). \end{aligned}$$

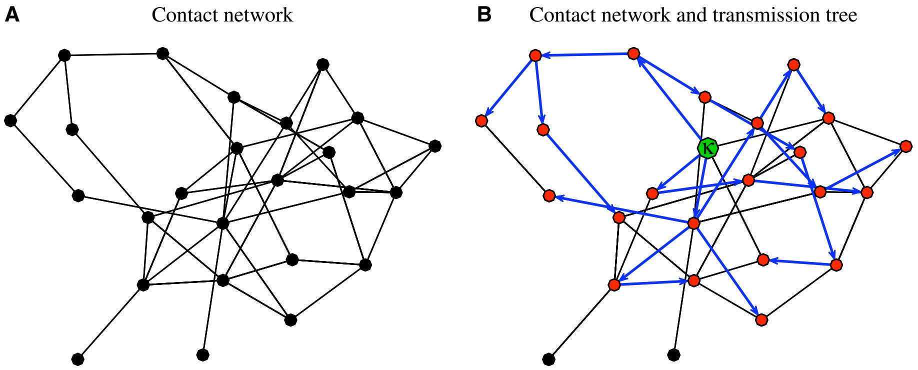

Fig. 1. (A) A realization of an Erdős-Rényi contact network ($\mathcal{G}$), representing the contacts between the individuals in the population. (B) A realization of a simulated SEIR epidemic across the network. The large green node represents the initial exposed individual (κ). Red nodes represent 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. $\mathcal{G}\backslash\mathcal{P}$).

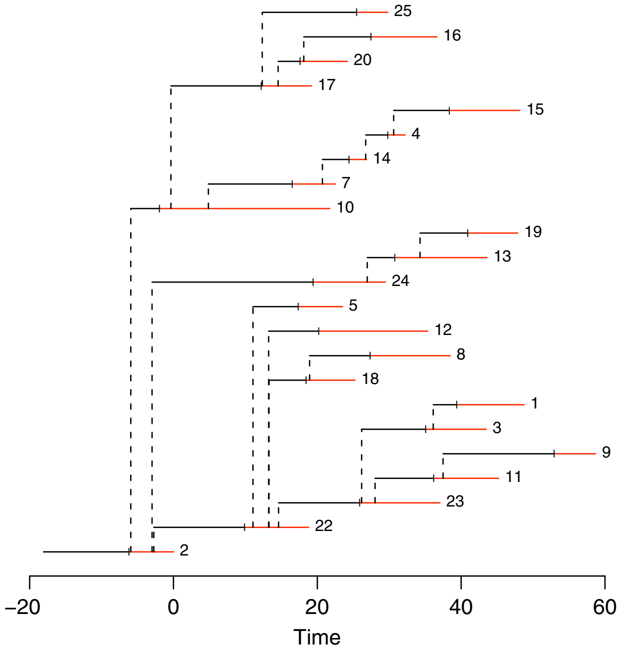

Fig. 2. Progression of the SEIR epidemic through time, as produced by function **plotepitree()** in **R** package **epinet**. Vertical dashed line segments show the infection pathway. Horizontal solid line segments show the time periods that the individuals were in the exposed (black) and infectious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

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

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

We can calculate the likelihood as $L_1L_2L_3L_4$, where L_1 is the contribution to the 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 ($\mathcal{G}\backslash\mathcal{P}$) and L_3 and L_4 are the contributions because of the transition (from exposed to infectious) 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

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

so that

$$L_1L_2 = \beta^{m-1} \exp[-\beta A],$$

where

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

Here, $\mathbf{1}(\cdot)$ 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 & Roberts (2005), is the form we use in our algorithm.

L_3 and L_4 are given by

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

and

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

where $B=\sum_{i=1}^{m}(I_i-E_i)$ and $C=\sum_{i=1}^{m}(R_i-I_i)$ are the total amounts of time spent by all individuals in the exposed and infectious states, respectively.

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 infectious times, we assign them uninformative (flat) prior distributions; when necessary, we assign a prior for κ that is uniform on $1,\ldots,m$. 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 performed to study the lengths of times that individuals infected with the measles virus spend in the exposed and infectious 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 next. 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 infectious 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 Britton & O'Neill (2002). 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 & 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 centred 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 $X\sim\text{Gamma}(a,b)$ indicate that X has a Gamma distribution with density $x^{a-1}b^{-a}\,\mathrm{e}^{-x/b}/\Gamma(a)$ for $x>0$; let $W\sim\text{IG}(c,d)$ indicate that W has an inverse Gamma distribution, that is, $1/W\sim\text{Gamma}(c,1/d)$; let $Y\sim\text{beta}(q,z)$ 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: $\pi_\beta(\beta)\sim\text{Gamma}(a_\beta,b_\beta)$, $\pi_{\theta_I}(\theta_I)\sim\text{IG}(a_I,b_I)$, $\pi_{\theta_E}(\theta_E)\sim\text{IG}(a_E,b_E)$ and $\pi_p(p)\sim\text{beta}(c,d)$, then the corresponding full conditional distributions of these parameters are:

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

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 to sample from the full conditional distribution of $\mathcal{G}$. We calculate the full conditional probability of the event $\{i,j\} \in \mathcal{G}$, which we denote by D_{ij}, assuming without loss of generality that $E_i < E_j$:

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

where $P(D_{ij} \mid \mathcal{P}, \beta, p) = p$, unless the edge (i,j) is in $\mathcal{P}$, in which case $P(D_{ij} \mid \mathcal{P}, \beta, p) = 1$. The values of $P(\mathbf{T} \mid D_{ij}, \mathcal{P}, \beta, p)$ and $P(\mathbf{T} \mid D_{ij}^c, \mathcal{P}, \beta, p)$ 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 $P(D_{ij} \mid \mathbf{T}, \mathcal{P}, \beta, p) = 1$, since $(i,j) \in \mathcal{P} \Rightarrow \{i,j\} \in \mathcal{G}$. Otherwise, if $(i,j) \notin \mathcal{P}$, then

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

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 aforesaid 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

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

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) 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 **E** and **I** are not fully known. To perform this update, we use a method similar to that described by Britton & O'Neill (2002). The typical prior assumption is that each of the m infected nodes is equally likely to be the initial infected, that is, $\pi_\kappa(i) = 1/m$ 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 **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 **E**. For each $j \neq \kappa$, we first identify the parent of j in $\mathcal{P}$, that is, the node that infected j. Since i must have been infectious (and not yet recovered) when j became exposed, and since j enters the exposed phase before the infectious 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 **I** is similar to that for **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, 2010), named **epinet**, containing software which implements the algorithm described before. This software is publicly available on the Comprehensive R Archive Network (CRAN; 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 *et al.*, 2010), 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 Britton & O'Neill (2002), rather than the algorithm detailed in Neal & Roberts (2005). The first reason is scalability. As noted by Neal & Roberts (2005), it is not necessary to know the entirety of the graph $\mathcal{G}$ to calculate the likelihood for our model. Neal & Roberts (2005) propose 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 $N \gg m$, that is, when only a small portion of a population is infected, this savings in computing resources may be significant.

In our experience, the algorithm described by Neal & Roberts (2005) ran significantly more slowly than did the algorithm described in section 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 situation will only worsen 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 parameters using the algorithm described before. Our primary interest here is 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 Britton & O'Neill (2002), the rapid spread of an epidemic throughout a population could be owing 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 10 different simulated Erdős-Rényi networks of $N = 40$ individuals with $p = 0.1, 0.2, \ldots, 1$. Over each of these 10 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 before with full data, assuming that the exposure, infectious and removal times were all known. We chose uniform priors for each parameter, specifically $\pi_\beta \sim \mathcal{U}(0, 1.5), \pi_p \sim$ 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 gives 50 and 90 per cent 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(β) 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.

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

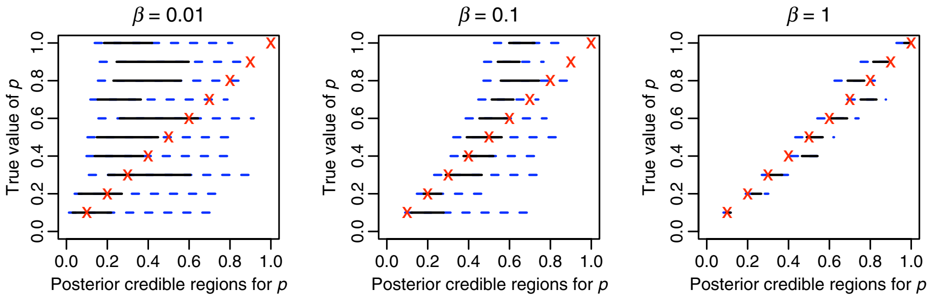

Fig. 3. Fifty per cent (black, solid) and 90 per cent (blue, dashed) posterior credible regions for p, for 10 different simulated Erdős-Rényi networks with $p=0.1, 0.2, \ldots, 1$, and simulated SEIR epidemics with $\beta=0.01$ (left), $\beta=0.1$ (centre), and $\beta=1$ (right).

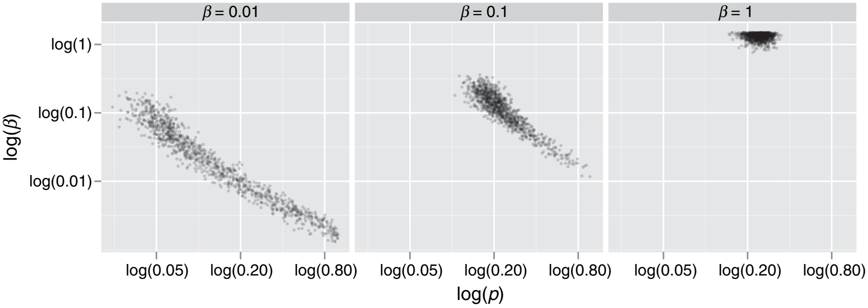

Fig. 4. Scatterplots of $\log(p)$ versus $\log(\beta)$ for a simulated Erdős-Rényi network with $p=0.2$ and three different simulated SEIR epidemics with $\beta=0.01$, 0.1 and 1. Each plot shows 1000 points sampled from the respective posterior samples for these parameters. Posterior correlations between $\log(p)$ and $\log(\beta)$ were estimated at -0.97, -0.90 and -0.05 for $\beta=0.01$, 0.1 and 1, respectively.

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 dataset, which contains data on 188 infected individuals, is notable for its completeness and depth of data (Pfeilsticker, 1863; Oesterle, 1992; Neal & Roberts, 2004). It contains for each individual the timing of the onset of various disease symptoms, and also identifies several other demographic variables for each individual, such as their school class, most likely source of infection and household information (including the identities of siblings and spatial location of the house). We do not attempt to include all these data in the present analysis, though in section 6.1 we consider extensions of the network model that would allow us to incorporate these types of nodal and dyadic covariates. Neal & Roberts (2004) and Britton *et al.*, (2010) analyse these data using more of the covariate information, but they do so in a non-network setting. We assume that the initial size of the susceptible population is the same as the number of individuals who were ultimately infected.

This is based on the fact that all of the infected were children born after the previous outbreak in Hagelloch and nearly all such children were infected. Detailed justification of this assumption is given in Neal & Roberts (2004). We assume that each individual's infectious period begins 1 day prior to the onset of prodromes and ends 3 days after the appearance of rash (or at death, if sooner). As the data do not contain any information about the exposure times of the individuals (**E**), we will treat these times as unknown and infer them.

We initially performed inference on this dataset 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 dataset who is recorded as showing symptoms of the disease approximately 1 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 5000 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 indicate 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 (where 'degree' refers to the number of contacts for an individual in a network) 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 infectious individual has in a totally susceptible population (Keeling & Rohani, 2008). Analogous to the expression derived by Britton & O'Neill (2002), 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 infectious state and Y is the time to the individual's removal after becoming infectious. 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

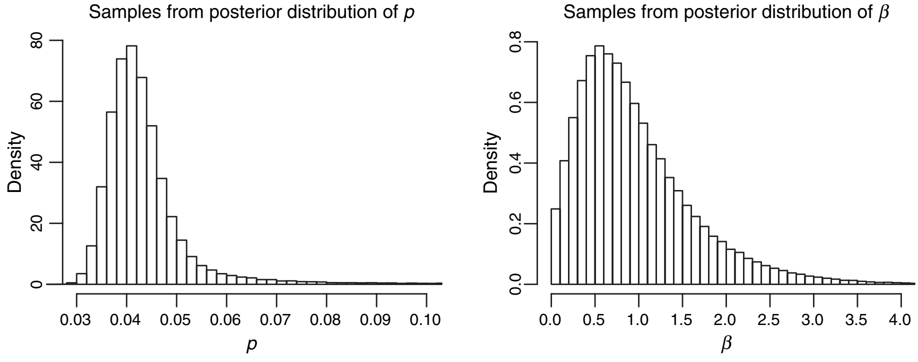

Fig. 5. Histogram of values sampled from the posterior density for parameters p and β, excluding the outlying data point and using conjugate prior distributions.

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

Note that this expression reduces to the formula given by Britton & O'Neill (2002) in the case that the length of the infectious periods are modelled by an exponential random variable, that is, $k_I = 1$. For the values of the β, θ_I and k_I that we consider in this example, (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 (3) using the posterior parameter samples for β, θ_I and k_I produced by the Hagelloch data analysis. The posterior mean and median are 7.7 and 7.6, respectively, and a 95 per cent posterior credible interval for R_0 is (6.4, 9.2). This is comparable with figures reported in the literature on measles. For instance, Huang (2008) gives a range of 5.8–14.3; and Edmunds *et al.* (2001) give a range of 6.1–10.2 for different outbreaks, using data that while much later than the Hagelloch data are still from prevaccination 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 per cent decrease in runtimes for the cases where the conjugate prior distributions were used; this was because of 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 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 per cent. 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

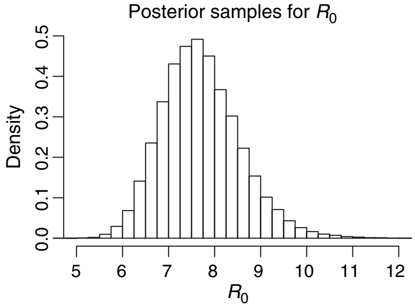

Fig. 6. Posterior samples of the basic reproduction number R_0 for the Hagelloch measles data, calculated using (3) using conjugate prior distributions and excluding the outlier.

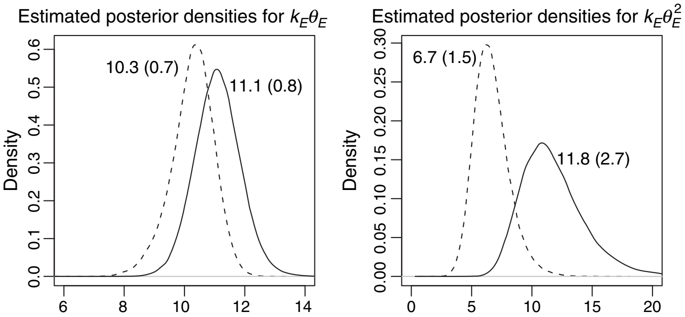

Fig. 7. Estimated posterior densities for the mean ($k_E\theta_E$, left panel) and variance ($k_E\theta_E^2$, right panel) of the length of the inferred exposed periods. The solid lines represent estimates based on all data points, while the dashed lines indicate estimates excluding the outlier. The 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.

have estimated the length of this period to be between 6 and 10.3 days for measles (Gough, 1977; Anderson & May, 1982; Schenzle, 1984).

Posterior predictive modelling

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. We then compared these simulated epidemics with the original Hagelloch measles epidemic data. One statistic of interest is the number of individuals in the infectious stage of the disease, as a function of time. Figure 8 gives a comparison of these epidemic curves for the simulations as compared with 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

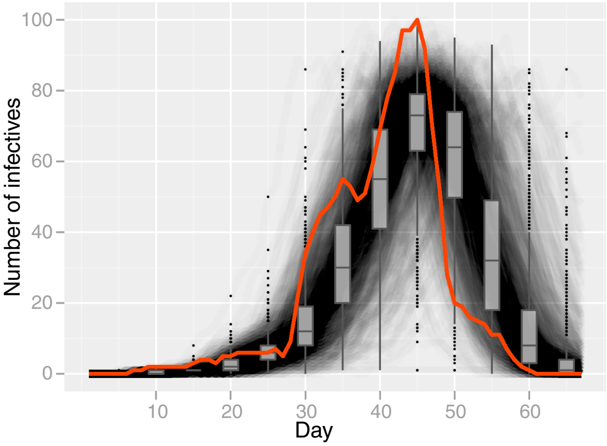

Fig. 8. Comparison of the number of individuals in the infectious stage of the epidemic, as a function of time. The individual epidemics simulated from samples taken from the joint posterior distribution of the parameters are traced in grey, while the original Hagelloch measles data is plotted in red. Summaries (in the form of boxplots) of the number of infectious individuals from across the 1000 simulations are given at intervals of 5 days.

used here: an Erdős-Rényi model may capture the correct mean degree of a network, but the degree distribution itself is fully determined 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 to capture effects because of household and classroom, as Neal & Roberts (2004) and Britton *et al.*, (2010) 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.

Incorporating additional information

We next consider a situation in which the data provide additional information beyond the infectious and removal times considered before. In particular, the Hagelloch dataset contains information about the actual transmission tree for this epidemic. For each individual, a ‘putative parent’ is given, that is, the data contain an indication of who is the most likely to have infected each individual (Oesterle, 1992). We use this information to construct a more informative prior distribution for the transmission tree $\mathcal{P}$. Rather than assuming a uniform prior for each node j, that is, $\pi_{\mathcal{P}_j}(r)$ 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 previous algorithm 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 eight 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.

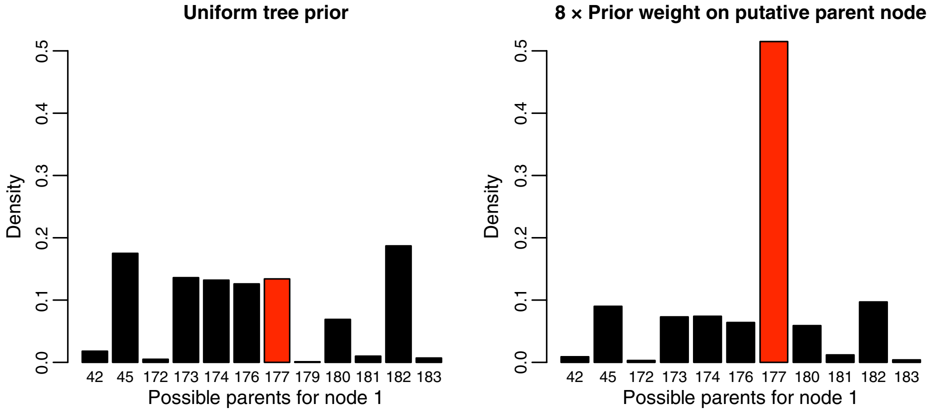

Fig. 9. Barplots of the samples for the posterior distribution of the parent of node 1 under a uniform prior assumption (left panel) and an assumption that places eight times as much prior mass on the putative parent node (right panel) as the other nodes. In the Hagelloch dataset, node 177 (highlighted 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 before, 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 infectious 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 article, we have extended the framework established by Britton & O'Neill (2002) by using a more flexible epidemic model and providing efficient computer code to run experiments. We ran experiments on simulated datasets to determine which areas of the parameter space are most likely to provide meaningful results and found that we were able to make good parameter estimates for a range of biologically plausible values. We have also performed inference for the network and epidemic parameters for an actual dataset 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 Britton & O'Neill (2002) is viable for datasets much larger than those considered previously in the literature (Britton & O'Neill, 2002; Ray & Marzouk, 2008). It suggests that extensions of the model should be explored (see section 6.1); 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 because of 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; in fact, for some applications, the samples of the posterior distributions of $\mathcal{P}$ (and perhaps also $\mathcal{G}$) that are produced by the MCMC algorithm may themselves be 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 aforegiven model 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 before. This broader class of models would allow for the inclusion of additional types of information that might be present in the data, as is the case for the Hagelloch dataset described in section 5.2, which includes several individual- and dyadic-level covariates.

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

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

We will have to modify the MCMC algorithm described in section 4 to reflect the more general ERGM case. For instance, updating the $\boldsymbol{\eta}$ parameter via a Metropolis–Hastings algorithm would be more difficult, since the Hastings ratio involves the ratio of ERGM normalizing constants for the current parameter value $\boldsymbol{\eta}^{(0)}$ and the new proposal $\boldsymbol{\eta}^*$. 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 (Snijders, 2002; Hunter *et al.*, 2008a). Hence, more complicated updating and estimation schemes such as those described by Snijders (2002) or Hunter *et al.* (2008b) may be necessary. Nonetheless, certain types of ERGMs, called dyadic independence models (Hunter *et al.*, 2008a), avoid the difficulties of estimating the ratio of normalizing constants while still incorporating useful statistics, such as geographic data on the individual nodes.

Acknowledgements

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. The authors also thank two anonymous reviewers for their helpful suggestions, which improved the quality of the article. This work was supported by a grant from the National Institutes of Health.

References

Anderson, R. & May, R. (1982). Directly transmitted infections diseases: control by vaccination. *Science* **215**, 1053–1060.

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

Barthelemy, M., Barrat, A., Pastor-Satorras, R. & Vespignani, A. (2005). Dynamical patterns of epidemic outbreaks in complex heterogeneous networks. *J. Theor. Biol.* **235**, 275–288.

Britton, T., Kypraios, T. & O'Neill, P. D. (2010). Inference for epidemics with three levels of mixing: methodology and application to a measles outbreak. Submitted.

Britton, T. & O'Neill, P. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scand. J. Statist.* **29**, 375–390.

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

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

Gough, K. (1977). The estimation of latent and infectious periods. *Biometrika* **64**, 559–565.

Handcock, M. S., Hunter, D. R., Butts, C. T., Goodreau, S. M., Morris, M. & Krivitsky, P. (2010). *ergm: a package to fit simulate and diagnose exponential-family models for networks*, version 2.2–3. Seattle, WA. Available on http://CRAN.R-project.org/package=ergm (accessed October 28, 2010).

Huang, S. (2008). A new SEIR epidemic model with applications to the theory of eradication and control of diseases, and to the calculation of R_0. *Math. Biosci.* **215**, 84–104.

Hunter, D. R., Goodreau, S. M. & Handcock, M. S. (2008a). Goodness of fit for social network models. *J. Amer. Statist. Assoc.* **103**, 248–258.

Hunter, D. R., Handcock, M. S., Butts, C. T., Goodreau, S. M. & Morris, M. (2008b). ergm: a package to fit simulate and diagnose exponential-family models for networks. *J. Stat. Softw.* **24**, 1–29.

Keeling, M. & Eames, K. (2005). Networks and epidemic models. *J. Roy. Soc. Interface* **2**, 295–307.

Keeling, M. & Rohani, P. (2008). Modeling infectious diseases in humans and animals. *Clin. Infect. Dis.* **47**, 864–866.

Kenah, E. (2009). Contact intervals, survival analysis of epidemic data, and estimation of R_0. *Arxiv preprint arXiv:0912.3330*.

Kermack, W. & McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* **115**, 700–721.

Meyers, L., Pourbohloul, B., Newman, M., Skowronski, D. & Brunham, R. (2005). Network theory and SARS: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81.

Neal, P. & Roberts, G. (2004). Statistical inference and model selection for the 1861 Hagelloch measles epidemic. *Biostatistics* **5**, 249–261.

Neal, P. & Roberts, G. (2005). A case study in non-centering for data augmentation: stochastic epidemics. *Statist. Comput.* **15**, 315–327.

Oesterle, H. (1992). Statistiche Reanalyse einer Masernepidemiie 1861 in Hagelloch, MD thesis, Eberhard-Karls Universität, Tübingen.

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

R Development Core Team (2010). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria; ISBN 3-900051-07-0.

Rahmandad, H. & Sterman, J. (2008). Heterogeneity and network structure in the dynamics of diffusion: comparing agent-based and differential equation models. *Manage. Sci.* **54**, 998–1014.

Ray, J. & Marzouk, Y. (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*, 3–7 August 2008. American Statistical Association, Alexandria, VA, USA.

Schenzle, D. (1984). An age-structured model of pre- and post-vaccination measles transmission. *Math. Med. Biol.* **1**, 169–191.

Snijders, T. A. B. (2002). Markov chain Monte Carlo estimation of exponential random graph models. *J. Soc. Struct.* **3**, 1–40.

Volz, E. (2008). SIR dynamics in random networks with heterogeneous connectivity. *J. Math. Biol.* **56**, 293–310.

Received April 2010, in final form September 2010

Chris Groendyke, Department of Statistics, Pennsylvania State University, 333 Thomas Building, University Park, PA 16802, USA.
E-mail: cxg928@psu.edu