Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure

TOM BRITTON

Uppsala University

PHILIP D. O'NEILL

University of Nottingham

ABSTRACT. A single-population Markovian stochastic epidemic model is defined so that the underlying social structure of the population is described by a Bernoulli random graph. The parameters of the model govern the rate of infection, the length of the infectious period, and the probability of social contact with another individual in the population. Markov chain Monte Carlo methods are developed to facilitate Bayesian inference for the parameters of both the epidemic model and underlying unknown social structure. The methods are applied in various examples of both illustrative and real-life data, with two different kinds of data structure considered.

Key words: Bayesian inference, epidemics, Markov chain Monte Carlo methods, Metropolis–Hastings algorithm, random graphs, stochastic epidemic models

1. Introduction

This paper is concerned with methodology for performing Bayesian statistical inference for stochastic epidemic models which include a simple kind of underlying unobserved social structure. This topic links two themes in which there is currently a great deal of interest, namely (i) stochastic models for epidemics in structured populations, and (ii) the use of Markov Chain Monte Carlo methods for inference for stochastic epidemic models. Before describing the present work in more detail, we focus briefly on these two areas.

In recent years there has been an increase in research activity regarding stochastic models for epidemics among populations with some kind of social structure. This work is motivated by a desire for model realism, and in particular by the fact that real-life human populations are themselves structured. In some cases, models are designed to include a fixed known social structure, such as simple household models (Becker & Dietz, 1995), or household models with two levels of mixing (Ball *et al*., 1997). An alternative approach is to regard the social structure itself as randomly generated within the model, as in the general social network models described in Andersson (1999). The social structure may also be considered at an individual level, for example with pair-formation models (Altmann, 1998). In the present paper, we will focus on the simplest case of randomly generated social structures, namely that a Bernoulli random graph will be used to describe potential contacts among a population of individuals. This particular model is analysed in Andersson (1998), although no attempt was made there to describe methods of statistical inference.

Undertaking statistical inference based on stochastic epidemic models and data from disease outbreaks is generally a non-standard problem. This is due to both the nature of the data, which is highly dependent and typically only partial, and also to the level of mathematical intractability of even the simplest stochastic epidemic models. Some novel approaches have been developed (Becker, 1989; Becker & Britton, 1999), although these often involve making

unrealistic modelling assumptions, which in turn affects the reliability of the conclusions. Recently, the use of Markov chain Monte Carlo (MCMC) methods has been explored (O'Neill & Roberts, 1999; O'Neill *et al.*, 2000). MCMC methods offer, at least in principle, important advantages over existing methods, most notable of which is the fact that they allow a much greater degree of modelling flexibility. However, the implementation of MCMC methods may be problematic, since algorithm convergence and mixing difficulties can arise due to the amount of missing data and correlation structures inherent within epidemic models. Consequently, algorithms often need to be designed with care.

As described above, in the present paper we shall consider an epidemic model among a population with unobserved social structure assumed to be described by a Bernoulli random graph. The epidemic model, described in detail in the next section, assumes potential infections occurring at the points of a Poisson process; exponentially-distributed infectious periods; no latent periods; and full immunity following the infectious period. It is clear that this model has limited application to the modelling of specific diseases. However, our objective is to develop methods of statistical inference, and it seems sensible to do so with a basic model before moving on to more complex situations. Furthermore, our focus in this paper is towards moderately-sized datasets, for the following reasons. First, real-life datasets of the kind that we shall consider, such as temporal data consisting of case-detection times, usually contain tens rather than hundreds of observations. Second, larger datasets, especially those collected over long periods of time, are often more appropriate for models which allow for interventions or changes in the environment under which the disease spreads. Finally, the design of efficient MCMC algorithms for larger datasets is a separate topic of interest in its own right.

The paper is organized as follows. The model is described in detail in section 2, as are the different kinds of data that we shall consider. In section 3 the likelihood is derived and used to define the posterior density of interest. Section 4 contains a description of an MCMC algorithm that is our main inferential tool. In section 5 we consider a number of examples to illustrate the performance and uses of the MCMC algorithm. We conclude in section 6 with an overview of progress, and suggestions for future work.

2. Model and data

In this section we describe the epidemic model that our analysis is based upon, review some known probabilistic results for this model, and indicate the kind of datasets that we shall consider.

2.1. Modelling assumptions

We shall model the social structure of a closed population using a random graph, G. Specifically, each individual in a closed population will be represented by a vertex in G. Given a particular realization of G, $\mathscr{G}$ say, the adjacency of two vertices represents regular social contact between the two corresponding individuals. Furthermore, a Markovian epidemic process can be defined on $\mathscr{G}$. We now describe the model in more detail.

Let $G = G(N, p)$ be a Bernoulli random graph defined on N labelled vertices, with p the probability that two vertices are joined by an edge, and let $\mathscr{G}$ be a given realization of G. Thus p represents the probability that two individuals have regular social contact with each other, with contacts between different pairs of individuals assumed to be independent of one another.

A Markovian epidemic process can now be defined on $\mathscr{G}$ as follows. Each vertex can be in one of three states, namely susceptible, infective, or removed. The susceptible state corresponds to a healthy individual who can contract the disease in question. The infective

state describes an individual who can pass the disease on to others. Finally the removed state describes a formerly infectious individual who now cannot be reinfected. Initially, there will typically be a small number of infectious individuals among an otherwise wholly susceptible population. Infective individuals remain so for a period of time that is exponentially distributed with mean γ^{-1} before becoming removed, where $\gamma > 0$ is known as the removal rate. During its infectious period, an infective makes infectious contacts with each adjacent susceptible according to a Poisson process of rate $\beta > 0$, where β is known as the infection rate. Each such contact results in the immediate infection of the susceptible in question. The Poisson processes governing different infective–susceptible pairs are assumed to be independent of one another. The epidemic continues until there are no more infectives left in the population.

2.2. *Review of known results*

We now briefly review some properties of the model described above (see Andersson, 1998). It is convenient to reparameterize the model by introducing $\lambda = Np$, where λ is the average number of social contacts of a single individual. Recall that 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. Then R_0 is given by $\lambda\beta/(\beta + \gamma)$ (see Andersson, 1998). This follows because λ is the expected number of social links and $\beta/(\beta + \gamma)$ is the probability of a contact occurring before the individual recovers. Denote by T the total number of infections that occur during the epidemic. As $N \to \infty$ the final proportion infected, T/N, converges in distribution to a 2-point distribution. If $R_0 \leq 1$ all the mass is concentrated at 0 while if $R_0 > 1$ then the limiting distribution also has positive mass at τ, where τ is the unique positive solution to the equation

$$1 - \tau = \exp(-R_0\tau).$$

In the latter case the amount of probability mass at τ can be derived using branching process theory. The initial stages of the epidemic can be approximated by a branching process with growth rate α, so that at (small) time t the number of infectives is approximately $\exp(\alpha t)$. Here α is the so-called Malthusian parameter, which is the solution to the integral equation $\int_0^\infty \exp(-\alpha t)\mu(t)dt = 1$, where $\mu(t)$ is the average rate at which an individual infects others t time units after he was infected. For the model presented above $\mu(t) = \lambda\beta\exp(-(\gamma + \beta)t)$. This is because an individual has λ social links on average, and while infectious will infect any given susceptible at rate β. Solving the equation for this choice of $\mu(t)$ yields that the Malthusian parameter α satisfies

$$\alpha = \beta\lambda - \gamma - \beta = (R_0 - 1)(\gamma + \beta).$$

In this paper we shall consider parameter estimation when the social network is not observed. Estimating γ is straightforward if the lengths of the infectious periods are known. The challenge comes in trying to distinguish β and λ, since different combinations of these parameters may lead to similar outbreak sizes. However, the results described above imply that for fixed R_0 (i.e. fixed expected outbreak size) and fixed γ, the speed at which an epidemic initially spreads will increase as β increases. In fact, as $\beta \to \infty$ the duration of the entire epidemic tends to 0 and all individuals that are socially linked with the initial infective become infected instantaneously. On the other hand, if β is small but λ is larger the final number infected might be the same, but the duration of the epidemic is likely to be longer. These results suggest that estimation of both β and λ is feasible.

If only the removal times are observed, then the heuristic reasoning in the previous paragraph still holds, although now estimation of γ is no longer straightforward. Finally, we note that explicit estimators for β and λ, given the kinds of data considered in the present paper, are not known.

2.3. Data

We will consider two possibilities for the kind of data that are available, namely (I) the times of infection and removal of each individual who ultimately becomes infected, and (II) the removal times only. In both cases the underlying social network is not observed. For simplicity we assume that the epidemic is known to have ceased, so that the number of observed removals equals the number of infections, T. In reality, it is far more likely that data of type (II) rather than (I) are actually observed, especially for human diseases. However, for some applications or for preliminary analyses it may be acceptable to impute the missing infection times, perhaps by simply assuming a fixed-length infectious period, so that the infection times are exactly specified by the removal times.

Our notation is as follows. Define $\mathbf{R} = (R_1, R_2, \ldots, R_m)$, where R_j is the removal time of individual j, $T = m$ is the number of observed removals, and $R_{\min} = \min_{1 \leqslant j \leqslant m}\{R_j\} = 0$, so that $R_{\min}$ has the role of time origin. We also define $\mathbf{I} = (I_1, I_2, \ldots, I_m)$, where I_j is the infection time of individual j. For data of type (II), the unknown infection times will be regarded as extra parameters whose inclusion in the model will greatly simplify the necessary likelihood calculations. The infection times themselves will be imputed as part of the MCMC algorithm described below.

3. Likelihood and posterior density

We start with some notation and definitions. It will be convenient to re-label the vertices of the graph such that vertices 1 to m correspond to the m ultimately infected individuals in the graph, with vertex j being associated with (I_j, R_j) $(j = 1, 2, \ldots, m)$. For convenience we define $I_j = R_j = \infty$ for $j = m+1, \ldots, N$. We shall assume that there is one initial infective, labelled κ, so that $I_\kappa < I_j$ for all $j \neq \kappa$, and we define $\tilde{\mathbf{I}}$ as $\mathbf{I} \setminus I_\kappa$. Note that κ is itself a parameter, and in particular is not simply fixed.

We shall say that $(i, j) \in \mathscr{G}$ if and only if the vertices i and j are adjacent in $\mathscr{G}$, where $1 \leqslant i, j \leqslant N$. We denote by P the random directed tree with labelled vertices whose root is the vertex corresponding to the initial infective, and in which a directed edge from vertex i to vertex j appears if and only if i infects j during the epidemic. Thus P denotes the pathway of infection. We denote a particular realization of P by $\mathscr{P}$, and say that $(i, j) \in \mathscr{P}$ if and only if the directed edge from i to j appears in $\mathscr{P}$. Notice that $(i, j) \in \mathscr{P}$ only if $(i, j) \in \mathscr{G}$ and $I_i < I_j < R_i$, and so there are constraints on the possible values of P. We denote by $|\mathscr{G}|$ and $|\mathscr{P}|$ the number of edges in $\mathscr{G}$ and $\mathscr{P}$, respectively; note that $|\mathscr{G}| \geqslant |\mathscr{P}| = m - 1$. Finally, we use the notation $\pi(\cdot|\cdot)$ to denote conditional densities (or mass functions, when appropriate).

Our objective is to make inferences about the model parameters β, γ and p given the data, which will be either $\mathbf{R}$ or $(\mathbf{I}, \mathbf{R})$. In a Bayesian framework we thus wish to explore the posterior density of the model parameters given the data under the assumption of some prior density $\pi(\beta, \gamma, p)$. However, the likelihood of the data given the parameters involves summation over all possible values of $\mathscr{G}$ and $\mathscr{P}$, and can be tedious to compute. We therefore include $\mathscr{G}$ and $\mathscr{P}$ as extra model parameters, since if they are known then the likelihood becomes far simpler to calculate. An additional benefit, as we shall see, is that certain posterior distributions become considerably simpler.

In fact, since the infection mechanism in our model is defined via Poisson processes, the likelihood will always be independent of the infection pathway $\mathscr{P}$. Specifically, the product L_1L_2, where L_1 and L_2 are defined by (2) and (3) below, can be shown to be independent of $\mathscr{P}$. However, there are certain benefits to retaining $\mathscr{P}$ as a parameter. For instance, keeping track of the value of $\mathscr{P}$ facilitates the implementation of the MCMC algorithm described below. Also, we could in principle consider more generalized infection mechanisms within the framework described here.

Denoting the likelihood by L, we have that

$$L = \pi(\tilde{\mathbf{I}}, \mathbf{R}|\beta, \gamma, \mathscr{G}, \mathscr{P}, p, I_\kappa) = \pi(\tilde{\mathbf{I}}, \mathbf{R}|\beta, \gamma, \mathscr{G}, \mathscr{P}, I_\kappa), \tag{1}$$

since $(\tilde{\mathbf{I}}, \mathbf{R})$ only depends on p via the value of $\mathscr{G}$. Note that the likelihood involves conditioning on the time of the start of the epidemic, I_κ. This is necessary so as to ensure a fixed time reference point, relative to which the density of the other infection and removal times can be calculated. Although L is dependent on several quantities ($\tilde{\mathbf{I}}, \mathbf{R}, \beta$, etc.), we shall often either suppress reference to all quantities, or only refer to some in order to emphasize particular dependencies (e.g. writing $L(\tilde{\mathbf{I}})$).

The likelihood L has three components. First, the contribution from the $m-1$ infections is given by

$$L_1 = \prod_{(j,k)\in\mathscr{P}} \beta \exp(-\beta(I_k - I_j)) = \beta^{m-1} \exp\left(-\beta \sum_{(j,k)\in\mathscr{P}} (I_k - I_j)\right). \tag{2}$$

Arguing in the same way, the contribution from infected individuals who fail to infect at least one of their neighbours in $\mathscr{G}$ is given by

$$L_2 = \exp\left(-\beta \sum_{\substack{1\leqslant j\leqslant m \\ (j,k)\in\mathscr{G}\setminus\mathscr{P}}} \{[(I_k \wedge R_j) - I_j] \vee 0\}\right). \tag{3}$$

Finally, the contribution due to the removal process is

$$L_3 = \gamma^m \exp\left(-\gamma \sum_{j=1}^{m} (R_j - I_j)\right). \tag{4}$$

The likelihood is thus given by

$$L = L_1L_2L_3,$$

and defined as zero for any impossible parameter choices (for example, if the infection times $\mathbf{I}$ are not possible given $\mathbf{R}$).

The posterior density of interest is obtained via Bayes' Theorem as proportional to the product of the likelihood and the prior. We assign independent priors to individual parameters. Thus

$$\pi(\beta, \gamma, p, \mathscr{G}, \mathscr{P}|\mathbf{I}, \mathbf{R}) \propto L\pi(\mathscr{P}|\beta, \gamma, \pi_\kappa, \mathscr{G})\pi(\mathscr{G}|p)\pi(\beta)\pi(\gamma)\pi(p)\pi(I_\kappa)\pi(\kappa), \tag{5}$$

where $\pi(\beta)$ is the prior density of β, etc., and where, since G is a Bernoulli random graph,

$$\pi(\mathscr{G}|p) = p^{|\mathscr{G}|}(1-p)^{\binom{N}{2}-|\mathscr{G}|}. \tag{6}$$

As will be described in the next section, $\pi(\mathscr{P}|\beta, \gamma, \pi_\kappa, \mathscr{G})$ is the uniform distribution on the set of all possible infection pathways. Also, for type II data we shall henceforth assume that the prior

$\pi(\kappa)$ is uniformly distributed. For type I data, both I_κ and κ are non-random, and so the relevant priors can be ignored.

In order to obtain information about the posterior density in (5) we use an MCMC algorithm. The algorithm allows us to generate approximate samples from the posterior density, and in particular from the marginal distributions of the parameters of interest. This is achieved by constructing a Markov chain whose stationary distribution is the same as the posterior distribution of interest. We now describe the algorithm in more detail.

4. MCMC algorithm

We now describe a single-component Metropolis–Hastings algorithm (see Gilks *et al*., 1996) to generate approximate samples from the posterior density. With the exception of the infection times and κ, we shall be able to sample each parameter directly from its full conditional distribution. Thus in the case where the data are given by (**I**, **R**), the MCMC algorithm will be a Gibbs sampler. In the following we denote by $\text{Gam}(\mu, \lambda)$ a Gamma-distributed random variable with density

$$f(x) = \frac{(\lambda x)^{\mu-1}\lambda \exp(-\lambda x)}{\Gamma(\mu)} \quad (x \geq 0),$$

and denote by $\text{Beta}(c, d)$ a Beta-distributed random variable with density

$$f(x) = \frac{x^{c-1}(1-x)^{d-1}}{B(c,d)} \quad (0 \leq x \leq 1).$$

Also, we write $\pi(\eta|\cdots)$ for the marginal density of a parameter η conditional upon the data and all other parameters.

Sampling β: From (2) and (3) it is immediate that

$$\pi(\beta|\cdots) \propto \pi(\beta)\beta^{m-1}\exp(-\beta A),$$

where

$$A = \sum_{(j,k)\in\mathscr{P}} (I_k - I_j) + \sum_{\substack{1\leqslant j\leqslant m \\ (j,k)\in\mathscr{G}\setminus\mathscr{P}}} \{[(I_k \wedge R_j) - I_j] \vee 0\}.$$

Note that A is the total time for which susceptibles in the population are exposed to infected individuals. It follows that if β has a $\text{Gam}(\mu_\beta, \lambda_\beta)$ prior, then

$$\pi(\beta|\cdots) \sim \text{Gam}(\mu_\beta + m - 1, \lambda_\beta + A).$$

Sampling γ: From (4) we find that if γ has a $\text{Gam}(\mu_\gamma, \lambda_\gamma)$ prior then

$$\pi(\gamma|\cdots) \sim \text{Gam}(\mu_\gamma + m, \lambda_\gamma + B),$$

where

$$B = \sum_{j=1}^{m} (R_j - I_j).$$

Sampling p: It follows from (6) that if p has a $\text{Beta}(d_1, d_2)$ prior then

$$\pi(p|\cdots) \sim \text{Beta}\left(|\mathscr{G}| + d_1, \binom{N}{2} - |\mathscr{G}| + d_2\right).$$

Note in particular that the choice $d_1 = d_2 = 1$ gives a Uniform(0,1) prior for p.

Sampling $\mathscr{G}$: In order to produce a realization of G conditional upon the data and all other parameters (including $\mathscr{P}$), it is sufficient to generate edges randomly according to their conditional probabilities of existence. Moreover, the independence properties of the Bernoulli graph mean that the event that any given edge exists is independent of the existence of other edges. We are thus concerned with $\Pr\{(i,j) \in \mathscr{G} | \cdots\} = \alpha_{ij}$, say, where $i, j = 1, \ldots, N$, $i \neq j$. It is straightforward to calculate α_{ij} conditional upon the various possibilities, namely: i infects j or vice-versa; both i and j remain susceptible; at least one of i and j becomes infected but neither infects the other. We thus obtain that

$$\alpha_{ij} = \begin{cases} 1 & \text{if } (i,j) \in \mathscr{P} \text{ or } (j,i) \in \mathscr{P}; \\ p & \text{if } i > m \text{ and } j > m; \\ \dfrac{u_{ij}p}{1 - p + u_{ij}p} & \text{otherwise,} \end{cases}$$

where

$$u_{ij} = \begin{cases} \exp\{-\beta[(R_i \wedge I_j) - I_i]\} & \text{if } 1 \leqslant i \leqslant m \text{ and } I_i < I_j, \\ \exp\{-\beta[(R_j \wedge I_i) - I_j]\} & \text{if } 1 \leqslant j \leqslant m \text{ and } I_j < I_i. \end{cases}$$

Sampling $\mathscr{P}$: In order to obtain a realization of P it is sufficient to consider each of the ultimately infected vertices in turn, and for non-initial infectives decide which other vertex was responsible for their infection. Specifically, consider a vertex j that is ultimately infected, so that $1 \leqslant j \leqslant m$. If j is not an initial infective then the set of vertices that could have infected j (we call such vertices suspects) is $\{i : (i,j) \in \mathscr{G}, I_i < I_j < R_i\}$. Since the infection mechanism is Markovian (with equal rates of infection β from each suspect i to j) it follows that each of the suspects is equally likely to have actually infected j. It is therefore sufficient to sample i uniformly from the set of suspects and then set $(i,j) \in \mathscr{P}$.

The algorithm described thus far is all that is required when the data consist of both infection and removal times. However, for the case where the data only consist of removal times, it is necessary to specify sampling schemes for the infection times $\mathbf{I}$, and the label of the initial infective, κ. In order to improve mixing we shall also describe an extra step which helps to prevent the algorithm from getting stuck in certain regions of the parameter space. In the following we shall write $\mathbf{I}^*$ to denote a new set of infection times proposed by the algorithm.

Sampling $\mathbf{I}$: Since the full conditional density of a given infection time is awkward to compute, we use a Hastings algorithm to update $\mathbf{I}$, as follows. All of the infection times, other than I_κ, are updated, but in a randomly selected order. The infection time for individual j is then updated as follows. Since $j \neq \kappa$, $(i,j) \in \mathscr{P}$ for some i. Then $I_j \in (I_i, \tau_j)$, where $\tau_j = \min(\{R_j\} \cup \{R_i\} \cup \{I_k : (j,k) \in \mathscr{P}\})$. A new infection time, I_j^* say, is then sampled according to a uniform density on the interval (I_i, τ_j), and this new value is accepted with probability

$$\frac{L(\mathbf{I}^*)}{L(\mathbf{I})} \wedge 1.$$

If the new value is not accepted then I_j remains unchanged.

Finally, I_κ is updated as follows. As before $I_\kappa < \tau_\kappa$, where now the R_i term in the definition of τ_κ is ignored, but now no lower bound exists on the possible value of I_κ. In this case a new infection time I_κ^* is sampled by setting $I_\kappa^* = \tau_\kappa - X$, where X is a sample from an exponential density with mean θ^{-1}. The new value is accepted with probability

$$\frac{\pi(I_\kappa^*) L(\mathbf{I}^*) \exp(-\theta(I_\kappa^* - I_\kappa))}{\pi(I_\kappa) L(\mathbf{I})} \wedge 1.$$

If the new value is not accepted then I_κ remains unchanged.

Sampling κ: The method of updating $\mathscr{P}$ does not allow the value of κ to change, since a given set of distinct infection times can only have one initial infective, whose label is by definition equal to κ. In order to allow κ to be updated, we again use a Hastings algorithm. Suppose that infective i is currently the initial infective, so that $\kappa = i$. A new value for κ is proposed as follows. Select an infective j at random satisfying $(i, j) \in \mathscr{P}$. Denoting by $v(i, \mathscr{P})$ the number of individuals that i infects according to $\mathscr{P}$, the probability that j is chosen is $v(i, \mathscr{P})^{-1}$. Next, swap the infection times of individuals i and j, so that now i is infected at time I_j and j is infected at time I_i. Thus the proposed new value of κ is j. In order to maintain a permissible value for $\mathscr{P}$, we propose an updated path $\mathscr{P}^* = (\mathscr{P} \setminus \{(i, j)\}) \cup \{(j, i)\}$. Accept the proposed new values with probability

$$\frac{L(\mathbf{I}^*, \mathscr{P}^*)v(i, \mathscr{P})}{L(\mathbf{I}, \mathscr{P})v(j, \mathscr{P}^*)} \wedge 1, \tag{7}$$

where as before I^* denotes the proposed set of infection times and labels. Note that although a new κ value is proposed, the actual value of the initial infection time is unchanged, and so the acceptance probability (7) does not depend on the prior for I_κ.

Mixing step: The algorithm thus far described can experience poor mixing in practice, for the following reason. Suppose that the current value of G is such that the infected individuals have no contact with those who are uninfected. In this case, if the values of β proposed are sufficiently large, then the current infection times will tend to cluster together near the initial infection time. Consequently, the algorithm will be extremely unlikely to propose values of G that allow uninfected individuals to contact infected ones, since any such proposed network would have very low probability under the assumption of a high β value. The algorithm would thus get stuck in a region where the infection times are very close together, β is large, and the uninfected individuals are unconnected to those who are infected.

It should be noted that this situation is partially a consequence of the fact that the model parameterization permits different explanations of the same outcome. Put crudely, the fact that an individual does not get infected could either be due to the values of the parameters controlling the disease spread, or instead because the individual is not connected to the infected part of the network.

In order to prevent the algorithm getting stuck it is therefore necessary to try and move the chain into different regions of the parameter space. The dependencies between parameters, such as those described above, suggest a blocking approach, so that highly correlated parameters are updated in a single block. We shall consider the block of parameters β, $\mathbf{I}$ and γ. We first propose a new value of $\mathbf{I}$, and then new values of β and γ, with the latter two values being proposed dependent on the first. This dependency in the proposal for β and γ is important since it helps to avoid proposing low-density regions of the parameter space.

The specific method we use is as follows. Let c be drawn from a uniform density on the interval $[r^{-1}, r]$, where $r > 1$ is constant. The value c is then used to rescale the current set of infection times so that $I_{\max} = \max_{1 \leqslant j \leqslant m}\{I_j\}$ remains fixed but the interval $I_{\max} - I_\kappa$ is scaled by c. Precisely, for $1 \leqslant j \leqslant m$ define

$$I_j^* = (1 - c)I_{\max} + cI_j.$$

Next, proposed values of β and γ are given by $\beta^* = \beta/c$ and $\gamma^* = \gamma/c$, respectively. Finally the proposed new values are accepted with probability

$$\frac{\pi(I_\kappa^*)\pi(\beta^*)\pi(\gamma^*)L(\beta^*, \gamma^*, \mathbf{I}^*)}{\pi(I_\kappa)\pi(\beta)\pi(\gamma)L(\beta, \gamma, \mathbf{I})} \wedge 1;$$

note that the likelihood term $L(\beta^*, \gamma^*, \mathbf{I}^*)$ is zero if $\mathbf{I}^*$ is incompatible with $\mathbf{R}$ (for example, if I_κ is not negative).

The rationale for rescaling the infection times is as follows. Roughly speaking, the lack of detail in the data makes it hard to distinguish between long infection periods with low β values and short infection periods with high β values. Consequently, a given configuration is likely to have a similar posterior density to a configuration in which the inter-infection times are either increased or decreased a little, with β and γ being updated in sympathy with the rescaling. In particular, increasing the infection times causes β to be reduced, which in turn makes the algorithm more likely to visit regions of the parameter space in which uninfected individuals may be connected to the infected part of the network.

The MCMC algorithm now proceeds as follows. Initial values for all unknown parameters are assigned, and prior distributions of the parameters chosen. Then, each parameter is updated in turn according to the schemes described in the preceeding paragraphs, with the current values of the other parameters being used in the conditioning. One entire update of all parameters, and the mixing step, is collectively known as a sweep. The process continues for a number of sweeps known as the burn-in period. The purpose of the burn-in period is to allow convergence of the Markov chain constructed by the algorithm. After the burn-in period, the parameters are sampled at regular intervals. These samples are, at least approximately, samples from the required posterior distribution.

Regarding convergence of the MCMC algorithm, diagnosis was performed informally, by monitoring the sample output chains of the parameters of interest and other relevant quantities. Examples of the latter include infection times in the case where they are not specified by the data, and certain network summaries such as the quantity S defined in example 1 below.

5. Examples

In the examples described below, the prior for p was always a Uniform (0, 1) distribution. Where required, the prior for I_κ was taken as the improper uniform distribution on $(-\infty, 0)$, and the value of θ used in the proposal density for I_κ was set at 0.5. The prior distributions for β and γ are described in each example. Note that the priors on p, β and γ induce a prior on R_0, although this will not in general have a standard distribution. In theory it would be possible to focus attention on choosing a prior for R_0, rather than its constituent parameters, although we do not consider this here. Finally, the value of the mixing parameter c was set on the basis of experimentation and exploratory initial runs, with a value chosen that appeared to allow reasonable mixing.

In each example, we present our results in terms of the model parameters β and p, and γ if appropriate. For the final two examples, with real outbreak data, we also give results concerning R_0. In general, sample-based estimates of any function of the model parameters can be easily obtained via the output of the simulated Markov chain.

Example 1. Test data, infection times unknown. We begin with a simple example that illustrates the behaviour of the algorithm. The data consist of the set of removal times $\{0, 1, 1\}$, and we set $m = 3$ and $N = 4$. Prior parameters were $\lambda_\beta = \lambda_\gamma = 0.001$, and $\mu_\beta = \mu_\gamma = 1$. It is far from straightforward to calculate posterior summary statistics exactly using numerical integration, primarily because this involves integrating over six variables (β, γ, p and the three unknown infection times). Attempts to perform the required calculations using MAPLE proved unsuccessful.

Table 1. *Posterior parameter summaries from MCMC algorithm using simple removal-times dataset, example* 1

	Parameter		
	β	γ	p
Mean	24.7	0.59	0.45
S.D.	95.2	0.52	0.21

Table 1 contains the means and standard deviations from the MCMC algorithm output. Posterior density estimates for β, γ and p were all found to be unimodal, with those for γ and p being approximately symmetric and that for β being right-skewed. However, it is also informative here to consider the inter-relationships between parameters, including the graph parameter $\mathscr{G}$. Regarding the latter, we now define a new quantity that will summarize an important aspect of $\mathscr{G}$.

Define S as the number of edges in $\mathscr{G}$ that are linked to the vertex corresponding to the uninfected individual. Thus, S may take the values 0, 1, 2 and 3. The reason that S is of interest is that it is related to β. Specifically, suppose the uninfected individual is labelled U. If $S > 0$ then there is at least one infective who fails to infect U. Thus, crudely, β needs to be large enough to allow two infections, but small enough so that it is not improbable that a third will not occur. In contrast, if $S = 0$ then the data, augmented by S, contain no information about the probability of an individual avoiding infection, and consequently the only constraint on β is that it should be large enough to permit two infections. Consequently, when $S = 0$ it is plausible that β can take values that are considerably larger than when $S > 0$. This effect can be seen in Fig. 1, where the pairwise scatterplots of β and p indicate that large β values only occur when $S = 0$.

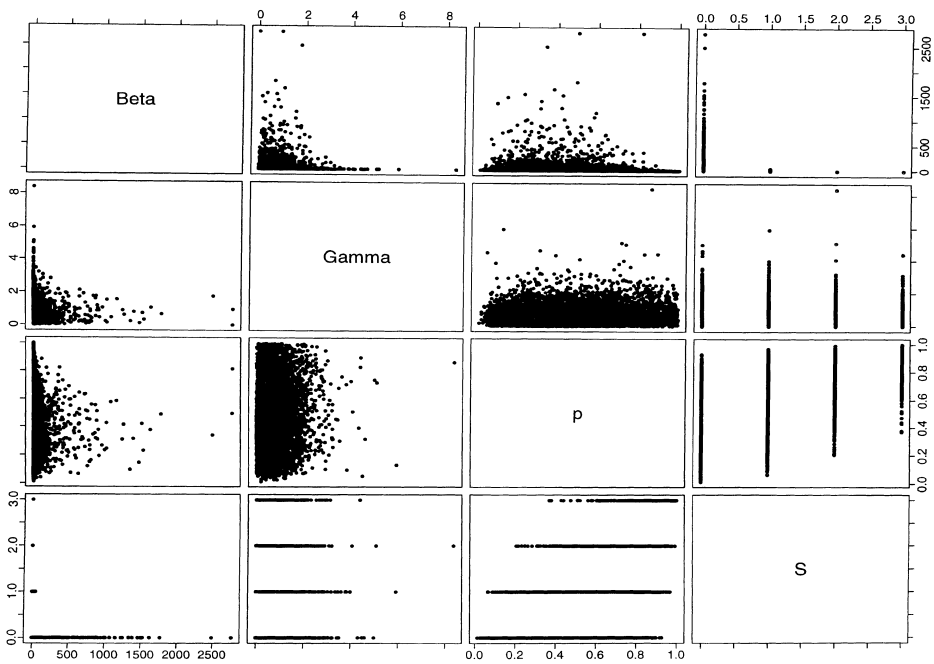

Fig. 1. Pair-wise scatterplots of β, γ, p and S, example 1.

Example 2. Distinguishing between different epidemics. We now show how our methodology can be used, at least informally, to distinguish between data from two epidemics with different parameter values. Unlike the previous example, the infection times are supposed known. As outlined in section 2.2, increasing the value of the infection rate β is likely to result in a faster-spreading epidemic. Two datasets, A and B (see Table 2), were constructed so that infections occurred more quickly in dataset B than A, suggestive of a higher β value in B, but where the final size of each outbreak was the same, namely $m = 15$. Each dataset consists of ordered pairs (I_j, R_j), for $j = 1, \ldots, 15$. Additionally, in the obvious notation, $R_j^{\mathrm{A}} - I_j^{\mathrm{A}} = R_j^{\mathrm{B}} - I_j^{\mathrm{B}}$ for each $j = 1, \ldots, 15$, so that the two datasets contain identical inferential information for γ. The value of N was set at 20.

Prior parameter values for β were chosen to be fairly uninformative; specifically, we set $\lambda_\beta = \mu_\beta = 0.001$, so that β has prior mean 1, and standard deviation $\sqrt{1000}$. Some results for the MCMC output for β and p are given in Table 3. The β values for the two datasets are markedly different, providing evidence that the β values underlying the two datasets are not the same, with the β value for dataset B appearing to be largest. Posterior density estimates for β and p were both found to be unimodal, with similar shapes to those in example 1.

Figure 2 shows a scatterplot that illustrates the relationship between β and p for dataset A; a very similar-looking plot is obtained by using dataset B instead. As can be seen, there is a clear correlation structure, so that as p decreases, β increases. This can essentially be interpreted as saying that the data could have arisen from a highly connected network with low infection rates, from a more sparse network with higher infection rates, or from intermediate situations.

Table 2. *Datasets A and B as used in example* 2

(I_j, R_j)	
Dataset A	Dataset B
(−1.2, 0.1)	(−0.9, 0.4)
(−0.7, 0)	(−0.7, 0)
(−0.1, 1.0)	(−0.5, 0.6)
(0.1, 0.9)	(−0.3, 0.5)
(0.3, 1.2)	(−0.2, 0.7)
(0.5, 1.6)	(−0.1, 1.0)
(0.6, 1.9)	(−0.1, 1.2)
(0.8, 1.1)	(0.0, 0.3)
(0.9, 2.1)	(0.2, 1.4)
(1.2, 2.1)	(0.3, 1.2)
(1.4, 2.6)	(0.4, 1.6)
(1.8, 2.8)	(0.5, 1.5)
(2.1, 2.9)	(0.6, 1.4)
(2.6, 3.5)	(0.7, 1.6)
(3.0, 3.8)	(0.9, 1.7)

Table 3. *Posterior parameter summaries for datasets A and B, example* 2

	Dataset A		Dataset B	
	β	p	β	p
Mean	0.27	0.55	0.46	0.51
Median	0.17	0.54	0.23	0.49
S.D.	0.28	0.26	0.62	0.27
Equal-tailed 95% C.I.	(0.077, 0.83)	(0.15, 0.95)	(0.093, 1.67)	(0.12, 0.95)

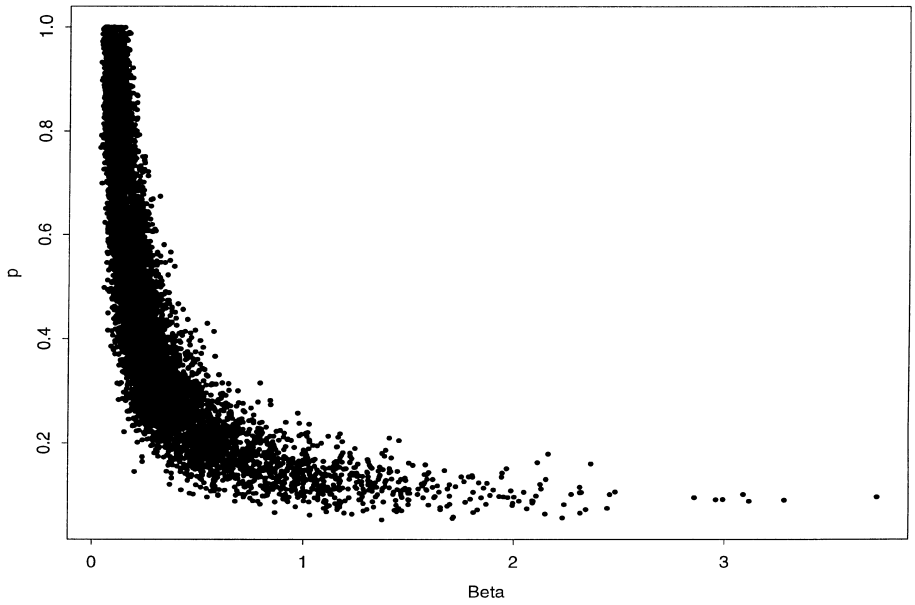

Fig. 2. Scatterplot of β and p, dataset A, example 2.

Table 4. *Posterior parameter summaries for datasets A and B with infection times deleted, example* 2

	Dataset A			Dataset B		
	β	γ	p	β	γ	p
Mean	0.45	1.17	0.52	1.74	1.43	0.36
Median	0.24	1.13	0.51	0.89	1.40	0.24
S.D.	0.55	0.31	0.27	2.01	0.37	0.27
95% C.I.	(0.096, 1.52)	(0.71, 1.73)	(0.13, 0.96)	(0.16, 5.9)	(0.87, 2.10)	(0.085, 0.90)

Next, the infection times in datasets A and B were deleted, and the MCMC algorithm applied to the two resulting removal times datasets. The prior density for γ was the same as that for β.

Table 4 contains summary results from the MCMC output. As can be seen there is clear evidence to suggest that the value of β for dataset A is smaller than that for dataset B. Although the difference in posterior values of β between the two datasets seems considerably larger than the corresponding difference when the infection times are known, it should also be noted that in this case the posterior values for p are rather different. In particular, the posterior mean and median of p are both markedly smaller for dataset B than for dataset A. Since β and p are related in a manner like that illustrated in Fig. 2, it seems reasonable to consider the value of βp as providing a crude comparison of the two datasets. However, we still find that βp has larger posterior values for dataset A. This suggests that the MCMC algorithm can still be used to distinguish between different β values, even in the absence of infection-time data.

Example 3. Gastroenteritis outbreak data. Our next example concerns an outbreak of gastroenteritis in a hospital ward in South Carolina, January 1996, as reported in Cáceres *et al.* (1998). Although viruses that cause gastroenteritis are commonly transmitted through contaminated food, on this occasion person-to-person spread was believed to have occurred.

Table 5. *Detection times of cases of gastroenteritis, example* 3

Day	0	1	2	3	4	5	6	7
Cases	1	0	4	2	3	3	10	5

Data were collected on the date of onset of symptoms for the 28 cases among 89 members of staff working on the ward during the study period, as well as 10 cases among 91 patients who were hospitalized for more than one day during the outbreak. Since the patient population was not closed, and only 10 patient cases occurred, we shall for simplicity restrict attention to the cases among staff members. The staff case data are given in Table 5. On the final day on which cases were recorded, the hospital ward was closed to new admissions, and no more cases occurred.

In order to perform inference for these data using our model we must, as with any form of modelling, make certain assumptions. However, our main purpose here is to illustrate methodology rather than perform a careful data analysis, and so we will be tolerant towards some of the less realistic assumptions. In addition to the fact that we have ignored cases among patients, our model takes no account of an incubation period, which for viral gastroenteritis is between 1 and 3 days (Benenson, 1990). The fact that the ward was eventually closed to the admission of new patients seems likely to have had some effect on the course of the epidemic, despite our only considering the epidemic among staff. Finally, there are also implicit assumptions associated with the model, such as the Bernoulli random graph social structure and exponentially-distributed infectious periods.

Prior distributions for β, γ and p were the same as for example 2. Posterior density summaries are given in Table 6, including information for the basic reproduction number $R_0 = Np\beta/(\beta + \gamma)$. Regarding the three basic model parameters, the marginal posterior densities for γ and p were reasonably symmetric, and the density for β right-skewed but fairly sharply peaked. Fig. 3 contains pairwise scatterplots for β, γ and p.

Regarding R_0, its marginal posterior density was found to be unimodal, with mean 1.17. As a very rough comparison, a martingale-based estimator of R_0 described in Becker (1989, p. 149) based only on the number infected (28) and the population size (89), and assuming homogeneous mixing ($p = 1$ in our framework) estimates R_0 as about 1.14. Although we would not expect this estimate to be the same as the posterior mean for our model, it is reassuring that they are fairly similar. In our model, the mean and standard deviation of the infectious period are both given by γ^{-1}, and this was found to have posterior mean 0.75 days. Although this seems quite short, it should be noted that here we are modelling the effective infectious period, since it is assumed that case detection corresponds to removal. Also, the posterior summaries for γ give in Table 6 seem compatible with the data at first sight. For example, there is only one day with no cases, which suggests that the unknown infectious periods are unlikely to be very long.

It would appear that the data and prior distributions used do not lead to strong posterior inference for the network parameter p. To investigate this further, alternative prior values

Table 6. *Posterior parameter summaries from MCMC algorithm using gastroenteritis dataset, example* 3

	Parameter			
	β	γ	p	R_0
Mean	0.061	1.47	0.54	1.17
Median	0.035	1.41	0.55	1.14
S.D.	0.0088	0.47	0.27	0.32
Equal-tailed 95% C.I.	(0.015, 0.19)	(0.81, 2.3)	(0.11, 0.96)	(0.73, 1.74)

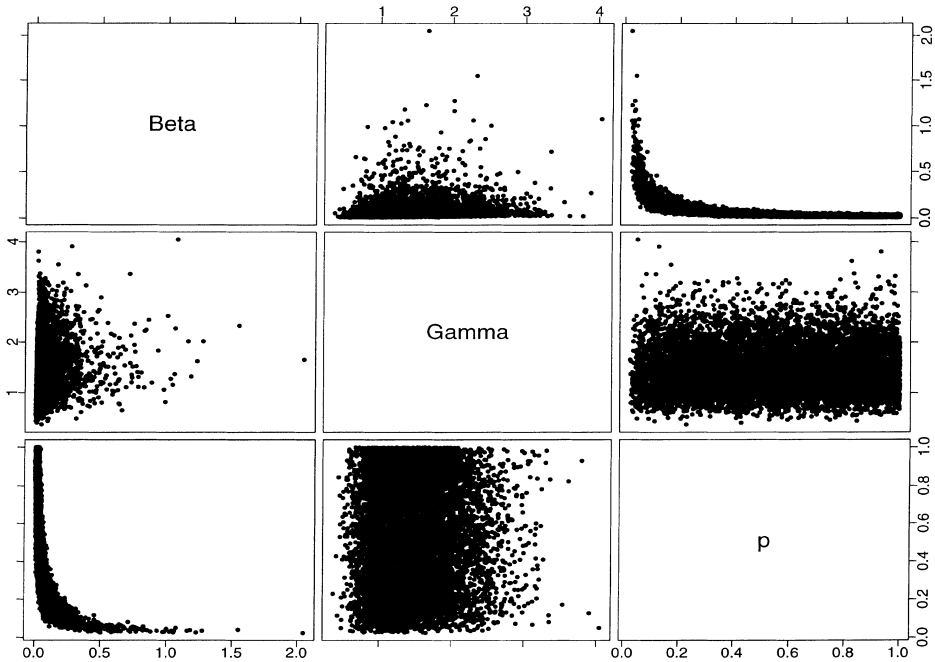

Fig. 3. Pair-wise scatterplots of β, γ and p, example 3.

$d_1 = 1$, $d_2 = 4$ were used instead, so that now p had prior mean 0.2 and standard deviation approximately 0.16. It was found that the resulting posterior density for p was right-skewed with mean 0.25 and median 0.21, while the posterior mean and median for β became 0.15 and 0.094, respectively. The values for γ and R_0 were virtually unchanged, suggesting that the effect of a less vague prior for p is (as expected) to restrict the posterior values of both p and β accordingly, while having less effect on inference for γ and R_0.

Example 4. Shigellosis outbreak data. Our final example is concerned with an outbreak of shigellosis in a shelter for the homeless between 27 December 1991 and 23 January 1992, as reported in L.A.D.H.S. Public Health letter (1992). The spread of disease was believed to have been propagated via person-to-person contact among 199 residents in the shelter, of whom 42 ultimately contracted the disease. The data, consisting of case-detection times, are given in Table 7. As in the previous example, we implicitly make a number of simplifying assumptions by using our model for these data. In particular, we assume that the epidemic ceased on 23 January although in reality a mass vaccination clinic was held on this date, after which no additional cases occurred. However, since the clinic was held on a date several days after the bulk of the recorded cases, our simplifying assumption does not seem too unreasonable. We

Table 7. *Detection times of cases of shigellosis, example* 4

Day	0	1	2	3	4	5	6	7	8	9	10	11	12	13
Cases	1	0	0	0	0	1	0	0	0	1	1	5	1	3
Day	14	15	16	17	18	19	20	21	22	23	24	25	26	27
Cases	0	2	3	4	7	4	3	2	1	0	0	0	2	1

note that a more complete analysis could be obtained by extending our approach along the lines in O'Neill & Roberts (1999) in which the epidemic is not necessarily assumed to have finished by the end of the observation period.

Prior distributions for β and γ were the same as for example 2. Posterior density summaries are presented in Table 8. The shapes of the marginal posterior densities for β, γ, p and R_0 essentially resembled those in example 3, and a corresponding scatterplot to Fig. 3 also looked similar. As for example 3, we can compare our estimate of the posterior mean for R_0, namely 1.12, with a martingale-based estimate given only final size data and assuming homogeneous mixing, and in this case the estimate is 1.09.

Regarding the algorithm, mixing was slower than for the previous example, although not prohibitively so. Although the actual data will affect algorithm performance, the main reason for slower mixing is likely to be the larger number of infections and larger population size in the current example. In particular, a larger number of infections increases the size of the parameter space that the algorithm explores, and this seems likely to have a greater impact on mixing than a larger population size.

Table 8. *Posterior parameter summaries from MCMC algorithm using shigellosis dataset, example* 4

	Parameter			
	β	γ	p	R_0
Mean	0.017	0.38	0.51	1.12
Median	0.0041	0.37	0.52	1.10
S.D.	0.031	0.096	0.30	0.24
Equal-tailed 95% C.I.	(0.0018, 0.042)	(0.24, 0.55)	(0.055, 0.96)	(0.77, 1.56)

6. Conclusions

We have described methodology for performing Bayesian statistical inference for a network epidemic model incorporating a simple unobserved social structure mechanism, given two types of temporal outbreak data. Inference can be made about the parameters governing the social structure as well as those governing the epidemic. For simplicity we have considered only a relatively basic model for both the underlying social structure and the epidemic transmission mechanism, using natural parameters (β, γ, p). This parameterization aids construction of a suitable MCMC algorithm, so that for example many of the full conditional distributions of parameters can be written down explicitly. It is possible to use alternative parameterizations, and in particular those which give lower posterior correlations, in order to improve convergence. However, the implementation details are rather more complicated, and so we have not considered this here. However, for large population analyses, reparameterizations might be worthwhile.

Our modelling framework can in principle be extended to more complex situations, particularly with a view to increased realism. Extensions are possible for both the underlying social structure and also the epidemic transmission model itself. In both cases, increased complexity may lead to additional parameters, in which case identifiability problems may occur, unless strong prior assumptions are employed. This is because the data may not be sufficiently detailed to allow separate estimation of individual parameters in a more complex model. However, this situation does not always arise: for instance, if a particular fixed social structure is assumed (perhaps on the basis of existing knowledge) then estimation is only required for the parameters governing epidemic spread. More complex modelling assumptions are in turn likely to lead to less straightforward MCMC algorithms, and in particular careful

design may be needed to ensure efficiently performing algorithms. This is especially likely to be the case if large datasets, perhaps with populations of several hundred, are considered.

Acknowledgements

PDO'N acknowledges the warm hospitality of the Mathematics Department at Stockholm University where this work was initiated, during a visit funded by the European Science Foundation initiative on Highly Structured Stochastic Systems. TB was partly supported by EPSRC grant GR/N09091. Both authors thank Håkan Andersson for his part in the early development of the work, and Arnoldo Frigessi and three anonymous referees for helpful comments.

References

Altmann, M. (1998). The deterministic limit of infectious disease models with dynamic partners. *Math. Biosci.* **150**, 153–175.

Andersson, H. (1998). Limit theorems for a random graph epidemic model. *Ann. Appl. Probab.* **8**, 1331–1349.

Andersson, H. (1999). Epidemic models and social networks. *Math. Sci.* **24**, 128–147.

Andersson, H. & Britton, T. (2000). *Stochastic epidemic models and their statistical analysis*. Lecture Notes in Statistics 151, Springer, New York.

Ball, F. G., Mollison, D. & Scalia-Tomba, G-P. (1997). Epidemic models with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89.

Becker, N. G. (1989). *Analysis of infectious disease data*. Chapman & Hall, London.

Becker, N. G. & Britton, T. (1999). Statistical studies of infectious disease incidence. *J. Roy. Statist. Soc. Ser. B* **61**, 287–307.

Becker, N. G. & Dietz, K. (1995). The effect of the household distribution on transmission and control of highly infectious diseases. *Math. Biosci.* **127**, 207–219.

Benenson, A. S. (ed.) (1990). *Control of communicable diseases in man*. 15th edn. American Public Health Association, Washington.

Cáceres, V. M., Kim, D. K., Bresee, J. S., Horan, J., Noel, J. S., Ando, T., Steed, C. J., Weems, J. J., Monroe, S. S. & Gibson, J. J. (1998). A viral gastroenteritis outbreak associated with person-to-person spread among hospital staff. *Infection Control Hospital Epidemiol.* **19**, 162–167.

Gilks, W. R., Richardson, S. & Spiegelhalter, D. J. (1996). *Markov chain Monte Carlo in practice*. Chapman & Hall, London.

Los Angeles County, Department of Health Services (L.A.D.H.S) Public Health Letter (1992) **14**, No. 4.

O'Neill, P. D. & Roberts, G. O. (1999). Bayesian inference for partially observed stochastic epidemics. *J. Roy. Statist. Soc. Ser. A* **162**, 121–129.

O'Neill, P. D., Balding, D. J., Becker, N. G., Eerola, M. & Mollison, D. (2000). Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. Roy. Statist. Soc. Ser. C* **49**, 517–542.

Received November 2000, in final form September 2001

Tom Britton, *Department of Mathematics, Uppsala University, P.O. Box 480, SE-751 06 Uppsala, Sweden.*
E-mail: tom.britton@math.uu.se