Biometrics 68, 755–765
September 2012

DOI: 10.1111/j.1541-0420.2012.01748.x

A Network-based Analysis of the 1861 Hagelloch Measles Data

Chris Groendyke,[1,2,*] David Welch,[1,3,4] and David R. Hunter[1,3]

[1]Department of Statistics, Pennsylvania State University, University Park, Pennsylvania 16802, U.S.A.
[2]Current address: Department of Mathematics, Robert Morris University, Moon Township,
Pennsylvania 15108, U.S.A.
[3]Center for Infectious Disease Dynamics, Pennsylvania State University, University Park,
Pennsylvania 16802, U.S.A.
[4]Current address: Department of Computer Science, University of Auckland, Auckland 1142, New Zealand
[*]*email:* groendyke@rmu.edu

Summary. In this article, we demonstrate a statistical method for fitting the parameters of a sophisticated network and epidemic model to disease data. The pattern of contacts between hosts is described by a class of dyadic independence exponential-family random graph models (ERGMs), whereas the transmission process that runs over the network is modeled as a stochastic susceptible-exposed-infectious-removed (SEIR) epidemic. We fit these models to very detailed data from the 1861 measles outbreak in Hagelloch, Germany. The network models include parameters for all recorded host covariates including age, sex, household, and classroom membership and household location whereas the SEIR epidemic model has exponentially distributed transmission times with gamma-distributed latent and infective periods. This approach allows us to make meaningful statements about the structure of the population—separate from the transmission process—as well as to provide estimates of various biological quantities of interest, such as the effective reproductive number, R. Using reversible jump Markov chain Monte Carlo, we produce samples from the joint posterior distribution of all the parameters of this model—the network, transmission tree, network parameters, and SEIR parameters—and perform Bayesian model selection to find the best-fitting network model. We compare our results with those of previous analyses and show that the ERGM network model better fits the data than a Bernoulli network model previously used. We also provide a software package, written in **R**, that performs this type of analysis.

Key words: Exponential family Random Graph Model; Hagelloch; Measles; Networks.

1. Introduction

Networks are now commonly used to model interactions between hosts that enable the spread of disease through a population. Estimating the parameters of these network models from data, however, remains a serious challenge. The focus of this article is on fitting a plausible network model to real data to demonstrate that rigorous statistical methods can feasibly be used with this class of models. The data we fit here—from a measles outbreak in Germany in 1861—were very well documented (Pfeilsticker, 1863) and, thus, provide an ideal testing ground for new methods.

There have been several previous analyses of this data set, with differing goals and utilizing various methods. Our analysis differs from most previous works in that we assume that the epidemic spreads across the edges of a contact network. Contacts, here, are assumed to be substantial enough to have a reasonable chance of transmitting the pathogen. We use the data to infer the properties of this contact network. These properties—the factors that influence the propensity of individuals to make infectious contacts with one another—are very important in the study of epidemiology, as the network structure is known to have a significant impact on both the spread of an epidemic (Anderson and May, 1991, Chapter 11–12, Wallinga, Edmunds, and Kretzschmar, 1999, Read and Keeling, 2003; Keeling and Eames, 2005; Meyers et al., 2005), as well as on the methods of containing the spread of these epidemics (Ball, Mollison, and Scalia-Tomba, 1997; Becker and Utev, 1998; Keeling et al., 2002; Hall and Becker, 2009). As argued in Welch, Bansal, and Hunter (2011), there is a lack of rigorous statistical work that fits network models to data; this article seeks to address that lack.

We extend the methodology set out in Britton and O'Neill (2002), Ray and Marzouk (2008), and Groendyke, Welch, and Hunter (2011), to deal with host covariate information. This means that known discrete or continuous properties of hosts, such as age, household, other group memberships, and spatial distribution can be incorporated into an analysis and the relative importance to the spread of the disease of each of these properties quantified. Further, we show that a Bayesian model selection algorithm can be used to find the subset of covariates that best explains the contact structure within the population. Thus, for the first time, an explicit network model of plausible complexity can be estimated from disease data allowing us to more fully understand the mechanisms at play in the spread of epidemics through populations, and potentially offer better means of testing alternate strategies for containing these epidemics.

In the remainder of this section, we introduce the Hagelloch measles data set and summarize previous analyses of these data. Section 2 describes the models and methods used in our analysis. Section 3 presents results including parameter estimates, model selection, estimates of the effective reproduction number, and an assessment of the model fit. Section 4 concludes with a discussion.

1.1 *Hagelloch Measles Data and Previous Work*

In 1861, a severe measles outbreak spread through the town of Hagelloch, Germany, ultimately infecting 188 children. Pfeilsticker (1863) recorded many pertinent details of this epidemic including, for each infected individual in the population, their household, school class, household, age, gender, dates of symptom onset, and various other items. Oesterle (1992) later augmented these data by mapping the spatial coordinates of each affected household and also inferred, for each infected individual, the person who was the putative source of infection. One hundred and eighty-eight children, aged fifteen and younger, were susceptible to measles during the time of this epidemic and each of these individuals was indeed infected over the course of this outbreak. Part of this data set is displayed in Web Figure 1. See Section 2 and Lawson and Leimich (2000) or Neal and Roberts (2004) for more detailed descriptions of this data set and population.

Lawson and Leimich (2000) analyze these data using a proportional hazards model. They are interested in the spatial and temporal effects of transmission (and their interaction), and thus, consider a spatio–temporal model. The authors use their model to estimate a parameter that measures the "spatial scale of spread" and find a weak spatio–temporal interaction in the data.

Neal and Roberts (2004) analyze the Hagelloch measles data by using a stochastic epidemic model that describes the transmission rate between two individuals (one infectious and one susceptible) as a function of the individuals' covariates. In particular, they consider the effect of belonging to the same household, attending the same school class, and the physical distance between the houses of the individuals. They seek to discover which factors are the most important in describing the transmission rate and use a reversible jump Markov chain Monte Carlo (MCMC) algorithm to choose among various models. Ultimately, they find that their full model (i.e., the model incorporating all of the effects mentioned earlier) best fits the data, and that there is very strong evidence that the classroom for younger children (6–10 years old) played a strong role in enabling the spread of the epidemic.

Britton, Kypraios, and O'Neill (2011) analyze these data by introducing a three-level mixing model (a generalization of the two-level mixing model of Ball et al., 1997) and an susceptible-exposed-infectious-removed (SEIR) epidemic model. Their model assumes a three-level structure, with each (susceptible) individual belonging to a household, a group (school class in this case), and to the community as a whole. In their model, an individual may transmit the disease to any individual within their household, group, or community; the corresponding frequencies of infectious contact for each type of transmission are modeled by independent Poisson processes with varying rates. The authors produce estimates for the transmission rates in their model, and also derive estimates of a threshold parameter (defined in Ball et al., 1997) for this epidemic. They compare the log-likelihood of their model to those of two different two-level mixing models (one which eliminates household-level mixing and another that eliminates group-level mixing) and conclude that the three-level mixing model offers a substantially better fit to the data and that both the group and household effects were important in the spread of this disease.

Groendyke et al. (2011) analyze these data by using a stochastic SEIR epidemic model to model the progression of the disease and an Erdős-Rényi random graph model (Erdős and Rényi, 1959; Gilbert, 1959) to describe the contact network in the population. Although the authors were successful in estimating the parameters for their models, the Erdős-Rényi model is likely an overly simple representation of the true interaction structure because it does not allow for the incorporation of the various factors that Neal and Roberts (2004) and Britton et al. (2011) found to be material in the transmission of this disease.

2. Methods and Models

2.1 *Model, Notation, and Assumptions*

We use an undirected random graph model to describe the contact network in the population of susceptible individuals. The nodes of the graph, which are labeled $1, \ldots, N$, correspond to the individuals, whereas the edges indicate the presence of a relationship sufficient to spread measles from one person to another. Note that not all forms of contact between individuals in the population will meet this criterion. We refer to a pair of nodes, i and j, $i \neq j$, as a dyad.

The specific type of random graph model we consider for this analysis is one in which the probability of an edge between individuals i and j is given by p_{ij}, where

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \sum_s \eta_s \mathbf{X}_{\{i,j\},s}, \tag{1}$$

$\mathbf{X}$ is a matrix of dyadic covariates, and $\vec{\eta} = \{\eta_s\}$ is the corresponding vector of parameters. We assume that the existence or not of edges within distinct dyads are mutually independent so that equation (1) fully specifies the probability distribution of the network. We denote a specific network by $\mathcal{G}$, an $\binom{N}{2}$ length vector of indicators of edges for the $\binom{N}{2}$ dyads in the network.

Most of the dyadic statistics we use for this analysis are binary in nature, taking a value of either 0 or 1, depending on whether the individuals in the dyad share a characteristic or not. We refer to these "matching" effects as homophily effects; they are also sometimes called "assortativity" effects, as in Cauchemez et al. (2011). These include effects for household, classroom, and gender homophily. Because preliminary analysis showed that the effects of being in the same classroom were likely to be different for Classroom 1 than for Classroom 2, we compute a separate statistic for each classroom. Similarly, we allow the gender homophily effect to vary according to whether the dyad consists of two males or two females. Two other dyadic statistics we consider are continuous in nature: absolute age difference between individuals (measured in units of 4 years), and the spatial distance (measured in units of 100 m) between the individuals' households. Finally, we also include a statistic whose value is 1 for every dyad, to measure the overall propensity of edge formation. This statistic is the analogue of the intercept term in a regression analysis. If we take $\mathbf{X}$ to consist of only the covariate whose value is 1 for every dyad, the resulting model is the Erdős-Rényi model. Interaction terms could easily be included in this model framework by introducing covariates for each subgroup. For example, for a classroom/gender interaction, we would include binary covariates for male/Classroom 1, female/Classroom 1, and so on. Due to the sparsity of the data here, we do not consider interactions.

These models belong to a class of random graph models known as exponential-family random graph models (ERGMs) or p^* models (Wasserman and Pattison, 1996), which have seen much use in the field of social network analysis. The particular type of model given in equation (1) is a dyadic independence model, in that the probability that any dyad will contain an edge is solely a function of the characteristics of the two individuals comprising the dyad, and is unaffected by any other dyads.

Although $\mathcal{G}$ contains all of the potential edges that the disease might travel across, in actuality, the disease only traverses a subset of these edges. This subset of edges forms a directed transmission tree, which we denote by $\mathcal{P}$. The root of this tree is the initially infected individual, whose identity is generally unknown, though we will assume that we know it for the Hagelloch measles data.

To model each person's progression through the course of the disease, we use a stochastic SEIR epidemic model that we describe briefly below (see Keeling and Rohani, 2008, for a thorough description of this model). We assume that the population initially consists of one infectious individual, whereas the remainder of the population is susceptible. Susceptible individuals may only become exposed via contact with people in the infectious class with whom they share an edge in $\mathcal{G}$. The time taken to transmit the disease along a given edge is assumed to follow an exponential distribution with mean $1/\beta$. We assume that exposed individuals remain in this category for a length of time described 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 class. They remain infectious for a length of time described by a gamma random variable with mean $k_I\theta_I$ and variance $k_I\theta_I^2$, after which time they are removed and play no further part in the epidemic.

The primary data for our model consist of the times at which each individual entered the exposed, infectious, and removed states. For an individual j, these times are denoted E_j, I_j, and R_j, respectively; the sets of all such times are denoted $\mathbf{E}, \mathbf{I}$, and $\mathbf{R}$, whereas the collective set of all times is denoted by $\mathbf{T} = (\mathbf{E}, \mathbf{I}, \mathbf{R})$. The Hagelloch measles data contain information that we can use to assign values to $\mathbf{I}$ and $\mathbf{R}$, but we will have to infer $\mathbf{E}$ as part of our inferential procedure. Because all $\mathbf{I}$ and $\mathbf{R}$ times are assumed known in this data set, statistical inference for the k_I and θ_I parameters is trivial and does not affect other parts of the analysis; however, this would no longer be true if not all of these times were observed, in which case the missing values would be treated as extra parameters. The times in the data are rounded to the nearest day but we treat them as exact; this assumption should introduce no systematic bias to our parameter estimates. Following Lawson and Leimich (2000) (also see Atkinson et al., 2011), we assume that each individual became infectious one day before the onset of the prodrome (the early, mildly symptomatic part of the disease) and that each individual entered the removed state 3 days after onset of rash (or at death, if sooner). We also remove one outlying data point from the data set as this individual's symptoms appeared nearly one month after the rest of the epidemic had subsided so its infection was probably contracted elsewhere; see Groendyke et al. (2011) for an analysis of the effects of this outlier.

2.2 *Inferential and Computational Methods*

Following Britton and O'Neill (2002), we treat $\mathcal{G}$ and $\mathcal{P}$ as extra parameters and estimate them along with the other model parameters to simplify the computational burden of updating the parameters in our MCMC algorithm. As in Groendyke et al. (2011), the likelihood function, expressed in a form that exploits the simplicity of the model when we condition on $\mathcal{G}$ and $\mathcal{P}$, is

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

By the independence assumption explained earlier, the term $f(\mathcal{G}|\boldsymbol{\eta})$ is simply a product of terms $p_{ij}^{I_{ij}}(1-p_{ij})^{1-I_{ij}}$ for all node pairs $i < j$, where p_{ij} is given by equation (1) and I_{ij} is the indicator of a contact in $\mathcal{G}$ between nodes i and j. The term $f(\mathcal{P}|\mathcal{G})$ is simply a uniform distribution that makes all $\mathcal{P}$ that are possible given a particular contact network $\mathcal{G}$ equally likely. Finally, because $\mathbf{T}$ depends on $\boldsymbol{\eta}$ only through $\mathcal{G}$, we may write the first term above as $f(\mathbf{T}|\beta, k_E, \theta_E, k_I, \theta_I, \mathcal{G}, \mathcal{P})$. As explained in Neal and Roberts (2005) and Groendyke et al. (2011), this term factors into four separate pieces, which give the contributions due to (a) contacts over which the epidemic was transmitted (i.e., edges in $\mathcal{P}$), (b) contacts over which epidemic was not transmitted ($\mathcal{G} \setminus \mathcal{P}$), (c) transitions from exposed to infectious, and (d) transitions from infectious to removed. If m represents the total number of individuals who eventually become infected, we thus obtain $f(\mathbf{T}|\beta, k_E, \theta_E, k_I, \theta_I, \mathcal{G}, \mathcal{P})$ as the following product:

$$\beta^{m-1}e^{-\beta A_1} \times e^{-\beta A_2} \times \frac{\left(\prod_{i=1}^{m}[I_i - E_i]\right)^{k_E - 1} \theta_E^{-m k_E} e^{-B/\theta_E}}{(\Gamma[k_E])^m} \times \frac{\left(\prod_{i=1}^{m}[R_i - I_i]\right)^{k_I - 1} \theta_I^{-m k_I} e^{-C/\theta_I}}{(\Gamma[k_I])^m},$$

where

$$A_1 + A_2 = \sum_{(a,b)\in\mathcal{P}} (E_b - I_a) + \sum_{(a,b)\in\mathcal{G}\setminus\mathcal{P}} ([\{E_b \wedge R_a\} - I_a] \vee 0)$$

equals the total "infectious pressure," that is, the total time spent by all infectious–susceptible pairs that share a contact, $B = \sum_{i=1}^{m}(I_i - E_i)$ is the total time spent by all individuals in the exposed state, and $C = \sum_{i=1}^{m}(R_i - I_i)$ is the total time spent in the infectious state.

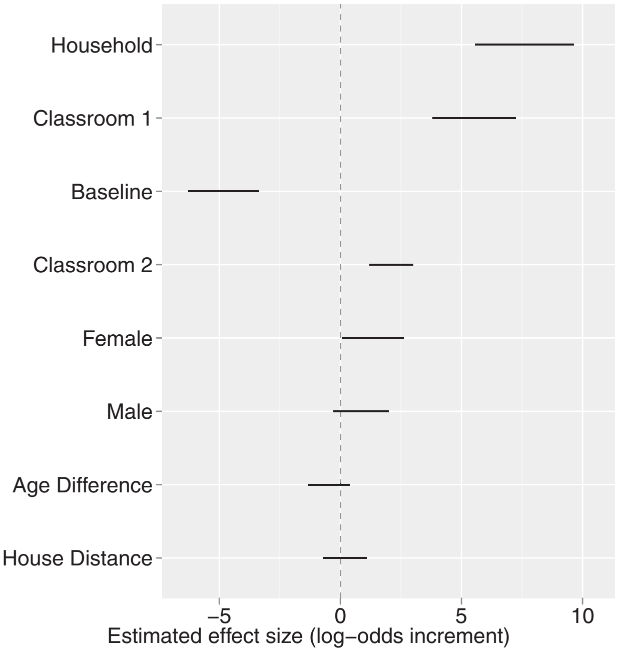

Figure 1. The points show the posterior mean whereas the lines show the 95% highest posterior density interval for each parameter. The house distance and age difference parameters correspond to continuous statistics measured in units of 100 m and 4 years, respectively. All of these network parameters have normal priors with means of 0 and standard deviations of 3. The primary model discussed in Section 3, model 4, excludes the house distance and age difference parameters (see Section 3.2).

We use a Bayesian inferential approach, assigning independent prior distributions to the parameters. For the epidemic parameters governing the lengths of the exposed and infectious periods (k_E, θ_E, k_I, θ_I), we assign uniform priors with hyperparameters governed by relevant known scientific information regarding measles: We assign $\pi_{k_E} \sim \text{Uniform}(8, 20)$, $\pi_{\theta_E} \sim \text{Uniform}(0.25, 1)$, $\pi_{k_I} \sim \text{Uniform}(15, 25)$, and $\pi_{\theta_I} \sim \text{Uniform}(0.25, 0.75)$. For β, we assign a uniform prior on (0, 4), encompassing the range of biologically plausible values; see Section 3.4 for further discussion. For the dyadic η parameters governing the network, we assign independent normal prior distributions with means of 0 and standard deviations of 3. These priors are quite diffuse because the η parameters are on the log-odds scale; experiments with more diffuse priors did not noticeably alter the resulting posteriors.

Inference is then based on the joint posterior distribution of the model parameters. To produce an approximate sample from this distribution, we use an MCMC algorithm similar to that described in Groendyke et al. (2011). The parameters $\beta, k_E, \theta_E, k_I, \theta_I, \mathbf{E}$, and $\mathcal{P}$ are updated exactly as described in Groendyke et al. (2011) using random-walk Metropolis-Hastings updates or Gibbs sampling; as per the assumptions described earlier, for the Hagelloch measles data, we assume that all values of $\mathbf{I}$ are fixed and known. However, the more complicated ERGM network structure used here necessitates different procedures for updating $\boldsymbol{\eta}$ and $\mathcal{G}$.

We update $\boldsymbol{\eta}$ using a Metropolis–Hastings step that proposes a new value for $\boldsymbol{\eta}$ from a mutivariate normal distribution centered at the current value of $\boldsymbol{\eta}$. The off-diagonal entries in the variance–covariance matrix of the proposal distribution are set to zero, whereas the diagonal entries are tuning parameters. We then accept the proposal according to the appropriate Hastings ratio.

Because the ERGM we use here is a dyadic-independence model, we can update $\mathcal{G}$ by considering each dyad separately. Thus, we cycle through each of the $\binom{N}{2}$ dyads, drawing from the appropriate full conditional distribution which, as a result of the ERGM network structure used here, will depend on $\boldsymbol{\eta}$ and $\mathbf{X}$. In Section 3, we describe a model selection procedure based on reversible jump MCMC (RJMCMC) to determine for which values of s the $X_{\{i,j\},s}$ statistics should be kept in the model given in equation (1) and which can be safely left out.

We provide a software package named **epinet** for the **R** language (R Development Core Team, 2009); this software is publicly available on the Comprehensive R Archive Network (cran.r-project.org). The **epinet** package includes the Hagelloch data set studied in this article, along with routines to perform the MCMC algorithm described here and various simulation and plotting functions. Through simulations studies and tests not reported here, this software has been shown to be able to successfully infer parameter values in many cases and is useful for data sets larger than the Hagelloch data analyzed here; see Groendyke et al. (2011) for further discussion of this software package.

We ran the algorithm for 50,000,000 iterations and thinned every 500 iterations to reduce the autocorrelation between samples and reduce memory requirements. The result is 100,000 samples containing at least 2,000 approximately independent samples (as measured by the integrated autocorrelation time) for each model parameter.

3. Results

3.1 *Network and Epidemic Parameter Estimates*

Here, we examine the results of the analysis of the Hagelloch measles data, highlighting the differences between our analysis and that of Groendyke et al. (2011), which used a similar inference approach, but employed a simpler Erdős-Rényi network model. Figure 1 summarizes the posterior distributions of the parameters in the network model, $\boldsymbol{\eta}$, via posterior means and 95% highest posterior density intervals. Each of these parameters can be interpreted as the incremental log-odds associated with a change of one unit in the corresponding covariate. We use the posterior distribution of these parameters along with equation (1) to estimate the probability, p_{ij}, of a contact existing between any two individuals i and j. For example, the model predicts that the posterior mean probability p_{ij} of a contact between two individuals ($\mathbf{X}_{\{i,j\},3} = -4.81$), both female ($\mathbf{X}_{\{i,j\},5} = 1$ and $\mathbf{X}_{\{i,j\},6} = 0$), both in Classroom 2 ($\mathbf{X}_{\{i,j\},2} = 0$ and $\mathbf{X}_{\{i,j\},4} = 1$), but not in the same household ($\mathbf{X}_{\{i,j\},1} = 0$), whose ages differ by 1 year ($\mathbf{X}_{\{i,j\},7} = 0.25$) and whose houses are 300 m apart ($\mathbf{X}_{\{i,j\},8} = 3$), satisfies

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = 7.36(0) + 5.14(0) - 4.81 + 2.10(1) \\ + 1.44(1) + 0.91(0) - 0.45(0.25) + 0.24(3),$$

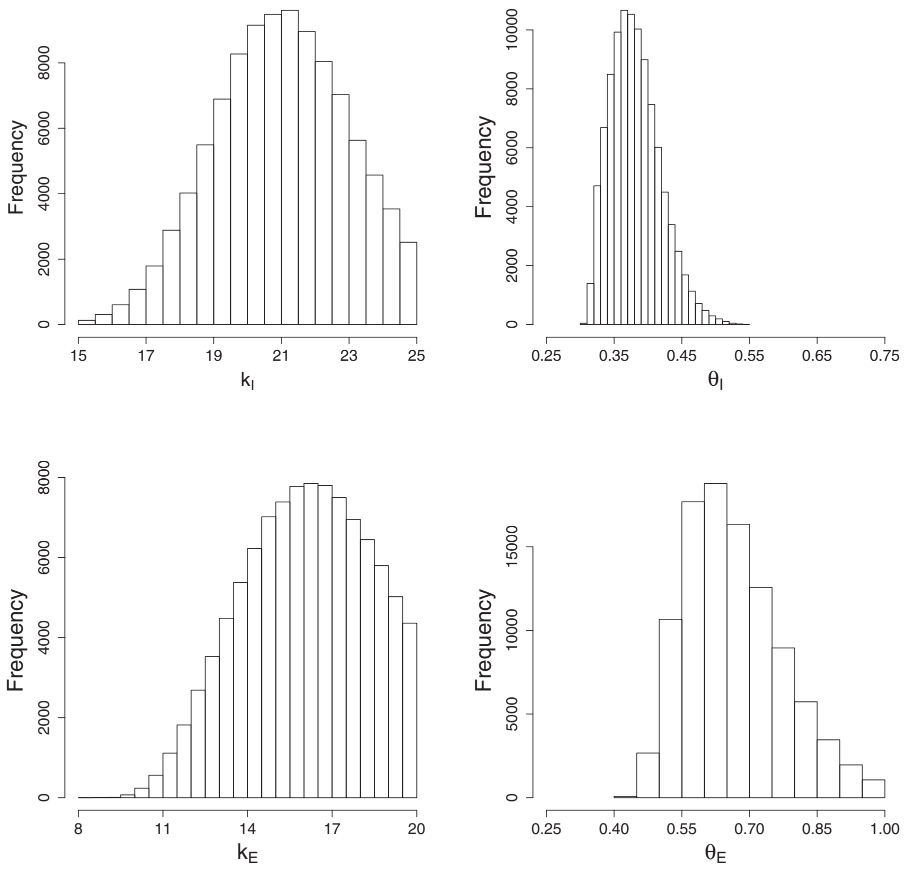

Figure 2. Estimated posterior densities for epidemic parameters k_I, θ_I, k_E, and θ_E. Uniform prior distributions were used for each of these parameters.

or $\hat{p}_{ij} = 0.34$. As only the marginal posteriors are presented in Figure 1, the credible interval for any p_{ij} cannot be read off directly, although it can be calculated from the MCMC output.

There are a few notable features of these parameter estimates. First, the household and classroom effects are overwhelmingly strong; any two individuals who are in both the same household but no other relationship have a mean posterior estimate $\hat{p}_{ij} = 0.93$ (ignoring the house distance and age difference terms), and that rises to $\hat{p}_{ij} = 1.00$ if they are also both in Classroom 1. (Classroom 1 corresponds to the classroom for younger children, whereas Classroom 2 is the classroom for the older group of children.) This is plausible considering the extremely contagious nature of measles. There is also a noticeable gender homophily effect, and furthermore, this effect appears to vary by gender, with females showing a stronger tendency to contact each other than males. There is also some evidence of an age effect; the posterior distribution for this coefficient falls largely below zero, indicating that increasing age differences result in decreasing odds of contact. The posterior distribution of the parameter related to spatial distance between houses is roughly symmetric and centered close to zero, indicating that the effect of this parameter is likely negligible.

The estimated posterior distributions for the epidemic parameters k_I, θ_I, k_E, and θ_E are shown in Figure 2. Based on these estimates, we find that the estimated mean length of the exposure period ($k_E \times \theta_E$) falls in the range of 9–12 days and the estimated mean length of the infectious period ($k_I \times \theta_I$) falls in the range of 7.5–8.5 days; these values are in concordance with known scientific information regarding measles (Gough, 1977).

3.2 *Model Selection*

The posterior distribution of the network parameters (see Figure 1 for summaries of the marginal distributions of these parameters) indicates that most of the parameters in the model are very likely to be substantially different from zero, and hence that the corresponding covariates have significant effects on the network structure. There are a couple of parameters, however, that deserve further discussion. Recall that the age difference parameter corresponds to a statistic measuring the absolute value of the age difference between two children. The marginal posterior distribution for this parameter is largely negative, suggesting that children who are closer in age will tend to be more likely to be in contact than those with larger age differences. Compared to many of the other parameters, though, the effect of this parameter appears to be rather weak. This is likely due to the inclusion of the two classroom effects in the model. As each classroom consists of children who are close in age, we might expect that much of the effect of age similarity would be captured in the classroom effects. Indeed, when we leave the age difference parameter out of the model, the estimates for both of the classroom parameters increase accordingly. Nonetheless, the fact that the majority of the marginal posterior distribution for this parameter is negative indicates that there may be some effect due to age difference beyond that which can be explained by the classroom effects.

We see a similar relationship between the house distance and household variables. The house distance parameter is estimated to be very close to zero in the presence of the household variable. However, when the household variable is excluded from the model, the house distance parameter is estimated to be substantially negative. Thus, house distance essentially acts as a proxy for household when the household parameter is excluded.

To help determine whether either or both of these parameters belong in our model, we use a RJMCMC algorithm to perform model selection among four candidate models. Model 1 contains parameters corresponding to all of the aforementioned factors: edges, household, Classroom 1, Classroom 2, house distance, male match, female match, and age difference. Model 2 contains all of these factors except for the house distance effect, model 3 contains all factors in model 1 except for age difference, and model 4 contains neither the house distance nor the age-difference effect. Each of the four models is assigned a prior probability of 1/4.

We implement the RJMCMC by augmenting our MCMC algorithm to include a model-switching step in each sweep through the parameters. We move among the four candidate models by proposing model changes that add or remove one parameter, with the specific proposals depending on the current state of the model. In all model states, there are two possible other models that we could move to and we propose moving to each of these two alternative models with probability 0.5.

With the exception of the parameter being added or dropped from the model, the proposed parameter values are all set equal to the current parameter values. If we are proposing switching from a model without the age-difference parameter to a model with this parameter, the proposed value of this parameter is drawn from a $N\left(0,\sigma_A^2\right)$ distribution. If we are proposing switching from a model without the house distance parameter to a model with this parameter, the proposed value of this parameter is drawn from a $N\left(0,\sigma_{HD}^2\right)$ distribution. Each proposed switch from one model to another is evaluated according to the general procedure outlined by Green (1995), though we omit the technical details here.

The Markov chain produced by this RJMCMC algorithm spent the majority (approximately 70%) of its time in model 4. There was also considerable evidence (about 26% posterior probability) for model 3. Models 1 and 2 each received very little (less than 3%) support. Based on these results and the discussion earlier, we proceed with our analysis using model 4, which we believe to be the best and most parsimonious of the candidate models.

Because RJMCMC model selection procedures can be influenced by the prior distributions of the parameters being investigated (Richardson and Green, 1997), we performed tests of the sensitivity of our model selection algorithm to the prior variances of the two parameters in question. We found that the procedure is relatively insensitive to the choice of priors, and that under a wide range of prior variances, model 4 is clearly favored.

3.3 *Degree Distribution*

We can also use the posterior distributions of the network parameters to construct estimates of the degree distribution (where "degree" refers to the number of contacts for an individual in a network) for this population. To do this, we sampled from the joint posterior distribution of the η parameters and used these sample values to construct simulated contact networks. The corresponding degree distributions of these networks are shown in Figure 3. We can clearly see a distinct difference in the pattern of the degree distribution generated by the Erdős-Rényi model, as opposed to model 4. Specifically, the Erdős-Rényi model produces a roughly symmetric distribution of degrees, whereas model 4 produces a degree distribution that is noticeably right-skewed. This right-skewed shape more closely resembles the shape of most typical social networks (Wasserman and Faust, 1994), indicating that the more general network model is likely to be a more realistic description of population interactions.

3.4 *Distinguishing Between Contact and Transmission*

One of the notable difficulties in using epidemic data to perform inference for the parameters in this model lies in separating the effects of the epidemic parameters from those of the network parameters. This problem was discussed in Britton and O'Neill (2002) and explored in Groendyke et al. (2011). One method of assessing the severity of this problem is to examine the correlations in the joint posterior distribution; as in Groendyke et al. (2011), we consider the correlation between $\log\left(p_{ij}\right)$ and $\log\left(\beta\right)$. For the Erdős-Rényi model, in which, for all i, j, $p_{ij} = p$, the correlation between $\log\left(p_{ij}\right)$ and $\log\left(\beta\right)$ is approximately -0.79, whereas for model 4, the posterior correlations between $\log\left(p_{ij}\right)$ and $\log\left(\beta\right)$ vary by dyad, but all fall between -0.31 and 0.05. In addition, Web Figure 2 shows the estimated posterior density for the parameter β for both the Erdős-Rényi model and model 4; clearly, the latter model yields a stronger signal for β, whereas in the Erdős-Rényi model, the posterior distribution of β is substantially influenced by the shape of its prior. Groendyke et al. (2011) similarly found that when using an Erdős-Rényi model to analyze this data, the posterior distribution of β was heavily influenced by its prior (a gamma distribution).

3.5 *Transmission Tree*

In some cases, the transmission tree itself—the sub-network containing information about who infected whom—may be of interest. Within the MCMC sampling procedure, the transmission tree is treated like any other unknown parameter and is sampled at each iteration of the algorithm. One such sample is shown in Figure 4. It is drawn in a style to display both temporal information (when individuals became exposed, infectious, etc.) and topological information (who infected whom). Inspection of the tree can be highly informative about the behavior of the epidemic. Here, for example, we see that individuals typically infected others as soon as they become infectious, suggesting that the virus is highly infectious and spread is limited by rapid exhaustion of susceptible contacts.

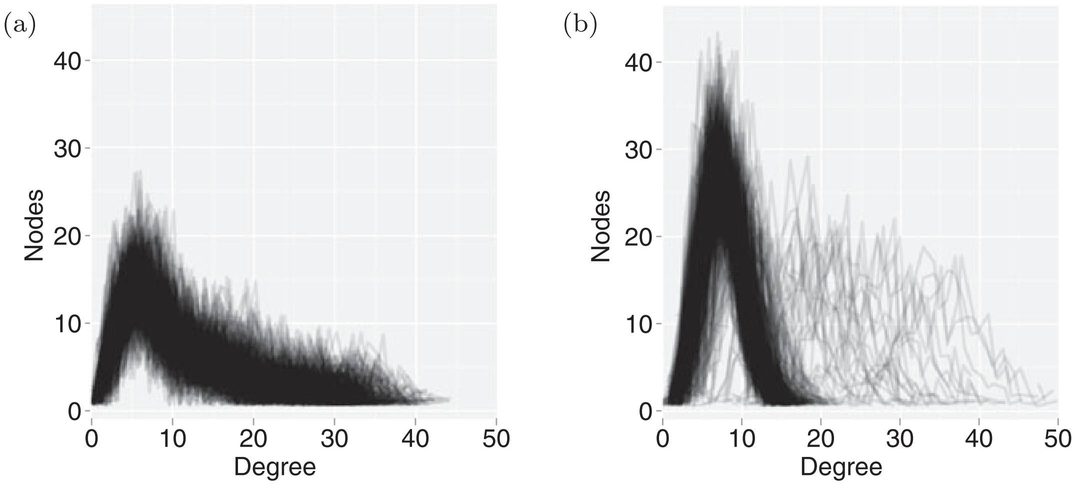

Figure 3. Degree distributions for contact networks simulated from the posterior distribution of the network model parameters for (a) model 4 and (b) the Erdős-Rényi model.

We caution, however, that this specific tree is just a single sample from the MCMC run and is not representative of the posterior distribution of transmission trees so must not be over interpreted. An example of the great variability in the topology of sampled trees is seen by looking at the case of Host 176, labeled in Figure 4. The tree shows this host causing 16 secondary infections, more than any other in the outbreak. Yet, the posterior distribution of secondary infections caused by Host 176, shown in Web Figure 3, indicates that there is little signal in the data for how many infections were caused by this host, with estimates ranging from 4 to 34. Oesterle (1992), who empirically assigned secondary infections to individual hosts, assigned none to Host 176. This is unsurprising given the data and the fitted model where, provided an edge is present in the contact network, the virus is equally likely to be transmitted over any edge from an infectious node to a susceptible one.

3.6 *Reproduction Number*

In the study of epidemics, one of the quantities that is commonly of interest is the basic reproductive number, R_0, defined as the mean number of secondary infections caused by a single infectious individual in a fully susceptible population (Keeling and Rohani, 2008; Anderson and May, 1991). Here, we are working with a mixed population of immune adults and susceptible children so consider the effective reproduction number, R, which is the actual number of secondary cases per primary case (Wallinga and Teunis, 2004) and provides a lower bound for R_0. R is largest at the beginning of the epidemic when depletion of susceptible individuals has had little impact. We use techniques for finding a mean estimate of R at the beginning of the epidemic by using methods for estimating R_0 in situations where a stochastic epidemic is assumed to spread over edges of a contact network that includes only susceptible individuals. Britton and O'Neill (2002) gives such a formula for a type of SIR epidemic and Erdős-Rényi network model; Groendyke et al. (2011) slightly modify this formula for use with an SEIR epidemic model. Although these formulas consider the mean degree of the contact network, they fail to take into account the shape of the degree distribution. Meyers (2007) describes an approach for calculating R_0 which depends on the first two moments of the degree distribution; Kenah (2011) discusses a similar formulation, originally proposed by Andersson (1998), incorporating the distribution of the length of time spent by individuals in the infectious state to produce a formula for R_0. Using this general framework, we can find an expression for the mean of R at the beginning of the outbreak corresponding to the network and epidemic models used here:

$$R = \left(\frac{E\left[D^2\right]}{E[D]} - 1 \right) \cdot \left(1 - \left[\frac{1}{1 + \beta\theta_I} \right]^{k_I} \right), \qquad (2)$$

where D is the random variable describing the degree distribution for the individuals in the population. Using equation (2) in conjunction with the joint posterior samples generated using our MCMC algorithm, we can approximate the posterior distribution of R. Doing so yields 95% posterior credible intervals of (6.2, 9.8) for the Erdős-Rényi model and (11.9, 18.9) for model 4. Thus, model 4 yields a substantially higher estimate of R, though both estimates seem reasonable (Anderson and May, 1991, gives estimates of R_0 for measles ranging between 5 and 18 for various outbreaks). For both models, the posterior distribution of R was roughly bell shaped and symmetric.

It is also interesting to further consider the posterior distribution of β in terms of its relationship with R, as per Equation (2). Seeing $\theta_I \approx 0.4$ and $k_I \approx 20$, the term involving β disappears quickly as β increases and is small for values of β greater than about 0.5. The posterior density for β (see Web Figure 2) has a large peak at approximately $\beta = 0.5$, but it is nearly flat for the values of β greater than 2. This implies that though the data indicate a strong signal for β, they are also unable to distinguish among the larger values of β. We might then expect epidemics with transmission rates of, say, $\beta = 2$ and $\beta = 4$ to look very similar, and simulation indicates that this is indeed the case.

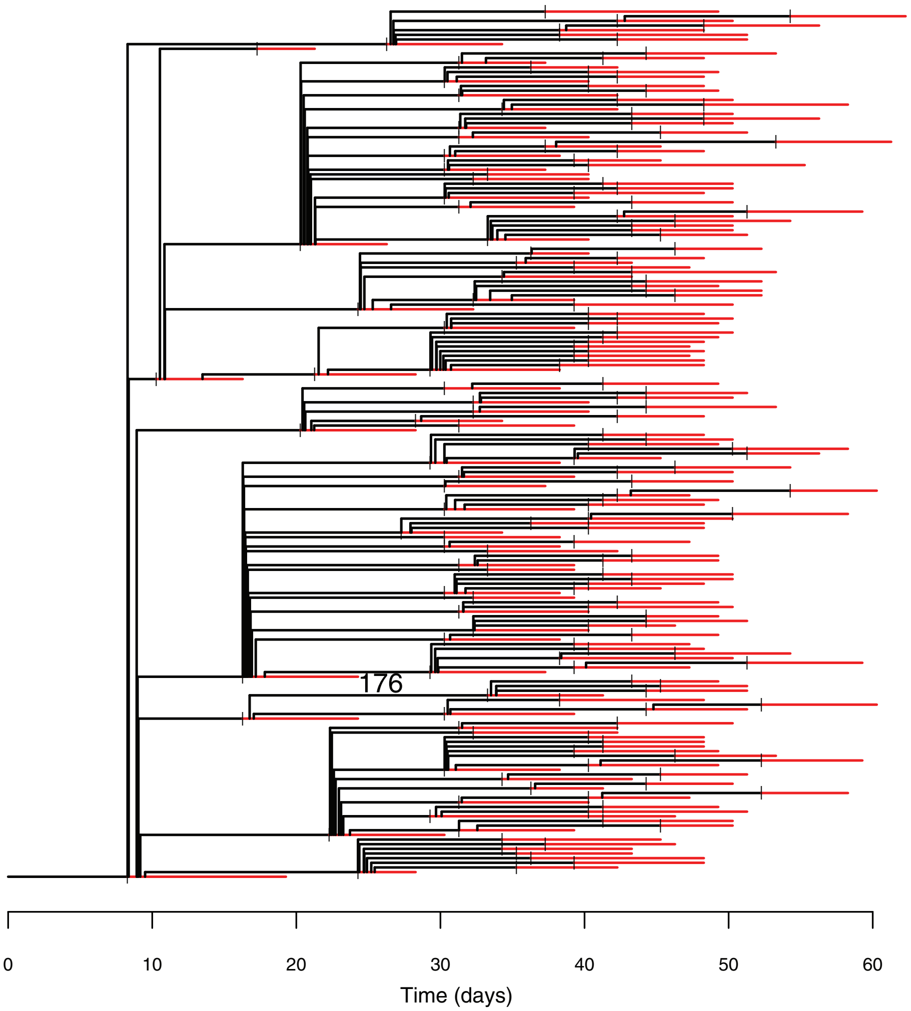

Figure 4. An example of a sampled transmission tree. The infection was introduced to the population at time zero. The horizontal lines represent exposed or infected hosts with a tick mark demarcating the transition from exposed to infectious (infectious hosts shown red in electronic version). The vertical lines show who infected whom. Thus, the branching points of the tree are infection events whereas the leaves of the tree are removal/recovery points. Host 176, labeled here at the leaf associated with its recovery time, is highlighted here as it is discussed in the text. This figure appears in color in the electronic version of this article.

3.7 *Assessing Model Fit*

We would also like to assess the quality of the network and epidemic models that we have employed here to describe the Hagelloch measles epidemic. To this end, we consider simulating 1000 contact networks from our network models and corresponding posterior parameter samples, and then simulating epidemics over these networks, again using epidemic parameter values sampled from the joint posterior distributions produced by our MCMC algorithm. We, then, assess the model fit by comparing the simulated epidemics with the actual original data. In particular, we compare the number of individuals in the infectious state over time as the epidemic progresses through the population. Figure 5 shows the actual data as compared to the simulated epidemics for model 4 as well as the Erdős-Rényi model. Overall, the simulated epidemics produced by model 4 appear to more closely match the original Hagelloch measles data than do those produced using the simpler Erdős-Rényi model. The epidemics produced by the Erdős-Rényi model spread through the population more slowly than did the actual outbreak. This is particularly noticeable at days 30–40 (where the simulated number of infectious individuals is fewer than those in the actual outbreak) as well as at days 50–60 (where the simulated number of infectious individuals is greater than those in the actual outbreak). In contrast, the more general ERGM matches the actual outbreak pattern much more closely, with the number of individuals in the infectious state rapidly increasing and then decreasing at roughly the same time points as in the actual outbreak. We believe that the main factor contributing to the ability to better match the actual outbreak pattern is the more complicated structure of the full network model—and the corresponding degree distribution pattern it produces. With the exception of β, the estimates of the epidemic model parameters were very similar between the two models; hence,

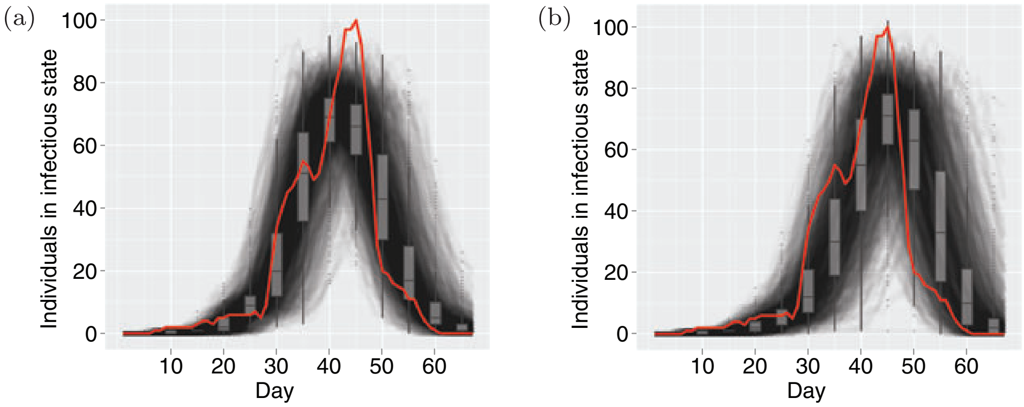

Figure 5. Number of individuals in the infectious state over time for the (a) ERGM model 4 and (b) Erdős-Rényi model. In each case, the actual Hagelloch measles outbreak data are given by the bold line (red in the electronic version), whereas simulated data are shown in multiple thin lines, with boxplots summaries shown at each 5-day time increment. This figure appears in color in the electronic version of this article.

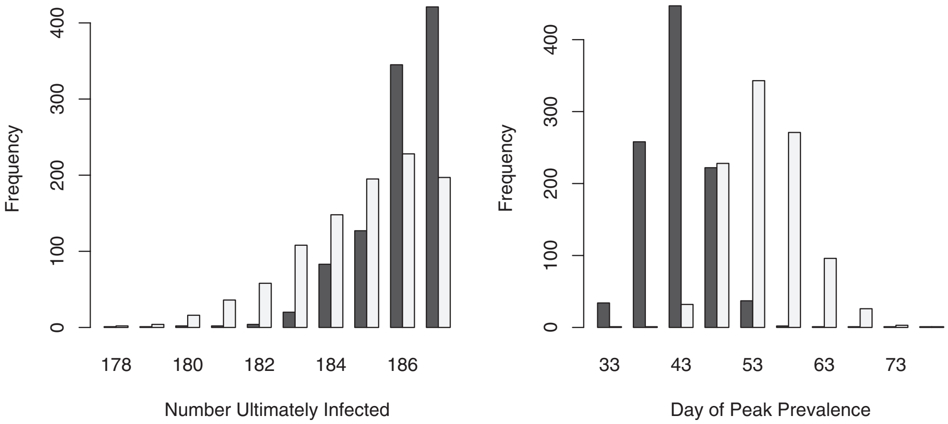

Figure 6. Summaries of simulated epidemic outbreaks in the absence (darker bars) and presence (lighter bars) of the school closing containment strategy. One thousand simulations were run for both cases. The left panel shows the distribution of outbreak sizes, that is, number of individuals who were ultimately infected during the course of the epidemic. The right panel shows the distribution of the time (measured in days since the beginning of the outbreak) that the infectious group reaches its maximal size.

we believe that the difference in network structure is the primary contributor to the improved model fit. We hasten to point out, though, that there are clearly aspects of this epidemic that even our improved model fails to adequately capture. For instance, the bulk of the simulated epidemics peak (in terms of number of infectious individuals) slightly (perhaps 3 days) before the peak of the actual outbreak. Also, the maximum number of infectious individuals in the actual outbreak is somewhat greater than that produced by the simulated epidemics.

It is also interesting to consider simulating the impact of containment strategies on the severity and rapidity of the epidemic. In particular, we consider the idea of shutting down the schools to try to contain the spread of the disease. This approach is perhaps overly simplistic, as the contact patterns of children outside of school would be modified, and probably increased, after a school closure (see, for example, Eames, Tilston, and Edmunds, 2011), yet we include it to demonstrate the ease with which control strategies can be modeled and tested within the framework we present here. As above, we simulate epidemics using parameter values sampled from the posterior distribution, but we set the η parameters corresponding to Classroom 1 and Classroom 2, both equal to zero. Examining the resulting simulated outbreaks, a couple of observations can be made. First, this containment strategy does little to diminish the ultimate size of the outbreak. The left panel of Figure 6 shows a histogram of the number of individuals infected in the simulated outbreaks. Although the outbreaks simulated under the containment strategy do indeed tend to be smaller, the difference is very small. (In fact, in both the presence and absence of the containment strategy, the epidemic affects almost all of the 187 susceptible individuals in virtually all of the simulations.) The right panel of Figure 6 gives a histogram showing the day on which the infectious group reaches its maximum size; we use this as a measure of the speed of the outbreak. We can see that the epidemic does spread considerably less rapidly in the presence of the containment strategy. These histograms indicate that this containment strategy, whereas only minimally effective in diminishing the ultimate size of the epidemic, is indeed able to significantly slow down the progression of the disease through the population.

4. Discussion

Although the problem of inferring the structure of a contact network using only epidemic data is a challenging one, our results suggest that it is indeed possible to utilize this type of data to make meaningful statements regarding which characteristics have significant influences on the propensity of individuals to make infectious contacts with one another.

In this article, we have extended previous works of Britton and O'Neill (2002), Ray and Marzouk (2008), and Groendyke et al. (2011) by considering a more general ERGM to describe the contact network in a population. This more flexible framework makes it possible to incorporate any number of nodal and dyadic covariates, any of which may be categorical or quantitative in nature. We have shown that we can not only distinguish the (biological) effects of the epidemic from the (sociological) effects of the population interactions, but we can also make meaningful statements regarding the contact structure of the population in question and which factors have substantial impacts on this structure. We demonstrated our procedure by analyzing a very rich data set describing a measles outbreak in the town of Hagelloch, Germany in 1861. The results of this analysis also suggest that this approach has the potential to provide more thorough information regarding population structure than has previously been considered.

We find that the results of our analysis of the Hagelloch measles data are broadly consistent with those of Neal and Roberts (2004) and Britton et al. (2011). Direct comparisons between these models is difficult, because the model structures used in the various analyses are quite different: Neal and Roberts (2004) used the covariates to model the transmission rate, Britton et al. (2011) used the household and classroom structures to define the levels in their three-level mixing model, and the present analysis uses the covariate information to model the network structure. We can nonetheless at least make some qualitative comparisons among the three sets of results.

Our analysis suggests that the household and classroom effects are the most substantial factors governing the network structure. Neal and Roberts (2004) and Britton et al. (2011) similarly found that these were likely significant factors in the spread of this disease; all three analyses find that the Classroom 1 effect was more substantial than the corresponding Classroom 2 effect. Our analysis found that the gender homophily factors also appear to affect the propensity of edge formation, whereas the evidence for the effect of age difference was much weaker; the other two analyses of these data did not include these factors in their model. Neal and Roberts (2004) finds a significant spatial effect in the transmission rate for this outbreak; they use three different forms for the spatial effect in their model and note that their results are robust to the choice of spatial model form. In the present analysis, other than the increase in infectious contact due to intrahousehold relationships, which we found to be the strongest effect, there does not appear to be much of a spatial effect in the data. Britton et al. (2011) did not include a spatial effect in their model. The choice of which model a researcher fits to data is largely down to what question the researcher wishes to answer. We believe there are many cases where it is of interest to separate the contact process from the transmission process as we have done here.

Although the network model considered here is much more general than those previously utilized for this type of inference, it could nonetheless be extended in several ways. Much social networks literature recommends using a dyadic dependence model rather than the independence model we use here to capture effects, such as clustering; however, we believe that the large number of covariates our independence model takes into account likely already captures clustering due to matching on these attributes, though in the absence of such covariate information it may be necessary to model clustering explicitly using a dyadic dependence model. It may also be useful in some cases to consider a more sophisticated model for the transmission rate than the simple model used here, which assumes that the transmission rate is constant, across both time and individuals. One might consider a rate that is a function of the length of time that an individual has been infected, because for many diseases, the level of infectiousness is known to vary throughout the infectious period. Further, we assumed that the infectious period began one day before the onset of prodromes and finished 3 days after the onset of rash. In work not reported here, we reanalyzed the data with a longer infectious period, finishing 5 days after the appearance of rash, as in Lawson and Leimich (2000). This change does not greatly affect parameter estimates (except for k_I and θ_I) but it changes the shape of the observed epidemic curves of the type shown in Figure 5, suggesting that this lag period could be estimated directly from the data.

We might also consider applying this type of inferential approach to data sets that are larger and more diverse than those that have been previously studied. Although previous studies that have statistically inferred network model parameters using epidemic outbreak data have mostly considered smaller and complete data sets (Britton and O'Neill, 2002; Ray and Marzouk, 2008), this approach is indeed viable for larger epidemics; our software easily allows for analysis of data sets containing up to 1000 infected individuals.

Most data sets in epidemiology are very incomplete, containing only a small fraction of the total infected population and having an unknown number of susceptible individuals to start with. In theory, missing data can be dealt with using our method by simply imputing the missing infections and times. In practice, large numbers of missing infections would drastically slow down the mixing of the MCMC algorithm and render the method unusable. Although we aim to extend the software to deal with small numbers of missing infections soon, working with larger numbers will require a more nuanced approach. Our approach does allow the possibility of incorporating different types of data into the analysis. Groendyke et al. (2011) and Welch et al. (2011) discuss potential methods for and benefits from including additional forms of data.

5. Supplementary Materials

Web Figures referenced in Sections 1.1, 3.4, 3.5, and 3.6 are available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgements

This work is supported by NIH grant R01-GM083603-01. The authors also thank the editors and two anonymous reviewers for their suggestions, which improved the quality of the article.

References

Anderson, R. and May, R. (1991). *Infectious Diseases of Humans.* Oxford: Oxford University Press.

Andersson, H. (1998). Limit theorems for a random graph epidemic model. *Annals of Applied Probability* **8,** 1331–1349.

Atkinson, W., Wolfe, S., and Hamborsky, J. (2011). *Epidemiology and Prevention of Vaccine-Preventable Diseases*, 12th edition. Washington DC: Centers for Disease Control and Prevention.

Ball, F., Mollison, D., and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *The Annals of Applied Probability* **7,** 46–89.

Becker, N. and Utev, S. (1998). The effect of community structure on the immunity coverage required to prevent epidemics. *Mathematical Biosciences* **147,** 23–39.

Britton, T., Kypraios, T., and O'Neill, P. (2011). Inference for epidemics with three levels of mixing: Methodology and application to a measles outbreak. *Scandinavian Journal of Statistics* **38,** 578–599.

Britton, T. and O'Neill, P. (2002). Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics* **29,** 375–390.

Cauchemez, S., Bhattarai, A., Marchbanks, T. L., Fagan, R. P., Ostroff, S., Ferguson, N. M., Swerdlow, D., and the Pennsylvania H1N1 working group, (2011). Role of social networks in shaping disease transmission during a community outbreak of 2009 H1N1 pandemic influenza. *Proceedings of the National Academy of Sciences* **108,** 2825–2830.

Eames, K. T., Tilston, N. L., and Edmunds, W. J. (2011). The impact of school holidays on the social mixing patterns of school children. *Epidemics* **3,** 103–108.

Erdős, P. and Rényi, A. (1959). On random graphs. *Publicationes Mathematicae* **6,** 290–297.

Gilbert, E. (1959). Random graphs. *The Annals of Mathematical Statistics* **30,** 1141–1144.

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

Green, P. J. (1995). Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika* **82,** 711–732.

Groendyke, C., Welch, D., and Hunter, D. R. (2011). Bayesian inference for contact networks given epidemic data. *Scandinavian Journal of Statistics* **38,** 600–616.

Hall, R. and Becker, N. (2009). Preventing epidemics in a community of households. *Epidemiology and Infection* **117,** 443–455.

Keeling, M. and Eames, K. (2005). Networks and epidemic models. *Journal of the Royal Society Interface* **2,** 295–307.

Keeling, M. and Rohani, P. (2008). Modeling infectious diseases in humans and animals. *Clinical Infectious Diseases* **47,** 864–866.

Keeling, M., Woolhouse, M., May, R., Davies, G., and Grenfell, B. (2002). Modelling vaccination strategies against foot-and-mouth disease. *Nature* **421,** 136–142.

Kenah, E. (2011). Contact intervals, survival analysis of epidemic data, and estimation of R0. *Biostatistics* **12,** 548–566.

Lawson, A. and Leimich, P. (2000). Approaches to the space-time modelling of infectious disease behaviour. *IMA Journal of Mathematics Applied in Medicine and Biology* **17,** 1–13.

Meyers, L. (2007). Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. *Bulletin of the American Mathematical Society* **44,** 63–86.

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

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

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

Oesterle, H. (1992). Statistiche Reanalyse einer Masernepidemie 1861 in Hagelloch. M.D.Thesis, Eberhard-Karls Universität, Tübingen.

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

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

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

Read, J. and Keeling, M. (2003). Disease evolution on networks: the role of contact structure. *Proceedings of the Royal Society of London. Series B: Biological Sciences* **270,** 699–708.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)* **59,** 731–792.

Wallinga, J., Edmunds, W., and Kretzschmar, M. (1999). Perspective: Human contact patterns and the spread of airborne infectious diseases. *TRENDS in Microbiology* **7,** 372–377.

Wallinga, J. and Teunis, P. (2004). Different epidemic curves for severe acute respiratory syndrome reveal similar impacts of control measures. *American Journal of Epidemiology* **160,** 509–516.

Wasserman, S. and Faust, K. (1994). *Social Network Analysis: Methods and Applications.* New York, NY: Cambridge University Press.

Wasserman, S. and Pattison, P. (1996). Logit models and logistic regressions for social networks: I. An introduction to Markov graphs and p^*. *Psychometrika* **61,** 401–425.

Welch, D., Bansal, S., and Hunter, D. R. (2011). Statistical inference to advance network models in epidemiology. *Epidemics* **3,** 38–45.

Received October 2011. *Revised December* 2011.
Accepted January 2012.