A Bayesian method for inferring transmission chains in a partially observed epidemic

Jaideep Ray and Youssef M. Marzouk
Sandia National Laboratories, Livermore, CA 94550-0969
{jairay,ymarzou}@sandia.gov [*]

Abstract

We present a Bayesian approach for estimating transmission chains and rates in the Abakaliki smallpox epidemic of 1967. The epidemic affected 30 individuals in a community of 74; only the dates of appearance of symptoms were recorded. Our model assumes stochastic transmission of the infections over a social network. Distinct binomial random graphs model intra- and inter-compound social connections, while disease transmission over each link is treated as a Poisson process. Link probabilities and rate parameters are objects of inference. Dates of infection and recovery comprise the remaining unknowns. Distributions for smallpox incubation and recovery periods are obtained from historical data. Using Markov chain Monte Carlo, we explore the joint posterior distribution of the scalar parameters and provide an expected connectivity pattern for the social graph and infection pathway.

Key Words: Epidemics, Bayesian inference, chains of transmission, social graphs, MCMC

1. Introduction

This paper is concerned with Bayesian inference for stochastic epidemic models that include a simple, but unobserved, underlying social network. It ties together two topics of current interest: (a) stochastic epidemic models in a structured population and (b) the use of Markov chain Monte Carlo (MCMC) samplers to explore complex high-dimensional posterior distributions. An efficient coupling of the two can benefit many problems of epidemiological interest. For example, *early* characterization of an outbreak of an emerging disease like H5N1 [1], based on *partial* observations, can profitably inform control measures. In this paper, we devise such an inference technique and apply it to the 1967 smallpox outbreak in Abakaliki, Nigeria [2].

There has been some recent interest in addressing problems of statistical inference, predicated on incomplete data, involving stochastic epidemics in a structured population. Typically, the structure involves clusters, most commonly a family or a household. The in-household rate of spread is assumed to be larger than the rate at which households themselves get infected. Cauchemez et al [3] performed such a study on the spread of influenza, as observed over a period of 15 days. Their recent work, however, structures the population into adults and children and infers the importance of children as vectors for the diseases, conditioned on Sentinel data [4]. Eichner and Dietz [2] divided the Abakaliki population into three groups and estimated inter- and intra-group spread rates of smallpox. In both of these studies, homogeneous mixing was assumed inside each group—i.e., there was no notion of a social graph.

Introduction of an unobserved social graph into an inference problem renders it high-dimensional, since the social graph itself becomes a model "parameter" to be inferred; Bayesian approaches in this context typically require the use of MCMC. MCMC has previously been used for inference in epidemiological models, even when social graphs were not involved [5, 6]. Britton and O'Neill [7] investigated a gastroenteritis and shigellosis outbreak where they explicitly introduced a social graph into a stochastic epidemic model. They assumed a SIR model and formulated a Bayesian inference problem for the dates of infection and the average contagious period of the disease (assuming the contagious period to be exponentially distributed). A closed population was assumed; a binomial graph, with an uncertain connection probability, modeled interpersonal relations. Disease transmission over a social link was modeled as a Poisson process, whose rate was also inferred as a part of the solution. Data consisted of the dates of removal of the victims of the epidemic and posterior exploration proceeded via MCMC; in particular, a mixture of Gibbs and Metropolis updates were used to sample the high-dimensional parameter field, which included both the social graph and the infection pathway. The size of the problem was generally small (10–40 patients in a population of roughly 100–200). Demiris and O'Neill [8] extended Britton and O'Neill's approach to address two-level mixing, i.e., a model in which the social graphs for inter-household and intra-houshold connections assumed different contact probabilities. However, they retained the SIR model, assumed that the contagious period was known, and continued to model the social connections with a binomial graphs.

The use of a binomial random graph to model social connections is rather restrictive; studies have shown that human contacts rarely follow such a pattern [9]. Britton's recent work has addressed the generation of random graphs that follow a given degree distribution [10], but has not yet been incorporated into an epidemic inference problem.

In this paper we extend Britton's technique to an SEIR (susceptible-exposed-infective-recovered) epidemic model. The unobserved social connections in the population are represented using two binomial graphs, for inter- and intra-household

[*] Corresponding author: jairay@sandia.gov; Published in the Proceedings of the Joint Statistical Meetings, Denver, CO. August 3–7, 2008; SAND2008-6579C

Table 1: Means and standard deviations, along with the corresponding shape and scale factors, of the Γ-distributed incubation, prodromal, and contagious/symptomatic periods for smallpox. These were obtained from [2].

Disease state	*Mean*	*Std. Dev.*	*Shape factor*	*Scale factor*
Incubation period, t_I	11.6	1.9	37.27	0.31
Prodromal period, t_P	2.49	0.88	8.0	0.31
Contagious period, t_C	16.0	2.83	31.96	0.50

connectivities. Data for the problem consist of the date of the start of the contagious phase, while dates of infection and recovery are objects of inference. The technique is tested on the Abakaliki outbreak and compared with the results (e.g., spread rates) obtained in [2]. In Section 2 we describe the epidemic in question and propose a model. In Sec 3, we formulate the statistical inference problem and describe the corresponding MCMC procedure. Results are presented in Section 4, and we conclude in Section 5.

2. Data and the Model

Between April and June 1967, a smallpox outbreak with 32 cases occurred in the town of Abakaliki, Nigeria (pop. 31,200) [11]. 30 of the 32 cases belonged to the Faith Tabernacle Church (FTC), a religious group that refused vaccination and medical treatment during the epidemic. The FTC members (a total of 74) lived in 7 large “compounds” or households, interspersed with non-FTC members (total non-FTC members over 7 compounds: 92). Social contacts between FTC and non-FTC members, however, were rare. The FTC members were closely related and visited each other often; four times a week they gathered at the church. The index case (case 0) in Abakaliki came to live in Compound 1 on April 2nd, 1967; by April 5th (day 0) she had developed a macular rash. Thereafter, smallpox spread through Compound 1. On day 25, a family seemingly free of smallpox moved to Compound 2, where on days 26 and 30, the children developed clinical signs of smallpox. Thereafter, smallpox spread in Compound 2. The last case occured on June 20 (day 76). No medical interventions were instituted until case 11; thereafter, smallpox cases were somewhat belatedly isolated at the hospital. This was the only concession FTC members made to the health authorities; they steadfastly refused to be vaccinated. 35 out of 74 FTC members had been vaccinated during their childhood, but the smallpox cases generally were not. Only the dates of appearance of symptoms were recorded, and the fates of the smallpox cases are unknown; strangely, no deaths were reported! A tabulation of the dates of appearance of symptoms, as well as the possible durations of the contagious period (since we do not know of the cases' fates) are available in [2]. The two non-FTC smallpox cases occurred in individuals associated with FTC members; one (case 20) operated a booth in the market opposite case 1, while the second (case 27), washed clothes for the people in Compound 1.

To test our technique, we will slightly idealize the problem. We will ignore the non-FTC population (including the two cases) and treat the FTC community as a closed population. We will assume that a social network, modeled as a binomial random graph with a link probability of p_{in}, exists within each compound. Cross-compound social links exist in a second binomial random graph, with a link probability of p_{out}. The spread of infection along a social link is modeled as a Poisson process, with rates β_{in} and β_{out} for in-compound and out-of-compound social links respectively. Each of 30 cases has unknown dates of infection and removal $I_i, R_i, i = 1 \ldots 30$, while their dates of exhibition of symptoms S_i constitute the data on which we condition all the unknown quantities. The other (unobserved) parameters of this stochastic epidemic model are the social network $\mathcal{G}$ and the infection pathway $\mathcal{P}$. We model the behavior of smallpox as “latent” period (non-contagious) which corresponds to the sum of the incubation and prodromal periods and the contagious period. The incubation, prodromal, and contagious/symptomatic periods are assumed to be gamma distributed, with the means and standard deviations in Table 1; these values were obtained from [2].

3. Formulation of the Inverse Problem

Let $\mathbf{I}$ be the set of infection dates I_i for the 30 cases. Let $\mathbf{R}$ be the set of dates of removal and $\mathbf{S}$ be the dates when the cases showed symptoms (the data). Then, if two cases j and k were known to have a social link between them, then the probability that j showed symptoms on S_j, conditioned on the infection date I_k and transmission rate β_{jk} is

$$\mathcal{L}^{(A)}_{(j,k)}(S_j|\beta_{jk}, I_k) = \beta_{jk} \exp(-\beta_{jk}(I_k - S_j))$$

where $\beta_{jk} = \beta_{\text{in}}$ or β_{out} depending upon whether j and k belong to the same or different compound. Thus the probability of observing $\mathbf{S}$, given the infection pathway (the directed graph of links $j \to k$ over which the disease has been transmitted) $\mathcal{P}$

$$\mathcal{L}^{(A)}(\mathbf{S}|\mathbf{I}, \vec{\beta}, \mathcal{P}) = \prod_{(j,k)\in\mathcal{P}} \exp(-\beta_{jk}(I_k - S_j)) \tag{1}$$

where $\vec{\beta} = \{\beta_{\text{in}}, \beta_{\text{out}}\}$ and (j,k) denotes the directed link $j \to k$.

$\mathcal{G}$ is the undirected graph of social contacts. $\mathcal{P}$ contains a subset of the links in $\mathcal{G}$ over which disease transmission occurred; the remaining links in $\mathcal{G}$ did not transmit the disease. These include links (j,k) between two nodes (people) who never contracted the disease, links where $j \in \mathcal{P}$ but k is not (and thus transmission over (j,k) never occurred) and $j,k \in \mathcal{P}$ but link $(j,k) \notin \mathcal{P}$. Given our Poisson model of transmission over links, the probability of escaping infection having been in contact for time τ_{jk} is $\exp(-\beta_{jk}\tau_{jk})$. Thus the probability of observing $\mathbf{S}$ given $\mathbf{I}, \mathbf{R}, \vec{\beta}, \mathcal{P}, \mathcal{G}$ is

$$\begin{aligned} \mathcal{L}^{(B)}(\mathbf{S}|\mathbf{I},\mathbf{R},\vec{\beta},\mathcal{P},\mathcal{G}) &= \prod_{(j,k)\in\mathcal{G}\backslash\mathcal{P}, j\in\mathcal{P}} \exp(-\beta_{jk}\tau_{jk}), \\ \tau_{jk} &= \max(\min(I_k, R_j) - S_j, 0) \end{aligned} \tag{2}$$

where $(j,k) \in \mathcal{G} \setminus \mathcal{P}$ are the set of links that exist in $\mathcal{G}$ but not in $\mathcal{P}$. For j who never contracted the disease, $I_j = \infty$ in Eq. 2.

Let a social network $\mathcal{G}$ contain $\mathbf{n} = \{n_1, n_2, \ldots n_7\}$ in-compound links in the 7 compounds and $\mathbf{m} = \{m_1, m_2, \ldots m_7\}$ out-of-compound links. Here, n_i refers to the total number of in-compound links that the residents of compound "i" have; m_i is corresponding number for out-of-compound links. Then given $\mathbf{p} = \{p_{\text{in}}, p_{\text{out}}\}$, the probability of observing $\mathcal{G}$ is

$$\mathcal{L}^{(E)}(\mathcal{G}|\mathbf{p}) = \prod_{i=1}^{7} p_{\text{in}}^{n_i}(1-p_{\text{in}})^{C(N_i)-n_i} p_{\text{out}}^{m_i/2}(1-p_{\text{out}})^{D(N_i,m_i)} \tag{3}$$

where

$$\begin{aligned} C(N_i) &= \begin{pmatrix} N_i \\ 2 \end{pmatrix}, \\ D(N_i, m_i) &= \frac{(N_{tot} - N_i)N_i - m_i}{2}, \end{aligned} \tag{4}$$

N_i is the FTC population in compound i, and $N_{tot} = 74$ is the total FTC population.

Since an infection pathway $\mathcal{P}$ is always contained in $\mathcal{G}$, and all the infection pathways supported by $\mathcal{G}$ are equally probable, we set the probability of observing $\mathcal{P}$ conditioned on $\mathcal{G}$ to a constant, i.e $\mathcal{L}^{(F)}(\mathcal{P}|\mathcal{G}) =$ constant.

The probability of observing a set of symptom times $\mathbf{S}$, conditioned on the set of infection and removal times, is given by

$$\mathcal{L}^{(D)}(\mathbf{S}|\mathbf{I},\mathbf{R}) = \prod_{i=1}^{30} p_l(S_i - I_i) p_R(R_i - S_i) \tag{5}$$

where $p_l(t)$ and the $p_R(t)$ are the probability densities of the latent and contagious periods of smallpox. These periods are modeled by gamma distributions, as prescribed in Section 2.

Thus the probability of observing $\mathbf{S}$ is

$$\mathcal{L}(\mathbf{S}|\mathbf{I},\mathbf{R},\vec{\beta},\mathbf{p},\mathcal{P},\mathcal{G}) = \mathcal{L}^{(A)}\mathcal{L}^{(B)}\mathcal{L}^{(D)}\mathcal{L}^{(E)}\mathcal{L}^{(F)} \tag{6}$$

Using Bayes rule, we write an expression for the joint posterior probability of $\mathbf{I}, \mathbf{R}, \vec{\beta}, \mathbf{p}, \mathcal{P}, \mathcal{G}$ conditioned on $\mathbf{S}$

$$p(\mathbf{I},\mathbf{R},\vec{\beta},\mathbf{p},\mathcal{P},\mathcal{G}|\mathbf{S}) \propto \mathcal{L}(\mathbf{S}|\mathbf{I},\mathbf{R},\vec{\beta},\mathbf{p},\mathcal{P},\mathcal{G})\pi^I(\mathbf{I})\pi^R(\mathbf{R})\pi^\beta(\vec{\beta})\pi^p(\mathbf{p}) \tag{7}$$

where $\pi(\cdot)$ are probability density functions (PDFs) representing prior information on the various parameters. Given the data in Section 2, it is realistic to assume that $\beta_{\text{in}} \geq \beta_{\text{out}}$ and $p_{\text{in}} \geq p_{\text{out}}$. We express these relations as

$$\begin{aligned} \beta_{\text{in}} &= (1+r)\beta_{\text{out}}, \quad r \geq 0, \beta_{\text{out}} \geq 0 \\ p_{\text{out}} &= \rho p_{\text{in}}, \quad \rho \leq 1, p_{\text{in}} \leq 1 \end{aligned} \tag{8}$$

and introduce them in Eq. 7 to derive an expression for the posterior $p(\mathbf{I},\mathbf{R},r,\beta_{\text{out}},p_{\text{in}},\rho,\mathcal{P},\mathcal{G}|\mathbf{S})$ as

$$\begin{aligned} & p(\mathbf{I},\mathbf{R},r,\beta_{\text{out}},p_{\text{in}},\rho,\mathcal{P},\mathcal{G}|\mathbf{S}) \propto \\ & \quad \mathcal{L}(\mathbf{S}|\mathbf{I},\mathbf{R},r,\beta_{\text{out}},p_{\text{in}},\rho,\mathcal{P},\mathcal{G})\pi^I(\mathbf{I})\pi^R(\mathbf{R})\pi^r(r)\pi^{\beta_{\text{out}}}(\beta_{\text{out}})\pi^{p_{\text{in}}}(p_{\text{in}})\pi^\rho(\rho) \end{aligned} \tag{9}$$

We employ uniform distributions for all the priors. Infection times $\mathbf{I}$ are believed to lie uniformly between 0 and 30 days before the date of exhibition of symptoms by the patient, while the removal times $\mathbf{R}$ lie between 0 and 40 after the date of appearance of symptoms. The prior on r is $\mathcal{U}(0, 40)$, while the prior distribution on β_{out} is $\mathcal{U}(0, 10)$. The priors on both ρ and p_{in} are $\mathcal{U}(0, 1)$.

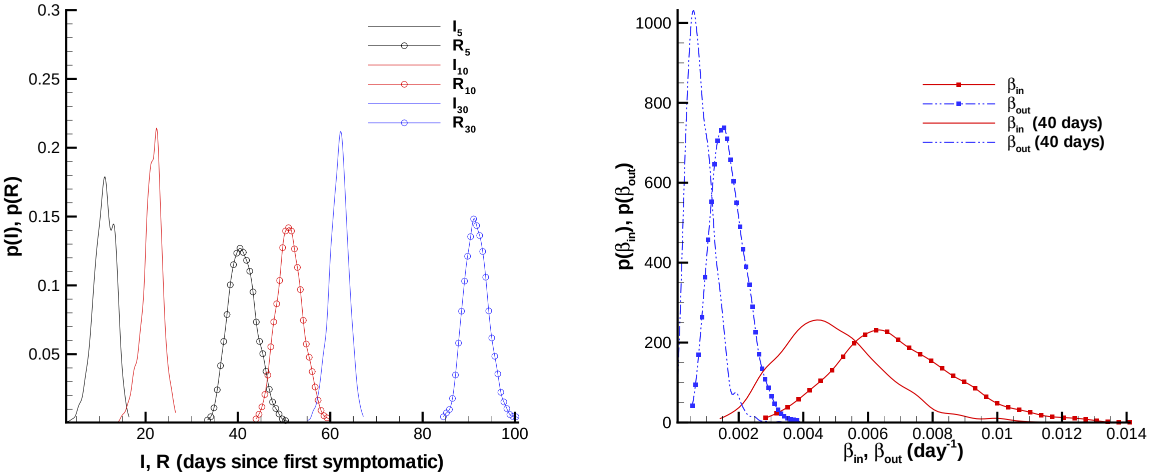

Figure 1: Left: Posterior PDFs for the infection and removal dates of cases 5, 10, and 30 in Problem I. Cases are denoted by distinct colors; plots of removal dates contain a symbol. Right: Posterior PDFs for the rates of spread β_{in} and β_{out} developed from data collected during the entire epidemic (plots with symbols) and data from the first 40 days (plots without symbols). MAP estimates of the spread rates drawn from the first 40 days may be affected by observational error.

Characterizing the epidemic now requires simulating from the posterior $p(\mathbf{I}, \mathbf{R}, r, \beta_{\text{out}}, p_{\text{in}}, \rho, \mathcal{P}, \mathcal{G}|\mathbf{S})$ in Eq. 9. The posterior distribution cannot be expressed in terms of canonical distributions, and hence we apply Metropolis-Hastings MCMC. In particular, we use an independence sampler for the duration of incubation and removal periods as well as for $\mathcal{P}$ and $\mathcal{G}$, in a manner similar to [7]. We also carry out the following transformations:

$$\begin{array}{rcll} \xi_{\text{in}} & = & \log(r), & -\infty < \xi_{\text{in}} < \infty \\ \xi_{\text{out}} & = & \log(\beta_{\text{out}}), & -\infty < \xi_{\text{in}} < \infty \\ \eta_{\text{in}} & = & \text{logit}(p_{\text{in}}), & -\infty < \eta_{\text{in}} < \infty \\ \eta_{\text{out}} & = & \text{logit}(\rho), & -\infty < \eta_{\text{out}} < \infty \end{array}$$

and use random-walk proposals for the remaining parameters $(\xi_{\text{in}}, \xi_{\text{out}}, \eta_{\text{in}}, \eta_{\text{out}}) = (\vec{\xi}, \vec{\eta})$. Our overall MCMC algorithm thus consists of componentwise Metropolis-Hastings updates for $\mathbf{I}$, $\mathbf{R}$, $\vec{\xi}$, $\vec{\eta}$, $\mathcal{P}$, and $\mathcal{G}$. Marginal distributions or estimates of interest are reconstructed from the chain. Note that the priors are expressed in terms of distributions on r, β_{out}, p_{in}, and ρ and are transformed to equivalent distributions on $\vec{\xi}$ and $\vec{\eta}$ for the purpose of computation.

4. Results

The Abakaliki data described in Section 2—i.e., the dates of symptoms $\mathbf{S}$—do not carry enough information to provide useful joint estimates of $\mathbf{I}$, $\mathbf{R}$, r, β_{out}, p_{in}, ρ, $\mathcal{P}$, and $\mathcal{G}$. There are simply too many parameters/variables to be inferred from relatively little data. However, the problem can be significantly simplified with a few fairly realistic assumptions. These assumptions result in alternate inferential problems which still extract useful information from the data. These problems are described below.

Problem I: In this case, we simplify the problem by assuming a fully connected population (i.e., $\rho = 1, p_{\text{in}} = 1$). However, in-compound and cross-compound social links differ in strength. The observed differential rates of spread are accommodated by having two distinct spread rates, β_{in} and β_{out}, on in-compound and cross-compound links respectively. This requires that we estimate r and β_{out} from the data. Further, while there is a unique $\mathcal{G}$ (given full connectivity), there are many possible $\mathcal{P}$. In this analysis, we will develop estimates for $\mathbf{I}$, $\mathbf{R}$, β_{in}, β_{out}, and the disease transmission chain $\mathcal{P}$.

The MCMC chain was run for 20000 iterations and thinned (for convenience in storage/post-processing) by saving samples every 10 iterations. Good mixing was observed for the last 15000 iterations (i.e., the last 1500 saved samples), and thus we used these samples to compute posterior estimates. In Figure 1 (left), we plot marginal posterior densities for the dates of infection and removal of 3 different cases (i.e., 3 infected individuals). While the ranges of the PDFs are rather wide ($\approx$ 10 days for the dates of infection and 15 for dates of removal), they are roughly symmetric. However, individual dates of infection and removal are largely nuisance variables and the specificity with which one can infer their values is not very important. In Figure 1 (right), we show posterior PDFs for both β_{in} and β_{out} (lines with symbols). The MAP (maximum a posteriori) estimate of β_{in} is roughly four times larger than the MAP estimate of β_{out}, corroborating the faster spread of smallpox inside a compound that was qualitatively observed in the data. Additional marginal posterior densities are available in [12]. For instance, we report the two-dimensional posterior marginal of $(\beta_{\text{in}}, \beta_{\text{out}})$; very little correlation was observed between the two spread rates.

In Figure 2, we plot the expected infection pathway $\langle \mathcal{P} \rangle$ as a directed graph. The 30 smallpox cases correspond to nodes in the graph and are colored per their compound affiliations. The color of each link indicates its probability of existence—all links with a probability of 0.3 or higher are in red and those between 0.1 and 0.3 are in blue. We see a clear transmission of the disease from the index case (Node 000) to other members of her compound; that is, most of the links originating from Node 000 are red and connect to other members of her compound. Furthermore, most of the red links exist between nodes of the same color, indicating cohabitation in the same compound. Cross-compound links are generally blue, indicating less certainty regarding their existence; however, this lack of certainty is balanced by their larger number, which enables inter-compound transmission. Also, higher-numbered nodes (which became infected and symptomatic later in time) have a larger proportion of blue links connected to them (e.g., Node 000 has only one, while Node 029 has four), indicating an inability to "pin down" the source of their infection as well as the identity of the individuals they infected. This is expected—as the number of infected individuals increased and became a substantial fraction of the total population of 74, the infection mechanism approached that of a homogeneously mixed population, i.e., where one has an equal probability of being infected by any contagious individual.

The results presented above were obtained with data collected during the entire 90-day duration of the epidemic. We now consider more restricted observations. By Day 40 of the outbreak, the disease had spread outside Compound 1 and thus there was some (slight) evidence of inter-compound transmission. Using the data collected by Day 40, we infer the same transmission parameters. In Figure 1 (right), we plot posterior PDFs for β_{in} and β_{out} using lines without symbols. While the widths of these PDFs are about the same as their 90-day counterparts, the MAP value of β_{in} is about a third smaller, as is the MAP value for β_{out}. The cause for this underestimation is being investigated, but may be due to errors in the *early* data. While the epidemic started in early April 1967, it was reported to the World Health Organization (which gathered the data) in the later part of May [11]. Further, the dates of appearance of symptoms were obtained by interviewing the families of infected cases rather than by documentary proof; thus, there was ample scope for the introduction of "observational" errors due to faulty memories.

Problem II: In this problem, we investigate a partially connected society. We assume, realistically, that the rate of spread along any social link is the same, i.e, $\beta_{\text{in}} = \beta_{\text{out}}$ (or equivalently $r = 0$), and that the slower spread of the disease across compounds is due to the paucity of "strong" social links across compounds. We assume that in-compound mixing is strong—that is, we have a fully connected social network inside a compound (i.e., $p_{\text{in}} = 1$)—and that a sparse social network exists between individuals across compounds (i.e., $\rho < 1$). In this case we develop posterior PDFs for $\beta = \beta_{\text{in}} = \beta_{\text{out}}$ and $p_{\text{out}} = \rho$. We also generate posterior samples of $\mathcal{P}$ as in the previous case, as well as $\mathcal{G}$. We perform the analysis using the first 40 days of data. We impose a $\mathcal{U}(0.25, 0.75)$ prior on ρ, thus ruling out a fully connected social network.

As in Problem I, we examined the MCMC chain to ensure good mixing, and used the last 15000 samples to characterize the posterior. Traces of the chain and autocorrelations are given in [12]. Figure 3 (left) shows marginal posterior densities for the dates of infection and recovery of cases 5 and 10. (Case 30 is not shown here because it did not exhibit symptoms in the first 40 days). These PDFs are quite similar in shape and scale to those obtained in Problem I. Modes of $p(I)$ and $p(R)$ in Problem II are shifted to the right by about 1 day. (n other words, Problem II suggests that infections and removals happened a day later, compared to Problem I. In relative terms, these two models thus show very little difference in the inferred infection and removal dates. In Figure 3 (right), we show the posterior PDF for β. We also plot the corresponding in-compound and out-of-compound transmission rates from Problem I, also inferred from the first 40 days of epidemic data. The MAP estimate of β lies between the MAP estimates for β_{in} and β_{out} from Problem I, though closer to β_{in}. If one compares median values, β from Problem II ($3.37 \times 10^{-3}\text{day}^{-1}$) is clearly closer to β_{in} as estimated in Problem I ($4.76 \times 10^{-3}\text{day}^{-1}$) than to β_{out} ($0.77 \times 10^{-3}\text{day}^{-1}$). This is intuitively correct, since the bulk of the transmission was in-compound. In Figure 4 we plot the expected infection pathway $\langle \mathcal{P} \rangle$. Only 13 cases were observed before day 40, so this graph is smaller than that of Figure 2. Again, transmission links can be inferred with high probability only for the first few cases; the links connected to the latter cases are generally blue. Overall, the transmission chains estimated using the two different models are quite comparable. A plot of the expected social network $\langle \mathcal{G} \rangle$ can be found in [12].

5. Conclusions

We have developed a Bayesian inference method that extracts disease spread rates and infers chains of transmission from partial observations of an epidemic. The method is novel in the sense that it explicitly accounts for the role of pathogenic transmissibility and mixing (i.e., social networks) in the spread of a disease and infers them from the data. This estimation capability can be applied in a variety of ways. For example, the posterior realizations of the social network could easily be used in a network epidemic model for any other disease (e.g., influenza). On the other hand, estimates of the transmissibility $\vec{\beta}$ can be used to predict the evolution of emerging infectious diseases; such efforts have already begun to appear [13, 1]. Furthermore, inference of a transmission chain can often be helpful in identifying the *mechanism* of transmission. For example, in-family

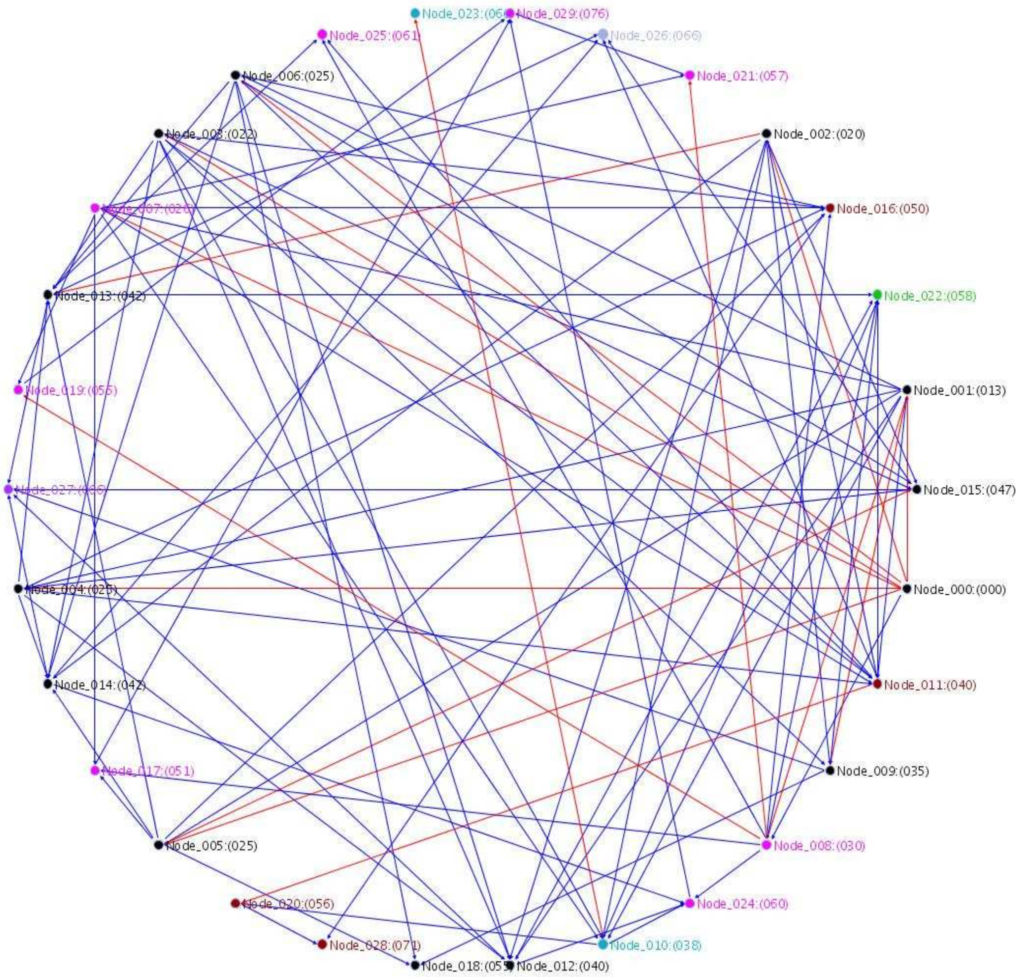

Figure 2: The expected infection pathway $\langle \mathcal{P} \rangle$, based on data collected during the entire epidemic for Problem I. Nodes represent the 30 cases of the outbreak and are colored by their compound affiliations. Edges in the graph are colored by their probability of existence—links with probability of 30% or higher are in red while those between 10% and 30% are in blue. Most of the transmission from the index case (Node 000) is in red, and connects individuals in the same compound. Infection transmissions between the later cases are almost completely in blue, indicating less heterogeneity in the transmission mechanism as a large fraction of the population became infected. The number in parentheses after each node's label is the day the case exhibited symptoms.

transmission is not very strong for Ebola (in sub-Saharan Africa), since it mainly spreads via contaminated bodily fluids, most commonly when preparing a contaminated corpse for final disposal. This task is generally performed by the women of a village [14], leading to strong *cross*-family transmission. Such a mechanism would be reflected in both $\vec{\beta}$ and $\langle \mathcal{P} \rangle$.

The method developed here, however, is not yet completely mature. Methodologically, we observe that mixing of the MCMC chain is difficult when $\langle \mathcal{G} \rangle$ has to be inferred, primarily because p_{out} seems to obey a bimodal distribution (with a peak near 1). A mode-hopping MCMC approach [15, 16] may ameliorate the problem and is currently being explored. Further, as mentioned in Section 1, the use of a binomial graph as a model for the social network involves a significant simplifying assumption; its effect on the overall validity of the method has to be determined. Adopting a more realistic social network model (e.g., a scale-free network) would be a welcome step, but may introduce further methodological challenges, particularly in the design of the MCMC algorithm. This is left for future work.

Acknowledgements

This work was supported by the Sandia National Laboratories LDRD (Laboratory Directed Research and Development) program, with funds allocated under the "Enabling Predictive Simulation" investment area. Sandia is a multiprogram laboratory operated by Sandia Corporation, a Lockheed Martin Company, for the United States Department of Energy's National Nuclear Security Administration under Contract DE-AC04-94-AL85000.

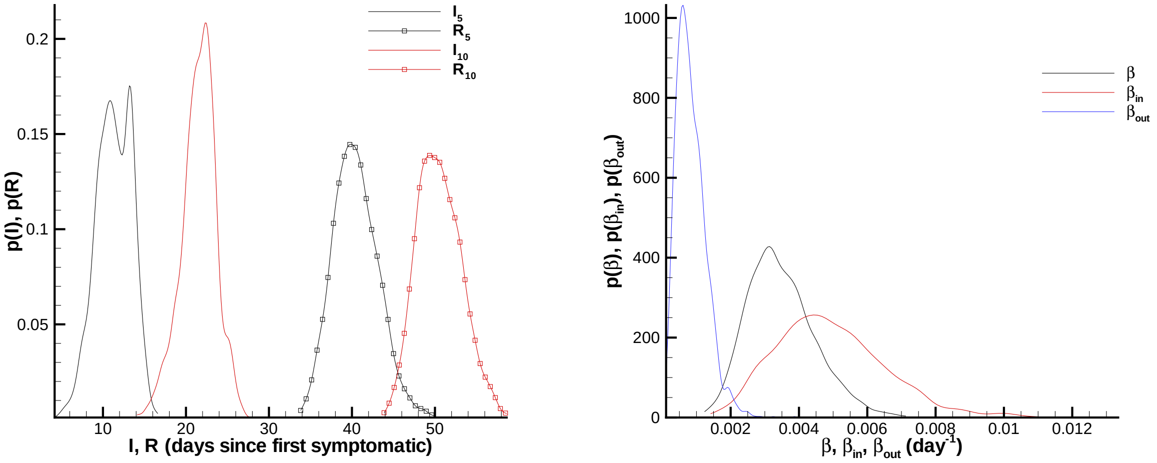

Figure 3: Left: Posterior PDFs for the dates of infection and removal of cases 5 and 10 in Problem II. The cases are denoted by distinct colors; plots of removal dates contain a symbol. Right: Posterior PDF for the rate of spread β in Problem II, along with posterior PDFs for β_{in} and β_{out} in Problem I, based on the first 40 days of data.

References

[1] Yang Yang, M. Elizabeth Halloran, Johathan D. Sugimota, and Ira M. Longini. Detecting human-to-human transmission of avian influenza A (H5N1). *Emerging Infectious Diseases*, 13(9):1348–1353, 2007.

[2] Martin Eichner and Klaus Dietz. Transmission potential of smallpox: Estimates based on detailed data from an outbreak. *American Journal of Epidemiology*, 158(2):110–117, 2003.

[3] S. Cauchemez amd F. Carrat, C. Viboud, A. J. Valleron, and P. Y. Boelle. A Bayesian MCMC approach to study transmission of influenza: application to household longitudinal data. *Statistics in Medicine*, 23:3469–3487, 2004.

[4] Simon Cauchemez, Alain-Jacques Valleron, Pierre-Yves Boelle, Antoine Flahault, and Neill M. Ferguson. Estimating the impact of school closure on influenza transmission from Sentinel data. *Nature*, 452:750–755.

[5] P. D. O'Neill and G. O. Roberts. Bayesian inference of partially observed stochastic epidemics. *Journal of the Royal Statistical Society, Series A*, 162:121–129, 1999.

[6] Michael Hohle, Erik Jorgensen, and Philip D. O'Neill. Inference in disease transmission experiments by using stochastic epidemic models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 54(2):349–366, 2005.

[7] Tom Britton and P. O'Neill. Bayesian inference for stochastic epidemics in populations with random social structure. *Scandinavian Journal of Statistics*, 29:375–390, 2002.

[8] Nikolaos Demiris and Philip D. O'Neill. Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(5):731–745, 2005.

[9] G. Chowell, J. M. Hyman, S. Eubank, and C. Castillo-Chavez. Scaling laws for the movement of people between locations in a large city. *Physical Review E*, 68(066102):1–7, 2003.

[10] Tom Britton, Maria Deijfen, and Anders Martin-Lof. Generating simple random graphs with prescribed degree distribution. *Journal of Statistical Physics*, 124(6):1377–1397, 2006.

[11] D. Thompson and William Foege. Faith tabernacle smallpox epidemic. Technical Report WHO/SE/68.3, World Health Organization, 1968.

[12] J. Ray, B. M. Adams, K. D. Devine, Y. M. Marzouk, M. M. Wolf, and H. N. Najm. Distributed micro-releases of bioterror pathogens: threat characterizations and epidemiology from uncertain patient observables. SAND Report SAND2008-6044, Sandia National Laboratories, Livermore, CA 94551-0969, October 2008. Unclassified unlimited release.

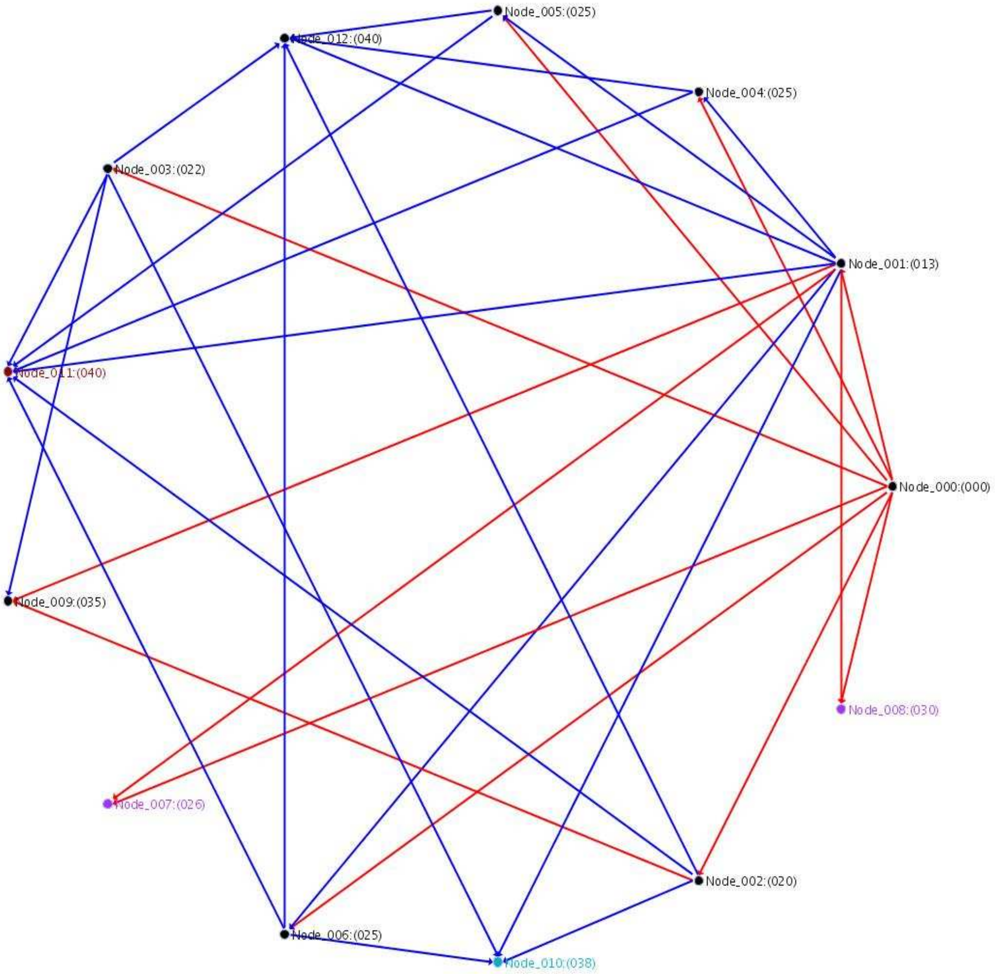

Figure 4: The expected infection pathway $\langle \mathcal{P} \rangle$, based on data collected during the first 40 days of the epidemic for Problem II. As in Figure 2, nodes represent the cases of the outbreak and are colored by their compound affiliations. Edges in the graph are colored by their probability of existence—links with probability of 30% or higher are in red while those between 10% and 30% are in blue. Most of the transmission from the index case (Node 000) is in red, and connects individuals in the same compound. Infection transmissions between the later cases are almost completely in blue, indicating a more homogeneous transmission mechanism as a large fraction of the population became infected. The number in parentheses after each node's label indicates the day the case exhibited symptoms.

[13] Gerardo Chowell, Hiroshi Nishiura, and Luis M. A. Bettencourt. Comparative estimation of the reproduction number for pandemic influenza from daily case notification data. *Journal of the Royal Society - Interface*, 4, 2007.

[14] G. Chowell, N. W. Hengartner, C. Castillo-Chavez, P. W. Fenimore, and J. M. Hyman. The basic reproductive number of Ebola and the effects of public health measures: The case of Congo and Uganda. Technical Report LA-UR-03-8189, Los Alamos National Laboratory, Los Alamos, NM, 2003.

[15] Hakon Tjelmeland and Jo Eidsvik. On the use of local optimizations within Metropolis-Hastings updates. *Journal of the Royal Statistical Society, B*, 66(2):411–427, 2004.

[16] Cristian Sminchisescu, Max Welling, and Geoffrey Hinton. A mode-hopping MCMC sampler. Technical Report CSRG-478, University of Toronto, 6 King's College Road, Pratt Building, Toronto, Ontario, CANADA, M5S 3G4, 2003. http://www.cs.toronto.edu/ crismin/.