A Semiparametric Bayesian Approach to Epidemics, with Application to the Spread of the Coronavirus MERS in South Korea in 2015

Michael Schweinberger[*] Rice University

Rashmi P. Bomiriya Penn State University

Sergii Babkin Rice University

[*] Corresponding author. Address: Department of Statistics, Rice University, 6100 Main St, Houston, TX 77005, USA. Email: m.s@rice.edu. Phone: + 1 713-348-2278. Fax: +1 713-348-5476.

A Semiparametric Bayesian Approach to Epidemics, with Application to the Spread of the Coronavirus MERS in South Korea in 2015

Abstract

We consider incomplete observations of stochastic processes governing the spread of infectious diseases through finite populations by way of contact. We propose a flexible semiparametric modeling framework with three advantages. First, it enables researchers to study the structure of population contact networks and their impact on the spread of infectious diseases. Second, it can accomodate both short- and long-tailed degree distributions and can detect potential superspreaders, which represent an important public health concern. Third, it addresses the important issue of incomplete data. Starting from first principles, we show when the incomplete-data generating process is ignorable for the purpose of likelihood-based inference for the parameters of the population model. We demonstrate the proposed framework by simulations and an application to the partially observed MERS epidemic in South Korea in 2015, which was driven by the coronavirus MERS, related to both SARS and COVID-19.

Keywords: Contact networks; Network sampling; Link-tracing; Missing data.

MSC subject classifications: 05C80, 05C81.

1 Introduction

The spread of infectious diseases (e.g., HIV, Ebola, SARS, MERS, COVID-19) through populations by way of contact represents an important public health concern.

A network-based approach to modeling the spread of infectious diseases is appealing, because the network of contacts in a population determines how infectious diseases can spread (e.g., Keeling and Eames, 2005; Danon et al., 2011; Welch, Bansal, and Hunter, 2011). One of the advantages of a network-based approach is that heterogeneity in the number of contacts can be captured (e.g., Danon et al., 2011), along with other features of population contact networks (e.g., Welch et al., 2011). Indeed, conventional models of epidemics—including classic and lattice-based Susceptible-Infectious-Recovered (SIR) and Susceptible-Exposed-Infectious-Recovered (SEIR) models (e.g., Andersson and Britton, 2000; Danon et al., 2011)—can be considered to be degenerate versions of network-based models, in the sense that such models postulate that with probability 1 the population contact network is of a known form: e.g., with probability 1 each population member is in contact with every other population member. Worse, the postulated form of the population contact network may not resemble real-world contact networks. A second advantage is that a network-based approach helps study the structure of a population contact network and its generating mechanism, helping generalize findings to populations with similar population contact networks.

Motivated by the shortcomings of classic and lattice-based SIR and SEIR models of epidemics, Britton and O'Neill (2002), Groendyke, Welch, and Hunter (2011, 2012), and others explored a network-based approach to epidemics. However, while a network-based approach is more appealing than classic and lattice-based SIR and SEIR models, existing network-based models of epidemics are either not flexible models of degree distributions or induce short-tailed degree distributions, as described in Section 3.1. Short-tailed degree distributions are problematic, because degree distributions of real-world contact networks are thought to be long-tailed (e.g., Jones and Handcock, 2003b, 2004) and the population members in the upper tail of the degree distribution represent an important public health concern: Population members with many contacts can infect many others and are hence potential superspreaders. Indeed, there is circumstantial evidence to suggest that superspreaders have played a role in the SARS epidemic in 2002–2003, the MERS epidemic in 2015, and the ongoing COVID-19 pandemic.

We introduce a flexible semiparametric modeling framework with three advantages. First, it shares with existing network-based approaches the advantage that it enables researchers to study the structure of population contact networks and their impact on the spread of infectious diseases. Second, in contrast to existing network-based approaches, it is a flexible model of both short- and long-tailed degree distributions and can detect potential superspreaders. Third, it addresses the important issue of incomplete data and can deal with a wide range of missing data and sample data. In fact, complete observations of epidemics are all but impossible, making it imperative to deal with incomplete data. Starting from first principles, we show when the incomplete-data generating process is likelihood-ignorable, that is, ignorable for the purpose of likelihood-based inference for the parameters of the population model. In addition, we discuss likelihood-ignorable sampling designs for collecting contact and epidemiological data, with a view to reducing the posterior uncertainty about the population contact network and its generating mechanism. We demonstrate the proposed framework by simulations and an application to the partially observed MERS epidemic in South Korea in 2015 (Ki, 2015). The MERS epidemic was driven by the coronavirus MERS, which is related to the coronaviruses SARS and COVID-19. We detect three superspreaders, who directly or indirectly infected all other 183 infected population members.

The remainder of our paper is organized as follows. We review existing parametric population models in Section 2 and introduce semiparametric population models in Section 3. Bayesian inference given likelihood-ignorable incomplete-data generating processes is discussed in Section 4. We present simulations in Section 5 and an application to the partially observed MERS epidemic in South Korea in Section 6. We conclude with an extended discussion of open questions in Section 7.

2 Parametric population models

We provide a short overview of selected network-based parametric population models of epidemics. To do so, we first describe a generic data-generating process in Section 2.1 and then review network-based parametric population models in Section 2.2.

2.1 Data-generating process

We consider a population with $N < \infty$ population members, who may be connected by contacts. In the simplest case, contacts among population members are time-invariant and are either absent or present. We discuss in Section 7.6 possible extensions to time-evolving population contact networks.

A basic data-generating process is shown in Figure 1 and can be described as follows:

- Generate a population contact network.
- Conditional on the population contact network, generate an epidemic.

The population contact network is generated by a random graph model as described in Section 2.2 (parametric) and Section 3.2 (semiparametric). Conditional on the population contact network, an infectious disease spreads through the population by way of contact, governed by a continuous-time stochastic process such as the SIR and SEIR model (Andersson and Britton, 2000; Britton and O'Neill, 2002; Groendyke et al., 2011, 2012). We focus on the network-based SEIR model, which can be sketched as follows (Britton and O'Neill, 2002; Groendyke et al., 2011, 2012). In the simplest case, the initial state of the stochastic process consists of a population with one infected population member and $N-1$ susceptible population members. Infected population members pass through three states: the exposed state; the infectious state; and the removed state. In the exposed state, population members are infected, but cannot infect others. In the infectious state, population members can infect susceptible population members by contact, with transmissions being independent across contacts. In the final state—the removed state—population members cannot infect others, either because they have recovered and are immune to re-infection or because they have died.

The epidemic continues until all infected population members are removed from the population. All of the described events—the event that an infectious population member infects a susceptible population member, and the transitions of infected population members from the exposed to the infectious state, and from the infectious state to the removed state—are independent and occur at random times. The waiting times until these events occur follow Exponential or Gamma distributions. More specific assumptions about the distributions of waiting times and the population contact network are detailed in Section 2.2 and in the monographs of Andersson and Britton (2000) and Britton and Pardoux (2019).

2.2 Parametric population models

Consider an epidemic that started at time 0 and ceased by time $0 < t < \infty$. We assume that the identities of infected population members are known and denoted by $1, \ldots, M$, where $M \in \{1, \ldots, N\}$. The population contact network is represented by $\boldsymbol{Y} = \{Y_{i,j}\}_{i<j}^{N} \in \mathcal{Y} = \{0,1\}^{\binom{N}{2}}$, where $Y_{i,j} = 1$ if population members i and j are in contact during the epidemic and $Y_{i,j} = 0$ otherwise. Since contacts are undirected and self-contacts are meaningless, we assume that $Y_{i,j} = Y_{j,i}$ and $Y_{i,i} = 0$ hold with probability 1. The transmissions are denoted by $\boldsymbol{T} = \{T_{i,j}\}_{i \neq j}^{N}$, where $T_{i,j} = 1$ if i infects j and $T_{i,j} = 0$ otherwise. Observe that $Y_{i,j} = 0$ implies $\max(T_{i,j}, T_{j,i}) = 0$ whereas $\max(T_{i,j}, T_{j,i}) = 1$ implies $Y_{i,j} = 1$ with probability 1. The starting time of the exposure, infectious, and removal period of population members are denoted by $\boldsymbol{E} = \{E_i\}_{i=1}^{N} \in \mathbb{R}_{+}^{N}$, $\boldsymbol{I} = \{I_i\}_{i=1}^{N} \in \mathbb{R}_{+}^{N}$, and $\boldsymbol{R} = \{R_i\}_{i=1}^{N} \in \mathbb{R}_{+}^{N}$, respectively, where $\mathbb{R}_{+} = (0, \infty)$ and $E_i < I_i < R_i$ holds with probability 1; note that E_i, I_i, R_i are undefined if population member i was not infected. We write $\boldsymbol{X} = (\boldsymbol{E}, \boldsymbol{I}, \boldsymbol{R}, \boldsymbol{T})$ and, in a mild abuse of language, we refer to $\boldsymbol{X}$ as an epidemic.

The complete-data likelihood function, given complete observations $\boldsymbol{x}$ and $\boldsymbol{y}$ of the epidemic $\boldsymbol{X}$ and the population contact network $\boldsymbol{Y}$, is of the form

$$L(\boldsymbol{\eta}, \boldsymbol{\theta};\, \boldsymbol{x}, \boldsymbol{y}) \;\;\propto\;\; L(\boldsymbol{\eta}_E;\, \boldsymbol{x})\; L(\boldsymbol{\eta}_I;\, \boldsymbol{x})\; L(\beta;\, \boldsymbol{x}, \boldsymbol{y})\; L(\boldsymbol{\theta};\, \boldsymbol{y}), \tag{1}$$

where $\boldsymbol{\eta} = (\boldsymbol{\eta}_E, \boldsymbol{\eta}_I, \beta) \in \Omega_{\boldsymbol{\eta}} \subseteq \mathbb{R}^{d_1}$ $(d_1 \geq 1)$ and $\boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}} \subseteq \mathbb{R}^{d_2}$ $(d_2 \geq 1)$ are the parameter vectors of the population model generating the epidemic $\boldsymbol{X}$ and the population contact network $\boldsymbol{Y}$, respectively. We describe each component of the complete-data likelihood function in turn, along with its parameters.

The components $L(\boldsymbol{\eta}_E;\, \boldsymbol{x})$ and $L(\boldsymbol{\eta}_I;\, \boldsymbol{x})$ of the likelihood function are of the form

$$L(\boldsymbol{\eta}_E;\, \boldsymbol{x}) \;\;\propto\;\; \prod_{i=1}^{M} p(I_i - E_i \mid E_i, \boldsymbol{\eta}_E) \tag{2}$$

and

$$L(\boldsymbol{\eta}_I;\, \boldsymbol{x}) \;\;\propto\;\; \prod_{i=1}^{M} p(R_i - I_i \mid I_i, \boldsymbol{\eta}_I), \tag{3}$$

where $p(. \mid E_i, \boldsymbol{\eta}_E)$ and $p(. \mid I_i, \boldsymbol{\eta}_I)$ are densities with suitable support parameterized by $\boldsymbol{\eta}_E$ and $\boldsymbol{\eta}_I$, respectively: e.g., the densities may be Gamma densities, with $\boldsymbol{\eta}_E \in \mathbb{R}^+ \times \mathbb{R}^+$ and $\boldsymbol{\eta}_I \in \mathbb{R}^+ \times \mathbb{R}^+$ being scale and shape parameters of Gamma densities.

Under the assumption that the waiting times until infectious population members infect susceptible population members are independent Exponential(β) random variables with rate of infection $\beta \in \mathbb{R}_+$, the component $L(\beta; \boldsymbol{x}, \boldsymbol{y})$ of the likelihood function is given by

$$L(\beta;\, \boldsymbol{x}, \boldsymbol{y}) \;\;\propto\;\; \beta^{M-1} \exp(-\beta\, a(\boldsymbol{x}, \boldsymbol{y})), \tag{4}$$

where $a(.) > 0$ is defined by

$$a(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{M} \sum_{j=1}^{M} y_{i,j}\, 1_{I_i < E_j} \max(\min(E_j, R_i) - I_i,\, 0) + \sum_{i=1}^{M} d_i(\boldsymbol{x}, \boldsymbol{y})\, (R_i - I_i), \tag{5}$$

where $1_{I_i < E_j}$ is 1 in the event of $I_i < E_j$ and 0 otherwise and $d_i(\boldsymbol{x}, \boldsymbol{y})$ is the number of non-infected population members in contact with population member i. The function $a(\boldsymbol{x}, \boldsymbol{y})$ was derived in Britton and O'Neill (2002) and Groendyke et al. (2011).

To represent existing (parametric) and proposed (semiparametric) population models along with possible extensions in a unifying framework, it is convenient to parameterize the distribution of the population contact network $\boldsymbol{Y}$ in exponential-family form (Brown, 1986). The component $L(\boldsymbol{\theta}; \boldsymbol{y})$ of the likelihood function can then be written as

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \;\;\propto\;\; \exp\left(\boldsymbol{\theta}^\top s(\boldsymbol{y}) - \psi(\boldsymbol{\theta})\right), \quad \boldsymbol{y} \in \mathcal{Y}, \tag{6}$$

where $\boldsymbol{\theta}$ is a d-vector of natural parameters, $s(\boldsymbol{y})$ is a d-vector of sufficient statistics, and

$$\psi(\boldsymbol{\theta}) \;\;=\;\; \log \sum_{\boldsymbol{y} \in \mathcal{Y}} \exp\left(\boldsymbol{\theta}^\top s(\boldsymbol{y})\right), \quad \boldsymbol{\theta} \in \Omega_{\boldsymbol{\theta}} = \{\boldsymbol{\theta} \in \mathbb{R}^{d_2} : \psi(\boldsymbol{\theta}) < \infty\} = \mathbb{R}^{d_2}. \tag{7}$$

Britton and O'Neill (2002) and Groendyke et al. (2011) assumed that contacts are independent Bernoulli(μ) ($\mu \in (0,1)$) random variables, which is equivalent to the one-parameter exponential family with natural parameter $\theta = \text{logit}(\mu) \in \mathbb{R}$ and sufficient statistic $s(\boldsymbol{y}) = \sum_{i<j}^{N} y_{i,j}$. Groendyke et al. (2012) extended the exponential-family framework to include predictors of contacts by assuming that contacts are independent Bernoulli($\mu_{i,j}$) ($\mu_{i,j} \in (0,1)$) random variables with $\text{logit}(\mu_{i,j}) = \sum_{k=1}^{d} \theta_k\, s_{i,j,k}(y_{i,j}) \in \mathbb{R}$, that is, the log odds of the probability of a contact between two population members is a weighted sum of functions $s_{i,j,k}(y_{i,j})$ of covariates and $y_{i,j}$, weighted by $\theta_k \in \mathbb{R}$ ($k = 1, \ldots, d$). A specific example is

given by $s_{i,j,1}(y_{i,j}) = y_{i,j}$ and $s_{i,j,2}(y_{i,j}) = c_{i,j}\, y_{i,j}$, where $c_{i,j} \in \{0, 1\}$ is a same-hospital indicator, equal to 1 if population members i and j were in the same hospital during the epidemic and 0 otherwise. The example model is equivalent to a two-parameter exponential family with natural parameters $\theta_1 \in \mathbb{R}$ and $\theta_2 \in \mathbb{R}$ and sufficient statistics $s_1(\boldsymbol{y}) = \sum_{i<j}^{N} s_{i,j,1}(y_{i,j})$ and $s_2(\boldsymbol{y}) = \sum_{i<j}^{N} s_{i,j,2}(y_{i,j})$. If $\theta_2 = 0$, the model of Groendyke et al. (2012) reduces to the model of Britton and O'Neill (2002) and Groendyke et al. (2011).

3 Semiparametric population models

We first describe shortcomings of parametric population models in Section 3.1 and then introduce semiparametric population models to address them in Section 3.2.

3.1 Shortcomings of parametric population models

While the network-based parametric population models reviewed in Section 2.2 are more flexible than classic and lattice-based SIR and SEIR models, these parametric population models have shortcomings. Chief among them is the fact that the induced degree distributions are short-tailed. Here, the degrees of population members are the numbers of contacts of population members.

For example, the model of Britton and O'Neill (2002) and Groendyke et al. (2011) assumes that contacts are independent Bernoulli(μ) random variables. As a consequence, the degrees of population members are Binomial($N-1$, μ) and approximately Poisson($N\,\mu$) distributed provided N is large, μ is small, and $N\,\mu$ tends to a constant, as one would expect in sparse population contact networks where the expected degrees of population members are bounded above by a finite constant and hence μ is a constant multiple of $1/N$. The model of Groendyke et al. (2012) allows degree distributions to be longer-tailed than the model of Britton and O'Neill (2002) and Groendyke et al. (2011)—depending on available covariates—but the induced degree distribution may nonetheless be shorter-tailed than the degree distributions of real-world contact networks. The degree distributions of real-world contact networks are thought to be long-tailed (e.g., Jones and Handcock, 2003a,b, 2004): e.g., in networks of sexual contacts arising in the study of HIV, some population members tend to have many more sexual contacts than most other population members. The population members in the upper tail of the degree distribution represent an important public health concern, because population members with many contacts can infect many others and are hence potential superspreaders.

Last, but not least, it is worth mentioning that scale-free networks with power law degree distributions (Barabàsi and Albert, 1999; Albert and Barabàsi, 2002) are known to induce long-tailed degree distributions. However, aside from the fact that the construction of scale-free networks is incomplete and ambiguous (Bollobás et al., 2001), these one-parameter models are not flexible models of degree distributions, and proper statistical procedures do not lend much support to informal claims that the degree distributions of many real-world contact networks are scale-free: see, e.g., the work of Jones and Handcock (2003a,b, 2004) on the degree distributions of sexual contact networks arising in the study of HIV, and the discussion of Willinger et al. (2009).

Therefore, more flexible population models are needed to accomodate both short- and long-tailed degree distributions.

3.2 Semiparametric population model

We introduce a semiparametric population model to accomodate both short- and long-tailed degree distributions and to detect potential superspreaders. The population model is semiparametric in that the prior of the epidemiological parameter $\boldsymbol{\eta}$ is parametric, while the prior of the network parameter $\boldsymbol{\theta}$ is nonparametric.

Let $s_1(\boldsymbol{y}), \dots, s_N(\boldsymbol{y})$ be the degrees of population members $1, \dots, N$, where the degree of population member i is defined by $s_i(\boldsymbol{y}) = \sum_{j=1:\, j \neq i}^{N} y_{i,j}$ $(i = 1, \dots, N)$. A simple model of the sequence of degrees $s_1(\boldsymbol{y}), \dots, s_N(\boldsymbol{y})$ is given by the exponential family of distributions

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}) \;=\; \exp\left(\sum_{i=1}^{N} \theta_i \, s_i(\boldsymbol{y}) - \psi(\boldsymbol{\theta})\right), \quad \boldsymbol{y} \in \mathcal{Y}, \tag{8}$$

where the degrees $s_1(\boldsymbol{y}), \dots, s_N(\boldsymbol{y})$ are the sufficient statistics and the weights of the degrees $\theta_1, \dots, \theta_N \in \mathbb{R}$ are the natural parameters of the exponential family, and $\psi(\boldsymbol{\theta})$ ensures that $\sum_{\boldsymbol{y} \in \mathcal{Y}} p_{\boldsymbol{\theta}}(\boldsymbol{y}) = 1$. The exponential-family form of (8) can be motivated by its maximum entropy property and other attractive mathematical properties (Brown, 1986). A convenient property is that the resulting likelihood function factorizes as follows:

$$L(\boldsymbol{\theta}; \boldsymbol{y}) \;\propto\; \exp\left(\sum_{i=1}^{N} \theta_i \, s_i(\boldsymbol{y}) - \psi(\boldsymbol{\theta})\right) \;\propto\; \prod_{i<j}^{N} \exp\left(\lambda_{i,j}(\boldsymbol{\theta}) \, y_{i,j} - \psi_{i,j}(\boldsymbol{\theta})\right), \tag{9}$$

where

$$\psi(\boldsymbol{\theta}) \;=\; \sum_{i<j}^{N} \psi_{i,j}(\boldsymbol{\theta}), \tag{10}$$

with $\psi_{i,j}(\boldsymbol{\theta})$ and $\lambda_{i,j}(\boldsymbol{\theta})$ given by

$$\psi_{i,j}(\boldsymbol{\theta}) \;=\; \log\left(1 + \exp\left(\lambda_{i,j}(\boldsymbol{\theta})\right)\right) \tag{11}$$

and

$$\lambda_{i,j}(\boldsymbol{\theta}) \;=\; \theta_i + \theta_j, \tag{12}$$

respectively.

To interpret the natural parameters $\theta_1, \dots, \theta_N$, observe that the model is equivalent to assuming that the contacts $Y_{i,j}$ are independent Bernoulli$(\mu_{i,j})$ ($\mu_{i,j} \in (0, 1)$) random variables with $\text{logit}(\mu_{i,j}) = \theta_i + \theta_j \in \mathbb{R}$. Thus, the log odds of the probability of a contact between population members i and j is additive in the propensities of i and j to be in contact with others. It is worth noting that the resulting model can be viewed as an adaptation of the classic p_1-model (Holland and Leinhardt, 1981) to undirected random graphs and is known as the β-model (Chatterjee et al., 2011; Rinaldo et al., 2013; Chen et al., 2019).

To cluster population members based on degrees and detect potential superspreaders, we assume that the degree parameters $\theta_1, \dots, \theta_N$ are generated by a Dirichlet process prior (e.g., Ferguson, 1973), that is,

$$\begin{aligned} \theta_1 \quad &\sim \; G \\ \theta_i \mid \theta_1, \dots, \theta_{i-1} \quad &\sim \; \frac{1}{\alpha + i - 1} \left(\alpha\, G + \sum_{h=1}^{i-1} \delta_{\theta_h} \right), \; i = 2, 3, \dots, \end{aligned} \tag{13}$$

where $\alpha > 0$ denotes the concentration parameter and G denotes the base distribution of the Dirichlet process prior, and δ_{θ_h} denotes a point mass at θ_h. A convenient choice of the base distribution is $N(\mu,\, \sigma^2)$ ($\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_+$). Draws from a Dirichlet process prior can be generated by generating the first draw from $N(\mu,\, \sigma^2)$, and the i-th draw with a probability proportional to α from $N(\mu,\, \sigma^2)$ and otherwise drawing one of the existing draws $\theta_1, \dots, \theta_{i-1}$ at random. Since degree parameters are resampled, some population members share the same degree parameters, with a probability that depends on the value of α. Thus, the Dirichlet process prior induces a partition of the population into subpopulations, where subpopulations share the same degree parameters.

Short- and long-tailed degree distributions. In addition to detecting potential superspreaders (i.e., subsets of population members with high propensities to form contacts), the model can accommodate short- and long-tailed degree distributions. Short-tailed degree distributions are obtained when, e.g., $\theta_1 = \cdots = \theta_N$. Then contacts are independent Bernoulli$(\mu_{i,j})$ ($\mu_{i,j} \in (0, 1)$) random variables with $\text{logit}(\mu_{i,j}) = 2\, \theta_1 \in \mathbb{R}$ and the degrees of population members are Binomial distributed, which implies that the degree distributions are short-tailed. Long-tailed degree distributions are obtained when most population members have low propensities to form contacts while some population members have high propensities. In fact, depending on the propensities $\theta_1, \dots, \theta_N$ of population members $1, \dots, N$ to form edges, many short- and long-tailed degree distributions can be obtained.

4 Bayesian inference given incomplete data

We discuss Bayesian inference for the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ of the population model. Since complete observations of epidemics and population contact networks are all but impossible, we focus on Bayesian inference given an incomplete observation of an epidemic and a population contact network. We give examples of incomplete data in Section 4.1. Then, starting from first principles, we separate the incomplete-data generating process from the complete-data generating process in Section 4.2, and discuss Bayesian inference given incomplete data in Section 4.3. Last, but not least, we review sampling designs in Section 4.4.

4.1 Incomplete data

In practice, data may be incomplete due to

- data collection constraints: e.g., it may be infeasible or expensive to collect some of the data, such as data on who infected whom;
- design-based incomplete-data generating processes: e.g., a sampling design determines which population members are included in the sample and hence which data are collected;
- out-of-design incomplete-data generating processes: e.g., population members refuse to share data when the data are considered sensitive;

and combinations of them.

4.2 Complete- and incomplete-data generating process

A principled approach to model-based inference separates the complete-data generating process from the incomplete-data generating process:

- The *complete-data generating process* is the process that generates the complete data, that is, the process that generates a realization $(\boldsymbol{x}, \boldsymbol{y})$ of $(\boldsymbol{X}, \boldsymbol{Y})$.
- The *incomplete-data generating process* is the process that determines which subset of the complete data $(\boldsymbol{x}, \boldsymbol{y})$ is observed.

A failure to separate the complete- and incomplete-data generating process can lead to misleading conclusions, as explained by Rubin (1976), Dawid and Dickey (1977), Thompson and Frank (2000), Koskinen et al. (2010), Handcock and Gile (2010, 2017), and Schweinberger et al. (2020).

We adapt here the generic ideas of Rubin (1976) to stochastic models of epidemics. Let $\boldsymbol{A} = \{\boldsymbol{A}_E, \boldsymbol{A}_I, \boldsymbol{A}_R, \boldsymbol{A}_T, \boldsymbol{A}_Y\}$ be indicators of which data are observed, where $\boldsymbol{A}_E = \{A_{E,i}\}_{i=1}^N \in \{0,1\}^N$, $\boldsymbol{A}_I = \{A_{I,i}\}_{i=1}^N \in \{0,1\}^N$, $\boldsymbol{A}_R = \{A_{R,i}\}_{i=1}^N \in \{0,1\}^N$, $\boldsymbol{A}_T = \{A_{T,i,j}\}_{i \neq j}^M \in \{0,1\}^{M^2-M}$, and $\boldsymbol{A}_Y = \{A_{Y,i,j}\}_{i<j}^N \in \{0,1\}^{\binom{N}{2}}$ indicate whether the values of $\{E_i\}_{i=1}^N$, $\{I_i\}_{i=1}^N$, $\{R_i\}_{i=1}^N$, $\{T_{i,j}\}_{i \neq j}^M$, and $\{Y_{i,j}\}_{i<j}^N$ are observed, respectively. The sequence of indicators $\boldsymbol{A}$ is considered to be a random variable, with a distribution parameterized by $\boldsymbol{\pi} \in \Omega_{\boldsymbol{\pi}} \subseteq \mathbb{R}^q$ ($q \geq 1$): e.g., $\boldsymbol{\pi} \in [0,1]^N$ may be a vector of sample inclusion probabilities, where $\pi_i \in [0,1]$ is the probability that population member $i \in \{1, \ldots, N\}$ is sampled and data on the contacts of population member i are collected. The parameter $\boldsymbol{\pi}$ may be known or unknown. We focus henceforth on the more challenging case where $\boldsymbol{\pi}$ is unknown. The observed and unobserved subset of the complete data $(\boldsymbol{x}, \boldsymbol{y})$ are denoted by $\boldsymbol{x}_{\text{obs}}$ and $\boldsymbol{x}_{\text{mis}}$ and $\boldsymbol{y}_{\text{obs}}$ and $\boldsymbol{y}_{\text{mis}}$, respectively, where $\boldsymbol{x} = (\boldsymbol{x}_{\text{obs}}, \boldsymbol{x}_{\text{mis}})$ and $\boldsymbol{y} = (\boldsymbol{y}_{\text{obs}}, \boldsymbol{y}_{\text{mis}})$.

In Bayesian fashion, we build a joint probability model for all knowns and unknowns and condition on all knowns. Let

$$p(\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}) \;\; = \;\; p(\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta})\, p(\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}) \tag{14}$$

be the joint probability density of $\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}$, where

$$p(\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}) \;\; = \;\; p(\boldsymbol{\pi} \mid \boldsymbol{\eta}, \boldsymbol{\theta})\, p(\boldsymbol{\eta} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta}) \tag{15}$$

is the prior probability density of $\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}$ and

$$p(\boldsymbol{a}, \boldsymbol{x}, \boldsymbol{y} \mid \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}) \;\; = \;\; p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\pi})\, p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta})\, p(\boldsymbol{y} \mid \boldsymbol{\theta}) \tag{16}$$

is the conditional probability density of $\boldsymbol{a}$, $\boldsymbol{x}$, $\boldsymbol{y}$ given $\boldsymbol{\pi}$, $\boldsymbol{\eta}$, $\boldsymbol{\theta}$; note that $p(\boldsymbol{y} \mid \boldsymbol{\theta}) \equiv p_{\boldsymbol{\theta}}(\boldsymbol{y})$. It is worth noting that all of these probability densities are with respect to a suitable dominating measure, but we do not wish to delve into measure-theoretic details, which would distract from the main ideas.

Since interest centers on the population model, it is natural to ask: Under which conditions is the incomplete-data generating process ignorable for the purpose of Bayesian inference for the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ of the population model?

Definition: likelihood-ignorable incomplete-data generating process. *Assume that the parameters $\boldsymbol{\pi}$, $\boldsymbol{\eta}$, $\boldsymbol{\theta}$ are variation-independent in the sense that the parameter space* *of* $(\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta})$ *is given by a product space of the form* $\Omega_{\boldsymbol{\pi}} \times \Omega_{\boldsymbol{\eta}} \times \Omega_{\boldsymbol{\theta}}$ *and that the parameters of the population model* $\boldsymbol{\eta}$ *and* $\boldsymbol{\theta}$ *and the parameter of the incomplete-data generating process* $\boldsymbol{\pi}$ *are independent under the prior,*

$$p(\boldsymbol{\pi} \mid \boldsymbol{\eta}, \boldsymbol{\theta}) \;=\; p(\boldsymbol{\pi}) \quad \textit{for all} \quad (\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta}) \in \Omega_{\boldsymbol{\pi}} \times \Omega_{\boldsymbol{\eta}} \times \Omega_{\boldsymbol{\theta}}. \tag{17}$$

If the probability of observing data does not depend on the values of the unobserved data,

$$p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\pi}) \;=\; p(\boldsymbol{a} \mid \boldsymbol{x}_{obs}, \boldsymbol{y}_{obs}, \boldsymbol{\pi}) \quad \textit{for all} \quad \boldsymbol{a},\ \boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{\pi} \in \Omega_{\boldsymbol{\pi}}, \tag{18}$$

then the incomplete-data generating process is called likelihood-ignorable and otherwise non-ignorable.

We give examples of likelihood-ignorable and non-ignorable incomplete-data processes and then show that likelihood-ignorable incomplete-data generating processes can be ignored for the purpose of Bayesian inference for the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ of the population model.

Example: likelihood-ignorable. *All infected population members visit hospitals, which record data on contacts, transmissions, exposure, infectious, and removal times of infected population members. To reduce the posterior uncertainty about the population contact network and its generating mechanism, investigators generate a probability sample of non-infected population members and record the contacts of sampled population members.*

We discuss sampling designs in Section 4.4. Some of them—e.g., link-tracing—exploit observed contacts of population members to include additional population members in the sample, which implies that the sample inclusion probabilities depend on observed contacts:

$$p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\pi}) \;=\; p(\boldsymbol{a} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{\pi}) \quad \text{for all} \quad \boldsymbol{a},\ \boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{\pi} \in \Omega_{\boldsymbol{\pi}}. \tag{19}$$

Example: non-likelihood-ignorable. *Suppose that there exists a constant* $\delta > 0$ *such that infected population members* i *with mild symptoms and fast recovery* ($R_i - I_i \leq \delta$) *do not visit hospitals, whereas infected population members* i *with severe symptoms and slow recovery* ($R_i - I_i > \delta$) *do visit hospitals. Hospitals collect data on infected population members who visit them, but no data are collected on other population members.*

Since the incomplete-data generating process excludes all infected population members with mild symptoms and fast recovery, it cannot be ignored. Statistical analyses ignoring it may give rise to misleading conclusions about the rate of infection β and other parameters of the population model.

4.3 Bayesian inference given incomplete data

We are interested in conditions under which the incomplete-data generating process can be ignored for the purpose of Bayesian inference for the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ of the population model.

The following result shows that, if the incomplete-data generating process is likelihood-ignorable and the prior is proper, then the incomplete-data generating process can be ignored for the purpose of Bayesian inference for the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ of the population model. The result adapts a generic result of Rubin (1976) to stochastic models of epidemics.

Proposition 1. *If the incomplete-data generating process is likelihood-ignorable and the prior* $p(\boldsymbol{\eta}, \boldsymbol{\theta}) = p(\boldsymbol{\eta} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})$ *is proper, then the parameter* $\boldsymbol{\pi}$ *of the incomplete-data generating process and the parameters* $\boldsymbol{\eta}$ *and* $\boldsymbol{\theta}$ *of the population model are independent under the posterior,*

$$p(\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{obs}, \boldsymbol{y}_{obs}) \;\; \propto \;\; p(\boldsymbol{\pi} \mid \boldsymbol{a}, \boldsymbol{x}_{obs}, \boldsymbol{y}_{obs})\, p(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{obs}, \boldsymbol{y}_{obs}), \tag{20}$$

and Bayesian inference for the parameters $\boldsymbol{\eta}$ *and* $\boldsymbol{\theta}$ *of the population model can ignore the incomplete-data generating process and can be based on the marginal posterior*

$$p(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{obs}, \boldsymbol{y}_{obs}) = \frac{\displaystyle\sum_{\boldsymbol{y}_{mis}} \int p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta})\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\eta} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\; d\boldsymbol{x}_{mis}}{\displaystyle\sum_{\boldsymbol{y}_{mis}} \int \int \int p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta})\, p(\boldsymbol{y} \mid \boldsymbol{\theta})\, p(\boldsymbol{\eta} \mid \boldsymbol{\theta})\, p(\boldsymbol{\theta})\; d\boldsymbol{x}_{mis}\, d\boldsymbol{\eta}\, d\boldsymbol{\theta}}.$$

The case of Britton and O'Neill (2002) and Groendyke et al. (2011, 2012), who considered Bayesian inference from observed infectious and removal times, is a special case of the incomplete-data framework considered here, with $\boldsymbol{x}_{\text{obs}} = \{\boldsymbol{I}, \boldsymbol{R}\}$ and $\boldsymbol{y}_{\text{obs}} = \{\}$. In general, when $\boldsymbol{x}$ or $\boldsymbol{y}$ or both are partially observed, Bayesian inference can be based on the marginal posterior of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ given the observed data as long as the incomplete-data generating process is likelihood-ignorable. Bayesian Markov chain Monte Carlo methods for sampling from the marginal posterior of $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ given the observed data are described in the supplement.

4.4 Sampling designs

To reduce the posterior uncertainty about the network parameter $\boldsymbol{\theta}$, it is advisable to sample from population contact networks. The reason is that epidemiological data may not contain much information about $\boldsymbol{\theta}$—and more so when the number of infected population members M is small relative to the total number of population members N and the transmissions are unobserved.

We describe two likelihood-ignorable sampling designs for generating samples of contacts and epidemiological data, ego-centric sampling and link-tracing. Some background on ego-centric sampling and link-tracing of contacts (but not epidemiological data) can be found in Thompson and Frank (2000), Handcock and Gile (2010), and Krivitsky and Morris (2017). Other network sampling designs are reviewed in Schweinberger et al. (2020). We adapt these ideas to sampling contacts along with epidemiogical data.

An ego-centric sample of contacts and epidemiological data can be generated as follows:

(a) Generate a probability sample of population members.

(b) For each sampled population member, record data on contacts and, should the population member be infected, data on transmissions, exposure, infectious, and removal times.

A probability sample of population members can be generated by any sampling design for sampling from finite populations (e.g., Thompson, 2012).

An interesting extension of ego-centric sampling is link-tracing. Link-tracing exploits the observed contacts of sampled population members to include additional population members into the sample. A k-wave link-tracing sample of contacts and epidemiological data can be generated as follows:

(1) Wave $l = 0$: Generate an ego-centric sample.

(2) Wave $l = 1, \dots, k$:

   (a) Add the population members who are linked to the population members of wave $l - 1$ to the sample.

   (b) For each added population member, record data.

Link-tracing implies that the sample inclusion probabilities depends on the observed contacts:

$$p(\boldsymbol{a} \mid \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{\pi}) \;=\; p(\boldsymbol{a} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{\pi}) \quad \text{for all} \quad \boldsymbol{a},\ \boldsymbol{x},\ \boldsymbol{y},\ \boldsymbol{\pi} \in \Omega_{\boldsymbol{\pi}}. \tag{21}$$

Ego-centric sampling can be considered to be a special case of k-wave link-tracing with $k = 0$. By construction, the probability of observing data does not depend on unobserved data, implying that both sampling designs are likelihood-ignorable.

5 Simulations

We explore the frequentist properties of Bayesian estimators and the reduction in statistical error due to sampling contacts by using simulations. We consider a population of size 187 and partition the population into three subpopulations $1, 2, 3$ by assigning population members to subpopulations $1, 2, 3$ with probabilities .4, .3, .3, respectively. We generate a population contact network according to model (8) with parameters $\theta_i = \gamma_{C_i} \in \mathbb{R}$, where $C_i \in \{1, 2, 3\}$ is an indicator of the subpopulation to which population member i belongs. Conditional on the population contact network, an epidemic is generated by the network-based SEIR model described in Section 2.1, assuming that $I_i - E_i$ and $R_i - I_i$ are independent Gamma$(\eta_{E,1}, \eta_{E,2})$ and Gamma$(\eta_{I,1}, \eta_{I,2})$ random variables, respectively. We assume that exposure, infectious, and removal times are observed, whereas transmissions and contacts are unobserved. The memberships of population members to subpopulations are likewise unobserved.

5.1 Parameter recovery

We generated 1,000 population contact networks and epidemics as discussed above. The data-generating values of parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, β, γ_1, γ_2, γ_3 are shown in Table 1. For each generated data set, we used a truncated Dirichlet process prior (Ishwaran and James, 2001) with $K = 3$ and $K = 5$ subpopulations. Truncated Dirichlet process priors are approximate Dirichlet process priors based on truncating Dirichlet process priors, which has computational advantages: see Ishwaran and James (2001) and the supplement.

Table 1 sheds light on the frequentist coverage properties of 95%-posterior credibility intervals. The simulation results indicate that the frequentist coverage properties of posterior credibility intervals are excellent in the case of the epidemiological parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, β, but less so in the case of the network parameters γ_1, γ_2, γ_3. Additional figures in the supplement show that the posterior medians of the network parameters are biased. These results underscore the challenge of estimating network parameters without observing contacts. We demonstrate in Section 5.2 that the statistical error can be reduced by sampling contacts.

5.2 Sampling contacts

To demonstrate the reduction in statistical error due to sampling contacts, we generate a population consisting of a low-degree subpopulation of size 127 with degree parameter $\gamma_1 = -3.5$, a moderate-degree subpopulation of size 50 with degree parameter $\gamma_2 = -1.5$, and a high-degree subpopulation of size 10 with degree parameter $\gamma_3 = 0.5$.

1,000 egocentric samples of size n = 25, 50, 75, 100, 125, 150, 187 are generated from the population of size $N = 187$. By construction of the model, estimators of the epidemiological parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$ are not expected to be sensitive to n—which determines how much information is available about the network parameter—and the mean squared error

(MSE) of the posterior median and mean of $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$ is indeed not sensitive to n (not shown). The MSE of posterior median and mean of the rate of infection β and the degree parameters $\gamma_1, \gamma_2, \gamma_3$ are shown in Figure 2. The figure demonstrates that samples of contacts reduce the MSE of posterior median and mean of β, $\gamma_1, \gamma_2, \gamma_3$: The MSE decreases as the sample size n increases. These observations suggest that in practice samples of contacts—or functions of contacts, such as degrees—should be collected.

6 Partially observed MERS epidemic in South Korea

We showcase the framework introduced in Sections 3 and 4 by applying it the partially observed MERS epidemic in South Korea in 2015 (Ki, 2015). The MERS epidemic was driven by the coronavirus MERS, which is related to the coronaviruses SARS and COVID-19. We retrieved the data from the website `http://english.mw.go.kr` of the South Korean Ministry of Health and Welfare on September 22, 2015; note that the website has been removed since, but the data can be obtained from the authors. The first MERS case was confirmed on May 20, 2015 and the last confirmed infection occurred on July 4, 2015. By September 22, 2015, 186 cases had been confirmed in 44 hospitals, of which 144 have recovered and 35 have died, all of which are considered to be removed from the population. 7 cases had not removed by September 22, 2015 despite the fact that the last confirmed infection occurred on July 4, 2015. We assume that the removal times of these 7 cases are unobserved and that the outbreak ceased by September 22, 2015, because the data base has not been updated by the South Korean Ministry of Health and Welfare since July 2015.

The data consist of infectious times and removal times, with 7 missing removal times. In addition, there are assessments by doctors on who infected whom. The exposure times and contacts are unobserved and inferred along with the 7 missing removal times.

Since there is no evidence to suggest that the incomplete-data generating process is non-ignorable, we estimate the population model under the assumption that the incomplete-data generating process is ignorable. We assume that $I_i - E_i$ and $R_i - I_i$ are independent Gamma$(\eta_{E,1}, \eta_{E,2})$ and Gamma$(\eta_{I,1}, \eta_{I,2})$ random variables, respectively. The priors of the epidemiological parameters η are based on the epidemiological characteristics of MERS described in Ki (2015) and are given by $\eta_{E,1} \sim \text{Uniform}(4, 8)$, $\eta_{E,2} \sim \text{Uniform}(.75, 3)$, $\eta_{I,1} \sim \text{Uniform}(1.5, 8)$, $\eta_{I,2} \sim \text{Uniform}(2.5, 7.5)$, and $\beta \sim \text{Uniform}(.1, 8)$. The assessments of doctors on who infected whom are used as a prior, that is, for each infected population member i, the population member j who infected population member i according to the doctors is assigned a prior probability of $100\, c > 0$ of being the infector of i and all other possible infectors are assigned a prior probability of $c > 0$ (as described in Groendyke et al., 2011). In addition, we assume $\alpha = 5$, $\mu = 0$, and $\sigma^2 = 10$.

We use the model introduced in Section 3 with a Dirichlet process prior truncated at $K = 2$ subpopulations as described in the supplement, because we expect that there are two subpopulations, one corresponding to potential superspreaders and one corresponding to all other population members. We compare the model with $K = 2$ subpopulations to the model with $K = 1$ subpopulation, which is equivalent to the classic Erdős and Rényi (1959) model used by Britton and O'Neill (2002) and Groendyke et al. (2011). We attempted to incorporate covariates as predictors of contacts (e.g., same-hospital indicators), but the results based on models with and without covariates turned out to be indistinguishable, therefore we focus on models without covariates. We sampled from the posterior by using the Bayesian Markov chain Monte Carlo methods described in the supplement. Markov chain Monte Carlo convergence diagnostics are provided in the supplement.

Figure 3 shows the posterior predictions of the number of population members in the infectious state plotted against time under the model with $K = 1$ and $K = 2$. The figure suggests that the model with $K = 2$ outperforms the model with $K = 1$ in terms of predictive power. In fact, the root mean squared deviation of the number of individuals in the infectious state (summed over the 61 days of the outbreak) is 674.4 and 486.76 under the models with $K = 1$ and $K = 2$, respectively.

Figure 4 shows the posterior probabilities of population members belonging to subpopulations under the model with $K = 2$ subpopulations. The figure shows that 3 population members belong with high posterior probability to the red-colored subpopulation, whereas all other population members belong with high posterior probability to the black-colored subpopulation. In fact, these 3 population members are suspected to have infected many other population members according to the assessments of doctors, as can be seen in Figure 4. The marginal posterior densities of γ_1 (black subpopulation) and γ_2 (red subpopulation) are shown in Figure 5 and suggest that the members of the red-colored subpopulation have a higher propensity to be connected than the members of the black-colored subpopulation. Taken together, these observations suggest that there was a small subpopulation consisting of 3 superspreaders who may have had a great impact on the outcome of the outbreak.

7 Discussion

We have introduced a semiparametric population model that can accomodate both short- and long-tailed degree distributions and can detect potential superspreaders, in addition to dealing with a wide range of missing and sample data.

That said, there are many open questions, some of which are related to the lack of data and others are related to computational and statistical challenges arising from the lack of data and the complexity of the models. We review a selection of open questions below.

7.1 Population

In the application to the partially observed MERS epidemic in South Korea, we applied the proposed framework to the 186 infected population members. In so doing, we assumed that the population of interest consists of those 186 infected population members. There is no denying that such an assumption is unappealing. In fact, the assumption was motivated by convenience rather than substantive considerations, including

- the challenge of determining the population of interest: e.g., is the population of interest the population of South Korea, the population of East Asia, or the population of the whole world?
- the lack of data on the population of interest, foremost the lack of data on the population contact network;
- associated computational and statistical challenges.

We discuss the data issue in Section 7.2, followed by computational challenges in Section 7.3, and conclude with some remarks on non-ignorable incomplete-data generating processes and models capturing additional features of population contact networks.

7.2 Data

As pointed out before, complete observations of population contact networks and epidemics are all but impossible. As a consequence, public health officials and researchers face a recurring question in the event of epidemics (whether outbreaks of Ebola viruses or coronaviruses such as SARS, MERS, and COVID-19): Which data to collect? We believe that, to learn the structure of population contact networks and their impact on the spread of an infectious diseases, investigators should attempt to collect contact and epidemiological data on all infected population members and collect samples of contacts from non-infected population members by likelihood-ignorable sampling designs. We discuss some of the challenges arising in practice along with possible solutions.

First, while it is challenging to collect data on transmissions, it is advisable to collect data that help reduce the posterior uncertainty about epidemiological parameters. One possible source of data are viral genetic sequence data, among other possible data sources. We refer interested readers to Bouckaert et al. (2019) for a recent review of possible data sources.

Second, it is prudent to sample from population contact networks to reduce the posterior uncertainty about network parameters. The two likelihood-ignorable sampling designs discussed above can be used to do so, though both require a sampling frame. If no sampling frame is available, an alternative would be respondent-driven sampling (e.g., Gile and Handcock, 2010; Gile, 2011), which is a form of link-tracing without a sampling frame. Location data collected by mobile phones and other electronic sources would be alternative sources of data, but raise data privacy issues (Fienberg and Slavković, 2010). In the past decade, substantial progress has been made on data privacy in the statistical literature: see, e.g., the work of Karwa and Slavković (2016) on data privacy in scenarios where network data are generated by the β-model (albeit without Dirichlet process priors and without epidemics). Studying data privacy for epidemics would be an important direction of future research.

7.3 Computational challenges

The lack of data has computational implications. Likelihood-based algorithms have to integrate over the unobserved data, so computational costs tend to increase with the amount of missing data. How to develop scalable statistical algorithms, with statistical guarantees, is an open question. One idea would be to develop a two-step estimation algorithm, first estimating the network parameters and then estimating the epidemiological parameters, leveraging computational advances in the statistical analysis of network data (e.g., Raftery et al., 2012; Hunter et al., 2012; Salter-Townshend and Murphy, 2013; Babkin et al., 2020) and epidemiological data (Bouckaert et al., 2019). A two-step estimation algorithm may require data on contacts, however, because without data on contacts the posterior correlations of the network parameters and the rate of infection can be high (Groendyke et al., 2011), in which case two-step estimation algorithms may not work well. A related problem is how to update posteriors efficiently as more data on infections and contacts come in.

7.4 Non-ignorable incomplete-data generating processes

We have considered here ignorable incomplete-data generating processes. If the incomplete-data generating process is non-ignorable, then either the incomplete-data generating process must be modeled or it must be demonstrated that Bayesian inference for the parameters of the population model is insensitive to the incomplete-data generating process. Both approaches require insight into the incomplete-data generating process and additional model assumptions, some of which may be untestable.

7.5 Population models capturing additional network features

We have focused here on the degrees of population members as network features, but there are many other important network features: e.g., if population contact networks exhibit closure (e.g., transitive closure, Wasserman and Faust, 1994), then infectious diseases may spread rapidly within subpopulations but may spread only slowly through the whole population, which has potential policy implications. There are two broad approaches to capturing closure in population contact networks: latent space models (Hoff et al., 2002; Handcock et al., 2007; Smith et al., 2019) and related latent variable models (e.g., Salter-Townshend et al., 2012; Rastelli et al., 2016; Fosdick and Hoff, 2015; Hoff, 2020); and exponential-family models of random graphs (Frank and Strauss, 1986; Snijders et al., 2006; Schweinberger and Stewart, 2020). Both classes of models can accomodate the degree terms we have used (see, e.g., Krivitsky et al., 2009) along with additional terms that reward closure in population contact networks. However, both of them come at computational and statistical costs. In fact, unless data on the population contact network are collected, closure in the population contact network may not be detectable at all, as pointed out by Welch (2011). This, too, underscores the importance of collecting data on contacts.

7.6 Time-evolving population contact networks

We have assumed that the population contact network is time-invariant, motivated by the lack of data on the population contact network and the desire to keep the model as simple and parsimonious as possible. In practice, the population contact network may evolve over time, because population members may add or delete contacts and because authorities may enforce social distancing measures. As a consequence, it would be natural to allow the population contact network to change over time. Extensions to time-evolving population contact networks could be based on temporal stochastic block and latent space models (e.g., Fu et al., 2009; Sewell and Chen, 2015, 2016; Sewell et al., 2016); temporal exponential-family random graph models (Robins and Pattison, 2001; Hanneke et al., 2010; Ouzienko et al., 2011; Krivitsky and Handcock, 2014); continuous-time Markov processes (Snijders et al., 2010); relational event models (Butts, 2008); and other models (e.g., Katz and Proctor, 1959; Durante and Dunson, 2014; Sewell, 2017). That said, these models would likewise come at computational and statistical costs, and would require samples of contacts over time.

Supplementary materials

The supplement describes Bayesian Markov chain Monte Carlo methods and convergence diagnostics.

Acknowledgements

We acknowledge support from the National Institutes of Health (NIH award 1R01HD052887-01A2) (MS, RB), the Office of Naval Research (ONR award N00014-08-1-1015) (MS), and the National Science Foundation (NSF awards DMS-1513644 and DMS-1812119) (MS, SB).

Disclosure statement

We do not have financial interests or other interests related to the results presented here.

Data availability

We retrieved the data from the website `http://english.mw.go.kr` of the South Korean Ministry of Health and Welfare on September 22, 2015; note that the website has been removed since, but the data can be obtained from the authors.

References

Albert, R., and Barabàsi, A. L. (2002), “Statistical mechanics of complex networks,” *Review of Modern Physics*, 74, 47–97.

Andersson, H., and Britton, T. (2000), *Stochastic Epidemic Models and Their Statistical Analysis*, New York: Springer-Verlag.

Babkin, S., Stewart, J., Long, X., and Schweinberger, M. (2020), “Large-scale estimation of random graph models with local dependence,” *Computational Statistics & Data Analysis*, to appear.

Barabàsi, A. L., and Albert, R. (1999), “Emergence of scaling in random networks,” *Science*, 286, 509–512.

Bollobás, B., Riordan, O., Spencer, J., and Tusnády, G. (2001), “The degree sequence of a scale-free random graph process,” *Random Structures & Algorithms*, 18, 279–290.

Bomiriya, R. P. (2014), "Topics in Exponential Random Graph Modeling," Ph.D. thesis, Department of Statistics, The Pennsylvania State University, https://etda.libraries.psu.edu/paper/22448.

Bouckaert, R., Vaughan, T. G., Barido-Sottani, J., Duchne, S., Fourment, M., Gavryushkina, A., Heled, J., Jones, G., Kuhnert, D., De Maio, N., Matschiner, M., Mendes, F. K., Moller, N. F., Ogilvie, H. A., du Plessis, L., Popinga, A., Rambaut, A., Rasmussen, D., Siveroni, I., Suchard, M. A., Wu, C.-H., Xie, D., Zhang, C., Stadler, T., and Drummond, A. J. (2019), "BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis," *PLOS Computational Biology*, 15, 1–28.

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

Britton, T., and Pardoux, E. (eds.) (2019), *Stochastic Epidemic Models with Inference*, Springer.

Brown, L. (1986), *Fundamentals of Statistical Exponential Families: With Applications in Statistical Decision Theory*, Hayworth, CA, USA: Institute of Mathematical Statistics.

Butts, C. T. (2008), "A relational event framework for social action," *Sociological Methodology*, 38, 155–200.

Chatterjee, S., Diaconis, P., and Sly, A. (2011), "Random graphs with a given degree sequence," *The Annals of Applied Probability*, 21, 1400–1435.

Chen, M., Kato, K., and Leng, C. (2019), "Analysis of networks via the sparse β-model," ArXiv:1908.03152.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V., and Vernon, M. C. (2011), "Networks and the Epidemiology of Infectious Disease," *Interdisciplinary Perspectives on Infectious Diseases*, 2011, 1–28.

Dawid, A. P., and Dickey, J. M. (1977), "Likelihood and Bayesian inference from selectively reported data," *Journal of the American Statistical Association*, 72, 845–850.

Durante, D., and Dunson, D. B. (2014), "Nonparametric Bayes dynamic modelling of relational data," *Biometrika*, 101, 125–138.

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

Ferguson, T. (1973), "A Bayesian analysis of some nonparametric problems," *The Annals of Statistics*, 1, 209–230.

Fienberg, S. E., and Slavković, A. (2010), *Data Privacy and Confidentiality*, Springer-Verlag, pp. 342–345.

Fosdick, B. K., and Hoff, P. D. (2015), "Testing and Modeling Dependencies Between a Network and Nodal Attributes," *Journal of the American Statistical Association*, 110, 1047–1056.

Frank, O., and Strauss, D. (1986), "Markov graphs," *Journal of the American Statistical Association*, 81, 832–842.

Fu, W., Song, L., and E., X. (2009), "Dynamic mixed membership blockmodel for evolving networks," in *Proceedings of the 26th Annual International Conference on Machine Learning*.

Gile, K. (2011), "Improved inference for respondent-driven sampling data with application to HIV prevalence estimation," *Journal of the American Statistical Association*, 106, 135–146.

Gile, K., and Handcock, M. H. (2010), "Respondent-driven sampling: An assessment of current methodology," *Sociological Methodology*, 40, 285–327.

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

— (2012), "A network-based analysis of the 1861 Hagelloch measles data," *Biometrics*, 68, 755–765.

Handcock, M. S., and Gile, K. (2010), "Modeling social networks from sampled data," *The Annals of Applied Statistics*, 4, 5–25.

— (2017), "Analysis of networks with missing data with application to the National Longitudinal Study of Adolescent Health," *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 66, 501–519.

Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2007), "Model-based clustering for social networks," *Journal of the Royal Statistical Society, Series A (with discussion)*, 170, 301–354.

Hanneke, S., Fu, W., and Xing, E. P. (2010), "Discrete temporal models of social networks," *Electronic Journal of Statistics*, 4, 585–605.

Hoff, P. D. (2020), "Additive and multiplicative effects network models," *Statistical Science*, to appear.

Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002), "Latent space approaches to social network analysis," *Journal of the American Statistical Association*, 97, 1090–1098.

Holland, P. W., and Leinhardt, S. (1981), "An exponential family of probability distributions for directed graphs," *Journal of the American Statistical Association*, 76, 33–65.

Hunter, D. R., Krivitsky, P. N., and Schweinberger, M. (2012), "Computational statistical methods for social network models," *Journal of Computational and Graphical Statistics*, 21, 856–882.

Ishwaran, H., and James, L. F. (2001), "Gibbs Sampling Methods for Stick-breaking Priors," *Journal of the American Statistical Association*, 96, 161–173.

Jones, J. H., and Handcock, M. S. (2003a), "An assessment of preferential attachment as a mechanism for human sexual network formation," *Proceedings of the Royal Society Series B*, 270, 1123–1128.

— (2003b), "Social Networks: Sexual Contacts and Epidemic Thresholds," *Nature*, 423, 605–606.

Jones, J. H. H., and Handcock, M. S. (2004), "Likelihood-Based Inference for Stochastic Models of Sexual Network Formation," *Population Biology*, 65, 413–422.

Karwa, V., and Slavković, A. B. (2016), "Inference using noisy degrees: Differentially private β-model and synthetic graphs," *The Annals of Statistics*, 44, 87–112.

Katz, L., and Proctor, C. H. (1959), "The configuration of interpersonal relations in a group as a time-dependent stochastic process," *Psychometrika*, 24, 317–327.

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

Ki, M. (2015), "2015 MERS outbreak in Korea: hospital-to-hospital transmission," *Epidemiology and Health*, 37, 1–4.

Koskinen, J. H., Robins, G. L., and Pattison, P. E. (2010), "Analysing exponential random graph (p-star) models with missing data using Bayesian data augmentation," *Statistical Methodology*, 7, 366–384.

Krivitsky, P. N., and Handcock, M. S. (2014), "A separable model for dynamic networks," *Journal of the Royal Statistical Society B*, 76, 29–46.

Krivitsky, P. N., Handcock, M. S., Raftery, A. E., and Hoff, P. D. (2009), "Representing Degree Distributions, Clustering, and Homophily in Social Networks With Latent Cluster Random Effects Models," *Social Networks*, 31, 204–213.

Krivitsky, P. N., and Morris, M. (2017), "Inference for social network models from egocentrically-sampled data, with application to understanding persistent racial disparities in HIV prevalence in the US," *Annals of Applied Statistics*, 11, 427–455.

Liu, J. S. (2008), *Monte Carlo Strategies in Scientific Computing*, New York: Springer-Verlag.

Ouzienko, V., Guo, Y., and Obradovic, Z. (2011), "A decoupled exponential random graph model for prediction of structure and attributes in temporal social networks," *Statistical Analysis and Data Mining*, 4, 470–486.

Raftery, A. E., and Lewis, S. M. (1996), "Implementing MCMC," in *Markov chain Monte Carlo in Practice*, eds. Gilks, W. R., Richardson, S., and Spiegelhalter, D. J., London: Chapman & Hall, Chap. 7, pp. 115–130.

Raftery, A. E., Niu, X., Hoff, P. D., and Yeung, K. Y. (2012), "Fast inference for the latent space network model using a case-control approximate likelihood," *Journal of Computational and Graphical Statistics*, 21, 901–919.

Rastelli, R., Friel, N., and Raftery, A. E. (2016), "Properties of latent variable network models," *Network Science*, 4, 407–432.

Rinaldo, A., Petrović, S., and Fienberg, S. E. (2013), "Maximum likelihood estimation in the β-model," *The Annals of Statistics*, 41, 1085–1110.

Robins, G., and Pattison, P. (2001), "Random graph models for temporal processes in social networks," *Journal of Mathematical Sociology*, 25, 5–41.

Rubin, D. B. (1976), "Inference and missing data," *Biometrika*, 63, 581–592.

Salter-Townshend, M., and Murphy, T. B. (2013), "Variational Bayesian inference for the latent position cluster model for network data," *Computational Statistics and Data Analysis*, 57, 661–671.

Salter-Townshend, M., White, A., Gollini, I., and Murphy, T. B. (2012), "Review of statistical network analysis: models, algorithms, and software," *Statistical Analysis and Data Mining*, 5, 243–264.

Schweinberger, M., and Handcock, M. S. (2015), "Local dependence in random graph models: characterization, properties and statistical inference," *Journal of the Royal Statistical Society, Series B*, 77, 647–676.

Schweinberger, M., Krivitsky, P. N., Butts, C. T., and Stewart, J. (2020), "Exponential-family models of random graphs: Inference in finite, super, and infinite population scenarios," *Statistical Science*, to appear.

Schweinberger, M., and Luna, P. (2018), "HERGM: Hierarchical exponential-family random graph models," *Journal of Statistical Software*, 85, 1–39.

Schweinberger, M., and Stewart, J. (2020), "Concentration and consistency results for canonical and curved exponential-family models of random graphs," *The Annals of Statistics*, 48, 374–396.

Sewell, D. K. (2017), "Network autocorrelation models with egocentric data," *Social Networks*, 49, 113–123.

Sewell, D. K., and Chen, Y. (2015), "Latent space models for dynamic networks," *Journal of the American Statistical Association*, 110, 1646–1657.

— (2016), "Latent Space Approaches to Community Detection in Dynamic Networks," *Bayesian Analysis*.

Sewell, D. K., Chen, Y., Bernhard, W., and Sulkin, T. (2016), "Model-based longitudinal clustering with varying cluster assignments," *Statistica Sinica*, 26, 205–233.

Smith, A. L., Asta, D. M., and Calder, C. A. (2019), "The geometry of continuous latent space models for network data," *Statistical Science*, 34, 428–453.

Snijders, T. A. B., Koskinen, J., and Schweinberger, M. (2010), "Maximum likelihood estimation for social network dynamics," *The Annals of Applied Statistics*, 4, 567–588.

Snijders, T. A. B., Pattison, P. E., Robins, G. L., and Handcock, M. S. (2006), "New specifications for exponential random graph models," *Sociological Methodology*, 36, 99–153.

Stephens, M. (2000), "Dealing with label-switching in mixture models," *Journal of the Royal Statistical Society, Series B*, 62, 795–809.

Thompson, S. (2012), *Sampling*, John Wiley & Sons, 3rd ed.

Thompson, S., and Frank, O. (2000), "Model-based estimation with link-tracing sampling designs," *Survey Methodology*, 26, 87–98.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions," *The Annals of Statistics*, 22, 1701–1728.

Warnes, G. R., and Burrows, R. (2010), *R package mcgibbsit: Warnes and Raftery's MCGibbsit MCMC diagnostic.*

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

Welch, D. (2011), "Is network clustering detectable in transmission trees?" *Viruses*, 3, 659–676.

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

Willinger, W., Alderson, D., and Doyle, J. C. (2009), "Mathematics and the internet: A source of enormous confusion and great potential," *Notices of the American Mathematical Society*, 56, 586–599.

A Proofs

We prove Proposition 1.

Proof of Proposition 1. Since the incomplete-data generating process is likelihood-ignorable,

$$\begin{aligned} p(\boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) \; &= \sum_{\boldsymbol{y}_{\text{mis}}} \int p(\boldsymbol{x}_{\text{mis}}, \boldsymbol{y}_{\text{mis}}, \boldsymbol{\pi}, \boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) \, \mathrm{d}\, \boldsymbol{x}_{\text{mis}} \\ &\propto \; p(\boldsymbol{a} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{\pi}) \, p(\boldsymbol{\pi}) \sum_{\boldsymbol{y}_{\text{mis}}} \int p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta}) \, p(\boldsymbol{y} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\eta} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta}) \, \mathrm{d}\, \boldsymbol{x}_{\text{mis}} \qquad (22) \\ &\propto \; p(\boldsymbol{\pi} \mid \boldsymbol{a}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) \, p(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}), \end{aligned}$$

where

$$p(\boldsymbol{\pi} \mid \boldsymbol{a}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) \; = \; \frac{p(\boldsymbol{a} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{\pi}) \, p(\boldsymbol{\pi})}{\displaystyle\int p(\boldsymbol{a} \mid \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}, \boldsymbol{\pi}) \, p(\boldsymbol{\pi}) \, \mathrm{d}\, \boldsymbol{\pi}}$$

$$p(\boldsymbol{\eta}, \boldsymbol{\theta} \mid \boldsymbol{a}, \boldsymbol{x}_{\text{obs}}, \boldsymbol{y}_{\text{obs}}) = \frac{\displaystyle\sum_{\boldsymbol{y}_{\text{mis}}} \int p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta}) \, p(\boldsymbol{y} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\eta} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta}) \, \mathrm{d}\, \boldsymbol{x}_{\text{mis}}}{\displaystyle\sum_{\boldsymbol{y}_{\text{mis}}} \int \int \int p(\boldsymbol{x} \mid \boldsymbol{y}, \boldsymbol{\eta}) \, p(\boldsymbol{y} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\eta} \mid \boldsymbol{\theta}) \, p(\boldsymbol{\theta}) \, \mathrm{d}\, \boldsymbol{x}_{\text{mis}} \, \mathrm{d}\, \boldsymbol{\eta} \, \mathrm{d}\, \boldsymbol{\theta}},$$

which implies that the parameter $\boldsymbol{\pi}$ of the incomplete-data generating process and the parameters $\boldsymbol{\eta}$ and $\boldsymbol{\theta}$ of the population model are independent under the posterior.

Table 1: *Coverage properties of 95% posterior credibility intervals: number of times 95% posterior credibility intervals covered true values of parameters* β, $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, γ_1, γ_2, γ_3 *in %, using a truncated Dirichlet process prior with* $K = 3$ *and* $K = 5$ *subpopulations, respectively.*

Parameter	β	$\eta_{E,1}$	$\eta_{E,2}$	$\eta_{I,1}$	$\eta_{I,2}$	γ_1	γ_2	γ_3
True value	2	8	.25	4	.25	-2	-1	0
$K = 3$	95.4%	95.4%	93.5%	95.2%	94.3%	83.4%	97.4%	89.4%
$K = 5$	93.0%	96.3%	96.3%	96.0%	95.7%	98.4%	100%	96.7%

Figure 1: Data-generating process: Conditional on contacts (undirected lines) among population members (circles), infectious population members (red) spread an infectious disease by contact (directed lines) to susceptible population members (white), which are exposed (gray) before turning infectious (red).

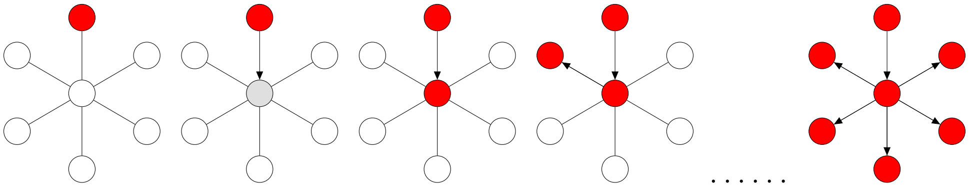


Figure 2: MSE of posterior median and mean of parameters β, γ_1, γ_2, γ_3 plotted against sample size n.

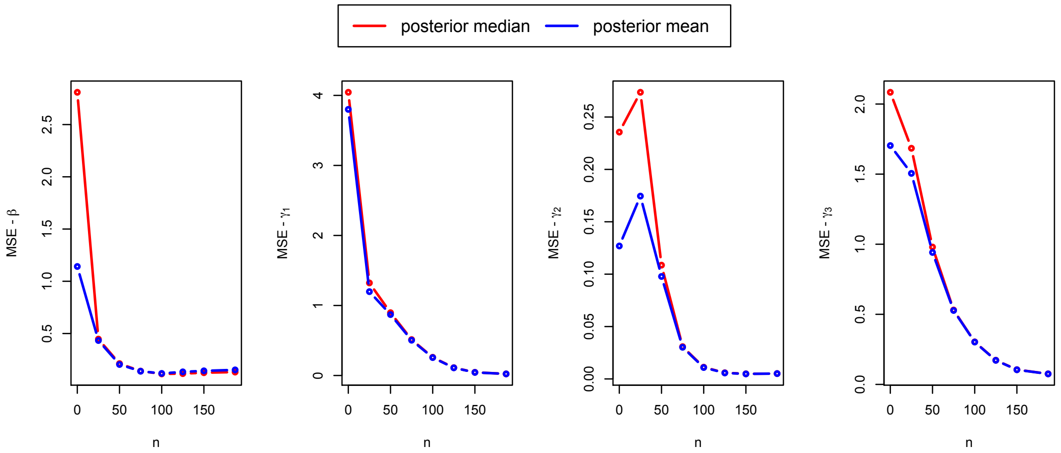


Figure 3: Posterior predictions of number of individuals in the infectious state as a function of time under the model with $K = 1$ (left) and $K = 2$ (right). The observed data are colored red and the predictions are colored black.

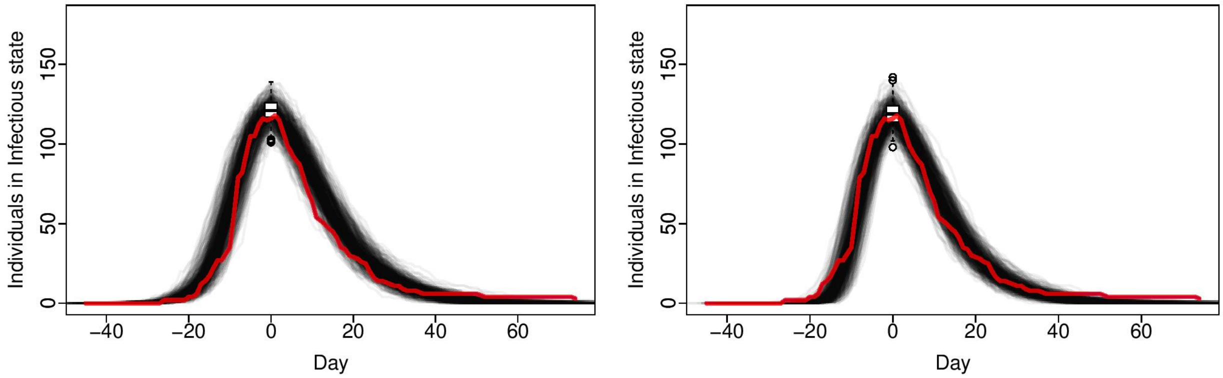


Figure 4: Posterior classification probabilities of population members represented by pie charts. Colors of slices indicate subpopulations and sizes of slices indicate posterior probabilities of belonging to the corresponding subpopulations. A directed edge between two population members indicates a transmission based on the assessments of doctors.

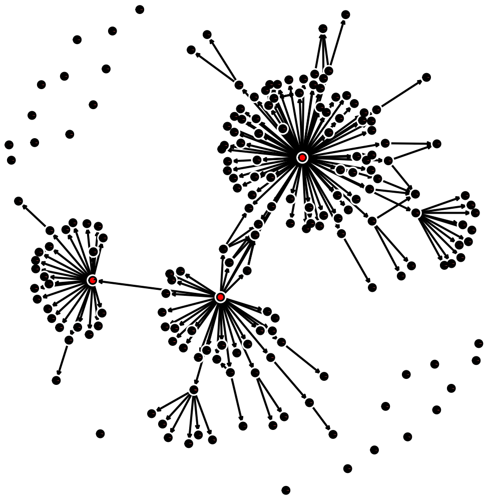
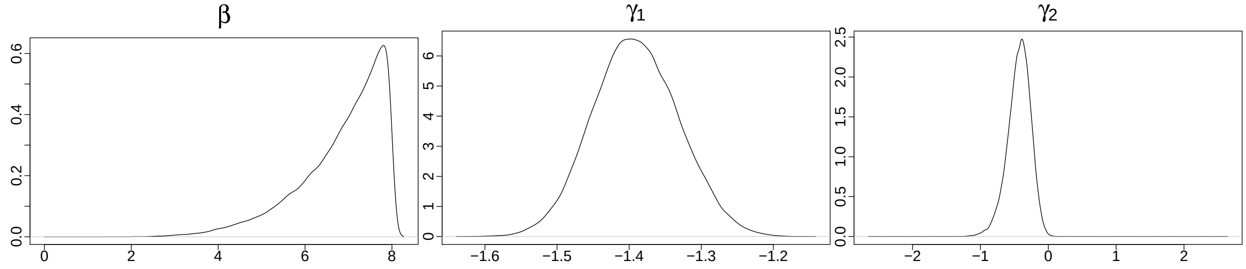


Figure 5: Marginal posterior densities of β, γ_1, γ_2 under the model with $K = 2$.



List of Figures

Supplementary Materials: A Semiparametric Bayesian Approach To Epidemics, with Application to the Spread of the Coronavirus MERS in South Korea in 2015

We follow the truncation approach of Ishwaran and James (2001) and truncate the Dirichlet process prior to facilitate Markov chain Monte Carlo sampling from the posterior. We first discuss the truncation approach and then discuss Bayesian Markov chain Monte Carlo methods and convergence diagnostics. In addition, we present figures that complement the tables and figures in Section 5 of the manuscript.

B Truncation of Dirichlet process priors

To facilitate sampling from the posterior, it is convenient to truncate Dirichlet process priors along the lines of Ishwaran and James (2001). A Dirichlet process prior can be truncated by choosing a large integer $K > 0$—which can be considered to be an upper bound on the number of subpopulations—and sampling

$$V_k \mid \alpha \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha),\ k = 1, 2, \ldots, K-1$$

and setting

$$\begin{aligned} \omega_1 &= V_1 \\ \omega_k &= V_k \prod_{j=1}^{k-1} (1 - V_j),\ k = 1, 2, \ldots, K-1 \\ \omega_K &= 1 - \sum_{k=1}^{K-1} \omega_k. \end{aligned}$$

Truncated Dirichlet process priors approximate Dirichlet process priors (Ishwaran and James, 2001) and imply that the parameters $\omega_1, \ldots, \omega_K$ are governed by a generalized Dirichlet distribution (Ishwaran and James, 2001). The memberships of population members to subpopulations are distributed as

$$\boldsymbol{Z}_i \mid \omega_1, \ldots, \omega_K \overset{\text{iid}}{\sim} \text{Multinomial}(1; \omega_1, \ldots, \omega_K),\ i = 1, \ldots, N.$$

The propensities of population members i to form contacts are given by $\theta_i = \boldsymbol{Z}_i^\top \boldsymbol{\gamma}$, where

$$\gamma_k \mid \mu, \sigma^2 \overset{\text{iid}}{\sim} N(\mu, \sigma^2),\ k = 1, 2, \ldots, K.$$

We use hyperpriors by assuming that α, μ, and σ^{-2} have conjugate Gamma, Gaussian, and Gamma hyperpriors, respectively.

C Bayesian Markov chain Monte Carlo methods

We sample from the posterior by combining the following Markov chain Monte Carlo steps by means of cycling or mixing (Tierney, 1994; Liu, 2008).

Concentration parameter α. If the hyperprior of concentration parameter α is $\text{Gamma}(A_1, B_1)$, we can sample α from its full conditional distribution:

$$\alpha \mid A_1, B_1, \omega_1, \ldots, \omega_K \;\sim\; \text{Gamma}(A_1 + K - 1, B_1 - \log \omega_K).$$

Mean parameter μ. If the hyperprior of mean parameter μ is $N(O, S^2)$, we can sample μ from its full conditional distribution:

$$\mu \mid O, S^2, \sigma^2, \gamma_1, \ldots, \gamma_K \;\sim\; N\left(\frac{S^{-2}\, O + \sigma^{-2} \sum_{k=1}^{K} \gamma_k}{S^{-2} + K\sigma^{-2}}, \frac{1}{S^{-2} + K\sigma^{-2}}\right).$$

Precision parameter σ^{-2}. If the hyperprior of precision parameter σ^{-2} is given by $\text{Gamma}(A_2, B_2)$, we can sample σ^{-2} from its full conditional distribution:

$$\sigma^{-2} \mid A_2, B_2, \mu, \gamma_1, \ldots, \gamma_K \;\sim\; \text{Gamma}\left(A_2 + \frac{K}{2}, B_2 + \sum_{k=1}^{K} \frac{(\gamma_k - \mu)^2}{2}\right).$$

Parameters $\omega_1, \ldots, \omega_K$. We sample $\omega_1, \ldots, \omega_K$ from the full conditional distribution by sampling

$$V_k^\star \mid \alpha, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N \;\overset{\text{ind}}{\sim}\; \text{Beta}\left(1 + N_k, \alpha + \sum_{j=k+1}^{K} N_j\right),\; k = 1, \ldots, K-1$$

and setting

$$\begin{aligned} \omega_1 &= V_1^\star \\ \omega_k &= V_k^\star \prod_{j=1}^{k-1} (1 - V_j^\star),\; k = 2, \ldots, K-1 \\ \omega_K &= 1 - \sum_{k=1}^{K-1} \omega_k, \end{aligned}$$

where N_k denotes the number of population members in subpopulation k.

Indicators $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_N$. We sample indicator $\boldsymbol{Z}_i$ from its full conditional distribution by sampling

$$\boldsymbol{Z}_i \mid \{\boldsymbol{Z}_j\}_{j \neq i}^{N}, \omega_1, \ldots, \omega_K, \gamma_1, \ldots, \gamma_K, \boldsymbol{y} \;\sim\; \text{Multinomial}(1; \omega_{i,1}, \ldots, \omega_{i,K}),$$

where

$$\omega_{i,k} = \frac{\omega_k \displaystyle\prod_{j:\, j \neq i}^{N} p(y_{i,j} \mid \boldsymbol{\theta}, Z_{ik} = 1, \{\boldsymbol{Z}_h\}_{h \neq i}^{N})}{\displaystyle\sum_{l=1}^{K} \left\{ \omega_l \prod_{j:\, j \neq i}^{N} p(y_{i,j} \mid \boldsymbol{\theta}, Z_{il} = 1, \{\boldsymbol{Z}_h\}_{h \neq i}^{N}) \right\}}.$$

Degree parameters $\gamma_1, \ldots, \gamma_K$. We update $\gamma_1, \ldots, \gamma_K$ by Metropolis-Hastings steps, where proposals are generated from random-walk, independence, or autoregressive proposal distributions (Tierney, 1994).

Contact network $\boldsymbol{Y}$. If the value of $Y_{i,j}$ is unobserved, we sample $Y_{i,j}$ from its full conditional distribution:

$$Y_{i,j} \mid \boldsymbol{x}, \beta, \boldsymbol{\theta} \overset{\text{ind}}{\sim} \text{Bernoulli}(q_{i,j}), \tag{23}$$

where

$$q_{i,j} = \frac{\exp(-\beta \max(\min(E_j, R_i) - I_i, 0))\, p_{i,j}(1)}{p_{i,j}(0) + \exp(-\beta \max(\min(E_j, R_i) - I_i, 0))\, p_{i,j}(1)}, \tag{24}$$

where $p_{i,j}(y_{i,j})$ is given by

$$p_{i,j}(y_{i,j}) = \exp(\lambda_{i,j}(\boldsymbol{\theta})\, y_{i,j} - \psi_{i,j}(\boldsymbol{\theta})) \tag{25}$$

and $\lambda_{i,j}(\boldsymbol{\theta}) = \theta_i + \theta_j$ with $\theta_i = \boldsymbol{Z}_i^\top \boldsymbol{\gamma}$.

Transmissions $\boldsymbol{T}$. If transmissions are unobserved, we use the approach of Groendyke et al. (2011, 2012) to update unobserved transmissions.

Exposure, infectious, and removal times $\boldsymbol{E}, \boldsymbol{I}, \boldsymbol{R}$. If exposure and infectious times are unobserved, we use the Metropolis-Hastings steps of Groendyke et al. (2011) to update unobserved exposure and infectious times. If removal times are unobserved, we use the Gibbs and Metropolis-Hastings steps described in the Ph.D. thesis of Bomiriya (2014, Section 4.6) to update unobserved removal times.

Parameter $\boldsymbol{\eta}$. We use Gibbs and Metropolis-Hastings steps to update the elements of $\boldsymbol{\eta}$ along the lines of Groendyke et al. (2011, 2012).

D Markov chain Monte Carlo convergence diagnostics

We sample from the posterior of all models by using the Bayesian Markov chain Monte Carlo algorithm above with 10,000,000 iterations, discarding the first 2,000,000 iterations as burn-in and keeping track of every 100-th iteration.

We used two means to detect possible non-convergence. First, we used trace plots of the all important epidemiological and network parameters, as shown in Figure 7. Second, we

Figure 6: First row: 95%-posterior credibility intervals of parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, β, γ_1, γ_2, γ_3 stacked on top of each other. Second row: histograms of posterior medians of parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, β, γ_1, γ_2, γ_3.

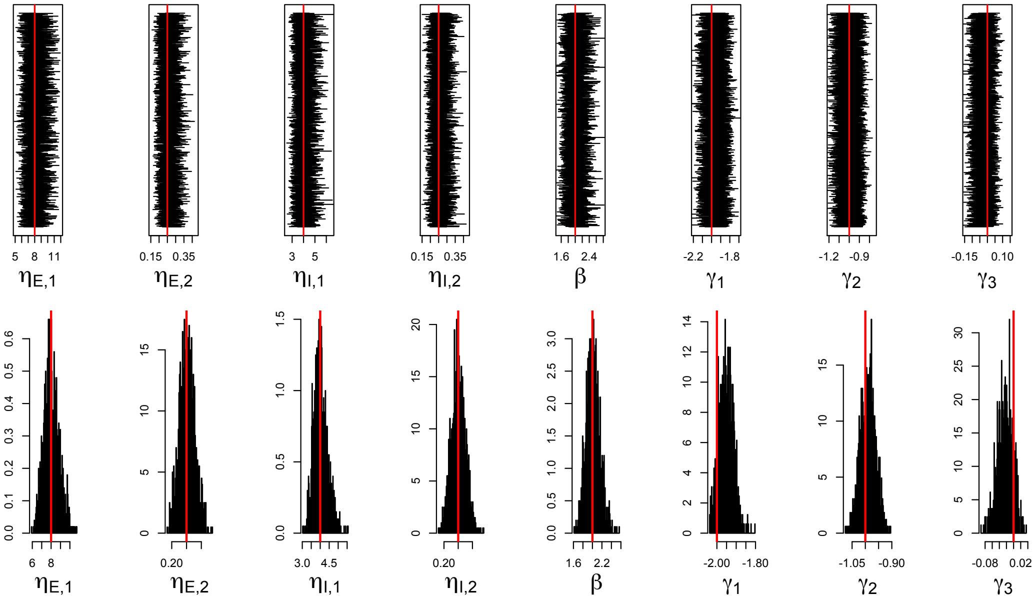


Figure 7: Trace plots of parameters under the model with $K = 1$ (left, β and θ) and the model with $K = 2$ (right, β, γ_1 and γ_2); note that $\theta = \text{logit}(p)$ is the natural parameter of the one-parameter exponential family of Bernoulli(p) distributions.

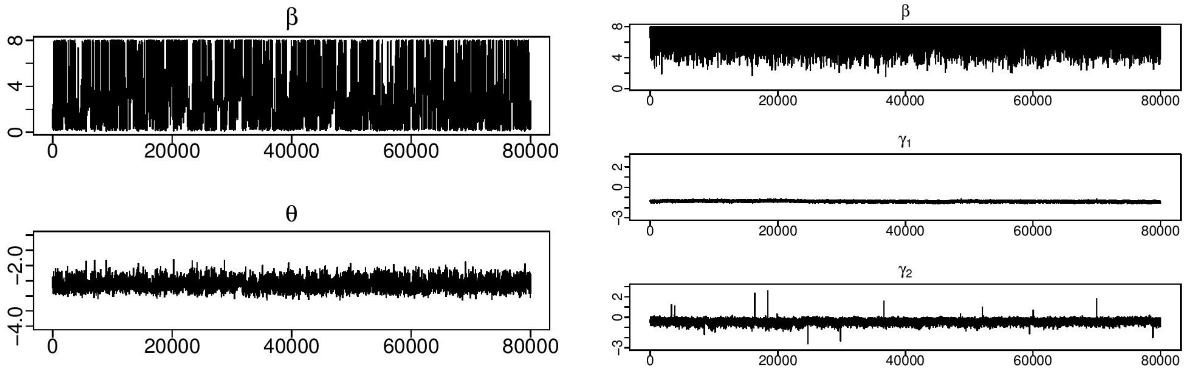
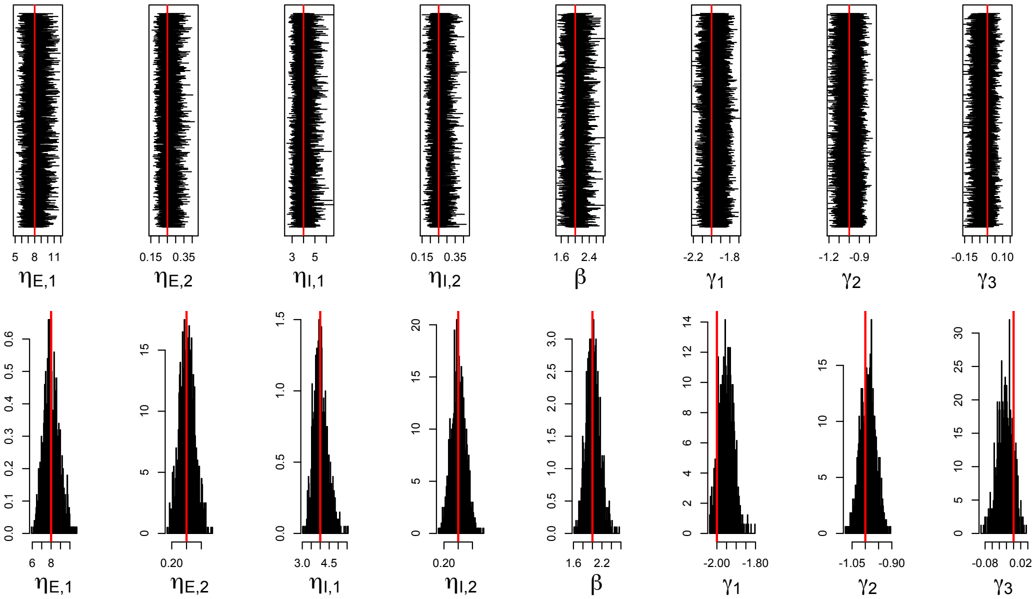


Figure 8: First row: 95%-posterior credibility intervals of parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, β, γ_1, γ_2, γ_3 stacked on top of each other. Second row: histograms of posterior medians of parameters $\eta_{E,1}$, $\eta_{E,2}$, $\eta_{I,1}$, $\eta_{I,2}$, β, γ_1, γ_2, γ_3.



used the convergence checks of Raftery and Lewis (1996) as implemented in the R package `mcgibbsit` (Warnes and Burrows, 2010). According to both, the burn-in and post-burn-in are long enough.

The so-called label-switching problem of Bayesian Markov chain Monte Carlo algorithms, arising from the invariance of the likelihood function to the labeling of subpopulations, is solved by following the Bayesian decision-theoretic approach of Stephens (2000), that is, by choosing a loss function and minimizing the posterior expected loss as described by Schweinberger and Handcock (2015, Supplement C) and implemented in the R package `hergm` (Schweinberger and Luna, 2018).

E Simulation results

Figure 8 shows 95%-posterior credibility intervals of parameters and posterior medians, which complements the tables and figures in Section 5 of the manuscript.