# Inferring contact network characteristics from epidemic data via compact mean-field models

Andrés Guzmán[1], Federico Malizia[1], Gyeong Ho Park[2], Boseung Choi[2,3,4*], Diana Cole[5], István Z. Kiss[1,6†]

[1]Network Science Institute, Northeastern University London, London E1W 1LP, United Kingdom
[2]Division of Big Data Science, Korea University Sejong Campus, Sejong, 30019, Korea
[3]Biomedical Mathematics Group, Institute for Basic Science, Daejeon, 34126, Korea
[4]College of Public Health, The Ohio State University, Columbus, OH 43210, USA
[5]School of Engineering, Mathematics and Physics, University of Kent, Canterbury, United Kingdom
[6]Department of Mathematics, Northeastern University, Boston, MA 02115, USA

June 16, 2025

**Abstract**

Modelling epidemics using contact networks provides a significant improvement over classical compartmental models by explicitly incorporating the network of contacts. However, while network-based models describe disease spread on a given contact structure, their potential for inferring the underlying network from epidemic data remains largely unexplored. In this work, we consider the edge-based compartmental model (EBCM), a compact and analytically tractable framework, and we integrate it within dynamical survival analysis (DSA) to infer key network properties along with parameters of the epidemic itself. Despite correlations between structural and epidemic parameters, our framework demonstrates robustness in accurately inferring contact network properties from synthetic epidemic simulations. Additionally, we apply the framework to real-world outbreaks—the 2001 UK foot-and-mouth disease outbreak and the COVID-19 epidemic in Seoul— to estimate both disease parameters and network characteristics. Our results show that our framework achieves good fits to real-world epidemic data and reliable short-term forecasts. These findings highlight the potential of network-based inference approaches to uncover hidden contact structures, providing insights that can inform the design of targeted interventions and public health strategies.

Keywords: Epidemics, Inference, Contact Networks

## 1 Introduction

The spread of infectious diseases is inherently tied to the structure of human interactions. Network theory provides a powerful framework for understanding how diseases propagate by capturing the complex web of contacts between individuals [1, 2, 3, 4, 5]. Studies have highlighted how structural properties such as heterogeneity [6, 7, 8, 9, 10], communities [11, 12, 13, 14], clustering [15, 16, 17, 18], and degree correlations [19, 20, 21, 22] play a significant role in shaping epidemic dynamics.

Epidemic models have traditionally been used to describe and predict disease spread based on assumptions about the underlying contact structure [2, 23]. Ultimately, their applicability to real-world processes depends on the availability and quality of data [24]. These models range in complexity, from classical mass-action

*Corresponding author. Division of Big Data Science, Korea University Sejong Campus, Sejong, 30019, Korea. E-mail: cbskust@korea.ac.kr

†Corresponding author. Network Science Institute, Northeastern University London, London E1W 1LP, United Kingdom. E-mail: istvan.kiss@nulondon.ac.uk.

approaches, where populations are assumed to mix homogeneously [25], to sophisticated network-based frameworks that explicitly incorporate individual-level connectivity patterns [26, 27, 28]. Simpler models offer easier tractability but may overlook key structural features, while more complex models provide richer descriptions but require more detailed input data [29, 30, 31, 24]. Finding a balance between these aspects is essential for effective epidemic modelling [32].

A particularly elegant and efficient modelling framework is the edge-based compartmental model (EBCM), which provides a compact yet powerful representation of epidemic processes on networks [33, 34, 35]. Unlike standard compartmental models where incorporating heterogeneity significantly increases model complexity, EBCM encodes network structure and characteristics through probability-generating functions, allowing epidemic dynamics to be described with only a few parameters and a reduced number of equations.

In many real-world scenarios, direct measurements of contact networks are unavailable or incomplete. Although collecting data from contact networks is feasible in certain cases, such as sexually transmitted infections [36, 37, 38, 39], it remains challenging for respiratory diseases [40, 41, 42]. While epidemic models are often used to simulate outbreaks given a known network structure, inferring the structure of the contact network from observed epidemic data represents an equally important challenge [43, 44, 45, 46, 47]. Since spreading dynamics inherently reflect network properties, they can be used to extract valuable information about the underlying structural information. Various methods have been proposed to reconstruct networks from data, including likelihood-based optimization approaches [48, 49, 50, 51, 52] and Bayesian inference techniques [53, 54, 55, 56, 57]. However, these methods often require detailed temporal data or strong prior assumptions, making them difficult to apply in real-world epidemic surveillance [58].

An alternative approach, Dynamical Survival Analysis (DSA), has been introduced to estimate epidemic parameters using infection and recovery time distributions [59]. Originally developed for mass-action models [60, 61], DSA was recently extended to network-based models [62], enabling parameter estimation while incorporating some aspects of network structure. However, existing applications remain limited in their ability to fully capture the heterogeneity of contact networks.

In this paper, we integrate the DSA approach with the edge-based compartmental model (EBCM) [33] to develop a Bayesian framework for inferring both disease and network parameters from epidemic data. This extends previous works [63], shifting from identifiability analysis to active inference in both synthetic and real-world scenarios. The manuscript is structured as follows: Section 2.1 introduces the EBCM framework, and Section 2.2 details the Bayesian inference procedure. Section 3 presents validation on synthetic and real data, specifically the first wave of COVID-19 in Seoul and the 2001 foot-and-mouth disease epidemic in the UK. Finally, Section 5 discusses the implications of our findings.

# 2 Methods

In this section, we outline the methodologies that form the foundation of our inference framework. First, we introduce the edge-based compartmental model (EBCM), which provides a compact representation of SIR processes on networks. This model serves as the backbone for describing the epidemic dynamics in structured populations.

Next, we present the complete inference process, detailing how these methods are integrated to estimate both epidemic and network parameters from observed outbreak data. Specifically, we employ Dynamic Survival Analysis (DSA) to construct the likelihood function, leveraging its ability to handle censored and aggregated epidemic data. Moreover, we describe the Robust Adaptive Metropolis (RAM) algorithm, a Markov Chain Monte Carlo (MCMC) technique designed for efficient exploration of the parameter space. RAM adapts to the local structure of the posterior distribution, improving convergence and robustness in high-dimensional settings. Together, these methods form a comprehensive framework for inferring epidemic dynamics and network structures from real-world outbreak data.

## 2.1 Edge-based compartmental model

We consider a Susceptible-Infected-Recovered (SIR) epidemic process, where individuals can be in one of three states: susceptible ($S$), infected ($I$), or recovered ($R$). Infection occurs at rate $\beta$ along a link between a susceptible and an infected node, while infected nodes recover independently of the network at rate $\gamma$.

In this study, we employ the Edge-Based Compartmental Model (EBCM) [33], which provides a compact and analytically tractable representation of epidemic dynamics on contact networks. The EBCM assumes that disease transmission occurs on a network generated by the configuration model (CM) [64, 65], which is characterized by a degree distribution $P(k)$. The key idea behind EBCM is to track the probability that a randomly chosen node remains susceptible rather than explicitly tracking individual infection events.

A central variable in the model is $\theta(t)$, defined as the probability that a randomly selected neighbor of a test node $u$ has not transmitted the disease to $u$ by time $t$. From now on, we omit the obvious time dependence. Given that a node $u$ has degree $k$, the probability that it remains susceptible is $s_u(k,\theta) = \theta^k$. Thus, the overall fraction of susceptible nodes in the population is given by:

$$S(t) = \sum_k P(k)\theta^k = \Psi(\theta), \tag{1}$$

where $\Psi(\theta)$ represents the probability generating function (PGF).

If a fraction $\rho$ of the population is initially infected at $t = 0$, we modify this expression as $S(t) = \hat{\Psi}(\theta) = \sum_k P(k)S(k,0)\theta^k$, where $S(k,0)$ is the probability that a node with degree $k$ is initially susceptible. Since initially infected nodes are selected at random, it follows that $S(k,0) = 1-\rho$. To fully characterize the system, we decompose $\theta$ into three probabilities, namely $\theta = \psi_S + \psi_I + \psi_R$, where $\psi_S$, $\psi_I$, and $\psi_R$ denote the probabilities that a randomly selected neighbor of node $u$ is, respectively, in the susceptible state at time $t$; infected but has not yet transmitted the disease to $u$ by time $t$; or recovered without having transmitted the infection to $u$ during their infectious period. Note that $\dot{\theta} = -\beta\psi_I$, where $\beta$ represents the rate at which an infected partner transmits the disease to the test node. Furthermore, we can express $\psi_I = \theta - \psi_S - \psi_R$, which leads to $\dot{\theta} = -\beta(\theta - \psi_S - \psi_R)$. Additionally, we express $\psi_R$ and $\psi_S$ as $\psi_R = \psi_R(0) + \gamma(1-\theta)/\beta$ and $\psi_S = \psi_S(0)\hat{\Psi}'(\theta)/\langle k\rangle$, where $\hat{\Psi}'(\theta)$ denotes the derivative of the probability generating function (PGF) with respect to $\theta$. Additionally, the average degree can be defined as $\langle k\rangle = \sum_k kP(k)S(k,0)$ which is also equivalent to the derivative of the PGF evaluated at $\theta = 1$. Finally, $\psi_R(0)$ and $\psi_S(0)$ are the probabilities of the test node being initially connected to a recovered or susceptible node respectively. Further details on the derivation of these expressions can be found in the supplementary material or in the original papers [66, 1]. By expressing $\psi_R$ and $\psi_S$ as functions of $\theta$, we can redefine $\dot{\theta}$ as a differential equation that depends only on $\theta$, $\beta$, $\gamma$, and the initial condition. With these considerations, the model is fully described by the following system of equations:

$$\begin{aligned} \frac{d\theta}{dt} &= -\beta\theta + \beta\psi_S(0)\frac{\hat{\Psi}'(\theta)}{\langle k\rangle} + \gamma(1-\theta) + \beta\psi_R(0), \\ \frac{dR}{dt} &= \gamma(1-S-R), \quad S = \hat{\Psi}(\theta). \end{aligned} \tag{2}$$

Typically, we assume $\psi_R(0) = 0$ and $\psi_S(0) = 1-\rho$. Solving Eqs. (2) provides the evolution of $S(t)$, $I(t)$, and $R(t)$. Moreover, the basic reproductive number ($R_0$) of the EBCM is defined as

$$R_0 = \frac{\beta}{\beta+\gamma}\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle}, \tag{3}$$

where $\langle k^2\rangle - \langle k\rangle = \sum_k k(k-1)P(k)S(k,0)$, corresponding to the derivative of the PGF evaluated at $\theta = 1$. For simplicity, from now on, we apply a change of variable $\mu \equiv \langle k\rangle$.

In this study, we consider two different degree distributions, which are summarized in Table 1 along with their parameters and probability generating functions. The Poisson distribution is characterized by a single parameter $\mu$, which defines both its mean and variance, resulting in a relatively homogeneous degree distribution. In contrast, the Negative Binomial distribution, parametrized by $\mu$, which represents the average degree, and $r$, which, together with $\mu$, determine the variance of the distribution. In particular, smaller values of $r$ lead to greater overdispersion. This flexibility makes the Negative Binomial distribution well-suited for modeling both homogeneous and heterogeneous network structures.

## 2.2 Statistical Inference framework

Accurate parameter estimation in epidemic modelling usually relies on optimizing a likelihood function that reflects both the underlying transmission dynamics and the nature of the available data. A common

| | Poisson | Negative Binomial |
|---|---|---|
| Parameter(s) | $\mu$ | $(r, \mu)$ |
| $\Psi(x)$ | $e^{\mu(x-1)}$ | $\left(\frac{r}{r+\mu(1-\theta)}\right)^r$ |
| $\Psi'(x)$ | $\mu e^{\mu(x-1)}$ | $\mu\left(\frac{r}{r+\mu(1-\theta)}\right)^{r+1}$ |

**Table 1:** Details of the probability generating functions used throughout the paper. The parameter $\mu$ for both distributions corresponds to the average degree given by $\sum_k kP(k)$.
**Alt text**: Table comparing the probability generating functions (PGFs) and their derivatives for Poisson and Negative Binomial degree distributions. It includes the associated parameters and analytical forms of $\Psi(x)$ and $\Psi'(x)$ used in the study.

approach involves fitting model-generated epidemic curves to observed data by minimizing discrepancies between them. However, this method is highly sensitive to noise, biases, and incomplete datasets, which can compromise inference accuracy. To address these challenges, we employ the *Dynamic Survival Analysis* (DSA) framework [67, 68, 69, 70], which provides a more robust approach by directly incorporating individual transition times between epidemic states into the likelihood function.

DSA was developed to overcome the limitations of traditional inference methods in infectious disease epidemiology by integrating dynamical systems theory with survival analysis techniques. Unlike conventional approaches that rely on aggregate epidemic curves, DSA leverages the mean-field ordinary differential equations (ODEs) governing population-level dynamics to model the probability distributions of transition times, such as the time of infection or recovery. This formulation allows DSA to construct likelihood functions for individual-level trajectories, making it particularly effective in handling censored, truncated, or incomplete data. In this framework, the susceptible fraction of the population, $S(t)$, is reinterpreted as a survival function, satisfying $S(0) = 1$. More generally, when a fraction $\rho$ of individuals is initially infected, we introduce three rescaled survival functions, which are defined as

$$\tilde{S}(t) = \frac{S(t)}{1-\rho} = \Psi(\theta), \qquad \tilde{I}(t) = \frac{I(t)}{1-\rho}, \qquad \text{and} \qquad \tilde{R}(t) = \frac{R(t)}{1-\rho}. \tag{4}$$

By substituting Eq. (4) in the system of equations for the EBCM, as given by Eqs. (2), we have

$$\begin{aligned} \dot{\tilde{S}}(t) &= \frac{d\tilde{S}}{d\theta}\frac{d\theta}{dt} = \Psi'(\theta)\dot{\theta} = \Psi'(\theta)\bigg[-\beta\theta + \beta(1-\rho)\frac{\hat{\Psi}'(\theta)}{\langle k\rangle} + \gamma(1-\theta)\bigg]. \\ \dot{\tilde{R}}(t) &= \gamma\tilde{I}(t) \qquad \text{and} \qquad \tilde{I}(t) = 1/(1-\rho) - \tilde{S}(t) - \tilde{R}(t), \end{aligned} \tag{5}$$

where, at $t = 0$ we have $\tilde{S}(0) = 1$, $\tilde{I}(0) = \rho/(1-\rho)$ and $\tilde{R}(0) = 0$.

DSA interprets the susceptible curve as an improper survival function representing the time of infection for a randomly chosen initially susceptible individual. That is, $\tilde{S}(t) = \mathsf{P}(T_I > t)$, where the random variable $T_I$ denotes the infection time. The density function of $T_I$ is given by $-\dot{\tilde{S}}(t)$, which is improper since $\lim_{t\to\infty}\tilde{S}(t) = \mathsf{P}(T_I = \infty) > 0$. We define $\mathsf{P}(T_I = \infty) = 1 - \tau$, where $\tau$ represents the final epidemic size. To obtain a proper survival function, we condition it on a final observation time $T \in (0, \infty)$ and the final epidemic size $\tau$ at time $T$. The resulting probability density function $f_\tau(t)$ on the interval $[0, T]$ is then given by:

$$f_\tau(t) = -\frac{\dot{\tilde{S}}(t)}{\tau}. \tag{6}$$

Note that DSA does not require knowledge of recovery times. However, if these times are available, they can be incorporated to enhance the quality of inference. Let $T_R$ represent the time of recovery of an infected individual. Given the infection time $T_I$, the infectious period $T_R - T_I$ follows an exponential distribution with rate $\gamma$. Using equation (6) and the density of the infectious period, we can define the density of the recovery time $T_R$ as:

$$g(t) = \int_0^t f_\tau(u)\gamma e^{-\gamma(t-u)}\mathrm{d}u. \tag{7}$$

Equation (7) represents the convolution of the density of the infection time $f_\tau(t)$ and an exponential distribution with rate $\gamma$, corresponding to the density of the infectious period. In practice, solving the system of ODEs (5) with respect to the observed recovery times is computationally more convenient.

Finally, the normalized density of the recovery time is given by:

$$\tilde{g}(t) = \frac{g(t)}{\int_0^T g(t)\mathrm{d}t}. \tag{8}$$

One of the advantages of the DSA method is the ability to build various likelihood functions based on the observed data. Let $N$ be the size of the population and $M$ be the initial number of infected individuals at the beginning time $t = 0$, and usually $N >> M$. We have $K$ individuals out of $N$, who are infected by time $T$ and $t_1, t_2 ..., t_K$ represent the time of infection for each infected individual. If $L$ individuals have recovered by time $T$ out of a total of $K + M$ infected, let $r_1, r_2, ..., r_L$ denote the time of recovery. If we can observe the time of infection and recovery exactly for each individual, then we can define the infectious period, $w_i = r_i - t_i$ or $w_i = r_i$ for initially infected, $i = 1, 2, \ldots, L$ respectively. Also, we could have $\tilde{L}$ infected individuals who have not recovered by time $T$, $\epsilon_j = T - t_j$ or $\epsilon_j = T$, $j = 1, 2, \ldots, \tilde{L}$, denote the censored infectious period respectively.

Given random samples of time of infection, $t_1, t_2, ..., t_K$,the log likelihood function is given by:

$$\ell_1 = K\log(\tau) + (N - (M + K))\log(1 - \tau) + \sum_{i=1}^{K}\log(f_\tau(t_i)). \tag{9}$$

The log likelihood function for the infectious period, $w_1, w_2, ..., w_L$ is given by

$$\ell_2 = L\log(\gamma) - \gamma\sum_{j=1}^{L} w_j. \tag{10}$$

The log likelihood function for the time of recovery, $r_1, r_2, ..., r_L$ is given by

$$\ell_3 = \sum_{j=1}^{L}\log(\tilde{g}(r_j)). \tag{11}$$

Finally, we can define the log likelihood for the censored infectious period, $\epsilon_1, \epsilon_2, \ldots, \epsilon_{\tilde{L}}$;

$$\ell_4 = -\gamma\sum_{k=1}^{\tilde{L}}\epsilon_k. \tag{12}$$

In some cases, we only know the number of infected individuals, $K$, by a given time, $T$, but lack information about the total population size $N$. This situation is common in real epidemic scenarios, where data about the segment of the population at risk of infection is often unavailable. For such cases, the likelihood $\ell_1$ can be reformulated as: $\ell_1 = \sum_{i=1}^{K} f_\tau(t_i)$, excluding terms related to the total population size. With this formulation, we can analyze the dynamics of the proportion of infected, susceptible, and recovered individuals. Furthermore, it is possible to estimate an effective population size using the following equation

$$N_{eff} = \frac{K}{1 - S_T}. \tag{13}$$

Depending on the data and particular context, we can use any of the four likelihoods described above or a combination of them. For example, suppose we just observe $K$ infections before a cut-off time $T$, then we could just use $\ell_1$ (9). If, besides the infection times, we also observe $L$ recovery times but not the specific nodes who underwent these changes, then we could use the combination of $\ell_1 + \ell_3$, see Eqs. (9) and (11). However, if the data is available at the node level and we have pairs of infection and recovery times for each node, we could use the combination $\ell_1 + \ell_2$, see Eqs. (9) (10), since we can calculate the infectious period for each node. The likelihood functions described above do not explicitly incorporate the network parameters, making it impossible to derive closed-form solutions for their maximum likelihood estimates. To address this limitation, we adopt a Bayesian approach, which allows us to sample from the posterior distribution and derive point estimates from it. To construct the sample, we employ a Markov Chain Monte Carlo (MCMC) method, specifically the Robust Adaptive Metropolis (RAM) algorithm. The RAM algorithm is more efficient than the standard Metropolis-Hastings algorithm [71], as it dynamically adjusts the variance-covariance matrix of the proposal distribution to maintain an optimal acceptance rate during the Metropolis steps. We assign vaguely informative prior distributions (little information is given about the parameters to be estimated) to the model parameters: a Gamma distribution, Gamma$(a, b)$, for the $\beta$ and $\gamma$ parameters, which is defined as

$$\text{Gamma}(x, a, b) \sim \frac{b^a x^{a-1} e^{-bx}}{\Gamma(a)}, \tag{14}$$

where $\Gamma(a)$ represents the gamma function. Additionally, for the parameter $\rho$ in the SIR model, we assume a Beta distribution, Beta$(a, b)$, where, in this context, $a = 1$ and $b = 1$ are chosen from a uniform distribution. Additionally, we assign a non-informative Gamma prior to the parameters of the degree distribution based on the support of these parameters.

Given a dataset $\mathbf{D}$, our goal is to fit the data with the EBCM, which is characterized by the set of parameters $\mathbf{X} = (\beta, \gamma, \rho, \boldsymbol{\Delta})$. Here, we use the vector $\boldsymbol{\Delta}$ to represent the parameters of the probability generating function of the network's degree distribution. Our objective is to infer the parameter set $\mathbf{X}$ such that the output from the EBCM matches the observed data. To achieve this, we utilize the RAM algorithm to generate one or more chains for the values of $\mathbf{X}$, which allows us to construct a sample of the posterior distribution. From this sample, we remove the burn-in and apply a thinning procedure to reduce autocorrelation. The inference scheme is illustrated in the left panels of Fig. 1.

Once a posterior distribution is obtained, we can find a point estimate for each parameter. This can be done by either calculating the *marginal mean*, *marginal median* or taking the *joint mode* of the posterior distribution, i.e., the point of highest probability in the full posterior. As shown in the top right panel in Fig. 1. Either of these can be used in conjunction with the EBCM to produce one single epidemic curve that can be compared to the data. However, to generate a credible interval around this single epidemic curve, we take a subsample $\mathbb{S}$ from the full posterior and use each element to solve the EBCM, thereby generating a set of solutions, as shown in the bottom right panel of Fig. 1. These can used to generate the desired credible interval over a specified period of time.

In subsequent sections, we show the application of this workflow to estimate network and epidemic dynamics parameters for two different scenarios. First, by the use of synthetic data corresponding to a controlled case where ground truth is available. Second, we consider two different datasets corresponding to real-world epidemics.

# 3 Inference from synthetic data

We begin by analysing synthetic data to verify the method's accuracy and effectiveness in a controlled environment. First, we consider 100 realizations of Gillespie simulations [72] with $\beta = 0.2$, $\gamma = 1$, on networks with $10^4$ nodes exhibiting Poisson ($POI(\mu = 10)$) and Negative Binomial ($NB(\mu = 10, r = 1)$) degree distributions. Since in a real epidemic process, complete data are rarely available, we consider four possible scenarios which reflect varying data availability, which are:

- $\ell_1 + \ell_2$: infection times $(t_1, t_2, \ldots, t_K)$ and infection periods $(w_1, w_2, \ldots, w_L)$

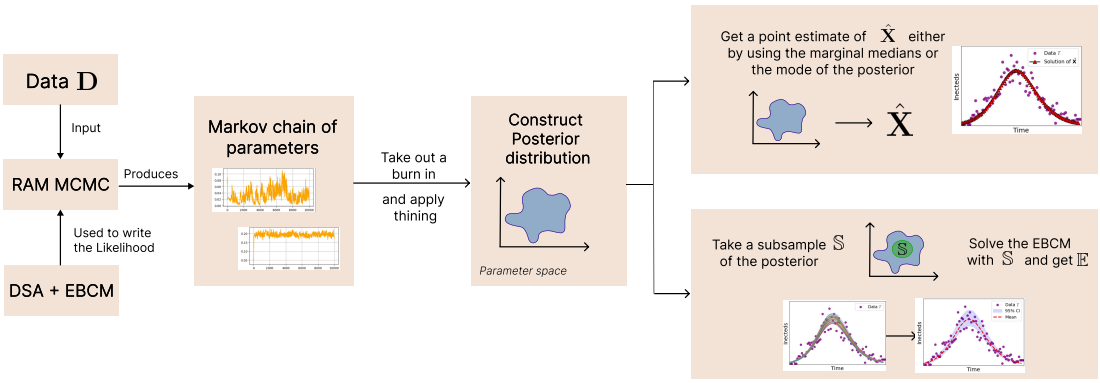


**Figure 1: Graphical representation of the inference process.** The procedure starts with epidemic data **D** as input for the Robust Adaptive Metropolis (RAM) algorithm, which employs the likelihood function derived from the DSA framework (see main text). The RAM algorithm generates a Markov chain of parameter samples, forming the posterior distribution. From this distribution, we obtain parameter estimates using two approaches: credible intervals and point estimates, as detailed in the main text.
**Alt text**: Flowchart of the inference process: epidemic data D feeds into the RAM algorithm, linked to a likelihood function, producing a Markov chain and posterior distribution, which yields credible intervals and point estimates.

- $\ell_1 + \ell_3$: list of infection times $(t_1, t_2, \ldots, t_K)$ and a decoupled list of recovery times $(r_1, r_2, \ldots, r_L)$.
- $\ell_1$: infection times only $(t_1, t_2, \ldots, t_K)$.
- $\ell_3$: recovery times only $(r_1, r_2 ... r_L)$.

Likelihood $\ell_4$ is not used for the synthetic data, as the cut-off time is set after the end of the epidemic. However, this likelihood is used in Section 4.2.

Following the procedure outlined in Section 2.2, we fit our models to all stochastic realizations. Initially, we consider the case where the data are fitted using the correct model. Specifically, data from simulations on networks with a Poisson degree distribution were fitted using the Poisson (Poi) model in the EBCM; we refer to this scenario as Poisson-Poi. Similarly, a match with the Negative Binomial is denoted by Negbin-NB. We also consider model mismatch, where data from networks with a Poisson degree distribution were fitted using the Negative Binomial (NB) model (Poisson-NB), and vice-versa, that is Negbin-Poi.

For all four possible combinations (Negbin-NB, Poisson-Poi, Negbin-Poi, Poisson-NB), we used the four different scenarios discussed above. Then, for each stochastic realization, we construct a sample of the posterior distribution and calculated the marginal median of $\beta$, $\gamma$, and $\mu$, and computed the basic reproduction number as defined in Eq.(3). The point estimates for $\hat{\beta}$, $\hat{\gamma}$, $\hat{\mu}$, $\hat{R}_0$ are obtained as the mean values of the distributions of the marginal medians. In Fig. 2, we show the density of the marginal medians and their mean, i.e., point estimates, for these four parameters and for both model match and mismatch. Additionally, we present these results for the four different data availability scenarios presented above.

As expected, the correct matches, Negbin-NB, first column, and Poisson-Poi, second column, yield accurate estimations for all parameters and for all different likelihoods. In the presence of data and model mismatch, we observe that fitting Poisson data with the Negative Binomial model (fourth column) often yields reasonable results. This can be attributed to the flexibility of the Negative Binomial distribution, i.e. one more free parameter compared to Poisson, which allows it to capture both homogeneous and heterogeneous degree distributions. In contrast, fitting data generated with the Negative Binomial with the Poisson model underperforms. This is especially evident in the estimates for $\beta$ and $\mu$ in the first and third rows, where the EBCM with a Poisson degree distribution tends to overestimate $\beta$ and underestimate $\mu$.

It is noteworthy that the mismatched Negbin–Poi model produced better estimates for $\beta$ and $\mu$ in the $\ell_1 + \ell_3$ and $\ell_3$ scenarios. However, even in these cases, $\gamma$ was substantially overestimated. These qualitative observations are further supported by the larger bias and mean squared error (MSE) of the estimates, as detailed in Supplementary Material (SM). For the estimation of $\gamma$, scenario $\ell_1 + \ell_2$ uses the infectious periods as data. Since the infectious period directly reflects $\gamma$ (the inverse of the mean infectious period), this scenario

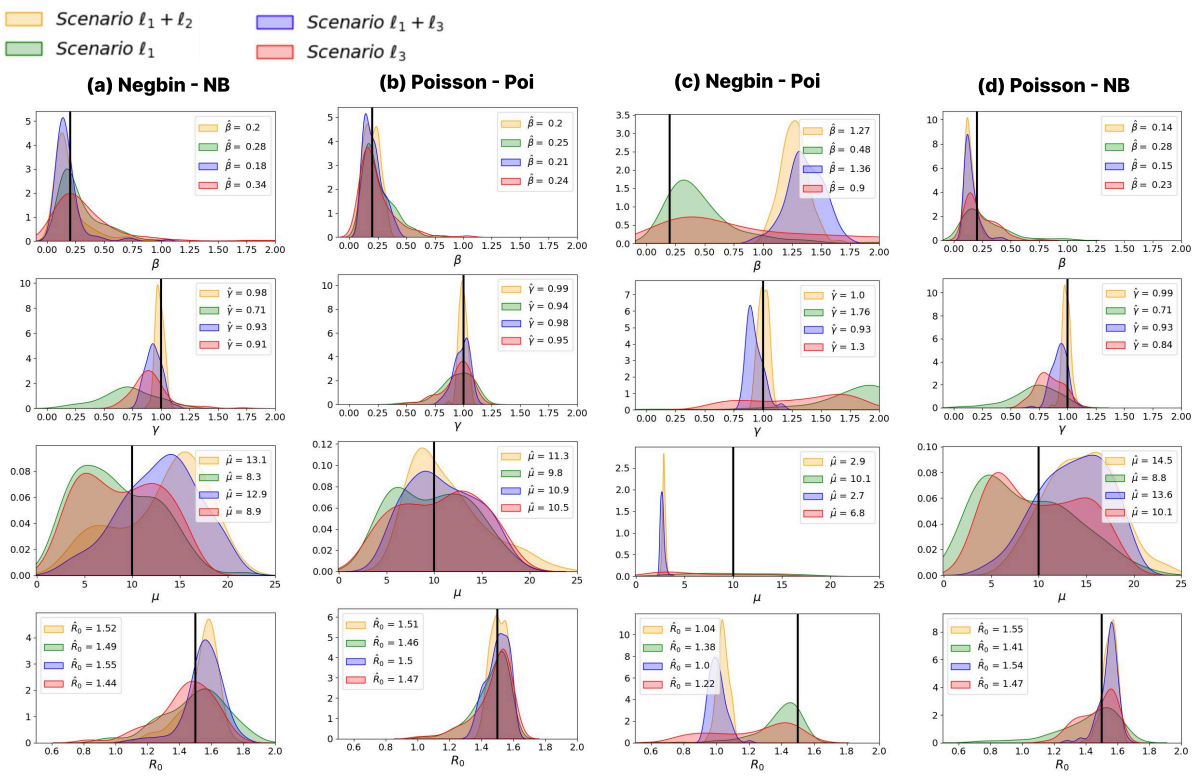


**Figure 2: Distributions of the marginal medians for model parameters and the basic reproduction number inferred from the DSA framework.** Each row represents the distribution of a different inferred parameter. The synthetic data were generated using either a Poisson or a Negative Binomial degree distribution, and inference was performed using both distributions for comparison. Each column corresponds to different inference cases: (a) and (b) show results where the data were fitted with the same degree distribution used for generation—(a) for Negative Binomial and (b) for Poisson. (c) and (d) show cases where the data were fitted using the opposite degree distribution—(c) for Negative Binomial data fitted with a Poisson model and (d) for Poisson data fitted with a Negative Binomial model. This analysis highlights the impact of assuming different degree distributions on parameter inference. The results are obtained by fitting 200 distinct realizations of Gillespie simulations with parameters $\beta = 0.2$, $\gamma = 1$, and an initial proportion of infected individuals set to $10^{-4}$. Networks with a Negative Binomial (Negbin) degree distribution were generated using parameters $\mu = 10$, $r = 1$, while networks with a Poisson degree distribution were generated with $\mu = 10$

**Alt text**: Panel of plots showing marginal median distributions for inferred model parameters and basic reproduction number. Rows represent different parameters; columns compare inference results under matching and mismatched degree distributions (Poisson or Negative Binomial).

yields highly accurate estimates, even for mismatched models and data. By contrast, estimates of $\gamma$ exhibit a larger bias in mismatched scenarios $\ell_1 + \ell_3$ and $\ell_3$, where the recovery times are used instead, but they do not align with the infection times. Interestingly, in the mismatch cases, in particular, see the panel on the third row and third column in Fig. 2, we note that the estimates of $\mu$ based on scenario $\ell_1 + \ell_2$ perform poorly when compared to estimates based on scenarios $\ell_1$ and $\ell_3$. This, at first, is surprising since $\ell_1 + \ell_2$ provides the most complete data. However, due to the model mismatch manifesting itself mainly in the infection part, this is likely to lead to a better estimation of $\gamma$, which in turn has a negative effect on the estimation of $\beta$ and $\mu$. In contrast, in scenarios $\ell_1$ and $\ell_3$, the estimation of $\mu$ improves at the expense of obtaining less accurate estimates of $\gamma$. These observations are further supported by the MSE values presented in SM, where this analysis is discussed in detail.

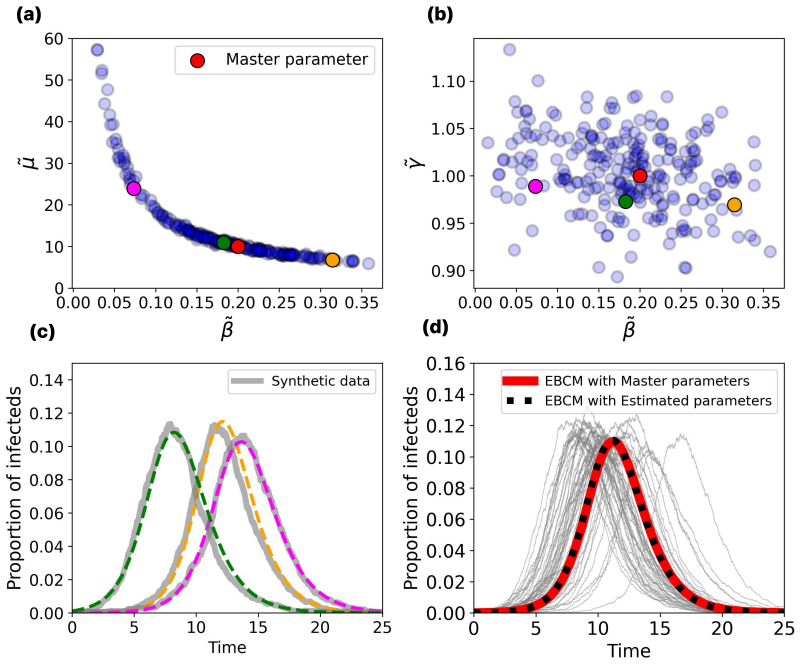


**Figure 3: DSA parameters estimation for each synthetic dataset.** (a) Projection of the 250 median estimates from the marginal posterior distributions of each dataset in the $\beta$–$\mu$ space. (b) Projection of the median estimates in the $\gamma$–$\beta$ space. (c) Epidemic curves from three selected datasets (gray lines) alongside the corresponding curves obtained by solving the EBCM using the median parameter estimates (pink, green, and orange points in panels (a) and (b)). Despite variations in parameter estimates, the model accurately reproduces the observed epidemic dynamics of the data. Panel (d) shows 50 of the 250 datasets (gray lines) with the solution of the EBCM obtained using the master parameters (dotted black line) and the solution from the final point estimates from the marginal posteriors (red line). Results are based on 250 synthetic epidemic datasets generated on a network with a Poisson degree distribution. The true parameters used were $\beta = 0.2$, $\gamma = 1$, $\mu = 10$, and $\rho = 1 \times 10^{-4}$. For inference, the cut-off time was set to $T = 25$.
**Alt text**: Multi-panel figure showing parameter estimates from 250 synthetic epidemic datasets. Panels (a) and (b) project median estimates in $\beta - \mu$ and $\gamma - \beta$ spaces. Panel (c) compares observed epidemic curves with model outputs using selected parameter estimates. Panel (d) shows epidemic curves from 50 datasets, alongside model solutions using master and inferred parameters, illustrating accurate recovery of epidemic dynamics.

Even though accurate and precise estimations can be obtained using the marginal medians of the posterior distributions, it is important to recognize that the posterior distribution is multivariate. This means that the marginals are not always fully representative of the entire distribution, especially when there is a correlation between parameters. When used for inference, network-based mean-field models have been noted to exhibit correlations between the infectivity $\beta$ and parameters of the degree distribution, such as the average degree $\mu$ [62]. For this reason, we now investigate the projections of the posterior distribution. We do this for Poisson-Poi case, and we consider the scenario where the data is complete, i.e., $\ell_1 + \ell_2$. In Fig. 3(a), we show the projection of the medians of the marginals in the $(\beta, \mu)$ space. We observe a strong inverse correlation between the infection rate $\beta$ and the average degree $\mu$.

Although we observe a considerable variance in both the average degree and the infection rate, it's important to highlight that most of the estimates of parameters are around the master parameter used in the simulation. Furthermore, no correlation is observed when exploring the $(\beta, \gamma)$ parameter space. As expected, the correlations between parameters can be mitigated by fixing either the infection rate or the average degree. This approach leads to a posterior sample without long tails, resulting in more accurate

estimates. Further discussions on the process of fixing parameters are provided in SM.

Up to this point, we have focused on parameter estimation with their validity based on comparison to their true values. However, another important consideration is to assess how the epidemic curve, such as new infecteds or prevalence, generated using the EBCM with the point estimates compares to the original outbreak data. To address this, we consider three different stochastic realizations of the epidemic, see grey curves in Fig. 3(c). We fit each of these separately with the corresponding point estimates reported as the pink, green, and orange points in Fig. 3(a) and Fig. 3(b). The EBCM with these point estimates leads to excellent agreement with the original stochastic realizations, see Fig. 3(c) and Fig. 3(d).

# 4 Inference from real-world data

In this section, we further validate our methodology by analyzing real epidemic data from two outbreaks. Based on the general procedure presented in Section 2.2, we apply it to two real epidemic datasets. The first one is the 2001 Foot-and-Mouth Disease (FMD) epidemic in the United Kingdom, which involved a highly contagious virus affecting farm animals. The second dataset captures the first wave of COVID-19 in Seoul, South Korea, documenting the onset of symptoms and confirmation of infection for approximately 500 individuals over the first 82 days following the appearance of the initial confirmed case.

For both datasets, we fit the EBCM using a probability-generating function corresponding to a Negative Binomial degree distribution, chosen for its flexibility in modelling both homogeneous and heterogeneous contact patterns. We adopt non-informative prior distributions, assigning $\beta$, $\gamma$, $\mu$, and $r$ a GAMMA$(a, b)$ prior, where $a$ is randomly selected from the parameter space and $b$ is fixed at $10^{-4}$. The initial number of infected individuals follows a BETA$(1, 1)$ prior, representing a uniform distribution and reflecting complete uncertainty about the starting conditions. A key distinction from our analysis of synthetic data is that the total population size is unknown. As discussed in Section 2.2, one approach is to estimate an effective population size using $N_{eff} = K/(1 - S[T])$ [68]. To obtain a denser posterior distribution, we performed multiple runs of the RAM algorithm, varying the initial conditions of the chain for each run. Consequently, we find point estimates for each parameter and evaluate the accuracy of the EBCM predictions by comparing them to real-world epidemic data. To quantify the discrepancy, we use the mean squared error (MSE), defined as

$$MSE = \frac{1}{T}\sum_{d=0}^{T}(\hat{J}(d) - J(d))^2, \tag{15}$$

where $d$ denotes the day index, $T$ is the cut-off time, $\hat{J}(d)$ represents the incidence predicted by the EBCM, and $J(d)$ corresponds to the epidemic incidence observed in the real-data.

Here, we compare the predictions of the EBCM with those obtained using the standard mass-action (MA) SIR model [59]. The MA model is incorporated into the inference framework described in Section 2.2 and serves as a benchmark for epidemic curve predictions. The governing equations for the SIR MA model are given by:

$$\dot{s}_t = -\sigma s_t \iota_t, \qquad \dot{\iota}_t = \sigma s_t \iota_t - \gamma \iota_t, \qquad \dot{r}_t = \gamma \iota_t, \tag{16}$$

where $\gamma$ is the recovery rate, and $\sigma$ is the infection rate. It is important to note that $\sigma$ differs from the infection rate $\beta$ in the EBCM: while $\beta$ represents the per-contact transmission rate, $\sigma$ describes the infection rate per infectious individual.

Unlike the EBCM, which explicitly accounts for the network structure, the MA model assumes homogeneous random mixing, meaning it does not incorporate any connectivity patterns. As a result, while the EBCM enables inference of both the epidemic dynamics and the underlying contact structure, the MA model can only be used to predict the epidemic curve. Consequently, our comparison is limited to the epidemic trajectory rather than network-related properties. For parameter inference, we follow the same numerical procedure as outlined earlier.

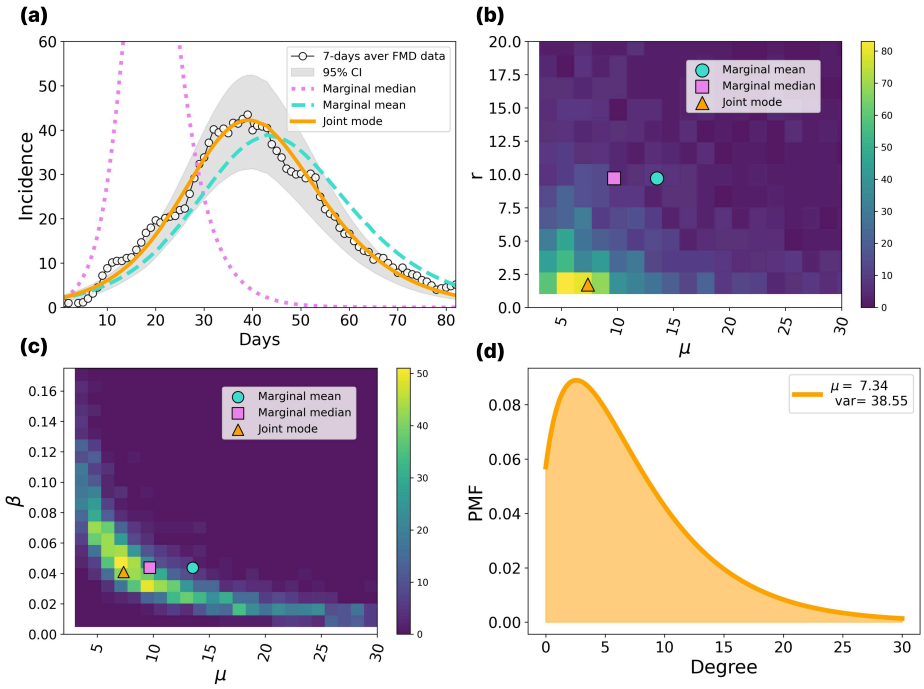


**Figure 4: Summary of inference results for the 2001 Foot-and-Mouth Disease outbreak in the UK.** (a) Evolution of the incidence based on parameter estimates from the three methods, alongside the 95% credible interval from the posterior distribution. The observed FMD incidence, smoothed using a 7-day moving average, is also shown for comparison. (b) and (c) Heatmaps of the posterior distribution projections, showing parameter density across the $(\beta, \mu)$ and $(r, \mu)$ spaces, respectively. Each projection includes point estimates obtained using the *marginal mean*, *marginal median*, and *joint mode*. (d) Inferred degree distribution obtained using the *joint mode* estimation method.
**Alt text**: Figure summarizing inference results for the 2001 UK Foot-and-Mouth Disease outbreak. Panel (a) shows modeled incidence curves from three estimation methods with 95% credible intervals, compared to original data. Panels (b) and (c) are heatmaps of posterior density in $\beta$ - $\mu$ and $r$ –$\mu$ parameter spaces, including key point estimates. Panel (d) displays the inferred degree distribution based on the joint mode estimate.

## 4.1 Foot-and-Mouth disease data

In this section, we analyze the Foot-and-Mouth Disease (FMD) dataset, which provides daily incidence data, $J(d)$, representing the number of newly infected individuals per day over a 200-day period. We focus on the first 82 days, corresponding to the initial wave of the outbreak. In the DSA framework, performing inference requires information about the temporal distribution of infections. To address this, we assume that infection times within each day are uniformly distributed and, based on this assumption, consider only the $\ell_1$ likelihood.

We generate five independent Markov chains, each running for 100,000 iterations, following the methodology outlined in Section 2.2. To construct the posterior distribution, we discard the first 50,000 iterations as burn-in and thin the remaining samples by selecting every 50th iteration to mitigate autocorrelation. From the resulting posterior, we estimate the five parameters $(\beta, \gamma, \mu, r, \rho)$ using three different methods: *marginal mean*, *marginal median*, and *joint mode*, as shown in 1. The *joint mode* is computed using the mean-shift algorithm, which approximates the kernel density based on the posterior sample.

In Fig. 4(a), we compare the EBCM solutions with real epidemic data using different parameter estimation methods. As shown, the parameters obtained via the *marginal mean* fail to capture the system's behaviour. While the EBCM integrated with parameters estimated through the *marginal median* provides a better fit, it still does not fully capture the original data. However, the EBCM solution obtained using parameters

estimated via the *joint mode* exhibits excellent agreement with the epidemic data. Additionally, we sample a subset of parameters, **S**, from the posterior distribution to construct a 95% credible interval for the daily new infections, as mentioned in the bottom right of 1. Notably, the prediction based on the *joint mode* lies at the centre of this interval. This is further illustrated in Fig. 4(b)-(c), where we present projections of the full posterior distribution for the $(\beta, \mu)$ and $(r, \mu)$ parameter spaces, respectively. These figures highlight that the *marginal mean* and *marginal median* estimates lie outside the region of highest posterior density. On the other hand, point estimates obtained using the *joint mode* are closer to the densest regions of the posterior projections when compared to the estimates from the marginal distributions. The discrepancy is likely due to the presence of long tails in the posterior distributions, which obscure important correlations between parameters and lead to suboptimal point estimates. In Table 2, we provide the point estimates for each method, along with the corresponding values of the $MSE$ calculated from the infeed incidence.

Additionally, in Fig. 4(d), we show the degree distribution with values of the mean and variance that are in line with expectations since highly connected markets were disproportionately affected at the beginning of the outbreak leading to a marked reduction in network heterogeneity [73]. While the model successfully captures the epidemic curve, the inferred contact network structure cannot be directly validated against ground truth data, as no empirical contact network is available for comparison.

Furthermore, we compare the predictions made by the EBCM, with the ones made by using the MA model. In this case, the parameters exhibit a lower correlation, resulting in a posterior distribution without long tails. Consequently, the *marginal median*, *marginal mean*, and *joint mode* produce similar estimates. In the table at the top-left corner in Figure 5, we show the estimates of $\beta$, $\gamma$, and $R_0$. Using these estimates, we can compare the MA and EBCM models in their ability to describe the original data. In Fig. 5, we plot the incidence along with the MSE calculated for each case. The EBCM, which accounts for an explicit contact structure, provides a closer fit to the original data compared to the classical MA model. This is clearly evidenced by the lower MSE values achieved by the EBCM, suggesting that the FMD outbreak is better described by a model with an explicit contact structure rather than a homogeneously mixed model.

## 4.2 COVID-19 data from Seoul, South Korea

In this section, we extend our analysis to the first wave of the COVID-19 pandemic in Seoul, South Korea, covering an 84-day period from January 26 to April 18, 2020 [74]. This dataset provides information on the time of symptom onset for each individual, as well as the date of their positive test result. For our analysis, we assume that the time of infection coincides with the onset of symptoms. Given the strict isolation measures in place, we treat the time of a positive test result as the effective recovery time, as individuals were promptly isolated upon testing positive. Using this dataset, we can aggregate the information to track the evolution of incidence (the number of new infections per day, as in Section 4.1), prevalence (the total number of currently infected individuals at any given time), and daily recoveries (the number of individuals recovering each day).

| Parameter | *marginal mean* | *marginal median* | *joint mode* |
|---|---|---|---|
| $\beta$ | 0.043 | 0.036 | 0.041 |
| $\gamma$ | 0.30 | 0.28 | 0.32 |
| $\mu$ | 13.51 | 9.70 | 7.34 |
| $r$ | 9.71 | 5.25 | 1.72 |
| Variance | 32.32 | 27.58 | 38.55 |
| $R_0$ | 1.89 | 1.32 | 1.31 |
| MSE Incidence | 3.44 | 1.03 | 0.76 |

**Table 2:** Estimated parameter values obtained using three inference methods: *marginal mean*, *marginal median*, and *joint mode*, for the the Foot-and-Mouth disease outbreak. The table also reports the Mean Squared Error (MSE) between the predicted incidence evolution and the observed data.
**Alt text**: Table showing estimated parameter values for the Foot-and-Mouth Disease outbreak using three inference methods: *marginal mean*, *marginal median*, and *joint mode*. Parameters include $\beta$, $\gamma$, $\mu$, $r$, variance, $R_0$, and the Mean Squared Error (MSE) of predicted incidence.

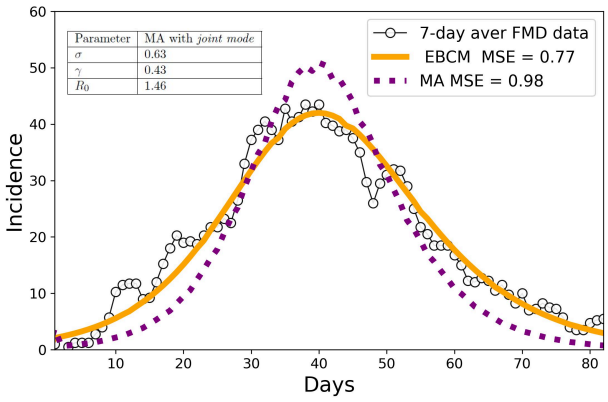


**Figure 5: Comparison of predicted incidence of infections using the EBCM and mass-action model for Foot-and-Mouth Disease (FMD).** The estimated incidence curves are presented for the DSA framework using the Edge-Based Compartmental Model (EBCM, shown as the orange curve) and the Mass-Action (MA) model (depicted as purple squares). Point estimates were derived using the joint mode for both models. The table in the top-left corner displays the results of the MA model, while the EBCM estimates can be found in Table 2. The black circled line represents the 7-day moving average of observed FMD cases.
**Alt text**: Predicted FMD incidence from EBCM (orange) and Mass-Action model (purple squares), compared to observed data (black line). Includes summary table for MA model.

Notably, the Seoul COVID-19 dataset also includes contact pattern data, recording the number of contacts each infected individual had between symptom onset and recovery. This allows us to use the contact data as ground truth for evaluating the network structure characteristics inferred by the EBCM. By comparing the inferred network properties with empirical contact data, we assess the capability of the model to recover meaningful structural information from epidemic observations. To perform our analysis, we consider the Negative Binomial degree distributions and follow the same procedure as in Section 4.1. Additionally, we consider the likelihood $\ell_1 + \ell_2$, which is able to accommodate the most complete type of data.

In Fig. 6, we show the predictions based on the EBCM for prevalence, incidence, and daily recoveries.

As with the FMD disease dataset (Section 4.1), the set of parameters estimated using the *joint mode* provides a temporal evolution of the system that aligns closely with the original data. In this case, the *marginal median* also performs well, yielding results that are closer to those obtained with the *joint mode*. In contrast, the *marginal mean* produces the least accurate results, failing to capture the behaviour of the system. These findings are illustrated in Fig. 6(a), 6(b), and 6(c), which show the evolution of incidence, prevalence, and daily recoveries for each method. Additional evidence for these results is provided by the Mean Squared Error (MSE) values in Table 3.

In Fig. 6(d), we compare the inferred degree distribution with the one obtained from the contact data. While the inferred distribution captures key characteristics of the empirical data, the point estimates derived from the *joint mode* yield a lower average degree and variance than those observed in the real network. Notably, the Negative Binomial distribution struggles to fully reproduce the long tail present in the empirical contact data, suggesting that higher-degree individuals may be under-represented in the inferred network structure.

Furthermore, we can use the contact data to fix the average degree. In this case, we fix $\mu = 7.61$, which is the average number of contacts entered in the survey. We can apply the inference procedure to find $\beta$, $\gamma$, and the parameter $r$ of the Negative Binomial distribution. Table 4 presents the point estimates obtained from the three methods: *marginal mean*, *marginal median*, and *joint mode*. In this scenario, the three methods produce very similar estimates. Furthermore, in Fig. 7(a), we observe the incidence predicted by the three methods, which are nearly identical. This indicates that fixing the average degree results in a much better-behaved posterior distribution. It is important to note that the contact data collected should be interpreted with caution. This data is highly dependent on an individual's perceptions and, as such, cannot be treated

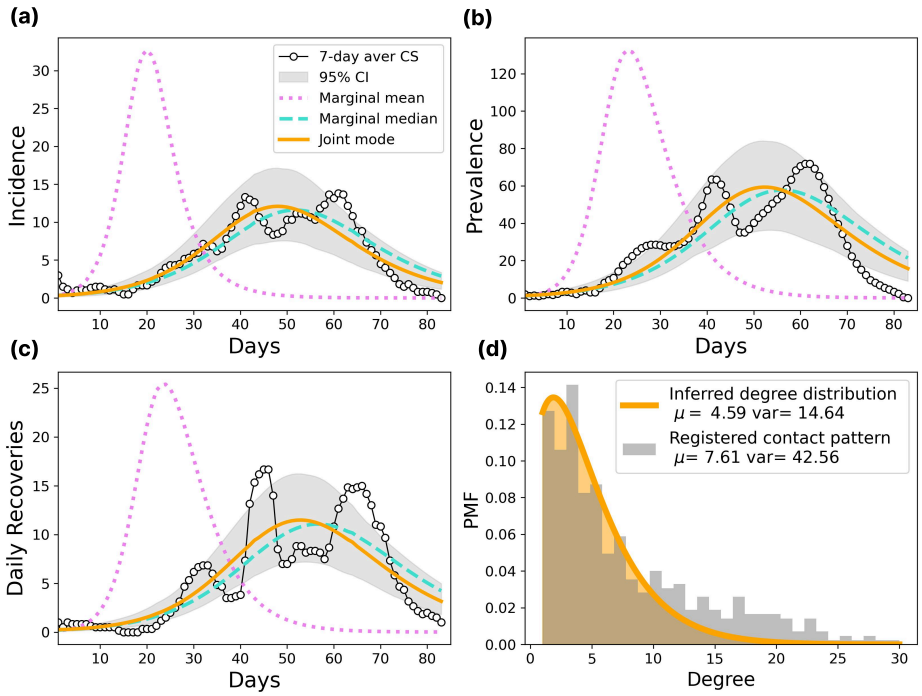


**Figure 6: Inference results for the first wave of COVID-19 in Seoul.** Panels (a), (b), and (c) depict the time evolution of prevalence, incidence, and daily recoveries, respectively, estimated using three inference methods: *marginal mean* (pink dotted line), *marginal median* (light blue dashed line), and *joint mode* (orange solid line). These estimates are compared to the 7-day moving average of observed data from the first wave of COVID-19 in Seoul (CS), represented by black empty circles. Panel (d) presents the inferred probability mass function derived from epidemic data, alongside the recorded contact data.
**Alt text**: Inference results for Seoul's first COVID-19 wave: (a–c) show estimated prevalence, incidence, and recoveries using three methods, compared to observed data. (d) Displays the inferred contact distribution alongside recorded contact data.

as a definitive ground truth for contact patterns.

Finally, to further demonstrate the potential of this framework, we performed parameter estimation, including the mean degree—using only partial data, specifically up to a cut-off time $T_c$ before the 84th day. In this case, we used the first 26 days of data to infer parameters and then forecasted the number of new infections for the following 10 days. As a result, the likelihood is now $\ell_1 + \ell_2 + \ell_4$, where $\ell_4$ accounts for the infection times of individual who did not recover before $T_c$. The 95% credible interval was obtained by sampling from the full posterior distribution, while point estimates were determined using the *joint mode*. For comparison, we performed the same analysis using the mass-action (MA) model. The results are shown in Fig. 7(b). The forecasted incidence was compared to the observed data using the mean squared error (MSE). The results indicate that the EBCM provides the most accurate forecast.

# 5 Discussion

Understanding the contact patterns of individuals during an epidemic remains a fundamental challenge in infectious disease modelling. Inferring the underlying network structure of an epidemic process is particularly difficult, even when the goal is not to reconstruct the entire network but rather to estimate key characteristics, such as the average degree or variance. A major obstacle in this inference process is the practical identifiability between connectivity and infectivity, a challenge previously highlighted in the literature [54, 62].

In this study, we introduced a framework for inferring network properties from epidemic data by integrating the Dynamic Survival Analysis (DSA) framework with the Edge-Based Compartmental Model (EBCM).

The EBCM provides a compact yet effective representation of the epidemic process, where the degree distribution is incorporated as a model parameter. By combining this with the flexibility of DSA, our approach enables the inference of crucial properties of the contact network that drive epidemic spreading.

We validated this framework using synthetic epidemic data generated via Gillespie simulations, considering networks with Poisson and Negative Binomial degree distributions. By applying the DSA-EBCM approach, we sampled from the joint posterior distribution of both disease and network parameters. Despite the inherent correlation between the infection rate, the recovery rate, and network properties such as the average degree, the posterior distributions consistently concentrated around the true parameter values. This allowed for an accurate reconstruction of both the epidemic dynamics and key network characteristics, despite relying solely on epidemic time-series data in which the network structure is only implicitly present.

Beyond synthetic data, we tested our methodology on two real-world outbreaks: the 2001 Foot-and-Mouth Disease (FMD) epidemic in the UK and the first wave of COVID-19 in Seoul, South Korea. In both cases, the framework successfully produced robust posterior distributions despite correlations between

| Parameter | *marginal mean* | *marginal median* | *joint mode* |
|---|---|---|---|
| $\beta$ | 0.042 | 0.035 | 0.052 |
| $\gamma$ | 0.192 | 0.192 | 0.193 |
| $\mu$ | 12.80 | 8.96 | 4.59 |
| Variance | 16.82 | 12.30 | 14.64 |
| $R_0$ | 2.35 | 1.43 | 1.44 |
| MSE Incidence | 1.14 | 0.48 | 0.41 |
| MSE Prevalence | 4.65 | 2.65 | 2.37 |
| MSE Daily Recovered | 1.07 | 0.65 | 0.6 |

**Table 3:** Estimated parameter values obtained using three inference methods: *marginal mean*, *marginal median*, and *joint mode*, for the 7-day moving average of the first wave of COVID-19 in Seoul (CS). The table also reports the Mean Squared Error (MSE) between the predicted epidemic trajectory and the observed data.
**Alt text**: Table presenting estimated parameter values for the first wave of COVID-19 in Seoul using three inference methods: *marginal mean*, *marginal median*, and *joint mode*. Parameters include $\beta$, $\gamma$, $\mu$, variance, and $R_0$. The table also reports the Mean Squared Error (MSE) for predicted incidence, prevalence, and daily recoveries compared to observed data.

| Parameter | *marginal mean* | *marginal median* | *joint mode* |
|---|---|---|---|
| $\beta$ | 0.044 | 0.044 | 0.043 |
| $\gamma$ | 0.196 | 0.196 | 0.193 |
| $Variance$ | 8.77 | 9.33 | 12.10 |
| $R_0$ | 1.42 | 1.44 | 1.47 |
| $MSE$ Incidence | 0.45 | 0.42 | 0.41 |
| $MSE$ Prevalence | 2.44 | 2.30 | 2.25 |
| $MSE$ New Recovered | 0.63 | 0.61 | 0.60 |

**Table 4:** Estimated parameter values obtained using three inference methods: *marginal mean*, *marginal median*, and *joint mode*, for the 7-day moving average of the first wave of COVID-19 in Seoul (CS) for a fixed average degree $\mu = 7.61$, obtained from the COVID-19 Survey data. The table also reports the Mean Squared Error (MSE) between the predicted epidemic trajectory and the observed data
**Alt text:** Table showing estimated values of $\beta$, $\gamma$, variance, and $R_0$ using three inference methods—*marginal mean*, *marginal median*, and *joint mode*—for the first wave of COVID-19 in Seoul with fixed average degree $\mu = 7.61$. Also includes Mean Squared Error (MSE) values for predicted incidence, prevalence, and new recoveries compared to observed data.

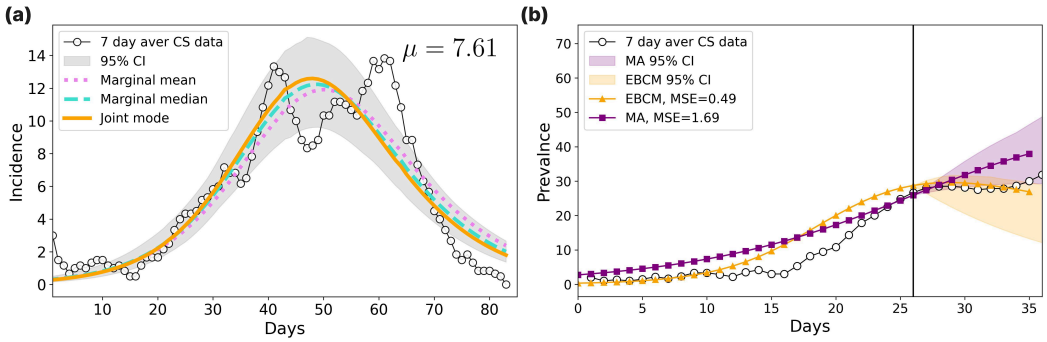


**Figure 7: Additional analysis of the first wave of COVID-19 in Seoul:** Panel (a) shows the evolution of the incidence obtained from the point estimates using *marginal mean*, *marginal median*, and *joint mode* for the case where the average degree is known and fixed at $\mu = 7.61$. Panel (b) shows the results obtained using partial data. Specifically, the first 26 days of data were used to forecast the following 10 days. In the figure, we show the prediction obtained using the EBCM in orange triangles, MA in square purple, and the 7 days moving average of the COVID-19 in Seoul (CS) in the black circles.
**Alt text**: COVID-19 incidence in Seoul: (a) model estimates using different inference methods with fixed average degree; (b) 10-day forecast using partial data, comparing EBCM, MA model, and observed data.

parameters. For the FMD dataset, we observed a multidimensional posterior distribution with long tails and a strong inverse correlation between $\beta$ and $\mu$. Nevertheless, as in the synthetic cases, a high-density region in parameter space provided the best description of the original data. Notably, we found that the *joint mode*, which represents the point of highest density in the sample of the posterior distribution, although more computationally demanding and less exact, yielded better point estimates compared to the *marginal median* or *marginal mean*, as it minimized the mean squared error.

For the Seoul COVID-19 dataset, we observed similar results, with the *joint mode* again emerging as the most accurate estimator of the epidemic process. Additionally, our inferred average degree closely matched independent contact data collected during the outbreak, reinforcing the validity of the approach. The framework also demonstrated its predictive capabilities by generating a short-term forecast with a 95% credible interval for the epidemic's progression over a 10-day period.

This study demonstrates that meaningful insights about underlying contact structures can be extracted solely from epidemic data without requiring explicit network observations. Future work could explore the extent to which this method can distinguish between homogeneous and heterogeneous network structures based only on outbreak dynamics. Furthermore, the approach could be extended to other spreading processes, such as information diffusion, where higher-order interactions may play a significant role. Comparing inferred network properties across different geographic regions—such as cities, counties, or states—could also provide insights into the diverse mechanisms that shape disease transmission.

Overall, our findings highlight the potential of integrating the EBCM and other network-based mean-field models with DSA to infer hidden contact structures from limited epidemic data. This approach provides a powerful tool for reconstructing essential network characteristics, improving epidemic forecasting, and enhancing our understanding of infectious disease spread in real-world settings.

## Acknowledgements

A.G. acknowledges the PhD studentship support from Northeastern University London. A.G. and I.Z.K. acknowledge useful discussion with Mauricio Santillana. B.C. and G.H.P. acknowledge the Basic Science Research Program through the NRF funded by the Ministry of Education (RS-202300245056). B.C acknowledges a grant of the project The Government-wide R&D to Advance Infectious Disease Prevention and Control (HG23C1629).

## Data availability statement

Code and synthetic datasets generated and analyzed during the current study are available from the corresponding author upon reasonable request.

## References


[1] István Z Kiss, Joel C Miller, Péter L Simon, et al. Mathematics of epidemics on networks. *Cham: Springer*, 598:31, 2017.

[2] Romualdo Pastor-Satorras, Claudio Castellano, Piet Van Mieghem, and Alessandro Vespignani. Epidemic processes in complex networks. *Reviews of modern physics*, 87(3):925, 2015.

[3] M. E. J. Newman. The structure and function of complex networks. 45(2):167–256.

[4] Matt J Keeling and Ken T.D Eames. Networks and epidemic models. 2(4):295–307.

[5] Leon Danon, Ashley P. Ford, Thomas House, Chris P. Jewell, Matt J. Keeling, Gareth O. Roberts, Joshua V. Ross, and Matthew C. Vernon. Networks and the epidemiology of infectious disease. 2011:1–28.

[6] Y. Moreno, R. Pastor-Satorras, and A. Vespignani. Epidemic outbreaks in complex heterogeneous networks. 26(4):521–529.

[7] Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. 10(4):1:1–1:26.

[8] Romualdo Pastor-Satorras and Alessandro Vespignani. Epidemic spreading in scale-free networks. *Physical review letters*, 86(14):3200, 2001.

[9] Robert M May and Alun L Lloyd. Infection dynamics on scale-free networks. *Physical Review E*, 64(6):066112, 2001.

[10] Romualdo Pastor-Satorras, Alessandro Vespignani, et al. Epidemics and immunization in scale-free networks. *Handbook of Graphs and Networks, Wiley-VCH, Berlin*, 2003.

[11] Clara Stegehuis, Remco Hofstad, and Johan Leeuwaarden. Epidemic spreading on complex networks with community structures. *Scientific Reports*, 6:29748, 07 2016.

[12] Chanchan Li, Guo ping Jiang, Yurong Song, Lingling Xia, Yinwei Li, and Bo Song. Modeling and analysis of epidemic spreading on community networks with heterogeneity. *Journal of Parallel and Distributed Computing*, 119:136–145, 2018.

[13] Matthieu Nadini, Kaiyuan Sun, Enrico Ubaldi, Michele Starnini, Alessandro Rizzo, and Nicola Perra. Epidemic spreading in modular time-varying networks. *Scientific reports*, 8(1):2352, 2018.

[14] Randi H Griffin and Charles L Nunn. Community structure and the spread of infectious disease in primate social networks. *Evolutionary Ecology*, 26:779–800, 2012.

[15] Víctor M. Eguíluz and Konstantin Klemm. Epidemic threshold in structured scale-free networks. *Phys. Rev. Lett.*, 89:108701, Aug 2002.

[16] Timo Smieszek, Lena Fiebig, and Roland W Scholz. Models of epidemics: when contact repetition and clustering should be included. *Theoretical biology and medical modelling*, 6:1–15, 2009.

[17] Joel C Miller. Percolation and epidemics in random clustered networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 80(2):020901, 2009.

[18] Emilie Coupechoux and Marc Lelarge. How clustering affects epidemics in random networks. *Advances in Applied Probability*, 46(4):985–1008, 2014.

[19] Marián Boguá, Romualdo Pastor-Satorras, and Alessandro Vespignani. Epidemic spreading in complex networks with degree correlations. *Statistical mechanics of complex networks*, pages 127–147, 2003.

[20] Yi Wang, Junling Ma, Jinde Cao, and Li Li. Edge-based epidemic spreading in degree-correlated complex networks. *Journal of theoretical biology*, 454:164–181, 2018.

[21] Marián Boguñá, Romualdo Pastor-Satorras, and Alessandro Vespignani. Absence of epidemic threshold in scale-free networks with degree correlations. *Phys. Rev. Lett.*, 90:028701, Jan 2003.

[22] Xuan-Hao Chen, Shi-Min Cai, Wei Wang, Ming Tang, and H Eugene Stanley. Predicting epidemic threshold of correlated networks: A comparison of methods. *Physica A: Statistical Mechanics and its Applications*, 505:500–511, 2018.

[23] Vito Latora, Vincenzo Nicosia, and Giovanni Russo. *Complex networks: principles, methods and applications*. Cambridge University Press, 2017.

[24] Federico Malizia, Luca Gallo, Mattia Frasca, Vito Latora, and Giovanni Russo. Individual-and pair-based models of epidemic spreading: Master equations and analysis of their forecasting capabilities. *Physical Review Research*, 4(2):023145, 2022.

[25] William Ogilvy Kermack and Anderson G McKendrick. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721, 1927.

[26] Erik Volz and Lauren Ancel Meyers. Epidemic thresholds in dynamic contact networks. *Journal of the Royal Society Interface*, 6(32):233–241, 2009.

[27] Frank Ball and Peter Neal. Network epidemic models with two levels of mixing. *Mathematical biosciences*, 212(1):69–87, 2008.

[28] Zhaoyang Zhang, Honggang Wang, Chonggang Wang, and Hua Fang. Modeling epidemics spreading on social contact networks. *IEEE transactions on emerging topics in computing*, 3(3):410–419, 2015.

[29] Denis Mollison. *Epidemic models: their structure and relation to data*. Number 5. Cambridge University Press, 1995.

[30] Gavin J Gibson, George Streftaris, and David Thong. Comparison and assessment of epidemic models. *Statistical Science*, 33(1):19–33, 2018.

[31] Jean Daunizeau, Rosalyn Moran, Jérémie Mattout, and Karl Friston. On the reliability of model-based predictions in the context of the current covid epidemic event: impact of outbreak peak phase and data paucity. *MedRxiv*, pages 2020–04, 2020.

[32] Matt J Keeling and Ken TD Eames. Networks and epidemic models. *Journal of the royal society interface*, 2(4):295–307, 2005.

[33] Joel C Miller, Anja C Slim, and Erik M Volz. Edge-based compartmental modelling for infectious disease spread. *Journal of the Royal Society Interface*, 9(70):890–906, 2012.

[34] Joel C Miller and Erik M Volz. Model hierarchies in edge-based compartmental modeling for infectious disease spread. *Journal of mathematical biology*, 67:869–899, 2013.

[35] Erik M Volz, Joel C Miller, Alison Galvani, and Lauren Ancel Meyers. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. *PLoS computational biology*, 7(6):e1002042, 2011.

[36] Thomas R Eng, William T Butler, et al. The hidden epidemic: confronting sexually transmitted diseases. 1997.

[37] Ken TD Eames and Matt J Keeling. Contact tracing and disease control. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, 270(1533):2565–2571, 2003.

[38] RB Rothenberg, PD McElroy, MA Wilce, and SQ Muth. Contact tracing: comparing the approaches for sexually transmitted diseases and tuberculosis. *The International Journal of Tuberculosis and Lung Disease*, 7(12):S342–S348, 2003.

[39] Geoffrey P Garnett and Roy M Anderson. Contact tracing and the estimation of sexual mixing patterns: the epidemiology of gonococcal infections. *Sexually transmitted diseases*, pages 181–191, 1993.

[40] Caroline Buckee. Improving epidemic surveillance and response: big data is dead, long live big data. *The Lancet Digital Health*, 2(5):e218–e220, 2020.

[41] Fatma Zubaydi, Assim Sagahyroon, Fadi Aloul, Hasan Mir, and Bassam Mahboub. Using mobiles to monitor respiratory diseases. In *Informatics*, volume 7, page 56. MDPI, 2020.

[42] Kin On Kwok, Arthur Tang, Vivian WI Wei, Woo Hyun Park, Eng Kiong Yeoh, and Steven Riley. Epidemic models of contact tracing: systematic review of transmission studies of severe acute respiratory syndrome and middle east respiratory syndrome. *Computational and structural biotechnology journal*, 17:186–194, 2019.

[43] O Arda Vanli and Davison Elijah Tsekeni. Inference of human contact networks based on epidemic data. *IISE Transactions on Healthcare Systems Engineering*, pages 1–17, 2024.

[44] Cécile Viboud, Kaiyuan Sun, Robert Gaffey, Marco Ajelli, Laura Fumanelli, Stefano Merler, Qian Zhang, Gerardo Chowell, Lone Simonsen, Alessandro Vespignani, et al. The rapidd ebola forecasting challenge: Synthesis and lessons learnt. *Epidemics*, 22:13–21, 2018.

[45] Anne Cori and Adam Kucharski. Inference of epidemic dynamics in the covid-19 era and beyond. *Epidemics*, page 100784, 2024.

[46] Vaiva Vasiliauskaite, Nino Antulov-Fantulin, and Dirk Helbing. On some fundamental challenges in monitoring epidemics. *Philosophical Transactions of the Royal Society A*, 380(2214):20210117, 2022.

[47] Ben Swallow, Paul Birrell, Joshua Blake, Mark Burgman, Peter Challenor, Luc E Coffeng, Philip Dawid, Daniela De Angelis, Michael Goldstein, Victoria Hemming, et al. Challenges in estimation, uncertainty quantification and elicitation for pandemic modelling. *Epidemics*, 38:100547, 2022.

[48] Srinivas Gorur Shandilya and Marc Timme. Inferring network topology from complex dynamics. *New Journal of Physics*, 13(1):013004, jan 2011.

[49] Mor Nitzan, Jose Casadiego, and Marc Timme. Revealing physical interaction networks from statistics of collective dynamics. 3(2):e1600396.

[50] Praneeth Netrapalli and Sujay Sanghavi. Learning the graph of epidemic cascades. *SIGMETRICS Perform. Eval. Rev.*, 40(1):211–222, June 2012.

[51] Federico Malizia, Alessandra Corso, Lucia Valentina Gambuzza, Giovanni Russo, Vito Latora, and Mattia Frasca. Reconstructing higher-order interactions in coupled dynamical systems. *Nature Communications*, 15(1):5184, 2024.

[52] Manuel Gomez Rodriguez, Jure Leskovec, David Balduzzi, and Bernhard Schölkopf. Uncovering the structure and temporal dynamics of information propagation. *Network Science*, 2(1):26–65, April 2014.

[53] Chris Groendyke, David Welch, and David R. Hunter. Bayesian Inference for Contact Networks Given Epidemic Data. *Scandinavian Journal of Statistics*, 38(3):600–616, 2011. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1467-9469.2010.00721.x.

[54] Tom Britton and Philip D. O'neill. Bayesian Inference for Stochastic Epidemics in Populations with Random Social Structure. *Scandinavian Journal of Statistics*, 29(3):375–390, 2002. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/1467-9469.00296.

[55] Fraser Lewis, Gareth J. Hughes, Andrew Rambaut, Anton Pozniak, and Andrew J. Leigh Brown. Episodic Sexual Transmission of HIV Revealed by Molecular Phylodynamics. *PLOS Medicine*, 5(3):e50, March 2008.

[56] 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, November 2005.

[57] Bastian Prasse and Piet Van Mieghem. Exact Network Reconstruction from Complete SIS Nodal State Infection Information Seems Infeasible. *IEEE Transactions on Network Science and Engineering*, 6(4):748–759, October 2019.

[58] Luca Gallo, Mattia Frasca, Vito Latora, and Giovanni Russo. Lack of practical identifiability may hamper reliable predictions in covid-19 epidemic models. *Science advances*, 8(3):eabg5234, 2022.

[59] Wasiur R. KhudaBukhsh, Boseung Choi, Eben Kenah, and Grzegorz A. Rempała. Survival dynamical systems: individual-level survival analysis from population-level epidemic models. 10(1):20190048.

[60] Francesco Di Lauro, Wasiur R. KhudaBukhsh, István Z. Kiss, Eben Kenah, Max Jensen, and Grzegorz A. Rempała. Dynamic survival analysis for non-markovian epidemic models. 19(191):20220124.

[61] Grzegorz A. Rempała and Wasiur R. KhudaBukhsh. Dynamical survival analysis for epidemic modeling. In Bharath Sriraman, editor, *Handbook of Visual, Experimental and Computational Mathematics : Bridges through Data*, pages 1–17. Springer International Publishing.

[62] Istvan Z. Kiss, Luc Berthouze, and Wasiur R. KhudaBukhsh. Towards Inferring Network Properties from Epidemic Data. *Bulletin of Mathematical Biology*, 86(1):6, December 2023.

[63] István Z. Kiss and Péter L. Simon. On Parameter Identifiability in Network-Based Epidemic Models. *Bulletin of Mathematical Biology*, 85(3):18, 2023.

[64] Michael Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2-3):161–180, 1995.

[65] Mark EJ Newman, Steven H Strogatz, and Duncan J Watts. Random graphs with arbitrary degree distributions and their applications. *Physical review E*, 64(2):026118, 2001.

[66] Joel C. Miller, Anja C. Slim, and Erik M. Volz. Edge-based compartmental modelling for infectious disease spread. *Journal of The Royal Society Interface*, 9(70):890–906, may 2012.

[67] Wasiur R. KhudaBukhsh, Boseung Choi, Eben Kenah, and Grzegorz A. Rempała. Survival dynamical systems: individual-level survival analysis from population-level epidemic models. *Interface Focus*, 10(1):20190048, 2020.

[68] Francesco Di Lauro, Wasiur R. KhudaBukhsh, István Z. Kiss, Eben Kenah, Max Jensen, and Grzegorz A. Rempała. Dynamic survival analysis for non-markovian epidemic models. *Journal of The Royal Society Interface*, 19(191):20220124, 2022.

[69] Harley Vossler, Pierre Akilimali, Yuhan Pan, Wasiur R. KhudaBukhsh, Eben Kenah, and Grzegorz A. Rempała. Analysis of individual-level data from 2018–2020 ebola outbreak in democratic republic of the congo. *Scientific Reports*, 12(1):5534, 2022.

[70] Wasiur R. KhudaBukhsh, Caleb Deen Bastian, Matthew Wascher, Colin Klaus, Saumya Yashmohini Sahai, Mark H. Weir, Eben Kenah, Elisabeth Root, Joseph H. Tien, and Grzegorz A. Rempała. Projecting COVID-19 cases and hospital burden in Ohio. *Journal of Theoretical Biology*, 561:111404, 2023.

[71] Matti Vihola. Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22, 11 2010.

[72] Daniel T. Gillespie. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of Computational Physics*, 22(4):403–434, 1976.

[73] Gareth Davies. The foot and mouth disease (fmd) epidemic in the united kingdom 2001. *Comparative immunology, microbiology and infectious diseases*, 25(5-6):331–343, 2002.

[74] Jin Ho Ha, Ji Yeon Lee, So Young Choi, and Sook Kyung Park. Covid-19 waves and their characteristics in the seoul metropolitan area (jan 20, 2020–aug 31, 2022). 2023.