scientific reports

OPEN

Network inference from population-level observation of epidemics

F. Di Lauro[1], J.-C. Croix[1], M. Dashti[1], L. Berthouze[2] & I. Z. Kiss[1]✉

Using the continuous-time susceptible-infected-susceptible (SIS) model on networks, we investigate the problem of inferring the class of the underlying network when epidemic data is only available at population-level (i.e., the number of infected individuals at a finite set of discrete times of a single realisation of the epidemic), the only information likely to be available in real world settings. To tackle this, epidemics on networks are approximated by a Birth-and-Death process which keeps track of the number of infected nodes at population level. The rates of this surrogate model encode both the structure of the underlying network and disease dynamics. We use extensive simulations over Regular, Erdős–Rényi and Barabási–Albert networks to build network class-specific priors for these rates. We then use Bayesian model selection to recover the most likely underlying network class, based only on a single realisation of the epidemic. We show that the proposed methodology yields good results on both synthetic and real-world networks.

Networks are an important tool for modelling systems with many interacting parts such as epidemics spreading within a population or neuronal activity in the brain. Indeed, the intricate interplay of many individual well-defined units can be captured by the links of a network, and this can be done with an unprecedented level of detail[5,15,18,29,35]. For instance, directed, weighted or temporal links can all be considered within this modelling paradigm. A main feature of network epidemic models is that the topology of the contact structure is treated separately from the characteristics of the pathogen (such as infectivity and typical recovery time), in contrast to mass action models such as Kermack–McKendrick[17]. The transmission dynamics of epidemic spreading on networks can be modelled as a continuous time Markov Chain process on a network[33]. Unfortunately, the literature only show few exact results, and these are mainly related to specific cases or small networks. Therefore, approximations are often introduced to simplify the exact model and derive quantitative results. Most notably, well-known and widely used theoretical approaches include mean-field and higher order approximations[12,18], edge-based dynamics[24,42], percolation[10,23,25] and generating functions[29,33].

These approaches have led to the realisation that the structure of the network has a profound impact on how diseases invade, spread and how to best control them. This impact is particularly well understood for degree heterogeneity and assortativity/disassortativity, and to a lesser extent, for clustering, the propensity of nodes that share a common neighbour to be connected[18,33].

However, depending on the field of application, the precision to which the underlying network is known can vary greatly, from absolute (when full description is available) to absent (when a description is entirely lacking). For example, whereas some technological networks can be mapped out to a great degree of detail, social networks can be challenging to query[2]. This has resulted in a significant amount of research aimed to develop methods for link prediction (for a survey, see[2]). Instead of assuming the availability of explicit information about nodes and edges, these methods rely on 'observables' from dynamical processes taking place on the network, under the assumption that these provide latent information about the missing underlying network structure. In the framework of epidemics on networks this suggests that it is possible to get insights about the structure of the network by observing quantities of interest at node and perhaps population level. Indeed, the inverse problem of inferring networks from epidemic data has been the subject of great scrutiny.

In particular, in the context of statistical inference, this task has been approached by either formulating it as a likelihood optimisation problem[6,11,12,26,28] or using Bayesian inference[1,7,13,31,41]. Compared to maximum likelihood optimisation methods (e.g. independent cascade model[12]), the Bayesian inference is usually based on a smaller number of observations of the epidemic[1,7,13]. However, both network inference approaches (explicit link inference and inferring parameters of a known network model) lead to good estimates for the network and parameters of

[1]Department of Mathematics, University of Sussex, Falmer, Brighton BN1 9QH, UK. [2]Department of Informatics, University of Sussex, Falmer BN1 9QH, UK. ✉email: i.z.kiss@sussex.ac.uk

the epidemic dynamics. Moreover, there is an interesting tradeoff between them. The former is able to identify the adjacency matrix, but requires the observation of a large number of cascades, whereas the latter can only infer some structural parameters (such as the probability of a link between two nodes), but relies on fewer observations. Recently, it has been conjectured[36] that an exact (link-by-link) reconstruction of networks might not be feasible due to requiring a subexponentially increasing number of observations and an exponentially increasing computation time with respect to the number of nodes in the network.

A common feature of the above mentioned work is their reliance on the availability of detailed data at node level, such as the complete temporal knowledge of all cascade trees in[12] or the observation of all the removal/infection times in the Bayesian framework of[1]. However, in most real-world scenarios, such detailed information is unlikely to be available. A more reasonable expectation is to be provided with population-level observations, that is, the number of infected nodes in the whole network at various times. Our aim in this paper is to establish the feasibility of inferring the class of the underlying network from population-level observations. Whilst a very recent paper[21] provides an algorithm to infer properties of a given network-type from prevalence data, we are not aware (for a survey, see[5]) of any research that specifically addresses the problem of network class inference based purely on population-level observations. We do so within the framework of continuous-time SIS epidemics on networks when only population-level data from a single realisation of the epidemic are available.

We treat this problem as an inverse problem and adopt a Bayesian approach which involves the following steps:

(a) propose a parametric forward model that reproduces network/population-level dynamics and reflects network structure;
(b) build a prior distribution for these model parameters on a network class basis;
(c) use the posterior measure to identify the most likely network class.

A complete description of the SIS dynamics on a network with N nodes requires to solve 2^N equations, one per possible state. The distribution of population-level statistics in time can be described via the count of the number of infected nodes in this dynamics; however, this process scales exponentially with the size of the network. Here, we take a different route and choose to use a surrogate model to represent the evolution of the count of infected nodes in the population. A reasonable candidate for this is a Birth-and-Death process (BD), see[27], characterised by only $(N + 1)$ equations and $2(N + 1)$ free parameters, the rates of infection and recovery, that need to be tuned to best represent the exact model. Whilst the rates of recovery are network independent and known exactly, the rates of infection in the surrogate model are more challenging to define.

In this work, the rates of infection in the surrogate model are provided by a simple parametric model, together with an estimation procedure based on extensive and detailed simulations of epidemics on three classes of well-known random networks: Regular, Erdős–Rényi and Barabási–Albert. This procedure leads to distinct rate models for the three classes of networks. These observations are encapsulated in a prior distribution for the rates of the BD process.

Finally, when one observes a single epidemic through population-level data, our prior and forward model can be used within a Bayesian model selection framework to identify the most likely underlying network class. It is worth noting that this framework is versatile enough to be used in conjunction with any set of population-level epidemic data, as it will still output the most likely network class, that is, the closest class (in terms of heterogeneity of the degree distribution) to that of the true underlying network.

The paper is structured as follows. In Sect. "The forward model" we describe the BD surrogate forward model together with a three-parameter model for its rates of infection. Section "Bayesian inference of network class from single epidemics" includes all aspects of the Bayesian approach we used, from building priors to model selection and model validation/stress testing. We conclude with a discussion and further research directions in Sect. "Discussion".

The forward model

A population of N individuals is considered with the contact structure between individuals described by an undirected network with adjacency matrix $G = (g_{ij})_{i,j=1,2,...,N}$ where $g_{ij} = 1$ if nodes i and j are connected and zero otherwise. Self-loops are excluded, so $g_{ii} = 0$ and $g_{ij} = g_{ji}$ for all $i, j = 1, 2, ... N$. The standard SIS epidemic dynamics on a network[18] is considered, which is driven by two type of events: (a) infection and (b) recovery from infection. Infection can spread from an infected and infectious node (I) to any of its susceptible neighbours (S) and this is modelled as a Poisson point process with per-link infection rate τ. Infectious nodes recover at constant rate γ, independently of their neighbours and become susceptible again. Initialization is made by randomly choosing I_0 nodes to be infected at the initial time, the others being susceptible. The resulting model is a continuous-time Markov Chain, and to fully specify its state we need an equation for each arrangement of length N with entries being either S or I, resulting in a state space of 2^N elements. While this is easy to formalise and write down theoretically, the numerical integration of the system becomes intractable even for modest values of N[5,18,39,40]. This motivates us to use a surrogate model, offering sufficient flexibility to approximate the time evolution of the number of infectious nodes in the network.

Birth-and-death approximation of SIS epidemics. We use a BD process, a continuous-time Markov chain with state space $\{0, ..., N\}$ and transitions of unit size, as the surrogate model. The up-jumps or infections are described by rates a_k, that is, the rates of infection in the presence of k infected nodes and encode the network structure. The down-jumps or recoveries are described by rates $c_k = \gamma k$. To understand why, we first observe that recoveries are independent events (since once a node is infected, its status no longer depends on other

nodes). An infected node recovers after a time drawn from an exponential distribution with rate γ. If k nodes are infected, the first recovery is going to happen according to the minimum of all recovery times, which is again exponentially distributed with rate γk. Hence, the transition probabilities of the surrogate process are given by the following forward Kolmogorov (or Master) equation:

$$\forall k \in \{0, \ldots, N\}, \; \dot{p}_{k_0,k}(t) = a_{k-1}p_{k_0,k-1}(t) - (a_k + c_k)p_{k_0,k}(t) + c_{k+1}p_{k_0,k+1}(t), \tag{1}$$

together with $a_{-1} = c_{N+1} = 0$ and an initial condition $k_0 \in \{0, \ldots, N\}$, with $p_{k_0,k}(t)$ the probability of being in the state with k infected at time t, given initial k_0 infected. The solutions of Eq. (1) and the rates of infection will be denoted by $p_{k_0,k}^\alpha$ and a_k^α, respectively, when the dependence on additional parameters α needs to be enforced.

The quality of the surrogate model, i.e., how well it approximates the exact model, depends strongly on the choice of infection rates a_k. The way a_k depends on k is determined by the underlying network structure. An analytic formula for a_k is only available for the fully connected network, namely: $a_k = \tau k(N - k)$, that is the number of S–I links (i.e., links connecting susceptible and infected nodes) in the network multiplied by the per-contact rate of infection τ.

In fact, in a stochastic simulation of the epidemic on a network, the rate of going from k to $k + 1$ infected nodes is exactly $\tau \times$ #S-I links. Hence, during a simulation it makes sense to keep track of the number of infected nodes, the number of S–I links and the time spent in each respective state. Further important observations can be made. The number of S–I links is a random variable and given a fixed number of infected nodes, say k, the number of S–I links can take different values. This is simply due to the stochasticity in how the infected nodes are laid out in the network. Thus a plausible choice for the rate a_k may be simply the average of the number of S–I links when there are exactly k infected nodes. However, some states are longer lived than others and this needs to be accounted for. Combining all the above, an empirical average rate of infection emerges, that is

$$\hat{a}_k = \tau \frac{\sum_i i t_{i,k}}{\sum_i t_{i,k}}, \; 1 \le k \le N, \tag{2}$$

where $t_{i,k}$ is the lifetime of a state with k infected nodes and i S–I links. We will use the notation $\hat{a}_k^{\theta,\tau,\gamma}$ to indicate the resulting estimate given the network class $\theta \in \Theta$:

$$\Theta := \{ \text{Reg}, \; \text{E-R}, \; \text{B-A} \}.$$

where we use Regular (Reg), Erdős–Rényi (E–R) and Barábasi–Albert (B–A) network classes. There are a number of reasons for this choice. First, these three classes are perhaps the most popular random network models, so they provide a good benchmark to test our model. Second, they can produce rich topologies in terms of degree heterogeneity, and therefore allow us to test the flexibility of our framework. Finally, the absence of higher-order structures (such as communities, or clustering) enables us to simplify the problem of fitting the (k, a_k) curves and thus to focus more specifically on the problem of network inference.

Hence, we can calibrate the infection rates a_k through a statistical analysis based on stochastic simulations of the SIS epidemics on networks. Namely, for a network class (with given average degree) and given disease parameters (τ, γ), we run 50 outbreaks on 50 different realisations of the network. We keep track of the states that the process visits along with the number of infected nodes, number of S–I links and lifetime of the states. This data feeds into Eq. (2) and leads to the value of $\hat{a}_k$ for all $0 \le k \le N$. To cover the entire range, $0 \le k \le N$, half of the outbreaks are started from $k_0 = 5$ infected nodes, chosen uniformly at random, and the other from $k_0 = N$ infected nodes. The former allows us to explore the curve up to the steady state, while the latter, although an artificial scenario, allows us to explore the curve from the steady state to N. Typical $(k, \hat{a}_k^{\theta,\tau,\gamma})$ curves are shown in Fig. 1. In what follows we assume that these rates are 'optimal' and that they lead to a surrogate model that agrees well with the exact one. This choice is motivated by the heuristics presented above which is further validated through extensive numerical simulations for three network classes and a large set of disease parameter values (see Sect. "Bayesian inference of network class from single epidemics").

Three-parameter model of infection rates. Consistent with results in[27], we notice that, although estimated $\hat{a}_k$ curves are distinct for different network classes, they all share some common features: specifically, they all satisfy $\hat{a}_0 = \hat{a}_N = 0$ and exhibit a single maximum in $[0, N]$. Perhaps the most important features that change between the three distinct network classes are the flatness and skewness of the $\hat{a}_k$ curves (see Fig. 1). It is clear that high heterogeneity in the degree distribution (i.e. Reg → E–R → B–A, displaying respectively no → medium → high heterogenity) increases the left skew.

The intuitive reason for these differences in the $(k, \hat{a}_k)$ curves is that epidemics on such different networks spread with distinct enough characteristics. In scale-free networks for example, the most exposed nodes are the hubs, so they get infected early on. This skews the $(k, \hat{a}_k)$ curve to the left, because once infected these hubs generate a disproportionately large number of S–I links. On the contrary, when all nodes have similar degrees, the $(k, \hat{a}_k)$ curves are more symmetric. Concerning E–R and Reg networks, the most important difference is that the former allows for some degree heterogeneity, whereas the latter does not. Degree heterogeneity plays an important role when it comes to disease transmission so it is no surprise that epidemics on E–R networks can affect a higher proportion of nodes in the initial stage of an outbreak when compared to epidemics on Reg networks[18].

This suggests that $\hat{a}_k$ curves could be parametrised with a low dimensional model. The departure from the fundamental assumption of homogeneous random mixing in epidemiological and ecological models has led to a myriad of models where bi-linear transmission terms proportional to $\sim I \times S$ or $\sim I \times (N - I)$ have been replaced by non-linear infection terms such as $I^p S^q$[14,20,37]. In particular it is noted that, in the context of classical compartmental and mean-field models, such terms can be inferred from the number of S–I links taken from
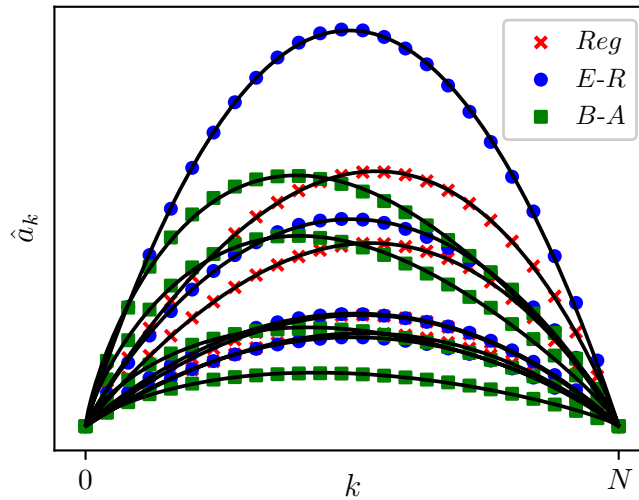
Figure 1. $\hat{a}_k$ curves (markers) along with the best fits from the C, a, p model (plain lines) on 12 different combinations of network classes and epidemic parameters. Parameters of the simulations considered are, from top to bottom: Reg (crosses), $(\langle k \rangle, \tau, \gamma) = \{(12, 1.43, 5.69)(6, 9.46, 4.23)(8, 4.47, 8.62)(13, 1.56, 9.18)\}$; E–R (circles), $(\langle k \rangle, \tau, \gamma) = \{(7, 5.96, 8, 07), (13, 5.8, 9.06), (6, 3.08, 7.61), (16, 0.99, 8.5)\}$; and B–A (squares), $(\langle k \rangle, \tau, \gamma) = \{(6, 3.09, 7.61)(8, 5.99, 7.01)(12, 0.79, 8.96)(16, 2.18, 5.81)\}$.

simulation and that they can lead to more exotic model behaviours. In the same spirit, we put forward the following model for the rates:

$$\forall k \in \{0, \dots, N\}, \; a_k := a_k^{(C,a,p)} = Ck^p(N-k)^p \left(a\left(k - \frac{N}{2}\right) + N \right), \tag{3}$$

where the three parameters C, a and p offer flexibility to adapt to various networks and epidemics of different severity. This choice is motivated by the heuristic thinking of how the epidemic unfolds on the network. The parameter $C > 0$ gives a general scaling, dealing with different infection intensities, $a \in [-2, 2]$ helps to shift the peak from the centre (e.g. $a < 0$ shifts the peak to the left), and $p > 0$ allows for different flatnesses (smaller p values leading to flatter curves). Note that, when $a = 0$, the model results in a particular case of the previously mentioned non-linear models. Immediately, one can note that this model fulfils a number of desirable properties: (a) it is low dimensional/parsimonious, (b) the model satisfies $a_0 = a_N = 0$ by construction, (c) it includes the complete network when $a_k = \tau k(N - k)$ and finally, (d) it has a single maximum within $[0, N]$.

The C, a, p values are obtained using a non-linear least-square fit (using a particle swarm algorithm[16]):

$$e(C, a, p; \mathcal{S}) = \sum_{k, \; \sum_i t_{i,k} > 0} \left(a_k^{(C,a,p)} - \hat{a}_k \right)^2. \tag{4}$$

Figure 1 showcases the flexibility of the model in fitting $\hat{a}_k$ curves coming from different network classes and confirms our observations about the rates being more left-skewed with increasing heterogeneity in node degree.

In the same figure, curves based on the (C, a, p) model are compared to the $(k, \hat{a}_k)$ curves. Systematic numerical investigations (not all plots shown) demonstrate that the proposed parsimonious three-parameter model fits the $(k, \hat{a}_k)$ curves well for all considered network classes, particularly Reg and E–R. For B–A networks small discrepancies between the $(k, \hat{a}_k)$ curves and the (C, a, p) model are possible.

Dataset. Proving that the behaviour of the exact system of 2^N equation is well approximated by our proposed system of $(N + 1)$ ordinary differential equations (1) is still an open question. Therefore, the validations that we provide in this paper are entirely based on extensive numerical simulations. Here, we discuss briefly the synthetic dataset $\mathcal{S}$ underpinning those numerical validations. For each network class, we varied the average degree ($5 \le \langle k \rangle < 20$). This covers a large number of scenarios and the networks remain relatively sparse. Regarding the epidemic parameters, we varied the infection and recovery rates ($(\tau, \gamma) \in (0, 10] \times (0, 10]$). Values for the rates were chosen via Latin hypercube sampling[22]. By doing so, we could observe many unique scenarios, providing a solid base upon which to test our methods.

However, there may be situations where the epidemic does not spread. Indeed, the behaviour of an epidemic is determined by the characteristics of both the network class and epidemic dynamics. The former includes quantities such as the average degree and higher-order moments, the latter includes per-link infection and recovery rates. All of this is captured by the reproduction number[18], R_0, which is the number of secondary infections caused by a typical infectious individual introduced into a fully susceptible population:
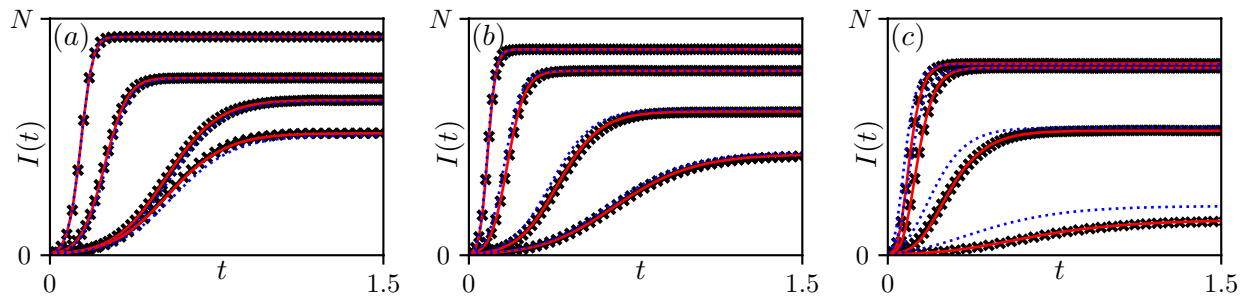
Figure 2. Average number of infected nodes from simulations (markers) and the numerical solutions of system (1) with rates a_k given either by the raw data $\hat{a}_k$ (continuous curve) or by the (C, a, p) model (dotted curves), with initial condition $k_0 = 5$. Three network classes are reported, each with $N = 1000$ nodes, from left to right, ordered by increasing heterogeneity, from Reg (**a**) and E–R (**b**) to B–A (**c**) networks. Networks and epidemic parameters are the same as in Fig. 1.

$$R_0 = \frac{\tau}{\gamma + \tau} \frac{\langle k^2 - k \rangle}{\langle k \rangle}. \tag{5}$$

If $R_0 \leq 1$ the infection will die out. However, if $R_0 > 1$, then an outbreak is expected. Since R_0 depends directly on the sampled network and disease parameters, we accepted only situations where $1 < R_0 \leq 10$. This led to 360 valid choices (network class, $\langle k \rangle, \tau, \gamma$), with 120 per network class. For all the 360 scenarios, data from the simulations were used to determine network class- and disease-parameter specific infection rates $\hat{a}_k^{\theta,\tau,\gamma}$ and the corresponding (C, a, p) models.

Numerical validation of the forward model. To validate our claim that the BD process is a good approximation of the true epidemic behaviour, we numerically integrated the master Eq. (1) with rates $a_k = \hat{a}_k$ and $c_k = \gamma k$, where $\hat{a}_k$ are the estimated rates via Eq. (2), for all 360 scenarios in $\mathcal{S}$. The master equation was also numerically integrated with rates given by the (C, a, p) model. The expected number of infected nodes from the numerical solution of both master equations was then compared to the average number of infected nodes based on simulations. Four representative examples of epidemics for each network class are shown in Fig. 2. For the vast majority of the tested cases (not all shown), the agreement between simulation and the (C, a, p) model is good to excellent. It is worth noting that, in the case of B–A networks there are a few parameter combinations where the agreement between the master equation with the rates given by the (C, a, p) model and simulation results is poorer, see Fig. 2c. This is despite the seemingly small discrepancy between $(k, \hat{a}_k)$ curve and the corresponding (C, a, p) model (not shown). However, the master equation with the $\hat{a}_k$-rates still leads to good agreement with simulations as shown in Fig. 2c (markers versus continuous line). Even so, it is reassuring to see that even when the agreement between the master equation with the (C, a, p) model breaks down, the agreement with the $\hat{a}_k$ holds. In[27], a similar surrogate model was used and the authors obtained good agreement between the BD model and simulations for an even wider range of network classes. This gives us confidence that the surrogate model is a viable model.

Bayesian inference of network class from single epidemics

In the framework presented so far, we proposed a surrogate model which approximates the evolution of the total number of infected nodes in a SIS epidemic on a network. The rates of infection in this forward model (i.e. $\hat{a}_k$) are parametrised by a three-parameter model (C, a, p) as detailed in Eq. (3). Early investigation shows that the $(k, \hat{a}_k)$ curves (thus the associated C, a, p triple) are distinct across the three different network classes that we considered. Hence, one may expect that discrete observations taken from a single epidemic spreading on an unknown network carry sufficient information to identify its most likely class.

To be more precise, we consider a population-level dataset $y = (k_1, \ldots, k_n)$ where $k_j \in \{0, \ldots, N\}$ for any $j = 1, \ldots, n$ is the number of infected nodes in the network at time $t_j \in [0, T]$, and we define the vector $s = (t_1, \ldots, t_n)$. Our objective is to predict the class $\theta \in \Theta$ of the underlying network from y. Figure 3 illustrates 10 distinct data sets for each of the three network classes. These data are obtained directly from Gillespie simulations[8,9] of the SIS epidemic on the respective networks. Observations are taken at regular times from the start of the epidemic to the point where the quasi steady-state is approached.

For each value of θ (that is a network class), we build a distribution $\pi_{0,\theta}$ over the parameters C, a, p based on offline simulations of SIS epidemics for a range of networks in each given class θ (see Sect. "Prior distributions for each network class"). By looking at the outcomes of our simulations, we observe that, for our chosen set of candidate classes Θ, the distributions $\pi_{0,\theta}(C, a, p), \theta \in \Theta$, cluster in distinct regions of the (C, a, p) parameter space. This is necessary for the inference to work, and it contributes to the validation of our model of choice for the rates $\hat{a}_k$. Assuming a non-informative uniform prior for θ, we derive a prior distribution $\pi_0(C, a, p, \theta)$ in the form of a mixture:

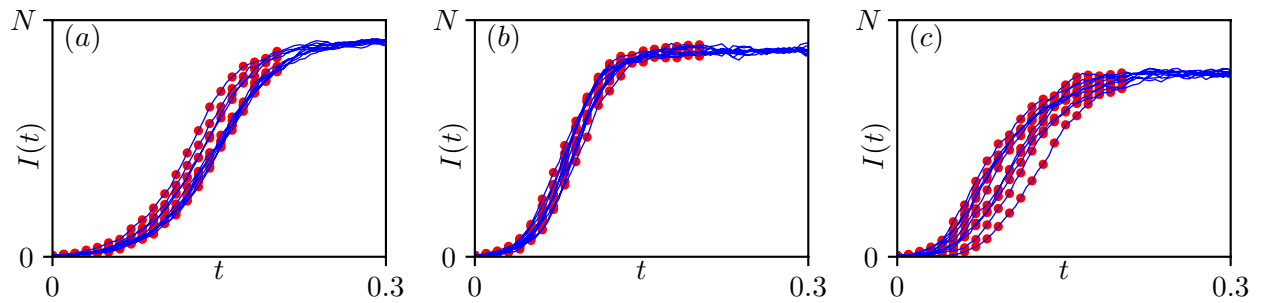$$\pi_0(C, a, p, \theta) = \frac{1}{3}\pi_{0,\theta}(C, a, p).$$

Figure 3. 3 examples on different network classes of 10 average epidemic paths, taken from dataset $\mathcal{S}$. Continuous curves represent the evolution of infectious counts and dots the observations y. Network and epidemic parameters for each panel are, from left to right, Regular (**a**) with $\langle k \rangle = 17, \tau = 2.62, \gamma = 4.03$, Erdős–Rényi (**b**) with $\langle k \rangle = 13, \tau = 5.80, \gamma = 9.06$, and Barabási–Albert (**c**) with $\langle k \rangle = 6, \tau = 8.16, \gamma = 8.23$. In each realisation, $k_0 = 5$ randomly selected nodes are infected at the beginning of the epidemic.

Our objective is the prediction of the underlying network θ given the data (y, s), which will be done using the posterior distribution $\pi(\theta|y, s)$ obtained by Bayes' rule:

$$\pi(\theta|y, s) = \int \pi(C, a, p, \theta|y, s) dC da dp$$

$$\propto \int \mathcal{L}^{C,a,p}(y, s) \pi_0(C, a, p, \theta) dC da dp, \tag{6}$$

$$\propto \int \mathcal{L}^{C,a,p}(y, s) \pi_{0,\theta}(C, a, p) dC da dp,$$

where, given C, a, p, the likelihood $\mathcal{L}^{C,a,p}$ can be expressed in terms of the solution operator of the forward model discussed above (see Sect. "The forward model"). This Bayesian classification methodology is also known as model selection, where the model is a particular class of networks. Once we have computed the posterior distribution $\pi(\theta|y, s)$ (see Sect. "Numerical method for posterior marginals computations"), we simply pick the most likely underlying network class (Maximum a posteriori estimator for θ given the data (y, s)).

Prior distributions for each network class. In this work, we consider prior distributions for each network class as a different density $\pi_{0,\theta}$ over the C, a, p space. To do this, we use the very same dataset that was used for numerical validation (see Sect. "The forward model"). Given the C, a, p values of each network class (see Fig. 4), we choose 100 triples to estimate a probability distribution and leave 20 for testing. The (C, a, p) values associated with the training scenarios are used to infer three Gaussian kernel density estimators[34] to be used as prior distributions. The bandwidth of these estimators is set by 10-fold cross-validation.

Numerical method for posterior marginals computations. Finally, to predict the underlying network class given a dataset (y, s), we need to compute the three marginals given in Eq. (6) (one per network class) and this is done by Monte-Carlo estimation. As already mentioned, the likelihood $\mathcal{L}^{C,a,p}(y, s)$ can be obtained using the forward operator associated with Eq. (1). Indeed, given a (C, a, p)-triple, the likelihood of a dataset (y, s) is given by:

$$\mathcal{L}^{C,a,p}(y, s) = \prod_{i=1}^{n-1} p_{k_i, k_{i+1}}^{C,a,p}(t_{i+1} - t_i),$$

using the fact that the BD process is time homogeneous. We choose to compute each term $p_{k_i, k_{i+1}}^{C,a,p}(t_{i+1} - t_i)$ where $1 \leq i \leq n - 1$ using the algorithm introduced in[3], allowing BD transition probabilities to be computed individually. This represents a significant reduction in computational time, when compared to matrix exponential since we are working with a network of size $N = 1000$ nodes.

Once we have an efficient numerical method to compute the likelihood, we use the corrected Arithmetic Mean estimator, recently introduced in[32] for the Monte-Carlo estimation of all marginals. Let A be a given subset of the (C, a, p) space, then it follows that:

$$\int \mathcal{L}^{C,a,p}(y, s) \pi_{0,\theta}(C, a, p) dC da dp = \frac{\pi_{\theta,0}(A)}{\pi_{\theta|(y,s)}(A)} \int_A \mathcal{L}^{C,a,p}(y, s) \pi_{A,\theta}(C, a, p) dC da dp, \tag{7}$$

where $\pi_{A,\theta}$ is the prior density of network class θ, conditional on $\theta \in A$. Each marginal is then estimated using the following procedure:
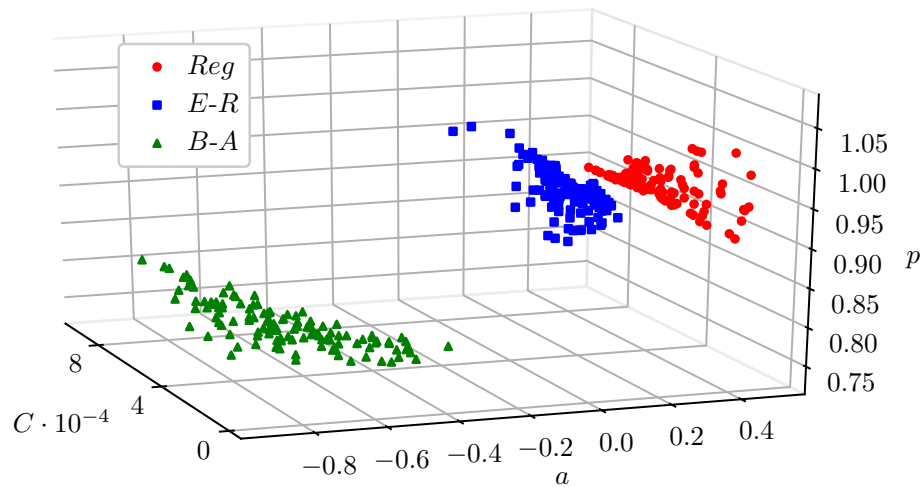
1. Find

Figure 4. Estimated C, a, p values from the dataset $\mathcal{S}$ (360 points in total, each coming from a unique combination of (network class, $\langle k \rangle$, τ, γ)). From left to right, we observe three distinct regions corresponding to Barabási–Albert (triangles), Erdős–Rényi (squares) and Regular networks (circles) networks.

$$(C^*, a^*, p^*) = \arg\max_{C,a,p} \left(\log \mathcal{L}^{C,a,p}(y,s) + \log \pi_{\theta,0}(C,a,p) \right).$$

This is done via a combination of global/local optimisation routines.

2. Sample the distribution $\pi_\theta(C, a, p|y, s)$ using a Random-Walk Metropolis-Hastings algorithm starting from (C^*, a^*, p^*) and denote the samples by $(C_i, a_i, p_i)_{1 \leq i \leq K}$ with $K = 500$.
3. Let H be the Fisher information evaluated at (C^*, a^*, p^*) and let $d(C, a, p)$ be defined as

$$d(C,a,p) := \left\langle (C_i, a_i, p_i) - (C^*, a^*, p^*), H\left[(C_i, a_i, p_i) - (C^*, a^*, p^*) \right] \right\rangle.$$

We then take $A := \left\{ (C, a, p) | d(C, a, p) \leq r \right\}$ where $r = \max_{1 \leq i \leq K} d(C_i, a_i, p_i)$. This choice was already suggested in[32]. In particular, it leads to $\pi_{\theta|(y,s)}(A) \approx 1$, which simplifies the right-hand-side of Eq. (7).
4. Use a Gaussian distribution $\mathcal{N}\left((C^*, a^*, p^*), H^{-1} \right)$ to estimate both $\pi_{\theta,0}(A)$ and the integral term on the right-hand-side of Eq. (7) by importance sampling.

Our complete Python implementation of this routine is available online, https://github.com/BayIAnet/NetworkInferenceFromPopulationLevelData.

Network class inference. In this section, we provide numerical results assessing the overall quality and applicability of our approach. We start by inferring networks from a testing dataset, where all data are simulated from either Regular, E-R or B-A networks, see Sect. "Inference based on the testing set". We then consider networks outside of our framework, namely synthetic networks with negative binomial degree distributions (Sect. "Inference of synthetic networks") and real-world networks (Sect. "Inference of real-world networks"). In all cases, we provide posterior probabilities for each network class across independent repetitions of the datasets to quantify uncertainty.

Inference based on the testing set. During the construction of the prior, we deliberately set aside 60 estimated (C, a, p) values to build a test set (20 per network class taken at random), meaning that they were not used in the calibration of the prior. In this section, we use this set to check if we can infer the known underlying network class.

The inference was performed as follows. For each of the (C, a, p) parameters in the testing set, we used the known underlying network and disease parameters (network class, $\langle k \rangle$, τ, γ) to simulate a dataset (y, s) with Gillespie's algorithm. We only generated a single network from the appropriate class and simulated a single epidemic. However, we generated 10 independent datasets, as shown in Fig. 3, and ran our inference model on each of them separately. The second step was to compute the 3 posterior probabilities corresponding to the different network classes, as detailed earlier. We thus obtained 3 posterior probabilities for each of the 60 elements in our test set and predicted the most likely underlying network class. To assess the uncertainty due to data sampling, we considered the results across all the independent datasets.

The quality of the inference is shown by the confusion matrix (Table 1), which provides the averaged posterior probabilities along with their standard deviation for each of the possible outcomes. The level of accuracy achieved in our tests is remarkable, with a score as high as 95% for Barabási–Albert, and a minimum of 79% for Erdős–Rényi. This also shows that there can be a moderate confusion between the Regular and Erdős–Rényi network classes, as their characteristics are quite similar w.r.t. (C, a, p) values (see Fig. 4) whereas Barabási–Albert

7

True/predicted	Regular	Erdős–Rényi	Barabási–Albert
Regular	85.5% (7.9%)	14.5% (7.9%)	0.0% (0.0%)
Erdős–Rényi	21.5% (10.7%)	78.5% (10.7%)	0.0% (0.0%)
Barabási–Albert	0.0% (0.0%)	5.0% (5.0%)	95.0% (5.0%)

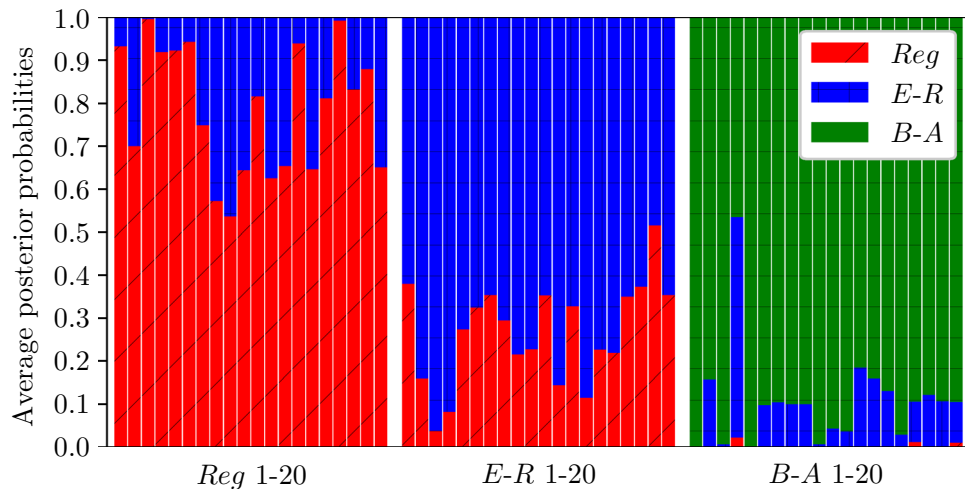**Table 1.** Averaged confusion matrix based on the test dataset (standard deviation is brackets).

Figure 5. Average posterior probability over the 60 tests (20 per network class and 10 realisations).

is rarely miss-classified. Further, the standard deviations show that these scores are stable across different data realisations, suggesting that our approach is consistent.

To get a more precise description of the classification results, we computed the average posterior probability for each of the 60 test elements, see Fig. 5. This revealed that the average posterior probability varies within each of the network class, probably due to differences in network or disease parameters. In some sense, this shows that for some network and disease parameters, the similarity between Regular and Erdős–Rényi is significant. For example, when the epidemic spreads fast and infects many nodes early on, the structure of the network is less important as the infection will be transmitted on. This means that the average degree is more important than the degree distribution. Nevertheless, our inference methodology returns a good classification in most cases. In fact, these tests show that our approach can successfully recover the network class from as little as 21 observations of a single epidemic.

Finally, we detail specificity and sensitivity for the 10 repetitions of the classification, offering per network class and global statistics in Fig. 6. We note that each marker has 10 occurrences but in some cases these are superimposed. Here again, one can see the stability and high efficiency of our approach for Barabási–Albert, with more confusion for Erdős–Rényi.

Inference of synthetic networks. We have shown that our methodology performs well when applied to the data generated on the networks that it was trained on. In this Section, we consider alternative network types for two reasons: (a) to stress-test the classification by using networks whose degree distributions do not come from the models used to build priors, and (b) to study the extent to which it can distinguish between different levels of heterogeneity in degree distribution.

To do this, we generated three synthetic networks using the configuration model[30] and a negative binomial degree distribution with parameters (p, n):

$$\forall k \in \{0, n\}, \ P(k) = \binom{k + n - 1}{k} p^k (1 - p)^{n-k}, \tag{8}$$

where p is the probability of success and n the number of failures. This choice is motivated by both the simplicity and the flexibility of this distribution. The average degree in all three networks was identical (i.e. fixed at $\langle k \rangle = 6$) but with different levels of heterogeneity depending on the variance, see Fig. 7a. To avoid the possibility of having disconnected components, the degree distributions were shifted so that the minimum was greater or equal to 3. Here, the degree distributions were chosen to exhibit different levels of heterogeneity, from low to a level comparable to those achieved in B–A networks. We then ran 10 independent epidemics with parameters $\gamma = 1$ and $\tau = 0.5$, starting from 5 infected nodes. As in Sect. "Inference based on the testing set", the inference was based on a dataset with 21 equally-spaced observations of the number of infected nodes. The results are shown
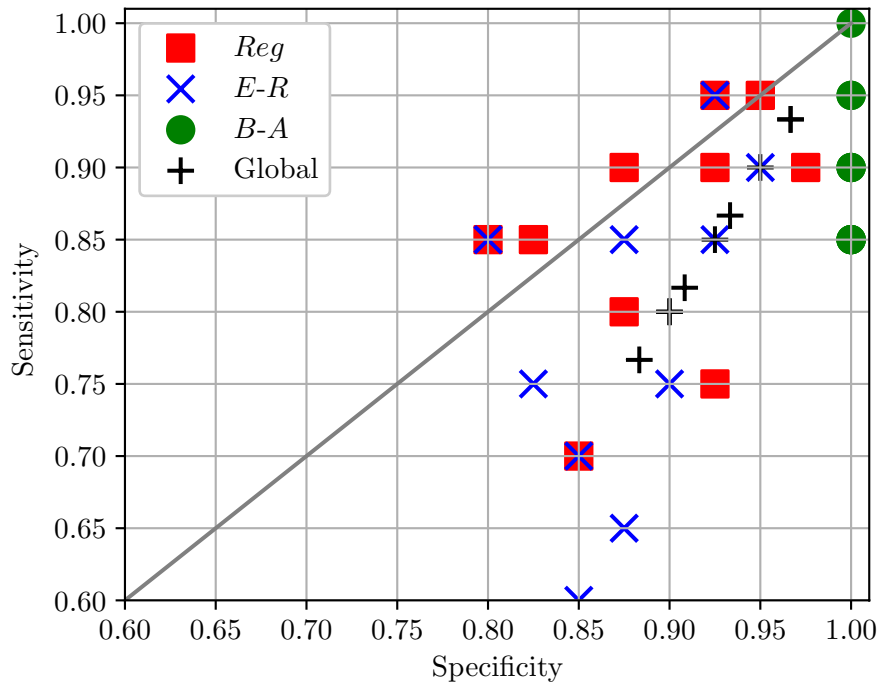
Figure 6. Specificity and sensitivity of the 10 independent classification at global and per-network levels.
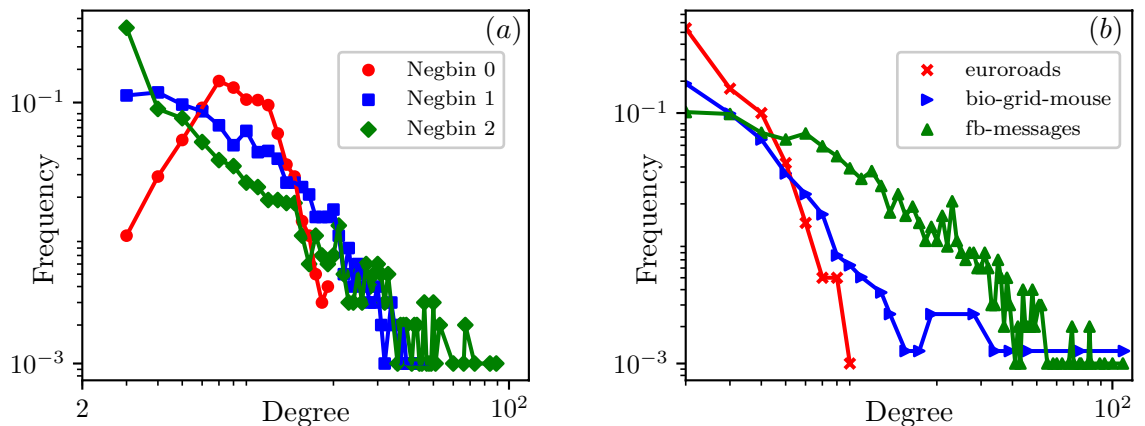

Figure 7. Degree distributions, ordered by variance, of three single negative binomials (**a**) following Eq. (8) and of the three real networks (**b**) used for the stress test. For (**a**), the average degree is $\langle k \rangle = 6$ for all networks. From low to high variance we have $\sigma = 8$ (Negbin 0), $\sigma = 40$ (Negbin 1), $\sigma = 120$ (Negbin 2). The values of (p, n) are $(0.23, 20)$, $(0.85, 1)$, $(0.95, 0.3)$, respectively. For (**b**) the basic metrics of these networks are $\{\langle k \rangle, \sigma^2, \text{Assortativity}, \text{Clustering}\} = \{2.53, 5.24, 0.102, 0.02\}, \{2.77, 40, -0.21, 0.04\}, \{12.30, 268.90, -0.08, 0.09\}$, respectively.

in Fig. 8, and confirm that our inference scheme is able to distinguish between networks with high/low levels of degree heterogeneity. In particular, by looking at Fig. 7a it is reasonable to expect that the first and third networks are going to be classified as E–R and B–A networks, respectively. Indeed, Fig. 8 shows that the first network in Fig. 7a is identified as E–R 80% of the time, whereas the third network in Fig. 7a is correctly classified as B–A for every single epidemic realisation.

When the degree distribution of the test network is such that its variance falls between typical variances observed in E–R and B–A networks (see the second network in Fig. 7a) our results are more sensitive to the individual realisation of the epidemic. However, even in this case, the network is identified to the closest type in terms of degree distribution. Moreover, heuristically at least, the B–A network seems to be favoured, which seems reasonable upon inspecting the degree distribution of the test network.

Inference of real-world networks. Finally, the last test we conducted was based on real world networks, which can exhibit higher-order structure beyond degree heterogeneity. We chose three real networks: the first is labelled
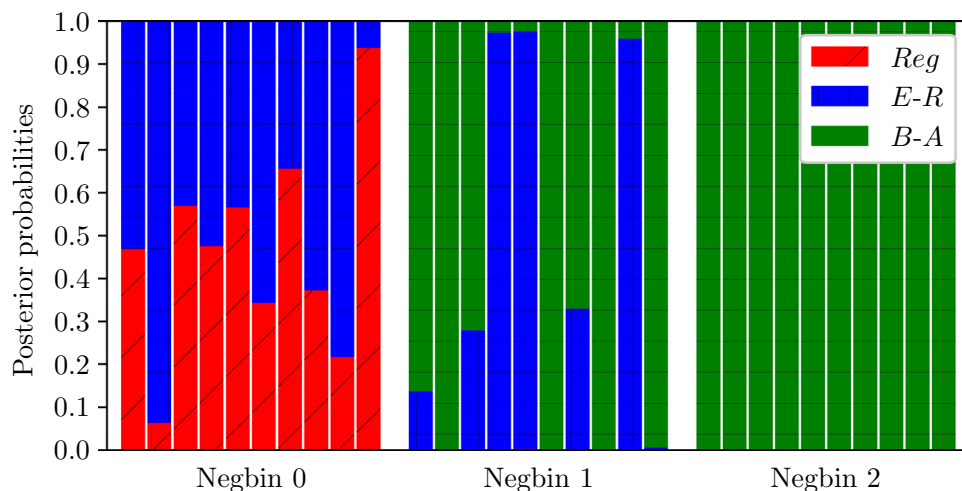
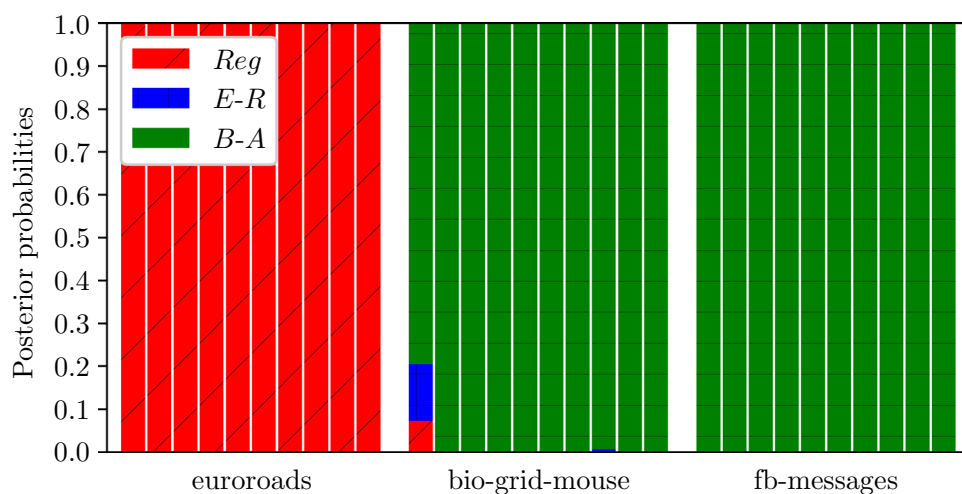**Figure 8.** Posterior probabilities for the 10 repetitions on each synthetic network.

Figure 9. Posterior probabilities for the 10 repetitions on each real-world network.

euroroads and is part of the KONECT collection[19], the second and third, bio-grid-mouse and fb-messages, are part of the network data repository Networkrepository[38]. The euroroads is an infrastructure network, bio-grid-mouse is a protein-protein network whilst fb-messages is based on the interactions of an online community of students at University of California. In Fig. 7b the degree distributions of these networks are shown. To keep the number of nodes equal to $N = 1000$, we only considered the largest connected component, and then, where necessary, removed peripheral, low-degree nodes such that the resulting network was still connected.

In line with Sect. "Inference of synthetic networks", we fixed $\gamma = 1$, and ran 10 distinct epidemics on each network in order to generate data for the inference. Values for the infection parameter were $\tau = \{1.5, 2.5, 0.4\}$ for euroroads, bio-grid-mouse and fb-messages, respectively. The posterior probabilities obtained from our approach are reported in Fig. 9 and are in line with our expectations based on the inspection of the respective degree distributions: the infrastructure network is very homogeneous, whilst the other two are scale-free, and hence correctly classified as B–A.

Discussion

In this paper, we proposed a new inference scheme that uses population-level incidence data at discrete regular times to infer the most likely network class over which the epidemic has initially spread. This is a challenging task because the exact epidemic model on a given network is forbiddingly high-dimensional meaning that even a numerical solution is out of reach. The key to carry out the inference is the approximation of the exact epidemic model by a BD process, whose rates not only encode the structure of the networks but also allow us to distinguish between the different network classes through a parsimonious three-parameter model. Whilst we have successfully numerically validated this surrogate model over a number of network classes and different
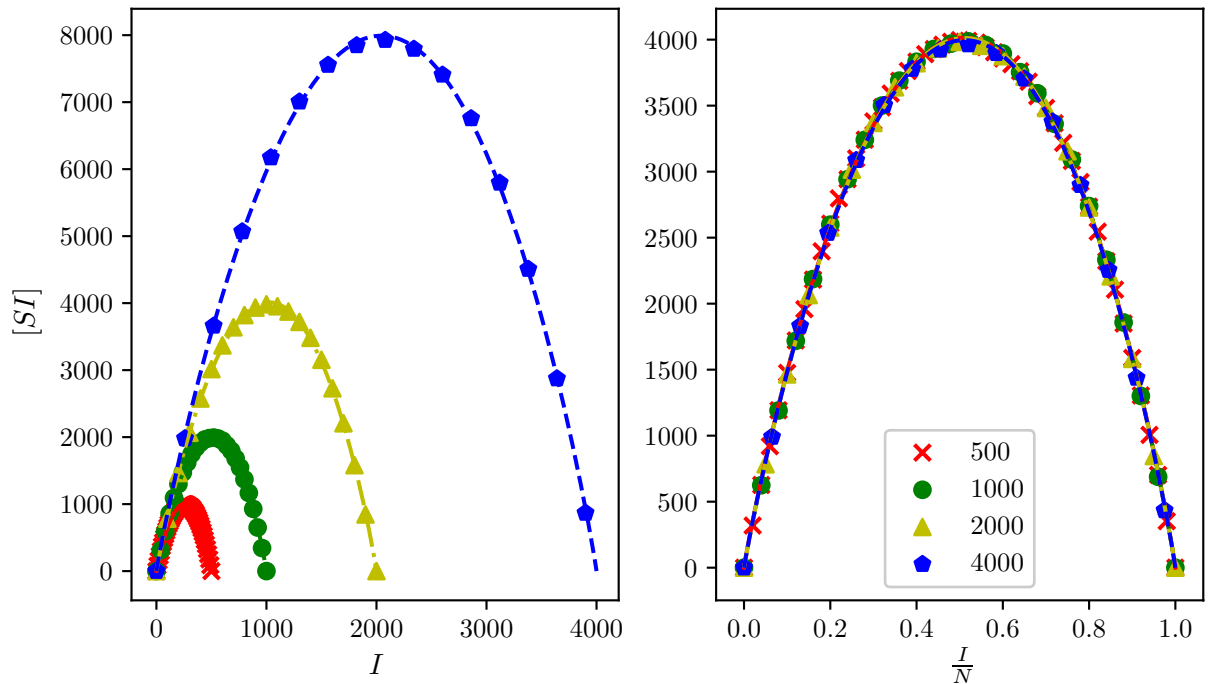
Figure 10. (Left) $(k, \hat{a}_k)$ curves based on Erdős–Rényi networks with $\langle k \rangle = 5, \tau = 1.793, \gamma = 3.785$ and $N = 500, 1000, 2000, 4000$. (Right) Scaled $(k, \hat{a}_k)$ curves relative to the $N = 2000$ case. Scaled version are obtained by plotting $(k/N, \frac{2000}{N}\hat{a}_k)$.

values of disease parameters, with further evidence in[27], a mathematical characterisation of the relation between the exact and this surrogate model remains an open problem.

Our analysis has focused on three classes of random networks: Regular, Erdős–Rényi and Barabási–Albert. This choice is motivated by the fact that such classes are well-known, simple to describe (depending only on one parameter) while differing in terms of degree heterogeneities. For each network class, the rates of infection in the corresponding BD approximation was obtained by using the time-weighted mean of S–I link counts. Despite these rates being network class-dependent, they all share some common features. This in turn allowed us to propose a parsimonious three-parameter model (C, a, p) that works across all network classes and, at the same time, can capture the differences in the rates of the approximating BD process. In addition to being robust to different values of τ, γ and average degree, these parameters exhibit a clear distinction between the three different network classes when plotted in the 3-dimensional (C, a, p) space. This knowledge is then encoded into prior distributions, constructed using kernel density estimators over the (C, a, p) space. Our Bayesian model selection procedure then consists in the numerical estimation of the relative marginal probabilities. Our results show that the inference scheme has good specificity and sensitivity, despite the simplicity of the model. These encouraging results lead to a number of questions and remarks. First of all, our choice of classes of random networks means that the main feature of the networks is their degree heterogeneity. We have yet to consider more complex networks, such as those exhibiting clustering or community structure. This would certainly lead to $(k, \hat{a}_k)$ curves of different shapes, potentially having other features such as multiple peaks for networks with multiple communities, and thus requiring either a more sophisticated or non-parametric model. Nevertheless, considering epidemics in terms of an approximate BD process appears to be a powerful approach if a tractable likelihood is desired. Moreover, once the most likely network class has been identified, one could continue and estimate τ, γ and the average degree.

We have used a fixed number of nodes ($N = 1000$) in all our numerical experiments. We do not expect major changes when the number of nodes is different. Preliminary numerical tests, see Fig. 10, suggest that there is a good degree of universality such that the $(k, \hat{a}_k)$ curves only differ by a scaling factor when the number of nodes changes, all other parameters being fixed. In this respect, our methodology could easily be adapted by directly considering the scaled epidemic (on $[0, 1]$) and repeating our tests for different values of N. Fortunately, our numerical method[4] scales well with N, since the transition probabilities in the likelihood are computed individually (with deeper continued fractions). The question of the limiting behaviour in the limit of large N can also be further investigated.

An interesting open question is that of the extent to which different network families are mapped onto distinct regions in the C, a, p space if networks are weighted, i.e., if the adjacency matrix has entries of magnitude $0 \le g_{ij} \le 1$. While a comprehensive answer to this query would require extensive simulations beyond the scope of this paper, there are a couple of points worth making. First, we already see that Regular and Erdős–Rényi classes are only really distinguishable when networks are sparse. If we keep τ, γ fixed and increase the average degree $\langle k \rangle$, we observe that both tend to the fully-connected network limit, where $C = \frac{\tau}{N}, a = 0, p = 1$. This is because the fully-connected network can be seen both as a regular network with degree $\langle k \rangle = N - 1$ and as an
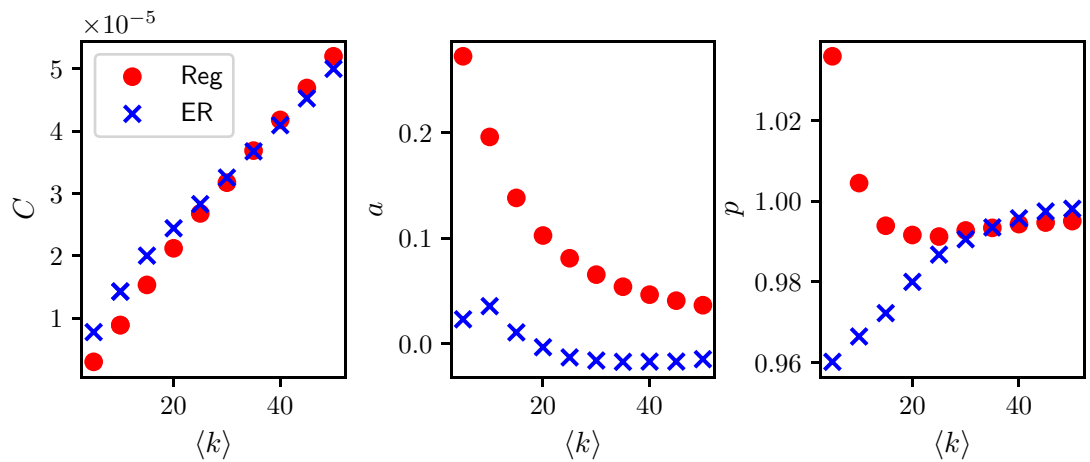
11

Figure 11. Behaviour of C, a, p when $\langle k \rangle$ increases for regular networks (circles) and Erdős–Rényi (crosses), all other parameters being equal. The maximum value of $\langle k \rangle$ explored is $\langle k \rangle = 50$. $\tau = 1$, $\gamma = 5$. The fully-connected limit, reached when $\langle k \rangle = N - 1$, is $C = \frac{1}{N}$, $a = 0$, $p = 1$; however, even at $\langle k \rangle = 50$, we can see how un-identifiability emerges between regular and Erdős–Rényi networks.

Erdős–Rényi with $p = 1$, see Fig. 11. This means that there is some degree of un-identifiability when network classes generate networks that are topologically almost identical to one another. Further, any unweighted network can be seen as a weighted fully-connected network, with weights either 0 or 1. For instance, an Erdős–Rényi network is a weighted complete network such that the element g_{ij} has weight 0 with probability p, 1 otherwise. With this consideration in mind, the question can be rephrased as: is it possible to use this framework to infer the weight distribution over a fully-connected network? Our conjecture is that the answer is yes. Provided that the weight distributions are distinct enough and the k, a_k curves can be captured by a model such as the C, a, p, we do expect to find similar scenarios to those shown by Fig. 4.

So far, we have used discrete data taken on a regular time grid covering the epidemic from its early stage (a few infectious nodes) up to its steady state. Increasing the frequency of data or restricting data to the very beginning of the epidemic are of significant practical interest. In the former case, one expects the discrete likelihood to converge to the simpler continuous one, enabling faster and easier analysis. In the latter case, it would lead to a model that does not require describing the whole epidemic as we currently do. Focusing on the initial stages of the epidemic, the most critical period in many cases, and upon solving a potential un-identifiability problem, such an approach could have an important real-world impact, making it possible to predict and control more accurately yet-to-be epidemics.

Finally, the proposed inference scheme could be improved by using more sophisticated models for the infection rates and by learning a larger number of different network classes, leading to a wide portfolio of data which can then be used for estimation. Of course, there is a trade-off in terms of what we can infer about networks using population-level discrete data. We cannot infer individual links for example but this is to be expected since the data we use for inference is not at the link- or node-level. Nevertheless, we believe that our approach could have practical implications, as the inference scheme is based on the kind of data that is most likely to be available in real-world scenarios (e.g. the number of infected people every day or week). Where such data is available but little is know about the contact pattern, our inference scheme may be able to provide some high-level information about the properties of the network which in turn could be exploited in the planning or implementation of control, in particular during the early stages of an epidemic.

References

1. Britton, T. & ONeill, P. D. Bayesian inference for stochastic epidemics in populations with random social structure. *Scand. J. Stat.* **29**(3), 375–390 (2002).
2. Brugere, I., Gallagher, B. & Berger-Wolf, T. Y. Network structure inference, a survey: motivations, methods, and applications. *ACM Comput. Surv.* **51**(2), 24:1-24:39 (2018).
3. Crawford, F. W., Minin, V. N. & Suchard, M. A. Estimation for general birth-death processes. *J. Am. Stat. Assoc.* **109**(506), 730–747 (2014).
4. Crawford, Forrest W. & Suchard, Marc A. Transition probabilities for general birth-death processes with applications in ecology, genetics, and evolution. *J. Math. Biol.* **65**, 553–580 (2012).
5. Danon, L. *et al.* Networks and the epidemiology of infectious disease. *Interdisc. Perspect. Infect. Dis.* **2011**, 28 (2010).
6. Du, N., Song, L., Yuan, M. & Smola, A. J. Learning networks of heterogeneous influence. *Adv. Neural Inf. Process. Syst.* **25**, 2780–2788 (2012).
7. Dutta, R., Mira, A. & Onnela, J. Bayesian inference of spreading processes on networks. *Proc. R. Soc. A Math. Phys. Eng. Sci.* **474**(2215), 20180129 (2018).

8. Gillespie, D. T. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comput. Phys.* **22**(4), 403–434 (1976).
9. Gillespie, D. T. Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**(25), 2340–2361 (1977).
10. Gleeson, J. P. Bond percolation on a class of clustered random networks. *Phys. Rev. E* **80**(3), 036107 (2009).
11. Gomez Rodriguez, M., Leskovec, J., Balduzzi, D. & Schölkopf, B. Uncovering the structure and temporal dynamics of information propagation. *Netw. Sci.* **2**(1), 26–65 (2014).
12. GomezRodriguez, M., Leskovec, J. & Krause, A. Inferring networks of diffusion and influence. In *Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'10* 1019–1028 (ACM, New York, NY, USA, 2010).
13. Groendyke, C., Welch, D. & Hunter, D. R. Bayesian inference for contact networks given epidemic data. *Scand. J. Stat.* **38**(3), 600–616 (2011).
14. Hethcote, H. W. & van den Driessche, P. Some epidemiological models with nonlinear incidence. *J. Math. Biol.* **29**(3), 271–287 (1991).
15. Keeling, M. J. & Eames, K. T. D. Networks and epidemic models. *J. R. Soc. Interface* **2**(4), 295–307 (2005).
16. Kennedy, J. & Eberhart, R. Particle swarm optimization. In *Proceedings of ICNN'95 - International Conference on Neural Networks* Vol. 4. 1942–1948 (Nov 1995).
17. Kermack, W. O. & McKendrick, A. G. A contribution to the mathematical theory of epidemics. *R. Soc. Lond. Proc. Ser. A* **115**, 700–721 (1927).
18. Kiss, I. Z., Miller, J. C. & Simon, P. L. *Mathematics of Epidemics on Networks: From Exact to Approximate Models* (Springer, Berlin, 2017).
19. Kunegis, J. Konect—the koblenz network collection. arXiv:1402.5500 (2017).
20. Liu, W., Levin, S. A. & Iwasa, Y. Influence of nonlinear incidence rates upon the behavior of sirs epidemiological models. *J. Math. Biol.* **23**(2), 187–204 (1986).
21. Ma, L., Liu, Q. & Van Mieghem, P. Inferring network properties based on the epidemic prevalence. *Appl. Netw. Sci.* **4**(1), 1–13 (2019).
22. McKay, M. D., Beckman, R. J. & Conover, W. J. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21**(2), 239–245 (1979).
23. Miller, J. C. Percolation and epidemics in random clustered networks. *Phys. Rev. E* **80**(2), 020901(R) (2009).
24. Miller, J. C., Slim, A. C. & Volz, E. M. Edge-based compartmental modelling for infectious disease spread. *J. R. Soc. Interface* **9**(70), 890–906 (2012).
25. Moore, C. & Newman, M. E. J. Epidemics and percolation in small-world networks. *Phys. Rev. E* **61**(5), 5678 (2000).
26. Myers, S. & Leskovec, J. On the convexity of latent social network inference. *Adv. Neural Inf. Process. Syst.* **23**, 1741–1749 (2010).
27. Nagy, N., Kiss, I. Z. & Simon, P. L. Approximate master equations for dynamical processes on graphs. *Math. Model. Natural Phenom.* **9**(2), 43–57 (2014).
28. Netrapalli, P. & Sanghavi, S. Learning the graph of epidemic cascades. *SIGMETRICS Perform. Eval. Rev.* **40**(1), 211–222 (2012).
29. Newman, M. E. J. The structure and function of complex networks. *SIAM Rev.* **45**(2), 167–256 (2003).
30. Newman, M. E. J., Strogatz, S. H. & Watts, D. J. Random graphs with arbitrary degree distributions and their applications. *Phys. Rev. E* **64**, 026118 (2001).
31. ONeill, P. D. & Roberts, G. O. Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc. Seri. A (Stat. Soc.)* **162**(1), 121–129 (1999).
32. Pajor, A. Estimating the marginal likelihood using the arithmetic mean identity. *Bayesian Anal.* **12**(1), 261–287 (2017).
33. Pastor-Satorras, R., Castellano, C., Van Mieghem, P. & Vespignani, A. Epidemic processes in complex networks. *Rev. Modern Phys.* **87**, 925 (2015).
34. Pedregosa, F. *et al.* Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
35. Porter, M. A. & Gleeson, J. P. *Dynamical Systems on Networks: A Tutorial* (Springer, Berlin, 2016).
36. Prasse, B. & Van Mieghem, P. Exact network reconstruction from complete sis nodal state infection information seems infeasible. *IEEE Trans. Netw. Sci. Eng.* (2018).
37. Roy, M. & Pascual, M. On representing network heterogeneities in the incidence rate of simple epidemic models. *Ecol. Complex.* **3**(1), 80–90 (2006).
38. Ryan, A. R. & Nesreen, K. A. The network data repository with interactive graph analytics and visualization. In *AAAI* (2015).
39. Simon, P. L. & Kiss, I. Z. From exact stochastic to mean-field ODE models: a new approach to prove convergence results. *IMA J. Appl. Math.* **78**(5), 945–964 (2013).
40. Simon, P. L., Taylor, M. & Kiss, I. Z. Exact epidemic models on graphs using graph-automorphism driven lumping. *J. Math. Biol.* **62**(4), 479–508 (2011).
41. Stack, J. C., Bansal, S., Anil Kumar, V. S. & Grenfell, B. Inferring population-level contact heterogeneity from common epidemic data. *J. R. Soc. Interface* **10**(78), 20120578 (2013).
42. Wang, Y., Cao, J., Li, X. & Alsaedi, A. Edge-based epidemic dynamics with multiple routes of transmission on random networks. *Nonlinear Dyn.* **91**(1), 403–420 (2018).

Acknowledgements

Author contributions

F.D-L., J.-C.C., L.B., M.D. and I.Z.K. planned and designed the study. F.D-L. performed most of the simulations and wrote first draft. J.-C.C. implemented, performed and wrote up the Bayesian inference part and commented on all drafts. M.D. and L.B. commented extensively on the inference part and edited the manuscript. I.Z.K. edited the first draft extensively and commented on all aspects of the work.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to I.Z.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.