PROCEEDINGS OF THE ROYAL SOCIETY A

rspa.royalsocietypublishing.org

Review

Cite this article: House T, Ross JV, Sirl D. 2013 How big is an outbreak likely to be? Methods for epidemic final-size calculation. Proc R Soc A 469: 20120436.
http://dx.doi.org/10.1098/rspa.2012.0436

Received: 17 July 2012
Accepted: 8 November 2012

Subject Areas:
applied mathematics, computational mathematics, mathematical modelling

Keywords:
epidemiology, infectious disease, Markov chain, susceptible-infectious-recovered(s)

Author for correspondence:
Thomas House
e-mail: t.a.house@warwick.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rspa.2012.0436 or via http://rspa.royalsocietypublishing.org.

Royal Society Publishing
Informing the science of the future

How big is an outbreak likely to be? Methods for epidemic final-size calculation

Thomas House[1], Joshua V. Ross[2] and David Sirl[3]

[1]Warwick Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK
[2]Stochastic Modelling Group, School of Mathematical Sciences, University of Adelaide, Adelaide, South Australia 5005, Australia
[3]Mathematics Education Centre, University of Loughborough, Loughborough LE11 3TU, UK

Epidemic models have become a routinely used tool to inform policy on infectious disease. A particular interest at the moment is the use of computationally intensive inference to parametrize these models. In this context, numerical efficiency is critically important. We consider methods for evaluating the probability mass function of the total number of infections over the course of a stochastic epidemic, with a focus on homogeneous finite populations, but also considering heterogeneous and large populations. Relevant methods are reviewed critically, with existing and novel extensions also presented. We provide code in MATLAB and a systematic comparison of numerical efficiency.

1. Introduction

(a) Motivation

Epidemic models are now widely used to inform policy on a range of issues, from childhood diseases [1] to bioterrorist smallpox [2], SARS [3], foot-and-mouth disease [4] and pandemic influenza [5]. These models have typically involved either numerical integration of ordinary differential equations that do not explicitly account for the underlying chance events in transmission, or Monte Carlo simulation of models that aspire to a high level of realism [1,6].

There has been growing interest, however, in consideration of the parsimonious stochastic epidemic models that were described during the earliest phase of mathematical epidemiology [7]. This is partly because of the impressive array of formal results that have accumulated over the years in this field [8], but also because of the role that these models can play in modern, computationally intensive inference of epidemiological parameters [9–11] and optimization problems.

The final size of an epidemic can be defined informally as the total number of people experiencing infection during the outbreak. This quantity is typically called the *attack rate* by applied epidemiologists, and expressed as a percentage of the population in question. The probability distribution of final sizes is of particular interest to statistical epidemiologists, owing to its use *inter alia* in the analysis of household data [12,13].

It turns out that there is a particularly large number of approaches to calculation of the probability mass function (PMF) for the final size of an epidemic (often called the *final-size distribution*). This paper aims to summarize these approaches, paying particular attention to: (i) numerical implementation, including a novel application for iterative methods; (ii) critical comparison of different methods; and (iii) consideration of how these methods can be applied to calculation of other epidemiological quantities of interest.

(b) Model definition

While we consider various generalizations, our starting point is the susceptible-infectious-recovered (SIR) epidemic model in a closed finite population. We consider a population of integer size N, and the state of the system is given by non-negative integer-valued stochastic variables $S(t)$ and $I(t)$, obeying $S + I \leq N$, which represent the number of individuals who are susceptible or infectious at time t, respectively. The number of recovered individuals R is given by $R = N - S - I$ due to our assumption that the population is closed (i.e. there are no births, deaths or migrations).

We assume that any pair of individuals makes contact at the points of a homogeneous Poisson process of rate β and that contacts between different pairs of individuals are mutually independent. Contact between an infectious and a susceptible individual results in the immediate infection of the susceptible individual, who experiences a random duration of infectiousness drawn from the *infectious period distribution* (also often called the recovery time distribution) and then recovers. Note that in these conventions, the overall rate of infection is βSI, so when comparing results across different values of N, it is often most instructive to hold the quantity $\beta/(N-1)$ constant.

The epidemic process starts in state (S_0, I_0) and must end when there are no more infectious individuals. Writing the final state in the form $(N - Z, 0)$ allows us to define the integer-valued stochastic variable Z as the *final size* of the epidemic. We are particularly interested in the PMF of this quantity; since this function has finite integer support, it is often conveniently represented as a vector of probabilities

$$\mathbf{p} = (p_k), \quad \text{for} \quad p_k = \Pr(Z = k). \tag{1.1}$$

In the rest of this paper, we will be interested in either the calculation of the components of the final-size distribution (often in vector representation) at machine precision or sampling from this distribution using Monte Carlo methods.

2. Material and methods

We now consider different ways to calculate the final-size distribution, starting with Monte Carlo methods that are simply described in §2*a*, before moving on to methods that run at machine precision in §2*b*,*c*. We have tried to use notation that is consistent within this manuscript, but in doing so, we may deviate from the conventions originally used for each technique.

(a) Monte Carlo methods

(i) Direct continuous-time simulation

In the event that the infectious period distribution is exponential with rate parameter γ (i.e. mean γ^{-1}) the system dynamics take the form of a continuous-time Markov chain. The events and their rates of occurrence for this Markov chain are

$$\left.\begin{aligned} &(S,I)\to(S-1,I+1) \quad \text{at rate } \beta SI\\ \text{and}\quad &(S,I)\to(S,I-1) \quad \text{at rate } \gamma I. \end{aligned}\right\} \tag{2.1}$$

This system can then be simulated directly using Monte Carlo methods, once initial conditions are specified. Stigler's Law [14] states that scientific discoveries are not named after their originators (and was, of course, discovered earlier by Robert K. Merton). It is therefore as would be expected that while the most common method for Monte Carlo simulation of a continuous-time Markov chain is typically called Gillespie's algorithm [15], this method was actually developed by probabilists some decades earlier [16–18].

The results of simulating the Markov chain defined by (2.1) are shown in figure 1*a*,*b*, for a relatively large and small outbreak, respectively. We provide the function `gil_mc.m` in the electronic supplementary material, (S1.1), as an implementation of this, based on the code from Keeling & Rohani [6]. Obviously, if one is interested in temporal dynamics of the epidemic, all of the information created by this method is useful. But for the purposes of sampling from the final-size distribution, there is a lot of unnecessary computational effort expended by this method.

Extension: waning immunity. Some of the methods presented in this paper rely on long-lasting immunity following recovery from natural infection—the SIR paradigm. Many diseases, for example respiratory synctial virus (RSV) [19] and rotavirus [20], exhibit significant waning immunity, which is readily incorporated in approaches based on direct simulation by adding the transition

$$(S,I)\to(S+1,I) \quad \text{at rate } \mu(N-S-I), \tag{2.2}$$

to (2.1), giving SIRS epidemic dynamics. The system as defined always ends in state $(N,0)$ and so what to calculate depends on the epidemiological problem in hand. In the context of SIRS epidemics with household structure, the total number of infection events is of interest due to its role in the calculation of epidemic thresholds [21,22], and this can be readily extracted from a realization of the full SIRS epidemic.

Extension: phase-type infectious periods. An assumption behind (2.1) is that the infectious period durations are exponentially distributed—this is justifiable for some diseases, but certainly not for others. Suppose we now introduce an integer index a between 1 and k for infectious individuals, and modify (2.1) to

$$\left.\begin{aligned} &(S,\ldots,I_a,\ldots)\to(S-1,\ldots,I_a+1\ldots) \quad \text{at rate } \beta_a S\sum_b I_b,\\ &(S,\ldots,I_a,\ldots,I_b,\ldots)\to(S,\ldots,I_a-1,\ldots,I_b+1,\ldots) \quad \text{at rate } q_{a,b}I_a\\ \text{and}\quad &(S,\ldots,I_a,\ldots)\to(S,\ldots,I_a-1,\ldots) \quad \text{at rate } \gamma_a I_a, \end{aligned}\right\} \tag{2.3}$$

so that the distribution of infectious periods is a phase-type distribution. Since phase-type distributions are dense in positive-valued distributions [23], any reasonable infectious period distribution can be approximated by a model of similar form to (2.3); however, this comes at potentially large computational cost. The most commonly used phase-type distribution is the hypo-exponential, where individuals pass from $I_0\to I_1\to\cdots\to I_k\to R$ linearly, which has

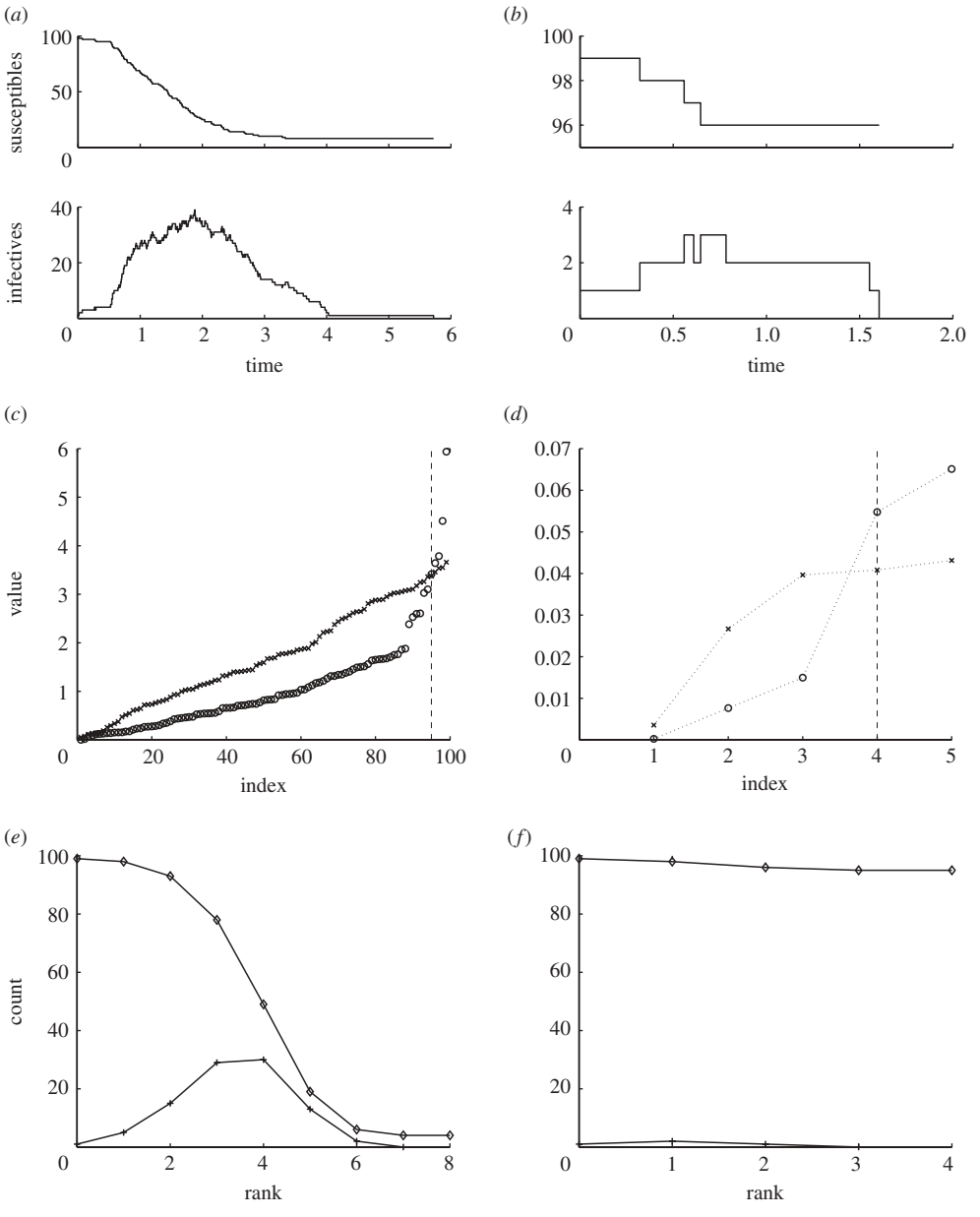

Figure 1. Realizations of different simulation schemes. Parameters are $N = 100$, $\beta = 3/(N-1)$, $I_0 = 1$, $S_0 = N - I_0$, exponential infectious period distribution with unit mean. (*a*) Gillespie, large outbreak, (*b*) Gillespie, small outbreak, (*c*) Sellke, large outbreak, (*d*) Sellke, small outbreak, (*e*) Ludwig, large outbreak and (*f*) Ludwig, small outbreak. (*c*,*d*) Open circles, sorted thresholds; cross symbols, cumulative infectivity; dashed line, final size. (*e*,*f*) Lines with diamonds, susceptibles; lines with crosses, new cases.

the effect of reducing variability compared with an exponential distribution, and leads to a *k*-Erlang infectious period distribution when the transition rates are all equal. Another frequent use of phase-type distributions is the hyper-exponential distribution, where all $q_{a,b}$ in (2.3) are zero meaning that each infective passes through a unique I_a state, which is often interpreted as individuals experiencing different disease severities, and which has the effect of increasing variability compared to an exponential distribution.

(ii) Sellke's method

Sellke's method provides a way to simulate final sizes of a stochastic epidemic with arbitrary infectious period distribution [24]. We suppose that each individual i has a stochastic variable T_i for its infectious period, picked from the infectious period distribution, and that susceptible individuals have a random *threshold* Q_i picked from an exponential distribution of mean 1 (in this work, we take the initial infectives to have zero threshold parameter). We arrange the labelling of individuals, without loss of generality, so that $Q_i \le Q_j$ when $i < j$. Then

$$Z = \min\left\{ i \middle| Q_{i+1} > \beta \sum_{j \le i} T_j \right\}. \tag{2.4}$$

The intuition behind this equation is that Q_i is essentially how 'resistant' individual i is to infection, and βT_i is the force of infection that individual i will contribute *if* ultimately infected. In this non-rigorous picture, the epidemic 'stops' when the total infectious pressure drops below the resistance/threshold of all remaining susceptibles.

To see more mathematically why this construction is equivalent to the model defined, we consider the argument of Andersson & Britton [8, §2.2]. If we let $I(t)$ be the number of infective individuals at time t, then the rate at which a given susceptible becomes infective is $\beta I(t)$ in the original model. We define the cumulative force of infection as

$$\Lambda(t) := \beta \int_0^t I(u)\,\mathrm{d}u \quad \Rightarrow \quad \lim_{t\to\infty} \Lambda(t) = \beta \sum_{i=1}^{Z} T_i. \tag{2.5}$$

Then considering a small unit of time δt, and a unit-mean exponentially distributed threshold Q we have that

$$\begin{aligned} \Pr(Q > \Lambda(t+\delta t) | Q > \Lambda(t)) &= \frac{\int_0^{\Lambda(t+\delta t)} \exp(-q)\,\mathrm{d}q}{\int_0^{\Lambda(t)} \exp(-q)\,\mathrm{d}q} \\ &= 1 - \beta I(t)\delta t + o(\delta t). \end{aligned} \tag{2.6}$$

Therefore, the rate at which Λ exceeds Q is $\beta I(t)$. Putting this together with (2.5) gives (2.4), showing that Sellke's method gives the same final size as direct simulation.

Realizations of Sellke's method for a relatively large and small outbreak are shown in figure 1*c*,*d*. Numerically, this is much more efficient and general than Gillespie's algorithm for sampling from the final-size distribution, but the function `sel_mc.m` given in electronic supplementary material, S1.2.1, does not provide the temporal dynamics of the relevant epidemic process, although such dynamics can be readily calculated. The fundamental approach of the Sellke construction is to keep track of total infection pressure and thresholds separately, which is quite generally applicable.

Extension: temporal dynamics. Once a realization of the Sellke construction has been made as described above, it is possible to construct the temporal dynamics of that realization without any further picking of (pseudo-)random numbers. This follows from (2.6) and is practically implemented by maintaining a list of recovery times for currently infectious individuals, infecting each individual in the order specified by the Q thresholds and interspersing recoveries at the appropriate times. We provide an implementation of this method in electronic supplementary material, S1.2.2, which has the major advantage that an arbitrary distribution of infectious period can be simulated efficiently without any approximation.

(iii) Ludwig's method

Another possibility for simulation comes from the construction of Ludwig [25]. This procedure does not provide any temporal information, and involves a set of discrete stages called 'ranks'. Ludwig's method starts with S_0 susceptibles, and places the I_0 initial infectives in rank 0. At a given rank g, we label the I_g infectives with a set of integers $\mathcal{I}_g = \{i | 1 \le i \le I_g\}$. Next, each infective

is cycled through in turn, picking an infectious period T_i from the relevant distribution, which then leads to an independent probability

$$\pi_i = 1 - \mathrm{e}^{-\beta T_i}, \tag{2.7}$$

of infecting each of the remaining susceptibles. The number of susceptibles S is therefore reduced due to infective i by an integer

$$\Delta S_i \sim \mathrm{Bin}(S, \pi_i). \tag{2.8}$$

Once each of the susceptibles in the current rank has been considered, all infectives in the current rank are removed, and the number of infectives in the new rank $g+1$ is $\sum_{i\in\mathcal{I}_g} \Delta S_i$. This process is continued until an empty rank is generated. Figure 1*e*,*f* shows a realization of this process for large and small outbreaks, respectively. We provide code `lud_mc.m` in electronic supplementary material, S1.3 to implement this algorithm.

Pellis *et al.* [26] have argued that while Ludwig's argument is intuitive, it is not always obvious for which generalizations of simple epidemic models it will still hold, leading these authors to provide additional detail about the approach. Indeed, the insights built up from Ludwig's argument can be applied to very complex populations incorporating several layers of structure—including households and networks—and for intrinsic varying severity, while the inclusion of infector-dependent varying severity invalidates it [27–29].

Perhaps the clearest way to confirm the validity of Ludwig's method is to make use of network theory. Using standard modern terminology [30], consider a random directed network where a link starting on individual i and ending on individual j is present with probability $\pi_{i,j}$ corresponding to the probability of infectious contact being made from i to j if i becomes infective, before i recovers. The key property required for use of Ludwig's method is that the probabilities of infectious contacts emanating from an individual must depend only on quantities that can be determined in advance, but cannot depend on (for example) the temporal behaviour of the epidemic before that individual is infected. Equation (2.7) is a simple example, but all that is needed is the calculability of probabilities of infectious contact in advance of the epidemic. A node j will be infected eventually if there is a path through the network from an individual i in the set of initially infective individuals $\mathcal{I}_0$ to j. Letting $\mathbf{D} = (D_{i,j})$ be the adjacency matrix for the random directed network (i.e. $D_{i,j}$ is 1 if i makes contact with j and is 0 otherwise) we can see that

$$j \text{ is eventually infected} \quad \Leftrightarrow \quad \exists n < \infty, i \in \mathcal{I}_0, \text{ such that } (\mathbf{D}^n)_{i,j} > 0. \tag{2.9}$$

Ludwig's method therefore involves finding the smallest n satisfying the right-hand condition of (2.9), which becomes j's rank. The assumption of independence of the links means that this can be simulated iteratively as described.

(b) Machine-precision, Markov chain methods

A continuous-time Markov chain such as that defined by the events and rates given in (2.1) can have its dynamics fully specified by a solution $\mathbf{p}(t)$ to the Kolmogorov forward equations

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathbf{p}(t) = \mathbf{p}(t)\mathbf{Q}, \tag{2.10}$$

for appropriate initial probability vector $\mathbf{p}(0)$ and rate matrix $\mathbf{Q} = (Q_{i,j})$. One way to define this matrix explicitly is by defining $\sigma(i)$ and $\iota(i)$ as the number of susceptibles and infectives associated with state i, respectively. Then

$$Q_{i,j\neq i} = \beta\sigma(i)\iota(i)\delta_{\sigma(j),\sigma(i)-1}\delta_{\iota(j),\iota(i)+1} + \gamma\iota(i)\delta_{\iota(j),\iota(i)-1}, \quad Q_{i,i} = -\sum_{j\neq i} Q_{i,j}, \tag{2.11}$$

where $\delta_{i,j}$ is the Kronecker delta. We write the jump matrix for this Markov chain

$$\mathbf{P} = (P_{i,j}), \quad \text{for} \quad P_{i,j} = \frac{Q_{i,j}(1-\delta_{i,j})}{\sum_{j\neq i} Q_{i,j}}. \tag{2.12}$$

We use $p_{S,I}(t)$ to stand for the element of $\mathbf{p}(t)$ corresponding to the probability of S susceptibles and I infectives at time t. We also use the notation $\boldsymbol{\mathcal{S}}$ for the state space of the Markov chain, which is composed of a set of absorbing states $\boldsymbol{\mathcal{A}}$ and an irreducible transient class $\boldsymbol{\mathcal{C}}$, and which has dimension $|\boldsymbol{\mathcal{S}}|$.

(i) 'Brute force' methods

In this framework, there are two 'brute force' methods for matrix-based calculation. Firstly, for a Markov chain with finite state space, equation (2.10) has a matrix exponential solution

$$\mathbf{p}(t) = \mathbf{p}(0)\mathrm{e}^{\mathbf{Q}t}, \tag{2.13}$$

which could be evaluated using, e.g. EXPOKIT [31] at a sufficiently large t. Alternatively, the probability vector after n events is, by definition of the jump matrix, $\mathbf{p}(0)\mathbf{P}^n$. For SIR dynamics, a maximum of $I_0 + 2S_0$ events can take place, and so the relation

$$\mathbf{p}(\infty) = \mathbf{p}(0)\mathbf{P}^{I_0+2S_0} \tag{2.14}$$

can be used to calculate the final-size distribution. Code for both of these methods is provided in electronic supplementary material, S2.1.

(ii) Bailey's method, with Neuts & Li's implementation

Bailey [7] notes that an integral representation is available for any final size probability:

$$\frac{\mathrm{d}}{\mathrm{d}t}p_{S,0}(t) = \gamma p_{S,1}(t) \quad \Rightarrow \quad p_{S,0}(t) = \gamma \int_0^t p_{S,1}(\tau)\,\mathrm{d}\tau. \tag{2.15}$$

Then by evaluating the Laplace transform of (2.10) at non-negative s, we see that

$$s\mathbf{q}(s) - \mathbf{p}(0) = \mathbf{q}(s)\mathbf{Q}, \quad \text{for} \quad \mathbf{q}(s) := \int_0^\infty \mathrm{e}^{-st}\mathbf{p}(t)\,\mathrm{d}t. \tag{2.16}$$

It is possible to solve for $\mathbf{q}(s)$ algebraically, and thus obtain

$$\Pr(Z = k) = \gamma \lim_{s\downarrow 0} q_{N-k,1}(s), \tag{2.17}$$

as a closed form solution. While Bailey's original text suggests many different forms for the algebraic solution of (2.16), Neuts & Li [32] considered the form most suitable for numerical implementation. We provide `bailey_fs.m` in electronic supplementary material, S2.2.1, as an implementation of their algorithm.

Extension: generalized transmission rates. The method as outlined is efficient due to the sparse, triangular structure of $\mathbf{Q}$ and does not depend sensitively on the actual transition rates. Neuts & Li [32] in fact considered models that had rates that have general functional dependence on S and I (but not, for example, time)

$$\left.\begin{aligned} &(S,I) \to (S-1, I+1) \quad \text{at rate } \lambda_{S,I} \\ \text{and} \quad &(S,I) \to (S, I-1) \quad \text{at rate } \eta_I, \end{aligned}\right\} \tag{2.18}$$

which could be useful in several contexts, and does not significantly alter the algorithmic procedure or the computational time.

Extension: maximum size. Neuts & Li's paper also considered calculation of the distribution for the maximum size of a stochastic epidemic, defined as the maximum value of $I(t)$ over the course of the epidemic and denoted I^*. This quantity would typically be called *peak prevalence* by epidemiologists. Neuts & Li consider implementation of the method of Daniels [33] and provide an iterative procedure for calculation of $\Pr(I^* \geq y)$. We implement this scheme as `NL_Imax.m` of electronic supplementary material, S2.2.2, which can be readily compared with the output of Gillespie or transient Sellke results.

(iii) Path integral/sum method

This approach has recently been used to evaluate the PMF of the number of secondary infections caused by an initially infected individual [34]. The method works by appending to the state of the Markov chain an indicator variable that takes the value one when the most recent transition corresponds to an infection event, and takes the value zero otherwise. In the SIR case, calling the indicator variable J gives the augmented Markov chain

$$\left.\begin{aligned} &(S,I,J)\to(S-1,I+1,1) \quad \text{at rate } \beta SI\\ \text{and}\quad &(S,I,J)\to(S,I-1,0) \quad \text{at rate } \gamma I. \end{aligned}\right\} \tag{2.19}$$

We then consider the jump chain of this modified Markov chain, a discrete-time Markov chain specifying the probabilities of jumping between states at the time of a jump; we label the transition probability matrix of the jump chain $\mathbf{P}$. The total number of infections over the course of an epidemic, Z, is then equal to the number of jumps which result in being in a state where $J=1$ up until there are no infectives remaining in the population.

We can evaluate the distribution of Z using the following result [34]. Let $(X(n), n\in\mathbb{Z}^+)$ be a discrete-time Markov chain taking values in $\mathcal{S}$ with transition probability matrix $\mathbf{P}=(P_{i,j})$. Define

$$\Omega=\sum_n c_{X(n)}, \tag{2.20}$$

where $c:\boldsymbol{\mathcal{S}}\to[0,\infty)$ is called the cost per visit, obeying $c_j=0, j\notin\boldsymbol{\mathcal{C}}$ and $\boldsymbol{\mathcal{C}}\subseteq\boldsymbol{\mathcal{S}}$ is an irreducible transient class. Let $\phi_i(z)=\mathbb{E}[z^{\Omega}|X(0)=i]$. Then $\phi(z)$ is the (maximal) solution to

$$\phi_i(z)=z^{c_i}\sum_{k\in\boldsymbol{\mathcal{S}}}P_{i,k}\phi_k(z), \quad (i\in\boldsymbol{\mathcal{C}}, z\in[0,1]), \tag{2.21}$$

with $\phi_k(z)=1$ for $k\notin\boldsymbol{\mathcal{C}}$. This result is directly derived by conditioning on the state first jumped to.

The result above allows us to evaluate the probability generating function (PGF) $\phi(z)$ of the distribution of Z. To evaluate the PMF, we can numerically invert the PGF considered as a general Laplace transform [35]. However, a more efficient procedure can be obtained by differentiating the equation (2.21) k times with respect to z and evaluating at $z=0$, resulting in a system of linear equations for $\{\phi_i^{(k)}(0), i\in\boldsymbol{\mathcal{C}}\}$. Solving this system and using the relation

$$\Pr(Z=k)=p_k=\frac{1}{k!}\left.\frac{\mathrm{d}^k\phi}{\mathrm{d}z^k}\right|_{z=0}, \tag{2.22}$$

allows computation of the kth element of $\mathbf{p}$. This method was employed in Ross [34], where code was also made available.

In addition to being applicable to the SIR model with exponential infectious period distribution, this method can be applied to any Markovian model, including SIRS models and with any phase-type infectious period distribution; of course, more complicated infectious period distributions and model structures will result in an increase in the size of the matrix $\mathbf{P}$ required, and hence the methodology will eventually become computationally unfeasible.

Extension: hitting times. For a discrete random variable such as the final size, the path sum is most appropriate. For continuous random variables, the path integral method can be used [36]. Using notation as above, we consider the Laplace transform of the total cost (rather than the PGF):

$$\Phi_i(s)=\mathbb{E}\left[\exp\left(-s\int_0^{\infty}c_{X(t)}\,\mathrm{d}t\right)\middle|X(0)=i\right]. \tag{2.23}$$

This is the maximal solution to the equations

$$\sum_{j\in\mathcal{S}}Q_{i,j}\Phi_i(s)=sc_i\Phi_i(s), \quad (i\in\boldsymbol{\mathcal{C}}), \tag{2.24}$$

where $\Phi_k(s)=1$ for $k\notin\boldsymbol{\mathcal{C}}$. For the hitting times of a stochastic epidemic (i.e. the first time at which there are zero infective individuals), the costs are simply $c_i=1$ for $i\in\boldsymbol{\mathcal{C}}$.

While moments of the hitting-time distribution can be simply obtained by differentiation of (2.24), calculation of the probability density at a given point in time requires numerical Laplace transform inversion, which is inherently numerically unstable. We recommend the Euler method of Abate & Whitt [35], with the roundoff error control proposed by Sakurai [37]. Our code for calculation of the Laplace transform is given in electronic supplementary material, S2.3.2.

(iv) Null space method

The null space method is based upon the spectral theory of matrices, in particular as appropriate for finite-state Markov chains and was used by Keeling & Ross [38]. We start with the Kolmogorov forward equations (2.10). Being a square matrix, $\mathbf{Q}$ is always diagonalizable in the sense of Jordan canonical form. The SIR model has $N+1$ absorbing states, and hence $N+1$ repeated eigenvalues of 0. We may write

$$\mathbf{Q} = \mathbf{T}^{-1}\mathbf{J}\mathbf{T}, \tag{2.25}$$

for some $\mathbf{T}$, where $\mathbf{J}$ is block diagonal with each *Jordan block* $\mathbf{J}_i$ of size n_i equal to the algebraic multiplicity (i.e. number of repetitions) of its eigenvalue $\lambda^{(i)}$. As the geometric multiplicity (i.e. the dimension of the null space) is equal to the algebraic multiplicity, each Jordan block $\mathbf{J}_i$ is simply diagonal with $\lambda^{(i)}$ on the diagonal. Perron–Frobenius theory applied to the stochastic (uniformized) matrix $\mathbf{M} = (\mathbf{Q}/q_\mathrm{m}) + \mathbf{I}_{|\mathcal{S}|}$, where $q_\mathrm{m} = \max_i\{-Q_{i,i}\}$, ensures that the eigenvalues lie within a disc in the negative half plane. Throughout this paper, we let $\mathbf{I}_n$ be the identity matrix of dimension n. For the SIR model, $\lambda_1 = \lambda_2 = \cdots = \lambda_{N+1} = 0 > \lambda_{N+2} \geq \mathrm{Re}(\lambda_{N+3}) \geq \cdots \geq \mathrm{Re}(\lambda_{|\mathcal{S}|})$. Now,

$$\mathbf{p}(t) = \mathbf{p}(0)\mathrm{e}^{\mathbf{Q}t} = \mathbf{p}(0)\mathbf{T}^{-1}\mathrm{e}^{\mathbf{J}t}\mathbf{T}, \tag{2.26}$$

since for the ith Jordan block, we have the form

$$\mathrm{e}^{\mathbf{J}_i t} = \mathrm{e}^{\lambda^{(i)}t}\mathbf{I}_{n_i}. \tag{2.27}$$

Hence as $t \to \infty$, we have that each Jordan block tends to a zero matrix with the exception of $\mathrm{e}^{\mathbf{J}_1} = \mathbf{I}_{N+1}$. Hence, we have

$$\lim_{t\to\infty} \mathbf{p}(t) = \mathbf{p}(0)\mathbf{U}_1\mathbf{I}_{N+1}\mathbf{T}_1, \tag{2.28}$$

where $\mathbf{T} = [\mathbf{T}_1; \mathbf{T}_2; \ldots; \mathbf{T}_d]$ and $\mathbf{T}^{-1} = [\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_d]$ with d equal to the number of distinct eigenvalues. We use the semi-colon (;) for vertical concatenation and comma (,) for horizontal concatenation of matrices, following the conventions of MATLAB. Let us write $\mathbf{T}_1 = [\mathbf{v}_1; \mathbf{v}_2; \ldots; \mathbf{v}_{N+1}]$, of dimensions $(N+1) \times |\mathcal{S}|$, with rows the left eigenvectors of $\mathbf{Q}$ (associated with the eigenvalue zero); and also write $\mathbf{U}_1 = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_{N+1}]$, of dimensions $|\mathcal{S}| \times (N+1)$, with columns the right eigenvectors of $\mathbf{Q}$ associated with the eigenvalue zero. Now, considering the left eigenvectors of $\mathbf{Q}$ (associated with the eigenvalue zero), we may set $\mathbf{v}_i = \mathbf{e}_i$, the vector with a one in the ith position and zeros elsewhere. Hence, we have

$$\lim_{t\to\infty} \mathbf{p}(t) = \mathbf{p}(0)[\mathbf{U}_1, \mathbf{0}_{|\mathcal{S}|\times(|\mathcal{S}|-(N+1))}] = [\mathbf{p}(0)\mathbf{U}_1, \mathbf{0}_{1\times(|\mathcal{S}|-(N+1))}], \tag{2.29}$$

where $\mathbf{0}_{d_1\times d_2}$ is the matrix of size $d_1 \times d_2$ consisting entirely of zeros. Now, by noting that if we start in an absorbing state (with probability one) then we must remain in that state, the columns of the matrix $\mathbf{U}_1$ span the null space of $\mathbf{Q}$ so that $\mathbf{v}_i\mathbf{u}_j = \delta_{ij}$. This is essentially a computationally efficient version of the jump-matrix-based equation (2.14), provided we know the null space. We use the `spspaces.m` code of Kowal [39] to compute the null space of $\mathbf{Q}$, and provide code in electronic supplementary material, S2.4, for application of this code to the SIR model.

Extension: maximum size. Markov chain-based methods, such as the null space and path integral can also be used to calculate maximum epidemic sizes by augmentation of the state space:

$$\left.\begin{aligned}&(S,I,I^*)\to(S-1,I+1,I^*) \quad \text{at rate } \beta SI \text{ if } I+1\le I^*,\\ &(S,I,I^*)\to(S-1,I+1,I^*+1) \quad \text{at rate } \beta SI \text{ if } I+1>I^*\\ \text{and}\quad &(S,I,I^*)\to(S,I-1,I^*) \quad \text{at rate } \gamma I.\end{aligned}\right\} \tag{2.30}$$

We do not provide code for this, since the Neuts & Li algorithm is certainly much more efficient, but note that Markov chain-based methods are extremely versatile, at the cost of state space expansion.

(c) Machine-precision, arbitrary infectious period methods

Analytic traction can be gained on the Sellke construction by the derivation of a Wald-type identity. This was done by Ball [40], who obtained a triangular system of linear equations for the probability p_k of observing k additional cases in a population of S_0 initial susceptibles and I_0 initial infectives:

$$\sum_{k=0}^{l}\frac{\binom{l}{k}p_k}{\binom{S_0}{k}(\Phi(\beta(S_0-l)))^{k+I_0}}=1, \quad (l=0,1,\ldots,S_0), \tag{2.31}$$

where Φ is the Laplace transform of the infectious period distribution, so for Markovian dynamics $\Phi(x)=\gamma/(x+\gamma)$. Note that $p_k=\Pr(Z=(k+I_0))$ in our conventions. Clearly, (2.31) can be written in the form

$$\mathbf{Bp}=\mathbf{1}, \tag{2.32}$$

where $\mathbf{1}$ is a column vector with all entries equal to one, and we call $\mathbf{B}=(B_{kl})$ the Ball matrix. We now consider three methods for numerical solution of (2.32).

(i) Direct substitution

The equation (2.32) can be solved directly, since the Ball matrix $\mathbf{B}$ is left triangular

$$p_k=1-\sum_{l<k}B_{kl}p_l. \tag{2.33}$$

Unfortunately, this procedure becomes numerically divergent for large population sizes. The essential reason for this is that the binomial coefficients in (2.31) become extremely large for large l and $k\approx l/2$. Figure 2*a* shows the large mass in the middle of the Ball matrix, which drives this numerical instability.

Extension: multiple-precision arithmetic. Demiris & O'Neill [41] showed how the problem of numerical divergence could be overcome through the use of multiple-precision arithmetic. We provide code in the electronic supplementary material, S3.1, to implement this using MATLAB's `vpa()` function, which can be easily modified to work at standard machine precision. A limitation of this method, however, is the large computational cost involved, since multiple-precision algorithms are typically extremely costly.

(ii) Iterative methods

A large number of numerical methods are available for the solution of equations of type (2.32) [42]. We find that the use of the conjugate gradient method on the equation

$$\mathbf{Ap}=\mathbf{b}, \quad \text{for } \mathbf{A}=\mathbf{B}^{\top}\mathbf{B}, \quad \mathbf{b}=\mathbf{B}^{\top}\mathbf{1}, \tag{2.34}$$

is stable for population sizes up to around 10^2. This problem is illustrated in figure 2*b*, which shows a concentration of the mass of $\mathbf{A}$ in one area of the matrix that is similar to that of $\mathbf{B}$ noted above. In contrast, iterative schemes that can be applied directly to non-symmetric equations as (2.32) such as biconjugate methods (e.g. MATLAB's `bicstab()`) or minimum-residual methods (e.g. MATLAB's `gmres()`) do not out-perform direct substitution.

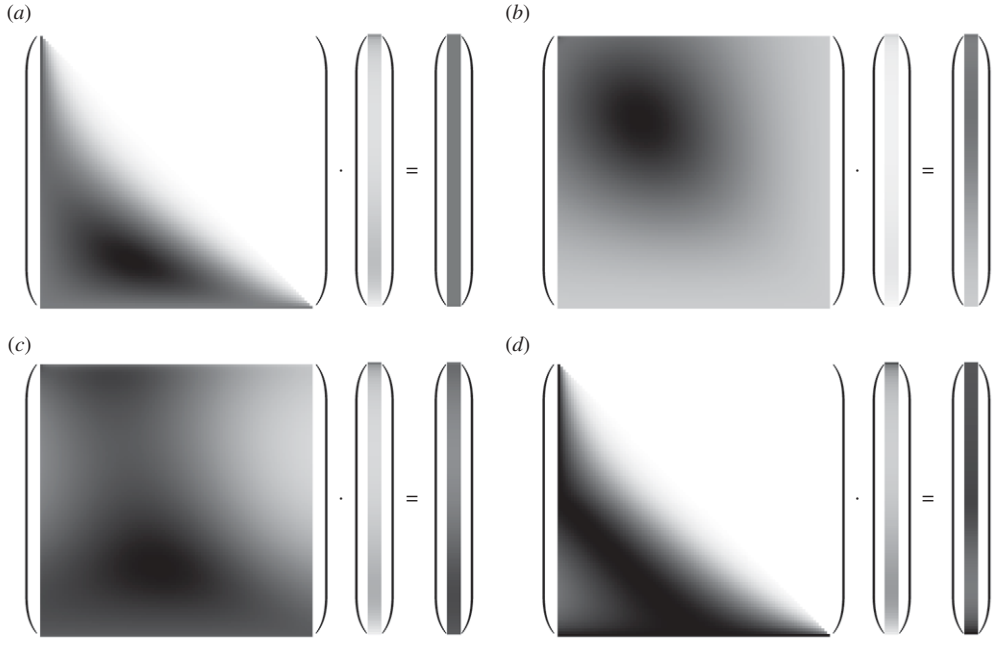

Figure 2. Visualization of the Ball equations in different matrix representations. In each case, shading intensity $\propto$ value$^{1/5}$, but with different constants of proportionality. Parameters are $N = 100$, $\beta = 2/(N-1)$, $I_0 = 1$, $S_0 = N - I_0$, exponential infectious period distribution with unit mean. Matrices are defined in §2*c* of the main paper. (*a*) Raw $\mathbf{Bp} = \mathbf{1}$, (*b*) symmetrized $\mathbf{Ap} = \mathbf{b}$, (*c*) preconditioned $\mathbf{Ep} = \mathbf{c}$ for PCG and (*d*) preconditioned $\mathbf{Gp} = \mathbf{h}$ for GMRES.

Extension: preconditioners. A major benefit of iterative methods is the potential to use preconditioners [42]. We find that the use of a Jacobi preconditioner together with an initial probability vector based on asymptotic results [8,41] can give accurate results even for $N = 10^3$. This preconditioner is formed through the matrix $\mathbf{D} = (D_{kl})$ for

$$D_{kl} = \begin{cases} A_{kl} \text{ if } k = l, \\ 0 \text{ otherwise.} \end{cases} \tag{2.35}$$

Then the preconditioned conjugate gradients method (PCG) effectively solves

$$\mathbf{Ep} = \mathbf{c}, \quad \text{for } \mathbf{E} = \mathbf{D}^{-1}\mathbf{A}, \quad \mathbf{c} = \mathbf{D}^{-1}\mathbf{b}. \tag{2.36}$$

This problem is visualized in figure 2*c*, which shows how this preconditioner evens out the density of mass in the matrix involved. Code is provided in electronic supplementary material, S3.2, to implement this method. There remains the possibility of further preconditioning based on the properties of $\mathbf{E}$ to enhance convergence further without significant computational cost, but we were unable to find an appropriate second preconditioner.

While we could not find a preconditioner that significantly improved biconjugate methods, the minimal residual method given by MATLAB's `gmres()` can be improved through the use of the preconditioner $\mathbf{F} = (F_{kl})$ where

$$F_{kl} = \begin{cases} \max\{B_{km}\}_{m=0}^{k} & \text{if } k = l, \\ 0 \text{ otherwise.} \end{cases} \tag{2.37}$$

This improved GMRES method effectively solves

$$\mathbf{Gp} = \mathbf{h}, \quad \text{for } \mathbf{G} = \mathbf{F}^{-1}\mathbf{B}, \quad \mathbf{h} = \mathbf{F}^{-1}\mathbf{1}, \tag{2.38}$$

and has comparable performance to direct substitution. This problem is visualized in figure 2*d*. As for PCG, the possibility of finding a second preconditioner is still open and could yield a significant improvement over direct substitution if an appropriate one is found.

(iii) Gontcharoff polynomials

Another way of computing the mass function of the final-size distribution is via the expression for the PGF of this distribution given slightly indirectly in theorem 2.6 of Ball [40] (Ball gives the PGF for $N - Z$, the number of susceptibles remaining at the end of the epidemic). We can use the fact that a mass function $(p_k,\ k \in \mathbb{Z}_+)$ can be recovered from its generating function $\phi(z) = \sum_k p_k z^k$ $(z \in [0,1])$ using the relationship (2.22) to find that, for $k = 0, 1, \ldots, n$,

$$\Pr(Z = (k + I_0)) = \sum_{l=0}^{k} \frac{S_0!}{(S_0 - k)!l!} q_{S_0-l}^{I_0+l} G_{l-k}(0|E^{S_0-k}U). \tag{2.39}$$

Here $q_i = \Phi(i\beta)$ $(i = 0, 1, \ldots)$ (the probability that an infective fails to infect any of a given set of i susceptibles), $U = (u_i = q_i,\ i = 0, 1, \ldots)$, $E^jU = (u_{j+i},\ i = 0, 1, \ldots)$ and the Gontcharoff polynomials $G_k(x|U)$ $(k = 0, 1, \ldots)$ are defined by

$$G_0(x|U) = 1$$

and

$$G_k(x|U) = \frac{x^k}{k!} - \sum_{i=0}^{k-1} \frac{u_i^{k-i}}{(k-i)!} G_i(x|U), \quad (k = 1, 2, \ldots).$$

See eqn (3.11) of Picard & Lefèvre [43]. Ball does not use Gontcharoff polynomials in his results, but rather a collection of polynomials $(\alpha_k(s),\ k = 0, 1, \ldots)$ defined as the solution to a triangular system of linear equations. Modulo some scaling, this system of equations is equivalent to the definition of the Gontcharoff polynomials given above: it is not hard to show that $\alpha_k(s) = k!G_k(s)$. Gontcharoff polynomials are useful in formal proofs, and do have something of a probabilistic interpretation when evaluated at 1 [44, §3.1]. It turns out that the scaling $\alpha_k(s) = k!G_k(s)$ can be better for computational purposes and is used in our implementation in electronic supplementary material, S3.3.

Extension: heterogeneous populations and generalized transmission. The approach of Picard & Lefèvre [43] generalizes the work of Ludwig [25] and Bahr & Martin-Löf [45], considering the very general case where 'each infective during [their infectious period] fails to transmit the infection within any given set of susceptibles with a probability depending only on the size of that set' [43: p. 269]. Polynomial-based equations of similar form to (2.39) can therefore be derived for more general models of transmission, and also for heterogeneous populations. These equation sets typically have similar numerical behaviour to each other.

(d) Asymptotic results

There is an extensive literature on asymptotic results for epidemic final sizes in large populations. This goes back to the earliest mathematical representations of the mean behaviour of epidemics [46], and also limiting distributions of simple epidemics [47], but more recently involves formal convergence proofs [45,48] and results obtained for quite complex population structures, including multiple types [49], households [50] and networks [28].

At the heart of most asymptotic results is a transcendental equation for the probability π that an individual avoids 'global' infection (defined in a model-specific way) of the form

$$\pi = F(\pi), \quad \text{for } F : [0,1] \to (0,1]. \tag{2.40}$$

Here, $\pi = 1$ (i.e. asymptotically vanishing levels of global infection) will always be a solution. If $F'(1) \leq 1$, then $\pi = 1$ is the desired result, but if $F'(1) > 1$ then we wish to find the largest $\pi < 1$

that satisfies the equation (2.40). For complex models, there may be many non-maximal solutions, rendering general root-finding algorithms, such as bisection, unreliable. Considering

$$\hat{\pi} = F^m(1-\varepsilon), \tag{2.41}$$

should, however, provide an accurate estimate for π for sufficiently small ε and a sufficiently large number m of iterations of F. In the case of household-structured models, there is then the question of derivation of the distribution of final-size proportions, which can often be done using the methods outlined elsewhere in this paper. Sample code for reproduction of fig. 2 of Ball *et al.* [51] using a Jump chain method and (2.41) is given in electronic supplementary material, S4.

(e) Epidemics on networks

There has been much recent interest in epidemic models where the population is connected on a network [52,53]. In these models, the population is made up of N individuals indexed by integers $i, j, \ldots$ and the contacts from i to j happen at the points of a Poisson process of rate $\beta_{i,j}$. If i is infectious and j susceptible at the point of contact, then j becomes infectious. Recovery happens, as before, after a time drawn from the infectious period distribution. This very general formulation can be somewhat simplified when there are several individuals with the same epidemiological characteristics (i.e. if there is a large discrete symmetry group for $\boldsymbol{\beta} = (\beta_{i,j})$). Alternatively, one may wish to preserve individual identity but make the simplifying assumption that $\boldsymbol{\beta} = \tau\mathbf{G}$, where $\mathbf{G} = (G_{i,j})$ is the adjacency matrix of an undirected, symmetrical, topological network without self-edges. Both simplifications are more commonly considered than the most general case [52,53].

(i) Monte Carlo methods

All Monte Carlo methods discussed so far are relatively simply adapted to populations with network structure. For the case of Gillespie's method, this involves direct simulation of the Markov Chain

$$\left.\begin{aligned} (I_i, S_j) &\to (I_i, I_j) \quad \text{at rate } \beta_{i,j} \\ \text{and} \qquad I_i &\to R \quad \text{at rate } \gamma. \end{aligned}\right\} \tag{2.42}$$

Code to perform direct network epidemic simulation is already available in a form that is readily adapted [6, program 7.7].

For the Sellke method on networks, thresholds $Q_i \sim \exp(1)$ can be generated at the start of the process, however, the ordering of thresholds to obtain an expression similar to (2.4) is not straightforward. Instead, we provide `sel_net.m` in electronic supplementary material, S5.1.1, in which new generations of infectives are created iteratively. This is somewhat similar to Ludwig's approach, but with the difference that the random numbers are generated for the susceptible nodes rather than the infectious contacts produced by infectives.

Ludwig's method is, given its links with random directed networks, particularly natural in the context of epidemics on networks. In this case, the ranks are constructed iteratively as before, but the picking with probability given by (2.7) is applied only to remaining susceptible neighbours of the individual concerned. We provide the function `lud_net.m` in electronic supplementary material, S5.1.2, that implements this algorithm.

(ii) Machine precision methods

There are two distinct meanings to the final-size distribution for network models. The first, as considered by Neal [54], is the final-size PMF averaged over realizations of a given random graph model. This has, to our knowledge, only been done so far for the Bernoulli/Erdös-Rényi random

graph where each link is present with probability π, independently of the presence or the absence of all other links. This work generates a set of equations very similar to (2.31):

$$(\mathbf{Np})_l = \sum_{k=0}^{l} \frac{\binom{l}{k} p_k}{\binom{S_0}{k} q_{S_0-l}^{I_0+k}} = 1, \quad (l = 0, 1, \ldots, S_0), \quad \text{where } q_k = \sum_{l=0}^{k} \binom{k}{l} \pi^l (1-\pi)^{k-l} \Phi(\beta l), \tag{2.43}$$

which can then be solved using similar methods to those discussed for the Ball equations above. Code that creates the matrix $\mathbf{N}$ is provided in electronic supplementary material, S5.2.1.

The second broad class of distributions concerns the probabilities of different outcomes for a given $\boldsymbol{\beta}$—typically the marginal probability for each node that it is ultimately infected. Here there is a 'multitype' formula also derived by Ball [40] that can be adapted. If we have a vector $\mathbf{n} = (n_i)$ whose elements are the initial number of susceptibles of type i, and also $\mathbf{m} = (m_i)$ with elements corresponding to the initial number of infectives of each type, then the relevant equations are

$$\sum_{\mathbf{u}=\mathbf{0}}^{\mathbf{v}} \frac{\binom{\mathbf{n}-\mathbf{u}}{\mathbf{v}-\mathbf{u}} p_{\mathbf{u}}}{\binom{\mathbf{n}}{\mathbf{v}} \prod_{i=1}^{N} (\Phi_i((\boldsymbol{\beta}(\mathbf{n}-\mathbf{v}))_i))^{u_i+m_i}} = 1, \quad \mathbf{0} \le \mathbf{v} \le \mathbf{n}, \tag{2.44}$$

where operations on vectors are defined in the natural way [8,40]. Code to generate the relevant matrix is given in electronic supplementary material, S5.2.2. It is also possible to consider a Markovian model as defined by (2.42). We provide code in electronic supplementary material, S5.2.3, to return an appropriate generator $\mathbf{Q}$ for this Markov chain when $\boldsymbol{\beta} = \tau \mathbf{G}$, which can then be analysed using methods for Markov chains already discussed.

An interesting recent development is the approach of Sharkey [55]. This method makes a 'closure' assumption that has been commonly used in pairwise epidemic network models [53] to derive a set of $O(N^2)$ differential equations describing the temporal behaviour of a network epidemic that is provably exact for loop-less networks (I. Z. Kiss 2012, personal communication). Since the non-network homogeneous models considered are essentially of fully connected cliques this makes the Sharkey model inappropriate for these applications, however, it can be compared with other methods when the network structure is loop-less, and code to evaluate the model is already available [55, appendix D].

3. Results and discussion

We compare and benchmark each method systematically in table 1. These were carried out on a Mac Pro with a dual quad-core 3.2 GHz chipset and 16 GB of RAM running MATLAB V. 7.9, 64-bit version. The machine precision $\epsilon \approx 2.2 \times 10^{-16}$, meaning that numbers close to 1.0 differing by this amount may be evaluated as equal by the machine. For variable-precision algorithms, the minimum number of digits (considering only powers of 2) needed to obtain accurate results is reported. For iterative matrix methods solving $\mathbf{Ax} = \mathbf{b}$, and using the vector norm $|\mathbf{v}| = \sqrt{\mathbf{v} \cdot \mathbf{v}}$, the standard diagnostic 'residual' is recorded, and is equal to $|\mathbf{b} - \mathbf{Ax}|/|\mathbf{b}|$. For machine precision methods, we quote times for evaluation of the entire PMF, while for Monte Carlo methods we quote times per sample. We also display the behaviour of each algorithm as a function of population size graphically in figure 3.

The exact times given in seconds will of course vary significantly between machines, and may be subject to significant modification with more optimized code—in addition, the exact part of the code that should be timed to make a fair comparison is not clear, for example, where sparse matrices need to be generated, this is not included in our time measurement because the sparse structure of these matrices can be stored in a parameter-independent manner. Despite these caveats, however, some strong signatures show up in the benchmarking that can be interpreted and generalized.

Table 1. Comparison of methods discussed in the main text. For simulation methods, times are mean per simulation averaged over 10^3 realizations. For all benchmarks, mean infectious period is 1, $\beta = 3/(N-1)$, $I_0 = 1$ and $S_0 = N - I_0$. For Benchmark I, $N = 10$ and infectious period distribution is exponential; for Benchmark II, $N = 10^3$ and infectious period distribution is exponential; and for Benchmark III, $N = 10^2$ and infectious period distribution is constant. Times are given in seconds; a dash '—' is displayed if the method is unable to return an accurate answer.

method	type	waning?	infectious period	transient?	Benchmark I	Benchmark II	Benchmark III
Gillespie	simulation	yes	phase-type	yes	3.7×10^{-4}	3.4×10^{-2}	—
Sellke	simulation	no	arbitrary	feasible	3.1×10^{-4}	4.5×10^{-4}	7.1×10^{-5}
Ludwig	simulation	no	arbitrary	no	6.2×10^{-4}	2.5×10^{-2}	3.1×10^{-3}
jump matrix	machine precision	yes	phase-type	no	7.3×10^{-5}	—	—
matrix exponential (EXPOKIT)	machine precision	yes	phase-type	yes	1.6×10^{-2}	9.7×10^{3}	—
Bailey	machine precision	no	exponential	feasible	6.3×10^{-5}	4.3×10^{-2}	—
Path sum	machine precision	yes	phase-type	no	5.2×10^{-4}	3.2×10^{2}	—
null space	machine precision	yes	phase-type	no	6.4×10^{-4}	1.9×10^{2}	—
Ball (direct)	machine precision	no	arbitrary	no	5.6×10^{-3}	—	7.6×10^{-1}
Ball (VPA)	arbitrary precision	no	arbitrary	no	1.9 (32 digit)	8.4×10^{5} (256 digit)	1.9×10^{4} (32 digit)
Ball (PCG, Jacobi)	machine precision	no	arbitrary	no	9.7×10^{-3} (residual 1.9×10^{-16})	2.0×10^{2} (residual 3.5×10^{-15})	3.9 (residual 8.6×10^{-16})
Ball (GMRES, preconditioned)	machine precision	no	arbitrary	no	9.2×10^{-3} (residual 2.5×10^{-16})	—	8.0×10^{-1} (residual 3.0×10^{-16})
polynomial (sca-led Goncharoff)	machine precision	no	arbitrary	no	1.4×10^{-3}	—	1.4×10^{-1}

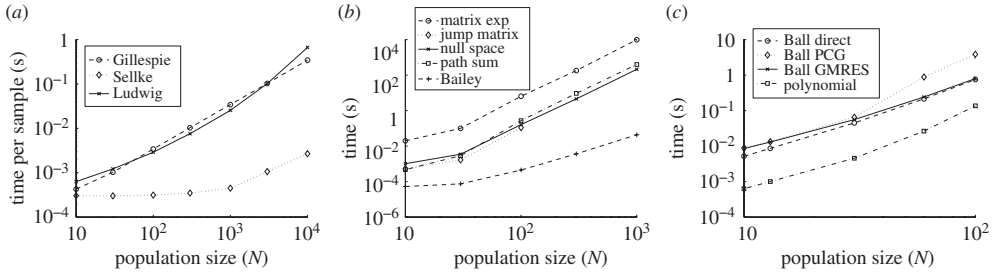

Figure 3. Scaling of time with population size. Parameters are $\beta = 3/(N-1)$ with unit mean (*a*,*b*) exponentially distributed, (*c*) constant, recovery. (*a*) Monte Carlo methods, (*b*) Markov chain methods and (*c*) arbitrary recovery.

(a) Numerical efficiency of Monte Carlo methods

If a Monte Carlo calculation is necessary, then Gillespie's method is typically the least efficient method. The Sellke construction is much faster for both final-size and temporal dynamics (in fact, it is the only approach that can generate temporal dynamics for an arbitrary infectious period distribution). Gillespie's method is, however, the only Monte Carlo approach that can deal straightforwardly with waning immunity. There is also the question of efficiency at different parameter values. For subcritical epidemics where the expected final size is much smaller than the population size, Ludwig's method can involve generation of many fewer pseudo-random numbers than Sellke's, for example.

When considering the usefulness of any Monte Carlo method, it is useful to know how many samples are required. Figure 4 shows three measures of convergence. Suppose the probability of final size z, p_z, and the associated cumulative probability $C_z = \sum_{w=1}^{z} p_z$ are known (in our examples through Bailey's method). Then if a proportion q_z of simulations have final size z and we define the empirical cumulative probability $E_z = \sum_{w=1}^{z} q_z$, the measures used are: (i) the *Kullback–Leibler (KL) divergence*

$$D_{\mathrm{KL}}(q||p) = \sum_{z=1}^{N} q_z \ln\left(\frac{q_z}{p_z}\right), \tag{3.1}$$

using the convention that $0\ln(0) = 0$; (ii) the *Kolmogorov–Smirnov (KS) D-statistic*

$$D_{\mathrm{KS}}(p,q) = \max\{\mathrm{abs}(C_z - E_z)\}_{z=1}^{N}; \tag{3.2}$$

(iii) the *summed absolute error*

$$D_{\mathrm{SAE}}(p,q) = \sum_{z=1}^{N} \mathrm{abs}(p_z - q_z). \tag{3.3}$$

Which of these measures is most relevant depends on the precise application. For example, the KL divergence is often used in estimation, while the KS D-statistic is used to assess model adequacy. Figure 4 indicates that the convergence with number of samples n is, respectively: (i) $O(n^{-1})$; (ii) $O(n^{-1/2})$; (iii) $O(n^{-1/2})$.

While we have considered 'exact' Monte Carlo methods, approximate simulation algorithms also exist, such as the τ-leaping method introduced by Gillespie [56]. This was motivated by the desire to improve the numerical efficiency of the 'Gillespie algorithm', and assumes that the transition rates of the model are held fixed for a period of time τ, implying that the number of events of each type that occur during that time period have independent Poisson distributions. The algorithm is appropriate to use and provides most benefit, when (i) the changes to the state of the system have no or minimal impact on the transition rates of *all* event types, and (ii) it is unlikely for the number of events to lead to a state which is inconsistent with the state space of the model. The τ-leaping methods therefore seem unlikely to provide substantial benefits

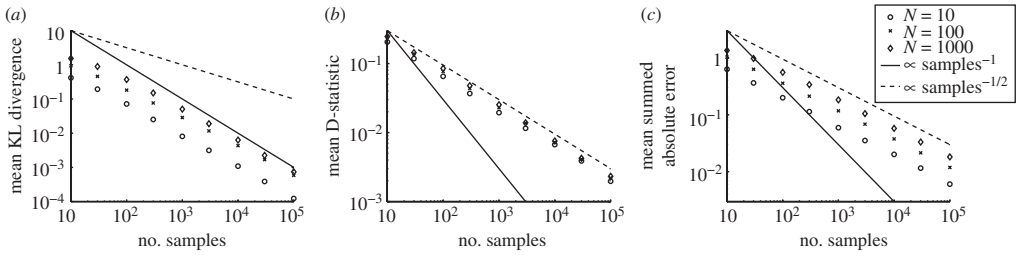

Figure 4. Rate of convergence of sampling from the epidemic final-size distribution with number of samples for three different measures of distribution proximity. Parameters are $\beta = 3/(N-1)$ with unit mean exponentially distributed recovery. (*a*) KL-divergence, (*b*) KS D-statistic and (*c*) summed average error.

for epidemic models: for example, requirement (i) is unlikely to be satisfied for homogeneously mixing models and (ii) will not be satisfied for network models. Nevertheless, a thorough analysis of this question with more sophisticated approximations [57] could be of significant interest.

(b) Numerical efficiency of Markov chain-based methods

For models based on Markov chains, multiple methods are available to calculate various quantities at machine precision. Of these, Bailey's method is the fastest, and is robust even for system sizes over 10^4 with the limitation on system size related to the resources available to store and process a dense $N \times N$ matrix, but does rely on special properties of the SIR model. Of the methods that apply to more general Markov chains, the exact calculation considered will determine the appropriate method: matrix exponentials can capture temporal dynamics; which of the null space and path sum/integral methods is faster is likely to be machine- and implementation-dependent, but both are efficient for calculation of a wide range of quantities. The system size limitations for these approaches are related to the resources available to store and process large sparse matrices, but as for Bailey's method there is no inherent numerical instability involved.

(c) Numerical efficiency of arbitrary infectious period methods

If a complex infectious period distribution is required then it may be impractical to use phase-type infectious distributions and Markovian approaches are unsuitable. Of the machine precision methods this leaves the Ball matrix equations and the use of Goncharoff polynomials. While these approaches are often much lower-dimensional than Markovian models, they are both inherently numerically unstable: the former because some elements of the Ball matrix are much larger than others; and the latter because solution involves summing many different positive and negative terms, leading to 'catastrophic cancellation'. Where possible, however, direct substitution of the Ball equations is numerically efficient; also Gontcharoff polynomials are comparable in numerical efficiency and stability to direct substitution.

When the system size becomes large enough to generate numerical instability, Jacobi-preconditioned conjugate gradients can be used to reach system sizes of 10^3 while retaining numerical efficiency. An important point about all iterative methods, however, is that small negative probabilities can be returned for any value of the PMF that is close to zero. While these do not lead to practical problems or major numerical instability, variable-precision should still be viewed as a gold-standard method, albeit one that can be prohibitively computationally costly. On the other hand, in the context of inference schemes such as those based on random-walk Metropolis–Hastings sampling over parameters θ, if a probability vector $\mathbf{p}(\theta)$ has been evaluated at one step and a change in parameters $\delta\theta$ is proposed, then $\mathbf{p}(\theta)$ can be used as the starting vector for iterative calculation of $\mathbf{p}(\theta + \delta\theta)$, and would be expected to converge more quickly than an initial vector based on asymptotic results if $\delta\theta$ is small.

Table 2. Comparison of methods discussed in the main text for final sizes on the network in figure 5. For simulation methods, times are mean per simulation averaged over 10^4 realizations. For all benchmarks, mean infectious period is 1 and infectious period distribution is exponential. Benchmark I is for a fixed loop-less network as shown in figure 5*c*, with $\tau = 1$. Benchmark II is an average over 20-node $p = 0.2$ Bernoulli graphs with $\tau = 3/((N-1)p)$. Times are given in seconds; a dash '—' is displayed if the method is unable to calculate the benchmark.

method	type	waning?	infectious period	transient?	topology	Benchmark I	Benchmark II
Gillespie	simulation	yes	phase-type	yes	all	3.0×10^{-4}	8.4×10^{-4}
Sellke	simulation	no	arbitrary	feasible	all	2.2×10^{-4}	2.4×10^{-4}
Ludwig	simulation	no	arbitrary	no	all	2.5×10^{-4}	3.1×10^{-4}
Matrix exponential	machine precision	yes	phase-type	yes	any fixed	3.0	—
Sharkey	machine precision	no	exponential	yes	loop-less	3.0×10^{-4}	—
Ball multitype	machine precision	no	arbitrary	no	any fixed	8.4×10^{-4}	—
Neal	machine precision	no	arbitrary	no	Bernoulli	—	3.0×10^{-2}

(d) Network methods

We also performed benchmarking of network methods, as shown in table 2. Here the conclusions are very similar to those for homogeneous models, but we note that for special topologies (averages over Bernoulli graphs or loop-less fixed networks) there are particularly efficient methods available. This is likely to be a general feature: for certain special networks, there will be techniques available that exploit the restricted topology; but in the most general case less efficient but more versatile approaches will be necessary. In this context, we note the efficiency of the Sellke construction, coupled with the possibility of reconstructing temporal dynamics as for the homogeneous case.

(e) Implications for inference

Since one of the main motivations for our study is the relevance of final-size calculations for statistical work, we now consider the implications of our results for estimation of epidemiological parameters. A theoretically attractive approach to inference is via exact evaluation of the likelihood. We have considered here machine-precision methods which facilitate achieving this goal. The inference itself is typically performed within a Bayesian framework, using Markov chain Monte Carlo (MCMC) methods [41,58]. However, all of the machine-precision methods we have considered become practically infeasible to implement as the population size, or the complexity of the epidemic model, increases, owing to growth in the number of equations to be solved and numerical instabilities. We have demonstrated in this work that deployment of appropriate numerical methods such as preconditioning or use of path sums can significantly extend the reach of machine-precision methods.

The MCMC inference framework is very flexible and so difficulties such as infeasible evaluation of the exact likelihood can often be overcome by some form of imputation or data augmentation [58,59]. In the context of final-size data for epidemics, an example is the random graph approach developed by Demiris & O'Neill [60] for stochastic multi-type epidemics in structured populations. Often such approaches are heavily reliant on being able to restrict the search space, or at least to be able to sample over it efficiently, by good choice of prior. This is typically not a trivial task. Another approach which is gaining in popularity is to use simulation-based methods, such as Approximate Bayesian Computation or pseudo-marginal methods [58,61–63]. This approach relies on the ability to simulate data efficiently, and hence our comparison of three Monte Carlo methods will facilitate the use of such simulation-based inference methods.

For small population sizes, there is a variety of methods which may be used to perform inference. The advantage of machine precision methods is that the PMF is computed without Monte Carlo errors, and hence inference based upon such calculations is reliable. However, these

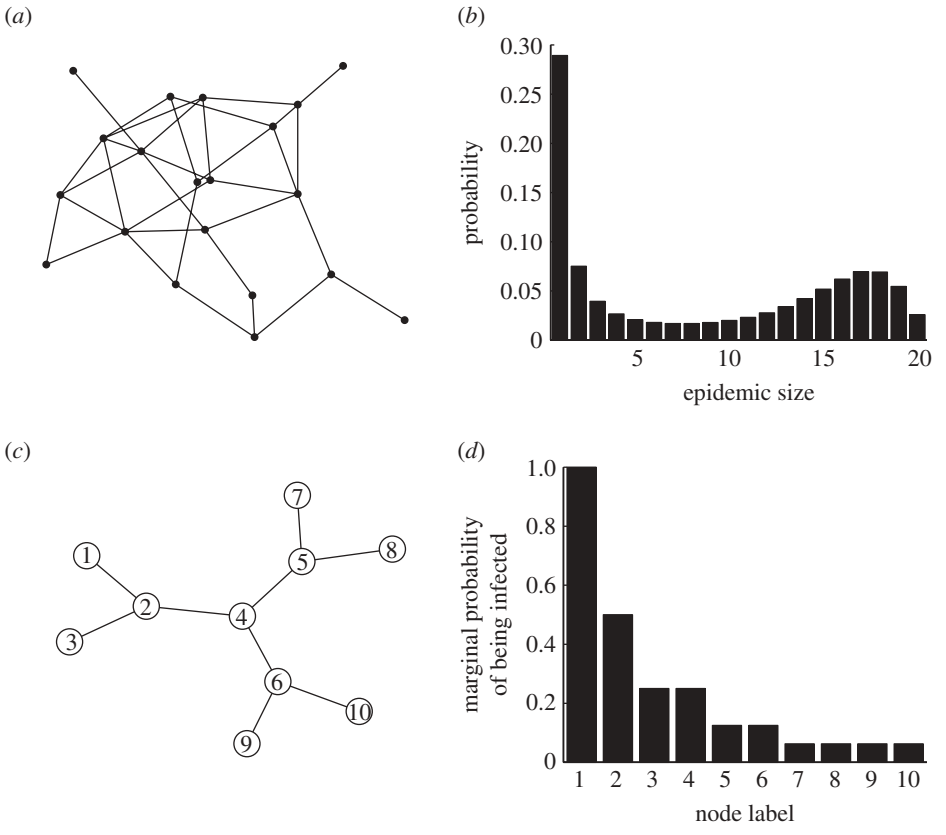

Figure 5. Topologies and epidemic probabilities for network benchmarks. (*a*) A typical Bernoulli random graph with independent probability $p = 0.2$ of each link. (*b*) The final-size probabilities of epidemic sizes averaged over realizations of $p = 0.2$ Bernoulli graphs with $\tau = 3/((N-1)p)$ and unit mean exponentially distributed recovery time. (*c*) Test network used for Benchmark II. (*d*) Marginal probabilities for infection on the test network starting on node 1 with $\tau = 1$ and unit mean recovery time. (*a*) Sample Bernoulli network, (*b*) final-size probabilities, (*c*) test network and (*d*) marginal probabilities.

methods may be time-consuming, and it may be more efficient to evaluate a Monte Carlo estimate. The difficulty is in determining how many simulations are required in order to estimate the mass function to *sufficient* accuracy for the problem in hand. To assist in determining this we have considered three measures of 'distance' between the machine-precision evaluation and the Monte Carlo estimate (figure 4). This informs us that to achieve very close to the same precision as the machine-precision methods, based upon the Benchmark cases running times in tables 1 and 2, that often the machine-precision methods are more computationally favourable. However, if one is happy to forgive some inaccuracy—possibly in bias and confidence—in parameter estimates, then some speed-up is possible using Monte Carlo methods. This trade-off will often be also dependent upon the application, in terms of the total running time required and available. In any case, as the population size increases Monte Carlo estimation methods become more appealing, and eventually become necessary. Additionally, whether the infectious period distribution can be well approximated by a sufficiently low-dimensional phase-type distribution can lead to additional questions about speed and accuracy. Nevertheless, certain general statements as discussed earlier will generally hold: for SIR epidemics, Sellke's method typically outperforms the Gillespie algorithm, while only Markovian models can deal with waning immunity. Temporal data will also restrict the use of Sellke's method to Monte Carlo simulation, while Markovian models can have temporal quantities evaluated at machine precision through the use of, for example, matrix exponentials.

4. Conclusions

Since the work of Bailey [7], much effort on stochastic epidemic models has focused on analytic results to enhance understanding [8]. Modern computational resources, however, mean that there are three particularly strong reasons to consider numerical algorithms. First, there is the possibility of making a fast sweep over a large region of parameter space to aid intuitive understanding. Secondly, there is improving the performance of computationally intensive inference. Thirdly, there is enhancement of the performance of other 'inverse problems' such as optimization of public health intervention strategies.

In this work, we have reviewed a fairly comprehensive selection of the existing methods for generation of the epidemic final-size distribution, and compared their numerical performance. We have shown how Jacobi-preconditioned conjugate gradients can be used to help alleviate the reported limitations of the Ball method; however, we consider it likely that problem-specific preconditioners and other numerical techniques can be developed with the aim of addressing the epidemiological questions above. We would furthermore encourage anyone reading this paper to contribute to such developments.

T.H. is supported by the UK Engineering and Physical Science Research Council. J.V.R. was supported under Australian Research Council's Discovery Projects funding scheme (project no. DP110102893). We are grateful to Lorenzo Pellis and three anonymous referees for helpful comments on this manuscript.

References

1. Anderson RM, May RM. 1991 *Infectious diseases of humans*. Oxford, UK: Oxford University Press.
2. Ferguson N, Keeling M, Edmunds W, Gant R, Grenfell B, Amderson R, Leach S. 2003 Planning for smallpox outbreaks. *Nature* **425**, 681–685. (doi:10.1038/nature02007)
3. Riley S *et al.* 2003 Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. *Science* **300**, 1961–1966. (doi:10.1126/science.1086478)
4. Tildesley MJ, Savill NJ, Shaw DJ, Deardon R, Brooks SP, Woolhouse MEJ, Grenfell BT, Keeling MJ. 2006 Optimal reactive vaccination strategies for a foot-and-mouth outbreak in the UK. *Nature* **440**, 83–86. (doi:10.1038/nature04324)
5. Baguelin M, Hoek AJV, Jit M, Flasche S, White PJ, Edmunds WJ. 2010 Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation. *Vaccine* **28**, 2370–2384. (doi:10.1016/j.vaccine.2010.01.002)
6. Keeling MJ, Rohani P. 2007 *Modeling infectious diseases in humans and animals*. Princeton, NJ: Princeton University Press. See www.modelinginfectiousdiseases.org.
7. Bailey NTJ. 1957 *The mathematical theory of epidemics*. London, UK: Griffin.
8. Andersson H, Britton T. 2000 *Stochastic epidemic models and their statistical analysis*. Lecture Notes in Statistics, no. 151. Berlin, Germany: Springer.
9. Gilks WR, Richardson S, Spiegelhalter DJ (eds) 1995 *Markov Chain Monte Carlo in practice*. London, UK: Chapman and Hall.
10. O'Neill P, Roberts G. 1999 Bayesian inference for partially observed stochastic epidemics. *J. R. Stat. Soc. A* **162**, 121–129. (doi:10.1111/1467-985X.00125)
11. Brooks S, Gelman A, Jones GL, Meng XL (eds) 2011 *Handbook of Markov Chain Monte Carlo*. Boca Raton, FL: CRC Press.
12. Longini IM, Koopman JS, Monto AS, Fox JP. 1982 Estimating household and community transmission parameters for influenza. *Am. J. Epidemiol.* **115**, 736–751.
13. Fraser C, Cummings DAT, Klinkenberg D, Burke DS, Ferguson NM. 2011 Influenza transmission in households during the 1918 pandemic. *Am. J. Epidemiol.* **174**, 505–514. (doi:10.1093/aje/kwr122)
14. Stigler SM. 2002 Stigler's lay of eponomy. In *Statistics on the table: the history of statistical concepts and methods*, ch. 14. Cambridge, MA: Harvard University Press.
15. Gillespie DT. 1977 Exact stochastic simulation of coupled chemical reactions. *J. Phys. Chem.* **81**, 2340–2361. (doi:10.1021/j100540a008)
16. Feller W. 1940 On the integro-differential equations of purely discontinuous Markoff processes. *Trans. Am. Math. Soc.* **48**, 488–515. (doi:10.1090/S0002-9947-1940-0002697-3)

17. Kendall DG. 1950 An artificial realization of a simple 'birth-and-death process. *J. R. Stat. Soc. B* (*Methodol.*) **12**, 116–119.
18. Bartlett MS. 1953 Stochastic processes or the statistics of change. *J. R. Stat. Soc. C* (*Appl. Stat.*) **2**, 44–64. (doi:10.2307/2985327)
19. White LJ, Waris M, Cane PA, Nokes DJ, Medley GF. 2005 The transmission dynamics of groups A and B human respiratory syncytial virus (hRSV) in England & Wales and Finland: seasonality and cross-protection. *Epidemiol. Infect.* **133**, 279–289. (doi:10.1017/S0950268804003450)
20. Atchison C, Lopman B, Edmunds, WJ. 2010 Modelling the seasonality of rotavirus disease and the impact of vaccination in England and Wales. *Vaccine* **28**, 3118–3126. (doi:10.1016/j.vaccine.2010.02.060)
21. Ball FG, Lyne OD. 2002 Optimal vaccination policies for stochastic epidemics among a population of households. *Math. Biosci.* **177 & 178**, 333–354. (doi:10.1016/S0025-5564(01)00095-5)
22. Ross JV, House T, Keeling MJ. 2010 Calculation of disease dynamics in a population of households. *PLoS ONE* **5**, e9666. (doi:10.1371/journal.pone.0009666)
23. Neuts MF. 1975 Probability distributions of phase type. In *Liber amicorum Professor emeritus* (ed. H Florin), pp. 173–206. Leuven, Belgium. Katholieke Universiteit Leuven, Departement Wiskunde.
24. Sellke T. 1983 On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Probab.* **20**, 390–394. (doi:10.2307/3213811)
25. Ludwig D. 1975 Final size distributions for epidemics. *Math. Biosci.* **23** 33–46. (doi:10.1016/0025-5564(75)90119-4)
26. Pellis L, Ferguson N, Fraser C. 2008 The relationship between real-time and discrete-generation models of epidemic spread. *Math. Biosci.* **216**, 63–70. (doi:10.1016/j.mbs.2008.08.009)
27. Ball F, Neal P. 2008 Network epidemic models with two levels of mixing. *Math. Biosci.* **212**, 69–87. (doi:10.1016/j.mbs.2008.01.001)
28. Ball F, Sirl D, Trapman P. 2009 Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Probab.* **41**, 765–796. (doi:10.1239/aap/1253281063)
29. Ball F, Britton T, Sirl D. 2011 Household epidemic models with varying infection response. *J. Math. Biol.* **63**, 309–337. (doi:10.1007/s00285-010-0372-6)
30. Newman M. 2010 *Networks: an introduction*. Oxford, UK: Oxford University Press.
31. Sidje RB. 1998 EXPOKIT. A software package for computing matrix exponentials. *ACM Trans. Math. Softw.* **24**, 130–156. (doi:10.1145/285861.285868)
32. Neuts MF, Li J. 1996 An algorithmic study of S-I-R stochastic epidemic models. In *Athens Conference on Applied Probability and Time Series Analysis: Applied probability, in honor of JM Gani* (eds RPCC Heyde, YV Prohorov, ST Rachev). Lecture Notes in Statistics, pp. 295–306. Berlin, Germany: Springer.
33. Daniels HE. 1974 The maximum size of a closed epidemic. *Adv. Appl. Probab.* **6**, 607–621. (doi:10.2307/1426182)
34. Ross JV. 2011 Invasion of infectious diseases in finite homogeneous populations. *J. Theor. Biol.* **289**, 83–89. (doi:10.1016/j.jtbi.2011.08.035)
35. Abate J, Whitt W. 1992 Numerical inversion of probability generating functions. *Oper. Res. Lett.* **12**, 245–251. (doi:10.1016/0167-6377(92)90050-D)
36. Pollett PK, Stefanov VE. 2002 Path integrals for continuous-time Markov chains. *J. Appl. Probab.* **39**, 901–904. (doi:10.1239/jap/1037816029)
37. Sakurai T. 2003 Computational techniques for solving stochastic models. PhD thesis, Department of Electrical and Electronic Engineering, The University of Melbourne, Australia.
38. Keeling MJ, Ross JV. 2008 On methods for studying stochastic disease dynamics. *J. R. Soc. Interface* **5**, 171–181. (doi:10.1098/rsif.2007.1106)
39. Kowal P. 2006 Null space of a sparse matrix. MATLAB File Exchange. See http://www.mathworks.fr/matlabcentral/fileexchange/11120-null-space-of-a-spa-rse-matrix.
40. Ball F. 1986 A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemics models. *Adv. Appl. Probab.* **18**, 289–310. (doi:10.2307/1427301)
41. Demiris N, O'Neill PD. 2006 Computation of final outcome probabilities for the generalised stochastic epidemic. *Stat. Comput.* **16**, 309–317. (doi:10.1007/s11222-006-8320-4)

42. Golub GH, van Loan CF. 1996 *Matrix computations*, 3rd edn. Baltimore, MD: Johns Hopkins University Press.
43. Picard P, Lefèvre C. 1990 A unified analysis of the final size and severity distribution in collective Reed–Frost epidemic processes. *Adv. Appl. Probab.* **22**, 269–294. (doi:10.2307/1427536)
44. Ball FG. 2000 Susceptibility sets and the final outcome of stochastic SIR epidemic models. Research Report 00-09, Division of Statistics, School of Mathematical Sciences, University of Nottingham, UK.
45. Bahr BV, Martin-Löf A. 1980 Limit theorems for some epidemic processes. *Adv. Appl. Probab.* **12**, 319–349. (doi:10.2307/1426600)
46. Kermack W, McKendrick A. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. Lond. A* **115**, 700–721. (doi:10.1098/rspa.1927.0118)
47. Daniels HE. 1967 The distribution of the total size of an epidemic. In *Proc. of the Fifth Berkeley Symp. on Mathematical Statistics and Probability, Berkeley, 21 June–18 July 1965 and December 1965–7 January 1966*, vol. 4, pp. 281–293. Berkeley, CA: University of California Press.
48. Scalia-Tomba G. 1990 On the asymptotic final size distribution of epidemics in heterogeneous populations. In *Stochastic processes in epidemic theory* (eds JP Gabriel, C Lefèvre, P Picard). Lecture Notes in Biomathematics, no. 86, pp. 189–196. Berlin, Germany: Springer.
49. Ball F, Clancy D. 1993 The final size and severity of a generalised stochastic multitype epidemic model. *Adv. Appl. Probab.* **25**, 721–736. (doi:10.2307/1427788)
50. Neal P. 2005 Compound poisson limits for household epidemics. *J. Appl. Probab.* **42**, 334–345. (doi:10.1239/jap/1118777174)
51. Ball F, Mollison D, Scalia-Tomba G. 1997 Epidemics with two levels of mixing. *Ann. Appl. Probab.* **7**, 46–89. (doi:10.1214/aoap/1034625252)
52. Bansal S, Grenfell BT, Meyers LA. 2007 When individual behaviour matters: homogeneous and network models in epidemiology. *J. R. Soc. Interface* **4**, 879–891. (doi:10.1098/rsif.2007.1100)
53. Danon L, Ford AP, House T, Jewell CP, Keeling MJ, Roberts GO, Ross JV, Vernon MC. 2011 Networks and the epidemiology of infectious disease. *Interdiscip. Perspect. Infect. Dis.* **2011**, 1–28. (doi:10.1155/2011/284909)
54. Neal P. 2003 SIR epidemics on a Bernoulli random graph. *J. Appl. Probab.* **40**, 779–782. (doi:10.1239/jap/1059060902)
55. Sharkey KJ. 2011 Deterministic epidemic models on contact networks: correlations and unbiological terms. *Theor. Popul. Biol.* **79**, 115–129. (doi:10.1016/j.tpb.2011.01.004)
56. Gillespie DT. 2001 Approximate accelerated stochastic simulation of chemically reacting systems. *J. Chem. Phys.* **115**, 1716–1733. (doi:10.1063/1.1378322)
57. Anderson DF, Ganguly A, Kurtz TG. 2011 Error analysis of tau-leap simulation methods. *Ann. Appl. Probab.* **21**, 2226–2262. (doi:10.1214/10-AAP756)
58. O'Neill PD, Balding DJ, Becker NG, Eerola M, Mollison D. 2000 Analyses of infectious disease data from household outbreaks by Markov chain Monte Carlo methods. *J. R. Stat. Soc. C* (*Appl. Stat.*) **49**, 517–542. (doi:10.1111/1467-9876.00210)
59. O'Neill PD. 2009 Bayesian inference for stochastic multitype epidemics in structured populations using sample data. *Biostatistics* **10**, 779–791. (doi:10.1093/biostatistics/kxp031)
60. Demiris N, O'Neill PD. 2005 Bayesian inference for stochastic multitype epidemics in structured populations via random graphs. *J. R. Stat. Soc. B* **67**, 731–745. (doi:10.1111/j.1467-9868.2005.00524.x)
61. Baguelin M, Newton JR, Demiris N, Daly J, Mumford JA, Wood JLN. 2010 Control of equine influenza: scenario testing using a realistic metapopulation model of spread. *J. R. Soc. Interface* **7**, 67–79. (doi:10.1098/rsif.2009.0030)
62. Neal P. 2010 Efficient likelihood-free Bayesian computation for household epidemics. *Stat. Comput.* 1–18. (doi:10.1007/s11222-010-9216-x)
63. McKinley TJ, Ross JV, Deardon R, Cook AR. Submitted. Simulation-based Bayesian inference for epidemic models.