The probability of epidemic fade-out is non-monotonic in transmission rate for the Markovian SIR model with demography

P. G. Ballard[1,*], N. G. Bean[2], J. V. Ross[3]

Abstract

Epidemic fade-out refers to infection elimination in the trough between the first and second waves of an outbreak. The number of infectious individuals drops to a relatively low level between these waves of infection, and if elimination does not occur at this stage, then the disease is likely to become endemic. For this reason, it appears to be an ideal target for control efforts. Despite this obvious public health importance, the probability of epidemic fade-out is not well understood. Here we present new algorithms for approximating the probability of epidemic fade-out for the Markovian SIR model with demography. These algorithms are more accurate than previously published formulae, and one of them scales well to large population sizes. This method allows us to investigate the probability of epidemic fade-out as a function of the effective transmission rate, recovery rate, population turnover rate, and population size. We identify an interesting feature: the probability of epidemic fade-out is very often greatest when the basic reproduction number, R_0, is approximately 2 (restricting consideration to cases where a major outbreak is possible, i.e., $R_0 > 1$). The public health implication is that there may be instances where a non-lethal infection should be allowed to spread, or antiviral usage should be moderated, to maximise the chance of the infection being eliminated before it becomes endemic.

Keywords: diffusion approximation, efficient algorithms, epidemic control, stochastic epidemic model

*Corresponding author

[1]School of Mathematical Sciences, The University of Adelaide, Adelaide SA 5005, AUSTRALIA. peter.ballard@adelaide.edu.au

[2]School of Mathematical Sciences, and ARC Centre of Excellence for Mathematical and Statistical Frontiers, The University of Adelaide, Adelaide SA 5005, AUSTRALIA. nigel.bean@adelaide.edu.au

[3]School of Mathematical Sciences, The University of Adelaide, Adelaide SA 5005, AUSTRALIA. joshua.ross@adelaide.edu.au

1. Introduction

The ultimate goal of modelling infectious disease dynamics is to gain insight into how to use resources best to eliminate infection. This may be achieved by making invasion difficult through minimising the probability of a major outbreak, for example through the use of prophylactic vaccination, antivirals or contact tracing [3, 4, 23].

For endemic diseases, with wide prevalence, once again the predominant focus is on reducing transmission as much as possible, and there have been a number of studies calculating the mean time to *endemic fade-out* [26, 20, 13].

Much less attention has been paid to what is the optimal approach to adopt when a major outbreak occurs. Typically, focus has been given to minimising the amount of infection – either the rate of new infections, or the total number of infections over the first wave of an outbreak – for example, through the use of antivirals, and once available, vaccination (e.g., [17, 7]). Here we instead focus on the probability of *epidemic fade-out* - that is, the probability of infection being eliminated between the first and second waves of infection.

In fact, a more comprehensive understanding of the probability of epidemic fade-out is named as one of the five challenges (for stochastic epidemic models involving global transmission) by Britton *et al.* [8], supporting earlier calls [1, 10]. The interest in this quantity for infection elimination is that following the first wave of an outbreak, the number of infectious individuals drops to a relatively low level. Then, if fade-out does not occur, it is likely that the disease will become endemic. Hence, this "*first trough*" of infection appears intuitively to be an ideal target for elimination.

We study a Markovian SIR model with demography ([26, 19, 2]), and in particular the probability of epidemic fade-out as a function of the effective transmission rate, recovery rate, population turnover rate, and population size parameter. We identify the ubiquity of a non-monotonicity property of the probability of epidemic fade-out as a function of the effective transmission rate (holding other parameters fixed). In fact, the probability of epidemic fade-out is very often greatest when the basic reproduction number, R_0, is approximately 2 (restricting consideration to cases where a major outbreak is possible, i.e., $R_0 > 1$). This means that there may be cases when, faced with an infectious outbreak, it would be beneficial to not take action to reduce R_0.

The identification of this phenomenon was achieved through the development of a numerical method which is highly accurate and efficient for computation of the probability of epidemic fade-out. To our knowledge, as supported by the paper [8], there have been only two existing methods proposed, both approximations, for evaluating this probability [25, 18]. These existing methods are asymptotic approximations, with accuracy improving in the limit as the population size parameter tends to infinity. Our method has the benefit of being highly accurate across a wider range of population sizes, including moderate-sized populations, whilst still using light computer resources and hence scaling well to large population sizes.

In the next section we introduce the Markovian SIR model with demography

that we study, before discussing deterministic and diffusion approximations of this model which are relevant to existing methods and our new method for evaluating the probability of epidemic fade-out. We then review the existing approximations. In Section 3 we detail our new method for computing the probability of epidemic fade-out. In Section 4.1 we validate its accuracy and efficiency, and in Section 4.2 we investigate the dependence of the probability of epidemic fade-out on the model parameters, identifying the ubiquity of a non-monotonicity property in the effective transmission rate. Finally, we conclude this work and discuss future research ideas.

2. Background

In this section we present the two existing methods for approximating the probability of epidemic fade-out [25, 18]. To achieve this, we first introduce the underlying model assumed in these earlier studies, and also two asymptotic approximations of this model. These are not only required for both existing methods but also for our new methods to be presented in Section 3.

2.1. The Markovian SIR model with demography

Following previous work [25, 18], we adopt the Markovian SIR model with demography [26, 19, 2]. However, we note that our methods can be easily modified to suit other SIR models which involve replenishment of susceptibles.

The well known SIR model puts every individual in the population into one of three classes: "S" for Susceptible, "I" for Infectious, and "R" for Recovered (or Removed). Let S, I and R denote the number of individuals in the respective states. Then, we assume that susceptible individuals become infectious at rate $\beta SI/N$, and infectious individuals recover at rate γI, where β is the effective transmission rate parameter, $1/\gamma$ is the average infectious period of an individual and N is the total population size. The population is closed, and hence of a constant size.

The SIR model with demography extends the SIR model by also having births (or immigration) of susceptibles, at a fixed rate μN, and deaths (or emigration) from each state at rates μS, μI and μR respectively, where μ is the population turnover rate parameter. We note that this means the actual population size, $S + I + R$, is no longer fixed, but that the birth rate is held constant (i.e., N, the population size parameter, is constant). A consequence of the latter, along with the fact that the number of recovered individuals, R, has no direct bearing on the other states, and that our interest herein is on the number of infectious individuals, is that we may describe the state of the system by (S, I) [13] with state space $\{(S, I) : 0 \leq S, I\}$. The Markovian SIR model with demography we consider herein is detailed in Table 1 and Figure 1.

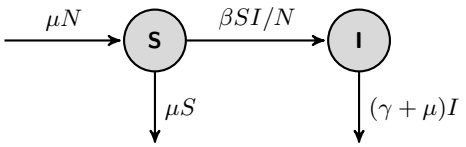


Figure 1: Diagram of the Markovian SIR model with demography. Note, the "R" state is redundant and has been removed.

Description	Transition	Rate
Infection	$(S, I) \to (S - 1, I + 1)$	$\beta SI/N$
Birth of Susceptible	$(S, I) \to (S + 1, I)$	μN
Removal of Susceptible	$(S, I) \to (S - 1, I)$	μS
Removal of Infectious	$(S, I) \to (S, I - 1)$	$(\gamma + \mu)I$

Table 1: Events, transitions and their rates for the Markovian SIR model with demography.

2.2. Asymptotic approximations: The density process

We now state two limiting results of the SIR model with demography, in the limit as N becomes large [15, 16, 22]. These approximations assist us in defining the *probability of epidemic fade-out*, and are furthermore made use of in the two existing methods for approximating the probability of epidemic fade-out, discussed in Section 2.3, and in our own methods to be introduced in Section 3.

Let $Y_N(t)$ be a process following the model defined in Section 2.1, with each value being an (S, I) pair, and with initial value (S_0, I_0). The associated *density process* is $X_N(t) = Y_N(t)/N$, with each possible value x being an (s, i) pair, where $s = S/N$ and $i = I/N$; and the initial value is $x_0 = (s_0, i_0) = (S_0/N, I_0/N)$. The density process is important because it allows us to analyse the limiting behaviour as $N \to \infty$.

Let $f(x, l)$ be the transition rate of the density process from state (x) to state $(x + l/N)$, where l can take on the possible 1-step transition values in Table 1: $(-1, 1)$, $(1, 0)$, $(-1, 0)$ and $(0, -1)$, respectively. Also define for the density process:

$$F(x) = \sum_l l f(x, l) = (-\beta si + \mu(1 - s), \quad \beta si - (\gamma + \mu)i)\,; \tag{1}$$

$B(x)$, a matrix whose $(j, k)^{th}$ element is given by $b_{j,k} = \dfrac{\partial f_j}{\partial x_k}$,

$$\Rightarrow B(x) = \begin{pmatrix} -\beta i - \mu & -\beta s \\ \beta i & \beta s - (\gamma + \mu) \end{pmatrix}; \tag{2}$$

and $G(x)$, a matrix whose $(j,k)^{th}$ element is given by $g_{j,k} = \sum_l \; l_j \; l_k \; f(x,l)$,

$$\Rightarrow G(x) = \begin{pmatrix} \beta si + \mu(1+s) & -\beta si \\ -\beta si & \beta si + (\gamma+\mu)i \end{pmatrix}. \tag{3}$$

Then by Theorem 3.1 of Kurtz [15] and Theorem 3.2 of Pollett [22], we have: *In the limit as $N \to \infty$, $X_N(t)$ weakly converges to a process which at time t is Gaussian with mean $X(t)$ and covariance $\Sigma(t)/N$; where $X(t)$ and $\Sigma(t)$ are the solutions to:*

$$\frac{dX(t)}{dt} = F(X(t)), \quad X(0) = (s_0, i_0); \tag{4}$$

$$\frac{d\Sigma(t)}{dt} = B(X(t))\Sigma(t) + \Sigma(t)B(X(t))^T + G(X(t)), \quad \Sigma(0) = \mathbf{0}. \tag{5}$$

The solution to (4) when multiplied by N, $NX(t)$, is also known as the *deterministic approximation* to $Y_N(t)$. A typical solution of this approximation is shown in Figure 2. The basic reproduction number corresponding to this approximation is,

$$R_0 = \beta/(\gamma+\mu) \quad . \tag{6}$$

Assuming $R_0 > 1$, the process begins at a point A, with a large number of susceptible individuals and a small number of infectious individuals (typically $A = (S,I) = (N-1,1)$). The outbreak rises through point B to point C, falls through point D to point E, and then rises again to point F. The cycle then repeats, on a smaller scale, as it spirals in towards the endemic value (S_e, I_e). Solving (1) for $F(x) = (0,0)$ and scaling by N gives

$$(S_e, I_e) = N\left(\frac{(\gamma+\mu)}{\beta}, \frac{\mu(\beta-\gamma-\mu)}{\beta(\gamma+\mu)}\right). \tag{7}$$

The maximum C and the minimum E of the deterministic infection curve occur when $S = S_e$, and we have chosen the points B, D and F to be at $I = I_e$.

2.3. Previous methods

The two main previous papers on this topic are by van Herwaarden [25] and Meerson and Sasorov [18]. Both of these methods share in common the first part of the analysis: they assume the outbreak follows the trajectory of the deterministic approximation until I falls below I_e (i.e., until point D in Figure 2), and then set up a two boundary hitting probability problem. There is the natural lower absorbing boundary at $I = 0$, and the probability of hitting this boundary before the upper boundary is used to approximate the probability of epidemic fade-out, here known as p_0. The choice of the upper absorbing boundary and method of solution of the two boundary hitting probability problem distinguishes the two methods.

van Herwaarden [25] chooses the upper absorbing boundary to be $I = I_e$. He then approximates p_0 as the probability of hitting the $I = 0$ boundary before the upper boundary. To do this, van Herwaarden assumes S and I are continuous, then simplifies analysis by using the Fokker-Planck equation,

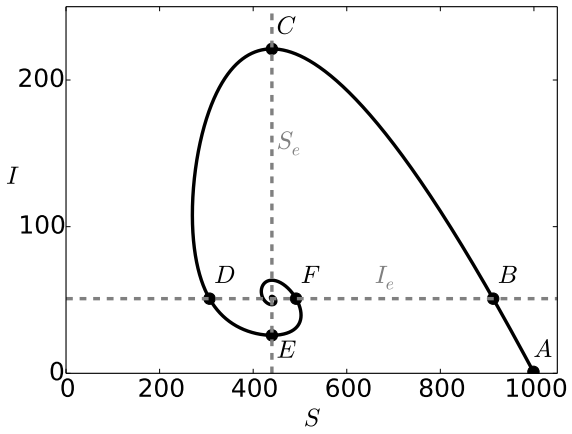


Figure 2: A deterministic trace with $N = 1000$, $\beta = 2.5$, $\gamma = 1$, $\mu = 0.1$, and initial point $(999, 1)$, with endemic values S_e and I_e (dashed lines). The outbreak starts at A, goes through points B, C, D and E through to F, and converges on (S_e, I_e). The *first trough* is between points D and F.

effectively assuming the diffusion approximation as presented in Section 2.2. He uses boundary layer analysis to obtain an approximation to the probability of absorption at the lower boundary.

This results in the following approximation of p_0 (where W_0 is the principal branch of the Lambert W function, and Γ is the gamma function), assuming $1/\sqrt{N} \ll \mu \ll 1$:

$$
\begin{aligned}
x_{1A} &= (-\gamma/\beta) W_0\left((-\beta/\gamma)\exp(-\beta/\gamma)\right), \\
C_3 &= -\ln\left(\frac{-\beta x_{1A}}{\beta x_{1A} - \gamma}\right) - \int_{x_{1A}}^{1} \left(\frac{x_{1A}}{1 - x_{1A}} \frac{\gamma\left(s - s\,\ln(s) - 1\right)}{\beta s^2\left(1 - s + (\gamma/\beta)\ln(s)\right)} + \frac{1}{s - x_{1A}}\right) \mathrm{ds} \quad , \\
K &= \frac{1}{\mu}\exp\left(\frac{\beta x_{1A} + (\beta - \gamma)\ln(1 - x_{1A})}{\mu} + C_3\right) \quad , \\
p_0 &= \exp\left(\frac{-KN\mu^2\left(\beta/\mu\right)^{(\beta-\gamma-\mu)/\mu}\exp\left(-\beta/\mu\right)}{(\gamma + \mu)\,\Gamma\left((\beta - \gamma - \mu)/\mu\right)}\right) \quad .
\end{aligned}
\tag{8}
$$

Meerson and Sasorov [18] instead use a slightly different upper absorbing boundary, namely a diagonal line from (S_e, I_e) to $(N, 0)$, and employ the WKB (Wentzel-Kramers-Brillouin) expansion method [6] in place of the Fokker-Planck equation to solve the two boundary problem. This results in the following approximation of p_0:

$$
\begin{aligned}
&K = \beta/\mu \quad , \\
&\delta = 1 - (\gamma+\mu)/\beta \quad , \\
&x_m = (-(\gamma+\mu)/\beta) W_0\left((-\beta/(\gamma+\mu))\exp(-\beta/(\gamma+\mu))\right) - 1, \\
&Q_1 = \int_0^{x_m} \left(\frac{s\,(s+\delta)}{(1+s)^2\,(s-(1-\delta)\ln(1+s))} - \frac{x_m}{(1+x_m)\,(s-x_m)} \right) ds \quad , \\
&y_m = \frac{(\delta+x_m)\,x_m}{1+x_m} \left(\frac{-x_m}{\delta} \right)^{K\delta} \exp\left(K\,(x_m+\delta) - \left(1+x_m^{-1}\right) Q_1 \right) \quad , \\
&C = \frac{y_m \delta}{2\pi\,(1-\delta)} \quad , \\
&S_0 = C\sqrt{\frac{2\pi}{K\delta}} \quad , \\
&p_0 = \exp\left(-N S_0\right) \quad .
\end{aligned}
\tag{9}
$$

The analysis assumes $NS_0 \gg 1$, and hence p_0 close to 0; however it turns out to be quite accurate in nearly all cases.

We assess the accuracy of these approximations in Section 4.1. However, we note that a nice property of these methods is that they give explicit mathematical expressions for p_0, with negligible computing time.

3. New algorithms

We now consider new algorithms for approximating the probability of epidemic fade-out. These methods are based upon reducing the dependency upon asymptotic approximations, yet still seeking to retain computational efficiency.

Similar to the existing methods, we decompose the problem into two parts. We first consider the state of the process upon its first entrance to the first trough. In place of the deterministic approximation used in the earlier work, we calculate an approximate distribution of the process based upon the diffusion approximation as presented in Section 2.2; further details are presented below in Section 3.2.

In the second part of our methods, we calculate the probability of reaching $I = 0$ before exiting the first trough. In this region, where I is relatively small and stochastic effects are important, we use discrete-state stochastic models. We present two alternative ways to do this: an exact computation in Section 3.3.1, and an efficient approximation in Section 3.3.2. This theoretically should further improve upon previous methods, which used asymptotic approximations (diffusion or WKB) for this calculation. In our discrete-state, stochastic representation of the system, the definition of the first trough can be unclear; for this reason, we commence by providing a precise definition of epidemic fade-out, which we adopt in our methods.

3.1. Definition of epidemic fade-out

Since we are dealing with a discrete system, we round up the endemic fixed point to the next highest integer pair, i.e. $S_d = \lceil S_e \rceil$ and $I_d = \lceil I_e \rceil$. Let $T = \{(S, I) : S < S_d, I = I_d\}$, i.e. the set of states on the dotted (green) line in Figure 3. We define the entrance to the first trough as the point when the system first enters T.

It is now tempting to define the end of the first trough as the point when the system next reaches $I > I_d$. However, due to stochastic effects, it is possible for the system to immediately jump up to $I > I_d$, but this certainly does not mean the end of the first trough. Therefore we need to set the upper boundary to a value greater than I_d in the $S < S_d$ region. So long as it is sufficiently far from I_d to avoid small fluctuations, this value is not critical. Hence $2I_d$ was chosen because it means that T is an equal distance from each absorbing I boundary.

So we define the end of the first trough to be when the system reaches either $I = 2I_d$, or both $S \geq S_d$ and $I \geq I_d$; i.e., the dashed (red) lines in Figure 3.

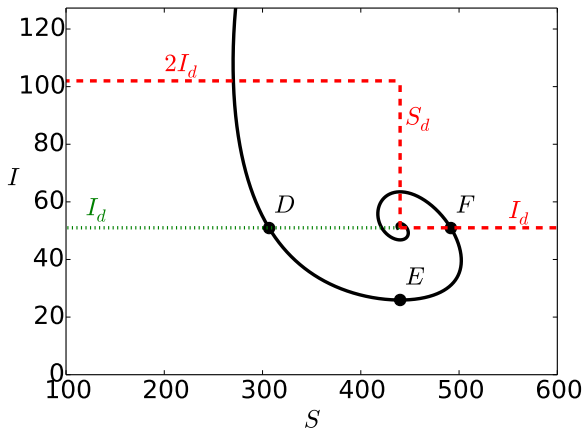


Figure 3: Deterministic trace with the same parameters, and same points D, E and F, as Figure 2. States in T, which denotes the start of the first trough, are shown by the dotted (green) line. The end of the first trough is shown by the dashed (red) line.

We then define p_0, the probability of epidemic fade-out, as the probability that the process is absorbed at $I = 0$ before it exits the first trough (that is, before it reaches the dashed red line), given that the process reaches the start of the first trough.

3.2. Part I: The distribution upon first entrance

In the first part of our method we approximate the distribution of the process upon its first entrance to T, i.e., the distribution of the process upon first

reaching the green dotted line in Figure 3. Results of Ethier and Kurtz [11] provide this distribution, corresponding to the diffusion approximation presented in Section 2.2.

Let τ be the time at which $X(t)$ first enters T, i.e. $\tau = \min\{t \geq 0 : X(t) \in T\}$. We may then approximate the "hitting distribution" of $X(t)$ when it first enters T, as now described.

If we use the subscript j to denote the j^{th} element of a vector, and subscript j,k to denote the row j, column k element of a matrix; then by applying Theorem 11.4.1 of Ethier and Kurtz [11] we have: *In the limit as $N \to \infty$, the distribution of the density process $X_N(t)$ when it first enters T is Gaussian, with mean $X(\tau)_1$ and variance*

$$H = \left(\Sigma(\tau)_{1,1} + \left(\frac{F(X(\tau))_1}{F(X(\tau))_2} \right)^2 \Sigma(\tau)_{2,2} - 2 \left(\frac{F(X(\tau))_1}{F(X(\tau))_2} \right) \Sigma(\tau)_{1,2} \right) /N \quad ; \tag{10}$$

where $F(x)$, $X(t)$ and $\Sigma(t)$ are as defined in equations (1), (4) and (5). We start the diffusion at point B in Figure 2 (that is, $\Sigma(t) = \mathbf{0}$ at point B) because we condition on a major outbreak occuring.

In our methods, we approximate the hitting distribution of S in T by discretising a Gaussian distribution with mean $NX(\tau)_1$ and variance N^2H, and renormalising such that $S \geq 0$ and $S < S_d$ (because $S \geq S_d$ corresponds to sample paths which never meet the criteria of entering the first trough). We call this discrete, renormalised distribution Δ, and it is the initial distribution for the calculations in Sections 3.3.1 and 3.3.2.

3.3. Part II: Modelling the behaviour within the first trough

For the second part of the computation we define the first trough Markov chain, with states arranged as in Figure 4. Each state is represented by an (S, I) pair. Column S has states $(S, 0)$ to $(S, 2I_d)$ for $S \leq S_d$, and has states $(S, 0)$ to (S, I_d) for $S > S_d$. There are two absorbing boundaries, representing the two possible outcomes: The lower absorbing boundary corresponds to epidemic fade-out occurring, and consists of the states $(S, 0)$ for all S; the upper absorbing boundary corresponds to epidemic fade-out not occurring, and consists of the states $(S, 2I_d)$ for $S < S_d$, (S, I_d) for $S > S_d$, and (S, n), for all $n \in [I_d, 2I_d]$, for $S = S_d$.

3.3.1. Exact model

We can evaluate p_0 exactly (for a given starting distribution, Δ). One way to solve this, using standard techniques [21], would be to simultaneously solve equations for all the approximately $(N + S_d) \times I_d$ points in the first trough. However we may take advantage of the fact that almost invariably $2I_d \ll N$, and solve it more efficiently by evaluating a column at a time, as now described.

For each column S, create a stochastic transition matrix, A_S, of the first exit from column S to its neighbouring columns. We partition A_S into two matrices F_S and B_S – such that $A_S = [F_S \; B_S]$ – which represent the first exit into the

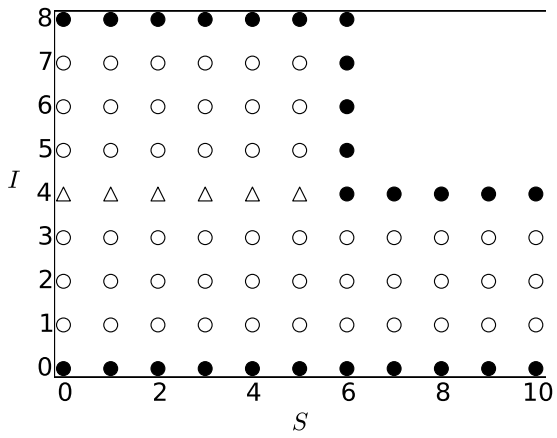


Figure 4: A two-dimensional representation of the states in the first trough Markov chain, with $S_d = 6$ and $I_d = 4$. The states in the absorbing boundaries are shown as solid circles. States in T (the start of the first trough) are denoted by triangles, and contain $\boldsymbol{\Delta}$, the distribution of the process when it enters the first trough. States continue infinitely to the right.

next and previous column, respectively. Provided state (S, m) is not in the upper absorbing boundary: then the $(m, n)^{th}$ element of F_S is the probability that the first exit from column S is into state $(S + 1, n)$, given that the process starts in state (S, m); and the $(m, n)^{th}$ element of B_S is the probability that the first exit from column S is into state $(S-1, n)$, given that the process starts in state (S, m). These probabilities are calculated exactly for the Markovian SIR model with demography, which has the transition rates illustrated in Figure 5(a).

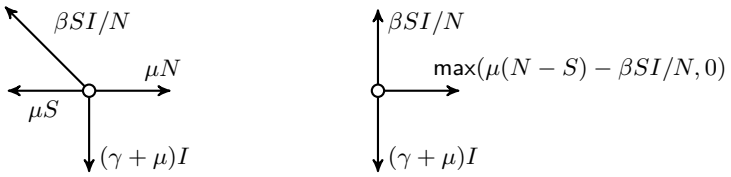
The upper absorbing boundary is an artificial absorbing boundary, so we need to treat it in a specific manner. We define F_S and B_S such that any probability mass in the upper absorbing boundary of column S moves to the upper absorbing boundary of column $S+1$. More specifically, if the state (S, m) is in the upper absorbing boundary, then row m of $[F_S \ B_S]$ is all zeros, except the $(m, n)^{th}$ element of F_S which is equal to 1, where $n = 2I_d$ if $S < S_d$, or $n = I_d$ if $S \geq S_d$.

We define P_S as the matrix of *first entry* into column $S + 1$ from column S. So the $(m, n)^{th}$ element of P_S is the probability that the first entry to column $S + 1$ is into state $(S + 1, n)$, given that the process starts in state (S, m).

With these definitions, we establish the recursive relation:

$$P_S = \begin{cases} F_0 & \text{if } S = 0 \\ (\boldsymbol{I} - B_S P_{S-1})^{-1} F_S & \text{if } S > 0, \end{cases} \tag{11}$$

where $\boldsymbol{I}$ (here only) is the identity matrix. The F_S and B_S matrices are straight-



(a) (b)

Figure 5: Transition rates from state (S, I) (except when state (S, I) is in the upper absorbing boundary. (a) shows the exact model (Section 3.3.1). (b) shows the approximate model (Section 3.3.2), in which every infection event (rate $\beta SI/N$) or removal of susceptible event (rate μS) is "paired" with a birth event (rate μN).

forward to calculate, so we can determine each P_S matrix.

Now define the vector D_S to be the distribution of Δ in column S. The only non-zero element of D_S is element I_d, and then only if $S < S_d$. Finally, we define E_S to be the distribution of all probability mass which first entered the first trough at column S or less, on the first trough Markov chain's first entry to column S. By definition, all probability mass enters the first trough at $S < S_d$. So for $S \geq S_d$, the definition of E_S simplifies to: the probability distribution on the first trough Markov chain's first entry to column S.

We then use P_S and D_{S+1} to calculate E_{S+1} through the recursion:

$$\begin{aligned} E_0 &= D_0, \\ E_{S+1} &= E_S P_S + D_{S+1} \quad \text{if } S \geq 0. \end{aligned} \tag{12}$$

As we increment S, eventually all but a vanishingly small amount of the probability mass is at one of the absorbing boundaries. That is, for any δ, there is an $S \geq S_d$ such that elements 0 and I_d of E_S sum to greater than $1-\delta$; when this occurs, p_0 is taken to be element 0 of E_S.

For a given Δ, this method gives an *exact* result (to the accuracy of the δ chosen). But (as we shall see in Section 4.1) it does not scale well to very large population sizes. This is because it requires the calculation and storage of four matrices (F_S, B_S, P_{S-1} and P_S) with a maximum size of $(2I_d{+}1) \times (2I_d{+}1)$. We now proceed to consider an approximate model which reduces the computational overheads.

3.3.2. Approximate model

Considering the method in Section 3.3.1, we can substantially reduce both evaluations and storage by making the following observation: whenever the system goes back from column S to column $S-1$, it returns to column S after an unknown number of intermediate events, followed by a "birth" event (because the birth events, at rate μN, are the only events which increase the number of susceptibles). So let us - as an approximation - assume that this

"unknown number of intermediate events" is in fact *no events*. In other words, every "death of a susceptible" event (at rate μS) is paired with a birth event (at rate μN); and every infection event (at rate $\beta SI/N$) is also paired with a birth event. So, with this assumption, we calculate the jump chain transition probabilities using the one step transition rates shown in Figure 5(b) rather than Figure 5(a).

If we were to follow the analysis of Section 3.3.1, that would mean $B_S = 0$ and so equation (11) reduces to $P_S = F_S$. However it is possible to avoid generating F_S (or any other large matrices) altogether.

Given E_S and the transition probabilities within column S, we can calculate the expected number of visits to each state before exiting the column. Since each state (S, I) only communicates with the adjacent states $(S, I-1)$ and $(S, I+1)$, this is a tri-diagonal series of simultaneous equations, which can be solved using an efficient technique such as the Thomas algorithm [12]. These expected numbers of visits multiplied by the transition probabilities to the right (determined from the transition rates in Figure 5(b)) give E_{S+1}. In other words, with reference to equation (12), we calculate the vector $E_S P_S$ directly without calculating P_S.

The elements of E_S corresponding to the absorbing boundaries ($I = 0$, and $I = 2I_d$ or I_d) accumulate probability mass as S increments. As in Section 3.3.1,

for sufficiently large $S \geq S_d$, all but a vanishingly small amount of probability mass is absorbed, allowing us to efficiently approximate p_0.

Given the simpler computation and low storage requirement, it is no surprise that this is much faster than the exact method in Section 3.3.1. But, as we shall show in the next section, this method is also very accurate.

4. Results

4.1. Accuracy and efficiency

In this section we compare the accuracy of all methods – the previous work of van Herwaarden [25] and Meerson and Sasorov [18] as presented in Section 2.3, and our new algorithms as detailed in Section 3.

As references, we also add an exact computation, and a Monte Carlo simulation. The exact computation uses a truncated state space, with an extra absorbing boundary at $S + I = (1.1)N$. The amount of probability mass absorbed at that boundary is extremely low (never greater than 10^{-5}) and does not affect the results significantly. For both the exact computation and Monte Carlo simulation, as well as our methods, we calculate p_0 using the definition in Section 3.1. For previous works, we use the expressions in Section 2.3.

Since we are interested in evaluating a probability which is only state (and not time) dependent, we may scale time and hence fix $\gamma = 1$, so time is in units of the average infectious period of an individual.

We chose seven values of the population size parameter, N: 1000, 3000, 10000, 30000, 100000, 300000 and 1000000; and, six effective transmission rate parameter values, β: 1.1, 1.2, 1.5, 2, 4 and 8. For each of these 42 pairs of

values, we chose three values of population turnover rate parameter μ, to give final p_0 values of approximately 0.1, 0.5 and 0.9 respectively. In a few cases (notably for low N and low β), an appropriate μ value could not be found. The μ values used are given in Table A.1 in Appendix A.

For $N \leq 3000$ all methods were compared against an exact computation. For larger N they were all compared against Monte Carlo simulations. The worst case and average errors for each N value are shown in Figure 6. Error bars are due to the uncertainty in the result from the Monte Carlo simulations, which require a very long run time for large N.

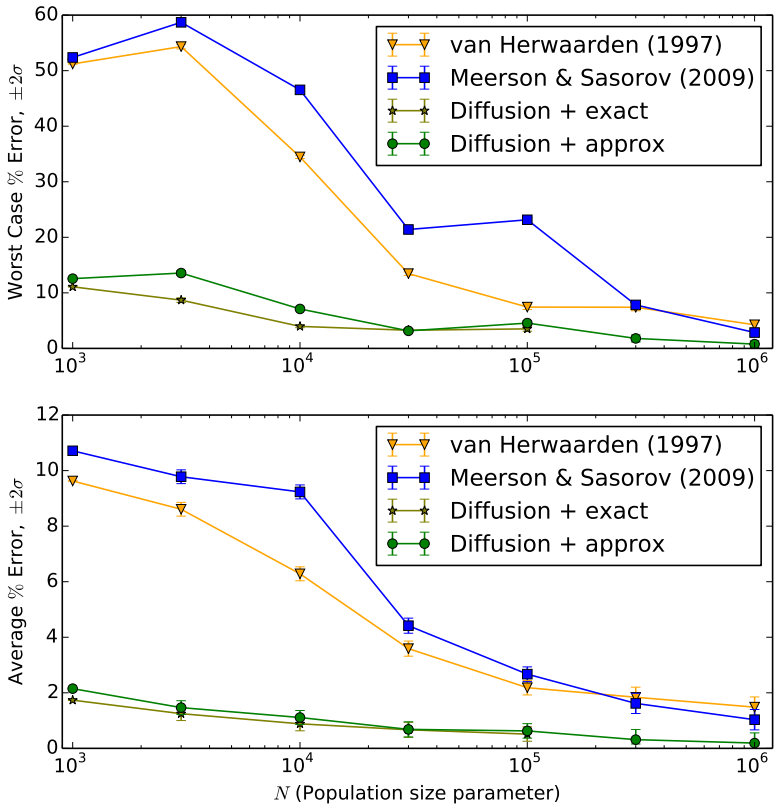


Figure 6: Plot of worst case and average error versus population size N, with $\pm 2\sigma$ error bars.

We see that both our methods are noticeably more accurate than the previously published results. We also see the accuracies of our two methods are comparable. This suggests that most of the error is in the diffusion approx-

imation of Δ (Section 3.2), rather than the Markov chain approximation of Section 3.3.2.

We also consider the efficiency of our methods, and compare them to exact computations and Monte Carlo simulations. To give comparable (though generally lower) accuracy, the Monte Carlo simulations were run long enough to give a standard deviation of $\sigma = 0.005 = 0.5\%$ in their approximations of p_0. All tests were run on a 2014 iMAC (Intel i5 core, 2.7 GHz, 8 GB RAM, Mac OS X) running Cython [5]. Computation time for different methods is shown in Figure 7. The methods of van Herwaarden and Meerson and Sasorov are not shown, because they take negligible time (in the milliseconds), and their times are independent of N. Therefore these methods remain the best for getting an approximate answer quickly.

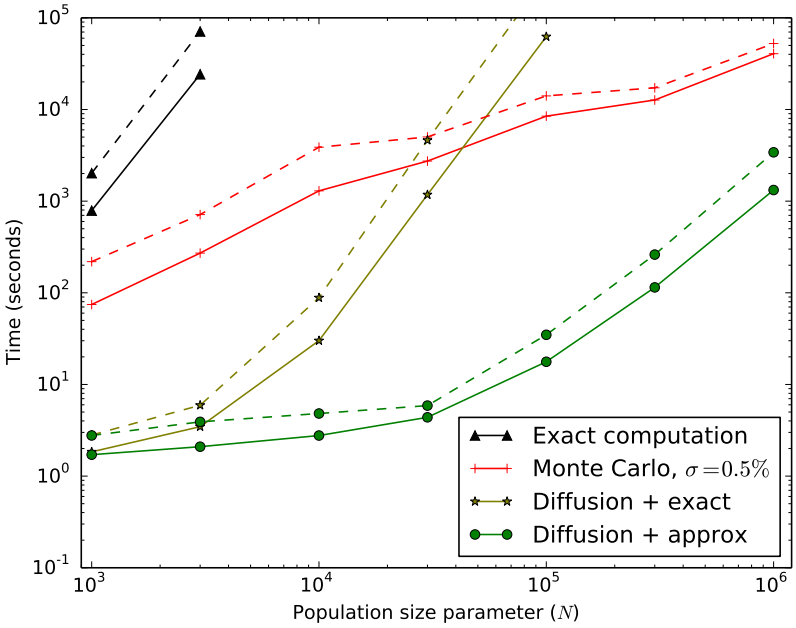


Figure 7: Plot of computation time versus population size N. Mean computation times are shown with solid lines. Slowest computation times are shown with dashed lines.

The time for the exact computation is approximately proportional to N^3, and quickly becomes impractical.

Our first method, based upon a diffusion approximation to the first entrance to the first trough, combined with an exact computation (Section 3.3.1) also has a computation time approximately proportional to N^3, and becomes impractical as N approaches 10^5.

Our second method, based upon a diffusion approximation to the first entrance to the first trough, combined with an approximate model (Section 3.3.2) has a time which is approximately proportional to N^2 and so is practical up to at least $N = 10^7$. The small memory overhead means even larger sizes may be computed if a long run time is not a concern.

The time for a set of Monte Carlo simulations is proportional to a little less than N^2, though our method is projected to be faster and more accurate up to at least $N = 10^7$.

So for a very wide range of N (from a few thousand, to the millions), the diffusion plus approximate model algorithm appears to be the most accurate of practical methods.

4.2. Analysis of the results

We used the method of Section 3.3.2 to run a larger set of tests, to explore how the probability of epidemic fade-out changes as a function of model parameters.

For $N = 1000$, 10000, 100000 and 1000000, we tested: $\gamma = 1$; 40 β values from 1.1 to 5, stepping in increments of 0.1; and μ values from 0.010 up to 0.089, 0.049, 0.029 and 0.019 for the respective values of N, stepping in increments of 0.001. Contour plots of the p_0 values are shown in Figure 8.

Figure 8 shows the interesting result that p_0 is generally non-monotonic in β. Naively, one might expect a higher infection rate β to cause the infection to be more persistent, and so give a lower p_0. What we instead see, in most cases, is a local maximum near $\beta = 2$. Since $\mu \ll 1$ and $\gamma = 1$, it follows that $R_0 = \beta/(\gamma + \mu) \approx \beta$ and so the local maximum is also near $R_0 = 2$.

The main reason for the non-monotonicity, and the peak near $R_0 = 2$, can be seen by considering the traces for $R_0 = 5, 2$ and 1.3 in Figure 9. Note these are the solutions to Equation (4) (scaled by N), and hence the deterministic approximations to the epidemic. We define I_m to be the minimum I value in the first trough of the deterministic approximation.

We may rearrange Equations (1) and (7) to give $dI/dt = (\beta I/N)(S - S_e)$ and hence $d(\ln(I))/dt = (\beta/N)(S - S_e)$. This means that for a given deterministic curve, the rate of change of $\ln(I)$ is proportional to $S - S_e$. It can also be shown that the minimum S occurs at $I > I_e$.

If we compare the $R_0 = 2$ curve to the $R_0 = 1.3$ curve, we see that the $R_0 = 2$ curve starts further from its endemic point; that is, it has a higher initial $S - S_e$ value. This means $S - S_e$ is higher in the early stages of the outbreak, which causes I to rise more steeply and for longer, so the maximum $I - I_e$ value is higher. This in turn gives S more time to fall, so the curve reaches a lower minimum $S - S_e$ value. Finally, this gives more time for I to fall, so the $R_0 = 2$ curve falls to a considerably lower I_m value then the $R_0 = 1.3$ curve. Biologically, the $R_0 = 2$ case indicates the infection "burning out" - it uses up so much resources (indicated by S falling low) that it is slow to re-establish itself, so it falls to a low I_m, giving it a higher probability of epidemic fade-out.

If we compare the $R_0 = 5$ curve to the $R_0 = 2$ curve, the $R_0 = 5$ curve has an even higher initial $S - S_e$ value, and so rises to a higher maximum $I - I_e$.

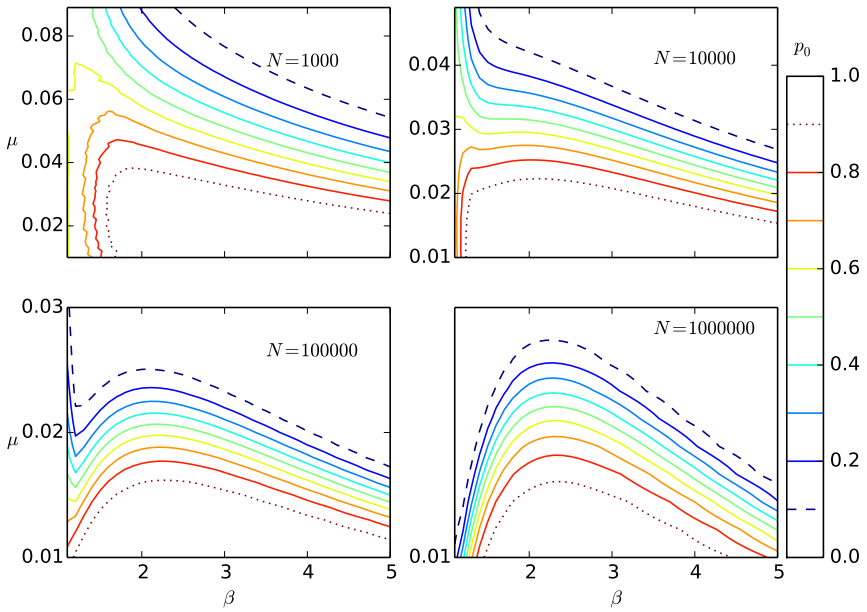


Figure 8: Constant p_0 contours for various β, μ and N values, with $\gamma = 1$ and $(S_0, I_0) = (N-1, 1)$. The $p_0 = 0.1$ contour is dashed, the $p_0 = 0.9$ contour is dotted, and the contours for $p_0 = 0.2$ to 0.8 in steps of 0.1 are solid. For most combinations of N and μ, p_0 peaks near $\beta = 2$.

But then, as S falls, it is limited by the condition that $S \geq 0$. So it does not fall to as low a minimum $S - S_e$ value as the $R_0 = 2$ curve does. This in turn means that its I_m is not as low as for $R_0 = 2$. It also only has to reach a comparatively low S before $S > S_e$, and the deterministic curve begins to rise again. Biologically, $R_0 = 5$ corresponds to a case where the infection rate is so high that the infection can re-establish itself from comparatively low resources.

So $S_e \approx N/2$ corresponds to a "sweet spot" where the curve can swing from a high $S - S_e$ to a low $S - S_e$, giving the most time for the curve to fall to a low I_m. And it follows from (6) and (7) that $R_0 = 2$ corresponds to $S_e = N/2$, and so this sweet spot occurs near $R_0 = 2$. This is illustrated in Figure 10, which plots I_m versus R_0 for the same parameters as used in Figure 9, with the lowest I_m occuring at $R_0 \approx 2.4$.

When we consider stochastic effects, a lower I_m corresponds to a greater probability of absorption at $I = 0$, and hence a higher p_0. However the probability of aborption also depends on the time the deterministic process spends near I_m. A longer time near I_m corresponds to a longer time near the $I = 0$ absorbing boundary. This gives the process more opportunity to be absorbed

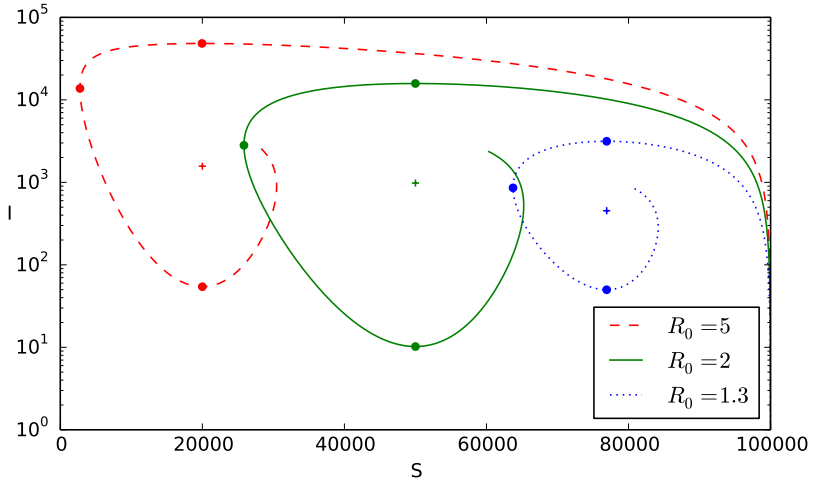


Figure 9: Comparison of deterministic traces, on a logarithmic I scale, for $N = 100000$, $\gamma = 1$, $\mu = 0.02$ and $(S_0, I_0) = (N-1, 1)$. The R_0 values of 5, 2 and 1.3 correspond to S_e values of approximately $N/5$, $N/2$ and $4N/5$ respectively. The plus signs show the endemic points (S_e, I_e) to which the curves converge. The dots mark the maximum I, minimum S and first trough minimum I (I_m).

due to stochastic effects, and so should lead to a higher p_0.

It is shown in Appendix B that in the first trough of the deterministic process, the time for which $I < I_m + \epsilon$, for sufficiently small ϵ, is monotonically decreasing in βI_m. This means it is also monotonically decreasing in $R_0 I_m$. We must handle this result with some care because I_m is itself dependent on R_0. But it means that, in the region where the I_m versus R_0 curve is relatively flat, a decrease in R_0 gives an increase in p_0. So we would expect the maximum p_0 to occur at an R_0 slightly lower than the R_0 which gives rise to the minimum I_m. This is also illustrated in Figure 10. For this particular case the minimum I_m occurs at $R_0 \approx 2.4$, but the lowest p_0 occurs at $R_0 \approx 2.2$.

5. Conclusion

We have presented a two stage method for calculating an accurate approximation for the probability of epidemic fade-out. Using an approximate model on the second stage gives an algorithm which is both fast and accurate. It is more accurate than the previously published formulae, and scales much better than exact computation methods. This technique can also be used in other SIR-type models with replenishment of susceptibles (for instance, those with a fixed population size).

A possible justification for why the approximate model of Section 3.3.2 retains such a high level of accuracy is that in the first trough (when I is low), the

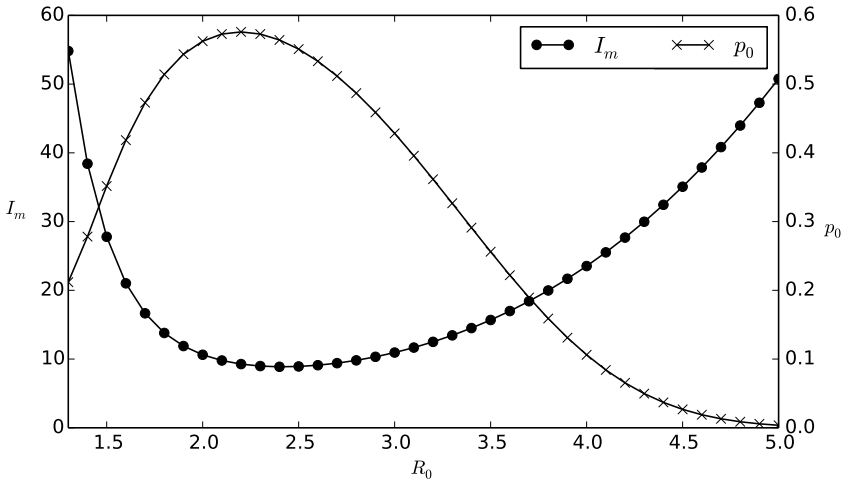


Figure 10: I_m and p_0 plotted against R_0, for $N = 100000$, $\gamma = 1$, $\mu = 0.02$, $(S_0, I_0) = (N-1, 1)$, and $\beta = R_0(\gamma + \mu)$. In this example, the minimum I_m is at $R_0 \approx 2.4$, while the maximum p_0 is at $R_0 \approx 2.2$.

birth events (μN) are almost always at a higher rate - and often a much higher rate - than infection events ($\beta SI/N$). Therefore there is a very small penalty (in terms of accuracy) for pairing every infection event with a birth event.

Further, comparing Figure 5(b) to Figure 5(a), we see that the approximation makes no change to the one step behaviour in the I (vertical) dimension. In the S (horizontal) dimension, the behaviour is simplified, but the average drift $(\mu(N-S)-\beta SI/N)$ is modelled correctly (except when $\mu(N-S)-\beta SI/N < 0$, but in those cases the Markov chain is near point F in Figure 3, so the computation is nearly complete). So the S dimension is modelled accurately in the first moment but not the second moment. It appears that this only introduces a small error because the I dimension is much more critical than the S dimension.

Using this fast and accurate method, we have found that the probability of epidemic fade-out often peaks when the basic reproduction number, R_0, is approximately 2 (restricting consideration to cases where a major outbreak is possible, i.e., $R_0 > 1$). This is because $R_0 \approx 2$ is high enough to use up a large proportion of resources, but not so high that the infection can easily recover from having few resources. The reason this occurs near $R_0 = 2$ appears to be due to the endemic point being near $S = N/2$.

A potential public health application is that there may be instances where action against an infection should be limited, to maximise the chance of infection being eliminated before it becomes endemic. We note there is some similarity here to the observations of Rozhnova, Metcalf and Grenfell [24], that decreasing R_0 by vaccination may sometimes lead to higher persistence, though their study was with respect to an already endemic infection, with seasonality.

The question of whether a peak near $R_0 = 2$ extends to other measures or models, is a topic for future research. Another avenue for future research is to determine methods which allow calculation of the probability of epidemic fade-out for models with seasonal forcing (i.e., a time-dependent effective transmission rate parameter) [14]. This in turn could aid understanding of the Critical Community Size for diseases such as measles in the pre-vaccine era [9].

6. Acknowledgements

The authors thank Nic Rebuli for helpful discussions on the diffusion approximation (Section 3.2). We also thank Guy Latouche for a suggestion which led us to the efficient and exact solution in Section 3.3.1. This work is supported by an APA Scholarship (PB), an Australian Research Council Future Fellowship (JVR; FT130100254), and the NHMRC (JVR; CRE PRISM^2).

7. References

[1] R. M. Anderson and R. M. May. *Infectious Diseases of Humans: dynamics and control*. Oxford University Press, 1991.

[2] H. Andersson and T. Britton. *Stochastic epidemic models and their statistical analysis*. Springer, 2000.

[3] F. G. Ball. The threshold behaviour of epidemic models. *Journal of Applied Probability*, 20(2):227–241, 1983.

[4] F. G. Ball and O. D. Lyne. Optimal vaccination policies for stochastic epidemics among a population of households. *Mathematical Biosciences*, 177:333–354, 2002.

[5] S. Behnel, R. Bradshaw, C. Citro, L. Dalcin, D. S. Seljebotn, and K. Smith. Cython: The best of both worlds. *Computing in Science & Engineering*, 13(2):31–39, 2011.

[6] C. Bender and S. Orszag. *Advanced Mathematical Method for Scientists and Engineers*. McGraw-Hill, 1978.

[7] A. J. Black, T. House, M. J. Keeling, and J. V. Ross. Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza. *Journal of the Royal Society Interface*, 10:20121019, 2013.

[8] T. Britton, T. House, A. L. Lloyd, D. Mollison, S. Riley, and P. Trapman. Five challenges for stochastic epidemic models involving global transmission. *Epidemics*, 10:54–57, 2015.

[9] A. J. K. Conlan, P. Rohani, A. L. Lloyd, M. Keeling, and B. T. Grenfell. Resolving the impact of waiting time distributions on the persistence of measles. *J. R. Soc. Interface*, 7:623–640, 2010.

[10] O. Diekmann and J. A. P. Heesterbeek. *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis and Interpretation*. John Wiley & Sons, 2000.

[11] S. N. Ethier and T. G. Kurtz. *Markov Processes*. John Wiley & Sons, 1986.

[12] J. D. Hoffman. *Numerical Methods for Scientists and Engineers*. Marcel Dekker, 2nd edition, 2001.

[13] A. Kamenev and B. Meerson. Extinction of an infectious disease: A large fluctuation in a nonequilibrium system. *Physical Review E*, 77:061107, 2008.

[14] M. J. Keeling, P. Rohani, and B. T. Grenfell. Seasonally forced disease dynamics explored as switching between attractors. *Physica D: Nonlinear Phenomena*, 148(3-4):317–335, 2001.

[15] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump Markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

[16] T. G. Kurtz. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. *Journal of Applied Probability*, 8(2):344–356, 1971.

[17] J. M. McCaw and J. McVernon. Prophylaxis or treatment? optimal use of an antiviral stockpile during an influenza pandemic. *Mathematical Biosciences*, 209:336–360, 2007.

[18] B. Meerson and P. V. Sasorov. WKB theory of epidemic fade-out in stochastic populations. *Physical Review E*, 80:041130, 2009.

[19] I. Nåsell. On the time to extinction in recurrent epidemics. *Journal of the Royal Statistical Society: Series B*, 61:309–330, 1999.

[20] I. Nåsell. Extinction and quasi-stationarity in the Verhulst logistic model. *Journal of Theoretical Biology*, 211:11–27, 2001.

[21] J. R. Norris. *Markov chains*. Cambridge University Press, 1998.

[22] P. K. Pollett. On a model for interference between searching insect parasites. *The Journal of the Australian Mathematical Society, Series B*, 32:133–150, 1990.

[23] J. V. Ross and A. J. Black. Contact tracing and antiviral prophylaxis in the early stages of a pandemic: the probability of a major outbreak. *Mathematical Medicine and Biology*, to appear.

[24] G. Rozhnova, C. J. E. Metcalf, and B. T. Grenfell. Characterizing the dynamics of rubella relative to measles: the role of stochasticity. *Journal of the Royal Society Interface*, 10:20130643, 2013.

[25] O. A. van Herwaarden. Stochastic epidemics: the probability of extinction of an infectious disease at the end of a major outbreak. *Journal of Mathematical Biology*, 35:793–813, 1997.

[26] O. A. van Herwaarden and J. Grasman. Stochastic epidemics: major outbreaks and the duration of the endemic period. *Journal of Mathematical Biology*, 33:581–601, 1995.

Appendix A

N	β	μ for $p_0 \approx 0.9$	μ for $p_0 \approx 0.5$	μ for $p_0 \approx 0.1$
1000	1.1	–	0.043	–
	1.2	0.053	–	–
	1.5	0.034	0.084	–
	2.0	0.033	0.060	0.112
	4.0	0.025	0.041	0.064
	8.0	0.015	0.025	0.039
3000	1.1	–	0.064	–
	1.2	0.026	0.050	–
	1.5	0.025	0.046	0.091
	2.0	0.026	0.041	0.062
	4.0	0.021	0.031	0.042
	8.0	0.012	0.018	0.025
10000	1.1	0.020	0.051	–
	1.2	0.017	0.038	0.095
	1.5	0.019	0.030	0.046
	2.0	0.021	0.031	0.041
	4.0	0.017	0.024	0.031
	8.0	0.010	0.014	0.018
30000	1.1	0.012	0.030	–
	1.2	0.012	0.022	0.041
	1.5	0.016	0.023	0.031
	2.0	0.018	0.025	0.031
	4.0	0.015	0.020	0.024
	8.0	0.008	0.011	0.014
100000	1.1	0.008	0.016	0.033
	1.2	0.009	0.015	0.022
	1.5	0.013	0.018	0.023
	2.0	0.015	0.020	0.025
	4.0	0.013	0.017	0.020
	8.0	0.007	0.010	0.012
300000	1.1	0.006	0.010	0.017
	1.2	0.008	0.012	0.016
	1.5	0.011	0.015	0.019
	2.0	0.014	0.017	0.021
	4.0	0.011	0.015	0.017
	8.0	0.006	0.008	0.010
1000000	1.1	0.005	0.007	0.010
	1.2	0.006	0.009	0.012
	1.5	0.010	0.013	0.015
	2.0	0.012	0.015	0.018
	4.0	0.010	0.013	0.015
	8.0	0.006	0.007	0.008

Table A.1: μ values used in Section 4.1

Appendix B

Theorem. *In the first trough of the deterministic process, the time θ for which $I < I_m + \epsilon$, for sufficiently small ϵ, is monotonically decreasing in βI_m.*

Proof. Consider the deterministic plot of I versus S, as in Figures 2 and 3. At the first trough minimum (point E), $dI/dS = 0$. For sufficiently small ϵ, we can therefore treat d^2I/dS^2 as constant, and in the region where $I < I_m + \epsilon$, I is parabolic when plotted against S. So the distance in the S dimension, for which $I < I_m + \epsilon$, is monotonically decreasing in the parabola curvature d^2I/dS^2.

The rate at which the deterministic process moves in the S direction is dS/dt, so θ is inversely proportional to dS/dt. This means that θ is monotonically decreasing in $(d^2I/dS^2)(dS/dt)$.

Substituting $I = Ni$ and $S = Ns$ into (1) gives

$$\frac{dS}{dt} = \mu N - \mu S - \beta SI/N \quad , \tag{B.1}$$

$$\frac{dI}{dS} = \frac{dI/dt}{dS/dt} = \frac{\beta SI/N - (\gamma + \mu)I}{\mu N - \mu S - \beta SI/N} \tag{B.2}$$

$$\Rightarrow \frac{d^2I}{dS^2} = \frac{(\beta I/N)(\mu N - \mu S - \beta SI/N) - [\beta SI/N - (\gamma + \mu)I]\,(-\mu - \beta I/N)}{(\mu N - \mu S - \beta SI/N)^2}$$

$$= \frac{I\mu(\beta - \gamma - \mu) - \beta I^2(\gamma + \mu)/N}{(\mu N - \mu S - \beta SI/N)^2} \quad . \tag{B.3}$$

At the first trough minimum of the deterministic curve we have defined $I = I_m$. Also $dI/dS = 0$, so it follows from (B.2) that $S = N(\gamma + \mu)/\beta$, and $N - S = N(\beta - \gamma - \mu)/\beta$. Substituting these into (B.1) and (B.3) gives

$$\frac{dS}{dt} = \mu N(\beta - \gamma - \mu)/\beta - I_m(\gamma + \mu) \quad ,$$

$$\frac{d^2I}{dS^2} = \frac{I_m\mu(\beta - \gamma - \mu) - \beta {I_m}^2(\gamma + \mu)/N}{[\mu N(\beta - \gamma - \mu)/\beta - I_m(\gamma + \mu)]^2} \quad ;$$

$$\Rightarrow \left(\frac{d^2I}{dS^2}\right)\left(\frac{dS}{dt}\right) = \frac{\beta I_m}{N} \quad . \tag{B.4}$$

Therefore θ is monotonically decreasing in βI_m.

□