



Heriot-Watt University
Research Gateway

Precise estimates of persistence time for SIS infections in heterogeneous populations

Citation for published version:

Clancy, D 2018, 'Precise estimates of persistence time for SIS infections in heterogeneous populations', *Bulletin of Mathematical Biology*, vol. 80, no. 11, pp. 2871-2896. <https://doi.org/10.1007/s11538-018-0491-6>

Digital Object Identifier (DOI):

[10.1007/s11538-018-0491-6](https://doi.org/10.1007/s11538-018-0491-6)

Link:

[Link to publication record in Heriot-Watt Research Portal](#)

Document Version:

Publisher's PDF, also known as Version of record

Published In:

Bulletin of Mathematical Biology

Publisher Rights Statement:

© The Author(s) 2018. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

General rights

Copyright for the publications made accessible via Heriot-Watt Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy

Heriot-Watt University has made every reasonable effort to ensure that the content in Heriot-Watt Research Portal complies with UK legislation. If you believe that the public display of this file breaches copyright please contact open.access@hw.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.



Precise Estimates of Persistence Time for SIS Infections in Heterogeneous Populations

Damian Clancy¹ 

Received: 12 February 2018 / Accepted: 24 August 2018
© The Author(s) 2018

Abstract

For a susceptible–infectious–susceptible infection model in a heterogeneous population, we derive simple and precise estimates of mean persistence time, from a quasi-stationary endemic state to extinction of infection. Heterogeneity may be in either individuals’ levels of infectiousness or of susceptibility, as well as in individuals’ infectious period distributions. Infectious periods are allowed to follow arbitrary non-negative distributions. We also obtain a new and accurate approximation to the quasi-stationary distribution of the process, as well as demonstrating the use of our estimates to investigate the effects of different forms of heterogeneity. Our model may alternatively be interpreted as describing an infection spreading through a heterogeneous directed network, under the annealed network approximation.

Keywords Stochastic epidemic models · Large deviations · Endemic fade-out · Stochastic networks · Superspreaders

1 Introduction

For models of infectious spread in which long-term quasi-stable endemicity is possible, a random variable of particular interest is the persistence time until infection dies out from the population. For a number of such models, it is known (Andersson and Djehiche 1998; Ball et al. 2016; van Herwaarden and Grasman 1995) that as the typical population size N tends to infinity, the expected persistence time τ for an infection that has become endemic in the population satisfies

$$\tau \sim \frac{C}{\sqrt{N}} \exp(N A) \quad (1)$$

✉ Damian Clancy
d.clancy@hw.ac.uk

¹ Department of Actuarial Mathematics and Statistics, Maxwell Institute for Mathematical Sciences, Heriot-Watt University, Edinburgh EH14 4AS, UK

where we use \sim to denote that the ratio of the two sides converges to 1 as $N \rightarrow \infty$, and A, C are constants whose values depend upon parameters of the process, but not upon N . We assume here, and from now on, that the basic reproduction number R_0 (the average number of new infections caused by a typical infected individual in an otherwise susceptible population) is greater than one, so that the process is supercritical, or ‘above threshold’.

A pioneering piece of work in this direction was van Herwaarden and Grasman (1995), where it was shown that the relationship (1) holds for a particular susceptible–infectious–removed (SIR) infection model. Evaluation of A required numerical solution of a system of ordinary differential equations, while no method for evaluating C was given. It should be noted that van Herwaarden and Grasman (1995) studied a diffusion approximation to the infection model, but it is now well known (see, e.g., Clancy and Tjia 2018; Doering et al. 2005) that such a diffusion approximation does not, in general, give correct leading-order asymptotics—that is, the value of the constant A computed from the diffusion approximation is not necessarily equal to the correct A value for the underlying discrete state-space model. Using rather different techniques, Andersson and Djehiche (1998) derived a result of the form (1), together with explicit expressions for the constants A, C , for the classic susceptible–infectious–susceptible (SIS) model of Weiss and Dishon (1971).

In recent years, a number of authors have applied techniques from statistical physics to study persistence times for a range of population models, including infection models. For models which are naturally one-dimensional, including the classic SIS model, a relationship of the form (1) together with explicit formulae for A, C can be obtained (Assaf and Meerson 2010, 2017). For multidimensional models (including most infection models), it is usually only possible to establish results of the cruder form $\lim_{N \rightarrow \infty} (\ln \tau)/N = A$, and to evaluate the leading-order constant A via numerical solution of a system of ordinary differential equations (Dykman et al. 1994; Elgart and Kamenev 2004; Hindes and Schwartz 2016; Kamenev and Meerson 2008; Lindley et al. 2014). One reason why the technique has not been more widely exploited is that for a k -dimensional model, it is necessary to solve a system of ordinary differential equations in $2k$ dimensions subject to boundary conditions at times $t = -\infty$ and $t = +\infty$. Progress has been made by a number of authors in the development of efficient numerical procedures (e.g. Forgoston et al. 2011; Lindley and Schwartz 2013), but implementation for models in dimensions $k > 2$ remains far from trivial.

Much of the above work on infection models assumes that individuals’ infectious periods are exponentially distributed. This is not biologically realistic for most infections and is done purely for reasons of mathematical tractability, so that it is of interest to understand the effect of this simplifying assumption upon the resulting persistence time estimates. Ball et al. (2016) extended the result of Andersson and Djehiche (1998) for the classic SIS model to allow for a quite general infectious period distribution, by applying a result on insensitivity in stochastic networks from Zachary (2007), finding that the leading-order constant A takes the same value regardless of the infectious period distribution (provided only that its mean is held constant), but that the prefactor constant C must be appropriately modified.

A different extension of the result of Andersson and Djehiche (1998) has recently been established in Clancy (2018). For an SIS model incorporating heterogeneity

in susceptible individuals' levels of susceptibility or infected individuals' levels of infectiousness, an explicit formula was found for the leading-order constant A in the relationship $\lim_{N \rightarrow \infty} (\ln \tau)/N = A$. Infectious periods were allowed to follow an Erlang distribution, and the value of A shown to depend only upon the mean of the distribution, under this assumption. In the current paper, we build on the work of Clancy (2018) to establish results of the much more precise form (1), including simple explicit formulae for the prefactor constant C . At the same time, we extend the model of Clancy (2018) in two ways: firstly, we allow for heterogeneity in individuals' infectious period distributions in addition to the heterogeneities in susceptibility and infectiousness of Clancy (2018); secondly, following the approach of Ball et al. (2016), we allow for infectious periods following quite general (not necessarily Erlangian) distributions. Clancy (2018) showed that for a sufficiently large population, greater heterogeneity (in the sense of majorization ordering, see Marshall et al. 2011), whether in susceptibility or infectiousness, leads to a reduction in mean persistence time of infection in the population. Using our more general model, we are also able to investigate the effect of heterogeneity in infectious period distributions; see Sect. 5.2.

We note that the only infection model for which such explicit formulae for A , C have previously been available is the SIS model in a homogeneously mixing population (with general infectious period distribution); in general, the leading-order constant A must be evaluated via numerical solution of a system of ordinary differential equations (a non-trivial exercise, as noted above), while no general method exists to evaluate the prefactor constant C for multidimensional models.

The remainder of the paper is structured as follows. In Sect. 2, we define our model and state our main result, Theorem 1. Section 3 recalls some standard approximations for infection models and general theory that we will require in the sequel. The proof of our results occupies Sect. 4. In Sect. 5, we demonstrate the accuracy of our approximations, both for mean persistence time and for the quasi-stationary distribution of the process; we apply our results to investigate the effects of different forms of heterogeneity; and we outline the application of our results to network models via the annealed network approximation (Hindes and Schwartz 2016, 2017). Finally, in Sect. 6, we present some concluding discussion and suggestions for further work.

2 The Model and Asymptotic Persistence Time Formulae

Consider a closed population of N individuals divided into k groups, with group i ($i = 1, 2, \dots, k$) consisting of N_i individuals. Denote by $f_i = N_i/N$ the proportion of the population belonging to group i , so that $\sum_i f_i = 1$. When a group i individual becomes infected, it remains so for a time distributed as a random variable T_i of mean $\alpha_i = E[T_i]$, after which it returns to the susceptible state. During this infectious period, the group i infective makes contacts with each individual in each group $j = 1, 2, \dots, k$ at the points of a Poisson process of rate $\beta \lambda_i \mu_j / N$, where β is some overall measure of infectiousness, λ_i represents the relative infectiousness of group i individuals and μ_j represents the relative susceptibility of group j individuals. (The assumption that the group i to group j infection rate factorises in this way is sometimes referred to as 'separable mixing'.) Without loss of generality, we scale the λ_i , μ_j values so

that $\sum_i \lambda_i f_i = \sum_j \mu_j f_j = 1$. These Poisson processes and infectious periods are all mutually independent. If a contacted individual is susceptible, then it becomes infected (and infectious); if the contacted individual is already infected, then the contact has no effect. We denote by $I_j(t)$ the number of infected individuals in group j at time $t \geq 0$, the corresponding number of susceptible individuals being $S_j(t) = N_j - I_j(t)$, and write $\mathbf{I}(t) = (I_1(t), I_2(t), \dots, I_k(t))$. We assume throughout that $\beta > 0$, and that $f_i, \alpha_i, \lambda_i, \mu_i > 0$ for all i . The basic reproduction number R_0 for our model is given by

$$R_0 = \beta \sum_{i=1}^k \alpha_i \lambda_i \mu_i f_i, \tag{2}$$

and we will assume throughout that parameter values are such that $R_0 > 1$.

In order to state our results, we require the following definitions. Define $D(\boldsymbol{\lambda}, \boldsymbol{\mu})$ to be the unique positive solution of

$$\beta \sum_{j=1}^k \frac{\alpha_j \lambda_j \mu_j f_j}{1 + \alpha_j \mu_j D(\boldsymbol{\lambda}, \boldsymbol{\mu})} = 1. \tag{3}$$

Denote by $\varphi_i(\theta) = E[e^{\theta T_i}]$ the moment-generating function of T_i , and define the function $\psi(\theta)$ to be

$$\psi(\theta) = \sum_{i=1}^k \mu_i f_i \varphi_i(-\beta \lambda_i (1 - \theta)). \tag{4}$$

Finally, define ω to be the unique solution in $[0, 1)$ of

$$\omega = \psi(\omega). \tag{5}$$

We are now in a position to state our main result as follows.

Theorem 1 *Consider the heterogeneous population SIS infection model defined above, and recall that τ denotes the expected time from quasi-stationarity to disease extinction.*

- (i) *Suppose that heterogeneity is in susceptibility but not infectiousness (so $\boldsymbol{\lambda} = \mathbf{1}$), and that $R_0 > 1$. Then as $N \rightarrow \infty$,*

$$\begin{aligned} \tau \sim & \frac{1}{\beta (1 - \omega) D(\mathbf{1}, \boldsymbol{\mu})} \sqrt{\frac{2\pi}{N \sum_i f_i \left(\frac{\alpha_i \mu_i}{1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})} \right)^2}} \\ & \times \exp \left(N \left(\sum_i f_i \ln \left(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}) \right) - \frac{D(\mathbf{1}, \boldsymbol{\mu})}{\beta} \right) \right). \end{aligned} \tag{6}$$

(ii) Suppose that heterogeneity is in infectiousness but not susceptibility (so $\mu = \mathbf{1}$) and that $R_0 > 1$. Provided that individuals' infectious periods are exponentially distributed (so the process $\mathbf{I}(t)$ is Markovian with transition rates given in Table 1), then as $N \rightarrow \infty$,

$$\tau \sim \frac{1}{D(\lambda, \mathbf{1})D(\mathbf{1}, \lambda)} \sqrt{\frac{2\pi}{N \sum_i f_i \left(\frac{\alpha_i \lambda_i}{1 + \alpha_i \lambda_i D(\mathbf{1}, \lambda)}\right)^2}} \times \exp\left(N \left(\sum_i f_i \ln(1 + \alpha_i \lambda_i D(\mathbf{1}, \lambda)) - \frac{D(\mathbf{1}, \lambda)}{\beta}\right)\right). \tag{7}$$

In the next section, we summarise some standard theory that we will require in order to prove Theorem 1; the proof itself occupies Sect. 4.

3 General Theory

We will make use of two standard approximations to the process $\mathbf{I}(t)$, valid during different phases of an outbreak. First, consider the early stages of an outbreak initiated by a small number of infected individuals in a large susceptible population. So long as the number of infected individuals remains small relative to the total population size, the process $\mathbf{I}(t)$ may be approximated by a multitype branching process in which each group i individual lives for a time distributed as T_i and during this time gives birth to new group j individuals ($j = 1, 2, \dots, k$) according to a Poisson process of rate $\beta \lambda_i \mu_j f_j$. Denoting by G_{ij} the number of type j offspring of a typical type i individual, the basic reproduction number R_0 is given by the dominant eigenvalue of the mean offspring matrix M with entries $m_{ij} = E[G_{ij}] = \beta \alpha_i \lambda_i \mu_j f_j$, leading to formula (2). Denoting by ω_i the probability that the branching process initiated by a single group i individual produces only a finite number of offspring (corresponding to a minor outbreak of infection), then for $R_0 \leq 1$ we have $\omega_1 = \omega_2 = \dots = \omega_k = 1$, while for $R_0 > 1$, $\omega = (\omega_1, \omega_2, \dots, \omega_k)$ is the unique solution in $[0, 1)^k$ of

$$\omega_i = \varphi_i \left(-\beta \lambda_i \left(1 - \sum_{j=1}^k \mu_j f_j \omega_j \right) \right) \quad \text{for } i = 1, 2, \dots, k,$$

see, for instance, Section 3 of Clancy and Pearce (2013).

We can alternatively treat the above multitype branching process as a single-type branching process in which the number of offspring produced by a typical type i individual is distributed as G_i where G_i follows a mixture distribution: $G_i = G_{ij}$ with probability $\mu_j f_j$ for $j = 1, 2, \dots, k$. For a process initiated by a single infected individual which belongs to group i with probability $\mu_i f_i$, the minor outbreak probability ω is given by $\omega = \sum_{i=1}^k \mu_i f_i \omega_i$, and for $R_0 > 1$, ω is the unique solution in $[0, 1)$ of $\omega = \psi(\omega)$, where $\psi(\theta)$ is defined by Eq. (4).

Given that a major outbreak occurs, then following the initial (branching process) phase, the growth of the epidemic towards an endemic equilibrium and the long-term quasi-equilibrium behaviour may be approximated by a deterministic process. Denoting by $\bar{F}_i(u) = \Pr(T_i > u)$ the survival function of T_i , then the scaled process $\mathbf{I}(t)/N$ is approximated by the deterministic process $\mathbf{y}(t)$ which evolves according to, for $i = 1, 2, \dots, k$,

$$y_i(t) = \int_{-\infty}^t \beta \left(\sum_{j=1}^k \lambda_j y_j(u) \right) \mu_i (f_i - y_i(u)) \bar{F}_i(t - u) du. \tag{8}$$

System (8) has a disease-free equilibrium point at $\mathbf{y} = \mathbf{0}$, and for $R_0 > 1$ a unique nonzero equilibrium point \mathbf{y}^* with components (Nold 1980; Clancy 2015)

$$y_i^* = \frac{\alpha_i \mu_i f_i D(\boldsymbol{\lambda}, \boldsymbol{\mu})}{1 + \alpha_i \mu_i D(\boldsymbol{\lambda}, \boldsymbol{\mu})} \quad \text{for } i = 1, 2, \dots, k, \tag{9}$$

where $D(\boldsymbol{\lambda}, \boldsymbol{\mu})$ is given by Eq. (3).

For the remainder of this section, we focus upon the case that infectious periods T_i are exponentially distributed. The process $\{\mathbf{I}(t) : t \geq 0\}$ is now a continuous-time Markov chain on the finite state space $S = \prod_{i=1}^k \{0, 1, \dots, N_i\}$ with transition rates given in Table 1. The moment-generating function of T_i ($i = 1, 2, \dots, k$) is $\varphi_i(\theta) = (1 - \alpha_i \theta)^{-1}$, and it follows that the minor outbreak probability ω is

$$\omega = 1 - (D(\boldsymbol{\mu}, \boldsymbol{\lambda})/\beta), \tag{10}$$

where $D(\boldsymbol{\mu}, \boldsymbol{\lambda})$ is given by Eq. (3) with the roles of $\boldsymbol{\lambda}, \boldsymbol{\mu}$ interchanged.

The process $\mathbf{I}(t)$ is a density-dependent process in the sense of chapter 11 of Ethier and Kurtz 2005; that is, the transition rates are of the form

$$P(\mathbf{I}(t + \delta t) = \mathbf{x} + \mathbf{l} \mid \mathbf{I}(t) = \mathbf{x}) = N W_{\mathbf{l}} \left(\frac{\mathbf{x}}{N} \right) + o(\delta t) \quad \text{for } \mathbf{x} \in S, \mathbf{l} \in L, \tag{11}$$

for some functions $W_{\mathbf{l}} : \mathbb{R}^k \rightarrow \mathbb{R}^+$, where L is the set of possible jumps from each state $\mathbf{x} \in S$. The scaled process $\mathbf{I}(t)/N$ converges almost surely over finite time intervals (Ethier and Kurtz 2005, Theorem 11.2.1), as $N \rightarrow \infty$, to the solution $\mathbf{y}(t)$ of the ordinary differential equation system

Table 1 Transition rates for the heterogeneous population SIS model with exponentially distributed infectious periods

Event	State transition	Transition rate
Infection in group j	$I_j \rightarrow I_j + 1$	$\frac{\beta}{N} \left(\sum_{m=1}^k \lambda_m I_m \right) \mu_j (N_j - I_j)$
Recovery in group j	$I_j \rightarrow I_j - 1$	$(1/\alpha_j) I_j$

$$\frac{dy}{dt} = \sum_{l \in L} l W_l(y).$$

That is, Eq. (8) may, in the case of exponentially distributed infectious periods, be written in the form

$$\frac{dy_i}{dt} = \beta \left(\sum_{j=1}^k \lambda_j y_j \right) \mu_i (f_i - y_i) - (1/\alpha_i) y_i \quad \text{for } i = 1, 2, \dots, k. \quad (12)$$

For $R_0 > 1$, the disease-free equilibrium point $y = \mathbf{0}$ is unstable and the endemic equilibrium point y^* given by (9) is globally asymptotically stable (Lajmanovich and Yorke 1976).

Denote by Q the transition rate matrix of the process, with entries given in Table 1. The state space S is made up of an absorbing state at $\mathbf{0}$ and a single transient communicating class C , and we denote by Q_C the restriction of Q to C . The infection will almost surely die out within finite time, and (Darroch and Seneta 1967) there exists a unique quasi-stationary distribution q with elements

$$q_x = \lim_{t \rightarrow \infty} \Pr(\mathbf{I}(t) = \mathbf{x} \mid \mathbf{I}(t) \in C) \quad \text{for } \mathbf{x} \in C.$$

The distribution q may be found as the unique solution of

$$q Q_C = -(1/\tau) q \quad \text{with} \quad \sum_{x \in C} q_x = 1, \quad (13)$$

where $-(1/\tau)$ is the eigenvalue of Q_C with largest real part, and the time to extinction from quasi-stationarity is exponentially distributed with mean τ .

For a process with transition rates of the form (11), Eq. (13) may be written as

$$\sum_{l \in L} \left(q_{x-l} W_l \left(\frac{x-l}{N} \right) - q_x W_l \left(\frac{x}{N} \right) \right) = -(\tau N)^{-1} q_x \quad \text{for } \mathbf{x} \in C, \quad (14)$$

with

$$\tau = \left(N \sum_{l \in L} q_{-l} W_l \left(-\frac{l}{N} \right) \right)^{-1}. \quad (15)$$

Writing $y = x/N$, then, following the methodology described in Assaf and Meerson (2017) and references therein, we adopt the WKB (Wentzel, Kramers, Brillouin) *ansatz* that

$$q_x = K_N \exp(-NV(y) - V_0(y) + o(1)) \quad (16)$$

for some functions $V(\mathbf{y})$, $V_0(\mathbf{y})$ that do not depend upon N , and some K_N that does not depend upon \mathbf{y} . Without loss of generality, we set $V(\mathbf{y}^*) = V_0(\mathbf{y}^*) = 0$. Then

$$q_{x-l} = q_x \exp \left(\mathbf{l}^T \frac{\partial V}{\partial \mathbf{y}} - \frac{1}{2N} \mathbf{l}^T \frac{\partial^2 V}{\partial \mathbf{y}^2} \mathbf{l} + \frac{1}{N} \mathbf{l}^T \frac{\partial V_0}{\partial \mathbf{y}} + o(1/N) \right),$$

and similarly

$$W_l \left(\mathbf{y} - \frac{\mathbf{l}}{N} \right) = W_l(\mathbf{y}) - \frac{\mathbf{l}^T}{N} \frac{\partial W_l}{\partial \mathbf{y}} + o(1/N).$$

Substituting into Eq. (14), and assuming that τ is sufficiently large for the right-hand side to be neglected, we obtain

$$\begin{aligned} & \sum_{l \in \mathcal{L}} W_l(\mathbf{y}) \left(\exp \left(\mathbf{l}^T \frac{\partial V}{\partial \mathbf{y}} \right) - 1 \right) \\ & + \frac{1}{N} \sum_{l \in \mathcal{L}} \exp \left(\mathbf{l}^T \frac{\partial V}{\partial \mathbf{y}} \right) \left(-\mathbf{l}^T \frac{\partial W_l}{\partial \mathbf{y}} + \mathbf{l}^T \frac{\partial V_0}{\partial \mathbf{y}} W_l(\mathbf{y}) - \frac{1}{2} \mathbf{l}^T \frac{\partial^2 V}{\partial \mathbf{y}^2} \mathbf{l} W_l(\mathbf{y}) \right) \\ & + o(1/N) = 0. \end{aligned} \quad (17)$$

Collecting together leading-order terms from Eq. (17), we have

$$\sum_{l \in \mathcal{L}} W_l(\mathbf{y}) \left(\exp \left(\mathbf{l}^T \frac{\partial V}{\partial \mathbf{y}} \right) - 1 \right) = 0. \quad (18)$$

That is, $V(\mathbf{y})$ satisfies the Hamilton–Jacobi equation $H \left(\mathbf{y}, \frac{\partial V}{\partial \mathbf{y}} \right) = 0$, where the Hamiltonian is defined to be $H(\mathbf{y}, \boldsymbol{\theta}) = \sum_{l \in \mathcal{L}} W_l(\mathbf{y}) \left(e^{\mathbf{l}^T \boldsymbol{\theta}} - 1 \right)$.

Collecting together second-order terms in Eq. (17) gives

$$\sum_{l \in \mathcal{L}} \exp \left(\mathbf{l}^T \frac{\partial V}{\partial \mathbf{y}} \right) \mathbf{l}^T \left(\left(\frac{\partial V_0}{\partial \mathbf{y}} - \frac{1}{2} \frac{\partial^2 V}{\partial \mathbf{y}^2} \mathbf{l} \right) W_l(\mathbf{y}) - \frac{\partial W_l}{\partial \mathbf{y}} \right) = 0, \quad (19)$$

a first-order linear partial differential equation to be solved for $V_0(\mathbf{y})$, once $V(\mathbf{y})$ has been found from Eq. (18). Equation (18) has been previously studied in numerous specific applications (e.g. Assaf and Meerson 2010; Clancy 2018; Dykman et al. 1994; Elgart and Kamenev 2004; Hindes and Schwartz 2016; Kamenev and Meerson 2008; Lindley et al. 2014). Equation (19) has been analysed in some detail in the $k = 1$ -dimensional case (Assaf and Meerson 2017), but does not seem to have been much studied in the context of multidimensional problems.

With the above standard results and general theory in hand, we now proceed to the proof of Theorem 1.

4 Proof of Asymptotic Persistence Time Formulae

4.1 Heterogeneous Susceptibilities and Exponentially Distributed Infectious Periods: The Main Body of the Quasi-Stationary Distribution

We will start by proving Theorem 1(i), corresponding to the case of heterogeneity in susceptibility but not infectiousness, so $\lambda = \mathbf{1}$, and focus initially on the case of exponentially distributed infectious periods. For the Markov chain model with transition rates given in Table 1, we aim to express the asymptotic form of the main body of the quasi-stationary distribution in the form (16). We thus need to find the constant K_N and the functions $V(\mathbf{y}), V_0(\mathbf{y})$. In this section and Sect. 4.2, we follow the approach described in Assaf and Meerson (2010) for $k = 1$ -dimensional processes, adapted to our multidimensional situation.

For the case $\lambda = \mathbf{1}$, the solution to Eq. (18) with boundary condition $V(\mathbf{y}^*) = 0$ is known, by a slight generalisation of formula (20) of Clancy (2018), to be

$$\begin{aligned}
 V(\mathbf{y}) = & \sum_{i=1}^k y_i (1 + \ln y_i - \ln(\beta \alpha_i \mu_i)) - \left(\sum_{i=1}^k y_i \right) \ln \left(\sum_{i=1}^k y_i \right) \\
 & + \sum_{i=1}^k (f_i - y_i) \ln(f_i - y_i) - \frac{D(\mathbf{1}, \boldsymbol{\mu})}{\beta} - \sum_{i=1}^k f_i \ln f_i \\
 & + \sum_{i=1}^k f_i \ln(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})). \tag{20}
 \end{aligned}$$

Next, to evaluate K_N , we consider the Taylor series expansion of formula (16) for \mathbf{y} in the vicinity of \mathbf{y}^* . Differentiating Eq. (20), we obtain

$$\frac{\partial V}{\partial y_i} = \ln \left(\frac{y_i}{\beta \alpha_i \mu_i (f_i - y_i) \left(\sum_{m=1}^k y_m \right)} \right), \tag{21}$$

so that in particular, $\left. \frac{\partial V}{\partial \mathbf{y}} \right|_{\mathbf{y}=\mathbf{y}^*} = \mathbf{0}$. This is as one would expect, since for large N the quasi-stationary distribution has its mode at $\mathbf{x} = N\mathbf{y}^*$. Recalling the boundary conditions $V(\mathbf{y}^*) = V_0(\mathbf{y}^*) = 0$, then for $|\mathbf{x} - N\mathbf{y}^*| = O(\sqrt{N})$, Taylor series expansion of (16) yields

$$q_{\mathbf{x}} = K_N \exp \left(-\frac{1}{2N} (\mathbf{x} - N\mathbf{y}^*)^T \left. \frac{\partial^2 V}{\partial y_i \partial y_j} \right|_{\mathbf{y}=\mathbf{y}^*} (\mathbf{x} - N\mathbf{y}^*) + o(1) \right). \tag{22}$$

Denoting by S the matrix with entries

$$s_{ij} = \left. \frac{\partial^2 V}{\partial y_i \partial y_j} \right|_{\mathbf{y}=\mathbf{y}^*},$$

then (22) represents a multivariate Gaussian distribution with variance matrix NS^{-1} , normalisation of which requires that

$$K_N = \sqrt{\frac{\det(S)}{(2\pi N)^k}}. \quad (23)$$

To evaluate the determinant of S , we differentiate Eq. (21) to obtain

$$\frac{\partial^2 V}{\partial y_i \partial y_j} = \left(\frac{f_i}{y_i(f_i - y_i)} \right) \delta_{ij} - \frac{1}{\sum_m y_m}, \quad (24)$$

where δ_{ij} is the Kronecker delta. In particular, recalling that the components of \mathbf{y}^* are given by formula (9) with $\boldsymbol{\lambda} = \mathbf{1}$, then S has elements

$$s_{ij} = \frac{\partial^2 V}{\partial y_i \partial y_j} \Big|_{\mathbf{y}=\mathbf{y}^*} = \left(\frac{(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}{\alpha_i \mu_i f_i D(\mathbf{1}, \boldsymbol{\mu})} \right) \delta_{ij} - \frac{\beta}{D(\mathbf{1}, \boldsymbol{\mu})}.$$

It follows that

$$\begin{aligned} \det(S) &= \frac{1}{D(\mathbf{1}, \boldsymbol{\mu})^k} \left(1 - \beta \sum_{i=1}^k \frac{\alpha_i \mu_i f_i}{(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2} \right) \prod_{i=1}^k \frac{(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}{\alpha_i \mu_i f_i} \\ &= \frac{\beta}{D(\mathbf{1}, \boldsymbol{\mu})^{k-1}} \left(\prod_{i=1}^k \frac{(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}{\alpha_i \mu_i f_i} \right) \sum_{i=1}^k f_i \left(\frac{\alpha_i \mu_i}{1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})} \right)^2, \end{aligned} \quad (25)$$

the last line above following from Eq. (3) with $\boldsymbol{\lambda} = \mathbf{1}$. Substituting from (25) into (23), we obtain

$$K_N = \sqrt{\frac{\beta}{(2\pi N)^k D(\mathbf{1}, \boldsymbol{\mu})^{k-1}} \left(\prod_{i=1}^k \frac{(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}{\alpha_i \mu_i f_i} \right) \sum_{i=1}^k f_i \left(\frac{\alpha_i \mu_i}{1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})} \right)^2}. \quad (26)$$

To find $V_0(\mathbf{y})$, substitute the derivatives (21, 24) into Eq. (19) to obtain, after some simplification, the partial differential equation

$$\sum_{i=1}^k \left(\beta \mu_i (f_i - y_i) \left(\sum_{j=1}^k y_j \right) - \frac{y_i}{\alpha_i} \right) \left(\frac{f_i - 2y_i}{2y_i(f_i - y_i)} + \frac{1}{2 \sum_j y_j} - \frac{\partial V_0}{\partial y_i} \right) = 0. \quad (27)$$

Equation (27) will be satisfied if we can find $V_0(\mathbf{y})$ such that

$$\frac{\partial V_0}{\partial y_i} = \frac{f_i - 2y_i}{2y_i(f_i - y_i)} + \frac{1}{2 \sum_j y_j} \quad \text{for } i = 1, 2, \dots, k,$$

and hence we find that the solution of (27) subject to the boundary condition $V_0(\mathbf{y}^*) = 0$ is

$$V_0(\mathbf{y}) = (1/2) \ln \left(\frac{\beta}{D(\mathbf{1}, \boldsymbol{\mu})^{k+1}} \left(\sum_{i=1}^k y_i \right) \prod_{i=1}^k \frac{y_i (f_i - y_i) (1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}{\alpha_i \mu_i f_i^2} \right). \tag{28}$$

Our analysis of the WKB approximation (16) is now complete, in that we have found explicit formulae for $V(\mathbf{y})$, K_N and $V_0(\mathbf{y})$ in Eqs. (20, 26, 28), respectively, giving an approximation to the main body of the quasi-stationary distribution. The WKB *ansatz* is thus justified, in that we have been able to exhibit a solution of the assumed form (16). However, $V_0(\mathbf{y}) \rightarrow -\infty$ as $\mathbf{y} \rightarrow \mathbf{0}$, and the asymptotic form (16) is not valid in the tail of the distribution, where $|\mathbf{x}| = O(1)$. We address this in Sect. 4.2.

4.2 Heterogeneous Susceptibilities and Exponentially Distributed Infectious Periods: The Asymptotic Persistence Time Formula

For the SIS model with heterogeneous susceptibilities and exponentially distributed infectious periods, the mean extinction time τ from quasi-stationarity is given, from formula (15), by $\tau = \left(\sum_{i=1}^k (1/\alpha_i) q_{\mathbf{e}_i} \right)^{-1}$, where \mathbf{e}_i denotes the unit vector with i th element equal to 1. We are not yet in a position to evaluate the quasi-stationary probabilities $q_{\mathbf{e}_i}$, since expression (16) is not valid in the range $|\mathbf{x}| = O(1)$. We will therefore derive an alternative asymptotic formula for $q_{\mathbf{x}}$ that is valid in the range $|\mathbf{x}| = O(1)$, but is un-normalised. Normalisation may be achieved by matching our two approximations in the region where their domains of validity overlap. In order that we can carry out this matching, we first consider the Taylor series expansion of our existing approximation (16) about $\mathbf{y} = \mathbf{0}$. Since $V_0(\mathbf{y})$ diverges at $\mathbf{y} = \mathbf{0}$, this is quite technical, as we shall see. To deal with this, we define

$$\begin{aligned} \phi(\mathbf{y}) &= V_0(\mathbf{y}) - \frac{1}{2} \ln \left(\sum_{i=1}^k y_i \right) - \frac{1}{2} \sum_{i=1}^k \ln y_i \\ &= (1/2) \ln \left(\frac{\beta}{D(\mathbf{1}, \boldsymbol{\mu})^{k+1}} \prod_{i=1}^k \frac{(f_i - y_i) (1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}{\alpha_i \mu_i f_i^2} \right), \end{aligned}$$

so that $\phi(\mathbf{y})$ is well behaved at $\mathbf{y} = \mathbf{0}$, and (16) may be rewritten as

$$q_{\mathbf{x}} = \frac{K_N}{\sqrt{(\sum_i y_i) (\prod_i y_i)}} \exp(-NV(\mathbf{y}) - \phi(\mathbf{y}) + o(1)).$$

Taylor series expansion in the range $|\mathbf{x}| = o(\sqrt{N})$, together with substitution for K_N from (26), gives

$$\begin{aligned}
q_x &= \frac{K_N}{\sqrt{(\sum_i y_i) \prod_i y_i}} \exp \left(-NV(\mathbf{0}) - N\mathbf{y}^T \frac{\partial V}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\mathbf{0}} - \phi(\mathbf{0}) + o(1) \right) \\
&= K_N \sqrt{\frac{D(\mathbf{1}, \boldsymbol{\mu})^{k+1}}{\beta (\sum_i y_i) \prod_i y_i} \prod_{i=1}^k \frac{\alpha_i \mu_i f_i}{(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu}))^2}} \\
&\quad \times \exp \left(-NV(\mathbf{0}) - N\mathbf{y}^T \frac{\partial V}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\mathbf{0}} + o(1) \right) \\
&= D(\mathbf{1}, \boldsymbol{\mu}) \sqrt{\frac{N}{(2\pi)^k (\sum_i x_i) \prod_i x_i} \sum_{i=1}^k f_i \left(\frac{\alpha_i \mu_i}{1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})} \right)^2} \\
&\quad \times \exp \left(-NV(\mathbf{0}) - N\mathbf{y}^T \frac{\partial V}{\partial \mathbf{y}} \Big|_{\mathbf{y}=\mathbf{0}} + o(1) \right).
\end{aligned}$$

A further difficulty arises here, since the derivatives $\partial V/\partial y_i$ given by (21) are not well defined at $\mathbf{y} = \mathbf{0}$. We therefore consider approach to $\mathbf{y} = \mathbf{0}$ along a specific trajectory $y_i = \hat{y} \xi_i$ as $\hat{y} \rightarrow 0$, where $\xi_1, \xi_2, \dots, \xi_k > 0$ are fixed with $\sum_{i=1}^k \xi_i = 1$. Along this trajectory, $\frac{\partial V}{\partial y_i} \rightarrow \ln(\xi_i/\beta \alpha_i \mu_i f_i)$ as $\hat{y} \rightarrow 0$, and so with $\hat{x} = N\hat{y}$,

$$\begin{aligned}
q_x &= \beta^{\hat{x}} D(\mathbf{1}, \boldsymbol{\mu}) \prod_{i=1}^k \left(\frac{\alpha_i \mu_i f_i}{\xi_i} \right)^{\hat{x} \xi_i} \sqrt{\frac{N}{(2\pi)^k \hat{x}^{k+1} \prod_i \xi_i} \sum_{i=1}^k f_i \left(\frac{\alpha_i \mu_i}{1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})} \right)^2} \\
&\quad \times \exp(-NV(\mathbf{0}) + o(1)). \tag{29}
\end{aligned}$$

We now seek an approximation for q_x valid for $|x| = O(1)$. For the SIS model with heterogeneous susceptibilities, with the convention that $q_x = 0$ for $x \notin C$, the exact balance equation (13) may be written as

$$\begin{aligned}
&\frac{\beta}{N} \sum_{i=1}^k \mu_i (N_i - x_i + 1) \left(\sum_{j=1}^k x_j - 1 \right) q_{x-e_i} + \sum_{i=1}^k (1/\alpha_i)(x_i + 1) q_{x+e_i} \\
&\quad - \left(\frac{\beta}{N} \sum_{i=1}^k \sum_{j=1}^k \mu_i (N_i - x_i) x_j + \sum_{i=1}^k (1/\alpha_i) x_i \right) q_x \\
&= -(1/\tau) q_x \text{ for } x \in C. \tag{30}
\end{aligned}$$

Assuming, as before, that τ is sufficiently large for the right-hand side of Eq. (30) to be neglected, and taking the linear approximation to the left-hand side, which is valid in the required range $|x| = o(\sqrt{N})$, we obtain the asymptotic balance equation

$$\beta \sum_{i=1}^k \mu_i f_i \left(\sum_{j=1}^k x_j - 1 \right) q_{\mathbf{x}-\mathbf{e}_i} + \sum_{i=1}^k (1/\alpha_i)(x_i + 1)q_{\mathbf{x}+\mathbf{e}_i} - \left(\sum_{j=1}^k (\beta + (1/\alpha_j))x_j \right) q_{\mathbf{x}} = 0 \tag{31}$$

for $\mathbf{x} \in C$. Equation (31) corresponds to the (linear) branching process approximation discussed in Sect. 2, and we seek an un-normalised stationary solution. One component of the solution may be found by solving the detailed balance equations (Kelly 2011, chapter 1) corresponding to (31), and the other component by analogy with the solution for the case $k = 1$ given in Assaf and Meerson (2010). The solution thus obtained may be written in the form

$$q_{\mathbf{x}} = \frac{1}{\sum_i x_i} \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_{i=1}^k (\beta \mu_i f_i)^{x_i} \left(\prod_{i=1}^k \alpha_i^{x_i} - \prod_{i=1}^k \left(\frac{\alpha_i}{1 + \alpha_i D(\boldsymbol{\mu}, \mathbf{1})} \right)^{x_i} \right) \frac{1}{D(\boldsymbol{\mu}, \mathbf{1})\tau}$$

where the normalising factor τ remains to be found. For $|\mathbf{x}|$ large, noting that $D(\boldsymbol{\mu}, \mathbf{1}) > 0$ and applying Stirling’s formula to the factorial terms, we obtain

$$q_{\mathbf{x}} \sim \frac{1}{\sum_i x_i} \frac{(\sum_i x_i)!}{\prod_i x_i!} \prod_{i=1}^k (\beta \alpha_i \mu_i f_i)^{x_i} \frac{1}{D(\boldsymbol{\mu}, \mathbf{1})\tau} \sim \left(\sum_i x_i \right) \prod_{i=1}^k \left(\frac{\beta \alpha_i \mu_i f_i}{x_i} \right)^{x_i} \sqrt{\frac{1}{(2\pi)^{k-1} (\sum_m x_m) \prod_m x_m}} \left(\frac{1}{D(\boldsymbol{\mu}, \mathbf{1})\tau} \right). \tag{32}$$

Along the previously considered trajectory with $x_i = \hat{x} \xi_i$, expression (32) reduces to

$$q_{\mathbf{x}} \sim \beta^{\hat{x}} \prod_{i=1}^k \left(\frac{\alpha_i \mu_i f_i}{\xi_i} \right)^{\hat{x} \xi_i} \sqrt{\frac{1}{(2\pi)^{k-1} \hat{x}^{k+1} \prod_i \xi_i}} \left(\frac{1}{D(\boldsymbol{\mu}, \mathbf{1})\tau} \right), \tag{33}$$

and we can now match expressions (29) and (33) to obtain

$$\tau \sim \frac{1}{D(\mathbf{1}, \boldsymbol{\mu})D(\boldsymbol{\mu}, \mathbf{1})} \sqrt{\frac{2\pi}{N \sum_i f_i \left(\frac{\alpha_i \mu_i}{1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})} \right)^2}} \exp(NV(\mathbf{0})). \tag{34}$$

Noting from formula (20) that $V(\mathbf{0}) = \sum_i f_i \ln(1 + \alpha_i \mu_i D(\mathbf{1}, \boldsymbol{\mu})) - (D(\mathbf{1}, \boldsymbol{\mu})/\beta)$, and recalling that the minor outbreak probability in the case of exponentially distributed infectious periods is given by Eq. (10), we have now established Theorem 1(i) for the case of exponentially distributed infectious periods. Note that our assumption that τ is sufficiently large for the right-hand side of Eq. (14) to be neglected is thus justified in retrospect. Next, in Sect. 4.3, we extend the result to cover the case of heterogeneous susceptibilities with more general infectious period distributions.

4.3 Heterogeneous Susceptibilities and General Infectious Period Distributions

To allow for infectious period distributions more general than exponential, we will follow the approach developed in Ball et al. (2016) for the classic SIS model (the case $k = 1$). That is, we consider a restarted version of our model and apply an insensitivity result for stochastic networks from Zachary (2007).

Consider our model of Sect. 2 in the case $\lambda = \mathbf{1}$. Following Hernández-Suárez and Castillo-Chavez (1999), we introduce a regeneration step as follows. Whenever the process reaches the state $\mathbf{I} = \mathbf{0}$, it remains there for an exponentially distributed period of mean 1, after which a randomly chosen individual becomes infected, and the process then continues as before. The newly infected individual at the regeneration step is chosen to belong to group i with probability ρ_i for some distribution $\rho = (\rho_1, \rho_2, \dots, \rho_k)$.

In the terminology of Zachary (2007), each occasion when an individual in group i becomes infected corresponds to a class i arrival with associated workload distributed as T_i/α_i . When the process is in state $\mathbf{x} \in S$, individuals arrive in class i ($i = 1, 2, \dots, k$) at rate

$$\phi_{0i}(\mathbf{x}) = \begin{cases} \rho_i & \text{for } \mathbf{x} = \mathbf{0}, \\ \frac{\beta}{N} \mu_i (N_i - x_i) \sum_{j=1}^k x_j & \text{for } \mathbf{x} \in C. \end{cases}$$

While $x_i > 0$, the workload of each class i individual reduces at rate $\phi_i(\mathbf{x})/x_i$, where

$$\phi_i(\mathbf{x}) = x_i/\alpha_i.$$

The framework of Zachary (2007) allows for movement of individuals between classes, which we do not require here, so that in the notation of Zachary (2007) we take, for $i = 1, 2, \dots, k$,

$$\phi_{ij}(\mathbf{x}) = \begin{cases} \phi_i(\mathbf{x}) & \text{for } j = 0, \\ 0 & \text{for } j = 1, 2, \dots, k. \end{cases}$$

That is, on completion of its workload, an individual of class i leaves the system with probability $\phi_{i0}(\mathbf{x})/\phi_i(\mathbf{x}) = 1$ and moves to class $j = 1, 2, \dots, k$ with probability $\phi_{ij}(\mathbf{x})/\phi_i(\mathbf{x}) = 0$.

In Theorem 2 of Zachary (2007), it is shown that if a distribution $\pi(\mathbf{x})$ satisfies a certain partial balance condition (Equations (10) of Zachary 2007), together with an integrability condition (Equation (11) of Zachary 2007), then $\pi(\mathbf{x})$ is the stationary distribution of the numbers of individuals present in each class, regardless of the distributions of T_1, T_2, \dots, T_k . In order to find the stationary distribution $\pi(\mathbf{x})$, we return to the case of exponentially distributed infectious periods. If the restarted process is reversible, then it is straightforward to find $\pi(\mathbf{x})$ from the detailed balance equations (Kelly 2011, chapter 1). In order that the restarted process be reversible, consider Kolmogorov's criterion (Kelly 2011, chapter 1) applied to the sequence of states $\mathbf{0} \rightarrow \mathbf{e}_i \rightarrow \mathbf{e}_i + \mathbf{e}_j \rightarrow \mathbf{e}_j \rightarrow \mathbf{0}$ for $i \neq j$. That is, we require the product of the transition rates in one direction around the loop to equal the corresponding product of transition rates in

the opposite direction. This may be achieved by taking $\rho_i = \mu_i f_i$ for $i = 1, 2, \dots, k$. We can then solve the detailed balance equations, thereby verifying that the restarted process with this choice of ρ_i values is indeed reversible, with stationary distribution $\pi(\mathbf{x})$ satisfying

$$\pi(\mathbf{x}) = \frac{1}{\beta} \frac{(\sum_{j=1}^k x_j)!}{\sum_{j=1}^k x_j} \prod_{j=1}^k \left(\frac{\beta \alpha_j \mu_j}{N} \right)^{x_j} \binom{N_j}{x_j} \pi(\mathbf{0}) \quad \text{for } \mathbf{x} \in C, \tag{35}$$

so that

$$\pi(\mathbf{0}) = \left(1 + \frac{1}{\beta} \sum_{\mathbf{x} \in C} \frac{(\sum_{j=1}^k x_j)!}{\sum_{j=1}^k x_j} \prod_{j=1}^k \left(\frac{\beta \alpha_j \mu_j}{N} \right)^{x_j} \binom{N_j}{x_j} \right)^{-1}.$$

It is now straightforward to check that the balance equations (10) of Zachary (2007) are satisfied by $\pi(\mathbf{x})$ given by (35). The integrability condition (11) of Zachary (2007) is trivially satisfied since our state space is finite. It therefore follows from Theorem 2 of Zachary (2007) that $\pi(\mathbf{x})$ given by (35) is stationary for the numbers of infected individuals in our restarted process, whatever the distributions of the infectious periods T_i .

Now $\pi(\mathbf{0})$ is the expected proportion of time spent in state $\mathbf{I} = \mathbf{0}$ in the long term, which is equal to the expected proportion of time spent in state $\mathbf{I} = \mathbf{0}$ during one regenerative cycle. Denoting by τ_0 the expected regeneration time, being the time from one entry into state $\mathbf{I} = \mathbf{0}$ until the following entry into state $\mathbf{I} = \mathbf{0}$, then $\tau_0 = 1/\pi(\mathbf{0})$. For $i = 1, 2, \dots, k$, denote by τ_i the expected time for the process to hit state $\mathbf{I} = \mathbf{0}$ after having been initiated with a single newly infected individual in group i . Notice that the values of $\tau_1, \tau_2, \dots, \tau_k$ are the same for the restarted process as for the original process and that

$$\tau_0 = 1 + \sum_{i=1}^k \mu_i f_i \tau_i,$$

so that

$$\sum_{i=1}^k \mu_i f_i \tau_i = \frac{1}{\pi(\mathbf{0})} - 1. \tag{36}$$

Recall from Sect. 3 that in the large population limit, the initial stage of an outbreak initiated by a single newly infected individual may be approximated by a multitype branching process. Recall that ω_i ($i = 1, 2, \dots, k$) denotes the probability that this branching process, initiated by a single group i individual, produces only a finite number of progeny. Denote by ζ_i the expected time to extinction of the process conditional upon a finite number of progeny being produced, and by σ_i the expected time taken to

attain quasi-stationarity, starting from a single group i individual, given that a major outbreak occurs. Then, in the large population limit, we have

$$\tau_i \sim \omega_i \zeta_i + (1 - \omega_i)(\sigma_i + \tau).$$

Now ζ_i, σ_i are negligible in comparison with τ , and so

$$\tau_i \sim (1 - \omega_i)\tau. \quad (37)$$

The above somewhat heuristic argument is made rigorous for the case $k = 1$ in Appendix B of Ball et al. (2014).

Recalling from Sect. 3 that $\omega = \sum_i \mu_i f_i \omega_i$, it follows from (37) that

$$\sum_{i=1}^k \mu_i f_i \tau_i \sim (1 - \omega)\tau.$$

Now from Eqs. (35) and (36),

$$\sum_i \mu_i f_i \tau_i = \frac{1}{\beta} \sum_{\mathbf{x} \in \mathcal{C}} \frac{\left(\sum_{j=1}^k x_j\right)!}{\sum_{j=1}^k x_j} \prod_{j=1}^k \left(\frac{\beta \alpha_j \mu_j}{N}\right)^{x_j} \binom{N_j}{x_j}, \quad (38)$$

and hence

$$\tau \sim \frac{1}{\beta(1 - \omega)} \sum_{\mathbf{x} \in \mathcal{C}} \frac{\left(\sum_{j=1}^k x_j\right)!}{\sum_{j=1}^k x_j} \prod_{j=1}^k \left(\frac{\beta \alpha_j \mu_j}{N}\right)^{x_j} \binom{N_j}{x_j}.$$

In the case of exponentially distributed infectious periods, we know from Eq. (10) that $\beta(1 - \omega) = D(\boldsymbol{\mu}, \mathbf{1})$, and so comparing with expression (34), we obtain our general result (6), and Theorem 1(i) is now proved.

For the homogeneous population case ($k = 1$), Ball et al. (2016) approximated the sum (38) directly, and from this derived an asymptotic approximation to the mean persistence time τ . For $k > 1$, it does not seem straightforward to approximate (38) directly, so that we have proceeded more indirectly via the methods of Sects. 4.1–4.2.

4.4 Heterogeneous Infectiousness and Exponentially Distributed Infectious Periods

As Clancy (2018) points out, see also Wilkinson and Sharkey (2013), it follows from the network duality results of Harris (1976) and Holley and Liggett (1975) that provided infectious periods are exponentially distributed, the value of τ is unchanged if we interchange the roles of $\boldsymbol{\lambda}, \boldsymbol{\mu}$. Theorem 1(ii) then follows immediately from Theorem 1(i).

5 Applications

5.1 Performance of the Mean Persistence Time Approximation

Figure 1 shows values of $\ln \tau$ computed from the exact formula (13) compared with values computed from our asymptotic formulae (6, 7), for the case of exponentially distributed infectious periods with heterogeneous susceptibility or infectiousness. We see that the approximation is quite accurate for the range of parameter values considered, and (although the effect is a little hard to make out from Fig. 1) the accuracy of the approximation improves as population size N increases, as expected. It is also apparent that the approximation performs better as R_0 increases. Since our methods are valid under the condition that $R_0 > 1$, it is perhaps not surprising that the accuracy of the approximation decreases as R_0 approaches 1. We note that the approximation appears to consistently err on the side of slightly underestimating the mean persistence time.

The effects of different infectious period distributions are illustrated in Fig. 2, where we consider both exponentially distributed infectious periods and infectious periods which are constant (non-random). For the case of exponentially distributed infectious periods, exact values of $\ln \tau$ are again computed from formula (13). For the case of constant infectious periods, we compare our asymptotic formula (6) with the results of Monte Carlo simulation. Specifically, for each N value we simulated 1000 realizations of the process started close to the deterministic endemic equilibrium, at the point $(\lfloor Ny_1^* \rfloor, \lfloor Ny_2^* \rfloor, \dots, \lfloor Ny_k^* \rfloor)$, where $\lfloor x \rfloor$ denotes the integer part of x . To

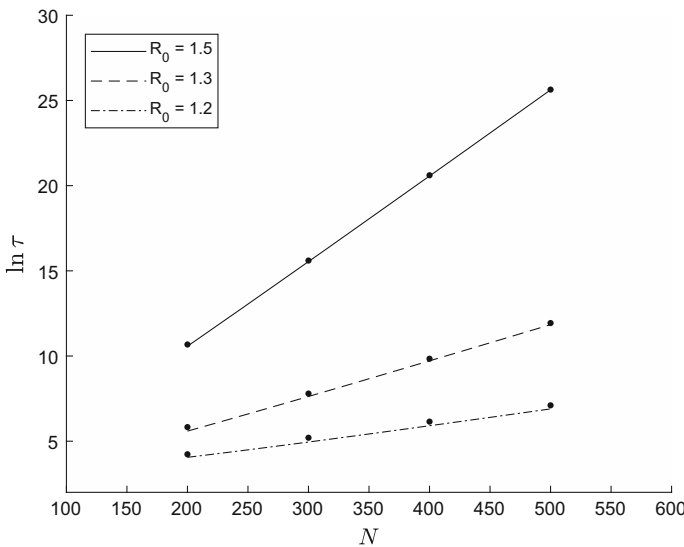


Fig. 1 Values of $\ln \tau$ plotted against population size N for the case of exponentially distributed infectious periods, with three different values for the basic reproduction number R_0 . Fixed parameter values $k = 2$, $f = (0.5, 0.5)$, $\alpha = (1, 1)$, $\lambda = (1, 1)$, $\mu = \frac{1}{3}(5, 1)$ (or equivalently $\lambda = \frac{1}{3}(5, 1)$, $\mu = (1, 1)$). Lines are computed from the asymptotic formula (6); dots are the true values of $\ln \tau$ computed from Eq. (13)

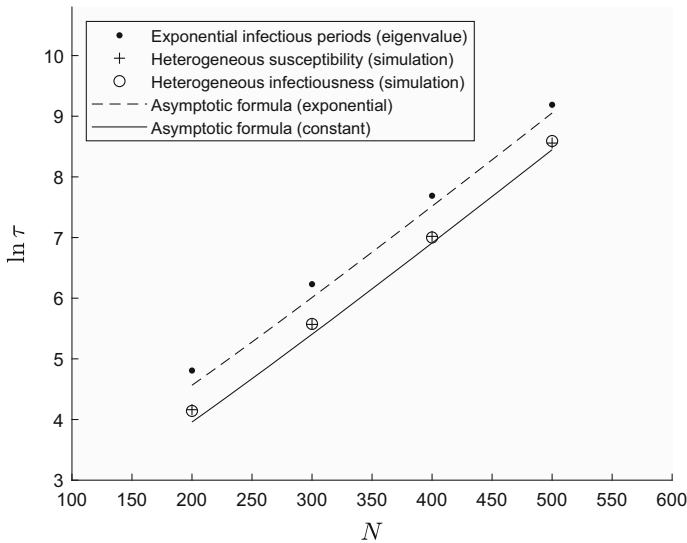


Fig. 2 Values of $\ln \tau$ plotted against population size N , showing the effect of the infectious period distribution upon the mean persistence time of infection τ . Fixed parameter values $k = 2$, $f = (0.5, 0.5)$, $\alpha = (1, 1)$, $R_0 = 1.3$. Dots computed from the eigenvalue equation (13) with $\lambda = (1, 1)$, $\mu = \frac{2}{51}(50, 1)$; crosses ('Heterogeneous susceptibility') computed via Monte Carlo simulation with $\lambda = (1, 1)$, $\mu = \frac{2}{51}(50, 1)$ and constant infectious periods; circles ('Heterogeneous infectiousness') computed via Monte Carlo simulation with $\lambda = \frac{2}{51}(50, 1)$, $\mu = (1, 1)$ and constant infectious periods; dashed line computed from formula (6) with $\lambda = (1, 1)$, $\mu = \frac{2}{51}(50, 1)$ and exponentially distributed infectious periods; solid line computed from formula (6) with $\lambda = (1, 1)$, $\mu = \frac{2}{51}(50, 1)$ and constant infectious periods

indicate the dependence of the expected persistence time upon parameters, we write $\tau = \tau_\alpha(\lambda, \mu)$ and denote by $\tilde{\tau}_\alpha(\mathbf{1}, \mu)$ the approximation to $\tau_\alpha(\mathbf{1}, \mu)$ given by the right-hand side of formula (6). For the case of heterogeneous susceptibility, we allowed a burn-in period of $t_0 = 0.3\tilde{\tau}_\alpha(\mathbf{1}, \mu)$ for the process to attain quasi-stationarity, after which the process was allowed to continue until either infection became extinct or time $t_{\max} = 1.8\tilde{\tau}_\alpha(\mathbf{1}, \mu)$ was reached. For the case of heterogeneous infectiousness, we took $t_0 = 0.3\tilde{\tau}_\alpha(\mathbf{1}, \lambda)$ and $t_{\max} = 1.8\tilde{\tau}_\alpha(\mathbf{1}, \lambda)$. We then computed the maximum likelihood estimate of $\ln \tau$ as described in Section 6 of Clancy (2018). Histograms of the observed extinction times were compared visually with the probability density function of the exponential distribution with rate parameter estimated by maximum likelihood and seen to fit reasonably well, providing reassurance that the burn-in period t_0 was sufficient. Although the network duality results of Harris (1976) and Holley and Liggett (1975) apply only to exponentially distributed infectious periods, we observe that, as noted in Clancy (2018), Monte Carlo estimates of τ are essentially identical for the cases of heterogeneous infectiousness and (corresponding) heterogeneous susceptibility, even with constant infectious periods. We observe that the error in our approximating formula is much the same when infectious periods are constant as when infectious periods are exponentially distributed, with mean persistence time again being consistently somewhat underestimated by our asymptotic formula (6).

5.2 Superspreaders and the Effects of Different Forms of Heterogeneity

For many outbreaks of infection, it is thought that a small number of infected individuals, often referred to as ‘superspreaders’, are responsible for a disproportionate amount of pathogen transmission (e.g. Lau et al. 2017; Plowright et al. 2017; Yates et al. 2006). This may arise through a variety of mechanisms. Two possibilities are that some individuals infect at a much higher rate than others (represented in our model by heterogeneity in λ); or that some individuals remain infectious for much longer than others (represented by heterogeneity in α). Our results allow us to study the effect of either of these forms of superspreading, comparing alternative forms of heterogeneity to one another, and also comparing with a matched (having the same value for R_0) homogeneous population, as follows.

To make explicit the dependence upon parameters, we now write $D(\lambda, \mu) = D_\alpha(\lambda, \mu)$, and similarly $\tau = \tau_\alpha(\lambda, \mu)$. We consider only the case of exponentially distributed infectious periods, in order that we can apply formula (7), and we impose the constraint $\sum_i \alpha_i f_i = 1$ (that is, we re-scale time so that the unit of time is the mean infectious period across the whole population). Let $\eta = (\eta_1, \eta_2, \dots, \eta_k)$ be any vector with positive components satisfying $\sum_i \eta_i f_i = 1$, representing the heterogeneity. Denoting by $\tilde{\tau}_\alpha(\lambda, \mathbf{1})$ the approximation to $\tau_\alpha(\lambda, \mathbf{1})$ given by the right-hand side of (7), then from Eq. (3), we have that $D_1(\eta, \mathbf{1}) = \beta - 1$ and $D_1(\mathbf{1}, \eta) = D_\eta(\mathbf{1}, \mathbf{1})$, and hence formula (7) implies that

$$\frac{\tilde{\tau}_1(\eta, \mathbf{1})}{\tilde{\tau}_\eta(\mathbf{1}, \mathbf{1})} = \frac{D_\eta(\mathbf{1}, \mathbf{1})}{\beta - 1}.$$

Now it follows from Jensen’s inequality applied to Eq. (3) that $D_\eta(\mathbf{1}, \mathbf{1}) \leq \beta - 1$, and so

$$\tilde{\tau}_1(\eta, \mathbf{1}) \leq \tilde{\tau}_\eta(\mathbf{1}, \mathbf{1}).$$

That is, for sufficiently large N , heterogeneity in levels of infectiousness leads to a shorter expected persistence time than corresponding heterogeneity in the lengths of infectious periods.

It was shown in Theorem 2(i) of Clancy (2018) that when $\alpha = \mu = \mathbf{1}$, the leading-order constant A in formula (1) is maximised, for a given value of R_0 , when $\lambda = \mathbf{1}$. This implies that for sufficiently large N , $\tau_1(\mathbf{1}, \mathbf{1}) \geq \tau_1(\eta, \mathbf{1})$, and since $D_1(\mathbf{1}, \eta) = D_\eta(\mathbf{1}, \mathbf{1})$ it also follows that for sufficiently large N , $\tau_1(\mathbf{1}, \mathbf{1}) \geq \tau_\eta(\mathbf{1}, \mathbf{1})$. That is, for a sufficiently large population, heterogeneity in either levels of infectiousness or infectious period durations reduces the expected persistence time of infection in the population, compared to a corresponding homogeneous population.

These effects are illustrated in Fig. 3, in which we take 10% of the population to generate up to 50 times more potentially infectious contacts per infectious period than the remaining 90% of the population. The leftmost point of each curve, at $\eta = (1, 1)$, corresponds to a homogeneous population. We see that as the degree of heterogeneity increases, the mean persistence time decreases, while the difference between the effects of the two types of heterogeneity increases. With maximal heterogeneity represented

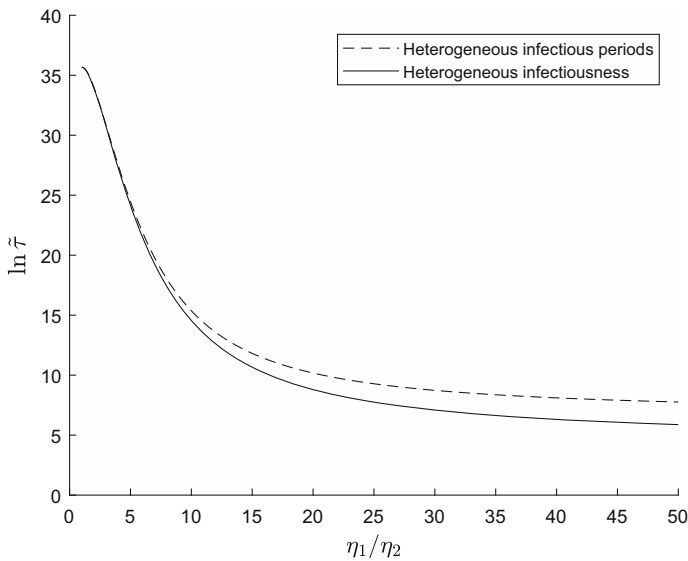


Fig. 3 Effect of different types of heterogeneity upon the mean persistence time τ . Plotted lines show values of $\ln \tilde{\tau}$, where $\tilde{\tau}$ denotes the approximation to mean persistence time given by formula (7). Fixed parameter values $k = 2$, $\mathbf{f} = (0.1, 0.9)$, $\boldsymbol{\mu} = (1, 1)$, $R_0 = 1.5$, $N = 500$. Solid line computed from formula (7) with $\boldsymbol{\alpha} = (1, 1)$, $\boldsymbol{\lambda} = \boldsymbol{\eta}$; dashed line computed from formula (7) with $\boldsymbol{\alpha} = \boldsymbol{\eta}$, $\boldsymbol{\lambda} = (1, 1)$

by $\hat{\boldsymbol{\eta}} = (1/f_1, 0)$, the limiting ratio is given by $\tilde{\tau}_{\hat{\boldsymbol{\eta}}}(\mathbf{1}, \mathbf{1})/\tilde{\tau}_1(\hat{\boldsymbol{\eta}}, \mathbf{1}) = (\beta - 1)/D_{\hat{\boldsymbol{\eta}}}(\mathbf{1}, \mathbf{1}) = 1/f_1$.

It is worth noting that when heterogeneity is in infectiousness, the fact that we have held R_0 constant across Fig. 3 implies that the endemic equilibrium point \mathbf{y}^* also remains fixed. In fact, from Eqs. (3, 9), it is immediate that when $\boldsymbol{\alpha} = \boldsymbol{\mu} = \mathbf{1}$, we have $y_i^* = (1 - (1/R_0))f_i$ for $i = 1, 2, \dots, k$, for any $\boldsymbol{\lambda}$. Now Theorem 11 of Clancy and Pearce (2013) demonstrates, via a multivariate normal approximation, that greater heterogeneity in infectiousness corresponds to greater variability in the quasi-stationary distribution. Thus, in this case, the decrease in mean persistence time observed across Fig. 3 (solid curve) corresponds to an increase in variability of the quasi-stationary distribution leading to larger fluctuations (around the same equilibrium point) and hence faster extinction of infection. However, the solid line in Fig. 3 could equally be interpreted as corresponding to heterogeneous susceptibilities, with exponentially distributed infectious periods and $\boldsymbol{\alpha} = \boldsymbol{\lambda} = \mathbf{1}$, $\boldsymbol{\mu} = \boldsymbol{\eta}$ in formula (6). When heterogeneity is in susceptibilities, it has been shown (Clancy and Pearce 2013, Theorem 10) that the overall endemic prevalence level $y^* = \sum_{i=1}^k y_i^*$ decreases with increasing heterogeneity. That is, when heterogeneity is in susceptibilities, the decrease in persistence time observed across Fig. 3 (solid curve) accompanies a corresponding decrease in overall endemic prevalence level. Furthermore, from formulae (3, 9), it is apparent that the endemic equilibrium point \mathbf{y}^* for the case $\boldsymbol{\alpha} = \boldsymbol{\eta}$, $\boldsymbol{\lambda} = \boldsymbol{\mu} = \mathbf{1}$ is the same as for the case $\boldsymbol{\mu} = \boldsymbol{\eta}$, $\boldsymbol{\alpha} = \boldsymbol{\lambda} = \mathbf{1}$. Consequently, when heterogeneity is in infectious period durations, the decrease in persistence time observed across Fig. 3 (dashed curve) again accompanies a corresponding decrease in overall endemic prevalence level.

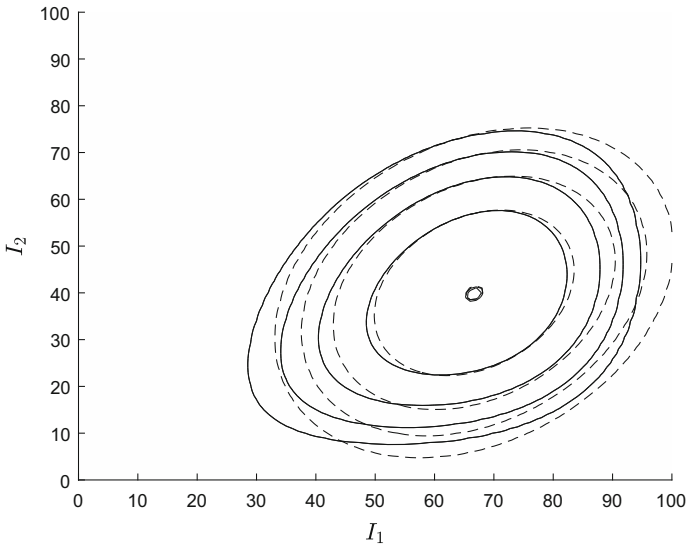


Fig. 4 Contour plots of the quasi-stationary distribution q and approximations. Contour levels correspond to probabilities $5 \times (10^{-3}, 10^{-5}, 10^{-7}, 10^{-9}, 10^{-11})$. Parameter values $k = 2$, $f = (0.5, 0.5)$, $\alpha = (1, 1)$, $\mu = \frac{1}{2}(3, 1)$, $\lambda = (1, 1)$, $R_0 = 2.5$, $N = 200$. Solid contours represent both the exact quasi-stationary distribution computed from Eq. (13) and the WKB approximation (16), which are indistinguishable. Dashed contours represent the Gaussian approximation (22)

5.3 Approximating the Quasi-Stationary Distribution

En route to our analysis of mean time to extinction, obtained via approximation of the tail of the quasi-stationary distribution, we also obtained an approximation for the main body of the quasi-stationary distribution, at least in the case of heterogeneous susceptibilities ($\lambda = \mathbf{1}$) and exponentially distributed infectious periods, from the WKB formula (16) with K_N , $V(y)$, $V_0(y)$ given by Eqs. (26, 20, 28), respectively. This new approximation may be regarded as a refinement of the multivariate Gaussian approximation (22), previously derived via an approximating diffusion (Ornstein–Uhlenbeck) process in Section 6 of Clancy and Pearce (2013). Figure 4 shows contour plots of the exact quasi-stationary distribution q , obtained from Eq. (13), and the two approximations (16, 22) for a population of size $N = 200$ consisting of two equal-sized groups, with group 1 individuals being three times as susceptible to infection as group 2 individuals. Our new approximation is clearly a great improvement upon the Gaussian approximation, particularly away from the endemic equilibrium point Ny^* , the contours of the WKB approximation being indistinguishable from those of the exact solution.

Note that unlike the mean persistence time τ , the quasi-stationary distribution q for the case of heterogeneous infectiousness cannot be obtained from the heterogeneous susceptibilities solution simply by interchanging the roles of λ , μ . In fact, even the location of the endemic equilibrium point Ny^* , corresponding to the mode of the quasi-stationary distribution, is not maintained under this transformation (Clancy and

Pearce 2013). On the other hand, for the case of heterogeneous susceptibilities, the WKB approximation (16) approximates the stationary distribution of the restarted process studied in Sect. 4.3 and hence may be used to approximate the quasi-stationary distribution of numbers of infected individuals even when infectious periods are not exponentially distributed.

5.4 Network Interpretation

The model of Sect. 2 may be interpreted as describing an infection spreading through a directed network under the so-called annealed network approximation, as outlined in Clancy (2018). Briefly, we suppose that individuals are assigned an in-degree d_{in} and out-degree d_{out} according to some joint probability mass function $p(d_{\text{in}}, d_{\text{out}})$ satisfying $E[d_{\text{in}}] = E[d_{\text{out}}]$. We suppose that there are a finite number k of $(d_{\text{in}}, d_{\text{out}})$ pairs having nonzero probability. We further assume that the network is uncorrelated, that is, with no correlations between degrees of neighbouring individuals. We define a bijective function $j(d_{\text{in}}, d_{\text{out}}) : \mathbb{Z}_+^2 \rightarrow \{1, 2, \dots, k\}$, so that any individual having degrees $(d_{\text{in}}, d_{\text{out}})$ belongs to group $j(d_{\text{in}}, d_{\text{out}})$. Denote by $d_{\text{in}}(j)$, $d_{\text{out}}(j)$ the in and out degrees, respectively, of a group j individual, and by κ the rate at which infection transmits along each link from an infectious individual to a susceptible individual. This network model may be approximated by our multigroup model of Sect. 2 by setting, for $j = 1, 2, \dots, k$,

$$\begin{aligned}\beta &= \kappa E[d_{\text{out}}], \\ f_j &= p(d_{\text{in}}(j), d_{\text{out}}(j)), \\ \mu_j &= d_{\text{in}}(j)/E[d_{\text{in}}], \\ \lambda_j &= d_{\text{out}}(j)/E[d_{\text{out}}].\end{aligned}$$

The undirected version of the above annealed network approximation (with $\lambda = \mu$) has been studied in Hindes and Schwartz (2016, 2017) in terms of the leading-order constant A in expression (1), evaluated via numerical solution of the Hamilton–Jacobi equation (18). Our results (6, 7), as well as being much quicker and more straightforward to evaluate, are thus considerably more precise, although applicable only to rather restricted classes of directed networks. Specifically, the assumption that $\lambda = \mathbf{1}$ corresponds to every individual having the same out-degree, whereas $\mu = \mathbf{1}$ corresponds to every individual having the same in-degree. Nevertheless, it is remarkable to be able to obtain such simple and precise results as formulae (6, 7) even for such a restricted class of networks.

6 Discussion and Further Work

The main contribution of this paper has been to provide simple explicit formulae (6, 7) for the mean persistence time, in the large population limit, of the heterogeneous population SIS infection model described in Sect. 2. The only infection model for which such a result has previously been available is the homogeneous population version of

this same model, corresponding to the case $k = 1$. Explicit formulae are particularly valuable here since numerical solution of the Hamilton–Jacobi equation (18) generally requires the solution of a high-dimensional system of ordinary differential equations subject to boundary conditions at times $t = -\infty$ and $t = +\infty$, so that even to obtain the leading-order constant A in formula (1) can be very challenging, in the absence of an explicit formula. We have shown, in Sect. 5.2, how such explicit formulae may be used to study qualitative features such as the effects of different types of heterogeneity upon the persistence time of infection. Additionally, in the course of our analysis, we have obtained a new and accurate approximation to the quasi-stationary distribution of the process, which determines long-term behaviour prior to eventual extinction of infection; see Sect. 5.3. Our model may also be interpreted as an approximate model for infection spreading on a directed network, as described in Sect. 5.4. While our results thus represent substantial progress, many open questions remain.

Firstly, while Theorem 1(i) for the case of heterogeneous susceptibilities allows for any infectious period distribution, Theorem 1(ii) for heterogeneous infectiousness applies only with exponentially distributed infectious periods. From numerical work, including that presented in Fig. 2, it seems likely that a result corresponding to formula (6), generalising formula (7) to allow for any infectious period distribution, does indeed apply when heterogeneity is in infectiousness, but we have not been able to prove this because the network duality results of Harris (1976) and Holley and Liggett (1975) apply only provided that infectious periods are exponentially distributed.

Secondly, although our model of Sect. 2 allows for heterogeneity in susceptibilities and infectiousness simultaneously, Theorem 1 requires that only one of these forms of heterogeneity be present. It would be of great interest to find a corresponding formula allowing for both forms of heterogeneity simultaneously. In particular, this would allow our result to be applied to infections spreading on a much more general class of networks, including undirected networks, as studied via the annealed network approximation in Hindes and Schwartz (2016), requiring $\lambda = \mu$. More generally, one could allow some quite general matrix of contact rates $\{\beta_{ij}\}$, rather than restricting as we have to contact rates that factorise as $\beta_{ij} = \beta\lambda_i\mu_j$, in order to study phenomena such as assortative/disassortative mixing (Clancy and Pearce 2013). Unfortunately, there is no reason to expect that explicit formulae such as (6, 7) exist at all in such cases, even under the assumption of exponentially distributed infectious periods. In particular, we have only been able to find an explicit solution $V(y)$ to the Hamilton–Jacobi equation (18) in the case $\lambda = \mathbf{1}$ (formula (20); see also Clancy 2018). Consequently, to evaluate the leading-order constant A in formula (1) generally requires numerical solution of Eq. (18), as implemented for the case $\lambda = \mu$ in Hindes and Schwartz (2016). Further, without an explicit formula for $V(y)$, Eqs. (19, 23) cannot be used to find explicit expressions for $V_0(y)$, K_N . One could, in principle, evaluate K_N and $V_0(y)$ numerically, as was done in Black and McKane (2011) for a particular SIR infection model, and thereby obtain the WKB approximation for the main body of the quasi-stationary distribution, corresponding to our result illustrated in Fig. 4. However, our asymptotic formulae for mean persistence time depend upon approximating the tail of the quasi-stationary distribution, and it is not clear how the matching procedure of Sect. 4.2 could be carried through numerically, without explicit formulae to match.

Finally, it would be of great interest to obtain explicit formulae such as (6, 7) for infection models incorporating features such as disease-induced immunity (SIR models), latent periods and demographic processes of birth, death and migration. For such more sophisticated models, a common strategy has been to resort to approximating the quasi-stationary distribution by a multivariate Gaussian distribution (obtained as the stationary distribution of an approximating Ornstein–Uhlenbeck diffusion process) and then substitute this Gaussian approximation into the right-hand side of Eq. (15) to obtain an approximation to the mean persistence time τ . For instance, Nåsell (1999) made use of this approach in studying an infection model incorporating demographic processes and disease-induced immunity (with exponentially distributed infectious periods), following on from which Andersson and Britton (2000) extended the model to include latency, with latent periods and infectious periods each being allowed to follow Erlang distributions. Unfortunately, while this approach can give some rough qualitative indication of the effect of model parameters upon persistence times, the numerical approximation to τ thus obtained is known to be extremely inaccurate (Clancy and Tjia 2018). Indeed, as pointed out in Nåsell (1999), this approximation does not yield correct N -dependence in the large population limit; specifically, the approximation which appears as equation (2.15) of Nåsell (1999) takes the form $\tau \approx c\sqrt{N} \exp(aN)$ for some constants a, c , in contrast to the asymptotic form (1) obtained in van Herwaarden and Grasman (1995). In fact, even the value of $\lim_{N \rightarrow \infty} (\ln \tau)/N$, given by the leading-order constant A in formula (1), is not correctly reproduced via this approach (Clancy and Tjia 2018; Doering et al. 2005). For these reasons, Andersson and Britton (2000) noted that their approximation ‘should only serve as a qualitative guidance and not be relied on in detail’. In view of this failure of the Ornstein–Uhlenbeck approximating diffusion approach, the approach that we have employed, via WKB approximation, can be seen to be of great potential value, yielding as it does the correct asymptotic behaviour. However, it is much more difficult to obtain explicit formulae via this approach, and indeed there seems no reason to expect that explicit formulae such as (6, 7) will exist in general. Consequently, much of the work to date employing this approach for models in dimensions $k > 1$, by many authors, has consisted essentially of numerical evaluation of the leading-order constant A in formula (1). One exception is the SEIS model in a homogeneous population—that is, the classic SIS model of Weiss and Dishon (1971), extended to allow for a latent (‘exposed’) period. It was shown in Clancy and Tjia (2018) that for this SEIS model, with latent periods and infectious periods each allowed to follow Erlang distributions, the value of $\lim_{N \rightarrow \infty} (\ln \tau)/N$ is given by $A = (1/R_0) - 1 + \ln R_0$, exactly as for the classic SIS model (Andersson and Djehiche 1998). Thus, while the presence of a latent period may impact substantially upon the mean persistence time, this impact is restricted to the prefactor constant C in formula (1), at least for this particular model. In general, while explicit formulae for the constants A, C in (1) may be too much to hope for, in cases where the leading-order constant A can only be evaluated numerically a natural next step may be to seek general methods for evaluating the prefactor constant C numerically. Even here, as mentioned in the previous paragraph, the difficulties to be overcome remain substantial.

Acknowledgements It is a pleasure to thank Frank Ball and Heiko Gimperlein for helpful discussions.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Andersson H, Britton T (2000) Stochastic epidemics in dynamic populations: quasi-stationarity and extinction. *J Math Biol* 41:559–580
- Andersson H, Djehiche B (1998) A threshold limit theorem for the stochastic logistic epidemic. *J Appl Probab* 35:662–670
- Assaf M, Meerson B (2010) Extinction of metastable stochastic populations. *Phys Rev E* 81:021116
- Assaf M, Meerson B (2017) WKB theory of large deviations in stochastic populations. *J Phys A Math Theor* 50:263001
- Ball FG, Britton T, Neal P (2014) On expected durations of birth–death processes, with applications to branching processes and SIS epidemics. [arXiv:1408.0641](https://arxiv.org/abs/1408.0641)
- Ball FG, Britton T, Neal P (2016) On expected durations of birth–death processes, with applications to branching processes and SIS epidemics. *J Appl Probab* 53:203–215
- Black AJ, McKane AJ (2011) WKB calculation of an epidemic outbreak distribution. *J Stat Mech Theor Exp* 2011:P12006
- Clancy D (2015) Generality of endemic prevalence formulae. *Math Biosci* 269:30–36
- Clancy D (2018) Persistence time of SIS infections in heterogeneous populations and networks. *J Math Biol* 77:545–570
- Clancy D, Pearce CJ (2013) The effect of population heterogeneities upon spread of infection. *J Math Biol* 67:963–987
- Clancy D, Tjia E (2018) Approximating time to extinction for endemic infection models. *Methodol Comput Appl Probab*. <https://doi.org/10.1007/s11009-018-9621-8>
- Darroch JN, Seneta E (1967) On quasi-stationary distributions in absorbing continuous-time finite Markov chains. *J Appl Probab* 4:192–196
- Doering CR, Sargsyan KV, Sander LM (2005) Extinctions times for birth–death processes: exact results, continuum asymptotics, and the failure of the Fokker–Planck approximation. *Multiscale Model Simul* 3:283–299
- Dykman MI, Mori E, Ross J, Hunt PM (1994) Large fluctuations and optimal paths in chemical kinetics. *J Chem Phys* 100:5735–5750
- Elgart V, Kamenev A (2004) Rare event statistics in reaction–diffusion systems. *Phys Rev E* 70:041106
- Ethier SN, Kurtz TG (2005) *Markov processes: characterization and convergence*. Wiley, New York
- Forgoston E, Bianco S, Shaw LB, Schwartz IB (2011) Maximal sensitive dependence and the optimal path to epidemic extinction. *Bull Math Biol* 73:495–514
- Harris TE (1976) On a class of set-valued Markov processes. *Ann Probab* 4:175–194
- Hernández-Suárez CM, Castillo-Chavez C (1999) A basic result on the integral for birth–death Markov processes. *Math Biosci* 161:95–104
- Hindes J, Schwartz IB (2016) Epidemic extinction and control in heterogeneous networks. *Phys Rev Lett* 117:028302
- Hindes J, Schwartz IB (2017) Epidemic extinction paths in complex networks. *Phys Rev E* 95:052317
- Holley A, Liggett T (1975) Ergodic theorems for weakly interacting infinite systems and the voter model. *Ann Probab* 3:643–663
- Kamenev A, Meerson B (2008) Extinction of an infectious disease: a large fluctuation in a nonequilibrium system. *Phys Rev E* 77:061107
- Kelly F (2011) *Reversibility and stochastic networks*. Cambridge University Press, Cambridge
- Lajmanovich A, Yorke JA (1976) A deterministic model for gonorrhea in a nonhomogeneous population. *Math Biosci* 28:221–236
- Lau MSY, Dalziel BD, Funk S, McClelland A, Tiffany A, Riley S, Metcalf CJE, Grenfell BT (2017) Spatial and temporal dynamics of superspreading events in the 2014–15 West Africa Ebola epidemic. *PNAS* 114:2337–2342

-
- Lindley BS, Schwartz IB (2013) An iterative action minimizing method for computing optimal paths in stochastic dynamical systems. *Physica D Nonlinear Phenom* 255:22–30
- Lindley BS, Shaw LB, Schwartz IB (2014) Rare-event extinction on stochastic networks. *Europhys Lett* 108:58008
- Marshall AW, Olkin I, Arnold BC (2011) *Inequalities: theory of majorization and its applications*. Springer, Berlin
- Nåsell I (1999) On the time to extinction in recurrent epidemics. *J R Stat Soc B* 61:309–330
- Nold A (1980) Heterogeneity in disease-transmission modelling. *Math Biosci* 52:227–240
- Plowright RK, Manlove KR, Besser TE, Páez DJ, Andrews KR, Matthews PE, Waits LP, Hudson PJ, Cassirer EF (2017) Age-specific infectious period shapes dynamics of pneumonia in bighorn sheep. *Ecol Lett* 20:1325–1336
- van Herwaarden OA, Grasman J (1995) Stochastic epidemics: major outbreaks and the duration of the endemic period. *J Math Biol* 6:581–601
- Weiss GH, Dishon M (1971) On the asymptotic behavior of the stochastic and deterministic models of an epidemic. *Math Biosci* 11:261–265
- Wilkinson RR, Sharkey KJ (2013) An exact relationship between invasion probability and endemic prevalence for Markovian SIS dynamics on networks. *PLoS ONE* 8:e69028
- Yates A, Antia R, Regoes RR (2006) How do pathogen evolution and host heterogeneity interact in disease emergence? *Proc R Soc B* 273:3075–3083
- Zachary S (2007) A note on insensitivity in stochastic networks. *J Appl Probab* 44:238–248