IOP Publishing Reports on Progress in Physics

Rep. Prog. Phys. 77 (2014) 026602 (51pp) doi:10.1088/0034-4885/77/2/026602

Review Article

Dynamics of infectious diseases

Kat Rock[1,2], Sam Brand[1,2], Jo Moir[1,2] and Matt J Keeling[1,2,3]

[1] WIDER Centre, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
[2] Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK
[3] School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry, CV4 7AL, UK

E-mail: M.J.Keeling@warwick.ac.uk

Received 6 March 2013, revised 21 November 2013
Accepted for publication 29 November 2013
Published 20 January 2014

Invited by Erwin Frey

Abstract

Modern infectious disease epidemiology has a strong history of using mathematics both for prediction and to gain a deeper understanding. However the study of infectious diseases is a highly interdisciplinary subject requiring insights from multiple disciplines, in particular a biological knowledge of the pathogen, a statistical description of the available data and a mathematical framework for prediction. Here we begin with the basic building blocks of infectious disease epidemiology—the SIS and SIR type models—before considering the progress that has been made over the recent decades and the challenges that lie ahead. Throughout we focus on the understanding that can be developed from relatively simple models, although accurate prediction will inevitably require far greater complexity beyond the scope of this review. In particular, we focus on three critical aspects of infectious disease models that we feel fundamentally shape their dynamics: heterogeneously structured populations, stochasticity and spatial structure. Throughout we relate the mathematical models and their results to a variety of real-world problems.

Keywords: epidemic, infection, vaccination, risk structure, spatial spread, stochasticity

1. Introduction

Infectious diseases pose a considerable challenge to modern life. In the 1950s and 1960s, with the development of safe vaccines and antibiotics it was largely felt that this age-old foe was beaten. However, recent experience with the evolution of drug-resistant pathogens, the emergence of new infections and the challenges of cost-effective control highlight that we are still a long way from eradicating most diseases. This review is focussed upon how tools from mathematics and theoretical physics can provide meaningful insights into infectious disease dynamics and hence provide both a general understanding of the patterns and processes as well as enable a thorough understanding of the impact of proposed control measures.

Infectious diseases are ubiquitous. It appears that being parasitic, and obtaining all your needs from another living thing, is an incredibly successful strategy. Even very simple organisms fall prey to infection, bacteria themselves can be infected by viruses (termed phage), while higher organisms often have a range of specific pathogens that depend on these organisms for their survival. Examples of high impact and high profile infections exist from every group of living organisms: agricultural crops are commonly infected with fungal or viral pathogens [223]; wild plant populations can be devastated by infection (such as Sudden Oak Death [94]); insects also suffer with honey bees being a notable example potentially related to Colony collapse disorder [86]; amphibians are likewise infected, with Chytrid fungus destroying populations of frogs worldwide [61]. However, infection in mammals has received by far the most attention both due to the impact on livestock production systems (e.g. Foot-and-Mouth disease [90, 147] or Classical Swine Fever [137]) and the effects on human health. Globally, it is estimated that one quarter of all human deaths are due to infectious diseases [253].

This review focuses on *infectious* diseases, those illnesses caused by an infecting organism that can be spread between individuals. This broad definition includes well-known viral infectious such as measles, influenza and HIV or bacterial infections such as meningitis and cholera, but also includes infection with fungal pathogens, protozoa (such as the organism that causes malaria) and larger parasites (such a helminths and other worms). Throughout this review, we generally focus on viral or bacterial pathogens, as these are most commonly studied in the literature, and we will generally use the term pathogen to refer to any infecting organism.

The study of infectious diseases naturally arises from two sources: the scientific desire to understand the patterns of mortality (death) and morbidity (illness) that surround us; and the desire to reduce illness and suffering. These two elements are epitomized by the work of John Snow in the 1840s and 1850s. In what is widely regarded as the birth of scientific epidemiology, Snow showed how clusters of cholera cases in London were related to the local water supply, and by eliminating this source (famously removing the handle of the Broad Street water pump) helped bring the epidemic under control. However, Snow's approach was predominantly statistical, looking for patterns in existing data. It was not until the pioneering work of Kermack and McKendrick [155] that epidemiology was placed on a strong mathematical foundation. Even so, it was only with the advent of computers, which could rapidly integrate the underlying ordinary differential equation (ODE) models, that the modelling of infectious disease dynamics became a practical tool [16]. The predictive power of these modelling techniques was first illustrated by work on Rubella using age-structured models (see section 3) when it was shown that although low levels of vaccination would reduce the overall prevalence of infection, it would delay infection to later in life so that more women would be infected during pregnancy. Hence a limited vaccination campaign would actually increase the amount of congenital rubella syndrome in unborn babies [8]. In more recent times, predictive mathematical models were used to inform control of the Foot-and-Mouth outbreak in 2001 (one of the first times models were used during an epidemic [90, 147]), helped inform planning for pandemic outbreaks [88, 175], and used for cost-effectiveness calculations before implementation of novel vaccine programmes [18, 244]. This review follows this chronological progress and basic philosophy, using simple models to develop an intuitive understanding of infectious disease dynamics before describing the complexities that would need to be included to make such models practically useful.

The epidemiology of infectious diseases is a complex and multi-factorial subject [16, 74, 151]. As such, no single review can hope to cover all aspects. Here we focus largely on the population dynamics of infectious diseases—how the number of individuals infected changes dynamically over time. This article therefore relies heavily on tools from mathematics and theoretical physics (the theory of differential equations and dynamical systems, statistical mechanics, and stochastic processes). However, there are multiple disciplines that feed into epidemiology that we do not cover. Statistics is one with which many readers will be familiar; statistics plays a vital role in both interpreting the observed infection data (accounting for the many biases in reporting) and in parameter inference when we attempt to fit models to data. Obviously, a range of biological sciences from microbiology to immunology to ecology are needed to inform our understanding of the pathogen, the host and their interaction; while we do not attempt a comprehensive review of such knowledge, the next section does provide sufficient biological background to motivate and justify the choice of model formulation. Finally, in recent years it has become apparent that other disciplines have a role to play: economics is vital to underpin cost-effectiveness studies that are key to assessing control programmes; sociology and psychology help to explain and predict human response to outbreaks or new treatments; while medical insights are needed to understand the link between the individual as a host for the pathogen and the individual as a patient that requires treatment. Therefore, when developing models for the spread of an infectious disease we not only need a range of mathematical skills but must account for the insights provided by many other disciplines. Throughout, we had attempted to draw on citations from the mathematical and biological literature whenever possible, so as to introduce the reader to this (possible novel) fields of scientific publication.

The remainder of this review is partitioned into four main sections. In the next section we review the basic mathematical models that underpin all work in this area, using the ODE models that date back to Kermack and McKendrick [155], and focus on how our understanding of pathogen biology is translated into a system of equations. To the quantitatively trained reader the formulation and solution of these early equations may appear trivial, but the insights provided are key to many public-health decisions and act as a building block for larger and more complex approaches. In sections 3–5 we build upon these foundations by considering three different aspects (population heterogeneity, stochastic dynamics and spatial structure) that fundamentally change the ways in which we model the infection dynamics, before finally discussing how such methods and insights could be combined to offer robust predictions and practical policy guidance.

2. Basic models

With only a very few exceptions, most mathematical descriptions of the population dynamics of infectious diseases rely on being able to partition the population into discrete non-overlapping compartments [15, 16, 74, 151, 180]. This act of compartmentalization allows us to study the dynamics of infection by capturing the transition of individuals between compartments. A simple example will illustrate the main points (figure 1): before the 2009 H1N1 (swine-flu) pandemic very few people in the population had seen a similar infection before and could therefore be categorized as *susceptible*, susceptible people that come in contact with the pathogen can become *infected* and pass the pathogen (influenza virus) on to other susceptible individuals, after a time infected people generally 'fight-off' the infection, are no longer able to transmit and are classified as *recovered*. This simply natural history of infection is familiar to us all, and the compartmentalization into the three categories Susceptible, Infected and Recovered allows us to develop simple mathematical models. Throughout the early sections of this review, we will illustrate such compartmental models using the caricature method shown in figure 1. However, before we write these models in terms of differential equations, there are two fundamental points that need emphasis.

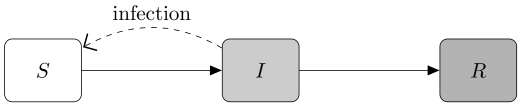

Figure 1. Compartmental caricature of the basic SIR model (solid lines show transition or movement between classes; dashed lines show the action of transmission).

Firstly, in the description above we have taken a pathogen-centric view, categorizing the host (person) by their status with respect to the pathogen—this is an epidemiological view-point and is necessary if we wish to study the pathogen dynamics. More commonly we are accustomed to a medical perspective where we think about the status of the patient and what symptoms they display. Sometimes these two view-points are aligned and severe illness coincides with the peak of infectivity, however many other scenarios are plausible; for example, with chickenpox (caused by the varicella zoster virus) the peak of infection generally occurs a few days before the onset of a rash and other symptoms [151].

Secondly, it must be noted that the susceptible, infected and recovered categorization ignores many important aspects of infection, and is at best an idealized paradigm for some infections. Several elements complicate this picture (such a carrier states, heterogeneous responses or vaccination) and we will deal with many of them later, but for now we will briefly mention two additional complexities. Not all infections follow the susceptible-infected-recovered pattern, there are many infections that can be caught multiple times; sexually transmitted infections (STIs) tend to follow this profile and individuals that are treated for infection generally return to the susceptible class, giving rise to SIS (susceptible, infected, susceptible) models. It is naive to think that the amount of pathogen an individual can transmit is constant throughout the time that they are infected, clearly following infection (with a low level of the infecting pathogen) there needs some time for the pathogen to build to sufficient levels for onward transmission to be likely. This effect can be captured by adding an additional *latent* (or *exposed*) class that individuals need to pass through before they become infectious (ie actively transmitting).

Finally, we stress that although we begin this exploration of infectious disease dynamics by considering simple systems of ODEs, their formulation may be motivated by considering the stochastic behaviour of individuals. In particular, following the work of [167] we can consider the ODEs as the limit of a stochastic process where a large (infinite) number of individuals are moving between the various model compartments. In this limit of infinitely many individuals, the stochastic movement between compartments can be conceptualized as a continuous flow. Section 4 provides more details of this stochastic approach and the connections to the deterministic models of sections 2 and 3.

2.1. A generic model

We can now put these elements together to create a generalized model for the dynamics of infection in a population [15, 16, 180] (see figure 2):

$$\begin{aligned}
\text{Susceptibles} \quad \frac{\mathrm{d}S}{\mathrm{d}t} &= B + \nu R - \lambda S - dS,\\
\text{Exposed} \quad \frac{\mathrm{d}E}{\mathrm{d}t} &= \lambda S - \alpha E - dE,\\
\text{Infectious} \quad \frac{\mathrm{d}I}{\mathrm{d}t} &= \alpha E - \gamma I - dI,\\
\text{Recovered} \quad \frac{\mathrm{d}R}{\mathrm{d}t} &= \gamma I - \nu R - dR.
\end{aligned} \tag{1}$$

Equation (1) contains one rate (B) which determines the rate at which new-born individuals enter the population, and five per capita rates (ν, d, α, γ and λ) that determine transitions between (or out of) categories. In this form the dynamics could be solved relatively trivially, however there is an additional non-linear feed-back that needs to be incorporated. The *force of infection* (λ) which quantifies the risk of infection is clearly related to the number of infectious individuals in the population. In particular it is standard to assume:

$$\text{Force of Infection } \lambda = \beta I/N, \tag{2}$$

where N is the total population size ($N = S+E+I+R$) and the parameter β captures both the rate at which epidemiologically relevant contacts are made and the probability that contact between an infectious and susceptible individual leads to the transmission of infection; the I/N term then accounts for the chance that the contact is infectious, assuming contacts are made randomly within the population. We will see later how the force of infection becomes modified in spatial models where we account for non-random mixing. In equation (2) we have assumed *frequency-dependent* transmission; this is derived by assuming that each individual has a fixed number of daily contacts that is independent of population size, hence the force of infection is related to the frequency of infectious individuals in the population. Frequency-dependent transmission tends to be the norm for human populations. The alternative is to assume *density dependent* transmission where more dense populations give rise to more contacts and hence the force of infection is related to the density of infectious individuals: $\lambda = \beta I$. However, in many theoretical settings it is often simpler to re-scale the variables such that the population size, N, is one; although clearly this is not feasible when dealing with changing population sizes. Under such a re-scaling the two transmission assumptions are equivalent.

Together equations (1) and (2) combine to generate a range of infection models that can either be studied analytically or be rapidly integrated numerically. We first highlight the three standard models that are contained within this more general formulation. The simple SIR model is recaptured by setting $B = d = \nu = 0$ and letting $\alpha \to \infty$; this is the ideal simple model for conceptualizing single epidemics which occur sufficiently quickly that natural births and deaths can be ignored. The SIS model is generated by setting $B = d = 0$ and letting $\alpha \to \infty$ and $\nu \to \infty$; this is the standard model for sexually transmitted diseases. Finally, the general equation itself would be described as an SEIRS model, with

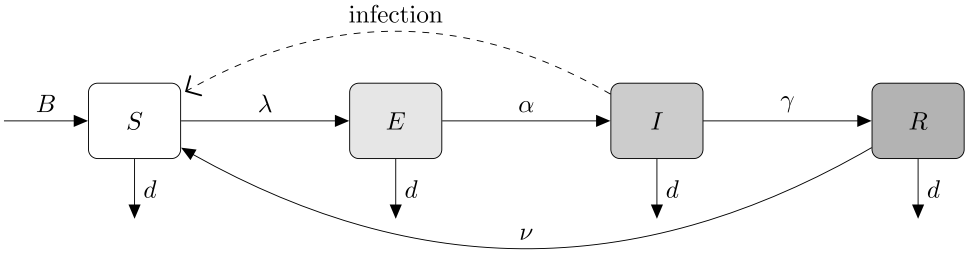

Figure 2. Compartmental caricature of the generic SEIRS model (solid lines show transition or movement between classes; dashed lines show the action of transmission; arrows entering or leaving the system correspond to births and deaths respectively).

demography and waning immunity; in particular the per capita rates are: d the natural death rate; α the rate of progressing from exposed to infectious, such that $1/\alpha$ (or more technically when accounting for the death rate $1/(\alpha+d)$) is the average exposed or latent period; γ is the rate of recovery, such that $1/\gamma$ (or more technically $1/(\gamma+d)$) is the average infectious period; and ν is the rate at which immunity is lost and the individual reverts to being susceptible.

2.2. Early behaviour

One of the most fundamental concepts in epidemiology is the *basic reproductive ratio* or *number*, represented by R_0. This quantity is defined as 'The average number of secondary cases produced by an average infected individual in a totally susceptible population', and in practice provides a quantitative guide to both invasion and control of an infectious disease [15, 16, 123, 180]. In particular, when R_0 is greater than one an infection invading a naïve population can spread whereas when R_0 is less than one chains of infection will always die out. These logical arguments can be made mathematically precise by calculating the Jacobian of the disease-free equilibrium ($S = N$, $E = I = R = 0$). For the general model defined by equations (1) and (2), the Basic Reproductive Ratio can be calculated in three parts considering the dynamics associated with a newly infected individual: the probability a newly infected individuals goes on to become infectious; the rate at which new cases are then generated; and the average duration over which the individual remains infectious.

$$\begin{aligned} R_0 &= \frac{\alpha}{\alpha+d} \times \beta\frac{S}{N} \times \frac{1}{\gamma+d} \\ &= \frac{\beta\alpha}{(\alpha+d)(\gamma+d)} \approx \frac{\beta}{\gamma}. \end{aligned} \tag{3}$$

Equation (3) has been gained by noting that it is calculated when everyone is susceptible ($S = N$) while the approximation holds if the natural death rate is small compared to the rates of transition through the exposed and infectious classes. Two points are immediately noteworthy in the calculation of R_0. The first is that many of the components of the full SEIRS model do not play a role in the value of R_0, in particular it is independent of both the birth rate and whether (or not) there is waning immunity; thus the SIR and SIS models have the same basic reproductive ratio for the same parameters despite the differences in underlying structure. Secondly, the value of R_0 depends on the transmission rate β, which in turn depends on the rate at which (epidemiologically relevant) contacts are made. We therefore find that R_0 depends on both the pathogen and the population. For example although measles and influenza are both spread through close-contact or local airborne interactions, R_0 for measles is around 17 [16] whereas R_0 for influenza is between 1.5 and 3.0 [239]. Considering HIV infection illustrates the role of the contacts, with values of R_0 during the 1980s estimated as in the range 2–5 in the UK, but 10–12 in sub-Saharan Africa [16].

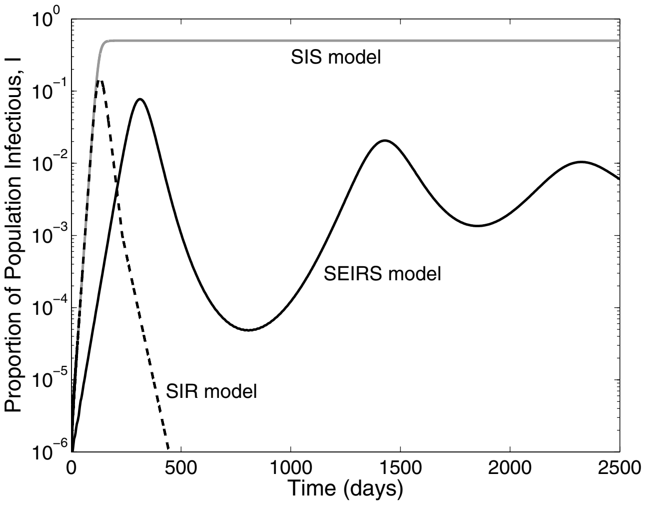

Figure 3. Comparison of the basic infection models derived from equations (78) and (2). The three curves show the proportion of infectious individuals in the population (I/N), for the standard SIS model, the standard SIR model and the general SEIRS model. (For all three models: $\beta = 0.2$, $\gamma = 0.1$. For the SIS model $B = d = 0$ and $\alpha = \nu \to \infty$; for the SIR model $B = d = \nu = 0$ and $\alpha \to \infty$; for the SEIRS model $B = d = 10^{-4}$, $\alpha = 0.1$ and $\nu = 10^{-3}$. All rates are given in days^{-1} ($I(0) = I_0 = 10^{-6}$, $N = 1$, $S(0) = N - I_0$.

Figure 3 shows the dynamics of equations (1) and (2), for three different cases: the SIS model, the SIR model and the general SEIRS model. In all three cases, the basic reproductive ratio is approximately 2 (for the SEIRS the true value is 1.996 due to the action of natural mortality). For the SIS and SIR models it is clear that they share the same initial growth rates, while the presence of an exposed (latent) class in the SEIRS model slows the dynamics. Thus, although R_0 informs about the ability of a pathogen to invade it does not completely determine the speed of invasion. Therefore in many applications we also need to consider the early (asymptotic) growth rate, r, as this can be more closely linked to available data. From consideration of the disease-free state of the general model, the early growth rate is given by

$$r = \frac{1}{2}\left[\sqrt{4S_0\beta\alpha + (\alpha-\gamma)^2} - (\alpha+\gamma+2d)\right],$$
$$r_{\mathrm{SIR}} = r_{\mathrm{SIS}} = S_0\beta - \gamma - d, \tag{4}$$

where $S_0(=S(0)/N)$ is the proportion of the population susceptible at the start of the outbreak. This raises an interesting and important applied issue, when we observe an exponential growth of cases in the early phase of an epidemic; there are currently no robust techniques for distinguishing between a small scale epidemic where all cases report symptoms of infection and a large-scale epidemic where only a small proportion report [18]. In general it is not until the epidemic peaks and we observe strong non-linear effects that this distinction can be made.

2.3. Long-term behaviour

Examining figure 3 also leads us to think about the long-term dynamics of this system. We first begin by considering a general solution before focussing on the specific cases of the SIS and SIR models. When considering long-term endemic behaviour, it is customary to set $B = d$ which effectively rescales the population size $N = 1$. Under this assumption the equilibrium dynamics are given by

$$\begin{aligned}
S^* &= \frac{(\alpha+d)(\gamma+d)}{\beta\alpha} \approx \frac{\gamma}{\beta},\\
E^* &= \frac{(\beta\alpha - (\alpha+d)(\gamma+d))(\nu+d)(\gamma+d)}{\beta\alpha(\alpha\nu+\nu\gamma+\alpha\gamma+\alpha d+\nu d+\gamma d+d^2)}\\
&\approx \frac{(\beta-\gamma)(\nu+d)\gamma}{\beta(\alpha\nu+\nu\gamma+\alpha\gamma)},\\
I^* &= \frac{(\beta\alpha - (\alpha+d)(\gamma+d))(\nu+d)}{\beta(\alpha\nu+\nu\gamma+\alpha\gamma+\alpha d+\nu d+\gamma d+d^2)}\\
&\approx \frac{\alpha(\beta-\gamma)(\nu+d)}{\beta(\alpha\nu+\nu\gamma+\alpha\gamma)},
\end{aligned} \tag{5}$$

where again the approximations hold whenever the natural death rate is small compared to the rates of transition through the exposed and infectious classes. In general this equilibrium is approached through a series of damped oscillations, which may interact with any seasonal forcing in the basic parameters [153]. It should be noted that the susceptible equilibrium is only viable (i.e. $S^* \leqslant N = 1$) when the basic reproductive ratio is greater than one; when the basic reproductive ratio is less than one only the disease-free equilibrium is a viable fixed point of the system. It is worth noting that at the equilibrium we have that $(S^* = 1/R_0)$, this is a common feature of many epidemiological models and it related to the fact that at equilibrium the average number of secondary cases produced per infected individual $(R_e = R_0 \times S/N)$ has to be equal to one, otherwise the number of case would change.

For the SIS model, the dynamics simplify enormously. In particular we now have the equilibrium solution:

$$S^*_{\mathrm{SIS}} = \frac{\gamma}{\beta} \qquad I^*_{\mathrm{SIS}} = 1 - \frac{\gamma}{\beta}.$$

For this model, the recovery of infected individuals back to the susceptible state is sufficient to maintain the presence of infection, and the fact that there are only two compartments means that $S + I = 1$. In fact, the SIS model is equivalent to the logistic growth model in ecology, and therefore has an algebraic solution:

$$I_{\mathrm{SIS}}(t) = \frac{(\beta-\gamma)I_0\mathrm{e}^{(\beta-\gamma)t}}{(\beta-\gamma) + I_0\beta\left[\mathrm{e}^{(\beta-\gamma)t} - 1\right]},$$

where I_0 is the proportion of infection in the population at $t = 0$.

In contrast, the simple SIR model without births and deaths does not have an endemic equilibrium; instead there is a single epidemic which eventually dies out. This extinction of infection is caused by a depletion of susceptibles to a level that cannot sustain the infection, without new births or waning immunity there is no way for the susceptibles to be replenished and so the infection is doomed to extinction. Mathematically, this is clear from the fact that the disease-free states are the only equilibria. However, for the simple SIR the key question is how many individuals become infected during the entire course of the epidemic, this addresses an issue of fundamental concern for any outbreak—how big is the epidemic likely to be. The solution to this problem dates back to the ground-breaking papers of Kermack and McKendrick [155], and arises from considering $(\mathrm{d}S/\mathrm{d}R) = (\mathrm{d}S/\mathrm{d}t)/(\mathrm{d}R/\mathrm{d}t)$. The *Final Epidemic Size*, R_∞ is given by the relation:

$$R_\infty = S_0\left[1 - \exp(-R_0R_\infty)\right], \tag{6}$$

where R_0 is the basic reproductive ratio, S_0 is the initial proportion of the population that is susceptible to infection and R_∞ is the proportion of the population that have passed through the infectious class (specifically $R_\infty = S(0) - S(\infty)$ assuming a small initial seed of infection). We find that an identical formula holds for SEIR-type models, all that is important is the lack of births, deaths and waning immunity. What is clear from this formula is that there will always be some individuals that escape an epidemic, although the proportion of the population that do escape becomes small as R_0 becomes large. We note that in practice epidemics rarely infect very high proportions of the population and this is largely due to heterogeneities in risk (see below) with some individuals with high risk contributing most to the value of R_0 whereas others with low risk more likely to escape infection. Hence, the final epidemic size should be viewed as an idealized calculation, although for the recent H1N1 (swine-flu) outbreak the value of $R_\infty \approx 0.37$ (derived from $R_0 \approx 1.25$ [18]) is in good qualitative agreement with finding from detailed serological studies [56].

2.4. Greater realism, more compartments

In the generic model above (equation (1)) there are four compartments that are used to characterize an individual's status with respect to the pathogen. Greater biological realism can be introduced by including more compartments and therefore greater heterogeneity in the dynamics. Later we discuss multiple compartments based on the host, allowing us to consider age-structured models or multiple host species. Here we focus on additional compartments due to the pathogen. Two approaches are worth highlighting in detail.

One way in which additional compartments may be required is when there is a more complex natural history of infection that needs to be captured. For example, we may wish to have a short-duration Maternal Immunity class that individuals are born into, individuals in this class would be protected against infection due to immunity gained from the mother. Alternative, there are infections (such as typhoid) where carrier individuals exist who may be weakly infectious for extremely long periods of time or who may intermittently revert to a highly infectious state; in such cases these additional compartments need to be included in any model to produce realistic results. However, the inclusion of these extra compartments is not conceptually challenging and uses the existing framework that specifies the rates of entering and leaving each compartment, although parametrization of such models may be much more complex.

An additional way in which more compartments can be included is to subdivide the infectious (and exposed) classes into more groups. Ignoring waning immunity this would lead to the following set of equations (figure 4):

$$\text{Susceptibles}\quad \frac{\mathrm{d}S}{\mathrm{d}t} = B - \lambda S - dS,$$

$$\text{Exposed}\quad \frac{\mathrm{d}E_1}{\mathrm{d}t} = \lambda S - \alpha_1 E_1 - dE_1,$$

$$\text{Exposed}\quad \frac{\mathrm{d}E_n}{\mathrm{d}t} = \alpha_{n-1}E_{n-1} - \alpha_n E_n - dE_n \qquad n = 2, \ldots, M-1,$$

$$\text{Infectious}\quad \frac{\mathrm{d}I_1}{\mathrm{d}t} = \alpha_m E_m - \gamma_1 I_1 - dI_1,$$

$$\text{Infectious}\quad \frac{\mathrm{d}I_n}{\mathrm{d}t} = \gamma_{n-1}I_{n-1} - \gamma_n I_n - dI_n \qquad n = 2, \ldots, M-1,$$

$$\text{Recovered}\quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma_M I_M - dR,$$

$$\text{Force of Infection}\quad \lambda = \sum_n \beta_n I_n / N. \tag{7}$$

This subdivision has two important implications. Firstly it allows us to control with far more finesse the amount of infection an individual sheds over time. This is an enhancement to the concept of splitting the infected component into two (the exposed and infectious components) and can be necessary to fully capture the known epidemiology of some infections. HIV is a notable example of when this can prove necessary as early and late stages of infection are estimated to be far more infectious than the intervening asymptomatic stages [124]. The second element becomes more obvious when we consider the transition of a single individual, which obviously brings in ideas of stochasticity (see section 4). When the exposed and infectious classes exist as a single compartment, the time spent in each compartment is exponentially distributed and hence some individuals recover rapidly while others are infectious for very long times. This is turn has implications for early growth rate of the infection in the population and the variability in the number of secondary cases produced by infected individuals, which in turn influences persistence [63, 152, 174]. In contrast, by having multiple exposed and infectious classes that an individual must transition through before recovery, the total time spent in all these classes is given by a gamma distribution. These multiple classes both act to speed the growth rate of infection in the population (see figure 5) and lead to less variability in individual behaviour. In practice there is rarely sufficient data to necessitate more than two or three exposed and infectious classes ($m = M = 3$). The individual-level implications of this additional structure is considered in more detail in section 4.4.2.

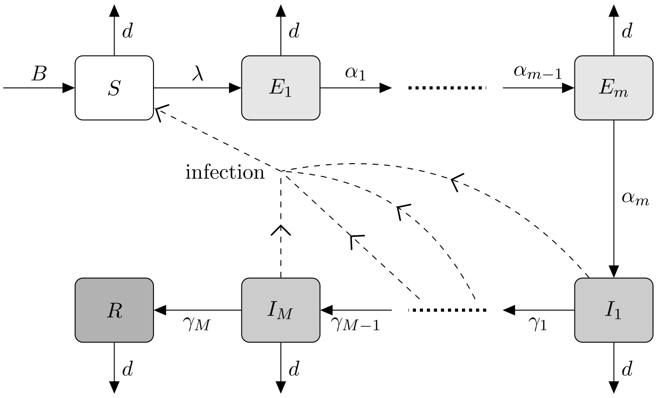

Figure 4. Compartmental caricature of the multi-compartmental $SE_1 \ldots E_m I_1 \ldots I_M R$ model (solid lines show transition or movement between classes; dashed lines show the action of transmission; arrows entering or leaving the system correspond to births and deaths respectively).

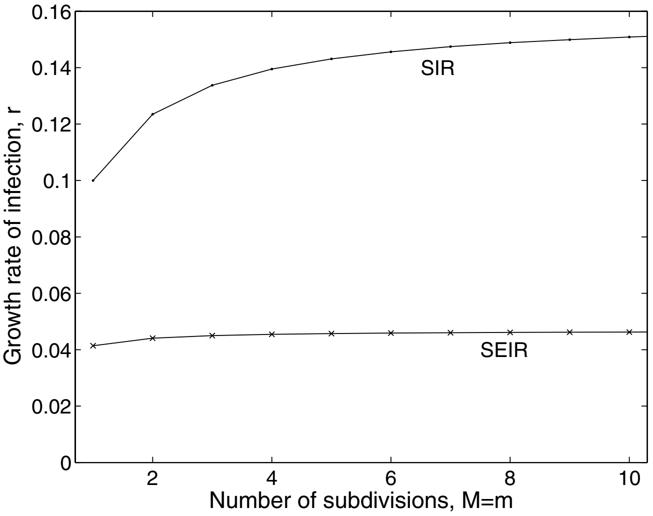

Figure 5. Comparison of the growth rate of infection, r, such that $I(t) \sim \exp(rt)$ for the SIR and SEIR model with additional compartments (M and m for infectious and exposed respectively), but ignoring births, deaths and waning immunity. The x-axis gives the number of sub-components that the infectious (and exposed) class has been divided into. (For all both models: $\beta_n = 0.2$, $\gamma_n = 0.1M$ and $\alpha_n = 0.1m$, $B = d = \nu = 0$; this scaling of the transition rates through the classes keeps the average time in the exposed and infectious classes independent of the number of classes.)

Here we have stressed the need to include multiple compartments (that reflect the underlying biology) in any attempt to generate reliable predictions. However, writing down such models in inherently unwieldy, and while necessary for accurate prediction it adds comparatively little to our deeper understanding of the dynamics. Therefore, in the sections that follow we often focus on how different elements (risk-structure, stochasticity or spatial structure) effect the simple SIR and SIS models, and we leave it to the reader to extrapolate in the obvious manner for models with more compartments.

2.5. Vaccination

Finally in this introduction to the basic epidemiological models, we introduce a vaccinated class to account for the effects of immunization. Here we need to be a little careful with our definitions: vaccinated generally means to have received the vaccine, whereas only a proportion of these will be immunized and protected against infection. With the best vaccines the proportion that are protected can be very high, often in excess of 99%, however with other vaccines, such as those against seasonal influenza where the vaccine is developed based on predictions of future strains, the effectiveness can be anywhere between 30% and 90% [38]. In the remainder of this section we therefore formulate models based on the level (rate or proportion of the population) of successful immunization, with the implicit assumption that this would necessitate a higher level of vaccination.

Vaccination is generally more common in diseases that naturally obey the S(E)IR-type paradigm (rather than SIS-type infections). This is because vaccines generally operate by triggering the host's own natural immunity to the infecting pathogen, without causing disease. A suitable equation for the dynamics of an SIR-type infection with vaccination would be (figure 6):

$$\begin{aligned}
&\text{Susceptibles} \quad \frac{\mathrm{d}S}{\mathrm{d}t} = B(1-p) - \lambda S - \mathrm{d}S - vS/N,\\
&\text{Infectious} \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \lambda S - \gamma I - \mathrm{d}I,\\
&\text{Recovered} \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - \mathrm{d}R, \qquad\qquad (8)\\
&\text{Vaccinated} \quad \frac{\mathrm{d}V}{\mathrm{d}t} = Bp + vS/N - \mathrm{d}V,\\
&\text{Force of Infection} \quad \lambda = \beta I/N.
\end{aligned}$$

Here vaccination takes two forms (although generally either one or the other dominates); firstly a proportion p of new-born individuals are vaccinated and successfully immunized and so enter the compartment V where they remain for life. Secondly we assume a random vaccination (and immunization) rate v, and if the randomly selected individual is susceptible then they are transferred to the V compartment. The first of these is most common and equates to the typical childhood immunization campaigns that have been so successful in reducing childhood mortality; the second of these is associated with vaccinating to control a novel outbreak (such as influenza pandemic [18] or foot-and-mouth [236]), or vaccinating wild animal populations [7, 57, 252]. It should be noted that p refers to the proportion of new-borns immunized and therefore lies between 0 and 1, whereas v is a rate and therefore is only constrained by resources and logistics.

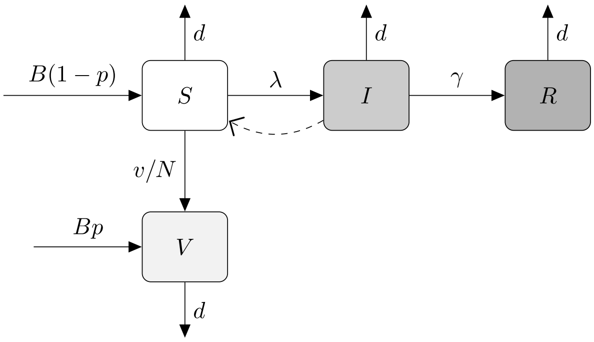

Figure 6. Compartmental caricature of the SIRV model (solid lines show transition or movement between classes; dashed lines show the action of transmission; arrows entering or leaving the system correspond to births and deaths respectively).

We first focus on the use of vaccination to eradicate infection. This can be approached in two ways: the first is to consider if (long-term) vaccination can prevent an infection invading and spreading from the disease-free equilibrium; the second is to consider the equilibrium state and how the prevalence of infection changes with vaccination. Looking at the disease-free equilibrium with vaccination (and assuming $B = d$ and hence $N = 1$) we find:

$$S^* = \frac{1-p}{1+v/d} \qquad V^* = 1 - \frac{1-p}{1+v/d}.$$

Examining the early growth rate of infection from this point gives:

$$\frac{\mathrm{d}I}{\mathrm{d}t} = (\beta S^* - \gamma - d)I$$

and hence the disease fails to invade (has negative growth rate) if

$$S^* < \frac{\gamma + d}{\beta} = \frac{1}{R_0}.$$

Therefore, we require immunity due to vaccination to be sufficient to drop the proportion of susceptibles in the population below their non-vaccinated equilibrium value in the presence of infection. What is important to realize is that it is not necessary to immunize everyone to protect the entire population, the action of immunising one person has additional benefits to other individuals who could have been infected by this person. This level of addition protection in the population is known as *herd-immunity* [16, 74, 151]. Therefore vaccination is an altruistic act which benefits both the individual and the population. Examining the two methods of vaccination separately gives critical thresholds:

$$p_c = 1 - \frac{1}{R_0} \qquad v_c = (R_0 - 1)d. \qquad (9)$$

Immunising above these thresholds prevents a pathogen from successfully invading.

Alternatively, we can consider the equilibrium state with both infection and vaccination acting to reduce the susceptible population, again assuming $B = d$. Two equilibria are possible depending upon whether the level of vaccination is sufficient to eradicate infection or not:

$$S^* = \frac{1}{R_0} \qquad \text{or} \qquad S^* = \frac{1-p}{1+v/d} \qquad (10)$$
$$I^* = \frac{d(R_0-1) - pdR_0 - v}{\beta} \qquad\qquad I^* = 0.$$

Hence we find that immunization has a linear impact on the prevalence of infection; the proportion of infected individuals drops linearly with both types of vaccination until it reaches zero at the critical thresholds. This is an important public-health observation, as it generally means that it is better to do some vaccination even if there is not sufficient resources (or demand) to reach the critical threshold. The exceptions to this rule are when there exists complex interactions between disease severity and age structure [8] as vaccination generally increases the average age of infection (see section 3.2).

The above analysis has focused on long-term dynamics, such that the impact of any vaccination scheme has had time to equilibrate. However, very often vaccination programmes are introduced at a given point in time, acting as a substantial perturbation to the system. One way of conceptualizing this process is that when the vaccination programme begins the population is likely to be at the non-vaccinated equilibrium point given by equation (5), which is likely to be far from the new appropriate equilibrium point of the vaccinated population given by equation (10). Convergence to this new vaccination equilibrium is via damped oscillations. Again this has important health implications as the number of infected individuals will first dip below the expected equilibrium level (and hence the vaccine programme is seen as a success)—this is known as the honeymoon effect [184]—before rising to much higher levels (when concerns about the vaccine programme are likely to be raised). Figure 7 illustrates this behaviour for two levels of vaccination given at birth below the critical threshold; not only do we observe large oscillations and substantial honeymoon effects, but also we note how long it can take for the dynamics to approach equilibrium. Given that this form of vaccination only occurs at birth the time-scales of the programme are comparable with the time-scales of the host life-cycle; this is why new vaccinate campaigns are often accompanied by catch-up programmes in older age-groups.

Again, there are multiple improvements that can be made to the standard model of vaccination (equation (8)). Clearly, adding an exposed class and subdividing the exposed and infectious classes into sub-compartments will increase the realism of the infection dynamics (although this will have limited impact on the critical eradication thresholds). When waning immunity is a feature of the infection dynamics ($\nu > 0$) then it becomes natural to also include this facet in the vaccination dynamics, with the possibility of a different rate of waning immunity to capture the fact that protection following infection may be more long-lasting than that due to vaccination. In addition, for many vaccines a single dose is insufficient to provide life-long immunity and therefore the impact of booster vaccines later in life must be incorporated into the standard equations [4]. However, two improvements are particularly interesting from a practical perspective: those concerning the use of vaccines in controlling novel outbreaks, and those that consider in more detail the action of the vaccine itself.

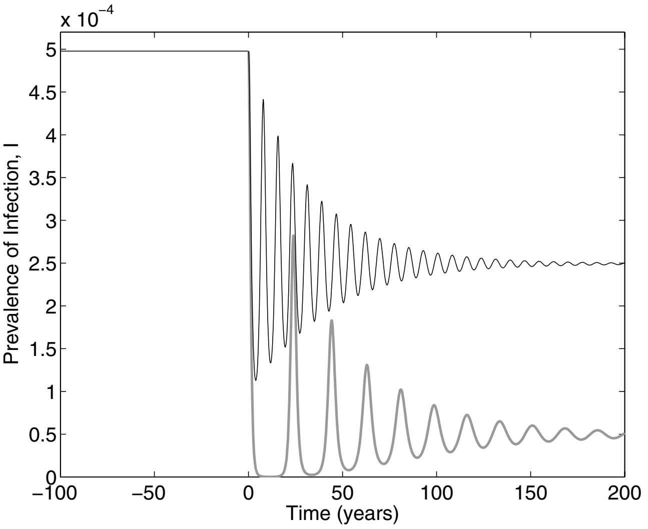

Figure 7. Illustration of the honeymoon effect for vaccination in the standard SIR model and two levels of vaccination. Vaccination begins at time 0, when the system is at its un-vaccinated equilibrium. Two levels of vaccination at birth are considered $p = 0.25$ (black line) and $p = 0.45$ (thick grey line), noting that the critical threshold occurs at $p_c \approx 0.5$. (The SIR model is given by: $\beta = 0.2$, $\gamma = 0.1$, $\alpha \to \infty$, $\nu = 0$ and $B = d = 10^{-4}$; with all rates being in days^{-1}).

Vaccination is often seen as a panacea against infection, and for many endemic diseases where there is a childhood vaccination campaign this is generally the case. However, vaccination can also be used reactively in the face of an on-going outbreak. Classic examples include the use of vaccination to control bioterrorist use of smallpox [92, 130] or pandemic influenza [88] in humans, or to limit the spread of livestock infections such as foot-and-mouth disease [236]. In these reactive situations two modifications are important. Firstly, there is generally a delay (of between a few days and several weeks) between receiving the vaccine and being protected, which in turn delays the impact of vaccination. Mathematically this can be captured by subdividing the vaccinated compartment into P classes and allowing susceptibility to infection to decrease as an individual moves through the classes; therefore immediately following vaccination protection in limited and it is possible to acquire the infection. Secondly, the assumption of random vaccination can be refined, only vaccinating individuals that have not yet received the vaccine. The new equations become (figure 8):

$$\text{Susceptibles} \quad \frac{\mathrm{d}S}{\mathrm{d}t} = B - \lambda S - \mathrm{d}S - vS/(S+I+R),$$

$$\text{Infectious} \quad \frac{\mathrm{d}I}{\mathrm{d}t} = \lambda S + \sum_n \sigma_n \lambda V_n - \gamma I - dI,$$

$$\text{Recovered} \quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - dR,$$

$$\text{Vaccinated} \quad \frac{\mathrm{d}V_1}{\mathrm{d}t} = vS/(S+I+R) - \mu_1 V_1 - \sigma_1 \lambda V_1 - dV_1,$$

$$\text{Vaccinated} \quad \frac{\mathrm{d}V_n}{\mathrm{d}t} = \mu_{n-1}V_{n-1} - \mu_n V_n - \sigma_n \lambda V_n - dV_n,$$
$$n = 2, \ldots, P-1,$$

$$\text{Vaccinated} \quad \frac{\mathrm{d}V_P}{\mathrm{d}t} = \mu_{P-1}V_{P-1} - \sigma_P \lambda V_P - dV_P. \qquad (11)$$

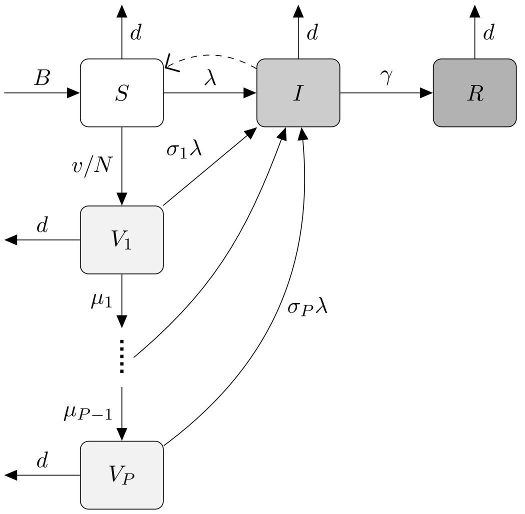

Figure 8. Compartmental caricature of the SIRV model (solid lines show transition or movement between classes; dashed lines show the action of transmission; arrows entering or leaving the system correspond to births and deaths respectively).

Here the delay from vaccination to immunity weakens the impact of this control, whereas the refinement in the vaccination procedure increases the speed at which herd-immunity can be reached.

Finally, in all we have discussed above, it has been assumed that being vaccinated either confers complete protection, or it leaves the individual completely susceptible. An alternative is to assume that vaccines offer a more limited form of protection (known as leaky vaccination [117]) but that this protection has three distinct forms: it can reduce an individuals susceptibility to becoming infected; it can reduce the onward transmission rate (or speed up recovery) if a vaccinated individual does become infected; and it can reduce the severity of disease. This latter measure, while important from a medical and public-health perspective, has no direct impact on the epidemiology and can be ignored in the mathematical models. We can now recast the model for vaccination at birth using this leaky vaccine concept (figure 9).

$$\text{Susceptibles}\quad \frac{\mathrm{d}S}{\mathrm{d}t} = B(1-p) - \lambda S - dS,$$

$$\text{Infectious}\quad \frac{\mathrm{d}I}{\mathrm{d}t} = \lambda S - \gamma I - dI,$$

$$\text{Recovered}\quad \frac{\mathrm{d}R}{\mathrm{d}t} = \gamma I - dR,$$

$$\text{Vaccinated, uninfected}\quad \frac{\mathrm{d}V}{\mathrm{d}t} = Bp - \lambda\sigma V - dV,$$

$$\text{Vaccinated, infectious}\quad \frac{\mathrm{d}J}{\mathrm{d}t} = \lambda\sigma V - \gamma\rho J - dJ,$$

$$\text{Force of Infection}\quad \lambda = \beta I/N + \beta\tau J/N. \tag{12}$$

where the effect on vaccination on susceptibility, transmission and recovery are captured by the parameters $\sigma(\leqslant 1)$, $\tau(\leqslant 1)$ and $\rho(\geqslant 1)$. By considering the invasion of the disease-free equilibrium we can again derive the critical eradication threshold:

$$p_c = \frac{R_0 - 1}{R_0 - R_{\mathrm{V}}} \qquad R_{\mathrm{V}} = \frac{\tau\sigma\beta}{\rho\gamma + d}$$

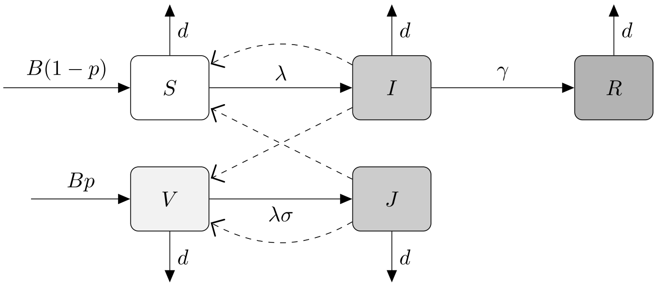

Figure 9. Compartmental caricature of the generic SIRVJ model (solid lines show transition or movement between classes; dashed lines show the action of transmission; arrows entering or leaving the system correspond to births and deaths respectively).

where R_{V} is the equivalent of the basic reproductive ratio when the entire population is vaccinated. From this formulation it is clear that any of these three actions (separately or in combination) can lead to eradication of the infection as long as $R_{\mathrm{V}} < 1$.

3. Heterogeneous populations

Heterogeneous populations are ones in which the total host population may be partitioned into two or more groups, classes or populations with distinct characteristics; this may be according to a variety of reasons including multiple species, ages groups or risk structure. In each of these there may be biological or behavioural differences which affect the spread or transmission of disease. Sometimes it may be easy to compartmentalize a population, not only according to disease status (e.g. S, E, I, R), but also by another natural partition of the population; for multiple species this is particularly easy, as each different species would form its own category. Conversely it may be hard to decide upon exactly *how* it is best to separate out a varying range of behaviour; for instance for STIs, where the number of sexual partners an individual has puts them at differential risks of contracting a disease, it is far from obvious how to determine where the boundary between 'low risk' and 'high risk' lies. Similarly when dealing with age structure, it is unclear how to partition the population into discrete age classes.

In the SIR modelling framework, instead of having just three states (according to infection status) there are now three states per population category. In the case of two populations (these can be thought of as two different species in this example) the SIR diagram (without demography) now looks a little different (see figure 10).

In the standard SIR model transmission of the pathogen between all individuals in the population is assumed to be governed by a single rate parameters (β) which is equivalence to assuming that all individuals mix randomly with each other. When the population is heterogeneous, we have the opportunity to allow different transmission rates within and between the classes. In particular, the transmission terms now depend upon who acquires infection from whom, and these new rates are now give by terms of the form $\beta_{\text{To From}}$ (i.e. β_{21} denotes the transmission rate of disease going to species 2 from species 1).

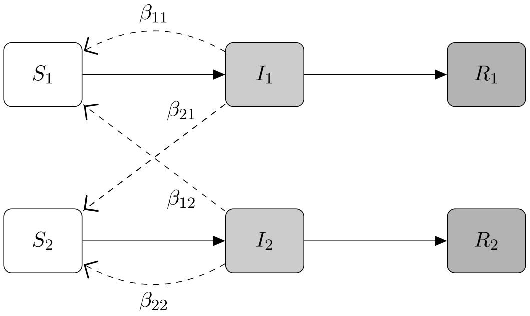

Figure 10. Compartmental caricature of the basic two-population SIR model.

A general multi-species model therefore takes much the same form as discussed above, but the rates are now species dependent and the force of infection is expressed as a weighted sum. For species i the SIR dynamics are given by

$$\begin{aligned} &\text{Susceptibles} \quad \frac{\mathrm{d}S_i}{\mathrm{d}t} = B_i - \lambda_i S_i - d_i S_i, \\ &\text{Infectious} \quad \frac{\mathrm{d}I_i}{\mathrm{d}t} = \lambda_i S_i - \gamma_i I_i - d_i I_i, \\ &\text{Recovered} \quad \frac{\mathrm{d}R_i}{\mathrm{d}t} = \gamma_i I_i - d_i R_i, \qquad (13) \\ &\text{Force of Infection} \quad \lambda_i = \sum_j \beta_i j I_j. \end{aligned}$$

This transmission interdependence becomes more obvious if we write in full a two-class model and SIR dynamics:

$$\begin{aligned} &\text{Class 1} \quad \begin{cases} \text{Susceptibles} & \frac{\mathrm{d}S_1}{\mathrm{d}t} = -\beta_{11}S_1I_1 - \beta_{12}S_1I_2, \\ \text{Infectious} & \frac{\mathrm{d}I_1}{\mathrm{d}t} = \beta_{11}S_1I_1 + \beta_{12}S_1I_2 - \gamma I_1, \\ \text{Recovered} & \frac{\mathrm{d}R_1}{\mathrm{d}t} = \gamma I_1, \end{cases} \\ &\qquad\qquad (14) \\ &\text{Class 2} \quad \begin{cases} \text{Susceptibles} & \frac{\mathrm{d}S_2}{\mathrm{d}t} = -\beta_{21}S_2I_1 - \beta_{22}S_2I_2, \\ \text{Infectious} & \frac{\mathrm{d}I_2}{\mathrm{d}t} = \beta_{21}S_2I_1 + \beta_{22}S_2I_2 - \gamma I_2, \\ \text{Recovered} & \frac{\mathrm{d}R_2}{\mathrm{d}t} = \gamma I_2, \end{cases} \end{aligned}$$

where the force of infection for each class λ_i has been included explicitly, and we have scaled the population size to one. This formulation for transmission can be represented more succinctly in the form of a transmission matrix known as the WAIFW (or Who Acquires Infection From Whom) matrix:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_{11} & \beta_{12} \\ \beta_{21} & \beta_{22} \end{pmatrix}. \qquad (15)$$

This allows the change of infectious individuals to be written in vector notation:

$$\frac{\mathrm{d}\boldsymbol{I}}{\mathrm{d}t} = \boldsymbol{S} \cdot \boldsymbol{\beta}\boldsymbol{I} - \gamma \boldsymbol{I}. \qquad (16)$$

The same heterogeneous population structure, transmission matrix and vectorization concepts can be applied in a similar fashion to SEIR, SIS and other models of disease progression.

In the unstructured SIR model (see section 2) parametrization of the infection rate, β, is relatively easy and is often derived from the equilibrium dynamics; in particular given that $S^* = \gamma/\beta$ and both the proportion seronegative (susceptible) and the duration of infection can be measured with relative ease, calculating β is not conceptually challenging. However, in heterogeneous populations where we have partitioned the host into n classes, estimating the matrix, $\boldsymbol{\beta}$ in much harder as there are now n^2 entries and often data for just n different data points. This means that often we are only able to estimate n parameters within the WAIFW matrix, and in general *any* matrix with n degrees of freedom can be made to fit the equilibrium dynamics. It therefore requires much more careful thought as to how to parameterise the matrix in these cases; how this is done depends on the specific population, the disease being modelled and the epidemiological understanding of the transmission routes [110]. This will be discussed in more detail below.

3.1. Heterogeneous reproductive ratio

As for single homogeneous population models, being able to estimate the basic reproductive ratio, R_0, for heterogeneous populations is highly desirable. As before this ratio may be derived from first principles by considering how individuals become infected and generate secondary cases. We will begin with a two-population model, although the argument generalizes to multiple populations. The basic reproductive ratio for an infected individual in population 1 or 2 (R_{01} and R_{02}, respectively) can be found as in section 2.2:

$$\begin{aligned} R_{01} &= \text{Average infectious period of 1s} \\ &\quad \times (\text{rate of generating infected 1s} \\ &\quad + \text{rate of generating infected 2s}) \\ &= \frac{1}{\gamma_1}(\beta_{11}N_1 + \beta_{21}N_2), \qquad (17) \\ R_{02} &= \frac{1}{\gamma_2}(\beta_{12}N_1 + \beta_{22}N_2). \end{aligned}$$

Note that the definition that everyone is susceptible, now becomes $S_1 = N_1$ and $S_2 = N_2$, where N_1 and N_2 are sizes of the two populations.

To calculate a population-level basic reproductive ratio, it is now necessary to average these two values. Here we return to the definition of R_0 which states that it is the number of secondary cases produced by an *average* infected individual; we therefore need to calculate what populations *average* infected individual belongs to during the early dynamics following disease invasion. To do this we return to the differential equations and the eigenvectors of the disease-free equilibrium. In particular, the eigenvector associated with the dominant eigenvalue will have the form $(-\widetilde{S}_1, \widetilde{I}_1, \widetilde{R}_1, -\widetilde{S}_2, \widetilde{I}_2, \widetilde{R}_2)$. The true population-level value of R_0 is then a weighted average of the individual R_0 values according to the ratio of early infection:

$$R_0 = R_{01}\frac{\widetilde{I}_1}{\widetilde{I}_1 + \widetilde{I}_2} + R_{02}\frac{\widetilde{I}_2}{\widetilde{I}_1 + \widetilde{I}_2}. \qquad (18)$$

This value of R_0 has many of the same properties as the value calculated for homogeneous (single-population) models: $R_0 = 1$ defines the critical threshold at which invasion is successful in a totally susceptible population, and $1 - 1/R_0$ determines the level of random immunization needed to eradicate an infection.

As well as different disease progression types (e.g. SIS, SEIR, etc, see section 2.1), heterogeneous populations may exhibit different types of transmission behaviour dependent upon the disease and population(s) in question. It is important to understand the underlying mechanisms which drive infection events, whether they are behavioural and/or biological. There is no set rule of partitioning populations; is not the same for every population, sometimes it may be straightforward and in others individuals may even start in one sub-population and end up in another. There are some key variants which highlight many of these features which are now discussed.

3.2. Age

Age structure is generally considered to be highly important, in particular in the transmission of disease amongst human populations. In general children of school age have very different mixing behaviour to adults or pre-school children; there are strong levels of assortative mixing (i.e. children meet a lot of other children at school whereas adults spend time with other adults at work). This means that it would be expected that the diagonal terms in the WAIFW matrix are dominant. In addition, there are important physiological differences between adults and children, meaning that children can respond very differently to infection; for example children and the elderly can often be the most severely affected by a disease. Also, children's behaviour (with regard to hygiene and physical interactions between individuals) is far different from adults; because of this and physiological effects, it is common that $\beta_{\text{Child Child}} > \beta_{\text{Adult Adult}}$.

Of more fundamental importance to both the modelling and the epidemiology, is that in addition to the standard progression through infectious states (that is left to right in figure 11) there is also movement between classes (top to bottom in figure 11). All individuals are born into the first class (children) and, after a period of time, move into the adult class. Movement between classes conserves epidemiological status, such that infected children that mature become infected adults. One important consequence of such age-structured models is that adults are older than children, and have therefore had longer to come into contact with any endemic infection. Therefore, at equilibrium, susceptibility declines with age.

Usually when considering long-term processes such as ageing, demography (births and natural deaths) are also incorporated, with individuals being born into the youngest age-class. Denoting children and adults by subscripts C and A, the two-class age-structure model becomes

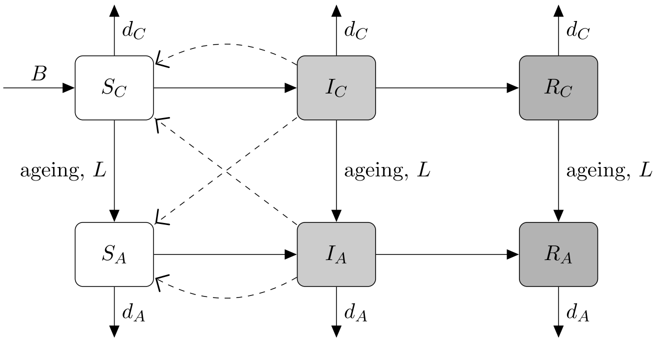

Figure 11. Compartmental caricature of the child/adult age-structured population SIR model.

$$
\begin{aligned}
\frac{\mathrm{d}S_\mathrm{C}}{\mathrm{d}t} &= B - \beta_\mathrm{CC}S_\mathrm{C}I_\mathrm{C} - \beta_\mathrm{CA}S_\mathrm{C}I_\mathrm{A} - LS_\mathrm{C} - d_\mathrm{C}S_\mathrm{C},\\
\frac{\mathrm{d}I_\mathrm{C}}{\mathrm{d}t} &= \beta_\mathrm{CC}S_\mathrm{C}I_\mathrm{C} + \beta_\mathrm{CA}S_\mathrm{C}I_\mathrm{A} - \gamma I_\mathrm{C} - LI_\mathrm{C} - d_\mathrm{C}I_\mathrm{C},\\
\frac{\mathrm{d}R_\mathrm{C}}{\mathrm{d}t} &= \gamma I_\mathrm{C} - LR_\mathrm{C} - d_\mathrm{C}R_\mathrm{C},\\
\frac{\mathrm{d}S_\mathrm{A}}{\mathrm{d}t} &= -\beta_\mathrm{AC}S_\mathrm{A}I_\mathrm{C} - \beta_\mathrm{AA}S_\mathrm{A}I_\mathrm{A} + LS_\mathrm{C} - d_\mathrm{A}S_\mathrm{A}, \qquad (19)\\
\frac{\mathrm{d}I_\mathrm{A}}{\mathrm{d}t} &= \beta_\mathrm{AC}S_\mathrm{A}I_\mathrm{C} + \beta_\mathrm{AA}S_\mathrm{A}I_\mathrm{A} - \gamma I_\mathrm{A} + LI_\mathrm{C} - d_\mathrm{A}I_\mathrm{A},\\
\frac{\mathrm{d}R_\mathrm{A}}{\mathrm{d}t} &= \gamma I_\mathrm{A} + LR_\mathrm{C} - d_\mathrm{A}R_\mathrm{A}
\end{aligned}
$$

with the parameter L capturing the per capita rate at which children leave the childhood class and move into the adult population (figure 12).

In fact, as age structure is so influential in some diseases, that the population can be partitioned into a whole range of age classes; examples include: pre-school, primary school, secondary school and adult [152]; different age classes for each school age year from 0 to 20 [53, 224]; or to match recorded age-related records [18]. It is worth noting that in the limit of infinitely many age classes this discrete age structured model becomes an integro-PDE of the form

$$
\begin{aligned}
\frac{\partial S(a)}{\partial t} &= B\delta(a) - S(a)\int_0^\infty \beta(a,a')I(a')\,\mathrm{d}a' - dS(a) - \frac{\partial S(a)}{\partial a},\\
\frac{\partial I(a)}{\partial t} &= S(a)\int_0^\infty \beta(a,a')I(a')\,\mathrm{d}a' - \gamma I(a) - dI(a) - \frac{\partial I(a)}{\partial a},\\
\frac{\partial R(a)}{\partial t} &= \gamma I(a) - dR(a) - \frac{\partial R(a)}{\partial a}, \qquad (20)
\end{aligned}
$$

where both ageing and time are continuous.

To return to the issue of parametrization mentioned earlier, we have already stated that because of child behaviour and physiology, it is expected that $\beta_\mathrm{CC} > \beta_\mathrm{AA}$ and that the WAIFW matrix is assortative (the diagonal terms dominate). However, in principle we have four transmission parameters that need to be estimated within these constraints (or more generally n^2 parameters for n populations). If we are faced with just information on the endemic equilibrium (or alternatively just on the early dynamics), then this at most allows us to fit two transmission parameters with some freedom in how the WAIFW matrix is constructed. Commonly assumed forms for

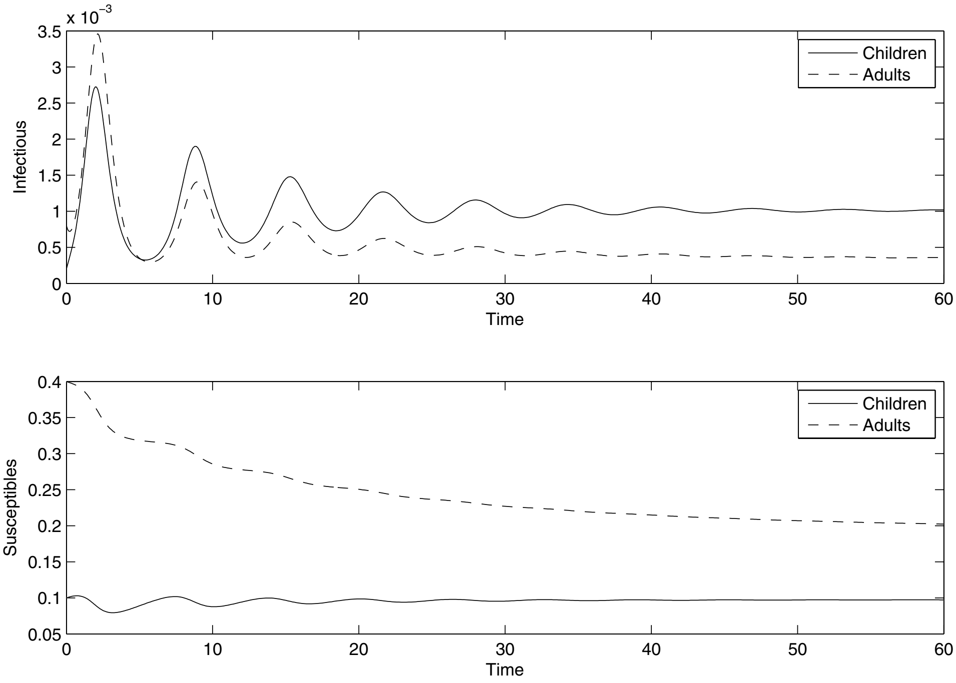

Figure 12. An example of a two-class age structured SIR model with parameters for a typical human population ($\beta = \begin{pmatrix} 100 & 10 \\ 10 & 20 \end{pmatrix}$, $\gamma = 10$, $B = d_C = 1/60$ years^{-1}, $d_A = 0$ and $L = 1/15$ years^{-1}). In this case 50% of each population start as susceptible and 0.01% of each are infectious. Initially it is seen that many naïve adults become infected (adults make up 80% of the population), however as time passes, the infection is predominantly seen in the child population even though it is smaller in size; this reflects that as children, individuals are at risk of becoming infected with the disease and if this occurs they will be not become a susceptible adult once they mature.

the WAIFW matrix are:

$$\beta = \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_2 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 & \beta_1 \\ \beta_2 & \beta_2 \end{pmatrix} \qquad \beta = \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_1 & \beta_2 \end{pmatrix}$$

$$\beta = \begin{pmatrix} \beta_1 & \beta_2 \\ \beta_2 & \beta_1 \end{pmatrix},$$

which correspond to special transmission between children, age-dependent susceptibility, age-dependent transmissibility, and strong assortativity. All of these matrices can be parametrized to generate the same endemic levels of infection and susceptibility; which form is chosen depends on additional information about the relative strengths of the transmission routes. This problem obviously becomes more complex as more age classes are included in the model formulation.

One method to address this uncertainty is to measure the mixing patterns between the age classes—although this does not directly quantify the degree of transmission. The pan-European POLYMOD study [199] surveyed 7,290 people in eight countries to determine age-dependent mixing patterns. The data generated has allowed far better parametrization of comparmentalized age-structured models, and is now a standard approach that has been used in a variety of setting from the shifting patterns in both the incidence of pertussis (whooping cough) [218], to control of pandemic influenza [185].

3.3. Super-spreader and super-shedders

Super-spreaders and super-shedders as the names suggest are individuals that are often responsible for relatively large numbers of secondary cases. Hence understanding the role of such individuals in an epidemic has clear implications for disease containment and control.

Biologically speaking a *super-shedder* is an individual that, due to some underlying biological or physiological mechanism, excretes (or 'sheds') a significant amount more infectious material than a 'normal' member of the population. Because of this, the individual is responsible for a higher than average proportion of the infections which occur; mathematically, if the transmission from super-shedders is f times higher than from normal individuals, then transmission from super-shedder to any other individual, regardless of class, will be $f\beta$ (figure 13). It is usually assumed that such individuals are born into the super-shedders class (so that if they become infectious they *will* be a super-shedder) and remain super-shedders for life.

Johne's Disease is a prime example of when super-shedders may be important. It is caused by Mycobacterium avium subspecies paratuberculosis and can infect a range of animals, although the main focus of attention is cattle. The disease causes wasting in animals, and eventually death. Recently there has been much interest in this disease as it is known that some dairy cows spread much more infectious material that others (in some cases this has been found to be up to 23 000 as much! [99]); many papers describe how identifying the super-shedder cows within a herd is difficult and costly [6] but without such measures the economic loss through reduced milk production and cattle death may be great [211]. It is clear that modelling work (using the types of heterogeneous model outlined in this section, potentially combined with stochastic models to account for small population sizes on

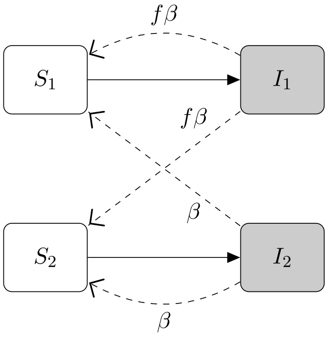

Figure 13. Compartmental caricature of super-shedder/non-super-shedder structured population SIR model. Population 1 is the super-shedders, and it is generally assumed that f much larger that 1, and that most individuals belong to population 2.

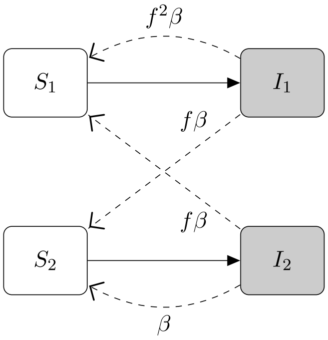

Figure 14. Compartmental caricature of super-spreader/non-super-spreader structured population SIR model. Population 1 is the super-spreaders, and it is generally assumed that f much larger that 1, and that most individuals belong to population 2.

farms) needs to be done to look at possible controls to reduce the spread of the disease and the economic impact of Johne's disease for farmers.

Although *super-spreaders* [151] (also referred to as *super-contactors* [182]) are responsible for way above average number of secondary cases, the underlying causes are very different to that of the super-shedders. These individuals shed 'normal' amounts of infectious material, however they have very high levels of mixing (they come into contact with a large number of other individuals) and hence are able to infect many more others. The disease transmission for super-spreaders, is therefore governed not by biology, but by behaviour (figure 14).

The most prominent examples of super-spreaders come from risk-structured models for STIs, which take into account the high variability in the number of sexual partners between individuals. People who have many different sexual partners are super-spreaders (high-risk), those with few partners are low-risk. High-risk people are much more likely to partner (mix with) other high-risk members of the population and so it would be expected that this is where a large proportion of transmissions occur (see section 3.6.3).

West Nile virus provides another example of a disease with prominent and important super-spreaders. West Nile virus entered the east coast of North America in 1999 and gradually tracked westwards, this is an avian infection that is spread by mosquito vectors (see section 3.5) but can also infect humans with often severe consequences. A full model would therefore need to include human, mosquito and bird populations, however the bird population is not homogeneous. The American Robin (*Turdus migratorius*) has been found to be the culprit for a disproportionate amount of disease propagation [179]. Kilpatrick *et al* [179] observe that mosquitoes have a preference for the American Robin over other species of bird and hence the Robin has much higher contact rate, acting as a super-spreading species.

Whilst the distinction between super-shedders and super-spreaders may appear slight, the effect of super-spreaders on a population is much greater than that of super-shedders due to the $f^2\beta$ transmission term (see figures 13 and 14). Figure 15 provides an illustrative example of invasion of infection into a totally susceptible population under the two different model assumptions; it is clear that the presence of super-spreaders greatly enhances the spread of infection.

These differences of model assumptions translate into structural difference in the WAIFW matrix and differences in the basic reproductive ratios:

$$\boldsymbol{\beta}_{\text{super-shedders}} = \begin{pmatrix} f\beta & \beta \\ f\beta & \beta \end{pmatrix} \qquad R_0 = \frac{\beta}{\gamma}(fN_1 + N_2),$$

$$\boldsymbol{\beta}_{\text{super-spreaders}} = \begin{pmatrix} f^2\beta & f\beta \\ f\beta & \beta \end{pmatrix} \qquad R_0 = \frac{\beta}{\gamma}(f^2N_1 + N_2). \qquad (21)$$

For super-spreaders the transmission matrix, $\boldsymbol{\beta}$, is symmetric as there is the same likelihood of transmission between both classes regardless who is infecting whom. Conversely, the same is not true of the super-shedders who are no more likely to become infected than any other individuals, but then transmit more readily. These matrices emphasize the slight but important differences in model structure between the biological and behavioural phenomena. The associated basic reproductive ratios highlight the quantitative implications, with the value for super-spreaders greater for all $f > 1$. It should be noted that given that super-spreaders are distinguished in terms of their behaviour, it may be possible to observe the interactions of a population and hence detect super-spreaders before an epidemic [182]. In contrast, it may be far more difficult to predict which individuals are likely to be super-shedders, and this may differ widely for different pathogens.

3.4. Multiple species

For two or more different species not only do we expect transmission parameters to differ but it is also likely that other rates, such as birth/death and recovery, may also be different due to inherent differences between species. Two species models are often needed when both livestock and a wildlife reservoir can act as hosts for the same pathogen; African buffalo are a clear example where control on foot-and-mouth disease in cattle is limited by the presence of infection in wildlife [39], while the controversy in the UK about the role of badgers in the transmission of bovine Tuberculosis [156, 162] largely hinges around estimating transmission parameters. In addition, there are a number of infectious diseases, known as zoonoses, (such as bubonic plague or West Nile virus) which can infect both humans and other animals, and so

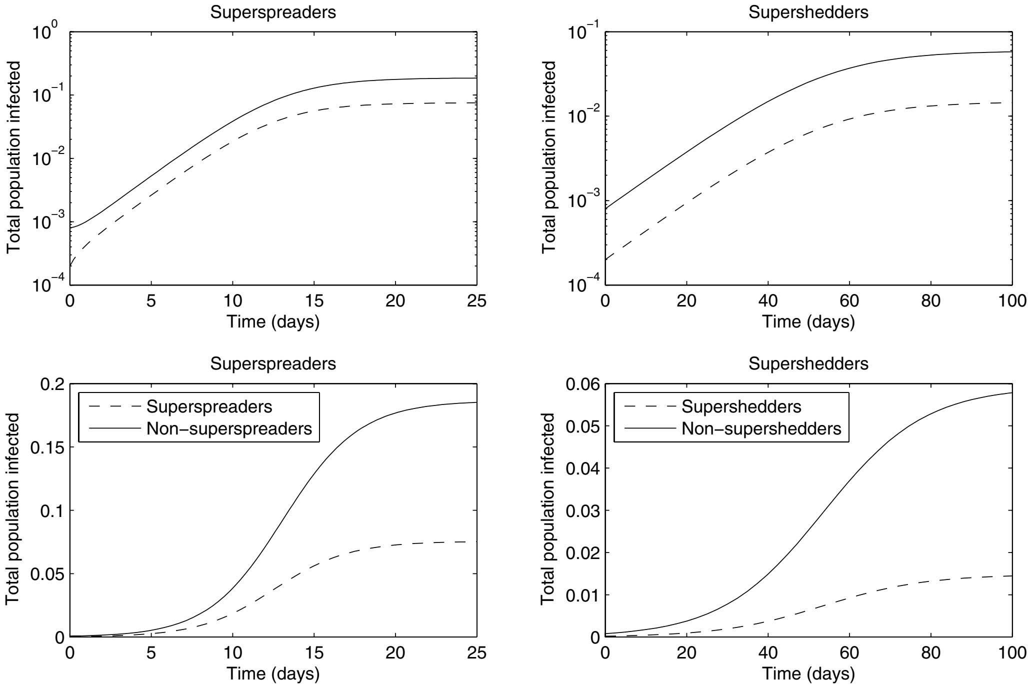

Figure 15. Example of the differences between the roles of super-spreaders and super-shedders during an epidemic (no demography) where $f = 2$, $\beta = 0.9$ and $\gamma = 1$. Super-shedders make up 20% of the total population and initially all individuals are susceptible excluding the 0.1% of each population which start infectious. Whilst the two epidemics look qualitatively similar, the outbreak of the epidemic across the population including super-shedders is slower and smaller than that of the super-spreader epidemic. This corresponds with the difference in reproductive ratios (R_0) of 1.08 and 1.44 for super-shedders and super-spreaders respectively.

models must take both species into account. However, the formulation of these two species models adds little novelty to the methods discussed above. One situation where there are subtle, yet important differences, is when individuals of one species (vectors) are required for transmission of infection between members of a different species; models for such vector-borne diseases are now explored in some detail with particular emphasis on the dynamics of malaria.

3.5. Case study: vector-borne disease

Some of the biggest world-wide health challenges at the moment (such as malaria) are vector-borne infections therefore accurate predictive models are vital to target controls effectively. In addition, vector-borne infections require a very specific type of multi-species model, to account for the cross-species interactions that are involved in transmission. We first outline some basic vector biology to motivate the transmission models.

3.5.1. Introduction to the biology of VB diseases. A vector-borne disease is one for which transmission in a primary (host) population only occurs via a secondary (vector) population. The host population may consist of a human population (for malaria or dengue infections), livestock (for bluetongue infections) or other wild populations such as birds (for West Nile virus). Vectors are generally blood-sucking arthropods and some of the most important vectors for the spread of disease are the mosquito (responsible for the spread of a multitude of diseases including malaria, west nile virus, dengue fever, etc), the sandfly (which notably causes leishmaniasis) and ticks (which can transmit Lyme disease). Many vector-borne diseases are deadly to humans and Africa is the continent with the high prevalence of vector-borne disease-induced mortality. In general the majority of vector-borne diseases occur in the tropics where the climate is most suitable for the vectors, although with climate change the range of vector-borne infections is increasing.

Malaria (caused by a protozoan parasite) is the deadliest vector-borne disease to humans; it kills around 1.2 million people per year [221] and is endemic in much of Africa and other regions with tropical climates. As well as persisting as an endemic disease in many locations, recently there have been epidemic outbreaks in areas which were formerly malaria free as well as sharp rises in the number of cases in endemic regions. The vector for human malaria is the female *Anopheles* mosquito which takes blood meals from humans in order to complete her reproductive cycle [115].

The specifics of the transmission does vary slightly across different vector-borne diseases, as do the host and vector populations, however the basic transmission cycle remains fairly consistent: the cycle begins with an infected host receiving a bite from a susceptible vector, as the vector take its blood meal the pathogen is transmitted to the vector and replicates until eventually the vector becomes infectious. At this stage upon biting a susceptible host the vector will transmit the disease back to the host population. This very specific biological transfer requires both the host and vector population and so in the absence of one or the other, the pathogen cannot spread through a population.

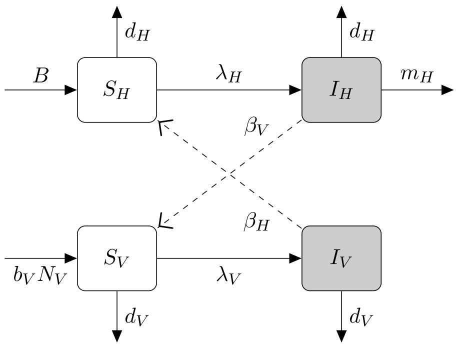

Figure 16. Compartmental caricature of Vector-borne infection model showing criss-cross transmission terms.

3.5.2. Outline of basic VB disease modelling. Modelling vector-borne infections is quite different to modelling directly transmitted pathogens due to the biological process involved; a human cannot directly infect another human with malaria (or in general a host cannot infect another host), and this needs to be taken into account with an appropriate model. A vector-borne disease model can be thought of as a specific type of heterogeneous population model, where the natural partition of the total population is into the two species: the host and the vector. Typically for each population SIR or SI-type dynamics are assumed, however in contrast to other models where a high level of assortativity between individuals in a class is often expected, there is no transmission *within* classes and hence a transmission (or WAIFW) matrix is of the form

$$\beta = \begin{pmatrix} 0 & \beta_{\rm H} \\ \beta_{\rm V} & 0 \end{pmatrix}.$$

For ease of notation the double subscript is dropped and β_i corresponds to transmission *to* species i. This type of transmission is know as criss-cross transmission (figure 16) as no assortative transmission occurs—individuals cannot contract the infection directly from another individual of the same species. Although the total population is partitioned into these two very separate categories, homogenous mixing between vectors and hosts is assumed. It is worth noting that this not the only example of criss-cross transmission, for example in the spread of STIs in a strictly heterosexual population, all individuals only mix with others of the opposite sex and so same-sex transmission does not occur directly.

Transmission of the vector-borne infections is generally considered to be frequency rather than density dependent; once the vector is sated from its blood-meal it will not bite again until necessary (for the female mosquito, one blood meal is usually taken per batch of eggs and feeding occurs on average once every four days [183]). The transmission dynamics are determined by this vector biting rate (a), the probability of a bite leading to infection for the host or vector ($p_{\rm H}$ and $p_{\rm V}$, respectively) and the numbers of susceptible and infectious hosts ($S_{\rm H}$ and $I_{\rm H}$) and vectors ($S_{\rm V}$ and $I_{\rm V}$). Where the number of vectors always refers to those of biting maturity (and gender). This gives the force of infection as:

$$\lambda_i = \beta_i I_j = \frac{a p_i I_j}{N_{\rm H}} \qquad \text{where } i \neq j.$$

The standard Ross–MacDonald Malaria model [177, 220] is usually given in the following form where parameters are as before, with the addition of a disease-induced per capita mortality rate, $m_{\rm H}$.

$$\begin{aligned} \text{Hosts} \quad \frac{{\rm d}S_{\rm H}}{{\rm d}t} &= B + \gamma_{\rm H} I_{\rm H} - \beta_{\rm H} S_{\rm H} I_{\rm V} - d_{\rm H} S_{\rm H}, \\ \frac{{\rm d}I_{\rm H}}{{\rm d}t} &= -\gamma_{\rm H} I_{\rm H} + \beta_{\rm H} S_{\rm H} I_{\rm V} - (d_{\rm H} + m_{\rm H}) I_{\rm H}, \\ \text{Vectors} \quad \frac{{\rm d}S_{\rm V}}{{\rm d}t} &= b_{\rm V} N_{\rm V} + \gamma_{\rm V} I_{\rm V} - \beta_{\rm V} S_{\rm V} I_{\rm H} - d_{\rm V} S_{\rm V}, \\ \frac{{\rm d}I_{\rm V}}{{\rm d}t} &= -\gamma_{\rm V} I_{\rm V} + \beta_{\rm V} S_{\rm V} I_{\rm H} - d_{\rm V} I_{\rm V}. \end{aligned} \tag{22}$$

Demography is included so that the long-term dynamics of the endemic disease can be studied. Parameters correspond not just to the disease in question but also to the populations, for example the natural death rate for humans, $d_{\rm H}$, is generally much smaller than that of the vector $d_{\rm V}$; for mosquitoes the average life expectancy is only 32 days [230]. In addition, there is no disease-induced mortality term for the vector population as the mosquito rarely suffers adverse effects from malarial infection. This system of equations (22) can easily be amended to incorporate a latency period by adding an additional *exposed*, E, class to both the host and vector populations.

R_0 may be again be determined from first principles using the expected duration of infection and the rates of transmission. However, for such vector-borne infections, it is standard to calculate the number of secondary host (human) cases generated by an average host (human) case, incorporating the cycle of transmission through the vector. As such we find the expected number of infected hosts from a single infected vector:

$$\begin{aligned} \text{infected hosts} &= \text{duration of infection} \times \text{rate of transmission} \\ &= \frac{1}{(d_{\rm V} + \gamma_{\rm V})} \times \lambda_{\rm H} \times N_{\rm H} \\ &= \frac{a p_{\rm H}}{(d_{\rm V} + \gamma_{\rm V})}, \end{aligned}$$

and similarly, the expected number of infected vectors from a single infected host:

$$\begin{aligned} \text{infected vectors} &= \text{duration of infection} \times \text{rate of transmission} \\ &= \frac{1}{(d_{\rm H} + m_{\rm H} + \gamma_{\rm H})} \times \lambda_{\rm V} \times N_{\rm V} \\ &= \frac{a p_{\rm V} N_{\rm V}}{(d_{\rm H} + \gamma_{\rm H} + m_{\rm H}) N_{\rm H}}. \end{aligned}$$

Therefore the expected number of cases generated around a complete cycle (R_0) is given by the product:

$$R_0 = \frac{a^2 p_{\rm H} p_{\rm V} N_{\rm V}}{(d_{\rm H} + \gamma_{\rm H} + m_{\rm H})(d_{\rm V} + \gamma_{\rm V}) N_{\rm H}} \tag{23}$$

(It should be noted that this value of R_0 that includes a complete cycle is the square of the value calculated using the eigenvector approach given in section 3.1; however, they agree on the invasion threshold).

This reproductive ratio yields some important information regarding the spread of vector-borne disease. Firstly, although

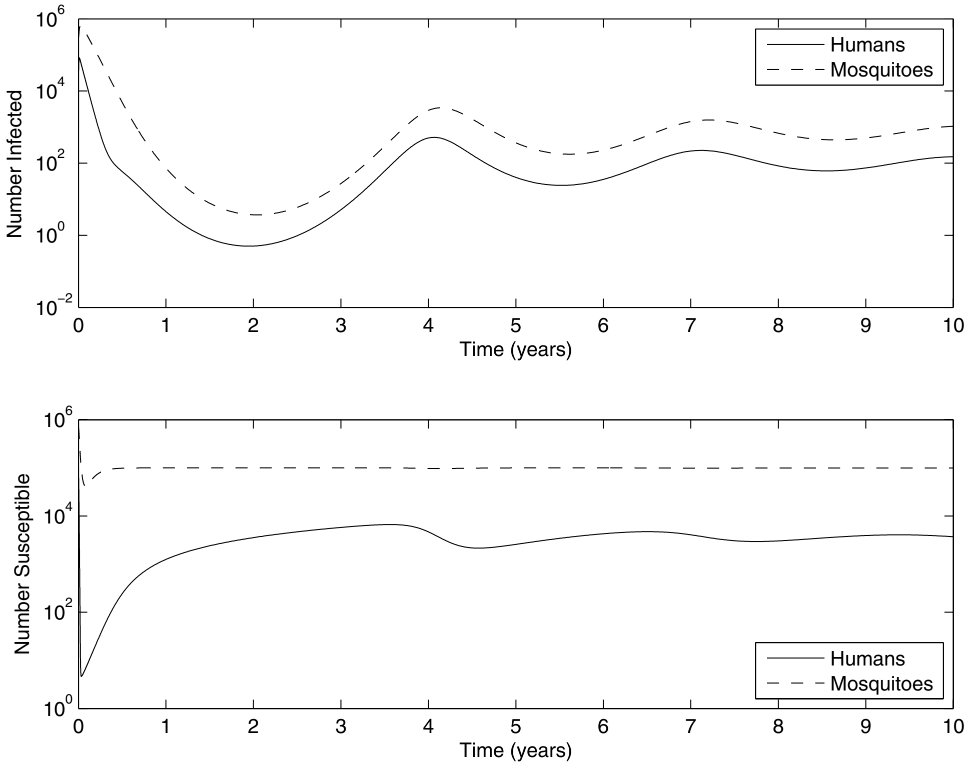

Figure 17. Human–mosquito malaria epidemic. ($N_\mathrm{H} = 10^5$, $N_\mathrm{V} = 10^6$, $a = 1/4\,\mathrm{days}^{-1}$, $p_\mathrm{H} = p_V = 0.9$, $\gamma_\mathrm{H} = 1/30\,\mathrm{days}^{-1}$, $B = b_\mathrm{H}N_\mathrm{H}$, $b_\mathrm{H} = d_\mathrm{H} = 1/42\,\mathrm{years}^{-1}$, $m_\mathrm{H} = 1/40\,\mathrm{days}^{-1}$, $b_\mathrm{V} = d_\mathrm{V} = 1/32\,\mathrm{days}^{-1}$).

there are several parameters which influence the value of R_0, births of either hosts or vectors do not play a part. Secondly (and somewhat intuitively) as the bite rate or transmission probabilities increase, so does R_0 although the bite rate enters in a squared form given that two bites are needed to complete a cycle. Finally, and most importantly, the ratio of vectors to humans is critical; if we are able to reduce the mosquito population to an appropriate level (compared to the size of the human population) it may be possible to reduce $R_0 < 1$; below the critical threshold. This result is equivalent to Ross's threshold theorem for malaria [220] which states that if the number of mosquitoes is below a certain figure then the amount of malaria in a given area will reduce to zero—disease extinction without the need to completely eradicate the mosquito.

3.5.3. Quasi-equilibrium assumption.

The *quasi-equilibrium assumption* is an approximation method which may be used successfully on this system of ODEs due to the nature of the time scales involved; since the vector rates are generally much more rapid to the corresponding host rates, the model becomes a multi-scale analysis problem. (In enzyme kinetics this approach is more commonly referred to as the quasi-steady-state assumption (QSSA) and in other disciplines as the method of multiple scales.)

The *quasi-equilibrium assumption* [151] states that due to the relatively short life expectancy of a vector compared to the host species (a human typically lives 500 times longer than a mosquito) and the time-scale of the disease, that an individual vector effectively sees a sustained level of infection within the host population for the duration of its life and so the dynamics can be approximated by assuming that $\frac{\mathrm{d}S_\mathrm{V}}{\mathrm{d}t} = \frac{\mathrm{d}I_\mathrm{V}}{\mathrm{d}t} = 0$ (and that the population is of constant size N_V). Hence the quasi-steady state vector populations are given by

$$S_\mathrm{V}^* = \frac{N_\mathrm{V}(b_\mathrm{V} + \gamma_\mathrm{V})}{\gamma_\mathrm{V} + \lambda_\mathrm{V} + d_\mathrm{V}},$$
$$I_\mathrm{V}^* = \frac{N_\mathrm{V}\lambda_\mathrm{V}}{\gamma_\mathrm{V} + \lambda_\mathrm{V} + d_\mathrm{V}}, \tag{24}$$

which are functions of S_H, I_H and N_H due to the dependence of λ_V on these variables. Combining these quasi-steady state solutions with the original host equations yields a new closed two-dimensional system:

$$\frac{\mathrm{d}S_\mathrm{H}}{\mathrm{d}t} = B + \gamma_\mathrm{H}I_\mathrm{H} - \lambda_\mathrm{H}S_\mathrm{H} - d_\mathrm{H}S_\mathrm{H}$$
$$\frac{\mathrm{d}I_\mathrm{H}}{\mathrm{d}t} = -\gamma_\mathrm{H}I_\mathrm{H} + \lambda_\mathrm{H}S_\mathrm{H} - (d_\mathrm{H} + m_\mathrm{H})I_\mathrm{H}, \tag{25}$$

where $\lambda_\mathrm{H} = \frac{ap_\mathrm{H}I_\mathrm{V}^*}{N_\mathrm{H}} \approx \frac{ap_\mathrm{H}}{N_\mathrm{H}} \frac{N_\mathrm{V}ap_\mathrm{V}I_\mathrm{H}}{(N_\mathrm{H}\gamma_\mathrm{V} + ap_\mathrm{V}I_\mathrm{H} + N_\mathrm{H}d_\mathrm{V})}$.

This quasi-equilibrium assumption may be justified more rigorously using the method of matched asymptotic expansions (for more details see [55, 106, 200]).

This method is advantageous as it halves the dimensionality of the system and so enables more simple analysis of the dynamics. Since it is usually only the host population which is of interest (from the point of view of reducing numbers of infections and subsequent deaths) this new system is very appealing to those in public health.

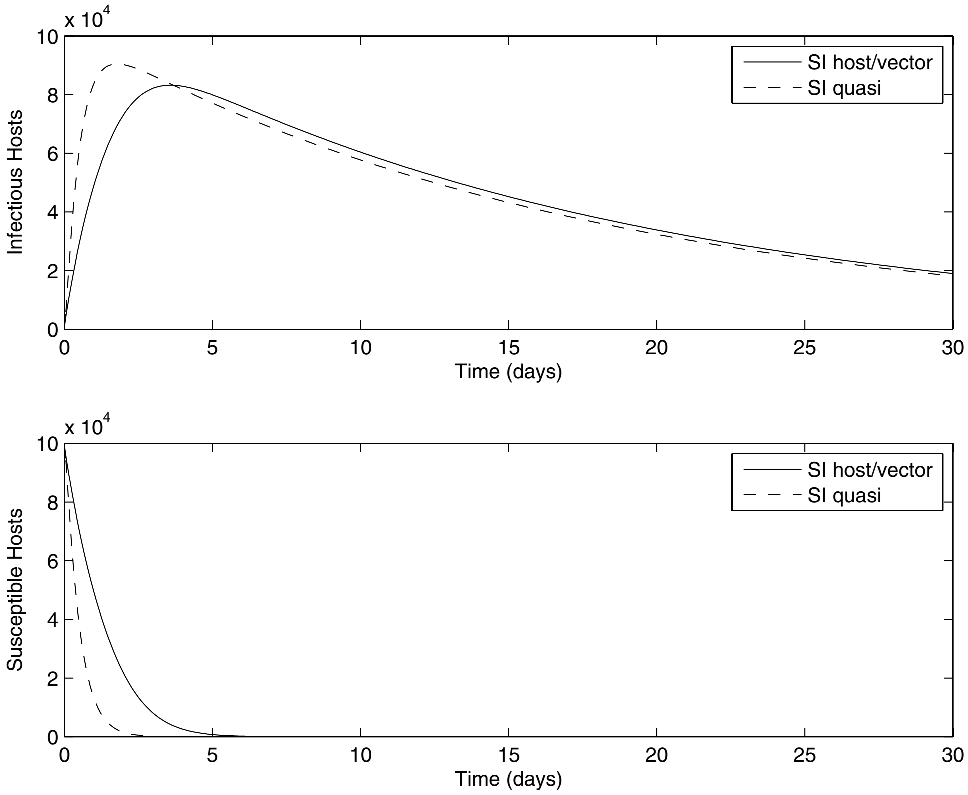

Figure 18. Example of how epidemic profiles may vary from the original deterministic system when using the quasi equilibrium assumption (parameters the same as in figure 17). Here the quasi-equilibrium assumption slightly overestimates the size of the epidemic and predicts that it will occur earlier.

The quasi-equilibrium assumption is exact for the equilibrium dynamics of the full system—as we have simply solved the equilibrium for the two species sequentially. Moreover the quasi-equilibrium assumption generally gives a good approximation to the full dynamics and whilst this should give a reasonable idea of the epidemic profile (with respect to amplitude and time scales), there will be slight temporal variation from the true behaviour as under the quasi-equilibrium assumption the vector population responds instantly to any changes in the host (see figure 18). This temporal variation is reflected in the reproductive ratio for the vector-borne model under the quasi-equilibrium assumption; R_0 is determined more simply than before as we now focus solely on the host dynamics:

$$\begin{aligned} R_0 &= \text{duration of infection} \times \text{rate of transmission} \\ &= \frac{1}{(d_H + \gamma_H + m_H)} \times \lambda_H \times N_H \\ &= \frac{(a^2 p_H p_V) N_V}{(d_H + \gamma_H + m_H)(N_H \gamma_V + a p_H + N_H d_V)}, \end{aligned}$$

which is always less that the value calculated from the full model (equation (23)), although in general the differences tend to be small.

3.5.4. Extensions to the vector-borne infection model. We have formulated the most basic model of a vector-borne disease such as malaria, however it would be naive to assume that all the inherent biology of the system has been captured within these simple equations. There are a number of key issues with this model which may lead to problems when trying to predicted outbreaks in specific populations.

Firstly, not all vector-borne diseases follow this exact pattern; for West Nile Virus evidence has been found that vertical transmission from females to eggs can occur in the mosquito (*Culex Univittatus*) [187]. Vertical transmission can be captured by a relatively simple modification to the vector equations:

$$\begin{aligned} \frac{dS_V}{dt} &= b_V S_V + (1-p) b_V I_V + \gamma_V I_V - \beta_V S_V I_H - d_V S_V, \\ \frac{dI_V}{dt} &= p b_V I_V - \gamma_V I_V + \beta_V S_V I_H - d_V I_V, \end{aligned} \tag{26}$$

where p is the probably that infection is transferred from mother to offspring. This leads to far greater persistence of infection in the vector population, and hence breaks the separation of time-scales needed for the quasi-equilibrium assumption. It should noted that there is no possibility of vertical transmission for either hosts or vectors for malaria.

The basic vector-borne disease model (in common with most simple epidemiological models) assumes that the death rate of both hosts and vectors is a constant per capita rate, leading to exponentially distributed life expectancies (see duration of infection in section 2.4). However the *age* at which a vector becomes infectious and hence its life expectancy whilst infectious determines the number of secondary infections that will result from this one vector. On average a vector which is infected at an early age will spend a longer time being infectious than a similar vector which was infected nearer to the end of its life, and hence will take more blood meals whilst infected and consequently spreads more infection to the host population. It has been suggested that logistically distributed life expectancies (and hence age-dependent death rates) are more realistic and may lead to better model predictions [41, 230]. To include this feature into the model we would need to combine the age-structured models (seen in section 3.2) with the vector models (of section 3.5.2), which would vastly increase the dimensionality of the system.

Climate is key for many vector-borne diseases. Mosquitoes in particular thrive in warm, humid climates. Generally mosquitoes need temperatures of 16 °C minimum [115] and pools of water to reproduce and, in even hotter conditions, the reproductive cycle shortens considerably. It is not just the mosquito that responds to climate, but also the malaria causing parasite which needs these high temperatures in order to develop. In regions where the temperature fluctuates during the year, peak transmission would be expected during the warm summer months when there is an abundance of mosquitoes and an appropriate temperature for parasite growth. There are many ways in which this seasonality may be incorporated, such as a temporally forced birth rate, biting rate or transmission probabilities. Taking vector birth rate as an example, usually such seasonal varying rates are give by a sinusoidal function of the form

$$b_V(t) = b_0(1 + b_1 \cos(\omega t)),$$

where ω is the period of forcing, b_0 is the average birth rate and b_1 is the amplitude of seasonality. However, this is not to suggest that there are no other potentially viable functions which could be based directly on data. There are clear parallels between this approach as the work on childhood infections in age-structured models, where seasonality is incorporated into the child-to-child transmission rate to capture the increase in mixing during school terms [53, 95, 224].

This basic deterministic model makes the assumption that hosts and vectors mix homogeneously, however the locations of a host and vector population may not overlap so neatly; in order to prevent desiccation of vectors and to reside a suitable environment for reproduction, mosquitoes are often situated around marshlands or over types of wet habitats. In general human towns would not coincide directly with such areas, and so human-mosquito interactions would be limited by this spatial aspect (the impact of spatial structure is discussed in more detail in section 5). In other cases, even when the mosquito is found in all areas, the spread of disease may not be consistent, particularly in mountainous regions where there is high temperature variation; this variation may lead to areas of low or no transmission, seasonal transmission and constant transmission with increasing temperature isoclines. Such a system may be modelled by partitioning the populations into regions of distinct transmission patterns (such as low/mid and high) and assume homogeneous mixing throughout a region with a small exchange between populations of different regions.

Some vectors are not so preferential in their choice of host; in the case of Lyme disease, the vector (*Ixodes* tick) usually feeds on a range of small mammal hosts and deer. However as can also be the case with other tick-borne diseases (such as virus which cause encephalitis and haemorrhagic fever [115]) once the tick population surpasses a threshold, the rise in numbers can cause a change from the ticks' enzootic (animal) hosts to human hosts, at which point zoonotic disease transmission occurs into the human population. Modelling vector-borne diseases with multiple host classes can be done in a compartmentalized structure to ensure these more complex dynamics are captured [169].

Recent years we have made huge scientific advantages due to technology (both computing power and physical capacities to conduct laboratory research). Despite this many biological phenomena, in particular within host pathogen-immune interaction and even basic vector biology, remain a mystery. Not only is there evidence to suggest that vectors may select specific hosts for their blood-meal dependent on host species [179, 227], but that they can also have a preference for infected hosts over susceptible ones [66]. Better understanding of such underlying biological mechanisms and this relationship between parasite, host and vector are necessary for improving the accuracy and effectiveness of mathematical models.

3.6. Case study: HIV

Since the first reported cases of AIDS (acquired immune deficiency syndrome) in the early 1980s, the disease has now claimed approximately 30 million lives worldwide [241] and there are an estimated 34 million people living with AIDS [242]. One approach to predict the future pattern of the epidemic is to use simple exponential or polynomial extrapolations of the observed trends to make predictions about the near future [58, 68, 122, 142, 195]. However, this approach has two main limitations: it may be sensitive to the type of functional form used to fit to historical data; it also provide no information or understanding of the HIV transmission mechanisms or the impact of interventions. Mathematical models parametrized to match epidemiological and clinical observations offer an alternative mechanistic approach.

Human immunodeficiency virus (HIV), the virus which causes AIDS, is transmitted through bodily fluids, notably blood and semen. As such, certain behaviours make an individual at greater risk of contracting HIV; most notably, unprotected sexual intercourse and intravenous drug use. Individuals regularly participating in such behaviours often form sub-populations of the general population and so it is natural to model HIV transmission amongst and between these sub-populations. Most of the early work modelling HIV transmission dynamics focused on male homosexual communities in developed countries where early cases were predominately detected [9, 14, 46–48, 134, 136, 181]. In developed countries, most new infections still occur in men that have sex with men or those sharing injecting equipment. In contrast, in developing nations, HIV is mainly spread through heterosexual contact, requiring a different set of modelling assumptions.

To illustrate the progression of HIV transmission modelling over the years we will start with a simple single-population compartment model, similar to the model described

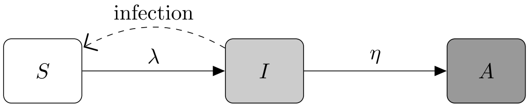

Figure 19. Generic SIA model (dashed lines show the direction of transmission).

in section 2.1. We will then add several complexities relevant to HIV transmission to the model and describe their impact on the predicted epidemic patterns. Given the huge public-health importance of HIV and the devastating effect that this pathogen can have on the population, there are a wealth of papers focussed towards predicting long-term dynamics. In many of these studies the aim is to generate accurate predictions and hence all known epidemiological and behavioural factors are included. Here we focus on specific elements of increased realism to better illustrate the effects of these heterogeneities.

3.6.1. One group: HIV in homogeneous male homosexual communities. The early work on modelling HIV transmission dynamics predominantly involved simple deterministic ODEs, similar to the SEIR equations (1) described in section 2.1. Male homosexual communities were the primary focus of these early models, as they were the primary focus of detected disease. For HIV the latent period following infection is very short compared to the infectious period and so the exposed class is often ignored [14]. However subsequent studies have found that the rates of transmission vary with the time since infection in a complex manner, often being high immediately after infection [48]. The basic model we adopt here consists of three compartments; susceptible individuals (S), infectious individuals (I) and infectious individuals who have developed AIDS (A), are symptomatic and assumed to no longer partake in sexual activity. Individuals transition between the compartments according to biologically and behaviourally defined rates (see figure 19). The full set of equations are

$$\begin{aligned}\text{Susceptible}\quad \frac{dS}{dt} &= -\lambda S,\\ \text{HIV positive}\quad \frac{dI}{dt} &= \lambda S - \eta I,\\ \text{AIDS}\quad \frac{dA}{dt} &= \eta I,\end{aligned}\tag{27}$$

where the force of infection λ is a product of the transmission probability per sexual contact τ, average rate of sexual partner change c, and proportion of the sexually active population who are infected, $I/(S+I)$ (individuals who have developed AIDS are assumed to be removed from the sexually active population)

$$\lambda = \tau c \frac{I}{S+I}.\tag{28}$$

As is the case for most infectious diseases in human populations, transmission is assumed to be frequency dependent and this is reflected by the form of λ; however, unlike many other infections the active populations size $(S+I)$ can be substantially reduced by the disease. Infectious individuals develop AIDS at rate γ. Clearly this model has parallels with the standard SIR model, with the A compartment playing a similar role to R.

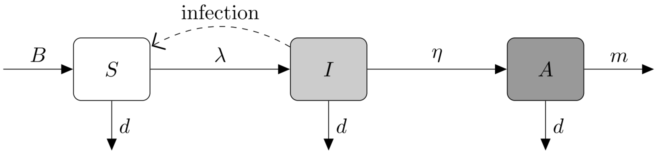

Figure 20. Generic SIA model with demography (dashed lines show the direction of transmission).

The time-scales of HIV infection are long, with an estimated infectious period of between 8 and 10 years even without treatment [194]. So for greater realism, demography must be included in the model (see figure 20). This is introduced in the form of an immigration rate B, with individuals entering the sexually active population as susceptible; in addition we include a disease-induced mortality rates m and a rate of leaving the sexually active population d:

$$\begin{aligned}\frac{dS}{dt} &= B - \lambda S - dS,\\ \frac{dI}{dt} &= \lambda S - (\eta + d)I,\\ \frac{dA}{dt} &= \eta I - (m+d)A.\end{aligned}\tag{29}$$

The terms B and d obviously mimic births and deaths in the standard SIR model, but refer to the sexually active population.

These early models assume that everyone in the population has the same number of partners and that the duration of partnerships can rapidly break and new ones reform. A natural modification is therefore to assume partnerships have a given duration [249]; explicitly including partnerships in this way slows transmission, as individuals in a susceptible-susceptible partnership are effectively protected against infection. However, if partnerships were allowed to overlap, such that one person could be in two concurrent partnerships, then the spread of infection is generally much faster was predicted to be far more rapid [81, 82, 178, 196, 249]. Concurrency also has a great impact when considering variable infectiousness over the infectious period (see section 3.6.6).

3.6.2. Multiple groups: HIV in heterogeneous male homosexual communities. In practice, not all individuals have the same number of sexual contacts over time leading to differences in the rate of change of sexual partners, c. In fact, a great deal of heterogeneity is observed in the number of sexual contacts of individuals in a population [171], with most people having very few partners in a given time period and a small number of individuals having very many. Many models have been formulated to account for this [14, 47, 140]. Following Jacquez *et al* [136] we partition the population into n groups based on sexual activity level, so that group i has on average c_i partners per unit time. Following similar methodology to before, we can calculate the transmission rates in terms of the underlying parameters. The total number of contacts per unit time of a susceptible individual in group i is c_i, the fraction of these contacts that are with group j is given by ρ_{ij}, $I_j/(S_j+I_j)$ is the probability that the contact with group j is with an infectious individual and τ_{ij} is the fraction of those contacts that result in transmission. Therefore the rate at which a susceptible individual in group i is infected by infectious individuals in group j is given by:

$$c_i \rho_{ij} \tau_{ij} \frac{I_j}{S_j + I_j}. \tag{30}$$

So the equations of this expanded heterogeneous population therefore become:

$$\begin{aligned} \frac{\mathrm{d}S_i}{\mathrm{d}t} &= B_i - \sum_{j=1}^{n} c_i \rho_{ij} \beta_{ij} S_i \frac{I_j}{S_j + I_j} - d_i S_i, \\ \frac{\mathrm{d}I_i}{\mathrm{d}t} &= \sum_{j=1}^{n} c_i \rho_{ij} \beta_{ij} S_i \frac{I_j}{S_j + I_j} - d_i I_i - \eta I_i, \\ \frac{\mathrm{d}A_i}{\mathrm{d}t} &= \eta I_i - d_i A_i - D A_i, \end{aligned} \tag{31}$$

where B_i is the recruitment rate of individuals into activity group i. This model is an example of partitioning a heterogeneous population into risk classes, in the same manner as was done for age structure or multi-species models (section 3).

3.6.3. Mixing patterns. It is not just the number of partners an individual has in a given time period that determines how risky their behaviour is, the type of partner chosen plays a major role as well. For example, a single sex act with a low-risk partner is far less likely to lead to infection than a sex act with a partner that is high-risk. Therefore it is not simply the number of sexual partners that determines an individual's risk of being infected, but also the risk-status of these partners. This can be captured using the mixing matrix ρ, where ρ_{ij} denotes the proportion of contacts from individuals with i contacts to individuals with j contacts. There are a three constraints on this mixing matrix: all elements must be non-negative; each row must sum to one; and the number of contacts of group i with group j must equal the number of contacts of group j with group i.

$$\begin{gathered} \rho_{ij} \geqslant 0, \\ \sum_j \rho_{ij} = 1, \\ c_i (S_i + I_i) \rho_{ij} = c_j (S_j + I_j) \rho_{ji}. \end{gathered} \tag{32}$$

There are two obvious forms for the mixing matrix that satisfy these conditions, and can be readily parametrized from the type of individual-level data that is commonly collected. Firstly, proportionate or random mixing, where individuals form partnerships at random but in proportion to their expected number of partners (this can be compared with configuration-type models of network generation [192]):

$$\rho_{ij} = c_j \frac{S_j + I_j}{\sum_k c_k (S_k + I_k)}. \tag{33}$$

This type of mixing was used in the early models of STIs as it does not require detailed information about partnerships [14]. However, proportionate mixing does not agree with studies of sexual mixing patterns (where high-partner individuals preferentially pair with other high-partner individuals) and models that use this assumption have been found to predict epidemic growth rates that are inconsistent with epidemiological data [134].

The second type of mixing that clearly satisfies the mixing matrix conditions (32) is known as restricted mixing, where all partnerships are made within risk group. The mixing matrix in this case is just the identity matrix, $\rho_{ij} = \delta_{ij}$. In network theory terminology, restricted mixing is usually referred to as fully assortative mixing, where individuals only form contacts with others who have the same (or similar) defining characteristics [116].

It is more likely that the reality is somewhere between these two extreme mixing patterns and fortunately linear combinations of these two extremes also satisfy equation (32). This intermediate behaviour can be captured by preferred mixing [136], where a fraction ρ_i of a group i's contacts are reserved for restricted mixing and the remainder for proportionate mixing:

$$\rho_{ij} = \rho_i \delta_{ij} + \frac{c_j (1 - \rho_i)(S_j + I_j)}{\sum_k c_k (1 - \rho_k)(S_k + I_k)}. \tag{34}$$

This formulation has allowed the investigation of the effect of mixing patterns on the rate of spread of infection and the overall proportion of the groups that ultimately become infected. In general, assortative mixing gives multi peaked epidemics—an early small peak from the rapid spread in the small but high activity classes and then a later but extensive peak due to the slower spread over a longer time in the low activity classes (figure 21).

In the case of HIV, AIDS related mortality and behavioural changes can result in changes in the composition of the sexually active population. For example, if infection is concentrated in the high-risk groups then their proportional representation in the population is reduced when infected individuals are removed from the sexually active population. Therefore the conditions (32) for the mixing matrices are no longer satisfied and so the mixing patterns ρ_{ij}, rate of partner change c_i of the different groups, or both must be altered [116].

In the search for ever more reliable and accurate descriptions, there are many more complexities that have been included in the above models to account for observed HIV transmission mechanisms. In practice, models are only limited by available data and our willingness to include multiple complicating factors.

3.6.4. HIV in heterosexual communities. HIV in heterosexual communities can be thought of as a specific case of the multiple group model described in the previous section. In this case there are two groups, males and females, and the mixing between them is strictly disassortative (just like the vector-borne disease transmission in section 3.5.2), so that males only form partnerships with females and vice versa. The two

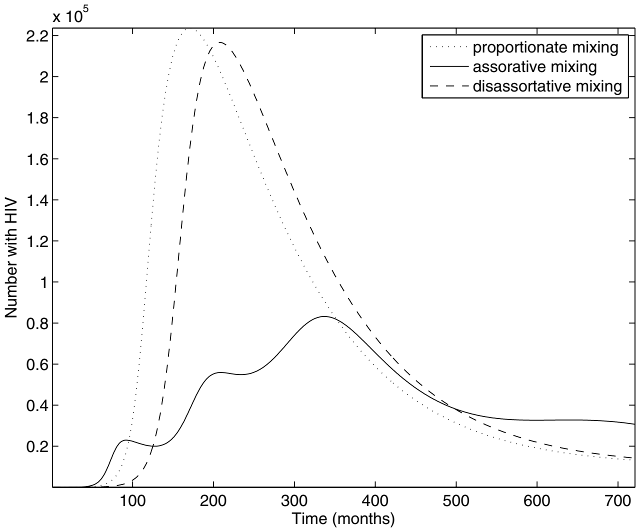

Figure 21. The change in total number infected in the population as predicted by equation (31) after introduction of one infected in the high-risk group, for random (equation (33)), assortative (equation (34) with $\rho_i = 0.99$) and disassortative mixing. The population structure can be represented in vector form with five risk groups; with the average number of contacts per year $c = (12, 24, 48, 96, 192)$ and the fraction of the population in each risk group $N = (0.24, 0.24, 0.43, 0.14, 0.06) = S(0)$ [136]. Other parameters are a birth rate $B = 100$ per month and $\eta = 0.2$ (years)$^{-1}$, $d_i = 0.012$ (years)$^{-1}\forall i$, $\beta_{ij} = 0.01 \forall ij$, $D = 1$ (years)$^{-1}$.

groups can be further stratified by age and sexual activity level and the relevant mixing parameters between groups can be approximated from empirical studies. Such models are required for developing countries where the majority of infection is in the heterosexual population; in such areas, it has long been realized that AIDS is capable of turning positive population growths negative [11, 12], driving the need for a rigorous understanding of HIV dynamics.

3.6.5. The impact of demography. The level of sexual activity is just one factor that can divide a population into different risk classes. It is also possible to stratify the population by age, type of sexual activity and intravenous drug use, amongst others, and to define the relevant mixing patterns between the groups, which often results in strong assortative mixing. When age is used in these models it is natural to allow individuals to age and change their behaviour as time progresses [1, 13], although such models require large amounts of data to parameterise this socio-demographic behaviour. Using mixing matrices involving age, gender and sexual activity level for a heterosexual population, it has been shown that the potential demographic impact of AIDS is enhanced by several known heterogeneities: a high level of sexual activity in the younger age classes compared to the older age classes; a male preference for younger female partners; and unequal transmission probabilities between males and females [10].

3.6.6. Variable infectivity. It has been found that a typical pathway for an individual infected with HIV consists of three stages. The initial short-duration stage, referred to as primary/acute infection, is characterized by a high viral load and high infectiousness. This is followed by an asymptomatic phase, where the viral load remains low and individuals are much less infectious. The final stage is the symptomatic stage including the onset of AIDS, where the viral load rises again. These three stages can be neatly captured by including more infectious compartments in the simple models (see section 2.4), with associated transmission rates and rates for moving between them parametrized to match observations [48]. As with the simple SIR model, the simple equations for HIV (29) show damped oscillations to equilibrium; however undamped oscillations in incidence may be generated if there is variable infectivity throughout the infectious period [232].

3.6.7. Variable incubation period. Studies have found that the incubation period from infection to the development of AIDS varies considerably between individuals, but does not fit the classic exponential distribution of standard models (section 2.1). This means that the rate of movement out of the infectious class is better captured as a function of time since infection, rather than a constant value. A reasonable function for this variable incubation period is $\eta(\tau) = k\tau$ [14], where τ denotes the time since an individual was infected and k is some constant. The infectious class is now given by $I(t, \tau)$, a function of both time t and duration of infection τ, and the system described by (29) becomes

$$\begin{aligned}\frac{\mathrm{d}S(t)}{\mathrm{d}t} &= B - dS(t) - \lambda(t)S(t),\\ \frac{\partial I(t,\tau)}{\partial t} + \frac{\partial I(t,\tau)}{\partial \tau} &= \lambda(t)S(t) - (\eta + d)I(t), \\ \frac{\mathrm{d}A(t)}{\mathrm{d}t} &= \eta I(t) - dA(t).\end{aligned} \tag{35}$$

The implications of this extended form of model are that the rise in cases of AIDS follows the rise in incidence of HIV infection but with a fairly clear delay. This is in contrast with the simple model (equation (27)) with constant incubation period, η, where there is a less marked lag between the rise in HIV cases and the rise in AIDS [14].

In conclusion models of HIV dynamics illustrate many elements of good mathematical modelling: a close integration of models, epidemiological data and clinical understanding; the use of models as a tool in better informing public-health; and the inclusion of multiple heterogeneities that may all play a key role in future dynamics. However, as with all attempts to model the real world, we are soon overwhelmed with the number of heterogeneities (age, gender, sexual activity, sexual preference, intravenous drug use, etc) many of which will be interdependent and interact. The skill of a modeller is therefore to assess and investigate which of these many factors are fundamental in addressing questions of epidemiological relevance. The history of modelling work in HIV is therefore the story of increasing sophistication as we both learned more about the infection and refined our modelling techniques.

4. Individuals and stochastic transmission

For models of infectious disease transmission the modelling scale can range from the truly microscopic, such as the within

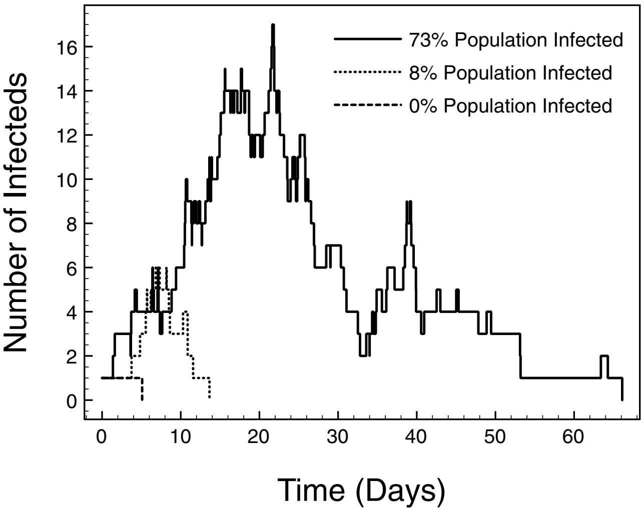

Figure 22. Three realizations of the stochastic SIR model initiated by a single infected (Recovery rate $\gamma = 0.2$ (days)$^{-1}$, reproductive number $R_0 = 2$, population size $N = 100$ plus the initial infected). Transmission and recovery occur at random times and lead to integer changes in the number of infecteds. Identical initial conditions can lead to very different outcomes, ranging from a large percentage of the population becoming infected at some point to a small percentage or even complete failure of the infection to invade the population.

host dynamics of disease [235], to the national and global scale [93]. The deterministic (ODE) epidemic models that formed the subject of the previous sections treated numbers of individuals experiencing each disease status as continuously differential quantities, which assumes an effectively infinite total population. This motivates formulating ODE models at the bulk population-level scale where the infinite population assumption is appropriate [44]. Whilst classical deterministic models for epidemic dynamics have a long and highly successful history in theoretical epidemiology, nonetheless formulating epidemic models at the scale of the discrete individual and taking into account the stochastic nature of transmission from individual to individual can capture phenomena that cannot be explained by a classical model. This is especially true when the number of infected hosts is small. A single infectious individual within a naive population has the potential to seed a large epidemic (if $R_0 > 1$) similar to that predicted by an ODE model. On the other hand by chance the infectious seed individual could fail to recruit secondary cases; the epidemic might fail to take hold despite $R_0 > 1$ (figure 22). This phenomenon cannot be explained using a deterministic epidemic model. The motivation behind stochastic models at the individual scale is clear, and indeed random models of disease date back to Daniel Bernoulli (c. 18th century). However, the numerical investigation of truly population-sized but individual scale epidemic models with stochastic transmission has only become feasible due to advances in electronic computational power.

The stochastic dynamics for compartmental epidemic models (e.g the SIR model) consist of individuals transitioning between disease states. In order to formulate a stochastic compartmental model the modeller has to specify the law in probability for the time and the type of each transition. In the literature many probabilistic mechanisms have been used to capture these transitions, including continuous time models [19, 20, 34, 36], discrete time models [147, 160] and generational time models such as the Reed-Frost chain binomial model [40] or the branching process epidemic model [111]. As an introduction to the characteristic dynamic properties of stochastic individual based models we will focus on modelling an outbreak of an infectious pathogen within the SIR modelling framework. The goals of this section will be to introduce the basic probabilistic properties of discrete event-based epidemic models, before describing in some detail methods for numerically simulating stochastic epidemics with the desired probabilistic structure, and the mathematical relationship between the stochastic epidemic model and related deterministic epidemic models. Special attention will be given to phenomena that cannot be adequately captured by continuous and deterministic models such as random early epidemic extinction, the emergence of persistent stochastically driven oscillations and occasional epidemic fade-outs.

4.1. Event-based stochastic epidemic models

Probabilistic compartmental epidemic models for individuals are usually introduced by considering the risk of disease transmission from currently infectious individuals to the currently susceptible ones, in addition to defining a probability law for the time spent within each disease compartment before recovery. Such transitions between compartments are examples of *stochastic events*. The set of possible stochastic events, $\mathcal{E}$, defines the possible transitions for epidemic states, such as the transmission of disease to each individual. This implicitly assumes that stochastic events such as transmission and recovery occur at a sufficiently quick time-scale as to be treated as instantaneous.

Although simpler approaches exist which allow the stochastic realization of some model types, we feel that it is important to be relatively precise in our definitions and approach both to allow sufficient generality and to better inform the link between stochastic and deterministic models. In order to simplify our discussion we will restrict attention to SIR-type epidemics amongst a homogenous populations. Population heterogeneity and alternative disease compartments can be introduced via population subdivision and the introduction of additional events with defined arrival rates. For the basic SIR model the epidemic state at time t is the vector,

$$X(t) = (S(t), I(t), R(t)), \tag{36}$$

where $S(t)$, $I(t)$ and $R(t)$ describe the integer number of susceptible, infected and removed individuals in the population class at time t. There are two possible events for the basic SIR model, transmission (e_I) and recovery (e_R), which are most conveniently described as vectors that model the impact of the event on the number of S, I and R individuals.

$$e_I = (-1, 1, 0), \tag{37}$$

$$e_R = (0, -1, 1). \tag{38}$$

The event-based stochastic epidemic model is an example of a *pure jump process* [158]. The advantage of treating the stochastic epidemic in this manner is that analysing and simulating the stochastic dynamics reduces to considering the random sequence of events and their time of arrival, $\{(e_k, T_k)\}_{k\geqslant 1}$, where the event-time pairs are ordered by their arrival time. At the arrival time for the kth event the epidemic state undergoes a transition

$$X(T_k) = X(T_k-) + e_k, \tag{39}$$

where $X(T_k-) = \lim_{t\uparrow T_k} X(t)$; that is the epidemic state instantaneously before the arrival time.

Stochastic models with many possible next events have competing waiting times; the next event that occurs is the one with the shortest waiting time. Denoting the random waiting time for the next event of type e, ΔT_e, (which is a random variable) we select the event with the shortest waiting time:

$$\Delta T = \min_{e\in\mathcal{E}}\{\Delta T_e\}. \tag{40}$$

with the next event type, e_{next}, being the one that has the minimum time:

$$e_{\text{next}} = \arg\min_{e\in\mathcal{E}}\{\Delta T_e\}. \tag{41}$$

The distribution of event waiting times is determined by the choice of model, based on any available data.

4.1.1. Arrival rates and the basic stochastic SIR model. The simplest probabilistic model for event arrivals is to postulate that each event occurs at some probabilistic rate $r_e(t)$, $e \in \mathcal{E}$, which may be time inhomogeneous and may depend on the population size, the current state of the system or even the history of the epidemic. A popular assumption is that the event arrival rates depend only on the current state of the epidemic $(X(t))$, this describes a *Markovian* jump process. Event rates represent the stochastic analogue to the rates of change familiar from deterministic models and hence can similarly be motivated by real epidemiological and demographic data. For special cases where the arrival rates are either deterministic or Markovian we will show that the distribution for waiting time until next event and the next event type can be given in a particularly simple form.

For any given event type e, the probability of the event occurring in the near future is proportional to the rate event arrival process, more precisely,

$$\mathbb{P}(\Delta e[t, t+h] = 1) = r_e(t)h + o(h), \qquad h > 0. \tag{42}$$

The 'work horse' of random models for the SIR process is the stochastic construction where each individual upon being recruited to the infected class remains infectious for the random length of time T, called the infectious duration [63]. During that time, the infectious individual makes random infectious contacts with the rest of the population, the standard model being contact at the points of a Poisson process [29]. When a susceptible is contacted that individual is infected and recruited to the infectious class. The most popular choice is to have the infectious durations exponentially distributed since the stochastic epidemic construction with Poisson process contacts and exponential durations can be described as a Markov process [84] with event arrivals governed by the rate of individual infectious contacts (β/N) and the rate parameter of the exponential duration time (γ). Since, individuals are assumed interchangeable within a homogeneous population, the arrival rate for *any* infection event (r_I) or recovery event (r_R) are,

$$r_I(t) = r_I(X(t)) = \frac{\beta S(t)I(t)}{N}, \tag{43}$$

$$r_R(t) = r_R(X(t)) = \gamma I(t). \tag{44}$$

The probabilistic dynamics for the *basic stochastic SIR model* [19] are therefore,

$$\mathbb{P}[(S, I)(t+h) - (S, I)(t) = (-1, 1) \mid (S, I)(t)]$$
$$= \beta\frac{S(t)}{N}I(t)h + o(h), \tag{45}$$
$$\mathbb{P}[(S, I)(t+h) - (S, I)(t) = (0, -1) \mid (S, I)(t)]$$
$$= \gamma I(t)h + o(h). \tag{46}$$

We do not explicitly include the number of removed individual $R(t)$, as it is a redundant variable when the population is fixed at size N.

4.1.2. Deterministic and Markovian arrival rates. The special case where the arrival rate of an event $r_e(t)$ is *deterministic* (independent of the stochastic dynamics) is important, both as a model for events occurring due to an external influence upon the epidemic and because of its analytical convenience. Any arrival process with a deterministic rate is a necessarily a *Poisson process* [158]. The number of new event arrivals for the Poisson process on an interval, $\Delta e[t_1, t_2]$, is independent for disjoint intervals and Poisson distributed,

$$\Delta e[t_1, t_2] \sim \text{Poisson}\left(\int_{t_1}^{t_2} r_e(s)\,\mathrm{d}s\right), \qquad t_1 \leqslant t_2. \tag{47}$$

Additionally, for *constant* rate arrivals the waiting time, ΔT_e, until the next event of a Poisson process is distributed exponentially with rate parameter given by the process rate [67],

$$\Delta T_e \sim \exp(r_e). \tag{48}$$

so that ΔT_e is drawn from an exponential distribution. For the basic stochastic SIR model the Markovian arrival rates given by (43) and (44) are piecewise constant between the arrival of new epidemic events, which implies that the waiting times for both the next transmission and the next recovery events are exponentially distributed. For the basic SIR model the properties of sets of exponential random variables applied to (40) and (41) giving the distributions of ΔT and e_{next} in simple form for any state $X(t)$,

$$\Delta T \sim \exp(r_I(t) + r_R(t)), \tag{49}$$

$$\mathbb{P}(e_{\text{next}} = e_I) = \frac{r_I(t)}{r_I(t) + r_R(t)}, \tag{50}$$

$$\mathbb{P}(e_{\text{next}} = e_R) = \frac{r_R(t)}{r_I(t) + r_R(t)}. \tag{51}$$

This can be straightforwardly generalized to time homogenous Markovian epidemics with more compartments and more event types.

4.1.3. Non-Markovian arrival times. For non-Markovian and time inhomogeneous arrival rates matters are more complicated, however a result due to Lipster and Shiryaev [158, 172] is useful for characterizing the waiting time between events. The idea is to make a change of time transformation $t \to \rho(t)$ so that in the new time variable $\rho(t)$ the number of new events $N_e(t) = \Delta \boldsymbol{e}[0, \rho(t)]$ arrive as a Poisson process with constant rate 1. This is achieved by relating the transformed time variable to t as,

$$\rho(t) = \inf_{\tau \geqslant 0} \left\{ \tau \,\middle|\, \int_0^\tau r_e(s)\, \mathrm{d}s = t \right\}. \tag{52}$$

Equation (52) describes an analytic trick; the transformed time progresses slower when the arrival rate is faster and vice versa in order to treat events as arriving at a uniform rate.

The first arrival time of a unit rate Poisson process is distributed $Z \sim \exp(1)$ and therefore the first arrival time for the event $\boldsymbol{e}$, T_e, is the solution to,

$$\int_0^{T_e} r_e(s)\, \mathrm{d}s = Z. \tag{53}$$

The first arrival time is particularly important in the context of the SIR epidemic, since an infection event arrives for each individual at most once. Equation (53), applied to the arrival of infection for an individual, was first described by Sellke [225] for the purpose of analysing the final size distribution of cases. In this context the set of events must distinguish between individuals, for example the infection event of each individual is considered distinct. The random variables Z then have an intuitive interpretation as a random 'resistance' for each individual which is eroded at a rate $r_I(t)$. Infection occurs for the individual at the time point where the resistance is eroded to zero. The final size analysis has been deepened by Ball for the single-population scenario [24] and also extended to include multi-type epidemics [25, 30]. For a thorough discussion of numerical methods, and potential pitfalls, for solving the resultant final size distributions see House *et al* [128]. The major drawback to using result (53) is that ΔT and $\boldsymbol{e}_{\text{next}}$ are represented implicitly in terms of a set of exp(1) distributed 'latent' random variables [64] rather than given directly. The ease of use of the implicit form will depend on the model specifics.

4.1.4. The Kolmogorov equation. The above description of stochastic rates and events naturally leads to the use of simulation methods which need to be performed multiple times to gain an understanding of the range of dynamics. However, it is possible to generate exact models for the probability of finding an epidemic in a given state. Consider the dynamics of the indicator function, $\mathbf{1}_x(\cdot)$, which is one if the epidemic is in some state x. The expectation of this state indicator function gives the time-dependent *probability* of finding the epidemic in that configuration; we denote this probability as $p(x, t)$,

$$\mathbb{E}[\mathbf{1}_x(X(t))] = \mathbb{P}(X(t) = x) = p(x, t). \tag{54}$$

For Markov processes the current epidemic state completely specifies the event arrival rates. Considering the rate at which $X(t)$ enters and exits the state x gives the complete set time-dependent state probabilities as the solution to a set of ODEs,

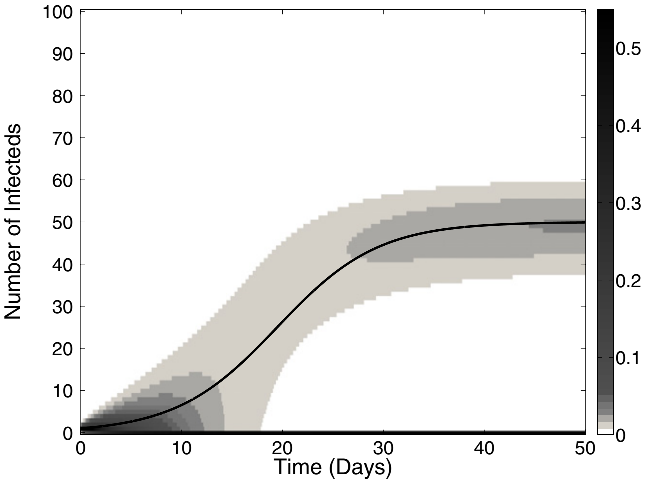

Figure 23. An example of the direct solution to the Kolmogorov equation using MATLAB exponential matrix solver. For simplicity an SIS epidemic is considered ($N = 100$, $\gamma = 0.2$ (days)$^{-1}$, $R_0 = 2$). Shading gives the time-dependent probability distribution of infecteds numbers. The solid curve gives the prediction of the related deterministic SIS model. Note that the infecteds probability distribution is bimodal between the endemic and disease-free states (probability of disease-free after 50 days is 0.5107).

$$\frac{\mathrm{d}p(x,t)}{\mathrm{d}t} = \sum_{e \in \mathcal{E}} [r_e(x - \boldsymbol{e}) p(x - \boldsymbol{e}, t) - r_e(x) p(x, t)]. \tag{55}$$

The equations (55) are variously called the *Kolmogorov*, *Chapman–Kolmogorov* or *master equations* for the probability of the stochastic epidemic being in state $x \in \mathcal{S}$ at time t given some initial distribution $p(x, 0)$ for each possible epidemic state x. The full set of probabilities for each epidemic state can be represented as a vector $\boldsymbol{p}(t)$. The advantage of vector notation is that the large set of equations (55) can be given compactly as a linear evolution equation,

$$\dot{\boldsymbol{p}}(t) = \boldsymbol{G}\boldsymbol{p}(t). \tag{56}$$

Where the matrix $\boldsymbol{G}$ is called the *generator* of the stochastic epidemic and encodes the relevant transition rates of stochastic infection and recovery events. The master equation (56) can be solved in terms of a matrix exponential,

$$\boldsymbol{p}(t) = \exp(\boldsymbol{G}t)\boldsymbol{p}(0), \qquad t \geqslant 0. \tag{57}$$

The numerical computation of matrix exponentials can be performed by a number of software packages, including MATLAB, see figure 23 for an example of the solution for an SIS-type epidemic. Where accessible the solution (57) can be used to investigate the stochastic behaviour of epidemics directly [146].

4.2. Simulating event-based stochastic epidemics

Although simple in formulation, equation (56) cannot be used in general due to the explosion in possible disease configurations as the population size and number of different population sub-groups becomes large; a phenomenon known as the *curse of dimensionality*. It is therefore convenient to resort to Monte Carlo techniques for large populations; that is to simulate realizations of the stochastic epidemic with identical probabilistic structure as the 'true' epidemic. Problems of interest can then be investigated via the Monte Carlo convergence of independently drawn random samples.

In this section we investigate techniques for numerically constructing sample epidemics. The first algorithm considered will be the *first reaction method* (FR) which focuses on sequentially solving for $(e_{\text{next}}, \Delta T)$ by directly drawing the ensemble of waiting times for each event $\{\Delta T_e\}_{e\in\mathcal{E}}$. The simulation progresses by implementing the soonest next event and time-stepping forward to its arrival time before generating the following next event. A more popular algorithm for sampling Markov Jump process realizations, Gillespie's direct method (GD) [103], is more efficient than the FR method but is only applicable when the arrival rates are time homogeneous and Markovian. Both algorithms are exact to machine precision, in the sense that any random event associated with stochastic epidemic model occurs with identical probability amongst the epidemic realizations sampled using the numerical scheme.

Both the GD and FR methods have a variable simulation time-step due to the stochastic variation in waiting times for epidemic events. For large-population sizes the arrival rate of any new event becomes fast, and so the waiting time between events becomes typically very short. Therefore the computational 'cost' of epidemic simulation over a fixed time horizon, or until a disease-free state is reached, typically increases significantly with the size of the population due to many more time-steps being required. A more computationally efficient, but approximate, epidemic simulation alternative for large-population sizes is the *Tau-leap method* (TL), also described by Gillespie [104], which can be thought of as a stochastic equivalent to the basic Euler time-stepping method for numerically solving ODEs. The main point of difference for the TL algorithm, compared to the GD and FR algorithms, is that the simulation time-step, or 'tau-leap', is a fixed constant τ. By assuming that the arrival rates for events are constant over the time-step multiple events can be generated as occurring at some time within the time-step. When τ is greater than the typical time-step of the exact algorithms the TL method greatly improves the speed at which an epidemic can be simulated, see [151] for a direct comparison between the numerical efficiency of the TL, FR and GD methods. For a schematic representation of the points of difference between the exact simulation methods and the Tau-leap method see figure 24.

4.2.1. Methods for generating waiting times. As we have seen for (piecewise) constant rates the waiting times are exponentially distributed, however this is not true generally. The most efficient method for generating each T_e can potentially differ from model to model, and indeed from event type to event type.

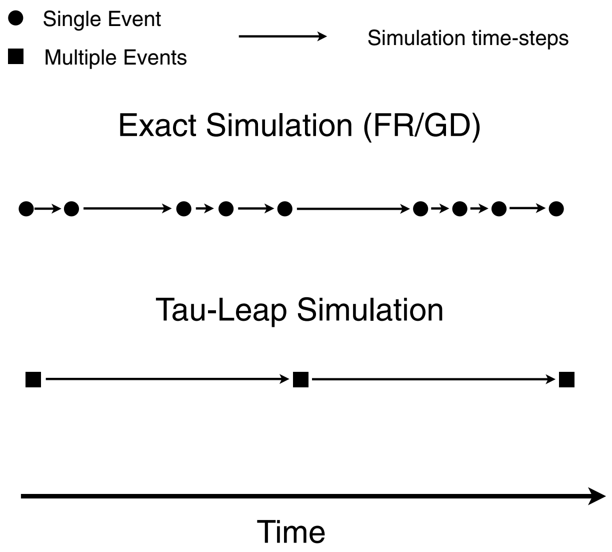

Figure 24. Schematic of exact stochastic simulation versus the approximate Tau-leap method. For the exact simulation methods, first reaction (FR) and Gillespie direct (GD), the time-steps are generated by the probabilistic arrival of single events according the epidemic model. FR simulation is flexible since non-Markovian and time inhomogeneous arrival rates are acceptable within the simulation framework. GD is significantly more computationally efficient than FR, but applies only to Markovian stochastic models. For tau-leap simulation time increases by a fixed amount τ at each time-step, multiple events of each type occur during each 'Tau-leap' and are realized at the end of the time-step according to a Poisson distribution. Here τ is greater than the typical time-step of the exact simulation and so TL is the most efficient method considered here, and can include history dependence, but at the cost of exactness.

A common variation from the basic SIR epidemic model is to consider non-exponentially distributed infectious periods [29], in this case the arrival time of the recovery event for an individual infected at time t_I is $T_R = t_I + T$ where T is distributed according to the desired infectious period. At time t, the waiting time until the recovery event is then,

$$\Delta T_R = T_R - t, \qquad t \leqslant T_R. \tag{58}$$

Other possibilities for calculating this waiting time are to use the Sellke-type construction (53) directly or to consider a thinned Poisson process for generating arrival events. It is worth noting that both the epidemic resistance and infectious duration are random variables that are drawn from their respective distributions independently of the overall epidemic dynamics and can therefore be pre-generated before simulation which can be advantageous for the retrospective comparison of epidemics [64] or for investigating epidemic spread analytically [24]. On the other hand they must be generated for each individual which effectively leads to considering each population model as a collection of N population sub-groups each of size 1.

4.2.2. Simulation algorithms. The complexity of the FR method is in the drawing of $\{\Delta T_e\}_{e\in\mathcal{E}}$, whether using pre-generated random variables or any other method, algorithmically it is comparatively simple.

First reaction algorithm.

```
1. Generate the initial epidemic state $X(0)$ from $\mathbb{P}_0$ and set time $t = 0$.
2. For each possible event $\boldsymbol{e} \in \mathcal{E}$ generate the waiting time $\Delta T_e$.
3. Update the current state by implementing the selected event, $X(t) \to X(t) + \boldsymbol{e}_{\text{next}}$.
4. Set the new time $t \to t + \Delta T_{e_{\text{next}}}$.
5. If end conditions have not been met return to 2, else stop.
```

The Gillespie direct algorithm is appropriate for simulating event arrivals with time homogeneous Markovian rates, which have exponentially distributed waiting times, and is justified directly by the properties of the minimum of a set of exponential random variables. This is an algorithmic description, for the basic stochastic SIR model the state dependent rates are given by (43) and (44).

Gillespie direct algorithm.

```
1. Generate the initial epidemic state $X(0)$ from $\mathbb{P}_0$ and set time $t = 0$.
2. For each possible event $\boldsymbol{e} \in \mathcal{E}$ calculate the rate $r_e(X(t))$ for the current state and the total exit rate $R = \sum_e r_e(X(t))$.
3. Randomly select the next event from the set of possible events. The event $\boldsymbol{e}_{\text{next}}$ being chosen with probability $r_{e_{\text{next}}}(X(t))/R$.
4. Update the current state by implementing the selected event, $X(t) \to X(t) + \boldsymbol{e}_{\text{next}}$.
5. Draw time increment $\Delta T \sim \exp(R)$ and set the new time $t \to t + \Delta T$. If *rand* is a random number between 0 and 1, then the waiting time can be computed as $\Delta T_e = -\ln(\text{rand})/R$.
6. If end conditions have not been met return to 2, else stop.
```

Both the FR method and the GD method are exact, however for large-population sizes they become numerically inefficient due the very great numbers of random events that occur in any fixed time interval.

The tau-leap algorithm [104, 105] approximates the number of random events that arrive in the short time interval $[t, t + \tau]$ by assuming the arrival rates are constant over that interval. As we have seen this implies that the number of event arrivals are Poisson distributed,

$$\Delta e[t, t + \tau] \sim \text{Poisson}(r_e(t)\tau). \tag{59}$$

Tau-leap algorithm.

```
1. Generate the initial epidemic state $X(0)$ from $\mathbb{P}_0$ and set time $t = 0$ and choose a time-step parameter $\tau$.
2. For each possible event $\boldsymbol{e} \in \mathcal{E}$ calculate the rate $r_e(t)$.
3. Draw the stochastic changes, $\Delta\boldsymbol{e}[t, t + \tau] \sim \text{Poisson}(r_e(t)\tau)$ for each $\boldsymbol{e} \in \mathcal{E}$.
4. Set $X(t) \to X(t) + \sum_{e\in\mathcal{E}}(\Delta e[t, t + \tau])e$.
5. Enforce population constraints as appropriate, i.e. if $S_i(t) < 0$, then set $S_i(t) = 0$.
6. Increase time $t \to t + \tau$.
7. If end conditions have not been met return to 2, else stop.
```

4.3. The large-population limit

The deterministic models presented in the previous sections have a long and successful history in predictive modelling of epidemic dynamics. Yet in many respects defining stochastic dynamics occurring at the level of discrete events seems a more natural starting point for a model of infectious disease dynamics spreading between individuals. In this section we make explicit the essential connection between models with intrinsic randomness and the deterministic models used in the majority of this review. We will demonstrate that deterministic ODE-type dynamics can be thought of as a type of central limit of the Markovian stochastic epidemic as the population size of the system N becomes large. This is analogous to classical results from statistical physics and chemistry such as the principle of mass action or the thermodynamic limit for physical systems [45]. The implication of this limiting result is that once a disease is established in a large population then predictions using a deterministic model will be very similar to those of a stochastic model.

Kurtz established conditions for Markovian dynamical process (say $X(t)$) with a state space of natural numbers with some size parameter N, to convergence to a deterministic processes in *density* or scaled form $x(t) = X(t)/N$ as the size parameter becomes large ($N \to \infty$). This is a variety of central limit result for Markov processes [164, 165], which has now been further generalized (cf [139] theorem 17.15). For the stochastic epidemics considered here the criterion for convergence are that for each event $\boldsymbol{e} \in \mathcal{E}$ we can find an $\mathcal{O}(1)$ function f_e, called the *parameter function* for $\boldsymbol{e}$ such that,

$$r_e(X(t)) = Nf_e(x(t)), \qquad \boldsymbol{e} \in \mathcal{E}. \tag{60}$$

This is effectively a criterion that none of the event occurrence rates for the stochastic process are faster than $\mathcal{O}(N)$, however identifying the appropriate parameter function is also important for assessing the appropriate limiting ODE for a given stochastic model. If (60) is satisfied for some set of parameter functions then, as $N \to \infty$, the deterministic dynamics for the scaled variables $x(t)$ is,

$$\dot{x}(t) = \sum_{e\in\mathcal{E}} \boldsymbol{e} f_e(x(t)), \tag{61}$$

$$x(0) = \lim_{N\to\infty} \frac{X(0)}{N}. \tag{62}$$

The type of convergence is uniform on compacts in probability. This is a particularly strong statement; it is not simply that the expectation of the stochastic model converges to the deterministic limiting dynamics, but that nearly every realization of the process converges in density (figure 25).

When N is large but stochastic dynamics are still considered important, it is often most convenient to use a diffusion process approximation. The stochastic density process $x(t)$ with parameter functions $\{f_e\}_{e\in\mathcal{E}}$ is very 'close' [165] to the diffusion process, written as an Itô form stochastic differential equation [141],

$$\mathrm{d}x(t) = \sum_{e\in\mathcal{E}} \boldsymbol{e} f_e(x(t))\,\mathrm{d}t + \sum_{e\in\mathcal{E}} \boldsymbol{e}\sqrt{\frac{f_e(x(t))}{N}}\,\mathrm{d}B_e(t), \tag{63}$$

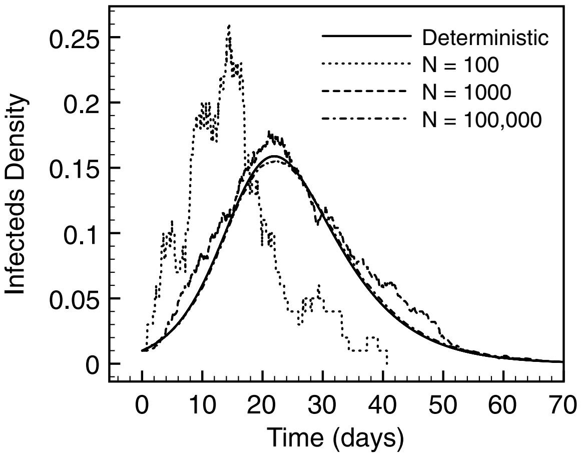

Figure 25. A comparison of the dynamics of infecteds population density between three realizations of the basic SIR epidemic model ($\beta = 2, \gamma = 1$) and the deterministic SIR model. As the population size N grows convergence of the stochastic model onto the deterministic model is evident.

where $\{B_e(\cdot)\}_{e\in\mathcal{E}}$ are a set of independent standard Brownian motions, one for each event type.

Hence, for the basic stochastic SIR model the diffusion approximation can be expressed as

$$\mathrm{d}s(t) = -\beta s(t)i(t)\,\mathrm{d}t - \sqrt{\frac{\beta s(t)i(t)}{N}}\,\mathrm{d}B_I(t), \tag{64}$$

$$\mathrm{d}i(t) = [\beta s(t)i(t) - \gamma i(t)]\,\mathrm{d}t + \sqrt{\frac{\beta s(t)i(t)}{N}}\,\mathrm{d}B_I(t) - \sqrt{\frac{\gamma i(t)}{N}}\,\mathrm{d}B_R(t), \tag{65}$$

where $s(t) = S(t)/N$, $i(i) = I(t)/N$, and stochasticity is captured by the independent standard Brownian motions $B_I(\cdot)$ and $B_R(\cdot)$. The details of Kurtz's approach to assessing the deterministic limit of stochastic processes heavily involves the use of Martingale theory, for an alternative approach based on taylor expansions of the Chapman–Kolmogorov equation for the epidemic process see Van Kampen [245] or Diekmann, Heesterbeek and Britton [74]. Note that while the Markovian event-driven model leads to the Kolmogorov Equations for the probabilistic behaviour of the complete system (see section 4.1.4), so the diffusion approximation model leads to Fokker–Plank equations [243].

4.4. Early extinction and infectious generations

A major departure for the theory of the stochastic epidemics compared to their deterministic counterparts is the threshold phenomenon for successful invasion must be reinterpreted as the probability of a 'large outbreak'. For stochastic models there are fluctuations in the number of infecteds, and since any disease-free epidemic state is a trapping point for the dynamics (in the absence of an external reintroduction mechanism), there will be a chance that the introduction of a small number of infecteds will fail to cause a significant sized epidemic. This is not a feature of the deterministic SIR model, where the numbers of infected individuals always increases when $R_0 > N/S(t)$. This risk of early extinction leads to a bimodal or *U-shape* [135] distribution of epidemic final sizes for the stochastic SIR model. In this section we concentrate on calculating the probability of a small epidemic occurring despite a small number of infected individuals being introduced to a large naive population; that is outbreaks that contribute to left-hand of the 'U' (figure 26).

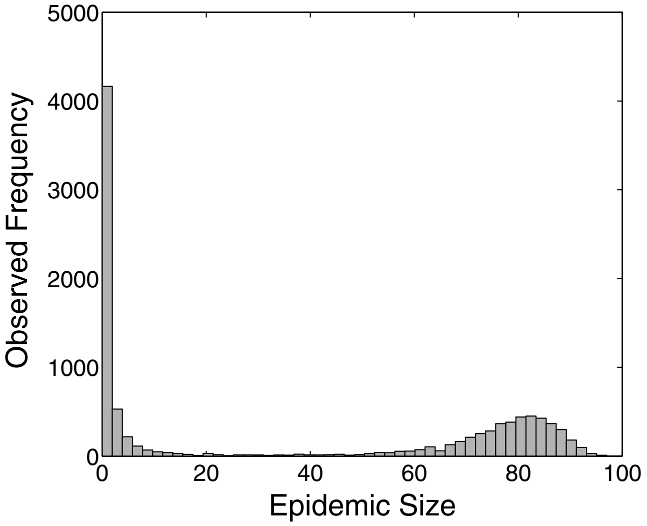

Figure 26. The characteristic U-shape of epidemic size distribution (from 100 000 replicate simulations) for the basic stochastic SIR model ($N = 100$, $R_0 = 2$) with a single initial infected. In the large N limit with generational dynamics branching process theory gives the probability of an epidemic being a member of the 'left side' of the U-shaped distribution.

4.4.1. Branching processes and early extinction.

For populations of finite size N without the possibility of disease reintroduction from some external source the epidemic ceases in finite time as it eventually reaches a state where no infected individuals remain. For SIR-type epidemics each individual that was infected at some point during the epidemic outbreak is called a *case*. If $Z_N(\infty)$ denotes the final number of cases for a population of size N then the probability of a 'true epidemic' or 'major epidemic' [23] is said to be,

$$\mathbb{P}(\lim_{N\to\infty} Z_N(\infty) = \infty). \tag{66}$$

This is best interpreted as a probability that a non-zero fraction of a large population becoming infected (i.e. a significant epidemic occurring) with the other option being that an effectively zero fraction of a large population becomes infected, which is called an *early extinction* of the epidemic.

In this setting the $N \to \infty$ limit is analytically convenient as the effects of susceptible depletion are ignorable during the early epidemic, such that there is no competition between infected individuals and hence all chains of transmission can be treated independently. Early in the epidemic, each infected individual recruits new cases from the total susceptible population according to a Poisson process with rate β; the probability of contacting an individual who has already been infected asymptoting to zero. Infected individuals continue recruiting over the duration of their infectiousness, which lasts a random period T. The new cases generated by an infected host are called the *offspring*, O, of the infected, and the distribution of the number of offspring for each infected is called the *offspring distribution*, denoted

$$P_O(k) = \mathbb{P}(O = k). \tag{67}$$

Since time is no longer an explicit factor; the epidemic dynamics are given as the number of infected in each *generation*. This type of process is called a *branching process* [120]. In the large N limit the offspring distribution of each infected become independent, and identically distributed, therefore it is sufficient to consider epidemics seeded by a single infected ($I_0 = 1$). Early extinction probabilities for $I_0 > 1$ can be calculated using the multiplication of independent probabilities.

The probability of early extinction can be analysed using the probability generating function (PGF) of the random number of offspring O ($G_O(z)$) defined as,

$$G_O(z) = \sum_k \mathbb{P}(O = k) z^k, \qquad z \in [0, 1]. \tag{68}$$

A classical result from branching process theory is that the probability of early extinction, P_{ext}, is given by the following intercept condition [120],

$$P_{\text{ext}} = \min\{z \in [0, 1] | G_O(z) = z\}. \tag{69}$$

This methodology allows us to compute extinction probabilities for a wide spectrum of model formulations.

For the basic Markov SIR model with infectious durations that are exponentially distributed ($T \sim \exp(\gamma)$), a more direct argument can be applied. Let p be the probability that the next event for any given infectious individual is to recruit (with rate β) rather than recover (with rate γ); this is given by considering competing waiting times,

$$p = \mathbb{P}(T_I < T_R) = \frac{\beta}{\beta + \gamma}. \tag{70}$$

Therefore, considering a single infected case, either the individual recovers, there are no cases and hence the infection goes extinct with probability 1, or the individual generates a new case and we need to consider the extinction probability given two infected individuals. Given the independence of the chains of transmission from these two infected hosts, the probability of total extinction is simply the product of each chain going extinct. Hence

$$P_{\text{ext}} = (1 - p) \times 1 + p \times P_{\text{ext}}^2 \tag{71}$$

Two solutions to this equation exist, $P_{\text{ext}} = 1$ or $P_{\text{ext}} = \gamma/\beta = 1/R_0$, with the latter being relevant when $R_0 > 1$. Numerical simulations of the basic stochastic SIR model reveal that this theoretical value is generally an over-estimate for finite N (as offspring are no longer independent and compete for susceptibles) but converges as N becomes large (figure 27).

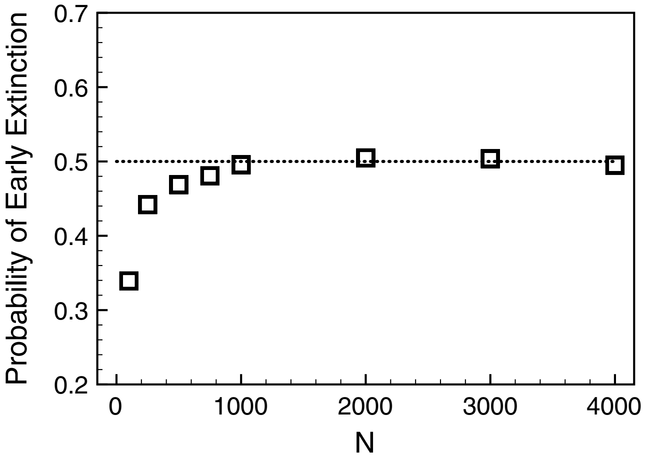

Figure 27. Simulated realizations of the basic stochastic SIR model ($\beta = 2,\ \gamma = 1$, Gillespie direct simulation method) with an epidemic seed of a single individual $I_0 = 1$ and increasing population size N, an early extinction for the finite N simulations is defined as a final number of cases $Z_N < 0.01N$. Branching process theory predicts an early extinction probability of 0.5 (dashed line), this is approached in the large N limit so that for $N > 1000$ the theoretical and numerical results become indistinguishable.

4.4.2. Early extinction for multi-compartmental models.

The choice of distribution for the infectious duration T has an important, but complex, effect on the dynamics of stochastic epidemics [62]. Since the assumption of exponential distributed T is often not supported by epidemiological data (e.g. [208]), other distributions need to be considered, with the use of multiple infected compartments being the most common (see section 2.4). Such models are commonly referred to as *multi-compartment* models. Here we will consider generalizing the single infectious duration to M sub-period durations, $\{T_i\}_{i=1}^M$ so that,

$$T = \sum_{i=1}^{M} T_i. \tag{72}$$

This is most compactly described as the $SI_{(1)} \ldots I_{(M)}R$ compartmental epidemic model.

As an illustration of the usefulness of the result (69) we consider a stochastic model with Poisson process transmission events and T given by (72) where the multi-stage infectious periods are distributed identically and independently, $T_i \sim \exp(M\gamma)$. In this case the infectious duration T is distributed according to an *Erlang* distribution; a special case of the gamma distribution [174], with exponentially distributed infectious durations ($M = 1$) and constant infectious durations ($M = \infty$) treated as special cases. Moreover, since the expected duration within each of the M infectious compartment is $1/M\gamma$ the value of R_0 is invariant to the choice of M.

The total number of offspring, O, is the sum of offspring in each stage, therefore the M-stage offspring distribution has the PGF:

$$G_O(z; M) = \left(\frac{p_M}{1 - (1 - p_M)z}\right)^M, \quad p_M = \frac{M\gamma}{\beta + M\gamma} \tag{73}$$

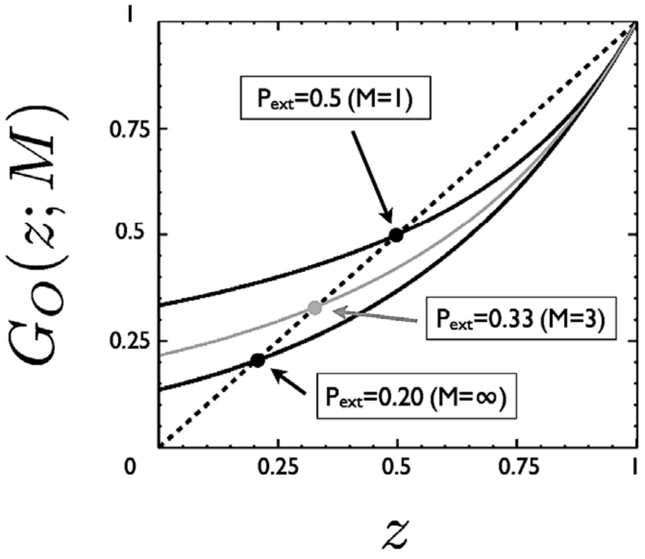

Figure 28. The graphical construction for the intercept criterion (69). Solid lines give the offspring PGFs for the stochastic epidemic with M infectious compartments ranging from geometric ($M = 1$) though negative binomial ($M = 3$) to Poisson ($M = \infty$). The minimum intercept (filled circles) with the unit gradient line gives the early extinction probability, P_{ext}. Graphical considerations and the convexity of G_O give that $P_{\text{ext}} < 1 \iff G'_O(1; n) > 1$.

The threshold for the probability of a significant outbreak being non-zero is $R_0 = 1$ which is independent of the choice of M. If $R_0 > 1$ the early extinction probability is strongly affected by the choice of infectious duration distribution with early extinction more likely in models with fewer compartments (figure 28).

4.5. The interplay between stochasticity and demography

There are two predominant mechanisms whereby SIR-type diseases can establish themselves for long durations in a population and avoid epidemic cessation due to the depletion of potential new hosts. Firstly, by recruiting new individuals born into the susceptible class and secondly, by external infectious pressure or imports of new infectious individuals. For deterministic epidemic models with demography (births and deaths) the persistence of an epidemic is characterized by the existence of endemic stable equilibrium points for the disease dynamics. For stochastic models this picture is complicated by both fluctuations about any potential endemic equilibrium and the realization that for (ergodic) stochastic models a disease-free trapping state will be reached eventually, albeit potentially after a very long duration. In this section we will discuss, with reference to the basic stochastic SIR model with demography, two phenomena that have been observed in real epidemic time series data: stochastic resonance and stochastic fade-out or repeated extinction-decolonization events.

To include demography for the basic SIR model we introduce a birth event e_{B} and death events for the susceptible, infected and removed individuals, $e_{\text{DS}}, e_{\text{DI}}, e_{\text{DR}}$. Following equation (1) in section 2.1 we assume new birth events arrive at constant rate BN^*, whereas the rate of death events are, respectively $\mathrm{d}S$, $\mathrm{d}I$ and $\mathrm{d}R$. By letting $B = d$ the initial population size $N(0) = S(0) + I(0) + R(0)$ converges rapidly to a fluctuating process about its equilibrium size, N^*.

$$\mathbb{E}[N(t)] = N^* + (N(0) - N^*)\mathrm{e}^{-Bt}.$$

With the simplest solution being to set the initial population size to N^*.

For modelling external imports of infection one possibility is to include a migration rate of infectious individuals from outside the population into the infected sub-group. However here we prefer to model external imports as due to some interaction (for example due to commuting between populations) between susceptibles and an external reservoir of infection. This acts as an additional term in the arrival rates of infection events through an import rate factor ξ,

$$r_I(t) = \frac{\beta}{N}S(t)I(t) \to \frac{S(t)}{N}\left(\beta I(t) + \xi\sqrt{N}\right). \tag{74}$$

Here we have assumed that the import rate per individual scales with the square-root of the population size ($\sqrt{N}$), which appears to be a generic feature of human diseases [35]. In the following examples we will use event-based stochastic models for the epidemic dynamics. However analysis will be performed using the diffusion approximation described in section 4.3. This approach gives good results due to the large-population sizes that will be considered.

A limitation of treating birth and deaths independently is that the state space for the epidemic technically becomes infinite due there being a possibility (albeit vanishingly small) of finding any population size. This inhibits an analysis based on solving the Kolmogorov matrix equation (56) directly such as in [146]. Solutions to the infinite state space problem in the literature include using a large N truncation [60] and constraining demographic effects to be 'one in, one out' and thereby fixing N [5].

4.5.1. Stochastic amplification. For deterministic dynamics the effect of perturbations can be analysed via linearization around the endemic equilibrium. For the SIR model with demography this reveals complex valued eigenvalues indicating that the approach to equilibrium will be oscillatory with a decaying amplitude governed by the real parts of the eigenvalues and frequency governed by the imaginary parts.

As we have seen in section 4.3 the stochastic dynamics of the epidemic for large N are well approximated by a diffusion process. The Brownian motion drivers of the diffusion dynamics continuously excite perturbations at each frequency and therefore the potential exists for resonant mode interactions between the Brownian noise source and the natural frequency of the approach to equilibrium predicted from the stability analysis. This causes sustained oscillations with a typical dominant frequency that are not predicted by related deterministic models (figure 29). In un-vaccinated populations, persistence of regular and irregular oscillatory dynamics have been observed for many childhood diseases [42, 109, 210]. Classically, oscillatory dynamics have been understood through the paradigm of deterministic dynamics

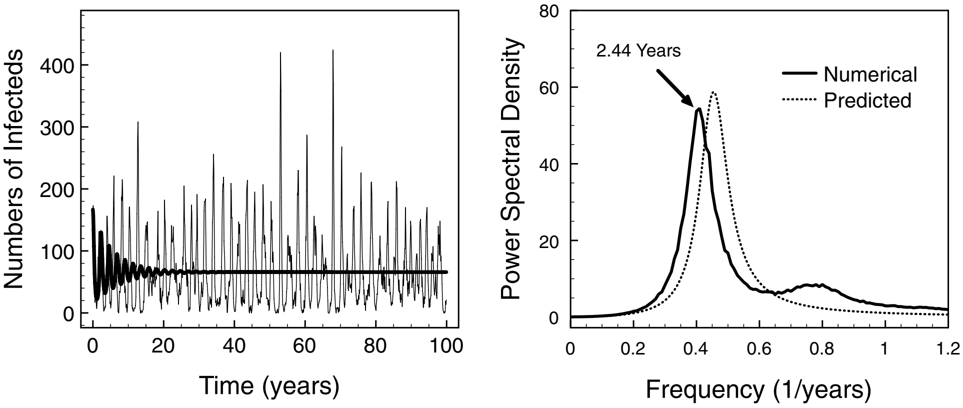

Figure 29. An SIR epidemic with demography amongst a population of equilibrium size $N^* = 10^5$ ($\beta = 1.175$, $\gamma = 0.077$, $B = d = 5.5 \times 10^{-5}$, $\xi = 3.16 \times 10^{-3}$ all rates given in (days)$^{-1}$). Left: a realization of the epidemic simulated (Tau-leap method $\tau = 0.05$ days) over a 100 year time window. Thick solid curve is the solution to the related deterministic model over the same period. Note that the stochastic realization sustains substantial oscillations despite the absence of temporal forcing whereas oscillations for the deterministic model are damped over time. Right: power spectral density spectrum for the numbers of infecteds with empirical mean removed $(I(t) - \bar{I})$ averaged over 1000 independent realizations. The spectrum reveals that the fluctuations about mean infecteds burden are dominated by persistent oscillations with a period of $\sim$ 2.44 years, in good agreement with the theoretical prediction of Alonso *et al* [5].

with temporal forcing [80], however stochastically induced resonant oscillations may also play a role [5, 168, 217].

These resonant interactions have been analytically investigated using multi-scale analysis [168] and via a perturbative expansion in powers of $\mathcal{O}(N^{-1/2})$ [5] which can be justified using a Van Kampen expansion of the Kolmogorov equation for the epidemic process, see [245] for more details. The outline of the perturbative expansion analysis is to expand the density variables around their deterministic equilibrium, $x(t) = x^* + \Delta x(t) + \mathcal{O}(N^{-1})$, where $\Delta x(t)$ represents the leading order correction due to stochasticity. By including the expansion into the equations for the s and i densities and including Brownian motion terms as in section 4.3 we generate the following Langevin equation:

$$\frac{\mathrm{d}}{\mathrm{d}t}(\Delta x(t)) = A(x^*)\Delta x(t) + \eta(t) + \mathcal{O}(N^{-1}). \qquad (75)$$

This is linear in the perturbation dynamics up to a correction of size $\mathcal{O}(N^{-1})$, Where $\eta(t)$ is a vector of Langevin white noise [245] with a cross-correlation structure determined by the expansion and $A(x^*)$ is the Jacobian matrix evaluated at the fixed point x^*. Using Fourier transforms we find that the spectral power of frequency ω takes the form:

$$P(\omega) = \frac{\alpha + B\omega^2}{[(\omega^2 - \Omega_0^2)^2 + \Gamma^2\omega^2]}. \qquad (76)$$

Where the unknown parameters (α, B, Ω and Γ) can be found via the van Kampen expansion with appropriate time re-scaling, see Alonso *et al* [5] for details. This approach is only capable of predicting the dominant mode of fluctuation, whereas numerical investigation reveals additional peaks in the power spectral density (figure 29).

4.5.2. Stochastic fade-out. The lack of persistence or the existence of fade-outs of disease in a population has been of interest to modellers since Bartlett [34–36] due to the wealth of time series data available, the connection between fade-out extinction and control, and the opportunity to gain testable insight into successful modelling approaches for epidemic models. Of particular importance is the concept of the *critical community size* (CCS), that is a population size above which stochastically driven extinction is not observed at the time scales relevant to human diseases and policy (i.e. centuries). For measles, extinctions (or fade-outs) can be observed from time series data available in England and Wales, USA and various isolated island communities, these demonstrate a remarkably consistent CCS estimate of between 3×10^5–5×10^5 [37]. Investigating the dynamical mechanisms that generate the CCS has been an important area of study for theoretical epidemiologists since it was noted that relatively simple seasonally forced stochastic SEIR models greatly over-estimated the observed CCS value [52]. Modifications to the basic model in order to correct this over-estimation have included adding age structure, spatial structure and non-exponential infectious durations [50, 145, 152]. The measure of disease persistence can also be controversial, if failed invasions are included (implicitly or explicitly) then Erlang distributed infectious periods with a large number of compartments can imply a lower rate of epidemic extinction analogous to the results shown in section 4.4 [145, 152]. On the other hand if statistics for an initial transient period are disregarded multi-compartmental infectious periods can be shown to destabilize an endemic disease and lead to higher rate of fade-out under certain circumstances [174]. In either case more realistic infectious durations are essential for matching stochastic epidemic models to data [62].

The various complex factors required to quantitatively compare stochastic numerical experimentation to real extinction data mitigate against analytic results. However, it is possible to establish at least approximate scaling arguments for the time to extinction in the simplest case when there are no external imports. We again assume that the diffusion approximation to the epidemic dynamics is a good one. Then analysing the time to extinction, T_{ext}, is equivalent to analysing

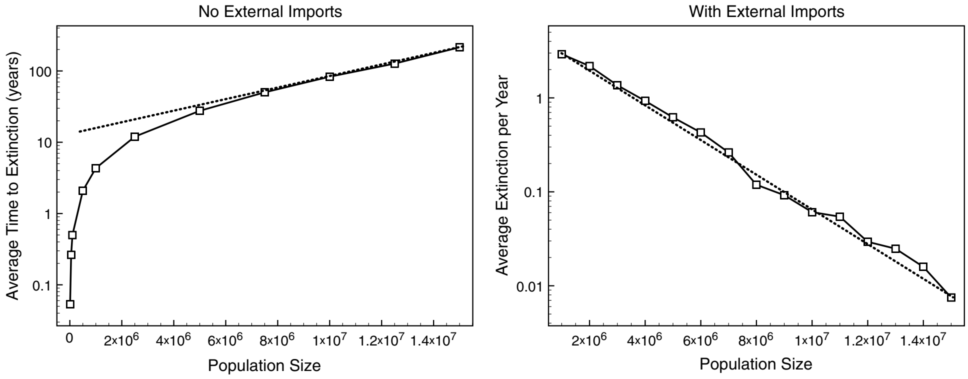

Figure 30. Fade-out statistics estimated from 1000 independent simulation replicates for a SIR epidemic with demography ($\beta = 1, \gamma = 0.2$, $B = d = 3.4 \times 10^{-5}$ all rates given in (days)$^{-1}$) generated using the tau-leap method ($\tau = 0.0584$ days). Left: average time to extinction with no external imports as N^* increases. Initial condition was given by (to nearest integer) the endemic equilibrium state x^* of the related deterministic model. The dashed line gives the exponential scaling α of the time to extinction $\mathbb{E}[T_{\rm ext}] \sim e^{\alpha N}$. Right: the average number of extinction events per year as N^* increases, calculated using 100 year time window with an import rate $\xi = 1.36 \times 10^{-4}$ (days)$^{-1}$. The exponential scaling is given as a black dashed line.

the first exit time problem of a diffusion with a small noise parameter, which has been extensively studied, classically by Freidlin and Wentzell [98, 214]. The key result is that the probability that the exit time is before any give time t obeys a large deviation principle (LDP) [166, 238],

$$\mathbb{P}(T_{\rm ext} < t) \sim e^{-N^* V(x,t)}, \qquad t \geqslant 0, \tag{77}$$

where $V(x,t)$ is the $\mathcal{O}(1)$ rate function for the LDP that depends on the initial state x and the time t [96]. Repeated simulation, with identical initial condition, shows that $\mathbb{E}[T_{\rm ext}] \sim e^{\alpha N^*}$ provides a good approximation for large N^* (figure 30). The presence of external imports greatly complicates any analysis since for low import rates the epidemic might spend a considerable time in the disease-free state, which through demographic turnover leads to an increase in the susceptible percentage of the population and therefore different epidemic conditions upon each successful reseeding. However, numerical investigation shows that the exponential scaling of average extinction events observed per year remains robust (figure 30).

5. Spatial heterogeneity and epidemic dynamics

Spatial structure and the spatial location of hosts or host populations are often crucial to the spread of a wide variety of pathogens, such as Foot-and-Mouth transmitted between farm-based commercial livestock [90, 147] or Sudden Oak Death spreading amongst and between woodlands [94, 216]. The reason spatial effects are so important is simply that transmission is generally a localized process, with greater risk of an infectious pathogen spreading between individuals that have frequent contact. In some cases frequent contact is associated with individuals abiding in close proximity, for example for diseases of plants spread by airborne fungal dispersal. In other scenarios frequent contact is due to the movement of individuals between population centres, this is the usual modelling paradigm for diseases affecting humans.

As with heterogeneity between individuals (section 3) it is often convenient when constructing a spatial epidemic model to divide the total population into a set of sub-groups. However for spatial epidemic models the motivation for subdividing the population is different from the previously considered scenarios. Rather than capturing the epidemiological variation between individuals the spatial epidemic model attempts to incorporate the effect of population aggregations such as towns, villages, farms or spatially proximate plants. It is typical for the frequency of transmission within each sub-group of a spatial model to be more intense than any cross sub-group interaction, due to the localized nature of transmission. The dynamical implications of a spatial division of the population compared to standard spatially homogeneous models has been repeatedly emphasized in the literature of theoretical population ecology, whether under the guise of *island biogeography* [176] or the *metapopulation* paradigm [118, 170]. The close analogy between these ecological investigations and epidemiology has become increasingly commented upon with the growth of models investigating joint host and parasite spatial population dynamics (e.g. [121]). Moreover, the ecological insight into metapopulation persistence has been increasingly seen as important to understanding the persistence of diseases such as measles [108, 144].

In this section we review some of the key dynamical features of epidemic models with spatial heterogeneity. We present these features as falling into broadly two categories, the *explicitly spatial* features that depend heavily upon the spatial positioning of the host populations and the *implicitly spatial* features that are due simply to the subdivision of the population. Often to capture these features requires specific model formulations, and often an individual-level stochastic approach to the dynamics. The discussion of implicit features will extend our previous investigation into stochastic epidemic fade-out (section 4.5.2) to a multi-population model where the interplay between global and local extinctions becomes important. Also the important model generalization of the stochastic SIR model to include household structure will be presented in some detail. Explicitly spatial features we will discuss include spatio-temporal invasion dynamics for PDE and spatially explicit metapopulation models. For spatially explicit models the dynamic and spatial variation in epidemic risk depends crucially on the seeding location of the epidemic and often is characterized by a travelling leading 'edge'.

5.1. Metapopulation dynamics of diseases

The key assumption behind a metapopulation model of population dynamics is that the entire habitable space can be described in terms of a set of discrete patches within which population members mix homogeneously but between which interactions occur at some other rate.

5.1.1. Classical metapopulations. Classically, the between patch interactions are treated as occurring on a significantly slower time scale than the within-patch dynamics. Consequently, a species successfully invading an unoccupied patch rapidly reaches its equilibrium population size at the local carrying capacity of the patch before any further invasion events are likely to have occurred [170]. On the fast time scale the within-patch population dynamics post-invasion are therefore effectively independent of other patch dynamics due to the infrequency of interactions. This assumption of time-scale separation leads to a description of the dynamics at the patch level as either 'occupied' or 'unoccupied'. A metapopulation model of this kind is commonly referred to as a *Levins-type metapopulation* [149].

By making the same time-scale separation assumption for disease invasion into a host population that is segregated into a metapopulation of inhabited patches we can generate analogous Levins-type invasion dynamics. Each inhabited patch can be described as containing either a susceptible population (S) where pathogen invasion is possible, an exposed population (E) within which the pathogen is established and further invasions do not alter the dynamics but also where infectious exports from the population are insignificant, an infectious population (I) where infectious exports actively recruit populations in susceptible patches and also a resistant (R) population which has reached herd-immunity either through the natural population dynamics of the disease or through some control policy such as the widespread vaccination of individuals abiding at the patch. As such we have substituted the SEIR compartmental description of individuals to one that approximates the state of discrete populations inhabiting patches.

A practical advantage to Levins-type models is that the presence or absence of a disease from a given population is often available information. More complex epidemic models which explicitly include disease dynamics within the host patches will generally require high quality data for effective parametrization, which might not be obtainable. As such the methods have been used to great success in wildlife populations where data are often limited [228]. Statistical inference methods for imputing model parameters for individual based epidemic models, including the Levins-type metapopulation model, have been widely discussed and are comparatively straight forward to implement, although potentially computationally intensive [72, 101].

5.1.2. Individual movement and metapopulation coupling. For some diseases within-population transmission due to homogeneous mixing is well founded and well understood and there is sufficient high quality data for parametrization, as is the case for measles in the UK. In this scenario it becomes more desirable to explicitly include both the within-patch transmission dynamics and between patch coupling within the same model framework. The exact modelling approach depends on the mechanism by which the disease is spread between different population sub-groups. In general, the model formulation follows that of any heterogeneous population:

$$\begin{aligned}\frac{\mathrm{d}S_i}{\mathrm{d}t} &= B_i - \lambda_i S_i - d_i S_i,\\ \frac{\mathrm{d}I_i}{\mathrm{d}t} &= \lambda_i S \; - \gamma_i I_i - d_i I_i,\\ \frac{\mathrm{d}R_i}{\mathrm{d}t} &= \gamma_i I_i - d_i R_i,\\ \text{where}\quad \lambda_i &= \sum_j \beta_{ij} I_j / N_j,\end{aligned} \tag{78}$$

where the transmission matrix β must be related to the interaction between sub-populations.

For infectious diseases of humans the natural mechanism for the introduction of a pathogen into a naive sub-population is via infectious commuters spending a period of time in the sub-group [150]. Therefore, the necessary ingredients for a mechanistic model of the spread of human diseases between populations (cities, towns, villages etc) should include temporary demographic movements as well as disease dynamics. The commuting individuals couple the local epidemic dynamics of the various sub-groups. Assuming that each person has a permanent home population, we write the number of people whose home location is the jth area but are temporarily located in the ith area as $N_{ij}(t)$, with similar notation for the number of susceptibles, $S_{ij}(t)$, and infecteds, $I_{ij}(t)$. Individual movement can then be modelled using matrices of rates for leaving the home location j and commuting to the location i (l_{ij}) and the rate of return, (r_{ij}) (figure 31):

$$\begin{aligned}\frac{\mathrm{d}N_{ii}}{\mathrm{d}t} &= -\sum_j l_{ji} N_{ii}(t) + \sum_j r_{ji} N_{ji}(t),\\ \frac{\mathrm{d}N_{ij}}{\mathrm{d}t} &= l_{ij} N_{jj}(t) \; - r_{ij} N_{ij}(t).\end{aligned} \tag{79}$$

Here N_{ij} can either be a continuous scale population for deterministic dynamics, or the model can be modified to account for stochastic dynamics. Generation of new infecteds from S_{ij} group can be modelled as frequency-dependent transmission between all people currently in the jth location. Hence, for the basic SIR model the force of infection on susceptibles in the S_{ij} group is

$$\lambda_{ij}(t) = \beta \frac{\sum_j I_{ij}(t)}{\sum_j N_{ij}(t)}. \tag{80}$$

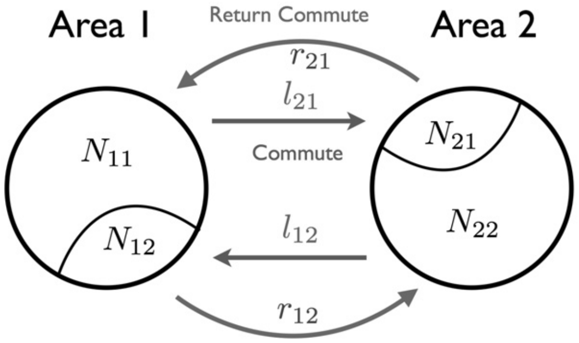

Figure 31. A schematic of the two area commuter model. The permanent population of each area are divided between those in their 'home' area and those temporarily visiting the other area. Epidemic spread occurs between individuals located in each area.

Due to its mechanistic nature the commuter model can be parametrized using any census data that includes commuting data. For example the UK 1991 census data has been so used for predicting the spread of influenza-type infections [70]. The mechanistic model can also serve as a comparison to more complex time series based models [109]. Commuter models have also been used to assess global level control measures, such as the role movement restriction on the global air travel network can play in reducing the potential impact of pandemic influenza [65, 125]. The predictions of such commuter models are pessimistic about the potential success of movement restrictions; 99% of movements need to be prevented in order to delay epidemic spread by just a few weeks [125], although this delay could provide critical time to muster a concerted public-health response.

The full commuter model can be simplified by assuming that all commuter movements are rapid. Rather than accounting directly for the location of each individual it is possible to treat each infectious individual, indexed only by their home area, as contributing to force of infection in each other area [150]. As an illustration of modelling using the rapid movement assumption we consider two sub-populations of equal size and demographic characteristics. The rapid commuter approximation implies that the force of infection on susceptibles in population i (λ_i) is

$$\lambda_i(t) = \frac{\beta}{N}\Big((1-\rho)I_i(t) + \rho I_j(t)\Big), \qquad i, j = 1, 2, \quad (81)$$

where the coupling parameter ρ is connected to the full commuter model through the relation, $\rho = 2q(1-q)$ where $q = l_{ij}/(r_{ji}+l_{ij})$ is proportion of time each individual spends away from their home area [150]. The coupled metapopulation dynamics implied by the rapid commuter approximation gives the force of infection at a spatial location as a weighted sum over the entire infectious population. This can also be good approximation when the underlying coupling mechanism is permanent immigration rather than temporary commuting such as for the population growth and dispersal model of Kot *et al* [161]. If there are n identically sized sub-populations and inter-commuting between them is equal the generalized force of infection is,

$$\lambda_i(t) = \frac{\beta}{N}\Big((1-n\rho)I_i(t) + \rho\sum_{j=1}^{n} I_j(t)\Big), \qquad i = 1, \ldots, n. \quad (82)$$

The coupling ranges $\rho \in [0, 1/n]$ where the maximum value corresponds to commuters spending equal time between their home location and the other locations. Because all population sizes are identical, and movements rapid, R_0 is invariant to the value of the coupling parameter.

Introducing stochastic dynamics into epidemic models of a single population can lead to important effects such as stochastically driven failure of disease to persist [37, 152] and the possibility of resonant interaction between demographic and epidemiological forces [5] (see sections 4.4.1 and 4.5). For spatial metapopulation models this picture is further complicated by the coupling between the sub-groups. If the metapopulation sub-groups are uncoupled then they can be treated as independent single populations. On the other hand if the sub-group dynamics are strongly coupled such that the dynamics are tightly correlated (i.e. synchronous [173]), then the metapopulation dynamics are effectively equivalent to those of a large single population. Interesting dynamics for stochastic spatial models lie in the intermediate coupling regime [50].

For a metapopulation encompassing two coupled identical populations we model a disease outbreak within the stochastic SIR framework including births and deaths. The epidemic is permanently sustained by external infectious imports. For this model the temporal correlation between the numbers of infecteds in the two populations, C_{12}, has been shown to take a particular approximate functional form [150],

$$C_{12} = \frac{\langle (I_1(t) - \langle I_1\rangle)(I_2(t) - \langle I_2\rangle)\rangle}{\sigma(I_1)\sigma(I_2)} \approx \frac{\rho}{\xi + \rho}, \quad (83)$$

where $\langle\cdot\rangle$ denotes an average over time and $\xi > 0$ is fitted from simulation. As the coupling between the two sub-populations increases the temporal correlation between their infected numbers saturates to unity. This indicates that the dynamics of infected numbers are highly synchronized when the populations are strongly coupled (figure 32). A important feature of metapopulation epidemic models is that they imply that the persistence of a disease at the local sub-population level and the global metapopulation level are undergoing different pressures [144]. At the sub-population level low levels of coupling decreases the frequency of epidemic rescue from chance extinctions due to infectious pressure from the other sub-populations. On the other hand, at the global level stronger coupling doesn't automatically imply a significantly lower rate of global extinction. Each sub-population is more tightly synchronized when coupling is stronger, implying that low infection burden in one sub-population is likely to be replicated across the metapopulation. Therefore for strongly coupled sub-populations local disease fade-out can become less frequent but more strongly correlated with global extinction, see figure 32 for a concrete example of this effect for a 10 sub-population metapopulation.

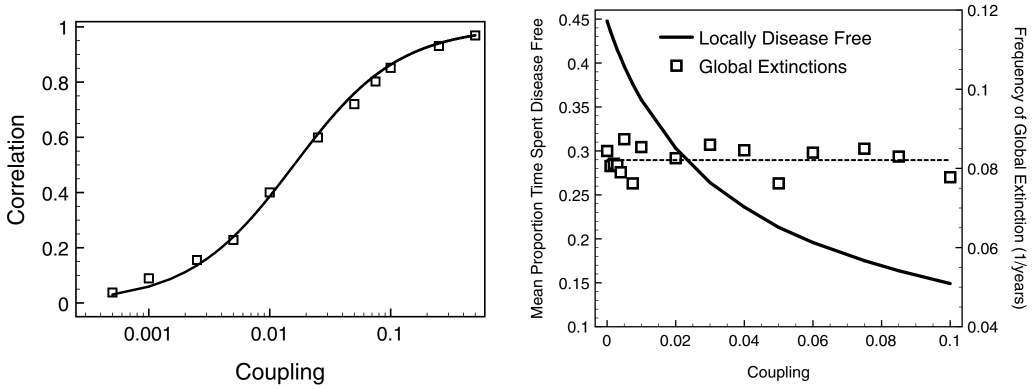

Figure 32. Synchrony and persistence for coupled metapopulation model with demography. The epidemiological parameters were $\gamma = 0.2$ (days)$^{-1}$, $R_0 = 10$, births and deaths occurred at rate $B = d = 5 \times 10^{-5}$, external infectious import rate was chosen as the standard $5.5 \times 10^{-5}\sqrt{N_i}$ (see Bartlett [34]). For the examples time averages and time proportions were calculated from a 3000 year simulation run with an initial 100 years discarded as burn-in. Left: 2 population sub-groups each of size $N_i = 10^6$. The temporal correlation, or synchrony, of the numbers of infecteds as a function of the coupling parameter ρ. Good agreement was found to a sigmoidal fit of the form $C = \rho/(\xi + \rho)$. Right: 10 population sub-groups each with population size $N_i = 3 \times 10^4$ and equal commuting force of infection (82) with $R_0 = 12.5$ (other parameters as before). The average (over sub-groups) proportion of time each local sub-group spends disease free (solid curve) is monotonically decreasing with coupling as rescue events reduce the frequency of stochastic fade-out, however the global persistence, measured by yearly frequency of global disease extinction, is nearly invariant to increased coupling (mean is dashed line). This is due to greater synchrony; local extinction events are less common but more strongly correlated with other local extinctions.

5.1.3. Household dynamics. An epidemiologically important special case of the metapopulation model is the *household* model of disease, which attempts to account for the greater transmission of infection between cohabiting humans compared to casual contacts made with the global population. From a metapopulation point of view a household model consists of a large number of small population sub-groups (local population sizes of less than 10) each of which represent a human cohabitation group or household. For infectious diseases of humans the household model can represent the ideal trade-off between capturing epidemiological details that are missed by classical epidemic models whilst retaining a degree of tractability in analysis [24, 26], parametrization [97] and control design [127]. The within-household transmission is overlaid by homogeneous transmission within the entire population, thus the household model can be thought of as a combination of a simple network cluster (the household) and the classical homogeneous transmission epidemic models. As such the force of infection in household i is generally expressed as:

$$\lambda_i = \widehat{\beta}(N_i)I_i(t) + \alpha\frac{\sum_j I_j(t)}{\sum_j N_j},$$

where $\widehat{\beta}$ and α capture within and between household transmission. It is generally assumed that between household transmission is frequency dependent, while the strength of within-household transmission may be a function of household size [128]. Generally, household models are formulated as stochastic since each household contains a small number of individuals. We continue this trend and take a particular form for the stochastic arrival rate of a transmission event within a household of size N_i,

$$r_I(t) = \beta\frac{I_i(t)}{N_i - 1} + \alpha\frac{I(t)}{N}, \qquad (84)$$

where $I(t)$ and N refer to the total number of infected individuals and the total population size respectively, and a simple frequency-dependent mixing assumption has been made for within-household transmission. Other disease events such as recovery or waning immunity are left unchanged.

Two key determinants of early pathogen invasion dynamics are the basic reproductive number R_0 and the early exponential growth rate r such that $I(t) \approx I(0)\exp(rt)$. For the basic SIR model with $\exp(\gamma)$ distributed infectious periods these can be derived as $R_0 = \beta/\gamma$ and $r = \beta - \gamma$. Naturally it is of interest to determine comparable values for the same basic SIR model but with incorporated household structure. It is common and analytically simpler to consider the household reproductive number R_* [27], defined as the expected number of households infected in a naive population by an average infected household. Importantly, R_* is also a threshold quantity; an epidemic cannot grow and infect a significant number of households if $R_* \leqslant 1$ [28].

If the number of households H is very large and the size of each household N_i is insignificant compared to the size of the population ($N_i \ll N$) then the probability that an infectious individual in the initially infected household recruits more than one individual from any naive household is effectively zero. Therefore during the early dynamics we can safely ignore the possibility multiple import events into a household. In the limit $H \to \infty$ and assuming all households are of equal size (n), the household reproductive number is the total secondary transmissions originating from an infected household,

$$R_* = \frac{\alpha}{\gamma}\mathbb{E}[Z_h], \qquad (85)$$

where Z_h is the final number of individuals infected in the seed household where there is a solitary initial infected and no external infectious pressure. As noted by Ross *et al* [219]

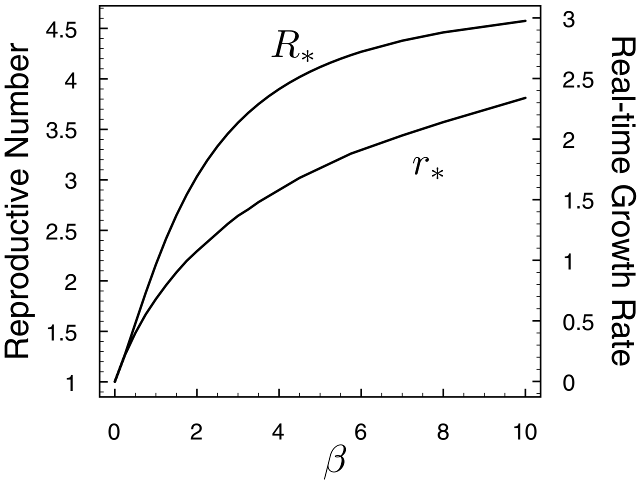

Figure 33. The household reproductive number R_* and real-time growth rate r_* for a population segregated into households of $n = 5$ individuals. The recovery rate and global transmission rate are unity, $\alpha = \gamma = 1$ hence the epidemic can only be sustained due to additional within-household infectious transmission occurring at individual to individual rate β. Due to both quantities increasing monotonically with β estimating the real-time (exponential) growth rate from public-health data also gives an estimate for R_* and the within-household transmission rate.

equation (85) is equivalent to the time integral formulation,

$$R_* = \mathbb{E}\left[\int_0^\infty \alpha I_h(t)\,\mathrm{d}t \,\middle|\, S_h(0) = n-1,\, I_h(0) = 1\right]. \quad (86)$$

There exists a considerable literature devoted to solving expressions of the type (86), which can be viewed as an uncontrolled version of a cost integral familiar from stochastic optimal control [96] and Markov decision processes [213]. Alternatively, Kolmogorov equations become feasible given that there are only $(n+1)(n+2)/2$ possible configurations for a household of size n. Methods based on Selke constructions are also practical [24]. House *et al* provide a comprehensive review and comparison of the methodologies for calculating final size distributions [127].

R_* provides a useful analogue to the classic R_0 in that it acts as a threshold parameter. However, for model parametrization calculating the real-time exponential growth rate could be more important since it is amongst the quantities available from public-health data during the early invasive stage of an epidemic outbreak. Ball *et al* [27] have demonstrated that the real-time exponential growth rate for households (r_*) is given by the unique positive solution to

$$1 = \mathbb{E}\left[\int_0^\infty \alpha I_h(t)\mathrm{e}^{-r_* t}\,\mathrm{d}t \,\middle|\, S_h(0) = n-1,\, I_h(0) = 1\right]. \quad (87)$$

Finding the unique r_* is therefore reduced to a root finding exercise [219]. Equation (87) was derived from a branching process approximation and therefore applies only at the early stages of an epidemic (figure 33).

In this discussion we have restricted ourselves to considering populations segregated into equally sized households, clearly this is a simplification of reality and neglects the differing role that large and small households can play in an epidemic outbreak. A larger household will generally imply more infected individuals once invaded by an infectious pathogen than a smaller household, therefore the larger invaded household will generally contribute greater infectious pressure. It is important to take into account that a larger household is also relatively more likely to be invaded than a smaller invaded household. At the total population level the fraction of all households of a given size n is typically described by a size distribution

$$\mathbb{P}(\text{Randomly chosen household size} = n) = p_n. \quad (88)$$

During the early epidemic phase (i.e. when the branching process approximation is valid) the probability that an individual infected via a global contact is a member of a size n household is given by the *size biased probability* $\tilde{p}_n$,

$$\tilde{p}_n = \frac{np_n}{\sum_{k\geqslant 0} kp_k}. \quad (89)$$

To calculate R_* and r_* for a population of heterogeneous households requires calculating quantities for each household size n and taking a size biased average [43].

The varying importance of household size to epidemic spread can be considered though the contribution of a given household size to R_*. It is clear that larger households provide a larger target for infection and therefore are more likely to be infected in the early epidemic, however the degree to which the household amplifies the infection depends on the assumptions about within-household transmission. For a household model with density dependent transmission; that is that the within-household transmission rate is $\beta I_h(t)$ rather than $\beta I_h(t)/(N_h - 1)$, Ball *et al* [27] have demonstrated that the smallest number of successful prophylactic vaccinations required for $R_* \leqslant 1$ is obtained by following an equalizing strategy. The equalizing strategy aims to leave all households with the same number of susceptibles before an outbreak, which effectively biases vaccine deployment to larger households. It is still unclear how these theoretical insights translate into practical control advise, when either there is insufficient vaccine to obtain $R_* = 1$ or when within-household transmission is frequency (rather than density) dependent.

5.2. Spatial positioning and invasion dynamics

The metapopulation ideal used in the previous section becomes unworkable when there is no natural means of partitioning the population, when potential hosts are dispersed fairly evenly over a geographic area any subdivision of the space is inherently arbitrary and continuous space models with dynamics given in terms of reaction–diffusion or integro-differential equations are more natural [201]. These dynamical formulations are generally deterministic, predicting continuous valued population densities across continuous space. An common alternative is to consider stochastic lattice-based models [215]. These models represent a hybrid between the discrete population sub-groups of a metapopulation model and the continuous space of reaction–diffusion systems. Finally, even when the underlying assumptions for a metapopulation model hold the rate of transmission between separate spatial units might well depend strongly on their relative spatial position [89, 147] in which case proximity becomes a significant risk factor and a useful method of targeting controls [236].

5.2.1. Reaction–Diffusion epidemic models. Capturing the movements of individuals is the key factor in the spatial transmission of infection. Reaction–diffusion models assume that individuals follow a Brownian random walk, which is often a reasonable approximation for dispersing animal populations [209]. In the absence of some attractive potential, a Brownian random walker describes a continuous path through space. Having observed the location of the random walker at some time t the change in location by future time $t + \Delta t$ will be Gaussian distributed with zero mean. In this section we will only consider random walkers on the 2D plane, where the position change in the two spatial dimensions over a period length Δt are independent. The probability density of the location of the random walker, $p(x,t)$ $x \in \mathbb{R}^2$ $t \geqslant 0$, evolves from its initial density according to the Fokker–Planck equation [158],

$$\frac{\partial p}{\partial t} = \frac{D}{2}\nabla^2 p, \tag{90}$$

where D is the diffusion coefficient and ∇^2 is the Laplacian operator in two dimensions.

It is possible to overlay epidemic models onto this random walker model, as done in the classic modelling work for the spatial spread of rabies [201] and bubonic plague [207]. For example we could model each random walker as having a disease state from the SIR compartment model and disease state dependent diffusion coefficients (D_S, D_I, D_R). The local densities for the population in each disease state are then given as $S(x,t)$, $I(x,t)$, $R(x,t)$. Epidemic transmission occurs locally, with the local reaction terms due to individual recovery rate γ and a force of infection given by the local frequency-dependent form

$$\begin{gathered}\lambda(x,t) = \beta\frac{I(x,t)}{N(x,t)},\\ N(x,t) = S(x,t) + I(x,t) + R(x,t),\\ x \in \mathbb{R}^2, \qquad t \geqslant 0.\end{gathered} \tag{91}$$

The spatial dynamics of this epidemic model are given by the non-linear reaction–diffusion partial differential equation (we refer the reader to Evans [87] for analytic details on these types of modelling equations),

$$\begin{aligned}\frac{\partial S}{\partial t} &= -\lambda S + \frac{D_S}{2}\nabla^2 S,\\ \frac{\partial I}{\partial t} &= \lambda S - \gamma I + \frac{D_I}{2}\nabla^2 I,\\ \frac{\partial R}{\partial t} &= \gamma I + \frac{D_R}{2}\nabla^2 R.\end{aligned} \tag{92}$$

The invasion dynamics of reaction–diffusion epidemics starting from infection at a single spatial location are characterized by a front of peak infection expanding isotropically away from the focal point. This is due to the epidemic front expanding preferentially into areas of greater local susceptible density. The emergence of an invasive travelling wave of infecteds raises some important epidemiological considerations, in particular the question of how rapidly the peak infectious front is travelling which has clear implications for the design of spatially targeted control measures. The formation of a moving epidemic front is not just a feature of reaction–diffusion models but is common to many epidemic models in continuous space, however the asymptotic velocity is easier to calculate analytically for reaction–diffusion models.

Since the underlying space and the movement of the random walkers are spatially isotropic we assume that the solution to the reaction–diffusion PDE can be written in travelling wave form, in a combined time and space variable $z = r - ct$,

$$S(x,t) = \hat{S}(z), \qquad I(x,t) = \hat{I}(z), \qquad R(x,t) = \hat{R}(z), \tag{93}$$

where $r = |x|$ is the distance from the initial point of infection, treated as the origin, and c is a wave speed. Introducing (93) into the reaction–diffusion dynamics gives the large radius ($r \gg 1$) expression for the travelling wave structure of the infected population,

$$\frac{D_I}{2}\hat{I}''(z) + c\hat{I}'(z) + \beta\frac{\hat{S}(z)}{\hat{N}(z)}\hat{I}(z) - \gamma\hat{I}(z) = 0, \tag{94}$$

Where we have neglect a term proportional to $1/r$ due to the large radius approximation. At the leading invasion edge of the travelling wave profile the population is naive ($\hat{S}(z) \approx \hat{N}(z)$). There will be a stable large radius travelling wave profile for the infected density whenever the characteristic equation of (94) has real roots, i.e. $(D_I/2)s^2 + cs + (R_0 - 1)\gamma = 0$, has a real solution. Since the invasion dynamics initially accelerate into the naive population the leading edge behaviour is asymptotically governed by the lowest wave velocity where (94) can predict a wave-like solution. Assuming $R_0 > 1$ the asymptotic wave speed is therefore given by

$$c = \sqrt{2D_I(R_0 - 1)\gamma}. \tag{95}$$

Although the asymptotic invasion speed is analytically accessible the full epidemic dynamics PDE does not admit an explicit solution; as is typical for non-linear dynamics over a wide variety of types of evolution. The efficient numerical solution of PDE models is a highly active area of research [240]. We give an example of a reaction–diffusion epidemic amongst animals where only the infected animals diffuse, with the remaining population static (figure 34). This follows the rabies model of Murray *et al* [201] where healthy foxes are territorial and remain stationary, whereas rabid foxes wander randomly occasionally transmitting to healthy foxes. Although for simplicity we do not use a parametrization and population structure entirely appropriate for rabies modelling.

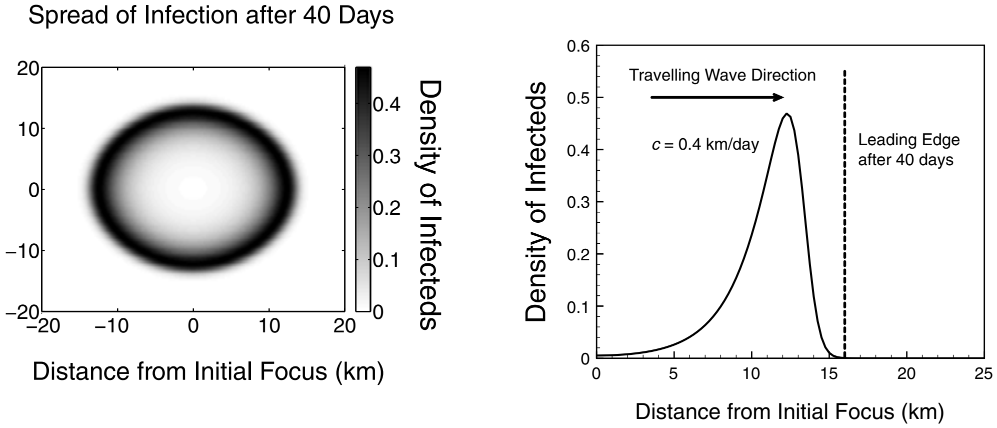

Figure 34. A reaction–diffusion epidemic amongst animals. Infected animals become rabid and walk randomly with coefficient $D_I = 0.1\,\mathrm{km}^2/\mathrm{day}$. The disease dynamics are characterized by a recovery rate $\gamma = 0.2\ (\mathrm{day})^{-1}$ and a basic reproductive number $R_0 = 5$. Analysis suggests that the asymptotic wave speed is $c = 0.4\,\mathrm{km/day}$, the leading edge of the epidemic is predicted to be 16 km from the initial focus after 40 days. Left: snapshot of spatial distribution of infected animal density after 40 days, solved numerically using a discretization $\Delta x = 0.25$ km and $\Delta t = 0.01$ days. Right: profile of numerical solution after 40 days demonstrating excellent agreement with the asymptotic invasion velocity prediction.

5.2.2. *Kernel-based spatial epidemic models.*

The underlying assumptions behind the reaction–diffusion spatial epidemic model were that the host-to-host transmission range is very short and that the mixing of individuals could be described by their diffusive random walk. These assumptions make the reaction–diffusion paradigm appropriate for modelling diseases of wild animals. However diseases of stationary individuals, e.g. plant-life, are usually better described by long-range transmission, for example by the wind-borne spore dispersal for fungal diseases [59]. Long-range dispersal of infectious pressure might also be a better modelling assumption for livestock-based epidemics where farms are fixed locations in the environment [89, 147, 197].

For transmission due to long-range mechanisms like spore dispersal we potentially have epidemiological interaction between a large proportion of the population, however the host-to-host transmission rate will typically decrease with the distance between the infectious and susceptible organisms. The dependence between spatial separation and transmission rate is usually specified by a *transmission kernel* K, which governs the force of infection $\lambda(x,t)$ experienced by the spatial point x due to the entire spatial distribution of infected hosts,

$$\lambda(x,t) = \beta \int I(y,t) K(x-y)\,\mathrm{d}y. \tag{96}$$

Equation (96) is the most basic form for a kernel-based force of infection expression, β sets the intrinsic transmission rate and $K(\cdot)$ encodes the effect of the spatial separation between a susceptible host at x and the infectious hosts at y. The force of infection form (96) is appropriate for both individual based models with stochastic dynamics [51] and for deterministic continuous population density models where the hosts are distributed uniformly across space [77]. Non-uniform spatial dispersal of host may require the inclusion of a local density $N(x,t)$ depending on the nature of transmission and interaction.

It is common to assume that the transmission kernel is invariant under translations and rotations; that is that no particular location is preferential for infection and there is no particular preferential direction of transmission. Therefore the kernel is only a function of separation distance $K(x) = K(r)$.

When the long-range transmission rate decays rapidly, i.e. $K(r) \to 0$ at least exponentially quickly as $r \to \infty$, then kernel-based stochastic transmission models predict that a pathogen invading into a large homogeneously dispersed host population spreads wave-like and obtains an asymptotic invasion velocity c that can be calculated analytically (see either [76, 77] for details). The invasion dynamics, from a spatial spread point of view, are essentially similar for both kernel-based and diffusion-based dispersal models. By contrast when $K(r)$ decays more slowly than exponentially with r it is possible for the pathogen to spread at an ever increasing velocity [161]. In this case the invasion dynamics cannot be summarized as a moving invasion front [191], simulation studies suggest that the invasion jumps forward generating new local foci [226] which reflects occasional dispersal which is very long range compared to ‘typical’ dispersal.

5.3. *Case study. Infectious diseases of farmed livestock: spatially explicit metapopulations*

In previous sections we have found the metapopulation approach to be a useful paradigm for spatial epidemic dynamics when the population can be segregated into groups determined by their location. Between sub-group transmission is usually considered to occur at a significantly lower rate compared to within group transmission, but otherwise is essentially arbitrary. In this section we marry the explicit distance dependence of kernel models with the fundamental metapopulation assumption that the population is segregated into a discrete set of local sub-groups, to produce a highly adaptable framework. The cost of this model flexibility is analytic intractability and we resort to simulation.

We focus on an infectious disease invading a spatially explicit Levins-type metapopulation using locally targeted vaccination as a control measure. This has proved to be a be a very successful modelling framework for the spread of Foot-and-mouth (FMD) amongst commercial livestock [89, 147, 197] as well as other diseases of livestock such as bluetongue virus [231]; implicit in the model is the assumption that the within-farm dynamics are sufficiently rapid that each can be treated as susceptible, exposed, infectious or removed (see section 5.1.1). For FMD there is a wealth of detailed historical data generated by the 2001 outbreak in the UK. This has allowed extensive retrospective analysis of the predictive power of spatially explicit modelling of FMD spread in the UK [237], statistical investigations of best parameter imputation [72, 229] and retrospective impact assessment of the proactive culling control measures used during the outbreak [234].

Our aim is to introduce the major model features required for FMD epidemic modelling and demonstrate that for a simplified model of FMD-like spread there exists an optimal radius around confirmed infected premises (IPs) within which vaccination should be targeted; this essentially recreates the result of Tildesley *et al* [236] albeit for a simplified FMD model. We closely follow the kernel-based transmission model used during the 2001 epidemic [147]. Farms, labelled $i = 1, \ldots, N$, are treated as point locations located respectively at $\{x_i \in \mathbb{R}^2\}_{i=1}^N$. Each farm has an epidemic status as either susceptible to disease (S), exposed (E), cryptically infectious (I) or removed (R); in this context removal denotes either that the farm livestock were culled once infection was detected or successfully vaccinated. The essential goal of this simulation study is to investigate the optimal distribution of vaccine so as to minimize numbers of total number of farms who have their livestock culled.

The force of infection on the susceptible farm i is given by the kernel-based transmission expression,

$$\lambda_i(t) = \sigma_i \sum_{j \in \text{inf. farms}(t)} \tau_j K(|x_i - x_j|), \qquad t = 0, 1, 2, \ldots, \tag{97}$$

where σ_i gives the susceptibility of farm i and τ_j the transmission rate from farm j. For accurate modelling of FMD these factors depend upon the numbers and species of the livestock at each farm [147], however for the simplified model presented here we assume homogeneity such that $\sigma_i = 1$ and $\tau_j = \beta$ for all farms. The kernel transmission model reflects the spatially localized nature of FMD transmission after a ban on livestock movement has been imposed, before any movement ban it is possible that the disease can be spread through farm-to-farm livestock relocations and hence a network approach might better reflect the outbreak dynamics [157]. Moreover, the FMD kernel used by both Keeling *et al* and Tildesley *et al* was directly estimated from veterinarian reports; for simplicity we use a Gaussian shaped kernel.

The dynamics are treated as stochastic and occur on a daily time scale; that is that any event that occurs during a day only effect the rate of future events on the next day. Therefore, the probability of the susceptible farm i becoming exposed on day i is,

$$\mathbb{P}(\text{Farm } i \text{ becomes exposed on day } t) = 1 - \mathrm{e}^{-\lambda_i(t)}. \tag{98}$$

Once exposed the new IP farm becomes actively, but cryptically, infectious with a daily probability given by the rate $\alpha > 0$. Clinical signs of infection in the livestock are subsequently detected on the infectious farms with rate $\gamma > 0$, and the presence of infection is realized. These detected IPs are promptly culled within 1 day, in agreement with the peak culling efficacy observed during the 2001 outbreak. Vaccination is deployed at all non-removed farms (although is only effective on susceptible farms) within a pre-defined vaccination radius, r_V of a detected IP up to a daily maximum capacity of deployment; priority is given to farms in the order that they were identified for vaccination. For the simple model considered here vaccine is treated as giving 100% immediate protection whenever deployed to a susceptible farm.

Upon detecting an IP it was standard practice for veterinarian assessment of dangerous contact (DC) between the IP and other farms to be performed, this was based upon such factors as recent vehicle movement between the premises etc. In this way, potentially infected farms were hoped to be identified before they could become infectious and generate subsequent cases. For the simple FMD model given here, the probability that a (non-removed) farm j is a dangerous contact of a newly detected IP i is again defined in terms of the transmission kernel,

$$\mathbb{P}(\text{Farm } j \text{ is a DC of farm } i) = 1 - \mathrm{e}^{-F\sigma_i\tau_j K(|x_i - x_j|)}. \tag{99}$$

DC farms are also promptly culled (within 2 days). The scaling parameter F should in principle be fitted from field data in order to reflect observed ratios of IP culls to DC culls. Given that transmission is a local process, additional culling of farms contiguous to an IP could be implemented [147]; we do not include this extra complication in our simplified model although there is evidence to suggest its effectiveness [234].

Clearly there is a significant trade-off between targeting vaccination locally to each IP and thereby protecting the farms most at risk and targeting more widely in order to get ahead of epidemic spread. Therefore one might expect a global minimum in the expected numbers of farms culled as a function of vaccination radius, which has been observed for the full model with more realistic vaccination assumptions and is a robust conclusion from simulation results of our simplified model (figure 35). However, if the policy decision is to minimize the epidemic duration (and hence costs on the economic) rather than minimize the number of farms affected, it might be preferable to increase the vaccination radius. Whether it is possible to gain analytic insight into the complex dependency of the optimal vaccination radius upon factors such as the maximum vaccination capacity [236], the spatial clustering of potential host farms [233] or underlying epidemiological parameters remains an open question. At the moment we are restricted to multiple stochastic simulations if we wish to include a degree of realism within our models.

5.4. Networks

Networks provide a highly informative way to characterize (spatial and/or social) interactions, where due to a realization of some underlying process there are a limited number of

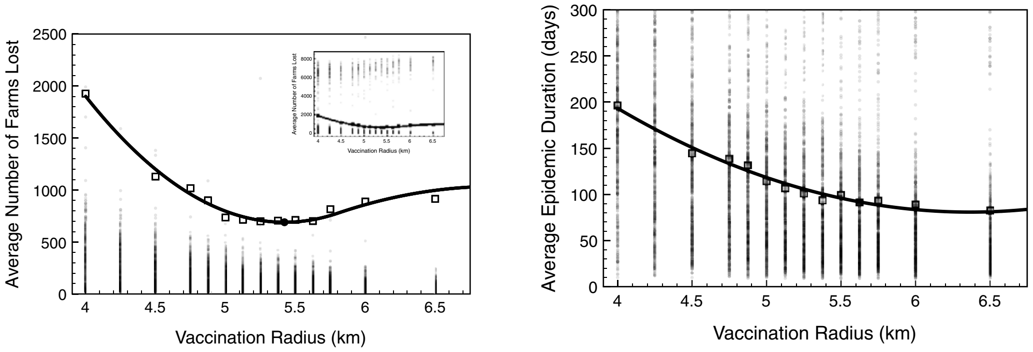

Figure 35. Results from a simplified Foot-and-Mouth like outbreak amongst $N = 10^4$ farms dispersed randomly over a 100×100 km area. Stochastic dynamics are given by Keeling-type daily transition probabilities [147]. Transmission is governed by a Gaussian shaped transmission kernel (width $L = 3$ km) with epidemiological parameters exposure multiplier $\beta = 0.6$, rate of becoming cryptically infectious $\alpha = 0.25$, and detection rate $\gamma = 0.2$ and dangerous contact scaling parameter $F = 6$. Farms with confirmed infection (IPs) and their dangerous contacts (DCs) are removed (quarantined and culled) within one day, other farms within the vaccination radius are vaccinated at a maximum number of 50 per day. The epidemic is initialized from a cryptically infectious farm at the centre of the space and its 9 nearest neighbours exposed to the disease. Left: the effect of vaccination radius on average epidemic outbreak (1000 simulation replicates with cubic spline smoothing). A clear trade-off exists between protecting farms ahead of disease spread and targeting more locally to detected IPs with a minimum in number of culled farms at a vaccination radius at 5.42 km. Dots show individual simulations. The inset shows the essential dichotomy in severity between the success and failure of early control effort. Right: average duration of the epidemic with vaccination radius. Dots show individual simulations.

interactions per pairs of individuals. As such networks are ideal for describing the spread of human infections (although they have been used in many other contexts), as humans tend to have a relatively low number of close-contact social interactions through which infection can spread, in comparison to population sizes. In fact there are strong links between networks and epidemiology; a network of interactions defines the possible transmission routes an epidemic can take, while the path taken by an epidemic naturally defines a network. It is therefore unsurprising that epidemiological applications of network theory abound [22, 69, 85, 159, 171, 250], with examples from theoretical physics, statistical mechanics and probability theory forming a theoretical backbone [2, 49, 113, 202, 205].

However the interplay between social contacts and transmission networks is not straightforward and within the same population multiple transmission networks may exist; different infectious diseases are associated with different network structures depending on the mode of transmission and characteristics of the pathogen. For example, highly contagious diseases that are spread by aerosol transmission will have a very dense transmission network with many contacts between individuals and a high degree of clustering. In contrast, diseases that require extremely close intimate contact, such as sexually transmitted diseases, will have a much sparser transmission network. In recent years, transmission networks have become a powerful and popular tool for investigating the spread of infection through a given population. An individual (or collection of individuals) in the population is represented as a node in the network, and a contact that could allow the transmission of infection between two individuals (for example sexual contact) is represented by a link or edge between the two nodes. These edges can be directed, meaning that infection can only be spread along the connection in one direction; an example of such directed transmission comes from the movement of hosts from one location to another, with the location acting as a node [22, 246]. However, in the vast majority of contexts, a contact between two individuals could allow infection to be passed in either direction, depending only on the infection status of the individuals involved. In this section we generally restrict our attention to such undirected networks.

For a population of size N, the undirected network of contacts through which a disease can spread can be represented by an adjacency matrix $G \in \{0, 1\}^{N^2}$,

$$G_{ij} = \begin{cases} 1 & \text{if } i \text{ and } j \text{ are connected,} \\ 0 & \text{otherwise.} \end{cases} \tag{100}$$

For undirected networks where infection can pass in either direction along a link, the adjacency matrix is symmetric. In principle the values in G could be extended to real values and represent the strength of connection between nodes. However, due to a lack of quantitative data such weighted networks are rarely used in practical settings, hence we generally assume that G is a binary variable.

From this matrix we can formulate a model for an SIR-type infection for the dynamics of node i,

$$\begin{aligned} \frac{\mathrm{d}S_i}{\mathrm{d}t} &= \lambda_i S_i, \\ \frac{\mathrm{d}I_i}{\mathrm{d}t} &= \lambda_i S_i - \gamma I, \\ \frac{\mathrm{d}R_i}{\mathrm{d}t} &= \gamma I_i, \\ \lambda_i &= \sum_j \tau G_{ij} I_j / N_j. \end{aligned} \tag{101}$$

Hence in this simple deterministic model, the adjacency matrix plays the same role as the who acquires infection from whom matrix in heterogeneous populations (equation (13)) or the transmission kernel is spatial models (equations (78), (97));

obviously to recapture these previous models we need to allow the connections between nodes to have a specified strength [69]. The above model therefore provides a representation of transmission between multiple large populations such that their internal dynamics can be treated as approximately deterministic. However, in the vast majority of cases network models are treated stochastically (see section 4) with each node representing a single individual that is either susceptible infected or recovered.

One advantage with networks is the speed with which they can be simulated. For spatial models (equation (97)), the force of infection acting on each individual (or sub-population) must be calculated by summing over all infected units and weighting by the appropriate kernel. In contrast, in network-based models there are generally only a small number of contacts and hence only a small number of potential sources of infection that need to be considered. The combination of this refinement with efficient stochastic simulation methods (see section 4.2.2) makes networks an ideal choice for large-scale simulation of infectious disease spread [85, 246], especially for infections with limited routes of transmission.

5.4.1. Network properties. Given that the network prescribes the epidemiological interaction, describing the underlying network structure in terms of a few basic properties can give important insights into the types of epidemic dynamics that could be expected. There is an ever growing number of network properties in the literature, but here we outline five that are key to the spread of infectious diseases. The first three of these are local properties that depend on the structure of the network around each individual, the last two are global properties that depend on the entire network and can be highly sensitive to its wider structure.

Degree distribution. The number of connections that a node has is known as its degree. Assuming an undirected network, the degree of node i is calculates as:

$$k_i = \sum_j G_{ij}.$$

This number gives an intuitive understanding of an individual's risk of acquiring an infection as well as their potential to cause further cases. Taken over the entire population these degrees form a distribution of probabilities; the degree distribution, $P(k)$, gives the probability that a randomly selected node has degree k. However, the degree of a node is not the only predictor of risk, the degree of their connected nodes and their position within the overall network also play important roles [100].

Degree correlations. In general, real networks are correlated with respect to degree such that the probability that a node has degree k depends on the degree of the neighbours of that node, k'. This is captured by the conditional probability $P(k'|k)$. A straightforward measurement to characterize degree correlations is the average nearest neighbour degree distribution, $k_{nn}(k)$, which measures the average degree of nodes connected to a node of degree k. Each node i in the network has an average nearest neighbour degree given by

$$k_{nn,i} = \frac{1}{k_i} \sum_j G_{ij} k_j. \tag{102}$$

The average nearest neighbour degree distribution can then be calculated by working out the average of k_{nn} for all nodes of a given degree k. This is related to the conditional probability $P(k'|k)$ since

$$k_{nn}(k) = \sum_{k'} k' P(k'|k). \tag{103}$$

The average nearest neighbour degree distribution provides a measure of the assortativity of the network (see section 3.6.3). When $k_{nn}(k)$ is increasing with k the network is considered to be assortative and high degree nodes tend to be connected to others with high degree. Strong correlations were shown to exist in sexual networks [171] and in other social networks, such as collaborations of networks of mathematicians, film actors and business people [204].

Clustering. In social networks it is reasonable to assume that any two of your friends are likely to know one another. This is known as clustering and is a property that has been widely observed. Given a network, a qualitative measure of the level of clustering is given by the clustering coefficient and is defined as the ratio of triangles to all triples in the network [143]:

$$\phi = \frac{\mathrm{trace}(G^3)}{||G^2|| - \mathrm{trace}(G^2)}. \tag{104}$$

Clustering is known to have substantial and conflicting epidemiological consequences. The presence of triangles within the network means that there is intensified competition between infected individuals for new susceptible hosts, which will slow epidemic spread. However, this is somewhat tempered by the fact that neighbouring susceptibles can be reached by multiple short-length paths. Understanding how these two factors effect dynamics and control at a local and population scale is an active area of research [131]. It should be noted that the household models explored in section 5.1.3, provide an example of extreme clustering as ever member of a household has connections to every other member.

Components. For infectious disease transmission, the most important global feature of a network is the presence of connected components—groups of nodes such that any node in the group can be reached from any other node in the group by following edges in the network. These connected components define the limit of disease spread in the network. (Note for directed networks, the definition of connected components becomes more involved [49].) A network has a Giant Component if a single component contains the majority of nodes in the network, so that most nodes are reachable from each other; in terms of infection this means that a highly infectious disease could reach the majority of the population. However, studies have shown that networks of sexual contacts are general made up of relatively small components. A study in Manitoba, Canada [254] found that a network of

4544 individuals consisted of 1503 components of size 2 to 82. Only 23 components had 19 or more persons and there were two types of component; 'linear components', where assortative mixing was present and all individuals had between 1 and 4 partners, and 'radial components', characterized by disassortative mixing with one highly central node. This obviously raises the question of how infections spread through entire populations; the answer is that the Manitoba study (and most studies of contact networks) focus on connections made in a fixed time window, the dynamic nature of sexual (and other) networks which are continually evolving helps to explain how entire populations are interconnected.

Path lengths. A path between two nodes in a network (or more precisely in a connected component of a network) is a series of steps following edges in the network to get from one node to the other. The shortest path length between two nodes (i and j) is the path with the fewest number of steps and we denote this as d_{ij}. The diameter of the network is defined as the longest shortest path length between any two pairs of nodes, $\max_{ij} d_{ij}$. Clearly, the average path length and diameter of the network will have a significant effect on the speed with which an infectious disease can reach all parts of the population. Average path length is affected by other network measures, with increased average degree and increased variance of the degree distribution leading to shorter path lengths, while increased clustering generally leads to an increase path lengths.

5.5. Network types

While simulation of network models is relatively straightforward, determining the appropriate structure for a network is highly complex. Two approaches dominate the literature, creation of synthetic networks based on a prescribed set of often simple rules or the use of data to describe network connections.

5.5.1. Synthetic networks.

Random and configuration networks. The most fundamental and widely studied random graph model is known as the Erdös-Rényi random graph [83]. The network consists of N nodes and between each distinct pair of nodes an edge is present with probability p, independent of all other edges. Therefore the mean degree is $c = (N-1)p$. The degree distribution p_k can be calculated by considering a node being connected to exactly k other nodes. Given that there are $\binom{N-1}{k}$ ways to choose the k nodes to connect, the degree distribution is given by,

$$\begin{aligned} p_k &= \binom{N-1}{k} p^k (1-p)^{N-1-k} \\ &= \binom{N-1}{k} \left(\frac{c}{N-1}\right)^k \left(\frac{N-1-c}{N-1}\right)^{N-1-k}. \end{aligned}$$

In the limit of large N

$$p_k = \frac{(N-1)^k}{k!} \left(\frac{c}{N-1}\right)^k \mathrm{e}^{-c} = \mathrm{e}^{-c} \frac{c^k}{k!},$$

which is the Poisson distribution.

An alternative random network formulation is the configuration model [192]. Here the degree k_i of each node is set initially (generally picked from given distribution) creating multiple unconnected nodes with k_i stubs. These stubs are then connected at random, and in a large population the chance of self-connection or multiple connections between two nodes becomes vanishing small. This can be compared to the proportionate or random-mixing assumption of section 3.6.3. Hence the configuration algorithm allows us to produce networks with arbitrary degree distributions, although with little or no clustering; it is closely related to the random-mixing assumptions used for earlier models of heterogeneous risk in STIs (section 3.6.3).

Lattices and small world networks. In keeping with other work in this section (see section 5.2) it is natural to assume that individuals have a spatial location and form contacts relative to that location. A simplified network that models this assumption is the lattice network, where nodes are placed on a regular grid and edges are present between adjacent individuals. Lattices possess high clustering (in terms of many short loops, although no triangles are present), two neighbours of a node are also likely to be neighbours of each other, and long path lengths, it takes many steps to move between two randomly chosen individuals. These features have a profound effect on the spread of an infectious disease [21, 119]. The high clustering has a strongly saturating effect, reducing the number of susceptible contacts each infected individual has. This slows down the transmission of infection, compared to the configuration models, as infection generally spreads in a wave-like manner.

Many observed social networks have been found to have what is called the 'small world' property. Such networks are characterized by short path lengths, it takes few steps to move between two randomly chosen individuals, and high clustering, two neighbours of a node are also likely to be neighbours of each other. Small world networks can be constructed from a regular lattice network by either randomly selecting two edges in the lattice and swapping their ends, thus preserving the degree of each node [251], or randomly selecting two nodes and connecting them [206]. These random edges provide long-range links which can allow an infection to jump to a new part of the network which may have a higher density of susceptible individuals. However, the transmission of infection will remain predominantly localized if long-range links are rare and saturation effects are observed [222]. Several authors have studied how local clustering and long-range contact influences the spread of infectious disease and vaccination strategies [27, 54, 154, 248].

Scale free. Studies of sexual networks [171] have found that the degree distributions are highly heterogeneous. This can be modelled by a class of networks known as scale-free networks, where most nodes have a small number of edges but a significant number of nodes have a large number of edges. These types of network can be formulated as configuration models from a given heterogeneous degree, however often they are formed by a dynamic process in which new nodes are

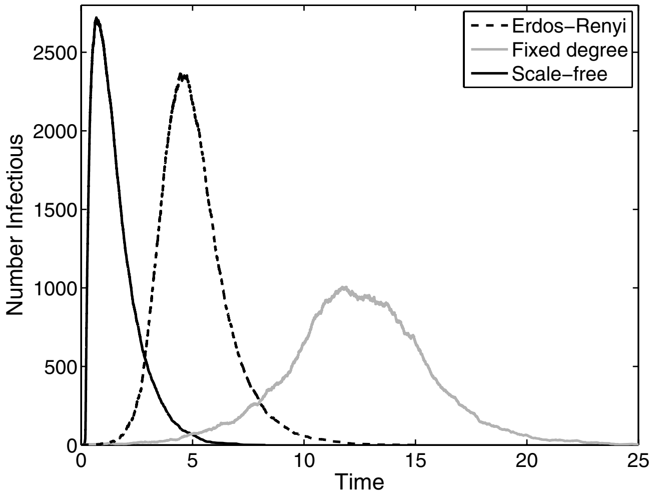

Figure 36. Examples of stochastic SIR-type simulations on three network types: Erdös–Rényi networks with Poisson distributed degree, Homogeneous networks where all nodes have equal degree, and scale-free networks where the degree distribution follows a heavy-tailed power law. In all three networks the mean degree is 3, the rate of transmission across a link is $\tau = 1.5$, the recovery rate $\gamma = 1$, while the population is ten thousand nodes; therefore all differences are due to heterogeneity in network degree. The Erdös-Rényi outbreak agrees with standard mean-field stochastic epidemic models, from which it is clear that network structure can either promote or retard transmission.

sequentially added to the network [32]. In these models new nodes preferentially connect to existing nodes in the network that already have a high number of contacts. This corresponds to the idea that individuals want to be friends with the most popular people. It should be noted that different methods of generating these scale-free networks are likely to lead to different higher order structures within the network.

Epidemics that spread across a truly scale-free networks lack a critical threshold in transmissibility of the infectious pathogen below which a significant outbreak cannot occur [212]. This is due to the possibility of arbitrarily highly connected nodes acting as epidemic amplifiers even when the rate of transmission along any edge is extremely slow (figure 36). On the other hand the existence of very highly connected nodes guides intervention based upon knowledge of network topology. It has been found that scale-free networks are very sensitive to the strategic removal of nodes [3, 126]. It is only necessary to remove a small number of the highly connected nodes to break the network into separate components, provided the mixing in the network is not assortative [204]. This insight could be very useful for designing control strategies, where selective targeting of control measures to dramatically impact of the potential for an infection to spread.

5.5.2. Realistic networks. Obtaining data on network connections that are epidemiologically relevant is a time-consuming and intrusive process. For STIs this requires asking individuals some highly personal information, although this information has been successfully gathered [138, 171, 254] and is often routinely collected as part of control measures. For social contacts, which can lead to the spread of airborne infections, similar personal issues can also arise although there is the added difficulty of trying to recall all the people we met over a given period.

Several approaches have been utilized to gather network data. Snowball sampling [107] uses the network structure of the population by getting individuals to name all their contacts, who are then traced and the process is repeated. However, there are problems with this approach: the data will form a single connected component and so large parts of the network could be missed; it may be difficult to get people to disclose sensitive information; respondents recall might not be perfect; it can be difficult to trace all contacts; and the sample will be biased towards individuals that like to make themselves known. Effectively snowball sampling is an adaptation of contact tracing, where links from all individuals (not just infected ones) are traced.

Another way to obtain information about the structure of the network is through the use of contact surveys [71, 198]. Contact surveys provide detailed local network (egocentric) data that can in principle be used to infer the larger network structure of the population. The data typically consists of information on a number of individuals (egos) and their contacts. The egos can then be connected in a similar way to generating a configuration network [193] to produce a realistic network for use in simulations. Alternatively, the egocentric data can be used to construct the who-acquires-infection-from-whom matrices for modelling purposes [198, 218].

A third way to obtain the transmission network is to use movement data, for example airline data [114, 133] or livestock movement data [190]. These data sources have the advantage that they are automatically collected in bulk, but the disadvantage that the network links sub-populations or groups of hosts rather than being a network between individuals. In addition, while it is clear that movements are often a primary mean of transmission, especially between livestock farms [102] alternative routes of transmission exist that are not captured by the movement network.

5.6. Exact results for particular networks

Many network properties are highly sensitive to the exact structure of the network, such that the presence/absence of one connection can dramatically change the values of any given property. It is therefore not surprising that a general relationship between the network configuration and the associated epidemic dynamics is still illusive. Therefore, there is considerable interest in trying to determine a low-dimensional set of network properties that provide quantitative or even qualitative understanding of the epidemic dynamics [17, 148, 186]. However, in the vast majority of cases we are constrained to having exact results for some specific network structures which provide general intuition.

5.6.1. Basic reproductive ratio. For large networks that can be formed from the configuration model [192] (including the Erdös-Rényi random graph [83]), the basic reproductive ratio can be explicitly calculated, in the limit that the network size tend to infinity [76]. In the early stages of an epidemic, the chance of an individual being infected is proportional to its degree. Given the random nature of the network, nodes with higher degree have a proportionally higher force of infection as each inward connection carries the same risk of transmission. For an infected individual of degree k, there are a potential $k-1$ contacts that it can infect, given that one of its contacts must have been its source of infection. For each of these contacts, the probability of transmission along an connections is $\tau/(\tau+\gamma)$—that is the chance of transmitting before recovery. Taking the average of all these quantities gives:

$$\begin{aligned} R_0 &= \frac{\tau}{\tau+\gamma}\,\frac{\mathrm{mean}(k(k-1))}{\mathrm{mean}(k)} \\ &= \frac{\tau}{\tau+\gamma}\left[\mathrm{mean}(k)+\frac{\mathrm{var}(k)}{\mathrm{mean}(k)}-1\right], \end{aligned}$$

where the mean(k) in the denominator is a normalizing factor and ensures that the probability of the infected individual being degree k sums to one. From this formulation it is clear that network structure has three main influences on early transmission dynamics: the mean degree is clearly central to the amount of onwards transmission; in addition variation about this mean also acts to increase transmission; finally the discrete structure of the network reduces transmission as one of the node's contacts must be its source of infection. The balance of these two conflicting influences (variation and the discrete number of contacts) determines how network structure will modify early dynamics. For networks with homogeneous degree (no variation) the removal of one contact can dramatically slow transmission [143]; in contrast for scale-free networks the variance in degree distribution dominates and R_0 can be infinite for all non-zero transmission rates (figure 36).

Through a more involved eigenvalue approach, similar calculations can be made when the network has degree correlations but is otherwise random with no additional structure [31]; this clearly mimics the risk-structured analysis of [75] with degree playing the role of risk structure. For more general networks, the calculation of R_0 is complicated by the presence of multiple short loops such that potential chains of transmission overlap and interfere. In such cases we must rely on numerical simulation methods that account for the definition of R_0 [131].

5.6.2. *Correspondence to mean-field models.*

For the Erdös–Rényi random graph [83], where the degree distribution is Poisson (with mean c), there exists a exact dynamic match to standard (mean-field) stochastic epidemic models (of the type explained section 4) as the number of nodes becomes large [33, 73]. In the standard stochastic models, the number of new infections per infected node is Poisson with degree β/γ which gives a exact match to the Erdös–Rényi random graph where the average number of new infections is Poisson with degree $\tau c/(\tau+\gamma)$. More over because of the random nature of the network depletion of susceptibles has an uncorrelated impact of each connection, so the match holds throughout the dynamics.

5.6.3. *Final epidemic size.*

When dealing with configuration networks with fixed homogeneous degree k, branching process theory allows us to determine the final epidemic size [73]:

$$R_\infty = 1-\left(1-\frac{\tau}{\tau+\gamma}\left[1-(1-R_\infty)^{\frac{k-1}{k}}\right]\right)^k.$$

Clearly this has many parallels with the standard form from [155], and tends to this standard result when as $k \to \infty$ but R_0 is kept finite.

5.6.4. *Percolation theory.*

Without doubt, percolation theory has had a huge impact on our understanding of potential dynamics and in particular the final size of epidemics on networks [205, 203]. In particular, percolation theory tells us about the probability of retaining a large connected component when each link in the network is deleted with probability $1-p$ [112]. There is therefore an intuitive link to epidemic models where p is considered as the probability of transmission across a link. Unfortunately there is a mis-match between the assumptions underlying percolation and epidemic models. In percolation models the probabilities associated with links are independent and this is crucial to many of the results. In epidemic models the probabilities of transmission from a node are frequently correlated as they will all depend on how long the node remains infected. Therefore only when all infected individuals remain infectious for a fixed period of time due the percolation theory results hold, although they provide an important bench-mark for spread on networks.

Unfortunately, very few exact results are known in the epidemiologically interesting case where networks are highly clustered, such that short loops are common. In these cases we generally rely on simulation results or approximations [131, 148]. Here the key questions are not simply how clustering changes the main epidemic features for constant parameter values, but how one epidemic feature varies as other features are held constant (as if fitted to observable data).

5.6.5. *Approximating network dynamics.*

Mean-field random-mixing models do not allow for higher order features of the transmission network, such as loops, correlations between cases and localized depletion of susceptibles. These features can have a significant impact on the dynamics of the disease spread. A natural extension of the earlier models is therefore to consider partnerships in the network (and the state of the two nodes involved) as the fundamental variables [78, 143, 163]. We will refer to these models as pairwise models.

Pairwise models offer a compromise between the random-mixing model and the explicit modelling the network in its entirety. We will illustrate the methodology with an SIS model as appropriate for STIs and assuming an equal degree k for all nodes; however it is straight forward to extend to more complex transmission and also to further stratify the susceptible and infectious classes in to different species/groups [78]. Denoting $[S]$ and $[I]$ as the number of susceptible and infectious individuals in the population, respectively, the SIS disease process can be represented by the following exact but unclosed differential equations

$$\begin{aligned}\frac{\mathrm{d}[S]}{\mathrm{d}t} &= -\tau[SI] + g[I],\\ \frac{\mathrm{d}[I]}{\mathrm{d}t} &= \tau[SI] - g[I].\end{aligned} \tag{105}$$

This model is exact, but to close the system, knowledge of $[SI]$, the number of partnerships between susceptible and infectious individuals, is required.

First order. Closing at the first order ignores pair level correlations and assumes that the number of pairs of a certain type is given by the independent product of the two constituent nodes:

$$[AB] \approx \frac{k[A][B]}{N}. \tag{106}$$

Substituting this approximation into the unclosed system 105 recovers the standard mean-field SIS equations derived in section 2.1, with $\beta = k\tau$.

Closing at second order. A better approximation to the epidemic behaviour on a network can be made by considering the time evolution of pairs. The number of susceptible-susceptible, susceptible-infectious and infectious-infectious pairs in the population can change due to infection within the pair, infection from outside the pair or recovery. This is captured by the following set of differential equations

$$\begin{aligned}\frac{\mathrm{d}[SS]}{\mathrm{d}t} &= -2\tau[SSI] + 2g[SI],\\ \frac{\mathrm{d}[SI]}{\mathrm{d}t} &= \tau([SSI] - [ISI] - [SI]) + g([II] - [SI]),\\ \frac{\mathrm{d}[SS]}{\mathrm{d}t} &= 2\tau([ISI] + [SI]) - 2g[II],\end{aligned} \tag{107}$$

where $[ABC]$ denotes three connected individuals with disease status A, B and C respectively. If we assume that the network is unclustered, such that there are no connections between A and C, then it is possible to close the system by approximating triples as two independent pairs that share a common node, leading to,

$$[ABC] \approx \frac{k-1}{k}\frac{[AB][BC]}{[B]}. \tag{108}$$

Although this triple approximation breaks down for the SIS model due to the strong correlations that develop between the ends of a triple that are not captured in this approximation, for the SIR model this triple approximation allows an exact formulation of the dynamics of infection [132]. Pairwise models closed at second order, have been used successfully in a number of situations, from the evolution of pathogen virulence [54], to the spread and control of STIs [79, 91], to the local spread of livestock infections [90].

In principle equations can be formulated for triples, with closure then requires to approximate the number of four-node states, however the complexity is close to overwhelming [129]. For example, with triples there are only two types that need consideration (triangles and open triples) however four nodes can be in six different interconnected configurations. Although the resulting triplewise equations equal or outperform the pairwise models in all situations considered [129], it is not clear that the increase in accuracy merits the additional effort.

More recently versions of these pairwise approximations have been produced that naturally account for degree heterogeneity at the cost of restricting attention to models of the SIR-type formulation [188, 189, 247]. Following the notation of [189], for a configuration-type network the epidemic dynamics are captured by the equation:

$$\frac{\mathrm{d}\theta}{\mathrm{d}t} = -\beta\theta + \beta\frac{\phi'(\theta)}{\phi'(1)} + \gamma(1-\theta), \tag{109}$$

where θ captures the total force of infection across an average link, and ϕ is the probability generating function of the degree distribution ($\phi(x) = \sum_k P(k)x^k$). Using this single variable θ the level of susceptibles over time can be determined:

$$S(t) = \phi(\theta(t))$$

and hence $I(t)$ can be found from the rate of change of S. This novel formulation allows for explicit calculation of many quantities of interest, such as early growth rates and final epidemic sizes [188], but is constrained to SIR models.

6. Future challenges

We hope this review has illustrated the wealth of problems and solutions that have been developed in the quantitative study of epidemiology. While much of this work is motivated by practical issues of public health, there are also a number of theoretical questions that are both scientifically interesting, but also would have wider impact. Here we outline a few future challenges in the hope that it generates thoughts and discussion.

Basic models. For even the simplest models there are still many unsolved problems; a large number reflect our lack of quantitative understanding of the basic biology but other are of a more technical nature. As an example of the former, it is still unclear the exact impact of vaccines in terms of the reduction in susceptibility and onward transmission that they afford, consequently it is unclear what could and should be measured to ascertain the levels of control offered by a vaccine. As an example of the latter, we hypothesized in section 2.4 that dividing the exposed and infectious class into two or three sub-compartments was generally viewed as sufficient, yet there is no clear understanding of how this division should be optimally achieved or parametrized to match available data. Finally, there is the question of how the wealth of new bio-medical data, including genetic and immunological measurements, should be incorporated into epidemiological modelling.

Heterogeneous populations. When the population is subdivided, there naturally arises the question of how the who acquires infection from whom matrix should be parametrized. When only endemic prevalences are available, then it is clear that there are more degrees of freedom than data and hence there is some flexibility in how parameters are assigned; however if we can observe fluctuations away from this endemic state then are there methods of calculating all the terms in the matrix? There is also the question of how many classes should a population be split into, for some such as age structure there are obvious groups while for others the available data may set the appropriate scales, however in general a population can always be subdivided along ever finer heterogeneities. When we consider the specific case studies of HIV and vector-borne diseases there are also the obvious practical issues of whether models can inform about key points in the transmission mechanisms where interventions will have maximum benefit, in particular the optimal targeting of control measures across multiple risk groups.

Stochastic transmission. Stochastic dynamics differ from their deterministic counterparts in terms of the variability in epidemic outcome and the risks of stochastic extinction; both of these would merit further investigation. We have already seen how a sudden start to a vaccination programmes can give rise to the 'honeymoon effect' with large oscillations; are there ways in which vaccination programmes should be designed to maximise this effect and the risk of subsequent stochastic extinction, hence leading to elimination of infection. Much of section 4 was focused toward understanding stochastic epidemics without the need for extensive simulations, this ideas can obviously be extended far further. There are clear questions as how large a population can be that will allow Kolmorgorov equations to be numerically integrated, and the type of additional understanding that can be derived from these machine precision models. There is also the issue over when the large-population size diffusion approximations are appropriate, and whether these can provide a meaningful assessment of the degree of variability we should expect to observe.

Spatial heterogeneity. The modelling of spatial epidemics is a relatively new field, and so there are a host of open questions. In the ever expanding field of networks, there are fundamental questions about the types of network structure that most reliably capture human social and sexual networks, and how the strength and dynamic nature of these connections should be incorporated. In addition, demographic processes of birth and death are largely ignored in network models, how to include these without changing the underlying network structure is another open question. When discussion foot-and-mouth disease we showed through simulation that there was an optimal ring-size for vaccination, but as yet there have been few attempts to consider the optimal spatial deployment of resources in a general context. Finally, while pairwise models provide a reliable and efficient means of approximating SIR-type infection on simple networks, these methods need extending to more complex network structures and SIS-type dynamics.

Model sufficiency. One overriding question that covers all aspects of epidemiological modelling is that of model sufficiency; how do we know which aspect of model structure we need to include in our formulation to generate accurate and reliable predictions. Clearly we are limited by computational resources, data, biological information and ultimately enthusiasm for generating ever more complex models. However the question remains, how do we know that our models are fit for purpose, are there guiding rules for when different types of heterogeneity are likely to play pivotal roles.

Capturing observed behaviour. One of the ultimate tests of any mathematical model is how well it captures the observed behaviour. In many scientific fields the is the possibility of matching models to experiments, where conditions are tightly controlled. In epidemiology this is rarely the case, there are often many confounding factors and the observed behaviour is seldom an accurate reflection of reality. Therefore there is a vital need for refined statistical techniques that can cope with noisy transmission processes, a low and possibly variable reporting rate of infection, and still allow for parameter inference.

Acknowledgments

The authors would like to thank all those in the WIDER (Warwick Infectious Disease Epidemiology Research) Centre for their help and support, in particular Lorenzo Pellis for his detailed comments. This work was funded by EPSRC (MJK, KR, SB and JM).

References

[1] Abramson P R and Rothschild B 1988 Sex, drugs and matrices: mathematical prediction of HIV infection *J. Sex Res.* **25** 106–22

[2] Albert R and Barabási A L 2002 Statistical mechanics of complex networks *Rev. Mod. Phys.* **74** 47–97

[3] Albert R, Jeong H and Barabási A-L 2000 Error and attack tolerance of complex networks *Nature* **406** 378–82

[4] Alexander M E, Moghadas S M, Rohani P and Summers A R 2006 Modelling the effect of a booster vaccination on disease epidemiology *J. Math. Biol.* **52** 290–306

[5] Alonso D, McKane A J and Pascual M 2007 Stochastic amplification in epidemics *J. R. Soc. Interface* **4** 575

[6] Aly S S, Anderson R J, Whitlock R H, Fyock T L, McAdams S C, Byrem T M, Jiang J, Adaska J M and Gardner I A 2012 Cost-effectiveness of diagnostic strategies to identify mycobacterium avium subspecies paratuberculosis super-shedder cows in a large dairy herd using antibody enzyme-linked immunosorbent assays, quantitative real-time polymerase chain reaction, and bacterial culture *J. Veterinary Diagnostic Investigation* **24** 821–32

[7] Anderson R M, Jackson H C, May R M and Smith A M 1981 Population dynamics of fox rabies in Europe *Nature* **289** 765–71

[8] Anderson R M and May R M 1983 Vaccination against rubella and measles: quantitative investigations of different policies *J. Hygiene* **90** 259–325

[9] Anderson R M, Blythe S P, Gupta S, Konings E, Anderson R M, Blythe S P, Gupta S and Konings E 1989 The transmission dynamics of the human immunodeficiency virus type 1 in the male homosexual community in the United Kingdom: the influence of changes in sexual behaviour *Phil. Trans. R. Soc. Lond.* B **325** 45–98

[10] Anderson R M, Gupta S and Ng W 1990 The significance of sexual partner contact networks for the transmission dynamics of HIV *J. Acquired Immune Deficiency Syndromes* **3** 417–29

[11] Anderson R M, May R M, Boily M C, Garnett G P and Rowley J T 1991 The spread of HIV-1 in Africa: sexual contact patterns and the predicted demographic impact of AIDS *Nature* **352** 581–9

[12] Anderson R M, May R M and McLean A R 1988 Possible demographic consequences of AIDS in developing countries *Nature* **332** 228–34

[13] Anderson R M, May R M, Ng T W, Rowley J T, Anderson R M, May R M, Ng T W and Rowley J T 1992 Age-dependent choice of sexual partners and the transmission dynamics of HIV in sub-Saharan Africa *Phil. Trans. R. Soc. Lond.* B **336** 135–55

[14] Anderson R M, Medley G F, May R M and Johnson A M 1986 A preliminary study of the transmission dynamics of the human immunodeficiency virus (HIV), the causative agent of AIDS *IMA J. Math. Appl. Med. Biol.* **3** 229–63

[15] Anderson R M and May R M 1979 Population biology of infectious diseases: I *Nature* **280** 361–6

[16] Anderson R M and May R M 1992 *Infectious Diseases of Humans* (Oxford: Oxford University Press)

[17] Aparicio J P and Pascual M 2007 Building epidemiological models from R0 an implicit treatment of transmission in networks *Proc. R. Soc.* B **274** 505

[18] Baguelin M, Van Hoek A J, Jit M, Flasche S, White P J and Edmunds W J 2010 Vaccination against pandemic influenza A/H1N1v in England: a real-time economic evaluation *Vaccine* **28** 2370–84

[19] Bailey N T J 1950 A simple stochastic epidemic *Biometrika* **37** 193–202

[20] Bailey N T J 1975 *The Mathematical Theory of Infectious Diseases and its Applications* 2nd edn (London: Charles Griffin & Co Ltd)

[21] Bak P, Chen K and Tang C 1990 A forest-fire model and some thoughts on turbulence *Phys. Lett.* A **147** 297–300

[22] Balcan D, Colizza V, Gonçalves B, Hu H, Ramasco J J and Vespignani A 2009 Multiscale mobility networks and the spatial spreading of infectious diseases *Proc. Natl Acad. Sci.* **106** 21484–9

[23] Ball F 1983 The threshold behaviour of epidemic models *J. Appl. Probab.* **20** 227–41

[24] Ball F 1986 A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models *Adv. Appl. Probab.* **18** 289–310

[25] Ball F and Clancy D 1993 The final size and severity of a generalised stochastic multitype epidemic model *Adv. Appl. Probab.* **25** 721–36

[26] Ball F and Lyne O D 2001 Stochastic multi-type SIR epidemics among a population partitioned into households *Adv. Appl. Probab.* **33** 99–123

[27] Ball F, Mollison D and Scalia-Tomba G 1997 Epidemics with two levels of mixing *Ann. Appl. Probab.* **7** 46–89

[28] Ball F, Sirl D and Trapman P 2010 Analysis of a stochastic SIR epidemic on a random network incorporating household structure *Math. Biosci.* **224** 53–73

[29] Ball F G and Lyne O D 2002 Optimal vaccination policies for stochastic epidemics among a population of households *Math. Biosci.* **177** 333–54

[30] Ball F, Britton T and Lyne O 2004 Stochastic multitype epidemics in a community of households: estimation and form of optimal vaccination schemes *Math. Biosci.* **191** 19–40

[31] Ball F, Britton T and Sirl D 2012 A network with tunable clustering, degree correlation and degree distribution, and an epidemic thereon *J. Math. Biol.* **66** 979–1019

[32] Barabasi A L and Albert R 1999 Emergence of scaling in random networks *Science* **286** 509–12

[33] Barbour A and Mollison D 1990 Epidemics and random graphs *Stochastic Processes in Epidemic Theory* ed J-P Gabriel *et al* (Berlin: Springer) pp 86–9

[34] Bartlett M S 1956 Deterministic and stochastic models for recurrent epidemics *Proc. 3rd Berkeley Symp. Math. Stat. Probab.* **4** 81–108

[35] Bartlett M S 1957 Measles periodicity and community size *J. R. Stat. Soc.* A **120** 48–70

[36] Bartlett M S 1960 *Stochastic Population Models in Ecology and Epidemiology* (New York: Wiley)

[37] Bartlett M S 1960 The critical community size for measles in the United States *J. R. Stat. Soc.* A **123** 37–44

[38] Basta N E, Halloran M E, Matrajt L and Longini I M 2008 Estimating influenza vaccine efficacy from challenge and community-based study data *Am. J. Epidemiol.* **168** 1343–52

[39] Bastos A D S, Boshoff C I, Keet D F, Bengis R G and Thomson G R 2000 Natural transmission of Foot-and-Mouth disease virus between African buffalo (*syncerus caffer*) and Impala (*aepyceros melampus*) in the Kruger National Park, South Africa *Epidemiol. Infection* **124** 591–8

[40] Becker N G 1989 *Analysis of Infectious Disease Data* (London: Chapman and Hall) 1st edn

[41] Bellan S E 2010 The importance of age dependent mortality and the extrinsic incubation period in models of mosquito-borne disease transmission and control *PLoS ONE* **5** e10165

[42] Bjørnstad O N, Finkenstädt B F and Grenfell B T 2002 Dynamics of measles epidemics: estimating scaling of transmission rates using a time series SIR model *Ecol. Monogr.* **72** 169–84

[43] Black A J, House T, Keeling M J and Ross J V 2013 Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza *J. R. Soc. Interface* **10** 20121019

[44] Black A J and McKane A J 2012 Stochastic formulation of ecological models and their applications *Trends Ecol. Evol.* **27** 337–45

[45] Blundell S J and Blundell K M 2012 *Concepts in Thermal Physics* (Oxford: Oxford University Press)

[46] Blythe S P and Anderson R M 1988 Distributed incubation and infectious periods in models of the transmission dynamics of the human immunodeficiency virus (HIV) *Math. Med. Biol.* **5** 1–19

[47] Blythe S P and Anderson R M 1988 Heterogeneous sexual activity models of HIV transmission in male homosexual populations *Math. Med. Biol.* **5** 237–60

[48] Blythe S P and Anderson R M 1988 Variable infectiousness in HFV transmission models *Math. Med. Biol.* **5** 181–200

[49] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 Complex networks: structure and dynamics *Phys. Rep.* **424** 175–308

[50] Bolker B and Grenfell B 1995 Space, persistence and dynamics of measles epidemics *Phil. Trans. R. Soc. Lond.* B **348** 309–20

[51] Bolker B M 1999 Analytic models for the patchy spread of plant disease *Bull. Math. Biol.* **61** 849–74

[52] Bolker B M and Grenfell B T 1993 Chaos and biological complexity in measles dynamics *Proc. R. Soc Lond.* B **251** 75–81

[53] Boltyanskii R V 1993 Chaos and complexity in measles models—a comparative numerical study *IMA J. Math. Appl. Med. Biol.* **10** 83–95

[54] Boots M and Sasaki A 1999 Small worlds and the evolution of virulence: infection occurs locally and at a distance *Proc. R. Soc. Lond.* B **266** 1933–8

[55] Britton N F 2003 *Essential Mathematical Biology* (Berlin: Springer)

[56] Broberg E, Nicoll A and Amato-Gauci A 2011 Seroprevalence to influenza A(H1N1) 2009 virus—where are we? *Clin. Vaccine Immunology : CVI* **18** 1205–12

[57] Brochier B, Kieny M P, Costy F, Coppens P, Bauduin B, Lecocq J P, Languet B, Chappuis G, Desmettre P and Afiademanyo K 1991 Large-scale eradication of rabies using recombinant vaccinia-rabies vaccine *Nature* **354** 520–2

[58] Brookmeyer R and Gail M H 1988 A method for obtaining short-term projections and lower bounds on the size of the AIDS epidemic *J. Am. Stat. Assoc.* **83** 301–8

[59] Brown J K M and Hovmøller M S 2002 Aerial dispersal of pathogens on the global and continental scales and its impact on plant disease *Science* **297** 537–41

[60] Cairns B J and Pollett P K 2005 Approximating persistence in a general class of population processes *Theor. Population Biol.* **68** 77–90

[61] du Preez L H, Hyatt A D, Muller R, Speare R and Weldon C 2004 Origin of the amphibian chytrid fungus *Emerging Infectious Diseases* **10** 2100

[62] Conlan A J K, Pej R, Lloyd Andrew L, Keeling Matt J and Grenfell Bryan T 2010 Resolving the impact of waiting time distributions on the persistence of measles *J. R. Soc. Interface* **7** 623–40

[63] Conlan A J K, Rohani P, Lloyd A L, Keeling M and Grenfell B T 2010 Resolving the impact of waiting time distributions on the persistence of measles *J. R. Soc. Interface* **7** 623–40

[64] Cook A R, Gibson G J, Gottwald T R and Gilligan C A 2008 Constructing the effect of alternative intervention strategies on historic epidemics *J. R. Soc. Interface* **5** 1203–13

[65] Cooper B S, Pitman R J, Edmunds W J and Gay N J 2006 Delaying the international spread of pandemic influenza *PLoS Med.* **3** e212

[66] Cornet S, Nicot A, Rivero A and Gandon S 2012 Malaria infection increases bird attractiveness to uninfected mosquitoes *Ecol. Lett.* **16** 323–9

[67] Cox D R and Isham V 1980 *Point Processes* (London: Chapman and Hall)

[68] Cox D R and Medley G F 1989 A process of events with notification delay and the forecasting of AIDS *Phil. Trans. R. Soc. Lond.* B **325** 135–45

[69] Danon L, Ford A P, House T, Jewell C P, Keeling M J, Roberts G O, Ross J V and Vernon M C 2011 Networks and the epidemiology of infectious disease *Interdisciplinary Perspectives on Infectious Diseases* **2011** 284909

[70] Danon L, House T and Keeling M J 2009 The role of routine versus random movements on the spread of disease in Great Britain *Epidemics* **1** 250–8

[71] Danon L, House T A, Read J M and Keeling M J 2012 Social encounter networks: collective properties and disease transmission *J. R. Soc. Interface* **9** 2826–33

[72] Deardon R, Brooks S, Grenfell B T, Keeling M J, Tildesley M J, Savill N J, Shaw D J and Woolhouse M E J 2010 Inference for individual-level models of infectious diseases in large populations *Stat. Sin.* **20** 239–61

[73] Diekmann O, De Jong M C M and Metz J A J 1998 A deterministic epidemic model taking account of repeated contacts between the same individuals *J. Appl. Probab.* **35** 462–8

[74] Diekmann O, Heesterbeek H and Britton T 2012 *Mathematical Tools for Understanding Infectious Disease Dynamics* (Princeton, NJ: Princeton University Press)

[75] Diekmann O, Heesterbeek J A and Metz J A 1990 On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations *J. Math. Biol.* **28** 365–82

[76] Diekmann O and Heesterbeek J A P 2000 *Mathematical Epidemiology of Infectious Diseases: Model Building, Analysis, and Interpretation* (New York: Wiley)

[77] Diekmann O 1978 Thresholds and travelling waves for the geographical spread of infection *J. Math. Biol.* **6** 109–30

[78] Eames K T D and Keeling M J 2002 Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases *Proc. Natl Acad. Sci. USA* **99** 13330–5

[79] Eames K T D and Keeling M J 2003 Contact tracing and disease control *Proc. R. Soc. Lond.* B **270** 2565–71

[80] Earn D J 2000 A simple model for complex dynamical transitions in epidemics *Science* **287** 667–70

[81] Eaton J W, Hallett T B and Garnett G P 2011 Concurrent sexual partnerships and primary HIV infection: a critical interaction *AIDS Behav.* **15** 687–92

[82] Epstein H and Morris M 2011 Concurrent partnerships and HIV: an inconvenient truth *J. Int. AIDS Soc.* **14** 13

[83] Erdős P and Renyi A 1960 Graphs with prescribed degrees of vertices *Publ. Math. Inst. Hung. Acad. Sci.* **5** 17–61 (in Hungarian)

[84] Ethier S and Kurtz T 1986 *Markov Processes: Characterization and Convergence* (New York: Wiley)

[85] Eubank S, Guclu H, Kumar V S A, Marathe M V, Srinivasan A, Toroczkai Z and Wang N 2004 Modelling disease outbreaks in realistic urban social networks *Nature* **429** 180–4

[86] Evans J D, Saegerman C, Mullin C and Haubruge E 2009 Colony collapse disorder: a descriptive study *PLoS One* **4** e6481

[87] Evans L C 2010 *Partial Differential Equations* 2nd edn (Providene, RI: American Mathematical Society)

[88] Ferguson N M, Cummings D A T, Cauchemez S, Fraser C, Riley S, Meeyai A, Iamsirithaworn S and Burke D S 2005 Strategies for containing an emerging influenza pandemic in Southeast Asia *Nature* **437** 209–14

[89] Ferguson N M, Donnelly C A and Anderson R M 2001 The Foot-and-Mouth epidemic in Great Britain: pattern of spread and impact of interventions *Science* **292** 1155–60

[90] Ferguson N M, Donnelly C A and Anderson R M 2001 Transmission intensity and impact of control policies on the Foot-and-Mouth epidemic in Great Britain *Nature* **413** 542–8

[91] Ferguson N M and Garnett G P 2000 More realistic models of sexually transmitted disease transmission dynamics: sexual partnership networks, pair models, and moment closure *Sexually Transmitted Diseases* **27** 600–9

[92] Ferguson N M, Keeling M J, Edmunds W J, Gant R, Grenfell B T, Amderson R M and Leach S 2003 Planning for smallpox outbreaks *Nature* **425** 681–5

[93] Ferguson N M, Cummings D A T, Fraser C, Cajka J C, Cooley P C and Burke D S 2006 Strategies for mitigating an influenza pandemic *Nature* **442** 448–52

[94] Filipe J A N, Cobb R C, Meentemeyer R K, Lee C A, Valachovic Y S, Cook A R, Rizzo D M and Gilligan C A 2012 Landscape epidemiology and control of pathogens with cryptic and long-distance dispersal: sudden oak death in northern Californian forests *PLoS Comput. Biol.* **8** e1002328

[95] Finkenstädt B F and Grenfell B T 2000 Time series modelling of childhood diseases: a dynamical systems approach *J. R. Stat. Soc.* C **49** 187–205

[96] Fleming W and Soner H M 2006 *Controlled Markov Processes and Viscosity Solutions* (Berlin: Springer)

[97] Fraser C 2007 Estimating individual and household reproduction numbers in an emerging epidemic *PLoS One* **2** 758

[98] Freidlin M I and Wentzell A D 1998 *Random Perturbations of Dynamical Systems* 2nd edn (Berlin: Springer)

[99] Fyock T, Whitlock R and Sweeney R 2009 Some Johne's cows do more harm than others *Hoard's Dairyman* **10** 615

[100] Ghani A C and Garnett G P 2000 Risks of acquiring and transmitting sexually transmitted diseases in sexual partner networks *Sexually Transmitted Diseases* **27** 579–87

[101] Gibson G J, Otten W, Filipe J A N, Cook A, Marion G and Gilligan C A 2006 Bayesian estimation for percolation models of disease spread in plant populations *Stat. Comput.* **16** 391–402

[102] Gilbert M, Mitchell A, Bourn D, Mawdsley J, Clifton-Hadley R and Wint W 2005 Cattle movements and bovine tuberculosis in Great Britain *Nature* **435** 491–6

[103] Gillespie D T 1977 Exact stochastic simulation of coupled chemical reactions *J. Phys. Chem.* **81** 2340–61

[104] Gillespie D T 2001 Approximate accelerated stochastic simulation of chemically reacting systems *J. Chem. Phys.* **115** 1716–33

[105] Gillespie D T and Petzold L R 2003 Improved leap-size selection for accelerated stochastic simulation *J. Chem. Phys.* **119** 8229–34

[106] Glendinning P 1994 *Stability, Instability, and Chaos: An Introduction to the Theory of Nonlinear Differential Equations* (Cambridge: Cambridge University Press)

[107] Goodman L A 1961 Snowball sampling *Ann. Math. Statist.* **32** 148–70

[108] Grenfell B 1997 (Meta)population dynamics of infectious diseases *Trends Ecol. Evol.* **12** 395–9

[109] Grenfell B, Bjørnstad O and Kappey J 2001 Travelling waves and spatial hierarchies in measles epidemics *Nature* **414** 716–23

[110] Grenfell B T and Anderson R M 1985 The estimation of age related rates of infection from case notifications and serological data *J. Hygiene* **95** 419–36

[111] Griffiths D A 1973 Multivariate birth-and-death processes as approximations to epidemic processes *J. Appl. Probab.* **10** 15–26

[112] Grimmett G 1999 *What is Percolation?* (Berlin: Springer)

[113] Grimmett G 2010 *Probability on Graphs: Random Processes on Graphs and Lattices* vol 1 1st edn (Cambridge: Cambridge University Press)

[114] Guimera R, Mossa S, Turtschi A and Nunes Amaral L A 2005 The worldwide air transportation network: anomalous centrality, community structure, and cities' global roles *Proc. Natl Acad. Sci.* **102** 7794–9

[115] Gullan P J 2005 *The Insects: an Outline of Entomology* (Chicago, IL: University of Chicago Press)

[116] Gupta S *et al* 1989 Networks of sexual contacts: implications for the pattern of spread of HIV *AIDS* **3** 807

[117] Halloran M E, Haber M and Longini I M 1992 Interpretation and estimation of vaccine efficacy under heterogeneity *Am. J. Epidemiol.* **136** 328–43

[118] Hanski I and Gaggiotti O E 2004 *Ecology, Evolution and Genetics of Metapopulations* (New York: Academic)

[119] Harris T E 1974 Contact interactions on a lattice *Ann. Probab.* **2** 969–88

[120] Harris T E 1963 *The Theory of Branching Processes* (Berlin: Springer)

[121] Hassell M P, Comins H N and May R M 1991 Spatial structure and chaos in insect population dynamics *Nature* **353** 255–8

[122] Healy M J R and Tillett H E 1988 Short-term extrapolation of the AIDS epidemic *J. R. Statist. Soc.* A **151** 50–65

[123] Heesterbeek J A P 2002 A brief history of $R_\circ$ and a recipe for its calculation *Acta Biotheor.* **50** 189–204

[124] Hollingsworth T D, Anderson R M and Fraser C 2008 HIV-1 transmission, by stage of infection *J. Infectious Diseases* **198** 687–93

[125] Hollingsworth T D, Ferguson N M and Anderson R M 2006 Will travel restrictions control the international spread of pandemic influenza? *Nature Med.* **12** 497–9

[126] Holme P, Kim B J, Yoon C N and Han S K 2002 Attack vulnerability of complex networks *Phys. Rev.* E **65** 056109

[127] House T and Keeling M J 2008 Household structure and infectious disease transmission *Epidemiol. Infection* **137** 654

[128] House T, Ross J V and Sirl D 2012 How big is an outbreak likely to be? Methods for epidemic final-size calculation *Proc. R. Soc.* A **469** 20120436

[129] House T, Davies G, Danon L and Keeling M J 2009 A motif-based approach to network epidemics *Bull. Math. Biol.* **71** 1693–706

[130] House T, Hall I, Danon L and Keeling M J 2010 Contingency planning for a deliberate release of smallpox in Great Britain—the role of geographical scale and contact structure *BMC Infectious Diseases* **10** 25

[131] House T and Keeling M J 2011 Epidemic prediction and control in clustered populations *J. Theor. Biol.* **272** 1–7

[132] House T and Keeling M J 2011 Insights from unifying modern approximations to infections on networks *J. R. Soc. Interface* **8** 67–73

[133] Hufnagel L, Brockmann D and Geisel T 2004 Forecast and control of epidemics in a globalized world *Proc. Natl Acad. Sci. USA* **101** 15124–9

[134] Hyman J M and Stanley E 1988 Using mathematical models to understand the AIDS epidemic *Math. Biosci.* **90** 415–73

[135] Isham V 2005 Stochastic models for epidemics: current issues and developments *Celebrating Statistics: Papers in Honour of Sir David Cox on his 80th Birthday* ed A C Davidson, Y Dodge and N Wermuth (Oxford: Oxford University Press) pp 27–54

[136] Jacquez J A, Simon C P, Koopman J, Sattenspiel L and Perry T 1988 Modeling and analyzing HIV transmission: the effect of contact patterns *Math. Biosci.* **92** 119–99

[137] Jalvingh A W, Nielen M, Maurice H, Stegeman A J, Elbers A R W and Dijkhuizen A A 1999 Spatial and stochastic simulation to evaluate the impact of events and control measures on the 1997–1998 classical swine fever epidemic in The Netherlands: I. Description of simulation model *Preventive Veterinary Med.* **42** 271–95

[138] Jolly A M, Muth S Q, Wylie J L and Potterat J J 2001 Sexual networks and sexually transmitted infections: a tale of two cities *J. Urban Health-Bull. New York Acad. Med.* **78** 433–45

[139] Kallenberg O 2002 *Foundations of Modern Probability* (Berlin: Springer)

[140] Kaplan E H and Lee Y S 1990 How bad can it get? bounding worst case endemic heterogenous mixing models of HIV/AIDS *Math. Biosci.* **99** 157–80

[141] Karatzas I and Shreve S E 1991 *Brownian Motion and Stochastic Calculus* (Berlin: Springer)

[142] Karon J M *et al* 1988 The projected incidence of AIDS and estimated prevalence of HIV infection in the United States *J. Acquired Immune Deficiency Syndromes* **1** 542

[143] Keeling M J 1999 The effects of local spatial structure on epidemiological invasions *Proc. R. Soc.* B **266** 859–67

[144] Keeling M J, Bjornstad O N and Grenfell B T 2004 Metapopulation dynamics of infectious diseases *Ecology, Genetics and Evolution of Metapopulations* ed I Hanski and O E Gaggiotti (Amsterdam: Elsevier)

[145] Keeling M J and Grenfell B T 1998 Effect of variability in infection period on the persistence and spatial spread of infectious diseases *Math. Biosci.* **147** 207–26

[146] Keeling M J and Ross J V 2008 On methods for studying stochastic disease dynamics *J. R. Soc. Interface* **5** 171–81

[147] Keeling M J, Woolhouse M E, Shaw D J, Matthews L, Chase-Topping M, Haydon D T, Cornell S J, Kappey J,

Wilesmith J and Grenfell B T 2001 Dynamics of the 2001 UK foot and mouth epidemic: stochastic dispersal in a heterogeneous landscape *Science* **294** 813–7

[148] Keeling M 2005 The implications of network structure for epidemic dynamics *Theor. Population Biol.* **67** 1–8

[149] Keeling M J, Danon L, Vernon M C and House T A 2010 Individual identity and movement networks for disease metapopulations *Proc. Natl Acad. Sci.* **107** 8866–70

[150] Keeling M J and Rohani P 2002 Estimating spatial coupling in epidemiological systems: a mechanistic approach *Ecol. Lett.* **5** 20–9

[151] Keeling M J and Rohani P 2008 *Modeling Infectious Diseases in Humans and Animals* (Princeton, NJ: Princeton University Press)

[152] Keeling M J and Grenfell B T 1997 Disease extinction and community size: modeling the persistence of measles *Science* **275** 65–7

[153] Keeling M J, Rohani P and Grenfell B T 2001 Seasonally forced disease dynamics explored as switching between attractors *Physica D—Nonlinear Phenom.* **148** 317–35

[154] Keeling M J 1999 The effects of local spatial structure on epidemiological invasions *Proc. R. Soc. Lond.* B **266** 859–67

[155] Kermack W O and McKendrick A G 1927 Contributions to the mathematical theory of epidemics: I *Proc. R. Soc.* A **115** 700–21

[156] King D 2007 Bovine tuberculosis in cattle and badgers: a report by the chief scientific adviser

[157] Kiss I Z, Green D M and Kao R R 2006 The network of sheep movements within Great Britain: network properties and their implications for infectious disease spread *J. R. Soc. Interface* **3** 669–77

[158] Klebaner F C 2005 *Introduction to Stochastic Calculus With Applications* 2nd edn (London: Imperial College Press)

[159] Klovdahl A S 1985 Social networks and the spread of infectious diseases: the AIDS example *Social Sci. Med.* **21** 1203–16

[160] Koelle K and Pascual M 2004 Disentangling extrinsic from intrinsic factors in disease dynamics: a nonlinear time series approach with an application to cholera *Am. Naturalist* **163** 901–13

[161] Kot M, Lewis M A and van den Driessche P 1996 Dispersal data and the spread of invading organisms *Ecology* **77** 2027–42

[162] Krebs J R 1997 *Bovine Tuberculosis in Cattle and Badgers* (London: Ministry of Agriculture, Fisheries and Food)

[163] Kretzschmar M and Dietz K 1998 The effect of pair formation and variable infectivity on the spread of an infection without recovery *Math. Biosci.* **148** 83–113

[164] Kurtz T 1970 Solutions of ordinary differential equations as limits of pure jump Markov processes *J. Appl. Probab.* **7** 49–58

[165] Kurtz T 1971 Limit theorems for sequences of jump Markov processes approximating ordinary differential processes *J. Appl. Probab.* **8** 344–56

[166] Feng J and Kurtz T G 2007 *Large Deviations for Stochastic Processes* (*Mathematical Survey & Monographs* N 131) (Providence, RI: American Mathematical Society)

[167] Kurtz T G 1971 Limit theorems for sequences of jump markov processes approximating ordinary differential processes *J. Appl. Probab.* **8** 344–56

[168] Kuske R, Gordillo L F and Greenwood P 2007 Sustained oscillations via coherence resonance in SIR *J. Theor. Biol.* **245** 459–69

[169] Levi T, Kilpatrick A M, Mangel M and Wilmers C C 2012 Deer, predator, and the emergence of Lyme disease *Proc. Natl Acad. Sci.* **109** 10942–7

[170] Levins R 1969 Some demographic and genetic consequences of environmental heterogeneity for biological control *Bull. Entomol. Soc. Am.* **15** 237–40

[171] Liljeros F, Edling C R, Amaral L A N, Stanley H E and Aberg Y 2001 The web of human sexual contacts *Nature* **411** 907–8

[172] Lipster R S and Shiryaev A N 2001 *Statistics of Random Processes I: General Theory* 2nd edn (Berlin: Springer)

[173] Liu Q X, Wang R H and Jin Z 2009 Persistence, extinction and spatio-temporal synchronization of SIRS spatial models *J. Stat. Mech.: Theory Exp.* **2009** P07007

[174] Lloyd A L 2001 Destabilization of epidemic models with the inclusion of realistic distributions of infectious periods *Proc. R. Soc.* B **268** 985–93

[175] Longini I M, Nizam A, Xu S F, Ungchusak K, Hanshaoworakul W, Cummings D A T and Halloran M E 2005 Containing pandemic influenza at the source *Science* **309** 1083–7

[176] MacArthur R H and Wilson E O 1967 *The Theory of Island Biogeography* (Princeton, NJ: Princeton University Press)

[177] Macdonald G 1957 *The Epidemiology and Control of Malaria* (Oxford: Oxford University Press)

[178] Mah T L and Halperin D T 2010 Concurrent sexual partnerships and the HIV epidemics in Africa: evidence to move forward *AIDS Behav.* **14** 11–6

[179] Marm Kilpatrick A, Daszak P, Jones M J, Marra P P and Kramer L D 2006 Host heterogeneity dominates West Nile virus transmission *Proc. R. Soc.* B **273** 2327–33

[180] May R M and Anderson R M 1979 Population biology of infectious diseases: II *Nature* **280** 455–61

[181] May R M 1988 The transmission dynamics of human immunodeficiency virus (HIV) *Phil. Trans. R. Soc. Lond.* B **321** 565–607

[182] McCaig C, Begon M, Norman R and Shankland C 2010 A symbolic investigation of superspreaders *Bull. Math. Biol.* **73** 777–94

[183] McClelland G A H and Conway G R 1971 Frequency of blood feeding in the mosquito Aedes aegypti *Nature* **232** 485–6

[184] McLean A R and Anderson R M 1988 Measles in developing-countries: II. The predicted impact of mass vaccination *Epidemiol. Infection* **100** 419–42

[185] Medlock J and Galvani A P 2009 Optimizing influenza vaccine distribution *Science* **325** 1705–8

[186] Meyers L A, Pourbohloul B, Newman M E J, Skowronski D M and Brunham R C 2005 Network theory and SARS: predicting outbreak diversity *J. Theor. Biol.* **232** 71–81

[187] Miller B R, Nasci R S, Godsey M S, Savage H M, Lutwama J J, Lanciotti R S and Peters C J 2000 First field evidence for natural vertical transmission of West Nile virus in Culex Univittatus complex mosquitoes from rift valley province, Kenya *Am. J. Tropical Med. Hygiene* **62** 240–6

[188] Miller J C, Slim A C and Volz E M 2012 Edge-based compartmental modelling for infectious disease spread *J. R. Soc. Interface* **9** 890–906

[189] Miller J C 2011 A note on a paper by Erik Volz: SIR dynamics in random networks *J. Math. Biol.* **62** 349–58

[190] Mitchell A, Bourn D, Mawdsley J, Wint W, Clifton-Hadley R and Gilbert M 2007 Characteristics of cattle movements in Britain—an analysis of records from the cattle tracing system *Animal Sci.* **80** 265–73

[191] Mollison D 1991 Dependence of epidemic and population velocities on basic parameters *Math. Biosci.* **107** 255–87

[192] Molloy M and Reed B 1995 A critical-point for random graphs with a given degree sequence *Random Struct. Algor.* **6** 161–79

[193] Molloy M and Reed B 1995 A critical point for random graphs with a given degree sequence *Random Struct. Algor.* **6** 161–80

[194] Morgan D, Mahe C, Mayanja B, Okongo J M, Lubega R and Whitworth J A G 2002 HIV-1 infection in rural Africa: is there a difference in median time to AIDS and survival compared with that in industrialized countries? *AIDS* **16** 597–603

[195] Morgan W M and Curran J W 1986 Acquired immunodeficiency syndrome: current and future trends *Public Health Rep.* **101** 459

[196] Morris M and Kretzschmar M 1997 Concurrent partnerships and the spread of HIV *AIDS* **11** 641–8

[197] Morris R S, Stern M W, Stevenson M A, Wilesmith J W and Sanson R L 2001 Predictive spatial modelling of alternative control strategies for the Foot-and-Mouth disease epidemic in Great Britain, 2001 *Veterinary Rec.* **149** 137–44

[198] Mossong J *et al* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases *PLoS Medicine* **5** e74

[199] Mossong J J *et al* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases *PLoS Med.* **5** e74

[200] Murray J D 2002 *Mathematical Biology: An Introduction* vol 1 3 edn (Berlin: Springer)

[201] Murray J D, Stanley E A and Brown D L 1986 On the spatial spread of rabies among foxes *Proc. R. Soc. Lond.* B **229** 111–50

[202] Newman M 2010 *Networks: An Introduction* (Oxford: Oxford University Press)

[203] Newman M E J 2003 Random graphs as models of networks *Handbook of Graphs and Networks* ed S Bornholdt and H G Schuster (New York: Wiley)

[204] Newman M E J 2002 Assortative mixing in networks *Phys. Rev. Lett.* **89** 208701

[205] Newman M E J 2002 Spread of epidemic disease on networks *Phys. Rev.* E **66** 016128

[206] Newman M E J and Watts D J 1999 Renormalization group analysis of the small-world network model *Phys. Lett.* A **263** 341–6

[207] Nobel J V 1974 Geographic and temporal development of plagues *Nature* **250** 726–8

[208] Okiro E A, White L J, Ngama M, Cane P A, Medley G F and Nokes D J 2010 Duration of shedding of respiratory syncytial virus in a community study of Kenyan children *BMC Infectious Diseases* **10** 15

[209] Okubo A and Levin S A 2002 *Diffusion and Ecological Problems: Modern Perspectives* 2nd edn (New York: Springer)

[210] Olsen L F, Truty G L and Schaffer W M 1988 Oscillations and chaos in epidemics: a nonlinear dynamic study of six childhood diseases in Copenhagen, Denmark *Theor. Population Biol.* **33** 344–70

[211] Ott S L, Wells S J and Wagner B A 1999 Herd-level economic losses associated with Johne's disease on US dairy operations *Preventive Veterinary Med.* **40** 179–92

[212] Pastor-Satorras R and Vespignani A 2001 Epidemic spreading in scale-free networks *Phys. Rev. Lett.* **86** 3200

[213] Puterman M L 2005 *Markov Decision Processes: Discrete Stochastic Dynamic Programming* 2nd edn (New York: Wiley)

[214] Ren J and Zhang X 2008 Freidlin–Wentzell's large deviations for stochastic evolution equations *J. Funct. Anal.* **254** 3148–72

[215] Rhodes C J and Anderson R M 1997 Epidemic thresholds and vaccination in a lattice model of disease spread *Theor. Population Biol.* **52** 101–18

[216] Rizzo D M and Garbelotto M 2003 Sudden oak death: endangering California and Oregon forest ecosystems *Frontiers Ecol. Environ.* **1** 197–204

[217] Rohani P, Keeling M J and Grenfell B T 2002 The interplay between determinism and stochasticity in childhood diseases *Am. Naturalist* **159** 469–81

[218] Rohani P, Zhong X and King A A 2010 Contact network structure explains the changing epidemiology of pertussis *Science* **330** 982–5

[219] Ross J V, House T and Keeling M J 2010 Calculation of disease dynamics in a population of households *PLoS One* **5** e9666

[220] Ross R 1911 *The Prevention of Malaria* (London: Murray)

[221] Sachs J and Malaney P 2002 The economic and social burden of malaria *Nature* **415** 680–5

[222] Sattenspiel L and Simon C P 1988 The spread and persistence of infectious diseases in structured populations *Math. Biosci.* **90** 341–66

[223] Savary S, Teng P S, Willocquet L and Nutter F W 2006 Quantification and modeling of crop losses: a review of purposes *Ann. Rev. Phytopathol.* **44** 89–112

[224] Schenzle D 1984 An age-structured model of pre- and post-vaccination measles transmission *IMA J. Math. Appl. Med. Biol.* **1** 169–91

[225] Sellke T 1983 On the asymptotic distribution of the size of a stochastic epidemic *J. Appl. Probab.* **20** 390–4

[226] Shaw M W 1995 Simulation of population expansion and spatial pattern when individual dispersal distributions do not decline exponentially with distance *Proc. R. Soc.* B **259** 243–8

[227] Simpson J E, Hurtado P J, Medlock J, Molaei G, Andreadis T G, Galvani A P and Diuk-Wasser M A 2012 Vector host-feeding preferences drive transmission of multi-host pathogens: West Nile virus as a model system *Proc. R. Soc.* B **279** 925–33

[228] Smith D L, Lucey B, Waller L A, Childs J E and Real L A 2002 Predicting the spatial dynamics of rabies epidemics on heterogeneous landscapes *Proc. Natl Acad. Sci. USA* **99** 3668–72

[229] Ster I C and Ferguson N M 2007 Transmission parameters of the 2001 Foot-and-Mouth epidemic in Great Britain *PLoS One* **2** e502

[230] Styer L M, Carey J R and Wang J L 2007 Mosquitoes do senesce: departure from the paradigm of constant mortality *Am. J. Tropical Med. Hyg.* **76** 111–7

[231] Szmaragd C, Wilson A J, Carpenter S, Wood J L N, Mellor P S and Gubbins S 2009 A modeling framework to describe the transmission of bluetongue virus within and between farms in Great Britain *PLoS One* **4** e7741

[232] Thieme H R and Castillo-Chavez C 1993 How may infection-age-dependent infectivity affect the dynamics of HIV/AIDS? *SIAM J. Appl. Math.* **53** 1447–79

[233] Tildesley M, House T, Bruhn M, Curry R J, O'Neil M, Allpress J L E, Smith G and Keeling M J 2010 Impact of spatial clustering on disease transmission and optimal control *Proc. Natl Acad. Sci. USA* **107** 1041–6

[234] Tildesley M J, Bessell P R, Keeling M J and Woolhouse M E J 2009 The role of pre-emptive culling in the control of foot-and-mouth disease *Proc. R. Soc. Lond.* B **276** 3239–48

[235] Tildesley M J, Gally D L, McNeilly T N, Low J C, Mahajan A and Savill N J 2012 Insights into mucosal innate responses to *Escherichia coli* O157 : H7 colonization of cattle by mathematical modelling of excretion dynamics *J. R. Soc. Interface* **9** 518–27

[236] Tildesley M J, Savill N J, Shaw D J, Deardon R, Brooks S P, Woolhouse M E J, Grenfell B T and Keeling M J 2006 Optimal reactive vaccination strategies for a Foot-and-Mouth outbreak in the UK *Nature* **440** 83–6

[237] Tildesley M J, Deardon R, Savill N J, Bessell P R, Brooks S P, Woolhouse M E J, Grenfell B T and Keeling M J 2008 Accuracy of models for the 2001 foot-and-mouth epidemic *Proc. R. Soc.* B **275** 1459–68

[238] Touchette H 2009 The large deviation approach to statistical mechanics *Phys. Rep.* **478** 1–69

[239] Truscott J, Fraser C, Cauchemez S, Meeyai A, Hinsley W, Donnelly C A, Ghani A and Ferguson N 2011 Essential epidemiological mechanisms underpinning the transmission dynamics of seasonal influenza *J. R. Soc. Interface* **9** 304–12

[240] Tveito A and Winther R 1998 *Introduction to Partial Differential Equations: A Computational Approach* (New York: Springer)

[241] UNAIDS 2010 *Global Report Fact Sheet*

[242] UNAIDS 2012 *2012 Report on the Global AIDS Epidemic*

[243] van Herwaarden O A and Grasman J 1995 Stochastic epidemics: major outbreaks and the duration of the endemic period *J. Math. Biol.* **33** 581–601

[244] van Hoek A J, Melegaro A, Zagheni E, Edmunds J W and Gay N 2011 Modelling the impact of a combined varicella and zoster vaccination programme on the epidemiology of varicella zoster virus in England *Vaccine* **29** 2411–20

[245] Van Kampen N G 1992 *Stochastic Processes in Physics and Chemistry* (Amsterdam: Elsevier)

[246] Vernon M C and Keeling M J 2009 Representing the UK's cattle herd as static and dynamic networks *Proc. R. Soc.* B **276** 469–76

[247] Volz E 2008 SIR dynamics in random networks with heterogeneous connectivity *J. Math. Biol.* **56** 293–310

[248] Wallinga J, Edmunds W J and Kretzschmar M 1999 Perspective: human contact patterns and the spread of airborne infectious diseases *TRENDS Microbiol.* **7** 372–7

[249] Watts C H and May R M 1992 The influence of concurrent partnerships on the dynamics of HIV/AIDS *Math. Biosci.* **108** 89–104

[250] Watts D J and Strogatz S H 1998 Collective dynamics of 'small-world' networks *Nature* **393** 440–2

[251] Watts D and Strogatz S 1998 The small world problem *Collective Dyn. Small-World Networks* **393** 440–2

[252] White P C and Harris S 1995 Bovine tuberculosis in badger (*Meles meles*) populations in southwest England: the use of a spatial stochastic simulation model to understand the dynamics of the disease *Phil. Trans. R. Soc. Lond.* B **349** 391–413

[253] WHO 1997 *Report on Infectious Diseases* (Switzerland: World Health Organization)

[254] Wylie J L and Jolly A 2001 Patterns of chlamydia and gonorrhea infection in sexual networks in Manitoba, Canada *Sexually Transmitted Diseases* **28** 14–24