Noname manuscript No.
(will be inserted by the editor)

Deterministic epidemiological models at the individual level

Kieran J. Sharkey

Received: date / Accepted: date

Abstract In many fields of science including population dynamics, the vast state spaces inhabited by all but the very simplest of systems can preclude a deterministic analysis. Here, a class of approximate deterministic models is introduced into the field of epidemiology that reduces this state space to one that is numerically feasible. However, these reduced state space master equations do not in general form a closed set. To resolve this, the equations are approximated using closure approximations. This process results in a method for constructing deterministic differential equation models with a potentially large scope of application including dynamic directed contact networks and heterogeneous systems using time dependent parameters. The method is exemplified in the case of an SIR (susceptible-infectious-removed) epidemiological model and is numerically evaluated on a range of networks from spatially local to random. In the context of epidemics propagated on contact networks, this work assists in clarifying the link between stochastic simulation and traditional population level deterministic models.

Keywords Master equations · ODE · heterogeneous contact networks · population dynamics · individual based models · deterministic models · epidemic · pair approximations

1 Introduction

Mathematical epidemiology is a rapidly evolving field of fundamental importance in understanding communicable diseases and for identifying strategies for their prevention and control. Historically, deterministic differential equation models have played a very important role in its development [16, 20, 1, 5], however they are usually applicable to very idealised systems in which a large degree of homogeneity is assumed. While

Kieran J. Sharkey
Department of Mathematical Sciences, The University of Liverpool, L69 7ZL, UK.
& Epidemiology Research Group, Department of Veterinary Clinical Sciences, The University of Liverpool, Leahurst, Neston, CH64 7TE, UK
Present address: Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester M1 7DN, UK
E-mail: kieran.sharkey@manchester.ac.uk

theoretically valuable for obtaining closed form expressions for threshold quantities such as the basic reproduction ratio R_0 [3,1,12,24] and the basic depression ratio D_0 [2], the domain of applicability of any given differential equation modelling approach remains quite limited.

By contrast, stochastic simulation methods [10] have been successfully employed in the description of very complicated systems to investigate threshold conditions and to evaluate the efficacy of methods of disease control [15,9,27]. In general, stochastic modelling approaches are found to be much more flexible than their deterministic counterparts. Nevertheless, there have been some promising recent developments in deterministic modelling that have focussed on the idea that the spread of infection is due to a network of contacts between individuals. This is a powerful construct because, if we incorporate the notion of a dynamic network of contacts with variable transmission strengths, essentially all communicable diseases can be viewed as being propagated in this way.

Fundamentally, infection is caused by a “contact” between an infectious and a susceptible individual irrespective of how that contact arises. This idea permits the description of the population dynamics in terms of pairs of individuals and produces a very natural extension of mean-field theory that incorporates some of the network structure [19,25,14,23,28,26]. However, although some attempts have been made to incorporate network heterogeneities [6,29] and group heterogeneities [8] into these pair-level models, they remain very idealised.

Motivated by these developments, this paper considers a class of individual based deterministic models with the potential to incorporate a large amount of heterogeneity and complexity. In the context of epidemics spread by contact networks, this development also helps to clarify the link between stochastic simulation and population level deterministic models.

We start with the observation that in principle, a complete deterministic description of an epidemiological system is obtained by integrating the master equations for the probabilities of the system states. This is impractical for all but the very simplest of systems because of the large number of equations that arise. However, by dividing the epidemiological system into smaller subsystems, the number of master equations can be reduced to a computationally feasible level without loss of accuracy. This comes at the considerable cost that the reduced set of equations is not closed and is therefore insoluble. By assuming statistical independence of the subsystems, we obtain a closed solvable system of equations. In this paper, these ideas are investigated in the case of the compartmental SIR model [16,1,11] where individuals are represented by three states: susceptible, infectious and resistant (or removed).

The next section introduces the reduced master equations for an arbitrary system. Section 3 applies them in the case of an SIR model utilising a closure based on statistical independence at the level of individuals. Section 4 relates this model to mean-field theory. Section 5 considers a closure based on statistical independence at the level of pairs of individuals and section 6 links this with population level pair-approximation models. Finally the resulting four models are compared with the results of stochastic simulation on a range of synthetic contact networks.

2 Reduced Master Equations

Here we consider a general system $\mathit{\Gamma}$ whose state is denoted by $\mathit{\Gamma}^\alpha$ where the index α is an integer lying between 1 and the total number of possible system states M. Using round brackets () to denote probabilities (to shorten equations, brackets are used throughout the paper to denote probabilities because in this context, the distinction between this and other uses of brackets should be clear), the probability $(\mathit{\Gamma}^\alpha)$ that the system $\mathit{\Gamma}$ is in state $\mathit{\Gamma}^\alpha$ is given by the master equations:

$$(\dot{\mathit{\Gamma}^\alpha}) = \sum_{\beta=1}^{M} \left[R^{\beta\alpha}(\mathit{\Gamma}^\beta) - R^{\alpha\beta}(\mathit{\Gamma}^\alpha) \right] \tag{1}$$

where $R^{\beta\alpha}$ denotes the transition rate from state $\mathit{\Gamma}^\beta$ to state $\mathit{\Gamma}^\alpha$. In principle, the solution of these equations provides the complete evolution of the probabilities of the states of the stochastic system $\mathit{\Gamma}$. However, this is not feasible for systems of any significant complexity.

Let us now split $\mathit{\Gamma}$ into a set of Z coupled subsystems ψ_i where i is an integer between 1 and Z. For each subsystem ψ_i, we can write master equations to describe the state probabilities:

$$(\dot{\psi_i^a}) = \sum_{b=1}^{m_i} \left[R_i^{ba}(\psi_i^b) - R_i^{ab}(\psi_i^a) \right] \tag{2}$$

where the indices a and b denote two of the m_i possible states of the ith subsystem and R_i denotes the matrix of transition rates between states for the ith subsystem and is, in general, dependent on the states of the other subsystems. For the case where there are $m_i = m$ states available to each subsystem, this results in $Z(m-1)$ equations (we only need $m-1$ equations per state because of the constraint that the probabilities must sum to 1; $\sum_a \psi_i^a = 1$). This can be far smaller than the M master equations for the complete system, but at the cost that these equations are not closed. In the next section we investigate this in the specific case of the SIR epidemiological model.

3 Individual-based SIR models

For an SIR model applied to N individuals, there are potentially $3^N - 1$ master equations. It is usually impractical to integrate these numerically unless N is very small. Here, an obvious set of subsystems is formed by the individuals themselves. Another possible set is formed by pairs of individuals. Treating individuals as a set of subsystems, $2N$ master equations are obtained. Solving this number of equations is feasible for reasonably large values of N.

Denoting the probability that the ith individual is in a susceptible or infectious state by (S_i) and (I_i) respectively, equation 2 becomes:

$$\begin{aligned} (\dot{S_i}) &= -R_i^{SI}(S_i) \\ (\dot{I_i}) &= R_i^{SI}(S_i) - R_i^{IR}(I_i) \end{aligned} \tag{3}$$

where here, resistant individuals are assumed to have lifelong immunity and so do not return to the susceptible class. Birth and death processes are also ignored.

It is useful to consider all infection events during an epidemic as being due to a "contact" between an infected and a susceptible individual. Here, contact is taken in its most general sense incorporating both direct contact and more indirect contacts such as by environment, vectors, air, water and transportation. These contacts can be represented by a matrix G where:

$$G_{ji} = \begin{cases} 1 \text{ if there is contact from individual } j \text{ to individual } i \\ 0 \text{ otherwise} \end{cases} \tag{4}$$

and $G_{ii} = 0$. It is also convenient to define T_{ji} to be to be the transmission rate from j to i when individual j is infectious and i is susceptible. In the case of a homogeneous transmission rate τ we have:

$$T_{ji} = \tau G_{ji} \tag{5}$$

Let us now look at the transition rates R_i^{SI} and R_i^{IR} in equation 3. The rate of becoming resistant (or removed) is usually assumed to be dependent only on the individual so we can write $R_i^{IR} = g_i$. The rate of becoming infectious is given by the total infectious pressure on a susceptible individual at a given time. This is:

$$R_i^{SI} = \sum_{j=1}^{N} T_{ji} \frac{(I_j S_i)}{(S_i)} \tag{6}$$

where $(I_j S_i)$ is the probability that individual j is infectious and individual i is susceptible. Equation 3 then becomes:

$$\begin{aligned} (\dot{S_i}) &= -\sum_j T_{ji}(I_j S_i) \\ (\dot{I_i}) &= \sum_j T_{ji}(I_j S_i) - g_i(I_i) \end{aligned} \tag{7}$$

where g_i and T_{ji} could be time dependent. Here and in subsequent equations, summations are assumed to be from 1 to N inclusive unless indicated otherwise.

In its current form, equation 7 is exact but not closed. It can be closed at the level of individuals by assuming statistical independence in the states of individuals: $(I_j S_i) = (I_j)(S_i)$. Hence:

$$\begin{aligned} (\dot{S_i}) &= -\sum_j T_{ji}(I_j)(S_i) \\ (\dot{I_i}) &= \sum_j T_{ji}(I_j)(S_i) - g_i(I_i) \end{aligned} \tag{8}$$

which is a closed, solvable system of equations. This is referred to as the "individual-based model" in what follows. In principle, this type of model has the scope to evaluate on an individual level the time evolution of complex epidemiological systems described by heterogeneous and time dependent contact networks.

In summary, both the full and reduced sets of master equations give the precise evolution of the probabilities of being infectious or susceptible during an epidemic. However, it is usually impractical to solve the full set, and the reduced set is incomplete and so has no solution. By assuming statistical independence at the level of individuals, an approximate set of master equations is obtained which is here termed the individual-based model. The accuracy of the individual-based model is entirely dependent on the validity of the independence assumption used to derive it. This is investigated numerically in section 7 by comparison with stochastic individual based models.

4 Connection with mean-field models

Most differential equation models of epidemiological systems describe population level dynamics. Here we explore how the individual-based SIR model of the previous section relates to mean-field population level SIR models.

Denoting the expectation values of the susceptible and infectious population sizes by $[S]$ and $[I]$ respectively [13,14], these quantities are be related to the probabilities in the reduced master equations by:

$$\begin{aligned}[S] &= \sum_i (S_i) \\ [I] &= \sum_i (I_i)\end{aligned} \tag{9}$$

It follows from equations 8 and 9 that:

$$\begin{aligned}[\dot{S}] &= -\sum_{ji} T_{ji}(I_j)(S_i) \\ [\dot{I}] &= \sum_{ji} T_{ji}(I_j)(S_i) - \sum_i g_i(I_i)\end{aligned} \tag{10}$$

Applying the mean-field assumption $(S_i) = [S]/N$ and $(I_i) = [I]/N$ gives the mean-field SIR model:

$$\begin{aligned}[\dot{S}] &= -\beta[I][S] \\ [\dot{I}] &= \beta[I][S] - \gamma[I]\end{aligned} \tag{11}$$

where:

$$\begin{aligned}\beta &= \frac{1}{N^2}\sum_{ji} T_{ji} \\ \gamma &= \frac{1}{N}\sum_i g_i\end{aligned} \tag{12}$$

When the transmission rates are homogeneous (equation 5) and the removal rates are also homogeneous ($g_i = g$), the mean-field theory becomes:

$$\begin{aligned}[\dot{S}] &= -\tau n[I][S]/N \\ [\dot{I}] &= \tau n[I][S]/N - g[I]\end{aligned} \tag{13}$$

where n is the average number of neighbours per node defined by $\|G\| = Nn$ where:

$$\|G\| \equiv \sum_{ji} G_{ji} \tag{14}$$

By applying the mean-field or probability averaging assumption to the individual-based model, we effectively smooth out the heterogeneity in the model and treat each site as identical. In turn, the individual-based model of the previous section follows from the master equations using the assumption of statistical independence at the level of individuals. Consequently the mean-field model depends on two assumptions: statistical independence of individuals and the mean-field assumption. Numerical comparisons of the mean-field model, individual-based model and stochastic simulation are made in section 7.

5 Pair-based SIR models

Both mean-field theory and the individual-based model make the assumption of statistical independence at the level of individuals. For population dynamics, differential equation models describing processes at the level of pairs of individuals have been investigated to avoid this assumption [19,25,14,23,28,26] and, in this section, we consider an analogous development at the individual level.

The complete reduced master equations for the individual pair dynamics in the SIR model are:

$$
\begin{aligned}
\dot{(S_j S_i)} &= -\sum_{k,k\neq i} T_{kj}(I_k S_j S_i) - \sum_{k,k\neq j} T_{ki}(S_j S_i I_k) \\
\dot{(I_j S_i)} &= \sum_{k,k\neq i} T_{kj}(I_k S_j S_i) - \sum_{k,k\neq j} T_{ki}(I_j S_i I_k) - T_{ji}(I_j S_i) - g_j(I_j S_i) \\
\dot{(R_j S_i)} &= -\sum_{k,k\neq j} T_{ki}(R_j S_i I_k) + g_j(I_j S_i) \\
\dot{(I_j I_i)} &= \sum_{k,k\neq i} T_{kj}(I_k S_j I_i) + \sum_{k,k\neq j} T_{ki}(I_j S_i I_k) + T_{ji}(I_j S_i) \\
&\quad + T_{ij}(S_j I_i) - (g_i + g_j)(I_j I_i) \\
\dot{(R_j I_i)} &= \sum_{k,k\neq j} T_{ki}(R_j S_i I_k) + g_j(I_j I_i) - g_i(R_j I_i) \\
\dot{(R_j R_i)} &= g_i(R_j I_i) + g_j(I_j R_i)
\end{aligned}
\tag{15}
$$

where the notation $(A_i B_j C_k)$ denotes the probability that individual i is in state A, individual j is in state B and individual k is in state C.

In connection with the discussion in section 2, pairs of individuals form sub-systems in addition to the individual level subsystems described in equation 7. The problem now is to find a closed set of equations. Firstly, instead of approximating the probability $(I_j S_i)$ in equation 7 by the independence assumption, the $(I_j S_i)$ expression from equation 15 is used. To generate a closed set of equations, the triples probabilities in this expression must be approximated in terms of pair level and individual level probabilities. It is possible to approximate triples in many ways and full details of the closure approximations used for the simulations in this paper are given in appendix A. These approximations are based on the assumption of statistical independence at the level of pairs. This model is referred to as the pair-based model in what follows.

While the pair-based model does not assume independence at the level of individuals or depend on a mean-field assumption, it does assume statistical independence at the level of pairs. The accuracy of this model therefore depends on the validity of this assumption.

The pair-based model for the symmetric contact networks considered in section 7 involves the solution of $(3n+2)N$ ordinary differential equations. This is considerably more than the $2N$ equations of the individual-based model, but is still numerically feasible for reasonably large values of N.

6 Connection with the population level pair-approximation models

In section 4, the individual-based model was related to the mean-field model. Similarly the pair-based model can be related to its corresponding "mean-field" model, where the averaged (or mean-field) quantities are pairs instead of individuals. Using the square bracket notation [13, 14] for population level pair equations on undirected contact networks, the expectation value for the number of pairs where node i is in state A and node j is in state B and for which there is a network link from i to j is given by:

$$[AB] = \sum_{ij} G_{ij}(A_i B_j) \tag{16}$$

When transmission rates are homogeneous (equation 5) and removal rates are homogeneous ($g_i = g$), the i index in equation 7 can be summed to give:

$$\begin{aligned} [\dot{S}] &= -\tau[SI] \\ [\dot{I}] &= \tau[SI] - g[I] \end{aligned} \tag{17}$$

For undirected networks, $G_{ij} = G_{ji}$ so $[AB] = [BA]$. Furthermore, this symmetry allows the following notation for population level triples to work unambiguously [13, 14]:

$$[ABC] = \sum_{ijk, k \neq i} G_{ij} G_{jk} (A_i B_j C_k) \tag{18}$$

For symmetric networks with homogeneous transmission and removal rates, the $S_j S_i$, $I_j S_i$ and $I_j I_i$ expressions in equation 15 give the following population level equations when the indices i and j are summed in conjunction with the definitions in equations 16 and 18:

$$\begin{aligned} [\dot{SS}] &= -2\tau[SSI] \\ [\dot{SI}] &= \tau([SSI] - [ISI] - [SI]) - g[SI] \\ [\dot{II}] &= 2\tau([ISI] + [SI]) - 2g[II] \end{aligned} \tag{19}$$

For symmetric contact networks with homogeneous transmission and removal rates, equations 17 and 19 provide a precise description of the expected population dynamics [19, 14, 23, 26]. For the more general case of asymmetric networks with homogeneous transmission and removal rates, the sum over the i and j indices in equation 15 produce a more general set of population level equations for which a more detailed notation is required [26].

Closure approximations are required to solve these population level equations. These population level closures can be derived by applying an averaging or "mean-field" assumption to the closure approximations for the pair-based model in appendix A. Full details of this are discussed in Appendix B. Equations 17 and 19 together with the closure approximations are referred to as the "pair-approximation" model in what follows.

To summarise, four models have been defined; two (mean-field and pair-approximation) are population level models and two (individual-based and pair-based) are individual level models. These models can be related to each other by the mean-field assumption and by the assumption of the statistical independence of individuals. These models are summarised with respect to their assumptions in table 1.

Model Name	Mean Field Assumption	Independence of individuals assumption
Mean-Field	Yes	Yes
Pair-Approximation	Yes	No
Individual-Based	No	Yes
Pair-Based	No	No

Table 1 Comparison of the mean-field, pair-approximation, individual-based and pair-based models with respect to the assumptions used to construct them.

7 Numerical results and discussion

Although there is considerable scope for introducing a large amount of complexity into both the individual-based and pair-based models, these investigations are left for future work. Here, to make a fair comparison with population level models, only static undirected contact networks with homogeneous transmission and removal rates are considered.

A range of networks from random to spatially local are constructed by using a two dimensional variant of the Watts & Strogatz small-world network [30,22]. A two dimensional construction is more relevant for epidemiological systems because local transmission typically occurs on a plane and not in one dimension as in the original Watts & Strogartz network.

To construct the networks, N individuals (or nodes) are distributed uniformly at random on the surface of a sphere. Each node is then connected via undirected links to its m nearest (Euclidean) neighbours. With probability p, each link is then completely removed and reassigned between two nodes chosen uniformly at random from the set of pairs of unconnected nodes. Self contact is also not permitted. By varying the reassignment probability p, it is possible to investigate a range of small-world networks from spatially local ($p = 0$) to Erdos-Renyi type random networks ($p = 1$)[7,22].

For each parameter set, 10,000 stochastic simulations of major outbreaks (defined here as infecting more than a quarter of the total population) are generated. Each of these simulations is initiated at the same individual. The infection of this same individual with all others being susceptible also defines the initial conditions for the individual-based and pair-based deterministic models. From the fraction F of simulations that produce major outbreaks, a naive estimation of R_0 can be obtained from $F = 1 - 1/R_0$ to give an indicator of the epidemiological nature of the systems being considered.

Figure 1 shows the comparison of mean-field, pair-approximation, individual-based and pair-based models for SIR epidemics on a random network ($p = 1$). Infectious (figure 1a) and susceptible (figure 1b) time series are shown as well as the variation with τ of the infectious population at a fixed point in time (figure 1c) and of the final size susceptible population (figure 1d). Here, $m = 6$ for which the above network construction produces an average number of neighbours per node of $n = 7.1$.

There is clear agreement between the pair-based model predictions and the simulated epidemics. The pair-approximation model also performs reasonably well here by keeping within the tolerances of the error bars. However, the individual-based and mean-field models do not perform as well.

The differences in the models can be understood in terms of the mean-field and independence assumptions used to construct them. These assumptions are summarised in table 1. This table shows that the poorer performance of the individual-based and

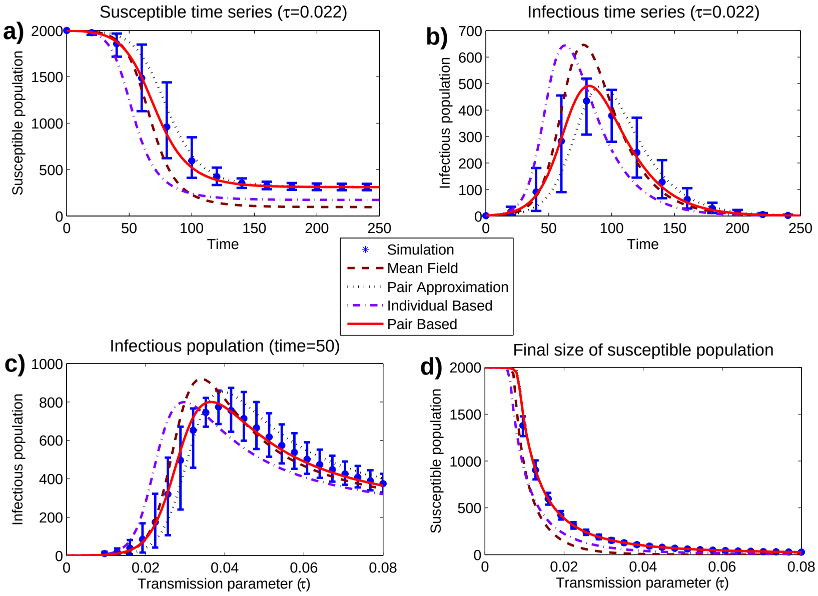

Fig. 1 Simulations and model predictions for SIR epidemics on a random network ($p = 1$) for which $m = 6$ and $N = 2000$. The removal rate is held fixed at $g = 0.05$. The mean of 10,000 simulations initiated at the same node is plotted with an astrix and the error bars indicate the 10th and 90th percentiles. Comparison is made with the mean-field (dashed line), pair-approximation (dotted line), individual-based (dot-dashed line) and pair-based (solid line) models. The graphs are a) Susceptible time series for $\tau = 0.022$ (corresponding to $R_0 \approx 5.4$), b) Infectious time series for $\tau = 0.022$, c) Number of infected individuals at time=50 as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 20$. d) Final size of susceptible population as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 20$.

mean-field models on the random network is attributable to the assumption of independence at the level of individuals. This failure of independence can be partly understood by observing that the network contacts of an infectious individual are more likely to be infectious than the network contacts of a susceptible individual; this correlation between pairs of nodes breaks the independence at the individual level [4]. For random networks with higher connectivity, the behaviour of the mean-field model improves because the independence assumption becomes more applicable [4,14]. Numerical evaluation (not reproduced here) confirms that the predictions of the individual-based model also improve for networks with higher connectivity. In the limit of complete connectivity ($n = N - 1$), the individual-based and mean-field models are almost equivalent for heterogeneous networks (appendix C).

The pair-based and pair-approximation models are also dependent on an independence assumption, but this is at the level of pairs of nodes instead of individuals (appendices A&B). The results in figure 1 indicate that for this random network, this assumption is accurate.

From table 1 it is evident that the small discrepancy in figure 1 between the pair-based and pair-approximation models and between the individual-based and mean-field models is attributable to the averaging or "mean-field" assumption. In general the heterogeneity that may cause the mean-field assumption to fail originates from

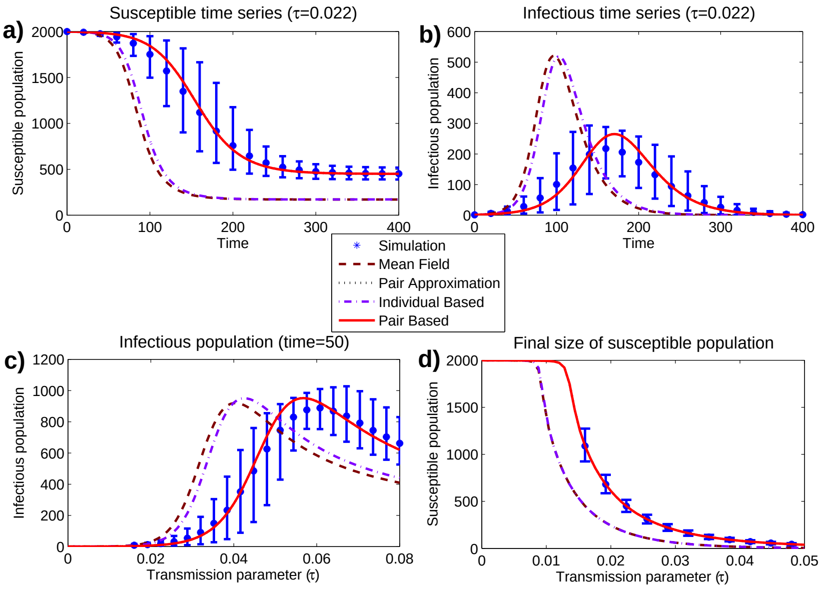

Fig. 2 Simulations and model predictions for SIR epidemics on a random symmetric network constructed by an iterative procedure to ensure that each of the 2000 nodes has exactly 6 neighbours. The removal rate is held fixed at $g = 0.05$. The mean of 10,000 simulations initiated at the same node is plotted with an astrix and the error bars indicate the 10th and 90th percentiles. Comparison is made with the mean-field (dashed line), pair-approximation (dotted line), individual-based (dot-dashed line) and pair-based (solid line) models. The graphs are a) Susceptible time series for $\tau = 0.022$ (corresponding to $R_0 \approx 2.2$), b) Infectious time series for $\tau = 0.022$, c) Number of infected individuals at time=50 as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 9.3$. d) Final size of susceptible population as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 5.3$.

two sources. Firstly there is the hardwired network based heterogeneity which, for this random network, is primarily due to the variation in the number of neighbours per node. Secondly there is heterogeneity that is induced by the dynamics of the epidemic. To investigate the distinction between network heterogeneity and dynamically induced heterogeneity further, figure 2 considers a randomly connected undirected network formed with the constraint that every node has exactly 6 neighbours. By design, this network is completely homogeneous at the level of pairs. Figure 2 show that at the resolution of the graph, there is now no discernable difference between the pair-based and pair-approximation models indicating that at the pair level, the mean-field assumption works very well. This also suggests that the discrepancy between the pair-based and pair-approximation models in figure 1 is attributable primarily to network based and not dynamically induced heterogeneity.

For the mean-field and individual-based models, the high degree of homogeneity in the network in figure 2 has resulted in very similar predictions, however a small discrepancy still remains for figures 2a,b&c. Although minimised, network heterogeneity may still play a part in this discrepancy. This is because in spite of the network possessing complete homogeneity at the pair level and being generally homogeneous because of its random construction, there will still be some degree of localised variation at the triples order or higher.

To evaluate the sensitivity to higher order network heterogeneity, the individual-based model was systematically initiated on each of the 2000 nodes in the network. The resulting variation in predictions was found to be very small and contained within the resolution of the lines for the individual-based model in figure 2. This indicates that the impact of any residual higher than pair-order localised heterogeneity is minimal for this network and that consequently the observed difference between the individual-based and mean-field models must be attributable to the heterogeneity induced by the dynamics. This heterogeneity will reduce with increasing connectivity in the network and for the extreme case of a fully connected network ($n = N - 1$), we have already noted the near equivalence of the individual-based and mean-field models (appendix C).

Perhaps the most extreme case of dynamically induced heterogeneity results from a spatially local network ($p = 0$). Epidemics on this type of network propagate outwards as a wave emanating from the first infected node with the wave front roughly marking a dividing line between the infectious and susceptible populations [20]. Consequently the homogeneous mixing of infected and susceptible individuals that can justify the mean-field and independence assumptions does not occur and, in fact, there is very little mixing of the populations.

The performance of the mean-field, pair-approximation, individual-based and pair-based models on a locally connected network is illustrated in figure 3. Here, all four models overestimate the speed of propagation of the epidemic to varying degrees. For this network, the individual-based model performs better than the pair-approximation model. This indicates that here, the mean-field assumption is less accurate than the independence assumption. This contrasts with the low-connectivity random networks in figures 1 and 2 where the independence assumption leads to greater inaccuracy than the mean-field assumption.

Clearly the best results on the spatially local network are obtained for the pair-based model and although not ideal, this provides a reasonable representation of the stochastic simulations on this network. The remaining discrepancy between the pair-based model and the stochastic simulation data is attributable to the assumption of independence at the level of pairs.

Notice from figure 3d that there is no discernable difference between the final size predictions of the mean-field and individual-based models and between the final size predictions of the pair-approximation and pair-based models in spite of very distinct time evolution (figures 3a,b&c). This is relatively unsurprising because final size predictions are often independent of the detailed space-time evolution of an epidemic [18,4]. However, for more complex heterogeneous networks this is not true because the final size of an epidemic depends on the localised cluster in which it is initiated. In the extreme case of initiating an epidemic at a node that has no neighbours, the individual-based and pair-based models will always produce a final size of 1, but this will not influence the predictions of the mean-field and pair-approximation models. As an example of this, the construction of the random network in figure 1 contains two nodes that are unconnected to any other node and there will also be small sub-clusters of nodes that have low-connectivity to the main giant cluster. This goes some way towards explaining the small discrepancy in the final size predictions between the individual-based and mean-field models in figure 1d which is not seen in the final size predictions in figures 2d&3d.

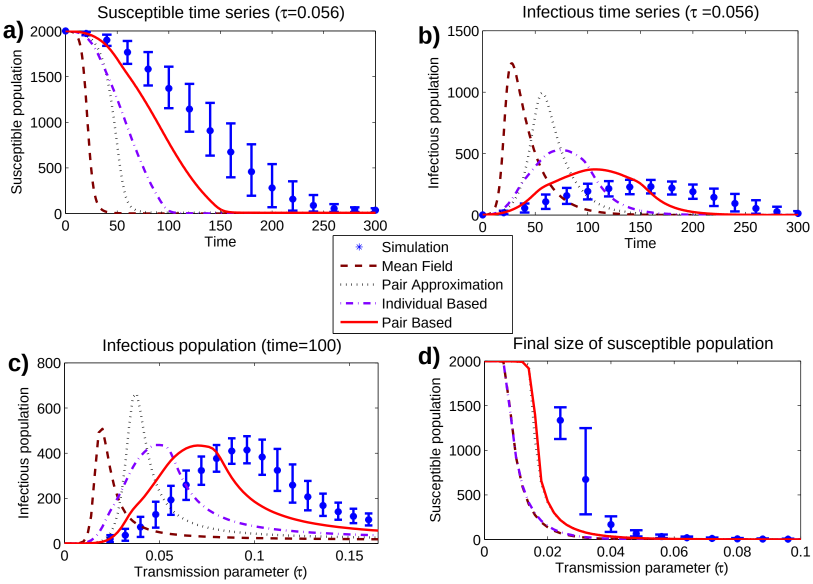

Fig. 3 Simulations and model predictions for SIR epidemics on a spatially local network ($p = 0$) for which $n = 6$ and $N = 2000$. The removal rate is held fixed at $g = 0.05$. The mean of 10,000 simulations initiated at the same node is plotted with an astrix and the error bars indicate the 10th and 90th percentiles. Comparison is made with the mean-field (dashed line), pair-approximation (dotted line), individual-based (dot-dashed line) and pair-based (solid line) models. The graphs are a) Susceptible time series for $\tau = 0.056$ (corresponding to $R_0 \approx 6$), b) Infectious time series for $\tau = 0.056$, c) Number of infected individuals at time=100 as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 20$. d) Final size of susceptible population as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 11.6$.

Figures 1 to 3 suggest that the accuracy of the four models decreases with increasing locality. Figure 4 investigates this explicitly by showing the variation of population size at specific times and for specific transmission rates as a function of the randomisation parameter p. These are not smooth functions of p because each value of p corresponds to a different randomly generated network. Two transmission rates are considered ($\tau = 0.056$) (figure 4a&b) which corresponds to $R_0 \approx 6$ and $\tau = 0.02$ (figure 4c&d) corresponding to $R_0 \approx 2$. For the two values of τ, two time points are chosen, one part way through the dynamics (figure 4a&c) and a final size result at the end of the epidemic (figure 4b&d).

The predictions of the mean-field model are seen to be independent of p. This is because p is a network rearrangement parameter which does not alter the value of $||G||$ in equation 14. This leaves the predictions of the mean-field model unchanged.

The mean-field and individual-based models do not provide a good match to any of the simulation results in figure 4.

The failure of the pair-based and pair-approximation models with decreasing p is seen most clearly for the final size predictions (figure 4b&d). The model predictions of the pair-based model fall outside of the tolerances of the error bars around $p = 0.1$. More detailed simulations on a $p = 0.1$ network are shown in figure 5.

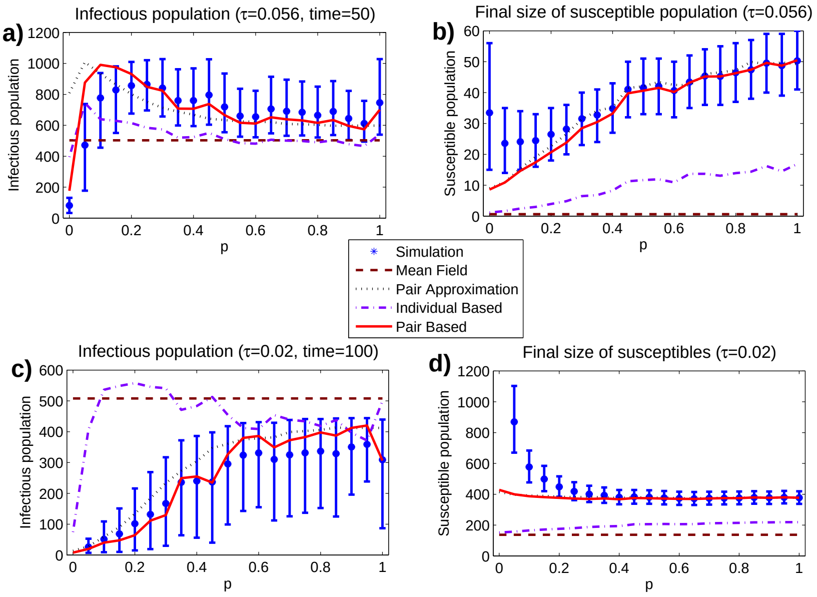

Fig. 4 Simulations and model predictions for SIR epidemics on a sequence of small-world networks ranging from spatially local ($p = 0$) to random ($p = 1$) where $n = 6$ and $N = 2000$. Each small world network is generated from the local network in figure 3 by applying different levels of randomisation p. The removal rate is held fixed at $g = 0.05$. The mean of 10,000 simulations initiated at the same node is plotted with an astrix and the error bars indicate the 10th and 90th percentiles. Comparison is made with the mean-field (dashed line), pair-approximation (dotted line), individual-based (dot-dashed line) and pair-based (solid line) models. The graphs are a) Infectious population at time=50 for $\tau = 0.056$. b) Final size of susceptible population for $\tau = 0.056$, c) Infectious population at time=100 for $\tau = 0.02$. d) Final size of susceptible population for $\tau = 0.02$.

Contrasting figure 5 with figure 3, it is seen that while there is very little difference between the dynamics of the pair-based and pair-approximation models for $p = 0.1$, there is a large difference for the local network $p = 0$. This divergence of behaviour in the interval $p = 0$ to $p = 0.1$ is also very evident in figure 4a.

8 Concluding remarks

A framework is introduced for the deterministic description of complicated epidemiological systems that represents a departure from previous work because of its focus on the deterministic evolution of the probability of events at an individual level instead of on population level dynamics. The potential benefits of this approach arise from its flexibility and scope of application. In particular, the framework encompasses dynamic and directed heterogeneous contact networks with time varying transmission and removal rates. These complicated systems could previously only be described by relatively ad hoc deterministic models or by stochastic simulation.

The advantages of this type of model over stochastic simulation remain to be determined. Possible benefits may arise from the lack of stochastic variation leading to just

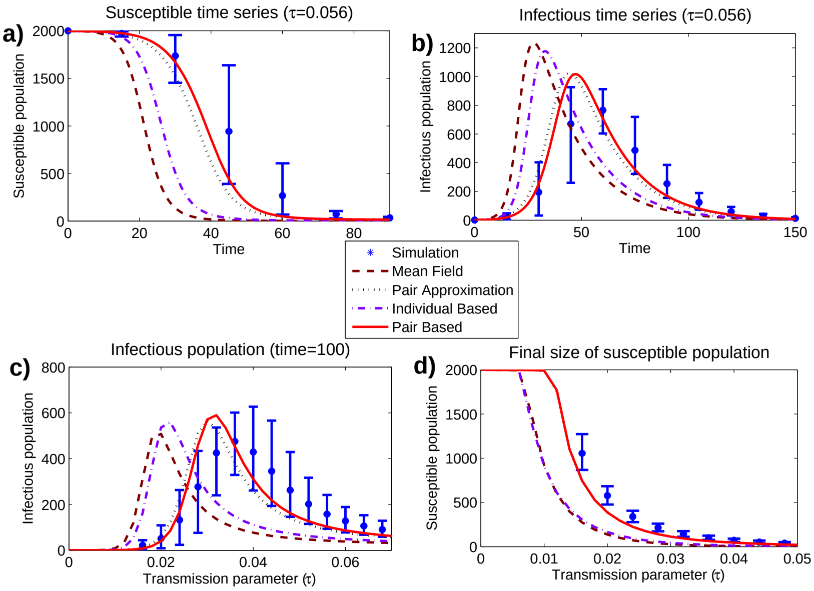

Fig. 5 Simulations and model predictions for SIR epidemics on a small-world network for which $p = 0.1$. The removal rate is held fixed at $g = 0.05$. The mean of 10,000 simulations initiated at the same node is plotted with an astrix and the error bars indicate the 10th and 90th percentiles. Comparison is made with the mean-field (dashed line), pair-approximation (dotted line), individual-based (dot-dashed line) and pair-based (solid line) models. The graphs are a) Susceptible time series for $\tau = 0.056$ (corresponding to $R_0 \approx 5.4$), b) Infectious time series for $\tau = 0.056$, c) Number of infected individuals at time=100 as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 6.7$. d) Final size of susceptible population as a function of τ ranging from $R_0 = 0$ to $R_0 \approx 4.7$.

one numerical evaluation for each parameters set. This may be helpful for identifying optimal control strategies by minimising cost functions with respect to, for example, the size of control zones, rapidity of culling and vaccination efficacy. This type of model also produces detailed information on the probability of infection for each individual as a function of time. This may find application in the construction of spatial visualisation tools for observing the propagation of infection probability in space.

Two types of individual level model were constructed, one assuming statistical independence at the level of individuals (the individual-based model) and one assuming statistical independence at the level of pairs (the pair-based model). The performance of these models was evaluated on a range of contact networks. Perhaps the most extreme type of network for these models (in view of the independence assumptions used to construct them) is the spatially local network where each individual is connected to its spatially nearest neighbours. For this network, the individual-based model is found to perform badly, although the pair-based model does a reasonable job of describing the epidemic. This is encouraging because for other networks, both the individual-based and pair-based models are expected to perform better than for this most extreme type of network.

In principle, models with greater accuracy could be obtained by assuming independence at the triples order or higher, however the number of equations required may prevent the practical application of this. Indeed, from an applied viewpoint, one possible limitation of individual level models, particularly in the case of the pair-based version, is the large number of ordinary differential equations that have to be solved. This may cause problems for very large or highly connected systems. For population level models, approximations at the level of triples may be feasible, although they may be cumbersome to write down. Population level models also have an advantage over individual level models when closed-form expressions are required for quantities such as R_0 and the final size of epidemics for simple epidemiological systems.

This paper considers a compartmental SIR model. Further developments would be needed to incorporate more complex aspects such as time dependent infectivity, waning immunity and latent states. In general, further work is required to determine the applicability of this type of model as a practical epidemiological tool and as a more general tool for evaluating dynamics on networks.

From a theoretical perspective, the construction of individual level deterministic models assists in understanding the link between stochastic processes on contact networks and population level deterministic models. To summarise this link, the master equations describing the probabilities of infection during a stochastic simulation are re-expressed in the form of reduced master equations describing subsystems. By applying the assumption of statistical independence of the subsystems, the individual-based and pair-based deterministic models are obtained. By applying a mean-field or averaging assumption, these individual level models are then related to the more traditional population-level mean-field and pair-approximation models. In the context of dynamics spread on contact networks, this clarifies the theoretical connection between stochastic simulation and population level deterministic models.

Appendix A: Pair-level closures

In general we approximate a triple probability $(A_iB_jC_k)$ by:

$$(A_iB_jC_k) \approx \frac{(A_iB_j)(B_jC_k)(C_kA_i)}{(A_i)(B_j)(C_k)}$$

Approximations of this form were originally constructed in theoretical physics [17] and have been used relatively recently to approximate triples quantities in population level epidemiological and ecological models. This type of expression can be justified by assuming statistical independence at the level of pairs [21,28,14,23,26].

This approximation applies to any set of three individuals i, j and k. However, for this work, the triples probabilities of interest are those for which there is a network link between i and j and also between j and k. For the symmetric contact networks considered in section 7, these triples are either “closed” for which there is an undirected link between individuals i and k or “open” for which there is no connection between i and k. For the closed case, the above expression is used in full. For the open case it is convenient to introduce the notation $(C_k \bowtie A_i)$ to signify that there is no network contact between i and k [26]. For the open triples we assume statistical independence between the unconnected nodes ($(C_k \bowtie A_i) = (C_k)(A_i)$) so that:

$$(A_iB_jC_k) \approx \frac{(A_iB_j)(B_jC_k)(C_k \bowtie A_i)}{(A_i)(B_j)(C_k)} \approx \frac{(A_iB_j)(B_jC_k)}{(B_j)}$$

The use of this expression for open triples means that the pair equations in equation 15 need only be solved for pairs of individuals that have a network link between them.

Appendix B: Pair-level closures for population-level models

The population level triples $[ABC]$ in equation 15 are approximated by:

$$[ABC] = N\zeta \frac{[AB][BC]}{[A][B][C]} \left(\phi \frac{[CA]}{n} + (1-\phi) \frac{[C \bowtie A]}{n_{\bowtie}} \right)$$

where N is the total population size, n is the average number of neighbours per individual and :

$$\begin{aligned} [C \bowtie A] &= [C][A] - [CA] \\ n_{\bowtie} &= N - 1 - n \end{aligned}$$

and:

$$\begin{aligned} \zeta &= \frac{\|G^2\| - Tr(G^2)}{Nn^2} \\ \phi &= \frac{Tr(G^3)}{\|G^2\| - Tr(G^2)} \end{aligned}$$

where the notation $\|G^2\|$ is defined by equation 14. This closure [26] was first suggested in a slightly different form by Morris and vanBaalen [21, 14, 28].

We now show that this population level approximation for triples follows from the approximation in appendix A when the mean-field assumption for pairs is applied.

For a triple $(A_iB_jC_k)$ with links between nodes i and j and between nodes j and k, we have from Eq 18:

$$[ABC] = \sum_{ijk,k\neq i} G_{ij}G_{jk}(A_iB_jC_k)$$

This can be split into a closed part with a network between i and k and an open part with no link by:

$$[ABC] = \sum_{ijk} G_{ij}G_{jk}G_{ki}(A_iB_jC_k) + \sum_{ijk} G_{ij}G_{jk}(G_{\bowtie})_{ki}(A_iB_jC_k)$$

where $G_{\bowtie}$ represents the "open network" of no links [26] and for symmetric networks is defined by:

$$(G_{\bowtie})_{ij} = 1 - G_{ij} - \delta_{ij}$$

where δ_{ij} is the Kronecker delta.

Applying the closure approximation from appendix A gives:

$$\begin{aligned} [ABC] \approx & \sum_{ijk} G_{ij}G_{jk}G_{ki} \frac{(A_iB_j)(B_jC_k)(C_kA_i)}{(A_i)(B_j)(C_k)} \\ & + \sum_{ijk} G_{ij}G_{jk}(G_{\bowtie})_{ki} \frac{(A_iB_j)(B_jC_k)(C_k{\bowtie}A_i)}{(A_i)(B_j)(C_k)} \end{aligned}$$

where here independence for the open pairs $(C_k{\bowtie}A_i)$ is not assumed. Applying the mean-field assumption for individuals $(A_i = [A]/N)$, network pairs $(A_iB_j = [AB]/Nn)$ and for the open pairs $((C_k{\bowtie}A_i) = [C \bowtie A]/Nn_{\bowtie})$ gives:

$$\begin{aligned}
[ABC] &\approx \sum_{ijk} G_{ij}G_{jk}G_{ki}\frac{[AB][BC][CA]}{n^3[A][B][C]} + \sum_{ijk} G_{ij}G_{jk}(G_{\bowtie})_{ki}\frac{[AB][BC][C \bowtie A]}{n^2n_{\bowtie}[A][B][C]} \\
&= Tr(G^3)\frac{[AB][BC][CA]}{n^3[A][B][C]} + \left(\|G^2\| - Tr(G^2) - Tr(G^3)\right)\frac{[AB][BC][C \bowtie A]}{n^2n_{\bowtie}[A][B][C]} \\
&= N\zeta\frac{[AB][BC]}{[A][B][C]}\left(\phi\frac{[CA]}{n} + (1-\phi)\frac{[C \bowtie A]}{n_{\bowtie}}\right)
\end{aligned}$$

as expected.

Appendix C: Near equivalence of mean-field and individual-based models for fully connected homogeneous networks

For a fully connected homogeneous network with homogeneous removal rates, the individual-based model (equation 10) becomes:

$$\begin{aligned}
[\dot{S}] &= -\tau Q \\
[\dot{I}] &= \tau Q - g[I]
\end{aligned}$$

where

$$Q = \sum_{ij}(1-\delta_{ij})(I_j)(S_i)$$

For an epidemic initiated at a single individual f within a totally susceptible population, symmetry requires the infection probability to diffuse identically across the other $N-1$ nodes. Consequently the probability of being susceptible is evenly distributed over these sites for the entire duration of the epidemic so:

$$\begin{aligned}
(S_{i\neq f}) &= \frac{[S]}{N-1} \\
S_f &= 0
\end{aligned}$$

The expression for Q then becomes:

$$\begin{aligned}
Q &= \sum_{ij}(1-\delta_{ij})(1-\delta_{if})\frac{(I_j)[S]}{N-1} \\
&= \sum_{ij}\frac{(I_j)[S]}{N-1} - \sum_{ij}\delta_{ij}\frac{(I_j)[S]}{N-1} - \sum_{ij}\delta_{if}\frac{(I_j)[S]}{N-1} + \sum_{ij}\delta_{ij}\delta_{if}\frac{(I_j)[S]}{N-1} \\
&= \frac{N[I][S]}{N-1} - \frac{2[I][S]}{N-1} + \frac{(I_f)[S]}{N-1}
\end{aligned}$$

The differential equation for (I_f) can be solved explicitly to give $(I_f) = \exp(-gt)$ where t is the time from the start of the epidemic. Hence:

$$Q = \frac{(N-2)}{N-1}[I][S] - \frac{[S]e^{-gt}}{N-1}$$

The mean-field result for a fully connected network is obtained by putting $n = N-1$ in equation 13:

$$[\dot{S}] = -\tau\frac{(N-1)}{N}[I][S]$$
$$[\dot{I}] = \tau\frac{(N-1)}{N}[I][S] - g[I]$$

which gives $Q = [I][S](N-1)/N$.

Provided that N is reasonably large, both the mean-field and individual-based models have $Q \approx [I][S]$. The small difference between the models is attributable to the explicit treatment of the first infected site in the individual-based model.

Acknowledgements I thank Roger Bowers for many helpful discussions and insights during the development of this work.

References

1. Anderson, R. M., May, R. M.: Infectious diseases of humans. Oxford University Press. (1992)
2. Bowers, R. G.: The basic depression ratio of the host: the evolution of host resistance to microparasites. Proc. R. Soc. B **268**, 243-250 (2001)
3. Diekmann, O., Heesterbeek, J. A. P., Metz, J. A. J.: On the definition and the computation of the basic reproduction ratio R_0, in models for infectious diseases in heterogeneous populations. J. Math. Biol. **28**, 365-382 (1990)
4. Diekmann, O., De Jong, M.C.M., Metz, J.A.J.: A deterministic epidemic model taking account of repeated contacts between the same individuals. J. Appl. Prob. **35**, 448-462 (1998)
5. Diekmann, O. & Heesterbeek, J. A. P.: Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. New York: John Wiley & Sons. (2000)
6. Eames, K.D., Keeling, M.J.: Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. PNAS **99**, 13330-13335 (2002)
7. Erdos, P., Renyi, A.: On random graphs. Publ. Math. Debrecen, **6**, 290-297 (1959)
8. Ferguson, N. M., Donnelly, C. A., Anderson, R. M.: The foot-and-mouth epidemic in Great Britain: Pattern of spread and impact of interventions. Science **292**, 1155-1160 (2001)
9. Ferguson, N. M., Cummings, D. A. T., Fraser, C., Cajka, J. C., Cooley, P. C., Burke, D. S.: Strategies for mitigating an influenza pandemic. Nature **442** 448-452 (2006)
10. Gillespie, D. T.: A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comp. Phys. **22**, 403-434 (1976)
11. Grenfell, B. T., Dobson, A. P.: Ecology of infectious diseases in natural populations. Cambridge University Press. (1995)
12. Heffernan, J. M., Smith, R. J., Wahl, L. M. Perspectives on the basic reproductive ratio. J. R. Soc. Interface **2**, 281-293 (2005)
13. Keeling, M.J., Rand, D.A., Morris, A.J.: Correlation models for childhood epidemics. Proc. R. Soc. B **264**, 1149-1156 (1997)
14. Keeling, M. J.: The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. B **B266**, 859-867. (1999)
15. Keeling, M. J.: Models of Foot-and-Mouth Disease. Proc. R. Soc. B **272**, 1195-1202 (2005)
16. Kermack, W. O. & McKendrick, A. G.: Contributions to the mathematical theory of epidemics. 1. Proc. R. Soc. Edinb. **A115**, 700-721. (1927)
17. Kirkwood, J. G.: Statistical Mechanics of Fluid Mixtures. J. Chem. Phys., **3**, 300-313 (1935)
18. Ludwig, D.: Final size distributions for epidemics. Math. Biosc. **23**, 33-46 (1975)
19. Matsuda, H. N., Ogita A., Sasaki, A., Satō K.: Statistical mechanics of population: The lattice Lotka-Volterra model. Prog. Theoret. Phys. **88**, 1035-1049. (1992)
20. Mollison, D.: Spatial contact models for ecological and epidemic spread. Journal of the Royal Statistical Society. **B 39** 283-326. (1977)

21. Morris, A. J.: Representing spatial interactions in simple ecological models. PhD thesis, Warwick University. (1997)
22. Newman, M. E. J.: The structure and function of complex networks. SIAM Review. **45**, 167-256. (2003)
23. Rand, D. A.: Correlation equations for spatial ecologies. In: Advanced ecological theory (ed. J. McGlade), Blackwell Scientific Publishing, 99-143. (1999)
24. Roberts, M. G.: The pluses and minuses of R_0. J.R. Soc. Interface **4**, 949-961 (2007)
25. Sato, K., Matsuda, H., and Sasaki, A.: Pathogen invasion and host extinction in lattice structured populations. J. Math. Biol. 32, 251-268 (1994)
26. Sharkey, K. J., Fernandez, C., Morgan, K. L., Peeler, E., Thrush, M., Turnbull, J. F., Bowers, R. G.: Pair-level approximations to the spatio-temporal dynamics of epidemics on asymmetric contact networks. J. Math. Biol. **53** 61-85. (2006)
27. Sharkey, K. J., Bowers, R. G., Morgan, K. L., Robinson, S. E., Christley, R. M.: Epidemiological consequences of an incursion of highly pathogenic H5N1 avian influenza into the British poultry flock. Proc Roy. Soc. B. **275** 19-28. (2008)
28. van Baalen, M.: Pair Approximations for different spatial geometries. In: The Geometry of Ecological Interactions: Simplifying Complexity, eds. Dieckmann, U., Law, R. & Metz, J. A. J., 359-387. (2000)
29. Volz, E.: SIR dynamics in random networks with heterogeneous connectivity. J. Math. Biol. **56** 293-310 (2008)
30. Watts, D. J., Strogatz, S. H.: Collective dynamics of “small-world” networks. Nature. **393**, 440-442 (1998)