Deterministic epidemic models on contact networks: Correlations and unbiological terms

Kieran J. Sharkey

1 Abstract

The relationship between system-level and subsystem-level master equations is investigated and then utilised for a systematic and potentially automated derivation of the hierarchy of moment equations in a susceptible-infectious-removed (SIR) epidemic model. In the context of epidemics on contact networks we use this to show that the approximate nature of some deterministic models such as mean-field and pair-approximation models can be partly understood by the identification of implicit anomalous terms. These terms describe unbiological processes which can be systematically removed up to and including the nth order by nth order moment closure approximations. These terms lead to a detailed understanding of the correlations in network-based epidemic models and contribute to understanding the connection between individual-level epidemic processes and population-level models. The connection with metapopulation models is also discussed. Our analysis is predominantly made at the individual level where the first and second order moment closure models correspond to what we term the individual-based and pair-based deterministic models, respectively. Matlab code is included as supplementary material for solving these models on transmission networks of arbitrary complexity.

2 Introduction

Epidemic dynamics are driven by processes which are typically stochastic in nature (Bartlett, 1956; Bailey, 1975). Nevertheless the probabilities of these processes can often be represented or approximated deterministically by a differential form of the Chapman-Kolmogorov equation known as the master equation. This comprehensive set of differential equations describes how the probabilities of the states of a system evolve in time. They are usually too numerous to evaluate numerically although they have been shown to be relevant for small homogeneous epidemic systems (Keeling and Ross, 2008). While numerical solutions remain problematic for systems of any significant size and complexity, master equations do permit exact stochastic realisations using the Gillespie algorithm (Gillespie, 1976; Renshaw, 1991), effectively regenerating the original stochastic epidemic processes.

Master equations are most adept at describing exponentially distributed stochastic processes. This, coupled with the close connection between master equations and other deterministic descriptions of epidemics underlies the almost ubiquitous use of "rates" such as the force of infection and the rate of removal in the design of deterministic epidemic models (Anderson and May, 1991). Although other distributions can be used in principle, they typically correspond to non-Markovian master equations and this can present significant implementation difficulties. It is worth noting that this is a relatively generic limitation of the deterministic approach.

Several deterministic methods for representing epidemics have been developed. These include the mean-field models (Kermack and Mckendrick, 1927; Anderson and May, 1991), pair-approximations (Matsuda et al., 1992; Keeling, 1999; Rand, 1999; van Baalen, 2000; Eames and Keeling, 2002; Murrell et al., 2004; Sharkey et al., 2006), and metapopulation models (Levins, 1969; Sattenspiel and Dietz, 1994; Keeling and Rohani, 2008). All of these attempt to approximate the average time course of an epidemic. Fundamentally, the average time course is implicit in the master equation, but the specific assumptions needed to relate this equation to particular deterministic epidemic models are not always clear. Identification of the relevant assumptions behind these classic deterministic models would certainly enhance our understanding of their domain of applicability and their relationship to the underlying stochastic processes. Indeed, the importance of understanding the basic connection between individual-level processes and population-level deterministic models has been emphasised several times (e.g. Levin and Durrett, 1996; Bansal et al., 2007).

One obvious approach is to start with the master equation and construct solvable deterministic models by applying specific assumptions to directly reduce the dimensionality of the state space. A classic example of this is the Fokker-Planck (or Kolmogorov-forward) equation (Risken, 1989) forming the theoretical justification for reaction-diffusion equation models of epidemics (Mollison, 1991; Murray, 2003). Another is the van Kampen linear noise approximation to the master equation (van Kampen, 2007) leading to a perturbative volume expansion approach (McKane and Newman, 2004; Ovaskainen and Cornell, 2006). More recently, other more computationally intensive methods of dimensional reduction have been investigated (Sharkey, 2008; Keeling and Ross, 2009; Simon et al., 2010).

In previous work (Sharkey, 2008), a systematic deconstruction of a deterministic SIR epidemic model on arbitrary transmission networks was used to illustrate the connections between the master equation, the network-based mean-field models and the network-based pair-approximation models. Additionally, the individual-based and pair-based deterministic models were constructed. Two assumptions connect these four types of model together with the master equations:- statistical independence and homogeneity. Here we extend this understanding by showing that implicit in the assumption of independence are anomalous terms describing unbiological processes. These terms enable us to understand some of the inaccuracies in network-based deterministic models in a more analytic way than was previously possible.

A side-product of this analysis is a systematic method for obtaining the hierarchy of moment equations at the individual level. The closed form of the first and second order moment equations correspond to the individual-based and pair-based models respectively. Matlab code is provided as supplementary material to solve these models on static network-based systems of arbitrary complexity.

In addition to master equations for systems, master equations for subsections of systems can also be written down. We refer to these subsections as subsystems (Sharkey 2008). For the present work we start by making a detailed investigation of the relationships between these equations. In particular, we show in the next section that the subsystem master equations follow as a consequence of the system master equation. We also show (section 4) that conversely, the master equation of a system can be obtained from the master equations of its subsystems provided that the subsystems are statistically independent and, collectively, fully specify the system state. The relevance of this construction for an epidemic system is then briefly introduced.

Section 5 elaborates on the main context for the present work which is a fixed-population susceptible-infectious-removed (SIR) compartmental model on a contact network. It puts this in the context of the general discussion of subsystems and systems, illustrating how moment equations can be derived as a consequence of the system master equation. It also highlights the assumption of pairwise statistical independence which is used to close the first order moment equations. Section 6 discusses the link between this construction and network-based mean-field models and metapopulation models.

Sections 7 and 8 use the results of section 4 to generate a better understanding of the problems with the pairwise independence assumption. In particular, we show that it generates implicit terms with no obvious interpretation and that these terms allow us to understand the failure of the assumption for certain contact networks. Sections 9 and 10 show how this analysis can be systematically extended beyond the pair level to all orders.

3 Systems, subsystems and master equations

Following prior work (Sharkey, 2008), we start by considering the state Γ^α of an arbitrary system Γ. The probability of the system being in state Γ^α is numerically equivalent to the expectation value $\langle\Gamma^\alpha\rangle$ where here, Γ^α represents a number which has value 1 when the system is in state Γ^α and zero otherwise. We will therefore use $\langle\Gamma^\alpha\rangle$ to denote both the probability of the state and/or its expectation value. In this notation, the master equation for Γ is:

$$\dot{\langle\Gamma^\alpha\rangle} = \sum_\beta \sigma^{\alpha\beta}\langle\Gamma^\beta\rangle - \sum_\beta \sigma^{\beta\alpha}\langle\Gamma^\alpha\rangle \tag{1}$$

where $\sigma^{\alpha\beta}$ denotes the transition rate from state Γ^β to state Γ^α and here and in what follows, the summations are over all possible system states. Note that this more conventional index ordering is opposite to that used in Sharkey (2008).

To avoid any ambiguity, the diagonal elements of all transition matrices in this paper are defined to be zero.

We suppose that within the system Γ, there exist well-defined smaller systems which we refer to as subsystems. We denote these subsystems by ψ_i where the index i distinguishes one subsystem from another. In the next section we will assume that the subsystems do not overlap and that they collectively specify the full system state without ambiguity. Presently we just need to suppose that at least one subsystem of Γ can be identified in an unambiguous manner. We can now write down a set of master equations for the individual subsystem states:

$$\langle \dot{\psi_i^a} \rangle = \sum_b \omega_i^{ab} \langle \psi_i^b \rangle - \sum_b \omega_i^{ba} \langle \psi_i^a \rangle \tag{2}$$

where ω_i^{ab} denotes the transition rate from state ψ_i^b to state ψ_i^a for the subsystem ψ_i. Here and throughout the paper, summations are assumed to be over all of the subsystem states available to ψ_i. We also denote system states by Greek superscripts and subsystem states by Roman superscripts.

Both the system and subsystem master equations must be valid and it is instructive to determine the conditions under which one can be derived as a consequence of the other. Let us attempt to obtain equation 2 from equation 1. We start with the probability of subsystem ψ_i being in state ψ_i^a which is given by the sum of the probabilities of the system states for which ψ_i is in state ψ_i^a:

$$\langle \psi_i^a \rangle = \sum_\alpha \langle \Gamma^\alpha \rangle D_i^{\alpha a} \tag{3}$$

where $D_i^{\alpha a}$ is a Kronecker-type delta in the states of the subsystems ψ_i such that it has value 1 if the system state Γ^α implies the subsystem state ψ_i^a and zero otherwise:

$$D_i^{\alpha a} = \begin{cases} 1 & \text{if } \Gamma^\alpha \Rightarrow \psi_i^a \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

Taking the derivative of equation 3 with respect to time and substituting from equation 1 gives:

$$\begin{aligned} \langle \dot{\psi_i^a} \rangle &= \sum_\alpha \langle \dot{\Gamma^\alpha} \rangle D_i^{\alpha a} \\ &= \sum_{\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha a} - \sum_{\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^\alpha \rangle D_i^{\alpha a} \\ &= \sum_{\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha a} \left[\sum_b D_i^{\beta b} \right] - \sum_{\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^\alpha \rangle D_i^{\alpha a} \left[\sum_b D_i^{\beta b} \right] \\ &= \sum_{b\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha a} D_i^{\beta b} - \sum_{b\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^\alpha \rangle D_i^{\alpha a} D_i^{\beta b} \end{aligned} \tag{5}$$

Swapping the dummy indices α and β in the second term on the right gives:

$$\langle \dot{\psi_i^a} \rangle = \sum_{b\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha a} D_i^{\beta b} - \sum_{b\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\beta a} D_i^{\alpha b} \tag{6}$$

Note that the quantity $\sum_{\alpha\beta}\sigma^{\alpha\beta}\langle\Gamma^{\beta}\rangle D_i^{\alpha a}D_i^{\beta b}$ gives the total flow of probability from state ψ_i^b to state ψ_i^a. By defining ω_i^{ab} by:

$$\omega_i^{ab}\langle\psi_i^b\rangle = \sum_{\alpha\beta}\sigma^{\alpha\beta}\langle\Gamma^{\beta}\rangle D_i^{\alpha a}D_i^{\beta b} \tag{7}$$

we obtain equation 2 as expected. Using equation 3 we can write the subsystem transition rates entirely in terms of the system transition rates and system states:

$$\omega_i^{ab} = \frac{\sum_{\alpha\beta}\sigma^{\alpha\beta}\langle\Gamma^{\beta}\rangle D_i^{\alpha a}D_i^{\beta b}}{\sum_{\alpha}\langle\Gamma^{\alpha}\rangle D_i^{\alpha b}} \tag{8}$$

It is no surprise that subsystem master equations are implicit in the system master equation, but it is valuable to formally relate the two. We return to this in the context of epidemic models in section 5 and show that this allows us to formally derive individual-level moment equations from the epidemic system master equation.

4 Statistical independence of subsystems

In the previous section we showed how to obtain subsystem master equations from the system master equation. Here we show that the converse problem of deriving the system master equation from the subsystem master equations can also be achieved in the specific case where the subsystems are assumed to be statistically independent.

We assume that the system Γ can be subdivided into a set of N subsystems $\psi_1, \psi_2, ...\psi_N$, each of which is entirely contained within Γ and defined such that the states of the N subsystems collectively specify the state of Γ without ambiguity.

Assuming statistical independence in the states of individual subsystems gives:

$$\langle\Gamma^{\alpha}(\psi_1^a, \psi_2^a ... \psi_N^a)\rangle = \langle\psi_1^a\rangle\langle\psi_2^a\rangle ... \langle\psi_N^a\rangle = \prod_{i=1}^{N}\langle\psi_i^a\rangle \tag{9}$$

where the subscript i should be taken to differentiate between the subsystem states. Here the system state dependence on the subsystem states is made explicit and this is implicitly assumed in equation 10 to equation 12 below. Differentiating equation 9 with respect to time gives:

$$\dot{\langle\Gamma^{\alpha}\rangle} = \sum_{j=1}^{N}\dot{\langle\psi_j^a\rangle}\prod_{i,i\neq j}\langle\psi_i^a\rangle \tag{10}$$

where the product is assumed to be over all N subsystems apart from j. Substituting from the individual subsystem master equations (equation 2) gives:

$$\dot{\langle\Gamma^{\alpha}\rangle} = \sum_{j=1}^{N}\left[\sum_b \omega_j^{ab}\langle\psi_j^b\rangle - \sum_b \omega_j^{ba}\langle\psi_j^a\rangle\right]\prod_{i,i\neq j}\langle\psi_i^a\rangle \tag{11}$$

By re-applying equation 9 we get:

$$\langle \dot{\Gamma^\alpha} \rangle = \sum_{j=1}^{N} \sum_b \omega_j^{ab} \langle \Gamma^\beta(\Gamma^\alpha, \psi_j^b) \rangle - \sum_{j=1}^{N} \sum_b \omega_j^{ba} \langle \Gamma^\alpha \rangle \tag{12}$$

regenerating the master equation for the full system. Here the states Γ^β only differ from the state Γ^α by the state of individual ψ_j and this is illustrated by a functional dependence on Γ^α and ψ_j^b. We can recast this equation in exactly the form of equation 1 with a general sum over all system states Γ^β where we specify:

$$\sigma^{\alpha\beta} = \begin{cases} \omega_j^{ab} & \text{if } \Gamma^\alpha \text{ differs from } \Gamma^\beta \text{ by (at most) the} \\ & \text{state of the single subsystem } \psi_j \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

This definition is equivalent to the following equation:

$$\sigma^{\alpha\beta} D_j^{\alpha a} D_j^{\beta b} = \omega_j^{ab} D_j^{\alpha a} D_j^{\beta b} \zeta_j^{\alpha\beta} \tag{14}$$

which holds for any combination of j, a, b, α, β. Here, $\zeta_j^{\alpha\beta}$ is introduced as a notational convenience such that:

$$\zeta_j^{\alpha\beta} = \begin{cases} 1 & \text{if all subsystems except for } \psi_j \text{ (which may change)} \\ & \text{remain the same under the transition } \Gamma^\beta \to \Gamma^\alpha \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

Equation 14 is valid when the subsystems are statistically independent, but it must also be consistent with the definition in equation 7 which is always true. We can demonstrate consistency by substituting equation 14 into the right-hand side of equation 7 to give:

$$\begin{aligned} \sum_{\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha a} D_i^{\beta b} &= \sum_{\alpha\beta} \langle \Gamma^\beta \rangle \omega_i^{ab} D_i^{\alpha a} D_i^{\beta b} \zeta_i^{\alpha\beta} \\ &= \omega_i^{ab} \sum_\beta \langle \Gamma^\beta \rangle D_i^{\beta b} \sum_\alpha D_i^{\alpha a} \zeta_i^{\alpha\beta} \end{aligned} \tag{16}$$

Notice that the identity:

$$\sum_\alpha D_i^{\alpha a} \zeta_i^{\alpha\beta} = 1 \tag{17}$$

is true for any i, a, β because there is only one system state (which may be Γ^β) which is identical to Γ^β but with ψ_i in state ψ_i^a. We therefore have:

$$\sum_{\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha a} D_i^{\beta b} = \omega_i^{ab} \sum_\beta \langle \Gamma^\beta \rangle D_i^{\beta b} \tag{18}$$

Applying equation 3 then gives equation 7 as expected.

This derivation of equation 1 from equation 2 represents a relationship between the subsystem and system master equations provided that the subsystems are statistically independent. Furthermore, it also provides a derivation of the master equations of subsections of the full system by replacing equation 9 with products of some (rather than all) of the subsystem states such as for doublets and triplets:

$$\begin{aligned} \langle \psi_i^a \psi_j^a \rangle &= \langle \psi_i^a \rangle \langle \psi_j^a \rangle \\ \langle \psi_i^a \psi_j^a \psi_k^a \rangle &= \langle \psi_i^a \rangle \langle \psi_j^a \rangle \langle \psi_k^a \rangle \end{aligned} \tag{19}$$

We shall see that this formalism is particularly valuable in understanding deterministic epidemic models and the connection between individuals and the full epidemic system. In particular, we can use this approach to obtain the moment equations of epidemic models from approximate subsystem master equations. In this context we suppose that Γ represents a self-contained epidemic system. While noting that other divisions into subsystems are possible, here and in what follows, we identify the subsystems ψ_i with individual units (such as people, farms, cities etc). According to this specification, the individuals themselves are systems containing a single subsystem, pairs of individuals are systems containing two subsystems and, in general, a collection of N individuals is a system composed of N subsystems. Hence any epidemic system of N individuals contains N subsystems and, furthermore, contains a hierarchy of systems within it containing between $n = 1$ and $n = N$ subsystems. The remainder of the paper elaborates on this application.

5 The SIR epidemic model with pairwise independence

The underlying premise of most epidemic models is that infection is spread by contact between infectious and susceptible individuals. The most comprehensive method for representing contact structures between individuals is the contact network (e.g. Newman, 2002; Meyers et al., 2003). These are networks of transmission routes by which infectious agents may pass from one individual to another. They may be static for the duration of an epidemic or dynamic whereby contacts can change in time as the epidemic evolves.

A contact network is conveniently represented as an adjacency matrix G whose elements G_{ij} have a value 1 if there is a transmission route from an individual j to an individual i and 0 otherwise (where the index-ordering convention from the master equations is replicated to avoid confusion). Additionally self-contact is not relevant so $G_{ii} = 0$. It is also convenient to define a transmission network T to be a weighted contact network where the elements T_{ij} represent the infectious pressure acting on individual i due to an infected individual j. Frequently, the rates in the matrix T describe a combination of the risk of transmission per contact and the rate of that contact. We can allow T to change in time provided that the rate of change is slow compared with the contact

rates. When the rate of change of T is so great that it no longer represents an averaged infectious pressure, the very notion of a contact network becomes lost; this scenario is outside the scope of the present work but see Volz and Meyers (2007; 2009) for work in this direction.

Our main interest will be the SIR compartmental model on a generic transmission network T. Here the subsystems ψ_i (or i for brevity) can be in one of the 3 states: $\psi_i^\alpha \epsilon \{\psi_i^S, \psi_i^I, \psi_i^R\}$, or $\psi_i^\alpha \epsilon \{S_i, I_i, R_i\}$ for brevity. The only transitions which are permitted are infection ($S_i \to I_i$) and removal ($I_i \to R_i$).

Assuming an exponentially distributed removal process for individual i with "rate" g_i, we have the following model of infection and removal (Sharkey 2008):

$$\begin{aligned} \langle \dot{S}_i \rangle &= -\sum_j T_{ij} \langle S_i I_j \rangle \\ \langle \dot{I}_i \rangle &= \sum_j T_{ij} \langle S_i I_j \rangle - g_i \langle I_i \rangle \end{aligned} \tag{20}$$

where $\langle S_i I_j \rangle$ denotes the probability that i is susceptible and that j is infectious and where the equation for the subsystem state $\langle R_i \rangle$ is implicit in the requirement $\langle S_i \rangle + \langle I_i \rangle + \langle R_i \rangle = 1$. Here, summations are between 1 and the total population size N and this is tacitly assumed for the summations in the equations that follow unless explicitly indicated otherwise. These equations are well known to be exact given the specified transmission and removal rates described. Nevertheless, it is instructive to see exactly how these moment equations follow from the linear system master equation.

We consider the epidemic system Γ which is composed of N subsystems which correspond to the individuals connected by a transmission network T. We identify the moment equations for the individuals with the subsystem master equations. For an individual ψ_i, we first consider the state $\psi_i^S = S_i$. From the last line of equation 5 we have, after performing the sum over b for the SIR model:

$$\langle \dot{S}_i \rangle = -\sum_{\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^\alpha \rangle D_i^{\alpha S} D_i^{\beta I} \tag{21}$$

The epidemic system Γ is assumed to obey a continuous-time Markov process such that the probability of two subsystem states changing simultaneously is zero. Consequently $\sigma^{\beta\alpha} = 0$ whenever Γ^α differs from Γ^β by more than the state of a single subsystem. This sparsity of the transition matrix has been used by some authors to recast it in a tridiagonal form by a suitable ordering of the system states (Keeling and Ross, 2008; Simon et al., 2010). In our case, each system state Γ^α typically has many neighbouring states into which it can jump corresponding to the number of ways in which its individual subsystems may change. It is not possible to write this in a tridiagonal form in the general case, but we can make use of our notational convenience $\zeta_i^{\alpha\beta}$ defined in equation 15 to perform a similar role. The quantity $\sigma^{\beta\alpha} D_i^{\alpha S} D_i^{\beta I}$ in equation 21 is zero unless it corresponds to the system transition where S_i becomes I_i and no other subsystem changes state. According to the contact transmission mechanism

discussed above, this rate is given by the sum of the infection rates (or total infectious pressure) acting on S_i in this system state:

$$\sigma^{\beta\alpha} D_i^{\alpha S} D_i^{\beta I} = \sum_j T_{ij} D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta I} \zeta_i^{\beta\alpha} \tag{22}$$

which holds for all α, β, i. Substituting this into equation 21 gives:

$$\begin{aligned} \langle \dot{S}_i \rangle &= -\sum_{\alpha\beta} \sum_j T_{ij} \langle \Gamma^\alpha \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta I} \zeta_i^{\beta\alpha} \\ &= -\sum_j T_{ij} \sum_\alpha \langle \Gamma^\alpha \rangle D_i^{\alpha S} D_j^{\alpha I} \sum_\beta D_i^{\beta I} \zeta_i^{\beta\alpha} \end{aligned} \tag{23}$$

Applying equation 17 then leads to:

$$\begin{aligned} \langle \dot{S}_i \rangle &= -\sum_j T_{ij} \sum_\alpha \langle \Gamma^\alpha \rangle D_i^{\alpha S} D_j^{\alpha I} \\ &= -\sum_j T_{ij} \langle S_i I_j \rangle \end{aligned} \tag{24}$$

Repeating this argument for state I_i gives (from equation 5 after the sum over b):

$$\langle \dot{I}_i \rangle = \sum_{\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^\beta \rangle D_i^{\alpha I} D_i^{\beta S} - \sum_{\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^\alpha \rangle D_i^{\alpha I} D_i^{\beta R} \tag{25}$$

Assuming an exponentially distributed removal process for individual i with "rate" g_i, we have:

$$\sigma^{\beta\alpha} D_i^{\alpha I} D_i^{\beta R} = g_i D_i^{\alpha I} D_i^{\beta R} \zeta_i^{\beta\alpha} \tag{26}$$

Swapping the dummy indices α and β in the first term on the right of equation 25 and substituting from equations 22 and 26 then leads to:

$$\langle \dot{I}_i \rangle = \sum_j T_{ij} \langle S_i I_j \rangle - g_i \langle I_i \rangle \tag{27}$$

by following analogous steps to those that led from equation 21 to equation 24. Population-level versions of these moment equations follow for the case of homogeneous infection and removal rates (Sharkey, 2008). The reader is also directed to an alternative derivation in Simon et al. (2010) of the population-level moment equations from the master equation for the case of homogeneous transmission and removal rates.

The only non-zero transitions for an individual subsystem in the SIR epidemic model are ω_i^{IS} and ω_i^{RI}. Comparison with equations 3, 8, 25 and 27 then gives the subsystem transition rates:

$$\begin{aligned} \omega_i^{IS} &= \sum_{j=1} T_{ij} \frac{\langle S_i I_j \rangle}{\langle S_i \rangle} \\ \omega_i^{RI} &= g_i \end{aligned} \tag{28}$$

Part of the price paid for writing master equations at the subsystem level is that they are non-linear. A more serious cost is that these equations do not form a closed set. However, if we assume pairwise statistical independence in the subsystems we have:

$$\langle S_i I_j \rangle = \langle S_i \rangle \langle I_j \rangle \tag{29}$$

giving the non-linear equations:

$$\begin{aligned} \langle \dot{S}_i \rangle &= -\sum_j T_{ij} \langle S_i \rangle \langle I_j \rangle \\ \langle \dot{I}_i \rangle &= \sum_j T_{ij} \langle S_i \rangle \langle I_j \rangle - g_i \langle I_i \rangle \end{aligned} \tag{30}$$

Note that while pairwise statistical independence implies these equations, the converse is not true. In general these equations are approximate and the interpretation of the quantities $\langle S_i \rangle$ and $\langle I_i \rangle$ as probabilities is only consistent with the master equation when pairwise independence holds.

Following Sharkey (2008), we refer to equation 30 as the "individual-based model" in the remainder of this work. We can relate these approximate probabilities to the expected susceptible and infectious populations by $[S] = \sum_i \langle S_i \rangle$ and $[I] = \sum_i \langle I_i \rangle$ respectively. Matlab code for solving these equations for arbitrary transmission networks is in the supplementary material.

6 Mean-field and metapopulation models

The form of the classic mean-field SIR model follows from the individual-based model by applying an assumption of homogeneity: $\langle S_i \rangle = [S]/N$, $\langle I_i \rangle = [I]/N$, so that equation 30 becomes:

$$\begin{aligned} [\dot{S}] &= -\beta [I][S] \\ [\dot{I}] &= \beta [I][S] - \gamma [I] \end{aligned} \tag{31}$$

where $\beta = (1/N^2) \sum_{ij} T_{ij}$ and $\gamma = (1/N) \sum_i g_i$. Defined in this way, this population-level mean-field model follows from the master equation by applying the assumptions of pairwise independence and homogeneity (Keeling 1999; Keeling and Eames 2005; Sharkey, 2008). Here and in what follows we use the term "mean-field" as a short-hand to refer exclusively to population-level mean-field models.

It is worth pausing to give these equations some context. Mean-field epidemic models are predicated on the law of mass action which is only precisely applicable in idealised circumstances such as evenly mixed gasses or large, fully connected contact networks. To some extent mean-field models can be applied away from these ideals by absorbing some of their inaccuracies into effective contact and removal rates to give optimal matches to epidemic incidence data. However, it is well known that the exponential epidemic form that results from

these models is not generically applicable (e.g. Keeling, 2005) and, furthermore, the connection between the individual-level transmission parameters and the population-level model parameters becomes obscure. Equation 31 gives us a slightly different perspective in which we do not lose touch with the individual processes (β is still defined in terms of individual transmission rates), but the model becomes inaccurate due to failures in the independence and homogeneity assumptions used to obtain it from the master equation. While the focus of this paper is only on the failure of the independence assumption, it does provide insights into the failure of the network-based mean-field model even if partially obscured by the additional assumption of homogeneity.

Metapopulation models are also relevant to our discussion of population-level models and the mean-field assumption. In particular, we note that the individual-based model as defined in the previous section is form-equivalent to some deterministic metapopulation models of epidemics. To investigate this connection in more detail, let us consider a model with M metapopulations such that the population dynamics of metapopulation α are:

$$\begin{aligned} [\dot{S_\alpha}] &= -\sum_{\beta=1}^{M} \rho_{\alpha\beta}[S_\alpha][I_\beta] \\ [\dot{I_\alpha}] &= \sum_{\beta=1}^{M} \rho_{\alpha\beta}[S_\alpha][I_\beta] - \gamma[I_\alpha] \end{aligned} \tag{32}$$

where here, $[S_\alpha]$ and $[I_\alpha]$ correspond to the number of susceptible and infectious individuals in metapopulation α respectively and $\rho_{\alpha\beta}$ is the rate at which infection is spread from metapopulation β to metapopulation α (e.g. Keeling and Rohini, 2008). Here we have assumed a fixed removal rate γ. The so-called "household" models can also have this form (e.g. Ball et al., 1997; Ball et al., 2009; Ross et al., 2010) where households are identified with metapopulations with one mode of infection (here represented by homogeneous mixing) and infection between households (here represented by a fixed contact network) are distinguished by a separate process or mechanism.

To identify a connection with the individual-based model, it is convenient to index individuals as α_i where the index i runs over all N_α individuals in metapopulation α. We can then write the expected populations as:

$$\begin{aligned} [S_\alpha] &= \sum_{i=1}^{N_\alpha} \langle S_{\alpha_i} \rangle \\ [I_\alpha] &= \sum_{i=1}^{N_\alpha} \langle I_{\alpha_i} \rangle \end{aligned} \tag{33}$$

Let us consider a network T of contacts between individuals such that all individuals within a specific metapopulation α are connected to each other with strength $T_{\alpha_i\alpha_j} = \rho_{\alpha\alpha}(1 - \delta_{\alpha_i\alpha_j})$ and all individuals in different metapopulations α and β are connected with strength $T_{\alpha_i\beta_j} = \rho_{\alpha\beta}$. Since the mean-field

and individual-based models are equivalent in sufficiently large fully connected populations (Sharkey, 2008), this effectively generates a mean-field model for the internal metapopulation dynamics. Using the individual-based model and equation 33, we can now write:

$$\begin{aligned}
[\dot{S_\alpha}] &= \sum_{i=1}^{N_\alpha} \langle \dot{S_{\alpha_i}} \rangle = -\sum_{i=1}^{N_\alpha} \sum_{\beta=1}^{M} \sum_{j=1}^{N_\beta} T_{\alpha_i \beta_j} \langle S_{\alpha_i} \rangle \langle I_{\beta_j} \rangle \\
&= -\sum_{i=1}^{N_\alpha} \sum_{\beta=1}^{M} \sum_{j=1}^{N_\beta} \rho_{\alpha\beta} \langle S_{\alpha_i} \rangle \langle I_{\beta_j} \rangle (1 - \delta_{\alpha_i \beta_j}) \\
&= -\sum_{\beta=1}^{M} \rho_{\alpha\beta} [S_\alpha][I_\beta] + \rho_{\alpha\alpha} \sum_{i=1}^{N_\alpha} \langle S_{\alpha_i} \rangle \langle I_{\alpha_i} \rangle
\end{aligned} \tag{34}$$

recovering the metapopulation model above except for the explicit removal of the self-interaction terms which becomes negligible in large metapopulations. Similarly we can obtain an equation for $[I_\alpha]$. For large populations, the stochastic error becomes less relevant provided that the epidemic has taken off and the expectation values $[I_\alpha]$ and $[S_\alpha]$ can become approximately identified with the actual population sizes during a single stochastic realisation. So, in the context of a particular type of static network, we can at least draw a connection with the form of some metapopulation models.

Clearly when the size of each metapopulation reduces, the probabilistic interpretation in terms of expectation values must return. When each metapopulation has a single individual, the interpretation is entirely probabilistic, similar to a Levins-type metapopulation model (Levins, 1969) but with a network structure between "populations". This connection is particularly apparent in the context of an SIS epidemic model as observed by Keeling and Ross (2008).

Some care should be taken not to push this discussion beyond its relevant context. In particular, metapopulation models of epidemics often treat the movement of individuals from one population into another as the cause of transmission between populations rather than contact between individuals in different populations. On an individual level this corresponds to a dynamic contact network which is coupled to the infection process. This type of process is clearly outside the scope of the individual-level models considered in this paper.

7 The anomalous terms at order $n = 2$

The individual-based model defined in section 5 is consistent with the system master equation provided that pairwise statistical independence holds. However, this is often a poor assumption; for example, the nearest neighbours of infectious individuals are obviously more likely to be infectious than average which immediately implies correlation between neighbouring sites.

In general the individual-based model is inconsistent with the interpretation of $\langle S_i \rangle$ and $\langle I_i \rangle$ as probabilities because it causes a departure from the master equation description. Here we show that the assumption of pairwise statistical independence of the states S_i and I_j in equation 29 gives rise to terms in the differential equation for the probability of the doublet state $\langle S_i I_j \rangle$ that have no obvious interpretation. Correction of these "anomalous" terms automatically induces correlation between pairs of sites, breaking pairwise independence and regaining consistency with the system master equation.

Before examining this further, we start by observing that the equation for the doublet state $\langle S_i I_j \rangle$ is given by:

$$\langle \dot{S_i I_j} \rangle = \sum_{k,k \neq i} T_{jk} \langle S_i S_j I_k \rangle - \sum_{k,k \neq j} T_{ik} \langle S_i I_j I_k \rangle - T_{ij} \langle S_i I_j \rangle - g_j \langle S_i I_j \rangle \qquad (35)$$

where quantities of the form $\langle A_i B_j C_k \rangle$ represent the probability that i is in state A, j is in state B and k is in state C. This equation can be shown to follow in a top-down fashion from the system master equation by a slightly laborious but straightforward argument along the same lines as for the single subsystem states in section 5 (see appendix A).

Here our aim is two-fold:- firstly to generate a better understanding of the failure of pairwise independence by looking at deviations from equation 35 and secondly to obtain higher order equations such as equation 35 in a systematic way from the equations for subsystems. The general basis for this analysis is described in section 4 where we showed that higher order moment equations can be derived from subsystem master equations provided that we assume that the subsystems are statistically independent. In the current context we use pairwise statistical independence to derive the corresponding equation for the doublet state $S_i I_j$ from the subsystem master equations.

With pairwise independence, the rate of change of the $n = 2$ (doublet) system state is given by:

$$\langle \dot{S_i I_j} \rangle = \langle \dot{S_i} \rangle \langle I_j \rangle + \langle S_i \rangle \langle \dot{I_j} \rangle \qquad (36)$$

Pairwise independence implies that the individual-based model becomes an exact representation of the subsystem master equations. Substituting from equation 30 into the above gives:

$$\langle \dot{S_i I_j} \rangle = \sum_{k} T_{jk} \langle S_i \rangle \langle S_j \rangle \langle I_k \rangle - \sum_{k} T_{ik} \langle S_i \rangle \langle I_j \rangle \langle I_k \rangle - g_j \langle S_i \rangle \langle I_j \rangle \qquad (37)$$

Careful consideration of the sums in this expression reveals that they contain terms that refer twice to the same site:

$$\begin{aligned} \langle \dot{S_i I_j} \rangle \quad = \quad & \sum_{k,k \neq i} T_{jk} \langle S_i \rangle \langle S_j \rangle \langle I_k \rangle - \sum_{k,k \neq j} T_{ik} \langle S_i \rangle \langle I_j \rangle \langle I_k \rangle - g_j \langle S_i \rangle \langle I_j \rangle \\ & + T_{ji} \langle S_i \rangle \langle S_j \rangle \langle I_i \rangle - T_{ij} \langle S_i \rangle \langle I_j \rangle \langle I_j \rangle \end{aligned} \qquad (38)$$

This equation represents the rate of change of the probability $\langle S_i I_j \rangle$ provided that the system states satisfy pairwise independence. More generally it represents the rate of change of the quantity $\langle S_i \rangle \langle I_j \rangle$ where $\langle S_i \rangle$ and $\langle I_j \rangle$ are defined as the solution of the individual-based model.

From a process point of view, the first term represents the generation of $S_i I_j$ by the infection of site j. The second term represents the destruction of $S_i I_j$ by infection of site i and the third term represents the destruction of $S_i I_j$ by the removal of site j. However, the fourth term has no relevant process counterpart. Indeed, the term $T_{ji}\langle S_i \rangle \langle S_j \rangle \langle I_i \rangle$ is inconsistent with the assumption of a compartmental model because it corresponds to a process in which site i is both infectious and susceptible at the same time. Additionally, the last term is clearly related to the self-destruction of $S_i I_j$ caused by the infection of site i by site j, however its form is not standard.

Comparison with the correct pair-level equation (equation 35) for the rate of change of the $n = 2$ system state probability $\langle S_i I_j \rangle$ highlights that pairwise statistical independence is inconsistent with a probabilistic interpretation and that this inconsistency manifests itself in two terms. The closer $\langle S_i \rangle \langle S_j \rangle \langle I_i \rangle$ is to zero and the closer $\langle S_i \rangle \langle I_j \rangle \langle I_j \rangle$ is to $\langle S_i \rangle \langle I_j \rangle$, the closer the individual-based model is to the underlying probabilistic description of the master equation. In this sense, these two anomalous terms quantify the accuracy of the individual-based model.

Although equation 38 is incorrect as an equation for $\langle S_i I_j \rangle$, it is not far from the correct form. Furthermore, it is obtained in a systematic way which does not demand the usual care to ensure that every possible process is accounted for. If we can systematically go from equations of the form of equation 38 to the correct equations then this would provide a potentially automated way of obtaining systems of moment equations. In this instance at least, the process is straightforward. We take the heuristic but systematic step of replacing the products of probabilities with combined probabilities:

$$\begin{aligned} \dot{\langle S_i I_j \rangle} &= \sum_{k,k\neq i} T_{jk}\langle S_i S_j I_k \rangle - \sum_{k,k\neq j} T_{ik}\langle S_i I_j I_k \rangle - g_j \langle S_i I_j \rangle \\ &\quad + T_{ji}\langle S_i S_j I_i \rangle - T_{ij}\langle S_i I_j I_j \rangle \end{aligned} \tag{39}$$

Now, the anomalous terms are readily identified as those terms in which the same site appears twice. Where the states contradict each other, we must set the probability to zero ($\langle S_i S_j I_i \rangle = 0$). Where they are same, they are contracted ($\langle S_i I_j I_j \rangle = \langle S_i I_j \rangle$) using the fact that the conditional probability of j being infectious given that j is infectious is 1. This results in equation 35 which we know to be exact. In section 10 we shall see that this systematic procedure applies beyond the pair level to all orders of n although it remains a conjecture that this always gives equations which are consistent with the system master equation.

The correction of the anomalous terms means that equation 35 is not the derivative of the product $\langle S_i \rangle \langle I_j \rangle$ since this is given by equation 38. Hence correction of the anomalous terms can be viewed as automatically breaking

the pairwise independence assumption and implies correlations between pairs of sites. To preserve these pairwise correlations and form a closed model at the pair level, we approximate the triples probabilities by:

$$\langle A_i B_j C_k \rangle = \frac{\langle A_i B_j \rangle \langle A_i C_k \rangle \langle B_j C_k \rangle}{\langle A_i \rangle \langle B_j \rangle \langle C_k \rangle} \tag{40}$$

This approximation is based on the statistical independence of the dynamics between pairs and preserves the pair-level correlations induced by the correction of the anomalous terms. It does not preserve correlations that may occur at the level of triples which may be implicit in the triplet states in equation 35. We delay consideration of correlations among triples until section 9. The form of closure given by equation 40 is discussed in more detail in appendix B and elsewhere (e.g. Kirkwood 1935; Rand 1999; Singer 2004). In subsequent sections we show that higher order correlations beyond the pair level can also be understood as being induced by the correction of higher order anomalous terms.

With this closure approximation, along with a refinement introduced for computational purposes (see appendix B), together with the pair-based equations:

$$\begin{aligned}
\langle \dot{S_i S_j} \rangle &= -\sum_{k,k\neq i} T_{jk} \langle S_i S_j I_k \rangle - \sum_{k,k\neq j} T_{ik} \langle S_i S_j I_k \rangle \\
\langle \dot{I_i I_j} \rangle &= \sum_{k,k\neq i} T_{jk} \langle I_i S_j I_k \rangle + \sum_{k,k\neq j} T_{ik} \langle S_i I_j I_k \rangle + T_{ij} \langle S_i I_j \rangle \\
&\quad + T_{ji} \langle I_i S_j \rangle - (g_i + g_j) \langle I_i I_j \rangle
\end{aligned} \tag{41}$$

and equation 20, we obtain a closed system of equations at the individual pair level which do not contain the two anomalous terms just identified. We refer to this system as the pair-based model (Sharkey 2008). The differential equations for $\langle S_i S_j \rangle$ and $\langle I_i I_j \rangle$ can be formally derived from the master equation in same way as equation 35 (see appendix A) or obtained by the heuristic approach introduced in this section. Matlab code is provided in the supplementary material for solving the pair-based model on transmission networks of arbitrary complexity. For the case of homogeneous transmission and removal rates the population-level pair equations follow from these equations by summing over all possible pairs and, using assumptions of homogeneity of pairs and of individuals, the population-level pair-approximation models can also be obtained (Sharkey, 2008).

8 The impact of the $n = 2$ anomalous terms

It is straightforward to see that the impact of the anomalous terms identified in the previous section is greater for networks with low connectivity. The two correct transmission terms in equation 39 contain a sum over k which, in conjunction with the transmission network T, corresponds to a sum from 1 to p where p is the number of immediate neighbours with a link towards j in the first

term and a link towards i in the second. Therefore, the relative significance of the anomalous terms on the differential equations diminishes as roughly $1/p$ so the performance of the individual-based model improves with increasing connectivity. We therefore conclude that the individual-based and pair-based models converge as the number of neighbours of each individual becomes large. This is consistent with well-known population-level results on the convergence of the mean-field and pair-approximation models with increasing network connectivity (e.g. Keeling 1999).

Notice that the fourth term on the right of equation 39 is only present for symmetric (undirected) links on the contact network (since it requires a contact from i to j and we already assume a contact from j to i to carry the infection dynamics). This term is positive which has the effect of increasing the rate of increase of $\langle S_i I_j \rangle$. This implies a greater exaggeration of the rate of spread of epidemics on symmetric networks than on asymmetric networks. We can also make a mathematical argument suggesting that the effect of this term should be minimised by a large infection to removal rate ratio corresponding to a large basic reproductive ratio R_0 (see appendix C for further elaboration).

The fifth term on the right of equation 39 is present for all networks whether symmetric or asymmetric. It is a negative term which is smaller than it should be. This term gives the individual-based model a tendency to exaggerate the severity of epidemics on all contact networks. In general, the error caused by this term depends on the approximation $\langle I_j \rangle \approx \langle I_j \rangle^2$ which is least accurate when $\langle I_j \rangle = 0.5$ and exact when $\langle I_j \rangle$ is zero or one. The impact of this term can be minimised (but not removed) by a large infection to removal rate ratio (see appendix C). Notice that both of the anomalous terms act in the same direction; they both lead to an exaggeration of the spread of an epidemic and so their effects do not cancel.

For illustrative purposes, these conclusions are demonstrated in figure 1 on a particular class of random network where every individual has the same number of incoming and outgoing connections. Although the analysis above is generic, these random k-regular networks are particularly simple and enable us to understand much of the epidemic behaviour by consideration of the two terms just identified.

Figure 1 shows a numerical comparison of the predictions of the individual-based and pair-based models with the "correct" expected infectious population obtained by stochastic simulation of the master equation for a set of six random k-regular networks. Three of these are undirected networks and three are directed fully asymmetric networks (similar but finite versions of those considered by Diekmann et al. (1998)). Also shown are the predictions of the population-level mean-field model defined in section 6 illustrating the effect of the assumption of homogeneity. The accuracy of the pair-based model indicates that the errors caused by higher order effects associated with correlations between triples which are ignored by equation 40 are minimal for these simulations and that the two anomalous terms identified above are at the root of the observed inaccuracies of the individual-based model.

It is clear from this figure that our qualitative conclusions are supported

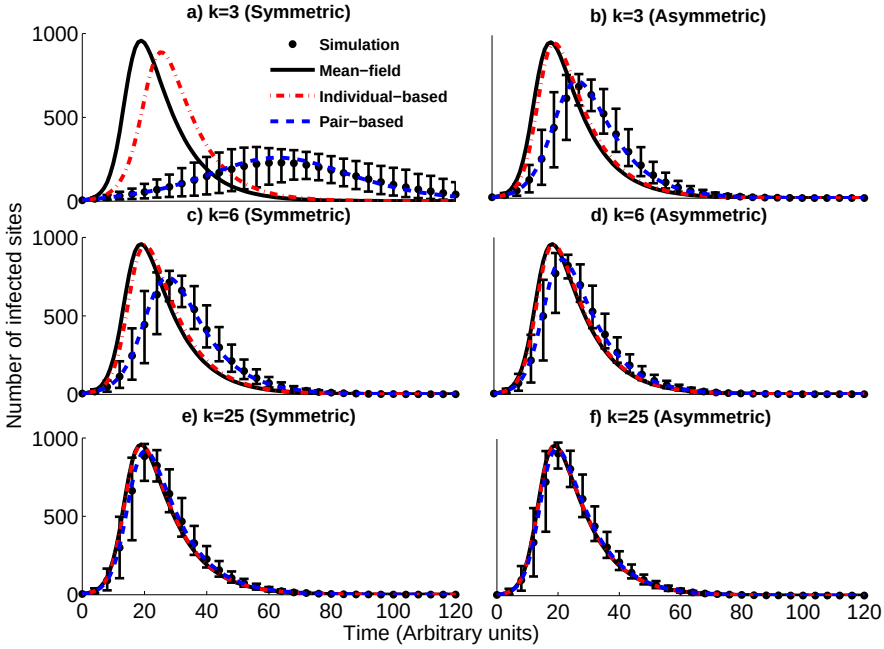

Figure 1: Comparison of the infectious population time series of the mean-field, individual-based and pair-based models against simulated epidemics on six random k-regular contact networks where three (a,c,e) are fully symmetric (undirected) networks, three (b,d,f) are fully asymmetric (directed) networks and where the number of neighbours per individual in subplots (a,b), (c,d) and (e,f) are 3, 6, and 25 respectively. The network size N is 2000 in each case. The mean of 1000 stochastic simulations is plotted and the error bars illustrate the 5th and 95 percentiles. The removal rate is held fixed at $g_i = 0.1$ for each simulation where we are using arbitrary units of time. The transmission rate τ is the same across each link and is chosen such that $N\beta = k\tau = 0.5$ so that the mean-field prediction is the same for each network. Each stochastic simulation is initiated on 5 initial infected sites that do not change. The individual and pair-based models are initiated on the same 5 sites.

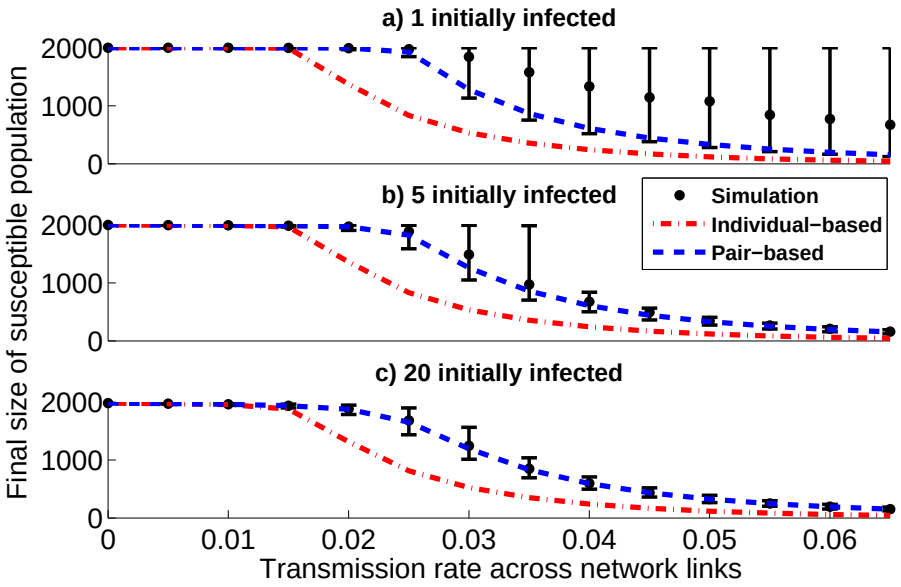

Figure 2: Comparison of the predictions of the individual-based and pair-based models with stochastic simulation close to the epidemic threshold. An undirected random 6-regular network is used with network size $N = 2000$. The epidemics are initiated with a) 1 infected, b) 5 infected and c) 20 infected sites. The mean of 1000 stochastic simulations is shown with error bars indicating the 5th and 95th percentiles. The removal rate is held fixed at $g_i = 0.1$.

on this class of network; the individual-based model performs badly on the low connectivity networks but appears to converge towards the pair-based model with increasing network connectivity, and, with all other aspects being equal, the performance of the individual-based model is substantially better for asymmetric networks than for symmetric networks.

All of the simulations considered so far have been initiated with five infected individuals and using relatively large transmission rates. Figure 2 illustrates the predictions for the final size of the susceptible population with respect to transmission rates for an undirected random k-regular network with $k = 6$ (corresponding to figure 1c). Here, low transmission rates around the epidemic threshold are investigated. Additionally, epidemics are initiated with 1, 5 and 20 infected sites. These plots are consistent with our analysis of the anomalous terms which suggested that the individual-based and pair-based models are most consistent when the transmission rate is large and, trivially, when there is no epidemic (appendix C).

The performance of the pair-based model in describing the average stochastic simulation is seen to decrease markedly when the epidemic is initiated with a single individual. This is expected to some extent as a consequence of the

bimodal response of either stochastic fade out or large epidemics. While the probability of stochastic fade out is implicit in the individual-based and pair-based models, additional errors will be present from neglecting higher order effects. The pair-based model accounts for the leading order errors, but small, higher order errors exist and are ignored by the approximation in equation 40. These errors are likely to be very small on a k-regular random network due to the low level of clustering (see the following two sections), however even a small error in the probability of an epidemic to take off will have a sizeable impact on the final size prediction due to the bimodal response. By increasing the number of initial infected sites, the probability of stochastic fade out is reduced, effectively returning a unimodal response and removing this error.

For other classes of network, the pair-based model also becomes inaccurate even without stochastic fade out (Sharkey, 2008). This is also caused by higher order anomalous terms which we discuss in the next two sections.

9 The anomalous terms at order $n = 3$

We have constructed a pair-based model which removes the effect of the anomalous terms at the pair level. However, further terms exist at higher orders which can cause this model to become inaccurate. This is because while equation 40 preserves the correlations between pairs induced by the correction of the $n = 2$ anomalous terms, it does not preserve any higher order correlations. We can examine these higher order correlations in more detail by considering triplewise independence:

$$\langle A_i B_j C_k \rangle = \langle A_i \rangle \langle B_j \rangle \langle C_k \rangle \tag{42}$$

Critically, triplewise independence implies and is implied by pairwise independence together with the additional requirement:

$$\langle A_i B_j C_k \rangle = \langle A_i \rangle \langle B_j C_k \rangle = \langle B_j \rangle \langle A_i C_k \rangle = \langle C_k \rangle \langle A_i B_j \rangle \tag{43}$$

For want of a better notation, we can refer to these additional relationships as the relative complement of 2wise independence with respect to 3wise independence: {3wise\2wise}; that is, the conditions implicit in 3wise independence which are not in 2wise independence.

With triplewise independence, we can again use the methodology developed in section 4 to obtain a master equation for the $n = 3$ system state $S_i S_j I_k$ which occurs in equation 35. Repeating the same procedure as for $n = 2$ of applying subsystem independence, differentiating and substituting from equation 30 (which holds because pairwise independence is implied by triplewise independence) we obtain:

$$\begin{aligned} \langle S_i \dot{S}_j I_k \rangle &= \sum_{l, l \notin \{i,j\}} T_{kl} \langle S_i \rangle \langle S_j \rangle \langle S_k \rangle \langle I_l \rangle - \sum_{l, l \notin \{k,i\}} T_{jl} \langle S_i \rangle \langle S_j \rangle \langle I_k \rangle \langle I_l \rangle \\ &- \sum_{l, l \notin \{k,j\}} T_{il} \langle S_i \rangle \langle S_j \rangle \langle I_k \rangle \langle I_l \rangle - g_k \langle S_i \rangle \langle S_j \rangle \langle I_k \rangle \end{aligned}$$

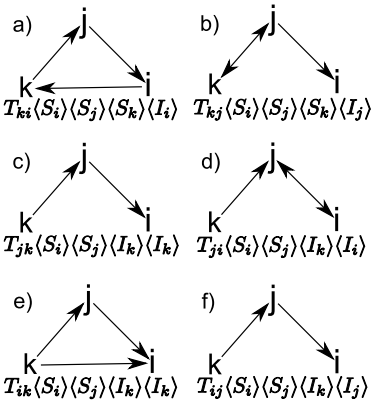

Figure 3: Graphs correspond to the six anomalous terms in equation 44 for the $n = 3$ moment equation of the state $S_iS_jI_k$. Graphs b,c,d and f are already corrected in the pair-based model. Terms a and e are implicit in the pair-based model and lead to errors in networks with a high density of sites connected as 3-cycles.

$$\begin{aligned} &+ \quad T_{ki}\langle S_i\rangle\langle S_j\rangle\langle S_k\rangle\langle I_i\rangle + T_{kj}\langle S_i\rangle\langle S_j\rangle\langle S_k\rangle\langle I_j\rangle \\ &- \quad T_{jk}\langle S_i\rangle\langle S_j\rangle\langle I_k\rangle\langle I_k\rangle - T_{ji}\langle S_i\rangle\langle S_j\rangle\langle I_k\rangle\langle I_i\rangle \\ &- \quad T_{ik}\langle S_i\rangle\langle S_j\rangle\langle I_k\rangle\langle I_k\rangle - T_{ij}\langle S_i\rangle\langle S_j\rangle\langle I_k\rangle\langle I_j\rangle \end{aligned} \tag{44}$$

Here the last six terms are readily identified as anomalous by the presence of duplicated sites. The type of network for which the impact of the anomalous terms will be greatest is dependent on the precise connectivity between the sites i, j and k. In terms of the dynamics, the triple $S_iS_jI_k$ is important when there is a contact from site k to site j and a contact from site j to site i. For a triple connected in this way, figure 3 illustrates the type of network structure corresponding to each of the six anomalous terms. The terms depicted in graphs b,c,d and f are not new and correspond to pairwise independence and are removed in the $n = 2$ moment closure model subject to the use of suitable pair-approximations which preserve pair-level correlations (see appendix B). The reoccurrence of these pairwise correlations is expected because pairwise independence is implicit in triplewise independence. Of particular interest here are the two new terms (graphs a and e). These distinguish an $n = 3$ moment closure model which accounts for correlations between triples from the $n = 2$ moment closure model which only accounts for correlations between pairs. They are a consequence of the independence assumption {3wise\2wise} and correspond to closed triples.

The process of correcting this equation is entirely systematic and identical to the $n = 2$ case. We first replace the products with combined probabilities. The anomalous terms are then identified by the presence of duplicated sites. These terms are removed if the states are contradictory or contracted if they

are the same. This results in the $n = 3$ moment equations:

$$\begin{aligned}\langle S_i\dot{S}_jI_k\rangle &= \sum_{l,l\notin\{i,j\}} T_{kl}\langle S_iS_jS_kI_l\rangle - \sum_{l,l\notin\{k,i\}} T_{jl}\langle S_iS_jI_kI_l\rangle \\ &- \sum_{l,l\notin\{k,j\}} T_{il}\langle S_iS_jI_kI_l\rangle - g_k\langle S_iS_jI_k\rangle \\ &- T_{jk}\langle S_iS_jI_k\rangle - T_{ik}\langle S_iS_jI_k\rangle \end{aligned} \quad (45)$$

This correction induces both correlations between pairs (graphs b,c,d and f) as well as new correlations at the triple level (graphs a and e). The errors resulting from these new anomalous terms cannot be removed at the pair level. Consequently, inaccuracy is expected in the pair-based model for networks with closed triples even when there is no clustering present at higher order. See Keeling (1999, figure 2) for an observation of this effect as well as the results at the end of section 10 on a network composed of a single closed triple (or 3-cycle). Here and in what follows we use the graph theory terms "path" and "cycle" to encompass weakly connected paths and cycles.

The two new anomalous terms act to increase the rate of increase of $\langle S_iS_jI_k\rangle$ which in turn increases the rate of increase of $\langle S_iI_j\rangle$ via equation 35. These terms therefore lead the pair-based model to exaggerate the spread of an epidemic as in the case of the terms at the previous order. However, it is not always the case that the anomalous terms act to exaggerate the rate of spread of an epidemic. Repeating the analysis for the other triple $\langle S_iI_jI_k\rangle$ occurring in equation 35, we find the two anomalous terms $T_{ki}\langle S_i\rangle\langle I_j\rangle\langle S_k\rangle\langle I_i\rangle$ and $-T_{ik}\langle S_i\rangle\langle I_j\rangle\langle I_k\rangle\langle I_k\rangle$ which both act to increase the rate of increase of $\langle S_iI_jI_k\rangle$ which in turn leads to a reduction in the predicted rate of spread of the epidemic. Hence, terms at higher order do not always act in the same direction and their effects may cancel. Nevertheless, experience suggests that the pair-based model normally exaggerates the spread of an epidemic suggesting that if higher order anomalies can be ignored, the anomalous terms associated with the triplet state $S_iS_jI_k$ dominate over those associated with the triplet state $S_iI_jI_k$.

With a suitable closure approximation for 4-tuples that preserves correlations between triples, a model can be constructed where these anomalous terms at the triple order are not present (see Bauch, 2005, House et al., 2009 and Appendix B for work in this direction). By proceeding instead to $n = 4$ moment equations and repeating the same procedure as above, 4-cycle graphs are uncovered contributing yet further anomalies on networks with this type of clustering such as square lattices.

For the individual-based model the leading order anomaly corresponds to essentially ubiquitous processes between two individuals. For the pair-based model the leading order anomalous terms correspond to 3-cycles with further anomalies from higher order n-cycles. However, these higher order errors only become relevant for networks with some type of clustering supporting existing intuition that the pair-level models work best for networks with little clustering (e.g. Keeling and Eames, 2005). The ubiquity of the leading order pair-level

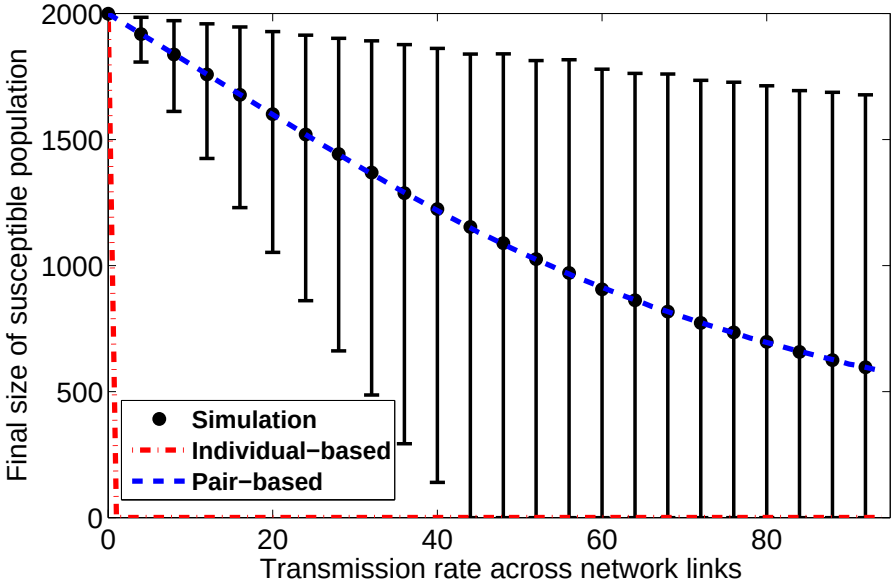

Figure 4: Evaluation of the final size prediction of the individual-based and pair-based models on a ring graph with nearest neighbour connectivity for transmission rates across each link varying from $\tau = 1$ to $\tau = 90$. The removal rate is held fixed at $g_i = 0.1$ and the network size is $N = 2000$. All epidemics are initiated with a single infected individual and comparison is made with the mean of 10,000 stochastic simulations with error bars indicating the 5th and 95th percentiles.

terms explains why pair-based models are a substantial improvement over the individual-based models for all networks with low connectivity.

To support our analysis, figure 4 shows the final size of the susceptible population for a ring graph in which each individual is arranged around a ring and connected to its two nearest neighbours. This network has no cycles except at the order of the network itself. It is clear from this figure that the pair-based model is essentially exact as we would expect in spite of the epidemics being initiated by a single infectious individual.

10 Systematic identification of higher order anomalous terms and the behaviour as n approaches N

We now extend this analysis to the heuristic derivation of moment equations of arbitrary order via the systematic identification of higher order anomalous

terms by repeating the procedure used for $n = 2$ and $n = 3$. Recall that we assume statistical independence at order n and that this incorporates all anomalous terms at order n and below. This assumption of statistical independence allows us to use the method outlined in section 4 for generating the equations. However, this is methodological and our interest is only in the nature of the new anomalous terms at order n which are not incorporated in an $n-1$ moment closure model. This corresponds to the impact of the statistical independence assumption $\{n\text{wise}\backslash(n-1)\text{wise}\}$.

We consider the master equation for the probability of the state of a system Λ composed of $n \leq N$ subsystems of the complete epidemic system Γ. Following the steps that led to equation 10 for the complete system, we have for Λ:

$$\dot{\langle \Lambda^\alpha \rangle} = \sum_{j=1}^{n} \langle \dot{\psi_j^a} \rangle \prod_{i,i\neq j} \langle \psi_i^a \rangle = \sum_{j=1}^{n} \lambda_j^\alpha \tag{46}$$

where the product is over all n subsystems except for j. For the SIR model, ψ_j^a can be susceptible, infectious or removed. Let us consider a term in the summation such that $\psi_j^a = S_j$. Substituting from the subsystem master equations with pairwise statistical independence (equation 30) gives:

$$\lambda_j^\alpha = -\sum_k \eta_{jk}^\alpha \tag{47}$$

where

$$\eta_{jk}^\alpha = T_{jk}\langle S_j \rangle \langle I_k \rangle \prod_{i,i\neq j} \langle \psi_i^a \rangle = T_{jk}\langle I_k \rangle \prod_i \langle \psi_i^a \rangle \tag{48}$$

When $\psi_j^a = I_j$ we obtain:

$$\lambda_j^\alpha = \sum_k \eta_{jk}^\alpha - g_j \langle I_j \rangle \prod_{i,i\neq j} \langle \psi_i^a \rangle = \sum_k \eta_{jk}^\alpha - g_j \langle \Lambda^\alpha \rangle \tag{49}$$

When $\psi_j^a = R_j$ we obtain:

$$\lambda_j = g_j \langle \Lambda^\alpha \rangle \tag{50}$$

There are two possibilities for each term η_{jk}^α:

1. Firstly the site k may be external to the system Λ and is therefore not one of the sites covered by the index i. In this case, we obtain a term describing a system of size $n+1$ and see that the master equation for Λ can depend on terms at an order one higher that the system itself. After applying our heuristic step of replacing products of probabilities with actual probabilities, this term is $T_{jk}\langle I_k \psi_1^a \psi_2^a ... \psi_n^a \rangle$. This is the general case of the observation that the pair-based equations contain terms at order $n = 3$ (the first two terms on the right of equation 39). This is also where the higher $\{n+1\text{wise}\backslash(n)\text{wise}\}$ correlations appear.

2. Alternatively, k may belong to the set of sites 1 to n. Suppose firstly that the system state Λ^α specifies ψ_k^a to be an infectious state; $\psi_k^a = I_k$. In this case:

$$\eta_{jk}^\alpha = T_{jk}\langle I_k\rangle\langle I_k\rangle \prod_{i,i\neq k}\langle\psi_i^a\rangle \longrightarrow T_{jk}\langle\psi_1^a ... \psi_n^a\rangle = T_{jk}\langle\Lambda^\alpha\rangle \tag{51}$$

where the systematic correction of contracting similar states is applied. This is precisely the correction procedure used for the specific cases $n = 2$ and $n = 3$. Now suppose that $\psi_k^a = S_k$:

$$\eta_{jk}^\alpha = T_{jk}\langle I_k\rangle\langle S_k\rangle \prod_{i,i\neq k}\langle\psi_i^a\rangle \longrightarrow 0 \tag{52}$$

where the term is again systematically corrected as in the $n = 2$ and $n = 3$ cases. Similarly, we obtain zero when $\psi_k^a = R_k$.

So in general, the process of correction appears systematic and the same as for $n = 2$ and $n = 3$ at all orders for this SIR model leading to the possibility of using this method to automatically generate systems of moment equations. We should note that this does not constitute a proof that the procedure always generates the correct equations, although it seems reasonably clear that this is the case for the present context of an SIR epidemic model.

This description neatly highlights the transition from the individual-level equations (equation 20) via a sequence of increasingly large moment equations to the complete master equation. Where n is small with respect to the full independent system size N, most of the terms are of type 1 containing terms at order $n + 1$ (e.g. equation 39). As n approaches N, most terms are of type 2 (at order n). Finally, in the case $n = N$, all terms are at order n (there are no external individuals) resulting in the complete self-contained master equation of the full system (or at least of its independent giant components) with no requirement for closure approximations.

For a particular moment equation at order n to be relevant to the dynamics, it must arise from a moment equation at order $n-1$ which in turn arises from a moment equation at order $n-2$, continuing until the individual-level equations (equation 20) are reached. This forms a weakly connected subgraph with n nodes and consequently the dynamics only depend on equations describing the states of these particular subgraphs. It is readily seen that these subgraphs must be acyclic since cycles do not need to be evaluated at a higher order.

At each order n, our procedure will produce anomalous terms from the pair level up to n by systematically joining two individuals within the subgraph subject to the transmission network T. However, provided that a consistent closure approximation is used to absorb the correlations at each order, the terms corresponding to cycles at order $n-1$ and less are already accounted for by lower order moment closure models (as in the $n = 3$ example in section 9). It is only the n-cycles which distinguish the nth order moment closure model from the $(n-1)$th order model. At order n, these terms join together the

end nodes of a simple path of length n between two individuals which we shall identify by $i = 1$ and $i = n$. In the context of equations 51 and 52, these terms correspond to ($k = 1$, $j = n$) or ($k = n$, $j = 1$). Hence at each order n, the equation for each n-tuple state has at most two anomalous terms which are not accounted for in the previous order moment closure model (subject to the use of a consistent closure approximation (Appendix B)). We saw examples of this in the two anomalous terms in equation 39 and in graphs a and e in figure 3.

In section 8 we argued that the order 1 and order 2 moment closure models should converge as the number of neighbours per individual becomes large. Since there are only two new anomalous terms per order (subject to there being cycles at that order), we expect that order n moment closure models converge to order $n - 1$ moment closure models as approximately the reciprocal of the number of terms at each order. At order n we have a summation over j (from equation 46) and a summation over k from equations 47 and 49. Hence, given closure approximations which preserve correlations at each order, the moment closure models converge towards each other with increasing n and with increasing network connectivity.

In spite of this result, we should be aware that high order cycles can sometimes have a noticeable impact in extreme circumstances as we now show. For the simulations on the $N = 2000$ ring graph shown in figure 4, it is clear that the higher order $N = 2000$ cycle did not have an observable effect. Figure 5a shows the final susceptible size predictions on a one-dimensional lattice with nearest neighbour connectivity with size varying from $N = 2$ to $N = 20$. The transmission rate across each link is scaled with network size as $\tau = N/5$ to ensure that on average, most of the population are infected. This network has no higher order cycles and the predictions of the pair-based model are exact even though epidemics are initiated with a single infected individual. A ring graph can be constructed from the one-dimensional lattice by connecting the end nodes together and this introduces a cycle at order N. Figure 5b shows simulations on the equivalent ring graph and illustrates a small discrepancy between the stochastic simulations and the pair-based model which we must attribute to this N-cycle. The "ring" graph of two nodes remains exact as it is unchanged from the one dimensional lattice. At 3 nodes, the network constitutes a single triangle and the small discrepancy is consistent with our conclusions in section 9 that closed triples will cause errors in the pair-based model. We also observe that under some circumstances, even when the first order of clustering is very high (here up to 20), the errors at this order (which are implicit in {3wise\2wise} statistical independence) can cascade down to have an impact on the pair-based model.

11 Discussion and conclusions

A central focus of this work was to develop a better understanding of the connection between stochastic simulation and deterministic models of epidemics propagated on contact networks. In general, the average behaviour of a stochas-

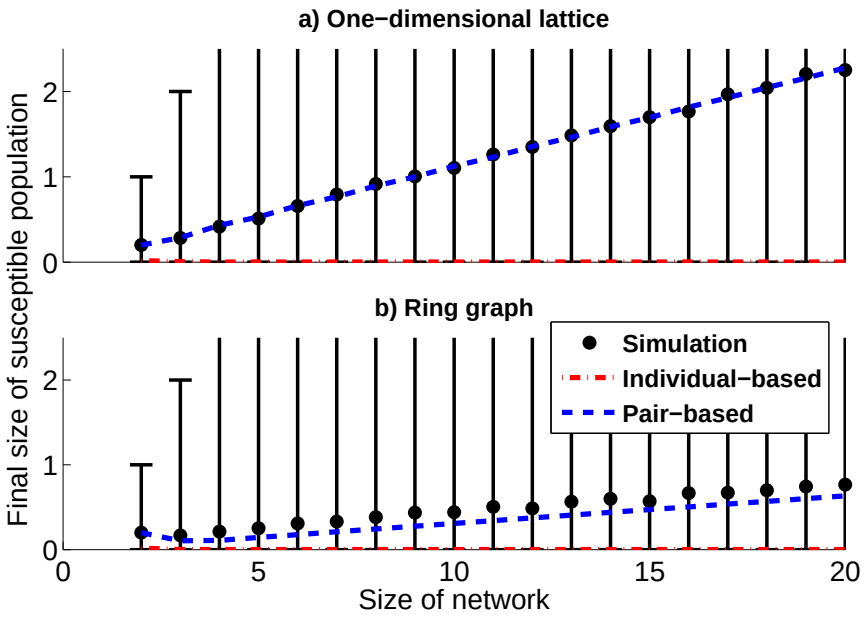

Figure 5: Evaluation of the final size predictions of the individual-based and pair-based models for a one-dimensional lattice with nearest neighbour connectivity and a ring lattice with nearest neighbour connectivity. Networks varying from $N = 2$ to $N = 20$ are considered. The transmission rate across each link is scaled with N as $\tau = N/5$ to ensure that a significant proportion of the network is infected in each case. The removal rate is $g_i = 0.1$. Comparison is made with the mean of 10,000 stochastic simulations with error bars indicating the 5th and 95th percentiles. All epidemics are initiated by a single infected individual.

tic epidemic system is, given initial conditions, a system characteristic with a fully deterministic time course. It is this average behaviour which deterministic models normally attempt to replicate. For certain classes of contact network, deterministic models can be obtained in a reasonably systematic way with well understood assumptions (e.g. Diekmann et al., 1998; Volz, 2008; Simon et al., 2010) but in the general case, the precise deterministic representation of a standard SIR epidemic on an arbitrary contact network of size N requires the solution of $3^N - 1$ equations. This normally necessitates the use of approximations.

Here we investigated a particular class of approximate deterministic individual-level moment closure models of epidemics propagated on arbitrary contact networks (Sharkey, 2008). We showed that a theoretical basis for these models can be provided by relationships which we derived between the master equations for arbitrary systems and those for their subsystems. This leads to sets of moment equations which are fully consistent with the master equations, and hence with stochastic simulation of the master equations.

We used closures of these moment equations which relied on assuming statistical independence at some order. The assumption of statistical independence can lead to a departure from the master equation description and to discrepancies with the expectation values of the stochastic epidemic variables. Our analysis focused on understanding the breakdown of these assumptions and on the resulting inaccuracies in the epidemic models. In particular, we showed that the failure of these individual-level moment closure models can be directly attributed to anomalous terms describing unbiological processes acting around closed loops in the contact network. This directly relates the errors in the models to networks with some degree of clustering (defined as low order cycles). This observation is well known in the population-level context of pair-approximation models (e.g. Keeling and Eames, 2005), although this gives a more systematic procedure for identifying, quantifying and correcting these anomalies.

These individual-level models exist in the domain between stochastic simulation (agent-based models) and population-based deterministic epidemic models; indeed, they can sometimes be viewed as an intermediate step between the two with the population-based models obtained from additional assumptions of homogeneity (Sharkey, 2008). Consequently, our conclusions at the individual level provide insights into the network-based population-level models. In this sense, the analysis is relevant to a range of commonly used epidemic models including network-based mean-field and pair-approximation models.

We considered a hierarchy of individual-level moment closure models which were closed at each order n by a Kirkwood-type closure (Kirkwood, 1935; Rand, 1999; Singer, 2004). We assumed that the Kirkwood-type closure preserves all correlations at order n and below but ignores correlations associated with $\{(n+1)\text{wise}\backslash n\text{wise}\}$ statistical independence, by which we mean the conditions implicit in $(n+1)$wise independence which are not also implicit in $n$wise independence. We showed that this type of statistical independence implies anomalous terms describing unbiological processes acting around loops or cycles in the network with path-length $n+1$ as well as implicit higher order errors. In this sense,

the anomalous terms quantify the errors associated with the closure approximation at each order and their analysis provides insights into the departure of these models from the underlying master equation. Correction of the anomalous terms at order n breaks the assumed independence which induces correlations and leads to more accurate models at order $n+1$.

Our choice of moment closure had the objective of preserving the correlations that are induced at each order by the correction of anomalous terms while ignoring higher order correlations. In this sense, closures based on statistical independence emerge as the obvious ones to use in this context. Many alternative closure approximations have been proposed to give improved results on specific classes of network by incorporating higher order effects (e.g. van Baalen, 2000). However, these do not fall naturally into the current context where we attempt to identify and isolate the cause of the errors at each order.

The lowest order moment closure models at orders $n = 1$ and $n = 2$ leads to the generation of what we termed the individual-based and pair-based deterministic epidemic models respectively. These models allow deterministic descriptions of epidemics at an individual level on transmission networks of arbitrary complexity. Code for solving these models is provided as supplementary material.

We showed that the individual-based model is fully consistent with the master equation of the stochastic system except for the single assumption of pairwise statistical independence. We traced the leading order errors in this model to the anomalous terms $T_{ji}\langle S_i\rangle\langle S_j\rangle\langle I_i\rangle$ and $-T_{ij}\langle S_i\rangle\langle I_j\rangle\langle I_j\rangle$ appearing in the equation for the doublet state probability $\langle S_iI_j\rangle$. The first of these describes a process in which individuals are in two states at the same time, directly contradicting the premise of a compartmental model. The second term is also inconsistent with the master equation description. These terms can be systematically corrected which induces correlations and leads to the pair-based models. As such, they describe the difference in behaviour between the individual-based and pair-based models as well as providing the leading order departure of the individual-based model from the average stochastic realisation of the master equation.

Analysis of these two terms therefore provides insights into the circumstances for which the individual-based model fails to produce good results. We showed that they qualitatively account for the observed behaviour of these models on networks for which higher order effects are relatively small (where pair-based models are accurate). Many of these behaviours are well known in the context of population-level models, although complicated by the additional assumption of homogeneity. However, the current approach permits an arguably more analytic understanding of the population-level correlations. Indeed, the error associated with the term $-T_{ij}\langle S_i\rangle\langle I_j\rangle\langle I_j\rangle$ has not been explicitly identified previously. We also identified terms at higher order giving further understanding of the types of network for which the deterministic population-level models fail.

While the error associated with the term $-T_{ij}\langle S_i\rangle\langle I_j\rangle\langle I_j\rangle$ is present for all networks, the term $T_{ji}\langle S_i\rangle\langle S_j\rangle\langle I_i\rangle$ only creates errors on symmetric (undirected) links. Consequently, with all other aspects the same, the individual-based model should perform better on directed networks. We confirmed this by simulation

(figure 1). This result is directly related to the frequently made point that during the course of an epidemic, any infectious individual must have at least one non-susceptible neighbour to have infected it. On a symmetric network, this reduces the number of potential susceptible contacts, slowing the progress of the epidemic and creating correlations between individuals (e.g. Diekmann et al., 1998; Eames and Keeling, 2002; May, 2006). We can see this connection by noting that the first term in equation 39 represents the creation of S_iI_j by a prior contact with another infectious individual k within the triple structure $S_iS_jI_k$. Since k cannot be the same site as i, this reduces the number of possible susceptible neighbours available to j. Part of the failure of the individual-based model on symmetric networks is then precisely because it allows site k and site i to overlap in the anomalous term $T_{ji}\langle S_i\rangle\langle S_j\rangle\langle I_i\rangle$. This insight is therefore not new, but it does enable the error to be uncovered and represented in a systematic way.

We found that the errors associated with these two anomalous terms do not cancel, but rather act in the same direction to over-estimate the severity of an epidemic. We therefore expect that the individual-based model will always predict a bigger epidemic than the pair-based model and provided that these leading order errors dominate, the individual-based model will exaggerate the severity of an epidemic. Additionally, the relative significance of these two anomalous terms increases when the number of neighbours per individual is small. Consequently the individual-based model is more accurate for networks with many connections per individual. Again, this replicates understanding at the population level (e.g. Keeling, 1999), but arguably makes this result more general and provides a different perspective. We also argued that the impact of these two terms is greatest for epidemics which propagate slowly with a basic reproductive ratio greater than but close to 1 implying that the individual-based and pair-based models should be in better agreement with increasing basic reproductive ratio.

The leading order anomalous terms can be corrected which, together with a suitable closure approximation based on $\{3\text{wise}\backslash 2\text{wise}\}$ statistical independence leads to the pair-based model. We demonstrated in section 10 that this process of identifying and correcting the anomalous terms is systematic at all orders; anomalous terms are first identified as those having repeated sites and are then contracted if the site states are the same or removed otherwise. The construction of higher order models such as the pair-approximation or pair-based models has traditionally involved the identification of all possible processes contributing to the rate of change of a quantity such as the S_iI_j pair. Ensuring that all of the possible processes are accounted for can be difficult and requires careful bookkeeping (van Baalen, 2000). Here, in the context of the SIR epidemic model, we have a systematic method for obtaining these equations. This may therefore enable the automated generation of systems of moment equations at any order. It seems reasonable, at least in the context of the SIR epidemic model that this procedure always leads to the correct moment equations although we did not prove this.

The pair-based model contains further anomalous terms describing unbio-

logical processes acting around triangles (or 3-cycles). Although we found that the $n = 2$ order terms always act in the same direction to exaggerate the rate of spread of an epidemic, at order $n = 3$, the terms can act in opposing directions which can reduce their net effects. Nevertheless, experience suggests that the pair-based model still exaggerates the spread of epidemics suggesting the dominance of the terms increasing the rate of spread.

The process of generating the higher order models is systematic. In principle, a model can be constructed without the $n = 3$ order anomalous terms, but then assumes $\{4\text{wise}\backslash 3\text{wise}\}$ statistical independence which corresponds to anomalous terms describing unbiological processes acting around 4-cycles and higher. In general we create an nth order moment equation model which depends on $\{(n+1)\text{wise}\backslash n\text{wise}\}$ statistical independence with leading order errors corresponding to unbiological processes acting around $(n+1)$-cycles. Moment closure models at order n can remove anomalous terms at order n and below.

In practice we note that individual-based moment closure models beyond the pair level are likely to be prohibitively complex and computationally expensive. However, consideration of the general case is interesting from a theoretical perspective. A novel point of this construction is that it allows us to understand the convergence of the moment closure models to the complete master equation description in the limit as n approaches the system size N. Additionally, we can see that each moment equation contains at most two anomalous terms corresponding to the two ways in which the end sites of a simple network path of length n can be closed to form an n-cycle (section 10). As a result, we found that the relative number of anomalous terms decreases at each order making the $n-1$th order moment closure model converge to the nth order moment closure model with increasing n. Nevertheless, we observed for the example of a ring graph that even very high order cycles can sometimes result in noticeable errors (figure 5b).

Our analysis of correlations is limited to contact networks which do not change rapidly in time with respect to the contact frequencies across network links. For networks whose links are very dynamic, the notion of a contact network becomes lost as infection is effectively permitted to travel directly between any two individuals. It is clear that the impact of the anomalous terms is reduced because the number of network contacts per individual is essentially increased (while their strength is decreased). While detailed analysis of this situation is outside the scope of the present work, for random networks, convergence to the mean-field model with increasingly rapid network dynamics is discussed in detail by Volz and Meyers (2007; 2009). This issue is also addressed in the context of a triple-approximation model by House et al. (2009). Additionally, metapopulations models frequently consider the movement of individuals from one population to another as the cause of transmission. At the individual level this corresponds to an extreme case of a dynamic network which is coupled to the infection process.

Finally, we found that the best agreement between the individual-level models and stochastic simulation was usually when epidemics are initiated with a reasonably large number of infected sites (figure 2). We argued that this is

because the chance of stochastic fade out of the epidemic is reduced. However, the probability of stochastic fadeout is intrinsically incorporated into the master equation description and therefore the moment closure models incorporate this as well, although in an approximate way. The significant error that we observed when the probability of stochastic fade out is not negligible is attributable to errors in predicting this probability. Even small errors can translate into a significant distortion of the average epidemic due to the bimodal response of either very small or very large outbreaks. For the example of a one-dimensional lattice for which the pair-approximation model is exact, no errors were observed, irrespective of the number of initial cases or the rate of transmission (figure 5a).

In conclusion, the analysis presented here enables a better understanding of deterministic models of epidemics on contact networks and their relationship to the master equation. It shows that the correlations at each order can be associated with cycles in the network of the same order and that this quantifies the departure from the master equation. This provides a more analytic perspective on the role of clustering in the behaviour of epidemic dynamics.

12 Acknowledgements

I thank Roger Bowers for helpful comments and insights and two anonymous reviewers for helpful suggestions.

13 Appendix A. Derivation of the S_iI_j moment equations from the system master equation

We can determine the equation for the probability of the pair state S_iI_j by a straightforward extension of the argument for singlet states in section 5. We first write down the probability for the doublet state:

$$\langle S_iI_j\rangle = \sum_\alpha \langle\Gamma^\alpha\rangle D_i^{\alpha S} D_j^{\alpha I} \tag{53}$$

Differentiating with respect to time and substituting from equation 1 gives:

$$\begin{aligned}
\langle \dot{S_iI_j}\rangle &= \sum_{\alpha\beta} \sigma^{\alpha\beta}\langle\Gamma^\beta\rangle D_i^{\alpha S} D_j^{\alpha I} - \sum_{\alpha\beta} \sigma^{\beta\alpha}\langle\Gamma^\alpha\rangle D_i^{\alpha S} D_j^{\alpha I} \\
&= \sum_{\alpha\beta} \sigma^{\alpha\beta}\langle\Gamma^\beta\rangle D_i^{\alpha S} D_j^{\alpha I} \left[\sum_a D_i^{\beta a}\right]\left[\sum_b D_j^{\beta b}\right] \\
&\quad - \sum_{\alpha\beta} \sigma^{\beta\alpha}\langle\Gamma^\alpha\rangle D_i^{\alpha S} D_j^{\alpha I} \left[\sum_a D_i^{\beta a}\right]\left[\sum_b D_j^{\beta b}\right] \\
&= \sum_{\alpha\beta ab} \sigma^{\alpha\beta}\langle\Gamma^\beta\rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta a} D_j^{\beta b}
\end{aligned}$$

$$- \sum_{\alpha\beta ab} \sigma^{\beta\alpha} \langle \Gamma^{\alpha} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta a} D_j^{\beta b} \tag{54}$$

Doing the sum on a and b (accounting for the fact that only one of them can change at a time with non-zero transition rate) gives:

$$\begin{aligned} \langle \dot{S_i} I_j \rangle &= \sum_{\alpha\beta} \sigma^{\alpha\beta} \langle \Gamma^{\beta} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta S} \\ &\quad - \sum_{\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^{\alpha} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta I} D_j^{\beta I} \\ &\quad - \sum_{\alpha\beta} \sigma^{\beta\alpha} \langle \Gamma^{\alpha} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta R} \end{aligned} \tag{55}$$

By implementing the transition processes at the system level we get:

$$\begin{aligned} \sigma^{\alpha\beta} D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta S} &= \sum_k T_{jk} D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta S} D_k^{\beta I} \zeta_j^{\alpha\beta} \\ \sigma^{\beta\alpha} D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta I} D_j^{\beta I} &= \sum_k T_{ik} D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta I} D_j^{\beta I} D_k^{\alpha I} \zeta_i^{\beta\alpha} \\ \sigma^{\beta\alpha} D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta R} &= g_j D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta R} \zeta_j^{\beta\alpha} \end{aligned} \tag{56}$$

Substituting into equation 55 gives:

$$\begin{aligned} \langle \dot{S_i} I_j \rangle &= \sum_{\alpha\beta} \sum_k T_{jk} \langle \Gamma^{\beta} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta S} D_k^{\beta I} \zeta_j^{\alpha\beta} \\ &\quad - \sum_{\alpha\beta} \sum_k T_{ik} \langle \Gamma^{\alpha} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta I} D_j^{\beta I} D_k^{\alpha I} \zeta_i^{\beta\alpha} \\ &\quad - \sum_{\alpha\beta} g_j \langle \Gamma^{\alpha} \rangle D_i^{\alpha S} D_j^{\alpha I} D_i^{\beta S} D_j^{\beta R} \zeta_j^{\beta\alpha} \\ &= \sum_k T_{jk} \sum_{\beta} \langle \Gamma^{\beta} \rangle D_j^{\beta S} D_k^{\beta I} \sum_{\alpha} D_i^{\beta S} D_i^{\alpha S} D_j^{\alpha I} \zeta_j^{\alpha\beta} \\ &\quad - \sum_k T_{ik} \sum_{\alpha} \langle \Gamma^{\alpha} \rangle D_i^{\alpha S} D_k^{\alpha I} \sum_{\beta} D_j^{\alpha I} D_j^{\beta I} D_i^{\beta I} \zeta_i^{\beta\alpha} \\ &\quad - g_j \sum_{\alpha} \langle \Gamma^{\alpha} \rangle D_j^{\alpha I} \sum_{\beta} D_i^{\alpha S} D_i^{\beta S} D_j^{\beta R} \zeta_j^{\beta\alpha} \end{aligned} \tag{57}$$

We now use a similar identity to equation 17:

$$\sum_{\alpha} D_i^{\beta b} D_i^{\alpha b} D_j^{\alpha a} \zeta_j^{\alpha\beta} = D_i^{\beta b} \tag{58}$$

This is clearly true when $D_i^{\beta b} = 0$. From equation 17, there must be a single state Γ^{α} for which $D_j^{\alpha a} \zeta_j^{\alpha\beta} = 1$. For this state, when $D_i^{\beta b} = 1$, this also

implies that $D_i^{\alpha b} = 1$ because ψ_i does not change state during this transition, establishing the identity. Applying this to equation 57 gives:

$$\begin{aligned}
\langle \dot{S_i} I_j \rangle &= \sum_k T_{jk} \sum_\beta \langle \Gamma^\beta \rangle D_i^{\beta S} D_j^{\beta S} D_k^{\beta I} - \sum_k T_{ik} \sum_\alpha \langle \Gamma^\alpha \rangle D_i^{\alpha S} D_j^{\alpha I} D_k^{\alpha I} \\
&\quad - g_j \sum_\alpha \langle \Gamma^\alpha \rangle D_i^{\alpha S} D_j^{\alpha I} \\
&= \sum_k T_{jk} \langle S_i S_j I_k \rangle - \sum_k T_{ik} \langle S_i I_j I_k \rangle - g_j \langle S_i I_j \rangle \\
&= \sum_{k,k \neq i} T_{jk} \langle S_i S_j I_k \rangle - \sum_{k,k \neq j} T_{ik} \langle S_i I_j I_k \rangle - T_{ij} \langle S_i I_j \rangle - g_j \langle S_i I_j \rangle \quad (59)
\end{aligned}$$

where the specification that $k \neq i$ in the first term of the last line is superfluous since this probability is zero but is required when we consider approximations of this probability. This completes the derivation of equation 35 from the system master equation.

14 Appendix B. General Kirkwood-type closure approximation

The Kirkwood-type closure approximation for the probability $\langle A_i B_j C_k \rangle$ is given by (Kirkwood, 1935; Rand, 1999; Singer, 2004; Sharkey, 2008):

$$\langle A_i B_j C_k \rangle = \frac{\langle A_i B_j \rangle \langle B_j C_k \rangle \langle C_k A_i \rangle}{\langle A_i \rangle \langle B_j \rangle \langle C_k \rangle} \quad (60)$$

When there is no link (in either direction) between two of the subsystems such as between C_k and A_i, we call this an open triple. To avoid evaluating a separate differential equation for this open pair, we use statistical independence to reduce the above approximation to:

$$\langle A_i B_j C_k \rangle = \frac{\langle A_i B_j \rangle \langle B_j C_k \rangle}{\langle B_j \rangle} \quad (61)$$

These two forms are the closure approximations used in the pair-based model (see supplementary material).

To close triple-level equations, we require an approximation of the generic 4-tuple probability $\langle A_i B_j C_k D_l \rangle$. A natural extension of the Kirkwood closure is a Fisher-Kopeliovich-type closure (Fisher and Kopeliovich, 1960):

$$\langle A_i B_j C_k D_l \rangle = \frac{\langle A_i B_j C_k \rangle \langle B_j C_k D_l \rangle \langle C_k D_l A_i \rangle \langle D_l A_i B_j \rangle \langle A_i \rangle \langle B_j \rangle \langle C_k \rangle \langle D_l \rangle}{\langle A_i B_j \rangle \langle A_i C_k \rangle \langle A_i D_l \rangle \langle B_j C_k \rangle \langle B_j D_l \rangle \langle C_k D_l \rangle} \quad (62)$$

Simplifications can also be made here by assuming statistical independence (analogous to equation 61) resulting in a "motif-based" approach (House et al., (2009)).

We can easily generalise this (Singer, 2004) and approximate the probability of the states of the n subsystems $\Lambda = \{\psi_1, \psi_2, ...\psi_n\}$ by:

$$\langle \Lambda^{\alpha} \rangle = \prod_{d=1}^{n-1} \prod_{1 \leq i_1 < i_2 < \cdots < i_d \leq n} \langle \psi_{i_1}^a \psi_{i_2}^a ... \psi_{i_d}^a \rangle^{(-1)^{n-d-1}} \tag{63}$$

Again, simplifications could be made in principle by applying independence assumptions across unconnected sites, but the resulting number of motifs would probably make this impractical.

We conjecture (although have not attempted to prove) that the correlations induced by the removal of anomalous terms at each order are fully preserved by this form of closure.

15 Appendix C. Analysis of the $n = 2$ anomalous terms with respect to rates of transmission and removal

The equations of the individual-based model cannot be solved analytically. However, if we assume that the infectious pressure acting on an individual during its infection and removal is approximately constant, we can solve the equations. We suppose that the infectious pressure acting on an initially susceptible individual behaves as the Heaviside function $\kappa H(t - t_0)$, switching on at some time $t = t_0$ with a constant "averaged" force of infection κ. This effectively treats the individual as existing in a constant sea of infectivity for the duration of its dynamics and decouples the behaviour of the individual from the rest of the network. The individual-based model then becomes:

$$\begin{aligned} \langle \dot{S}_i \rangle &= -\kappa H(t - t_0) \langle S_i \rangle \\ \langle \dot{I}_i \rangle &= \kappa H(t - t_0) \langle S_i \rangle - g \langle I_i \rangle \end{aligned} \tag{64}$$

Rewriting this in terms of the dimensionless parameters $\omega = \kappa/g$, $\tau = gt$ gives:

$$\begin{aligned} \frac{d\langle S_i \rangle}{d\tau} &= -\omega H(\tau - \tau_0) \langle S_i \rangle \\ \frac{d\langle I_i \rangle}{d\tau} &= \omega H(\tau - \tau_0) \langle S_i \rangle - \langle I_i \rangle \end{aligned} \tag{65}$$

For simplicity we let $\tau_0 = 0$ (not to be confused with the start of the epidemic) and solve these equations to give:

$$\begin{aligned} \langle S_i \rangle &= e^{-\omega\tau} \\ \langle I_i \rangle &= \frac{\omega}{1 - \omega} \left(e^{-\omega\tau} - e^{-\tau} \right) \end{aligned} \tag{66}$$

We can use this to investigate the impact of the anomalous terms $T_{ji} \langle S_i \rangle \langle S_j \rangle \langle I_i \rangle$ and $-T_{ij} \langle S_i \rangle \langle I_j \rangle \langle I_j \rangle$ in equation 38. It is clear that $\langle I_i \rangle$ becomes zero when

$\omega \to 0$. This rather uninteresting scenario corresponds to no infection and in this case both anomalous terms are clearly zero at all times. Note that ω is the ratio between infection rate and removal rate and is therefore effectively the basic reproductive ratio R_0. On a network-wide scale, $R_0 < 1$ will also imply $\langle I_i \rangle = 0$ for most individuals in the system.

More generally, for the term $T_{ji}\langle S_i \rangle \langle S_j \rangle \langle I_i \rangle$, while we cannot comment on the state of the node j, the parts of the expression involving the node i provide a weighting factor for the errors contributed by this term. The overall contribution from this factor is then:

$$\int_0^\infty \langle I_i \rangle \langle S_i \rangle d\tau = \frac{1}{2(1+\omega)} \tag{67}$$

provided that $\omega \neq 0$ and $\omega \neq 1$. This term is then clearly minimised by large ω.

The error from the term $-T_{ij}\langle S_i \rangle \langle I_j \rangle \langle I_j \rangle$ relates to the accuracy of the approximation $\langle I_j \rangle \approx \langle I_j \rangle^2$. The error contributed from this term is then related to:

$$\int_0^\infty \langle I_j \rangle - \langle I_j \rangle^2 d\tau = 1 - \frac{\omega}{2(1+\omega)} \tag{68}$$

provided that $\omega \neq 0$ and $\omega \neq 1$. This has a minimum value of $1/2$ as $\omega \to \infty$. Hence this term is minimised by $\kappa >> g$ but not removed altogether. We note that this is implicitly tied to the use of an exponentially distributed removal process and that a more realistic removal process would not necessarily cause this residual. Additionally, very rapid infection and a low removal rate will likely lead to the situation in which the value of $\langle S_i \rangle$ becomes zero before the approximation breaks down in the removal phase thus minimising the contribution from the anomalous term.

Although the assumption of a constant sea of infection is somewhat contrived and certainly inaccurate, it enabled us to make some analytic progress. There is no immediately obvious reason why the qualitative conclusions we obtained should change for a more dynamic infectious pressure.

References

[1] Anderson RM, May RM (1991) *Infectious diseases of humans.* Oxford University Press.

[2] Bailey NTJ (1975) *The mathematical theory of infectious diseases.* Griffin, London.

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

[4] Ball F, Sirl D, Trapman P (2009) Threshold behaviour and final outcome of an epidemic on a random network with household structure. *Adv. Appl. Prob.* **41**, 765-796.

[5] Bansal S, Grenfell BT, Meyers LA (2007) When Individual behaviour matters: homogeneous and network models in epidemiology. *J. Roy. Soc. Interface* **4**, 879-891.

[6] Bartlett MS (1956) Deterministic and stochastic models for recurrent epidemics. *Proc. Third Berkley Symp. Math. Statist. Prob.* **4**, 81-108.

[7] Bauch C (2005) The spread of infectious diseases in spatially structured populations: An invasory pair approximation. *Math. Biosci.* **198**, 217-237.

[8] Diekmann O, De Jong MCM, Metz JAJ (1998) A deterministic epidemic model taking account of repeated contacts between the same individuals. *J. Appl. Prob.* **35**, 448-462.

[9] Eames KD, Keeling MJ (2002) Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. *Proc. Natl. Acad. Sci. USA* **99**, 13330-13335.

[10] Fisher IZ, Kopeliovich BI (1960) On a refinement of the superposition approximation in the theory of fluids *Dokl. Akad. Nauk. SSSR* **133**, 81-83.

[11] Gillespie DT (1976) A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *J. Comp. Phys.* **22**, 403-434.

[12] House T, Davies G, Danon L, Keeling MJ (2009) A motif-based approach to network epidemics. *B. Math. Biol.* **71**, 1693-1706.

[13] Keeling MJ (1999) The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. B* **266**, 859-867.

[14] Keeling MJ (2005) The implications of network structure for epidemic dynamics. *Theor. Popul. Biol.* **67**, 1-8.

[15] Keeling MJ, Eames, KTD (2005) Networks and epidemic models. *J. Roy. Soc. Interface* **2**, 295-307.

[16] Keeling MJ, Ross J (2008) On methods for studying stochastic disease dynamics. *J. Roy. Soc. Interface* **5**, 171-181.

[17] Keeling MJ, Rohani, P (2008) *Modeling infectious diseases in humans and animals.* Princeton University Press.

[18] Keeling MJ, Ross JV (2009) Efficient methods for studying stochastic disease and population dynamics. *Theor. Popul. Biol.* **75**, 133-141.

[19] Kermack WO, McKendrick AG (1927) Contributions to the mathematical theory of epidemics. *Proc. R. Soc. Edinb. A* **266**, 859-867.

[20] Kirkwood JG (1935) Statistical mechanics of fluid mixtures. *J. Chem. Phys.* **3**, 300-313.

[21] Levin SA, Durrett R (1996) From Individuals to Epidemics. *Phil. Trans. R. Soc. Lond. B* **351**, 1615-1621.

[22] Levins R (1969) Some demographic and genetic consequences of environmental heterogeneity for biological control. *Bull. Entomol. Soc. Am.* **15**, 237-240.

[23] Matsuda HN, Ogita A, Sasaki A, Sato K (1992) Statistical mechanics of populations: The lattice Lotka Volterra model. *Prog. Theor. Phys.* **88**, 1035-1049.

[24] May RM (2006) Network structure and the biology of populations. *Trends in Ecology and Evolution* **21**, 394-399.

[25] McKane AJ, Newman TJ (2004) Stochastic models in population biology and their deterministic analogs. *Phys. Rev. E* **70**, 041902.

[26] Meyers LA, Newman MEJ, Martin M, Schrag S (2003) Applying network theory to epidemics: control measures for Mycoplasma pneumoniae outbreaks. *Emerg. Infect. Dis.* **9**, 204-210.

[27] Murray JD (2003) *Mathematical Biology.* Springer.

[28] Murrell DJ, Law R, Dieckmann U (2004) On moment closures for population dynamics in continuous space. *J. Theor. Biol.* **229**, 421-432.

[29] Mollison, D. (1991) Dependence of epidemic and population velocities on basic parameters. *Math. Biosci.* **107**, 255-287.

[30] Newman MEJ (2002) Spread of epidemic disease on networks *Phys. Rev. E* **66**, 016128.

[31] Ovaskainen O, Cornell SJ (2006) Space and stochasticity in population dynamics. *Proc. Natl. Acad. Sci. USA* **103**, 12781-12786.

[32] Rand DA (1999) Correlation equations and pair approximations for spatial ecologies. In *Advanced ecological theory: principles and applications* eds McGlade J (Blackwell, Oxford), pp 100-142.

[33] Renshaw, E (1991) *Modelling biological populations in space and time.* Cambridge University Press.

[34] Risken, H (1989) *The Fokker-Planck Equation Methods of Solutions and Applications.* Springer Series in Synergetics Springer-Verlag, Berlin.

[35] Ross JV, House T, Keeling MJ (2010) Calculation of disease dynamics in a population of households. *PLoS ONE* **5**, e9666.

[36] Sattenspiel L, Dietz K (1994) A structured epidemic model incorporating geographic mobility among regions. *Math. Biosci.* **128** 71-91.

[37] Sharkey KJ, Fernandez C, Morgan KL, Peeler E, Thrush M, Turnbull JF, Bowers RG (2006) Pair-level approximations to the spatio-temporal dynamics of epidemics on asymmetric contact networks. *J. Math. Biol.* **53**, 61-85.

[38] Sharkey KJ (2008) Deterministic epidemiological models at the individual level. *J. Math. Biol.* **57**, 311-331.

[39] Simon PL, Taylor M, Kiss IZ (2010) Exact epidemic models on graphs using graph-automorphism driven lumping. *J. Math. Biol.* doi: 10.1007/s00285-010-0344-x

[40] Singer A (2004) Maximum entropy formulation of the Kirkwood superposition approximation. *J. Chem. Phys.* **121**, 3657-3666.

[41] van Baalen M (2000) Pair Approximations for different spatial geometries. In *The geometry of ecological interactions: Simplifying complexity*, pp 359-387.

[42] van Kampen (2007)*Stochastic Processes in Physics and Chemistry.* Elsevier.

[43] Volz E, Meyers LA (2007) Susceptible-infected-recovered epidemics in dynamic contact networks. *Proc. R. Soc. B* **274**, 2925-2933.

[44] Volz E (2008) SIR dynamics in random networks with heterogeneous connectivity. *J. Math. Biol.* **56**, 293-310.

[45] Volz E, Meyers LA (2009) Epidemic thresholds in dynamic contact networks. *J. Roy. Soc. Interface* **6**, 233-241.