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

Approximating quasi-stationary behaviour in network-based SIS dynamics

Christopher E. Overton · Robert R. Wilkinson · Adedapo Loyinmi · Joel C. Miller · Kieran J. Sharkey

Received: date / Accepted: date

Abstract Approximations to Susceptible-Infectious-Susceptible stochastic dynamics typically predict a stable endemic steady-state when above threshold. This can be hard to relate to the underlying stochastic dynamics, which has no endemic steady-state but can exhibit approximately stable behaviour. Here we relate the approximate models to the stochastic dynamics via the definition of the quasi-stationary distribution (QSD), which captures this approximately stable behaviour. We develop a system of ordinary differential equations that approximate the number of infected individuals in the QSD for arbitrary contact networks and parameter values. When the epidemic level is high, these QSD approximations coincide with the existing approximation methods. However, as we approach the epidemic threshold, the models deviate, with these models following the QSD and the existing methods approaching the all susceptible state. Through consistently approximating the QSD, the proposed methods provide a more robust link to the stochastic models.

Keywords moment-closure · graph · epidemic model · stochastic · pair approximation

1 Introduction

The Markovian network-based Susceptible-Infectious-Susceptible (SIS) model (also referred to as the contact process [15,24]) is a stochastic model describing how pathogens spread on a host contact network [5,13,20,34,36,37,41]. In these dynamics, individuals can flip back and forth between two states: susceptible and infected. When an individual is infected, its neighbours in the network (or graph) that are susceptible are directly at risk of becoming infected. Infected individuals eventually return to the susceptible state and are again at risk. When all individuals are susceptible, they remain so for all future time and the pathogen is said to have died out. The all-susceptible state is thus an absorbing state. The model is sometimes considered to be a reasonable mathematical representation for the propagation of sexually transmitted diseases and computer viruses [10].

Approximations to stochastic SIS models, such as mean-field models [23,37,46–48] and pair-approximation models [12,13,17,18,20,25,44], can characterise important features of the stochastic dynamics. One example is the epidemic threshold, below which the pathogen quickly goes extinct, and above which large

Christopher E. Overton
Department of Mathematics, University of Liverpool, UK
Present address: Department of Mathematics, University of Manchester, UK
Clinical Data Science Unit, Manchester University NHS Foundation Trust
E-mail: c.overton@liverpool.ac.uk
christopher.overton@manchester.ac.uk

Robert R. Wilinson
Department of Applied Mathematics, Liverpool John Moores University, UK

Adedapo Loyinmi
Tai Solarin University of Education, Nigeria

Joel C. Miller
Department of Mathematics and Statistics, La Trobe University, Australia

Kieran J. Sharkey
Department of Mathematics, University of Liverpool, UK

outbreaks can occur. However, above threshold, these approximate models reach a stable endemic steady-state solution [34] which is not observed in the stochastic dynamics. The stable endemic steady-state which emerges means it is not always clear how to relate these results back to the underlying stochastic process, since the only stable solution to the stochastic model is the disease-free state.

Sufficiently above threshold, the stochastic system may exhibit apparently stationary behaviour, since the probability of extinction over any finite time period can be made very small. Indeed, this apparently stationary behaviour is often observed, with extinction just a theoretical certainty which almost never occurs over reasonable timescales. The quasi-stationary distribution (QSD) is commonly used to define, quantify and understand the long-term behaviour of finite Markov chains with absorbing states. Examples include: modelling the spread of a computer virus across a network with cure and reinfection [19,27,37, 49], chemical reactions in which materials or catalysts can be exhausted [6,7,31,35,38], and wildlife management models [16,21,26,33,40,39,43]. Within Markovian SIS dynamics, various statistics have been derived using the concept of the QSD [2–4,14]. This includes use by Wilkinson and Sharkey [50] to derive a measure of the invasion probability, by Ferreira and colleagues [11] to approximate the epidemic threshold, and by Nåsell [30] to account for the influence of epidemic and demographic forces on the time to extinction.

The calculation of the QSD can require a large number of stochastic simulations, and therefore it is necessary to derive approximation methods. Thus far, approximations have mainly focused on well-mixed populations. Kriscio and Lefevre [22] used a conditional birth-and-death process to approximate the QSD of Markovian SIS epidemic dynamics, which has since been extended by Nåsell [28,29]. Allen and Burgin [1] used a system of ordinary differential equations that approximate the expected prevalence in the QSD when the epidemic severity is high. Dickman and Vidigal [9] developed a pair approximation for the QSD on circles, which the model derived in this paper yields as a special case.

In arbitrary network-structured populations, van Mieghem [46] has shown that the endemic steady-state of the "N-intertwined" individual-based SIS model, which is akin to the network-based mean-field approximation, leads to a "meta-stable state", which is consistent with the quasi-stationary distribution, when sufficiently above the epidemic threshold. This behaviour has also been observed for pair-based SIS approximations [13,20]. These approximations to the stochastic dynamics are typically obtained by making statistical independence assumptions.

Here our objective is to clarify the link between stochastic SIS dynamics and the approximate models by relating them via the QSD. Well above threshold, the stochastic model exhibits stationary-like behaviour and the conditioning of the QSD has minimal impact over short timescales, yielding a meaningful connection between the stochastic model and its approximations. Closer to the threshold and below threshold, the mapping according to the QSD becomes more important because the unconditioned approximate model with its steady state no longer approximates the stochastic process and its absorbing state. This leads to greater numerical correspondence between the models in this regime.

The paper is structured as follows. In Section 2, we define the Markovian network-based SIS modelling framework and the master equation that describes the expected behaviour, followed by defining the QSD in Section 3. Section 4 develops approximation methods that capture aspects of the QSD in a numerically feasible way and we prove the existence of endemic equilibrium solutions for the node-level pair-based SIS approximation (often referred to as the pair-quenched-mean-field approximation). We then analyse the performance of the proposed methods on different contact networks in Section 5.

2 Markovian SIS dynamics on a contact network

We consider a finite set $\mathcal{V}$ of individuals, who are labelled via an arbitrary bijection to $\{1, 2, \ldots, |\mathcal{V}|\}$. Let $N = |\mathcal{V}| < \infty$. Individuals can be in one of two states: susceptible, denoted by S, or infected, denoted by I. An individual $j \in \mathcal{V}$, while infected, makes infectious contacts to an individual $i \in \mathcal{V} \setminus \{j\}$ according to a Poisson process with rate $T_{ij} \geq 0$. If a susceptible individual $k \in \mathcal{V}$ receives an infectious contact, it immediately becomes infected for an exponentially distributed time period with mean $1/\gamma_k$, after which it immediately becomes susceptible again. We define the neighbourhood of an individual j, denoted $\mathcal{N}_j$, as the set of individuals that can either infect or be infected by j; i.e. $i \in \mathcal{N}_j$ if $T_{ij} > 0$ or $T_{ji} > 0$. We assume that the transmission matrix T is strongly connected; i.e. every individual is at risk of future infection if at least one individual is infected. The matrix T can either represent a directed or undirected contact network.

This model is described by a continuous-time Markov chain $\{\Sigma(t) : t \geq 0\}$ with finite state space $\{S, I\}^N$, parametrised by an irreducible square matrix T with non-negative entries and a vector γ with

positive entries. Let $\sigma_\alpha \in \{S, I\}^N$ denote a state of the population. We assume throughout that state σ_1 corresponds to the all susceptible state. Let $\Sigma_i(t)$ denote the status of individual i at time t, and for a given state σ_α, let $\sigma_{\alpha i}$ denote the status of individual i in that state.

From a given state σ_α, the process can transition to a new state in which one individual has changed state from S to I or from I to S. If the status of individual i is changing, we denote the new state by $\sigma_\alpha^{i\to X}$, where $X \in \{S, I\}$ is the new status of i. The transition rates for the Markov chain are given in Table 1, where δ is the Kronecker delta.

Table 1 Transitions for the Markovian network-based SIS model

from	to	at rate
$\sigma_\alpha : \sigma_{\alpha i} = S$	$\sigma_\alpha^{i\to I}$	$\sum_{j\in\mathcal{V}} T_{ij}\delta_{I\sigma_{\alpha j}}$
$\sigma_\alpha : \sigma_{\alpha i} = I$	$\sigma_\alpha^{i\to S}$	γ_i

The time evolution of the Markov chain is captured by the master equation

$$\frac{\mathrm{d}P(t)}{\mathrm{d}t} = QP(t), \tag{1}$$

where $P_\alpha(t) = P(\Sigma(t) = \sigma_\alpha)$ is the probability that the system is in state σ_α at time $t \geq 0$, and Q is a matrix of transition rates (obtained from Table 1). In particular, $P_1(t)$ denotes the probability that all individuals are susceptible at time t. Although this can be solved to determine the future behaviour, in many cases this is infeasible since the matrix Q grows rapidly with N.

3 The quasi-stationary distribution

Let us construct a vector $\rho(t)$, such that its components $\rho_\alpha(t)$, indexed by α, represent the conditional probability that the system is in state σ_α at time t given that at least one individual is infected; i.e. $\rho_\alpha(t) = P(\Sigma(t) = \sigma_\alpha | \Sigma(t) \neq \sigma_1)$, where σ_1 is the disease-free state. We have

$$\rho_\alpha(t) = \frac{P_\alpha(t)}{1 - P_1(t)}, \tag{2}$$

for $\alpha \neq 1$. For $\alpha = 1$, we set $\rho_1(t) = 0$ for all t. Here we have assumed that $P_1(t) \neq 1$ for all $t \geq 0$, which is satisfied whenever $P_1(0) \neq 1$, though as $t \to \infty$ the limit tends to 1, which can eventually make it numerically unstable to calculate the conditional probability this way. Using Equation (2) and the master equation (1), the time derivative of $\rho_\alpha(t)$ is given by

$$\frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} = \begin{cases} 0 & \text{if } \alpha = 1 \\ \frac{(QP)_\alpha}{1-P_1} + \frac{P_\alpha (QP)_1}{(1-P_1)^2} & \text{if } \alpha = 2, 3, \ldots, 2^N, \end{cases} \tag{3}$$

where we suppress the explicit time dependence of P and ρ in favour of compactness. The right-hand side can be expressed in terms of ρ by using Equation 2. However, we opt to keep this in terms of P since this form is used when developing the approximate models.

The state space for the Markov chain is finite and consists exhaustively of one absorbing state and a communicating class of transient states. The non-absorbing states form a communicating class of transient states because the contact network is strongly connected and the vector γ of recovery rates is positive. Thus, there exists a unique quasi-stationary distribution (QSD), independent of initial conditions, which is equivalent to the limiting conditional distribution [8]. This QSD, denoted by ρ^*, is a stationary distribution of the conditional probability and an equilibrium of Equation (3). Since ρ^* is unique, if we find some distribution P^* over all 2^N system states which satisfies

$$\frac{(QP^*)_\alpha}{1 - P_1^*} + \frac{P_\alpha^*(QP^*)_1}{(1-P_1^*)^2} = 0 \qquad \alpha = 2, 3, \ldots, 2^N,$$

then ρ^* is given by

$$\rho_\alpha^* = \begin{cases} 0 & \text{if } \alpha = 1 \\ \frac{P_\alpha^*}{1-P_1^*}, & \text{if } \alpha = 2, 3, \ldots, 2^N. \end{cases} \tag{4}$$

Here, for convenience, we define the QSD such that it assigns probability zero to the absorbing state, as opposed to leaving it undefined. Finding the QSD directly is in many cases infeasible since the size of the state space grows geometrically with the population size.

To go from the system-level master equation to node-level equations, we sum Equation (3) over all states in which individual $i \in \mathcal{V}$ is infected. Through this (see Appendix A.1), we arrive at an expression for the rate of change of the probability that i is infected conditioned on non-extinction

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=I)\right) = \frac{\sum_j T_{ij}\langle S_i I_j\rangle - \gamma_i\langle I_i\rangle}{1-P_1} + \frac{\langle I_i\rangle}{(1-P_1)^2}\sum_j \gamma_j\langle I_j S\rangle, \tag{5}$$

where $\langle S_i I_j\rangle$ is shorthand for $P(\Sigma_i(t)=S, \Sigma_j(t)=I)$, $\langle I_i\rangle$ is shorthand for $P(\Sigma_i(t)=I)$ and $\langle I_j S\rangle$ is shorthand for the probability that node j is infected and all other nodes are susceptible. As above, P_1 is the probability that all nodes are susceptible. Finding a steady-state solution such that

$$0 = \frac{\sum_j T_{ij}\langle S_i I_j\rangle - \gamma_i\langle I_i\rangle}{1-P_1} + \frac{\langle I_i\rangle}{(1-P_1)^2}\sum_j \gamma_j\langle I_j S\rangle, \tag{6}$$

the probability that node i is infected in the QSD can be calculated as

$$\langle I_i\rangle^{\mathrm{QSD}} = \frac{\langle I_i^*\rangle}{(1-P_1^*)}. \tag{7}$$

To find an exact solution to Equation (6) requires constructing a hierarchy describing how different states, ranging from pairs up to full system size, change in time, which is computationally no more efficient than solving Equation (4) directly. However, in this form we can employ moment-closure techniques to approximate these higher order terms. Such approaches are commonly used for approximating the standard probability distribution for epidemic models [12,13,17,18,20,25,44]. One approach is to assume statistical independence at the level of indidividuals in Equation (5). Alternatively, we can construct exact equations describing how the pair probability $\rho(\Sigma_i(t)=S, \Sigma_j(t)=I)$ changes in time, which we can approximate by assuming statistical independence at the level of pairs.

Although on the left-hand side of Equation (5) we define the conditional distribution, we retain the standard distributions on the right-hand side. It is possible to express the right-hand side in terms of conditional probabilities. However, through keeping the standard distributions, the approximations obtained in the later sections were found to be more reliable (not shown). By finding approximations that would cause the right-hand side to be zero, we can then transform these into approximations to the conditional distribution by using Equation (7), where both $\langle I_i^*\rangle$ and P_1^* will also need be approximated.

4 Approximating the QSD

In this section, we use moment closure methods to approximate the solution to Equation (5). The first approach is to assume that the states of neighbouring individuals are statistically independent, resulting in a relatively simple model that scales computationally with the number of nodes in the network. We then remove this assumption, and instead assume statistical independence at the level of pairs. This results in a more complex model that scales computationally with the number of pairs of nodes, but should capture the correlations between neighbouring nodes.

4.1 Individual-based approach

Approximating Equation (6) by assuming that the states of individuals are independent gives

$$0 \approx \frac{\sum_j T_{ij}\langle S_i\rangle\langle I_j\rangle - \gamma_i\langle I_i\rangle}{1-\prod_k\langle S_k\rangle} + \frac{\langle I_i\rangle}{(1-\prod_k\langle S_k\rangle)^2}\sum_j \gamma_j\langle I_j\rangle\prod_{k\neq j}\langle S_k\rangle.$$

To find the approximation to the probability that node i is infected in the QSD ($\langle I_i\rangle^{\mathrm{QSD}}$) under this independence assumption, we need to find vectors $\langle X\rangle^*$ and $\langle Y\rangle^*$ satisfying,

$$0 = \frac{\sum_j T_{ij}\langle X_i\rangle^*\langle Y_j\rangle^* - \gamma_i\langle Y_i\rangle^*}{1-\prod_k\langle X_k\rangle^*} + \frac{\langle Y_i\rangle^*}{(1-\prod_k\langle X_k\rangle^*)^2}\sum_j \gamma_j\langle Y_j\rangle^*\prod_{k\neq j}\langle X_k\rangle^*, \tag{8}$$

for all i. In the exact case, we need to scale the steady-state by the density remaining in the transient states (Equation (7)) to obtain the QSD probability. Following a similar procedure, from $\langle X\rangle^*$ and $\langle Y\rangle^*$, the probability that i is infected in the QSD is approximated by computing

$$\langle I_i\rangle^{\mathrm{QSD}}_{\mathrm{approx}} = \frac{\langle Y_i\rangle^*}{1-\prod_k \langle X_k\rangle^*}. \tag{9}$$

However, we are only interested in solutions of Equation (8) that are feasible; i.e. $\langle Y_i\rangle^* \in [0,1], \langle X_i\rangle^* = 1-\langle Y_i\rangle^*$ for all i. To obtain such a solution, define

$$\frac{\mathrm{d}\langle Y_i\rangle}{\mathrm{d}t} = \sum_j T_{ij}\langle X_i\rangle\langle Y_j\rangle - \gamma_i\langle Y_i\rangle + \frac{\langle Y_i\rangle \sum_j \gamma_j \langle Y_j\rangle \prod_{k\neq j}\langle X_k\rangle}{1-\prod_k \langle X_k\rangle} \quad , \quad \langle X_i\rangle = 1-\langle Y_i\rangle. \tag{10}$$

Equation (10) is positively invariant in $[0,1]^N$ (see Appendix A.2). This gives a system of N coupled equations, which can be numerically integrated until a steady state is reached. Alternatively, other fixed point analysis approaches can be applied. Starting from initial conditions satisfying $\langle X_i\rangle \in [0,1]$ and $\langle Y_i\rangle \in [0,1]$ at $t=0$ for all i, this process will give a feasible solution to Equation (8). From Equation (9), we can approximate the expected number infected in the QSD as

$$[I]^{\mathrm{QSD}}_{\mathrm{approx}} = \sum_i \langle I_i\rangle^{\mathrm{QSD}}_{\mathrm{approx}}.$$

We refer to this as the *node-level individual-based model*.

Theorem 1 *For a (strongly connected) K-regular graph, with homogeneous transmission and recovery rates, the node-level individual-based model yields a feasible approximation of the expected prevalence in the QSD such that* $\langle I_i\rangle^{\mathrm{QSD}}_{\mathrm{approx}} \in (0,1)$*, this being the same for all* $i \in \mathcal{V}$*, and* $[I]^{\mathrm{QSD}}_{\mathrm{approx}} \in (1,N)$*. On any strongly connected network, provided a solution exists such that* $\langle I_i\rangle^{\mathrm{QSD}}_{\mathrm{approx}} \in (0,1)$ *(which is found to hold numerically in all instances investigated), then* $[I]^{\mathrm{QSD}}_{\mathrm{approx}} \in (1,N)$*. Therefore, the number of infected individuals in the QSD is lower bounded by 1, a property which is shared by the true QSD.*

Proof Appendix A.3

As a further approximation to the expected number of infected individuals in the QSD, we can treat all individuals of a given degree equally by assuming

$$\langle I_i\rangle \approx \frac{[I_{k_i}]}{|C_{k_i}|} \quad , \quad T_{ij} = T_{k_i k_j} \quad , \quad \gamma_i = \gamma_{k_i} \qquad (i\in\mathcal{V}, j\in\mathcal{N}_i),$$

where k_i is the degree of node i, $[I_{k_i}]$ is the expected number of infected individuals with degree k_i, and $|C_k|$ is the number of degree k nodes. After summing over all i of a given degree, assuming statistical independence at the level of individuals, and setting the left hand side to zero, Equation (10) becomes a system of M equations in as many variables, where M is the number of unique node degrees in the network (see Appendix D). We refer to the resulting model as the *population-level individual-based model*. In the special case of a circle network, this population-level model yields a model developed by [9].

4.2 Pair-based approach

Assuming independence at the level of individuals may be unrealistic, since infection spreads through contact between individuals. Here, we keep Equation (5) without approximation, and also sum Equation (3) over all states in which individual $i \in \mathcal{V}$ is susceptible and individual $j \in \mathcal{N}_i$ is infected, so that we arrive at the equation for the rate of change of the probability that i is susceptible and j is infected conditioned on non-extinction (following a similar derivation to Equation (5)):

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=S, \Sigma_j(t)=I)\right) = \frac{\langle \dot{S_i I_j}\rangle}{1-P_1} + \frac{\langle S_i I_j\rangle}{(1-P_1)^2}\sum_j \gamma_j \langle I_j S\rangle \qquad (i\in\mathcal{V}, j\in\mathcal{N}_i), \tag{11}$$

where $\langle \dot{S_i I_j}\rangle$ is the rate of change in the probability that i is susceptible and j is infected under the standard distribution. The rate $\langle \dot{S_i I_j}\rangle$ depends on triple-probabilities (see Appendix B), which can be approximated in terms of individual-probabilities and pair-probabilities using

$$\langle A_i B_j C_k\rangle = \langle A_i C_k|B_j\rangle\langle B_j\rangle \approx \frac{\langle A_i B_j\rangle\langle B_j C_k\rangle}{\langle B_j\rangle}, \tag{12}$$

which assumes that the states of nodes i and k are independent when given the state of node j. Guided by this approximation for triple-probabilities, and following [12,45], we then approximate

$$\langle S\rangle \approx \frac{\prod_{i,j\in\mathcal{N}_i:j<i}\langle S_i S_j\rangle}{\prod_i \langle S_i\rangle^{n_i-1}} \quad , \quad \langle I_j S\rangle \approx \frac{\prod\limits_{x\in\mathcal{N}_j}\langle I_j S_x\rangle \prod\limits_{y\neq j}\prod\limits_{x\in\mathcal{N}_y:x<y,x\neq j}\langle S_y S_x\rangle}{\prod\limits_{x\neq j}\langle S_x\rangle^{k_x-1}\langle Y_j\rangle^{k_j-1}}. \tag{13}$$

Setting the left hand sides of equations (5) and (11) to zero, applying the above approximations, and imposing

$$\begin{aligned}\langle S_i\rangle &= 1-\langle I_i\rangle,\\ \langle S_i S_j\rangle &= \langle S_i\rangle - \langle S_i I_j\rangle,\\ \langle I_i I_j\rangle &= \langle I_i\rangle - \langle I_i S_j\rangle,\end{aligned}$$

yields a system of $N+\sum_i k_i$ equations in as many variables (see Appendix E). We refer to this as the *node-level pair-based model*.

Again, as a further approximation we can treat all individuals of a given degree equally, and all pairs of neighbours of given degrees equally, by assuming

$$\langle I_i\rangle \approx \frac{[I_{k_i}]}{|C_{k_i}|} \quad , \quad \langle S_i I_j\rangle \approx \frac{[S_{k_i} I_{k_j}]}{|C_{k_i k_j}|} \quad , \quad T_{ij} = T_{k_i k_j} \quad , \quad \gamma_i = \gamma_{k_i} \qquad (i\in\mathcal{V}, j\in\mathcal{N}_i),$$

where $|C_{k,l}|$ is the number of pairs between a degree k node and a degree l node, and $[S_{k_i} I_{k_j}]$ is the expected number of pairs involves a susceptible degree k_i node and an infected degree k_j node. After applying approximations (12) and (13), summing over all i of a given degree and over all pairs i and j of given degrees, and setting the left hand sides to zero, equations (5) and (11) become a system of $M+M^2$ equations (see Appendix F). We refer to this as the *population-level pair-based model*.

5 Numerical results

Here we determine how the methods developed in this paper perform when used to approximate the expected number of infected individuals in the QSD for various networks and parameter values.

We assume that: the transmission rate for any pair of connected individuals is equal (taking $T_{ij}=\tau$ whenever $T_{ij}>0$ and zero otherwise), the network is undirected, and infected individuals recover at the same rate; i.e. $\gamma_i=\gamma$ for all $i\in\mathcal{V}$. In the case of an evenly-mixed population, represented by a complete network, the epidemic threshold of the standard individual-based model is given by $(N-1)\times\tau/\gamma=1$. We therefore choose to plot the expected number of infected individuals against $\tau\times\bar{d}/\gamma$, where $\bar{d}$ is the average degree of the graph, to ensure that all networks are tested over a comparable range of epidemic severity. This is a rough approximation for epidemic severity, since in reality it depends on the degree distribution and correlations rather than just the average degree [17]. We assume $\gamma=1$ throughout, so that the ratio can be changed by changing τ.

The standard individual-based models (Appendix B) have been proven to possess a non-zero steady-state solution in the region of parameter space where the epidemic severity is large [20,46]. For the standard node-level pair-based model (Appendix B), we prove existence of a non-zero steady-state solution in Appendix C. When the transmission rate is sufficiently large, we observe that the models proposed in this paper converge to the standard models, so these endemic steady-states approximate the expected number of infected individuals in the QSD. To demonstrate this, the dynamics for these standard models are compared to the QSD approximation methods (Section 4). We are particularly interested in how our methods perform for low values of $\tau\times\bar{d}/\gamma$, where the standard models will not capture the QSD.

As a baseline for comparison, we simulate the stochastic SIS model using the Gillespie algorithm. To calculate the expected prevalence in the QSD, we average over all simulations that have not gone extinct. 100,000 simulations are run until $t=300$, since by this point all cases reached a steady-state. We compare the expected number of infected individuals in this solution with the steady-state of the QSD approximation methods and the standard models, solved using the Runge-Kutta method. For both the stochastic simulations and the approximation methods, the population is initiated with every node infected. This is to improve accuracy of the stochastic simulations, since a higher proportion will attain the QSD.

5.1 Impact of network structure

To test the methods, consider three networks: the complete network, the (NxN) square-lattice (with fixed boundaries), and Zachary's karate club [51]. The complete network represents a well-mixed population, in which all individuals are connected to each other. The square-lattice is a commonly used network when adding structure to population dynamics. We consider the variant with fixed boundaries, so the interior nodes have four neighbours, edge nodes have three neighbours and corner nodes have two neighbours. Although there is a lot of symmetry across the network, the regular structure with multiple loops can prove challenging for moment closure approximation methods. Zachary's karate club is an example of a real world network, formed from interactions between members of a karate club.

Figure 1(a) compares the node-level individual-based model (Equation (10)) with stochastic simulations. Below the epidemic threshold (where the standard model switches from zero to an endemic steady-state), the QSD method captures the behaviour reasonably accurately. As $\tau \times \bar{d}/\gamma$ increases, the approximation diverges, with differing levels of performance on each of the graphs tested. This individual-based method performs best on the complete network, on which it provides a good approximation to the expected number of infected individuals for all parameter values. Some level of accuracy is also observed on Zachary's Karate club. However, on the square-lattice this method does not perform well when above the epidemic threshold, significantly overestimating the expected number of infected individuals in the QSD. This is because the structure of the lattice results in significant local correlations which makes the assumption of statistical independence of individual nodes unrealistic.

Using the population-level individual-based model (see Section 4.1 and Appendix D), little accuracy is lost (Figure 1(b)). The same pattern of performance occurs across the three networks, and by overlaying the results, the population-level model is almost indistinguishable from the node-level model on the resolution of the graph. This suggests that the QSD approximation is mainly determined by the degree distribution, though there is likely to be some minor variations for graphs with the same distribution but differing in other network properties.

Since the assumption of individual-level statistical independence can lack accuracy, we developed a node-level pair-based model for the QSD (see Section 4.2 and Appendix E). Figure 2(a) shows the accuracy of this approximation, which is significantly improved over the individual-based models on all networks. On the complete network and Zachary's karate club, this approximation is very accurate, and on the lattice it loses some accuracy but significantly outperforms the individual-based approximation. The loss of accuracy on the lattice is expected, since pair-approximation methods are generally considered to perform weakly on such structures.

Although the pair-based model is computationally feasible, for large graphs it can be slow. Therefore, we derived a population-level pair-based model (see Section 4.2 and Appendix F). Again, little accuracy is lost for all networks (Figure 2(b)), with the result being indistinguishable from the node-level model.

For each of the methods proposed, a stationary solution is reached for all parameter values on all networks. These solutions appear to be unique and lower bounded by 1. Therefore, the proposed methods satisfy the basic properties of the QSD. Sufficiently above the epidemic threshold, our models and the standard (unconditioned) models coincide (Figures 1 and 2), showing that the standard models approximate the expected number of infected individuals in the QSD in this region. However, as the transmission rate decreases, the steady states of the standard models deviate from this, eventually tending to the disease-free steady-state. Therefore, the standard models are not a reliable measure of the expected prevalence in the QSD since they do not capture this for all parameter values, and the endemic steady-state in the intermediate range (between the disease-free steady-state and coinciding with the QSD model) is hard to relate to any properties of the underlying stochastic process. The models we propose are more robust for providing insight into the stochastic epidemic model.

5.2 Impact of network size

We now investigate how increasing the size of the population affects the accuracy of the results, testing a 100 node (10x10) lattice, 225 node (15x15) lattice and 400 node (20x20) lattice. Here the square-lattice is chosen because this presented itself as the worst case, with other networks expected to perform better. The lattice is expected to perform badly because the strict structure leads to very high local correlations, which may not be captured by the moment-closure approximations.

Since the population-level models perform similarly to the node-level models at capturing the expected number of infected individuals, with significantly reduced computational cost, in this section we

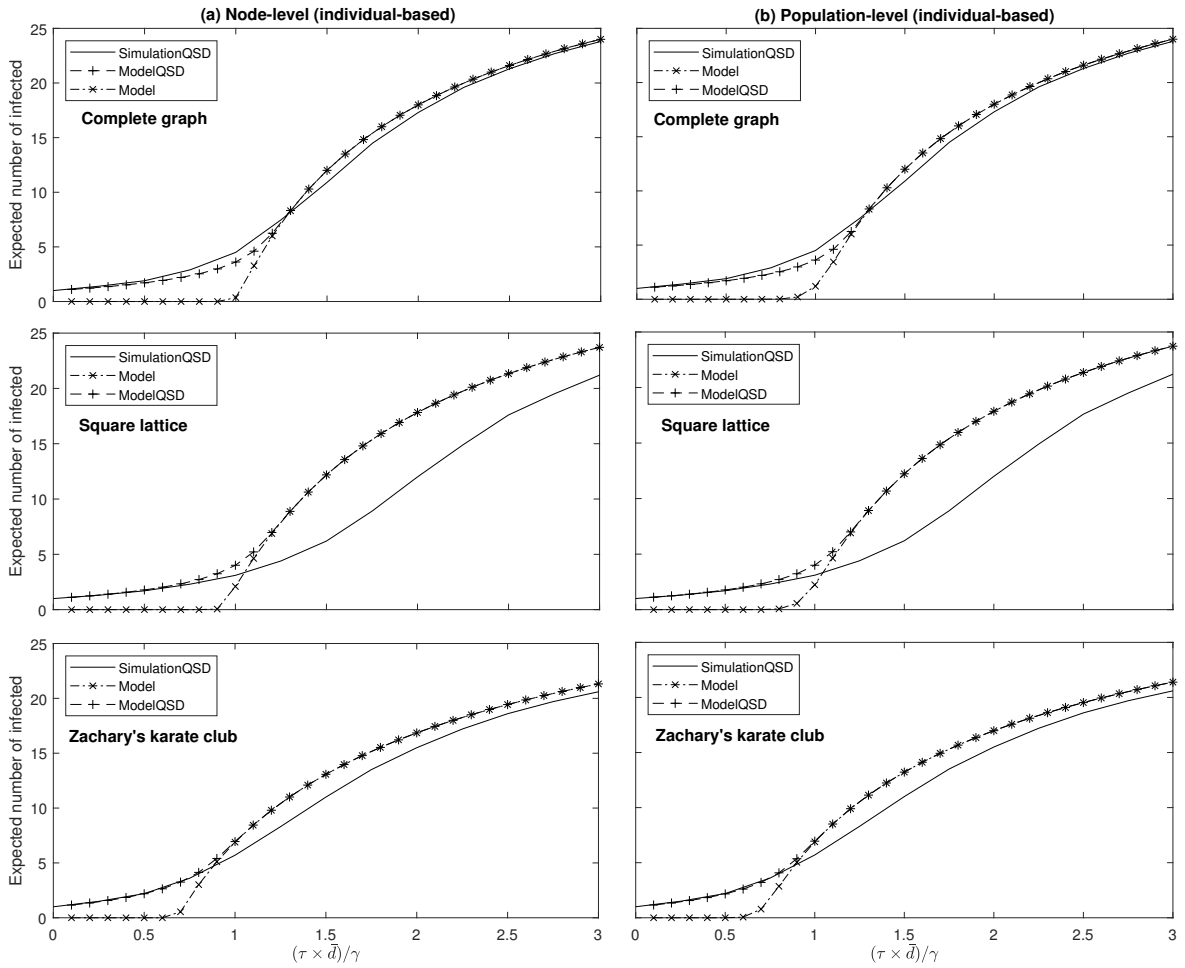

Fig. 1 The expected number of infected individuals in the QSD as calculated by the individual-based model versus stochastic simulation on a 36 node complete network, 36 node (6x6) square-lattice and the 34 node karate club network, for a range of parameter values. The left plot shows the node-level methods (Equation (10)) and the right shows the population-level methods (Equation (D3)). The solid lines represent the average of 10,000 stochastic simulations conditioned against extinction, the dashed line (plusses) represents the proposed QSD approximation method and the dash-dotted line (crosses) represent the standard unconditioned model. The simulated QSD is accurate to within the resolution of the line.

only use these models to approximate the dynamics. Comparing the QSD method to the simulation results (Figure 3), we see good agreement for low transmission parameters for both individual-based and pair-based methods. However, once the individual-based methods pass the epidemic threshold, where the standard method reaches a non-zero steady state, both the standard method and the QSD method diverge significantly from the simulation results, overestimating the true expected number of infected in the QSD, echoing what we observed in Figure 1. For the pair-based models, once the parameters exceed the epidemic threshold, we still see some deviation from the simulation results for both the standard and QSD methods. However, this is much smaller than for the individual-based methods. For higher relative transmission rates, the model solutions provide a reasonable approximation to the expected number of infected individuals in the QSD. For the three lattice sizes considered, in the regions below and sufficiently above the epidemic threshold, the relative magnitude of the discrepancy between the approximations and simulation results does not change with population size, for both individual-based and pair-based models. However, in the intermediate region there is some sensitivity to population size. Below the epidemic threshold, the standard models do not capture the dynamics of the QSD, regardless of population size, whereas the QSD approximation models are accurate.

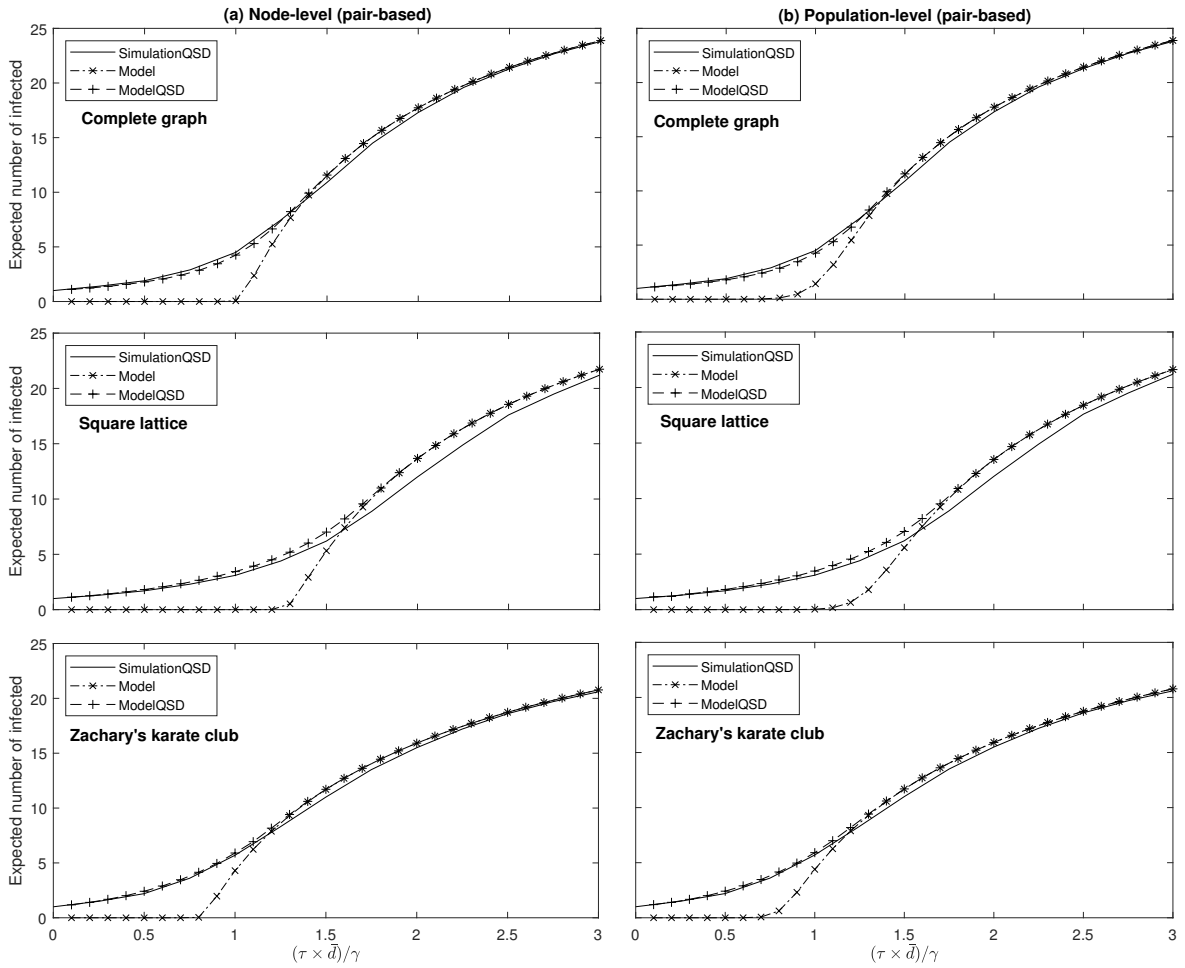

Fig. 2 The expected number of infected individuals in the QSD as calculated by the pair-based model versus stochastic simulation on a 36 node complete network, 36 node (6x6) square-lattice and the 34 node karate club network, for a range of parameter values. The left plot shows the node-level methods (Equation (E4)) and the right shows the population-level methods (Equation (F3)). The solid lines represent the average of 10,000 stochastic simulations conditioned against extinction, the dashed line (plusses) represents the proposed QSD approximation method and the dash-dotted line (crosses) represent the standard unconditioned model. The simulated QSD is accurate to within the resolution of the line.

6 Discussion

The standard deterministic SIS model [23,46–48] exhibits an epidemic threshold below which the pathogen will go extinct and above which the pathogen will reach an endemic steady-state solution [23,46]. More complicated 'deterministic' models have been developed, such as pair-approximations models [12,13,17, 18,20,25,44], in which this threshold behaviour is also observed [17,25]. However, no steady-state solution exists in the stochastic SIS model, making it hard to relate the deterministic and stochastic models in finite populations.

When the time to absorption (extinction of the pathogen) is long enough, the endemic equilibrium of the network-based deterministic SIS models has been observed to approximate the expected number of infected individuals over short enough time scales. This is quantified by the quasi-stationary distribution (QSD) of the stochastic models [1,20,46]. Although the unconditioned models numerically approximate the expected prevalence in the QSD when sufficiently above threshold, this is not the case as the system approaches the threshold (and below threshold). In particular, comparison of a model with a genuine steady state with one without such a state is not well-defined. To correct this discrepancy, we constructed approximate models which are formally related to the stochastic dynamics via the QSD. The resulting models capture properties of the QSD at all levels of epidemic severity.

Our first approach assumed that the states of neighbouring nodes are independent, at both node-level and population-level. Although this assumption is not particularly realistic, on the complete network this

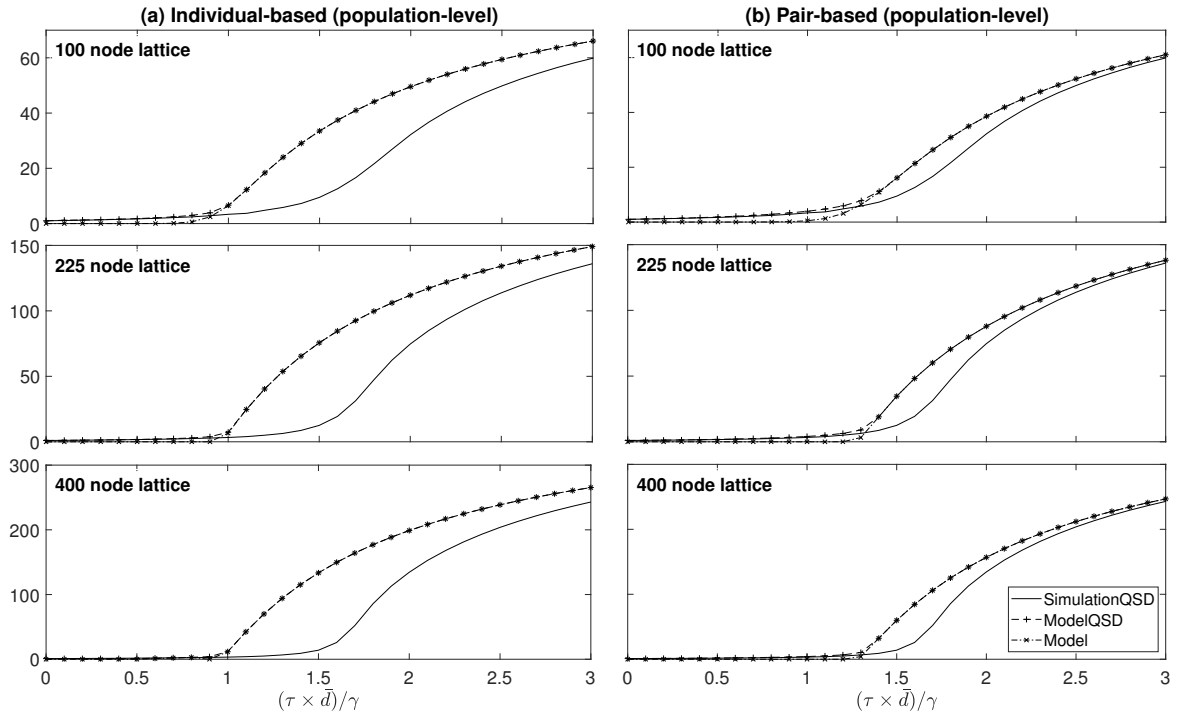

Fig. 3 The expected number of infected individuals in the QSD as calculated by the population-level models versus stochastic simulation on a 10x10 node square-lattice, 15x15 node square-lattice, and 20x20 node square-lattice for a range of parameter values. The left plot shows the individual-based methods (Equation (D3)) and the right shows the pair-based methods (Equation (F3)). The solid lines represent the average of 100,000 stochastic simulations conditioned against extinction, the dashed lines (plusses) represent the proposed QSD approximation method and the dash-dotted lines (crosses) represent the standard unconditioned model. The simulated QSD is accurate to within the resolution of the line.

provides a good approximation to the expected number of infected individuals in the QSD (Figure 1). However, for more structured networks the accuracy decreased. Little accuracy was lost when computing the expected number of infected individuals compared to the node-level model, with a significant reduction in computational cost. To improve accuracy, we developed node-level and population-level methods based on assuming independence at the level of pairs, which performed well on all networks tested (Figure 2). Again, little accuracy was lost in the population-level model. With the significant reduction in computational cost, the population-level models are therefore superior to the node-level models for capturing the expected prevalence in the QSD. However, one advantage of the node-level models is the insight these can give into the dynamics of individual nodes in the population, which the population-level models lose.

With the standard unconditioned approximation methods, it is not inherently clear what the models are capturing, since the stochastic model does not exhibit a stable steady-state. By developing conditioned approximation models that capture the quasi-stationary distribution of the stochastic model, we have presented an approximation framework that is directly related to the underlying stochastic process. Sufficiently above the epidemic threshold, the unconditioned standard models coincide with the conditioned QSD models, demonstrating, as expected, that the standard models approximate the QSD when above threshold. Through directly approximating the QSD, the conditioned models are consistent in approximating the QSD for all parameter values. This consistency makes them a more robust method for capturing quasi-stationary behaviour of stochastic epidemic models.

This paper has focussed on the theoretical insights this model grants, and we have shown that the models can be reasonably accurate on a variety of networks. In particular, we show that the pair-based model can perform well on a square-lattice, which is expected to be one of the worst cases for moment-closure approximations. The accuracy and deterministic nature of the models makes them more amenable to analysing how different network structures can alter the statistics of the QSD than the use of stochastic simulation. This is valuable for characterising the likelihood and severity of the epidemic, for example through the invasion probability [50], which can be calculated directly from the node-level models proposed, and the expected prevalence, which we presented in the numerical results. The potential future applications of this work include applying the methods to investigate how network structure, such

as the degree variance, affects the properties of the QSD, as well as extending the model to other epidemic and population dynamics models.

Appendices

Appendix A Node-level individual-based QSD model

A.1 Derivation of node-level conditional distribution equation

The rate of change in the probability that node i is infected in the QSD is given by the sum of the rates of change in the full system state probabilities for which node i is infected. That is, we have

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=I)\right) &= \sum_{\alpha:\sigma_{\alpha i}=I}\frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t}\\ &= \frac{\sum_{\alpha:\sigma_{\alpha i}=I}(QP)_\alpha}{1-P_1}+\frac{(QP)_1}{(1-P_1)^2}\sum_{\alpha:\sigma_{\alpha i}=I}P_\alpha,\end{aligned}$$

where the terms are defined in Section 3. The numerator of the first term on the second line corresponds to the rate of change in the probability that node i is infected, which is given by $\langle\dot{I}_i\rangle$ in Equation (B3) in Appendix B. The summation in the second term corresponds to the probability that node i is infected, $\langle I_i\rangle$. Therefore, we can write

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=I)\right)=\frac{\langle\dot{I}_i\rangle}{1-P_1}+\frac{(QP)_1}{(1-P_1)^2}\langle I_i\rangle.$$

Here $(QP)_1$ is the rate at which the system enters the absorbing state. The system can only reach the absorbing state from a state with a single infected individual, in node j for example, which transitions to the all susceptible state at rate γ_j. Therefore $(QP)_1=\sum_j\gamma_j\langle I_jS\rangle$, where we use $\langle I_jS\rangle$ to denote the probability that node j is infected and all other nodes are susceptible. Using this along with Equation (B3), we obtain

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=I)\right)=\frac{\sum_j T_{ij}\langle S_iI_j\rangle-\gamma_i\langle I_i\rangle}{1-P_1}+\frac{\langle I_i\rangle}{(1-P_1)^2}\sum_j\gamma_j\langle I_jS\rangle.$$

A.2 Proof that the individual-based node-level QSD model is invariant on $[0,1]^N$.

Proof To prove that the model in Equation (10) is invariant we use the method from [23]. Along the boundaries to the set we are interested in, we either have $\langle Y_i\rangle=0$ and $\langle X_i\rangle=1$ or $\langle Y_i\rangle=1$ and $\langle X_i\rangle=0$. To show the system is invariant, we need to show that along these boundaries the trajectories do not point away from this set.

First consider $\langle Y_i\rangle=0$. At this boundary, we have

$$\langle\dot{Y}_i\rangle=\sum_j T_{ij}\langle Y_j\rangle.$$

If $\langle Y_i\rangle\in[0,1]$, this cannot be negative, and therefore at $\langle Y_i\rangle=0$ the trajectory in the i direction cannot leave the set $[0,1]^N$. Now consider $\langle Y_i\rangle=1$. We have

$$\langle\dot{Y}_i\rangle=-\gamma_i+\gamma_i\prod_{k\neq i}\langle X_k\rangle.$$

The product in this equation is in $[0,1]$ if $\langle X_k\rangle\in[0,1]$ for all k. Therefore, this equation can never be positive, so along this boundary the trajectory cannot leave the set $[0,1]$. Therefore, this model is invariant on $[0,1]^N$.

A.3 Proof of Theorem 1

Proof Consider the node-level individual-based model (Equation (10)) on a k-regular network with homogeneous transmission and recovery. If we start with a fully infected population, $\langle Y_i \rangle$ will be equal for all i at every time point. Therefore we can denote $\langle I_i \rangle = a$ for all $i \in \mathcal{V}$. We can write the rate of change in the node probabilities as

$$\dot{a} = -\gamma a + \tau(k)(1-a) + a\frac{\gamma N a(1-a)^{N-1}}{1-(1-a)^N}. \tag{A2}$$

In the steady state $\dot{a} = 0$. If we rule out $a = 0$, since Equation (A2) is undefined for $a = 0$, then we obtain

$$(1-a)(k)\left(\frac{\tau}{\gamma} + \frac{N}{k}\frac{a(1-a)^{N-2}}{1-(1-a)^N}\right) = 1.$$

We are therefore interested in solutions to $f(a) = 0$ with $a \in [0,1]$, where

$$f(a) = (1-a)(k)\left(\frac{\tau}{\gamma} + \frac{N}{k}\frac{a(1-a)^{N-2}}{1-(1-a)^N}\right) - 1.$$

To see if a solution exists within this interval we check the signs at the end points.

At $a = 1$

$$f(1) = -1 < 0$$

the function is negative.

As a goes to zero

$$\lim_{a\to 0} f(a) = (k)\frac{\tau}{\gamma} - 1 + \lim_{a\to 0} N\frac{a(1-a)^{N-1}}{1-(1-a)^N}.$$

$$\lim_{a\to 0} N\frac{a(1-a)^{N-1}}{1-(1-a)^N} = \lim_{a\to 0} N\frac{(1-a)^{N-1} + (N-1)a(1-a)^{N-2}}{N(1-a)^{N-1}} = 1.$$

$$\implies f(0) = (k)\frac{\tau}{\gamma} > 0 \text{ if } \frac{\tau}{\gamma} > 0.$$

Therefore as long as the transmission rate τ is greater than zero there exists a solution to $f(a) = 0$ in the open interval $(0,1)$, since $f(a)$ is non-singular on $(0,1)$.

We now need to show that our approximation to the expected number of infected individuals in the QSD is bounded below by one. This proof holds for all networks provided a solution exists satisfying $\langle Y_i \rangle \in (0,1)$ for all i, which we have proven for k-regular networks. Consider the node-level individual-based model; i.e.

$$\langle \dot{Y}_i \rangle = -\gamma_i \langle Y_i \rangle + \sum_j T_{ij}\langle X_i \rangle \langle Y_j \rangle + \frac{\langle Y_i \rangle}{1 - \prod_k \langle X_k \rangle}\sum_j \gamma_j \langle Y_j \rangle \prod_{k\neq j}\langle X_k \rangle \tag{A3}$$

To approximate the QSD we calculate $\langle Y_i^* \rangle / (1 - \prod_k \langle X_k^* \rangle)$, where $\langle Y^* \rangle$ and $\langle X^* \rangle$ are steady-state solutions to (A3).

Let S be the sum of N independent Bernoulli random variables with success probabilities given by the vector $\langle Y_i^* \rangle$ for $i \in \{1, 2, \ldots, N\}$, which is a feasible solution of Equation (A3). It is straightforward then that $\mathbb{E}[S] = \sum_i \langle Y_i^* \rangle$, and we can write

$$\begin{aligned}\sum_i \langle Y_i^* \rangle =& \mathbb{E}[S] = \sum_{x=1}^{x=N} \mathrm{P}(S = x)x \geq \sum_{x=1}^{x=N} \mathrm{P}(S = x) = \mathrm{P}(S \geq 1) \\ =& 1 - \prod_j (1 - \langle Y_j^* \rangle)\end{aligned}$$

So when we approximate the expected number infected in the QSD as

$$\frac{\sum_i \langle Y_i^* \rangle}{1 - \prod_j (1 - \langle Y_j^* \rangle)}$$

this cannot be less than 1. Therefore, provided a non-zero solution exists to Equation (A3), the approximation to the expected number of infected individuals in the QSD is not less than 1.

Appendix B Standard approximate models

Due to the prohibitive computational cost of solving the master equation (Equation (1)), approximation methods are useful. In this section, we give an overview of the heterogeneous mean-field and pair-approximation methods, which can be interpreted as approximating the expected behaviour of the stochastic model. For detailed derivations and analysis of these models see [20].

Under the heterogeneous mean-field model, we assume that: all individuals with the same degree can be treated identically, the status of neighbouring individuals are independent, $\gamma_i = \gamma$ for all $i \in \mathcal{V}$, and $T_{ij} = \tau$ for all $i, j \in \mathcal{V}$ with $T_{ij} > 0$ or $T_{ji} > 0$ (the network is assumed undirected for simplicity). The rate of change in the expected number of susceptible and infected individuals, stratified by the degree of the individual, is then approximated by [20]

$$\begin{aligned}[\dot{S_k}] &\approx -\tau \sum_{l\in\mathcal{M}} |C_{k,l}| \frac{[S_k]}{|C_k|}\frac{[I_l]}{|C_l|} + \gamma[I_k] \\ [\dot{I_k}] &\approx \tau \sum_{l\in\mathcal{M}} |C_{k,l}| \frac{[S_k]}{|C_k|}\frac{[I_l]}{|C_l|} - \gamma[I_k],\end{aligned}$$

where $[S_k]$ is the expected number of susceptible individuals of degree k at time t, $|C_k|$ is the number of degree k nodes, $|C_{k,l}|$ is the number of pairs involving a degree k node and a degree l node, and $\mathcal{M}$ is the set of unique degrees on the network. Above, and throughout, we use 'dot' notation for derivatives with respect to time. Whilst the assumption of neighbouring individuals being independent is unrealistic, the resulting model has low computational cost, and hence it is popular to study.

Instead of assuming statistical independence between individuals, models have been derived by writing down exact equations for the expected number of individuals and pairs:

$$\begin{aligned}[\dot{S_k}] =& \gamma[I_k] - \sum_{l\in\mathcal{M}} \tau[S_k I_l] \\ [\dot{I_k}] =& -\gamma[I_k] + \sum_{l\in\mathcal{M}} \tau[S_k I_l] \\ [\dot{S_k I_l}] =& \gamma([I_k I_l] - [S_k I_l]) + \tau(\sum_{m\in\mathcal{M}} [S_k S_l I_m] - \sum_{m\in\mathcal{M}} [I_m S_k I_l] - [S_k I_l]) \\ [\dot{S_k S_l}] =& \gamma([S_k I_l] + [I_k S_l]) - \tau(\sum_{m\in\mathcal{M}} [S_k S_l I_m] + \sum_{m\in\mathcal{M}} [I_m S_k S_l]) \\ [\dot{I_k I_l}] =& \tau([S_k I_l] + [I_k S_l] - 2\gamma[I_k I_l] + \tau(\sum_{m\in\mathcal{M}} [I_m S_k I_l] + \sum_{m\in\mathcal{M}} [I_k S_l I_m]),\end{aligned} \tag{B1}$$

where $[A_k B_l]$ is the expected number of pairs at time t, between degree k and l individuals in states A and B respectively, and $[A_k B_l C_h]$ is the expected number of triples at time t, between degree k, l and h individuals, in states A, B and C respectively.

Solving this system exactly involves deriving a full hierarchy of equations describing triples and quads and so on [10], and therefore we wish to approximate this system by closing the hierarchy early. This can be done by expressing triples as some function of pairs and individuals. To approximate the triples, we analyse the number of edges starting from a susceptible node, following [10,20]. The total number of SA edges (for $A \in \{S, I\}$) from a degree k node to a degree l node are $[S_k A_l]$. Since we have $[S_k]$ susceptible degree k nodes, we have approximately $[S_k A_l]/(k[S_k])$ edges leading from a given susceptible degree k node to a given degree l node in state A. Therefore, for a chosen susceptible degree k node the probability that two neighbours, with degree l and m, are in states A and B is given by $[A_l S_k][S_k B_m]/k^2[S_k]^2$. We have $k(k-1)$ choices of the two neighbours, and $[S_k]$ choices of the susceptible node, and therefore we can approximate the expected number of triples $[A_l S_k B_m]$ as

$$[A_l S_k I_m] \approx \frac{k-1}{k}\frac{[A_l S_k][S_k I_m]}{[S_k]}. \tag{B2}$$

This approximation makes the homogeneity assumption that the neighbours of susceptible degree k nodes are interchangeable and the states of pairs are independent. Using this expression, the system of equations (B1) is closed at the level of pair terms, which allows the system to be solved with reasonably low computational cost.

These two models act at the population level, since they describe how the expected number of individuals with certain traits change. Following the motivation behind these models, node-level models have been developed that describe how the probability of individual nodes being infected change with time. Such models have been referred to as individual-based models [44,45], node-level models [32], propagation models [20] or quenched-mean field [11,25]. The advantage of such models over the population-level models is that we do not need to make any homogeneity assumptions about the underlying populations, and therefore properties such as clustering, directed edges and degree heterogeneity are naturally captured. The downside however is that the computational cost scales with at least the number of nodes.

Under Markovian network-based SIS, the dynamics of individual nodes are given by [44]

$$\begin{aligned}\langle \dot{S}_i \rangle &= -\sum_j T_{ij} \langle S_i I_j \rangle + \gamma_i \langle I_i \rangle, \\ \langle \dot{I}_i \rangle &= \sum_j T_{ij} \langle S_i I_j \rangle - \gamma_i \langle I_i \rangle,\end{aligned} \tag{B3}$$

where $\langle A_i \rangle$ represents the probability $P(\Sigma_i(t) = A)$ with $A \in \{S, I\}$, and $\langle A_i B_j \rangle$ represents the probability $P(\Sigma_i(t) = A, \Sigma_j(t) = B)$ with $A, B \in \{S, I\}$.

This equation exactly describes the rate of change for individual nodes in terms of pairs. Pairs of nodes are exactly described by

$$\begin{aligned}\langle \dot{S_i I_j} \rangle =& \sum_k T_{jk} \langle S_i S_j I_k \rangle - \sum_k T_{ik} \langle I_k S_i I_j \rangle \\ & - (T_{ij} + \gamma_j) \langle S_i I_j \rangle + \gamma_i \langle I_i I_j \rangle, \\ \langle \dot{S_i S_j} \rangle =& - \sum_k T_{jk} \langle S_i S_j I_k \rangle - \sum_k T_{ik} \langle I_k S_i S_j \rangle, \\ \langle \dot{I_i I_j} \rangle =& \sum_k T_{jk} \langle I_i S_j I_k \rangle + \sum_k T_{ik} \langle I_k S_i I_j \rangle - (\gamma_i + \gamma_j) \langle I_i I_j \rangle \\ & + T_{ij} \langle S_i I_j \rangle + T_{ji} \langle I_i S_j \rangle,\end{aligned} \tag{B4}$$

where $\langle A_i B_j C_k \rangle$ represents the probability $P(\Sigma_i(t) = A, \Sigma_j(t) = B, \Sigma_k(t) = C)$ with $A, B, C \in \{S, I\}$. To solve this requires a hierarchy of equations up to full system size. Following similar logic to the population-level equations, this system can be approximated by making assumptions of statistical independence. Assuming that the states of individuals are independent, $\langle S_i I_j \rangle \approx \langle S_i \rangle \langle I_j \rangle$, we can close the hierarchy at the level of individuals. Alternatively, we can assume independence at the level of pairs. The natural assumption of statistical independence to apply to pairs is that, given three nodes in a line, if the state of the central node is known then the state of the outer two nodes are independent. For all triples in the system above, the central node in the configuration is always the centre node of a line between the two outer nodes. Therefore, if we consider the triple $\langle A_i B_j C_k \rangle$, this can be approximated as a function of lower order terms by using conditional probabilities and assuming statistical independence. By the definition of conditional probabilities, we obtain

$$\langle A_i B_j C_k \rangle = \langle A_i C_k | B_j \rangle \langle B_j \rangle.$$

Assuming that the states of nodes i and k are independent given the state of node j, this becomes

$$\langle A_i B_j C_k \rangle \approx \langle A_i | B_j \rangle \langle C_k | B_j \rangle \langle B_j \rangle = \frac{\langle A_i B_j \rangle \langle B_j C_k \rangle}{\langle B_j \rangle}, \tag{B5}$$

which closes the hierarchy at the level of pairs. Other methods to approximate triples in terms of pairs and individuals have been proposed [17,42,44], however we do not consider them in this paper.

The population-level methods described above can be derived rigorously from the node-level methods [44]. In the exact case, we have

$$[A_k] = \sum_{j:k_j=k} \langle A_j \rangle$$

and

$$[A_k B_l] = \sum_{i:k_i=k} \sum_{j:k_j=l} \langle A_i B_j \rangle$$

where $A, B \in \{S, I\}$ and k_i is the degree of node i. Using this, the rate of change for the population-level terms can be derived. From this, we can also approximate the node-level quantities as

$$\langle A_i \rangle \approx \frac{[A_{k_i}]}{|C_{k_i}|}, \tag{B6}$$

and

$$\langle A_i B_j \rangle \approx \frac{[A_{k_i} B_{k_j}]}{|C_{k_i,k_j}|}.$$

The models described here exhibit an epidemic threshold, above which the pathogen persists and below which the pathogen dies out (illustrated in Figure B1 for the node-level pair-based model). For the population-level models and individual-based node-level model, above these thresholds a unique, globally stable steady-state exists [17,18,20,23,46]. For the node-level pair-based model, the disease-free solution has been shown to become unstable as the transmission rate increases [25], at which point we have shown that an endemic steady-state solution exists (Appendix C). Numerically, this endemic equilibrium appears to be unique and globally attracting, similar to the endemic solutions in the other models.

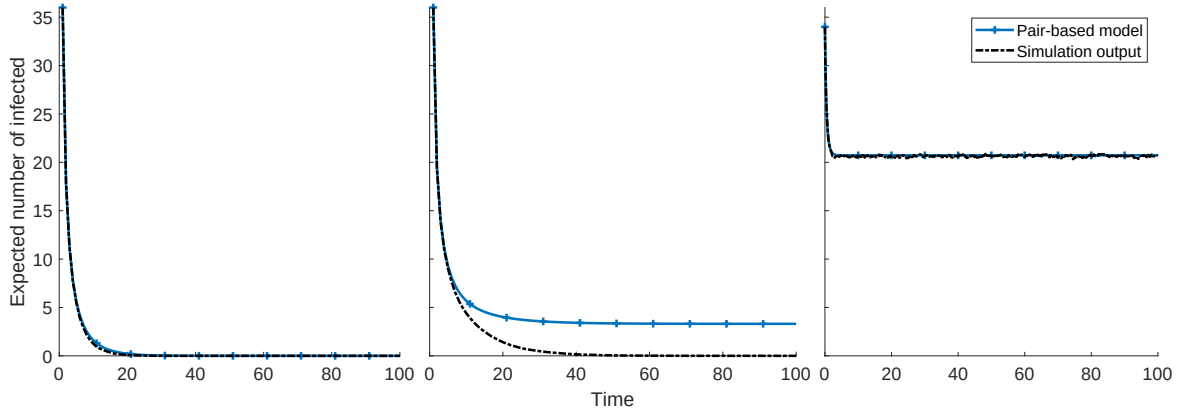

Fig. B1 Comparing the standard pair-based model (equations (B3) and (B4) with the closure from Equation (B5)) with closures with the output of stochastic simulations on Zachary's karate club network. We plot the expected number of infected individuals against time for each of the methods. As the figures move from left to right the transmission rate increases. In the right-most figure, steady-like behaviour is observed in the stochastic model, since the expected time to extinction is very long.

When comparing these models to the underlying stochastic process (e.g. Figure B1), below the epidemic threshold the models accurately capture the expected number of infected individuals in the stochastic process. However, as the transmission rate increases (or recovery rate decreases), we pass the epidemic threshold, and observe an endemic equilibrium that does not correspond to the stochastic process. Eventually, when the parameters are sufficiently above the epidemic threshold, the endemic steady-state solutions of these models can approximate the behaviour of the stochastic model for a long time, since the time to extinction of the pathogen is very long. Here, the stochastic process behaves similarly to the quasi-stationary distribution of the model; i.e. the expected long-term behaviour if extinction has not occurred.

Appendix C Proof of existence of an endemic steady-state for the standard pair-based model

Proof In [23], a theorem is proven regarding the existence of stable endemic solutions for ordinary differential equation epidemic models. Here we demonstrate that the standard pair-based SIS model (equations (B3) and (B4) with the closure from Equation (B5) [25]) satisfies the requirements for this proof, and therefore has a stable endemic steady-state.

Consider an ODE of the form

$$\frac{\mathrm{d}y}{\mathrm{d}t} = Ay + N(y). \tag{C1}$$

If the following statements hold, then there exists a threshold above which an endemic steady-state exists.

1. A compact convex set C on the domain of N is positively invariant, with $y = 0 \in C$.
2. $\lim_{y\to 0} ||N(y)||/||y|| = 0$
3. There exists $r > 0$ and a real eigenvector w or A^T such that $(w \cdot y) \geq r||y|| \quad \forall y \in C$
4. $(w \cdot N(y)) \leq 0 \quad \forall y \in C$
5. $y = 0$ is the largest positively invariant set contained in $H = \{y \in C | (w \cdot N(y)) = 0\}$

The first step is to write the pair-based model in the form (C1). The pair-based model is given by

$$\begin{aligned}\langle \dot{I}_i \rangle =& \sum_j^N T_{ij} \langle S_i I_j \rangle - \gamma \langle I_i \rangle \\ \langle \dot{S_i I_j} \rangle =& \sum_{k\neq i}^N T_{jk} \frac{\langle S_i S_j \rangle \langle S_j I_k \rangle}{\langle S_j \rangle} - \sum_{k\neq j}^N T_{ik} \frac{\langle I_k S_i \rangle \langle S_i I_j \rangle}{\langle S_i \rangle} - \langle S_i I_j \rangle - \gamma \langle S_i I_j \rangle + \gamma \langle I_i I_j \rangle,\end{aligned}$$

where $\langle S_i \rangle = 1 - \langle I_i \rangle$, $\langle I_i I_j \rangle = \langle I_j \rangle - \langle S_i I_j \rangle$ and $\langle S_i S_j \rangle = \langle S_i \rangle - \langle S_i I_j \rangle$.

This can be rewritten as

$$\begin{aligned}\langle \dot{I}_i \rangle =& \sum_j^N T_{ij} \langle S_i I_j \rangle - \gamma \langle I_i \rangle \\ \langle \dot{S_i I_j} \rangle =& -(T_{ij} + 2\gamma) \langle S_i I_j \rangle + \gamma \langle I_j \rangle + \sum_{k\neq i}^N T_{jk} \langle S_j I_k \rangle \\ & - \sum_{k\neq i}^N T_{jk} \frac{\langle I_i S_j \rangle \langle S_j I_k \rangle}{\langle S_j \rangle} - \sum_{k\neq j}^N T_{ik} \frac{\langle I_k S_i \rangle \langle S_i I_j \rangle}{\langle S_i \rangle}.\end{aligned}$$

Defining $y_i = \langle I_i \rangle$ for $1 \leq i \leq N$ and $y_i = \langle S_1 I_{i-N} \rangle$ for $N+1 \leq i \leq 2N$, $y_i = \langle S_2 I_{i-2N} \rangle$ for $2N+1 \leq i \leq 3N$, and so on, we can write the pair-based model in the form of Equation (C1). Compiling the linear terms into the matrix A, we see that A is only negative on the diagonal. The remaining non-linear terms define the function $N(y)$, which only assigns negative values to each input. Now it is required to check if the properties hold.

Property (1.) holds because the system is invariant on the set $C = \{0 \leq \langle I_i \rangle \leq 1; 0 \leq \langle S_i I_j \rangle \leq 1\}$. Property (2.) holds because as $y \to 0$ the denominator of all terms, $1 - \langle I_i \rangle$, goes to one, and the numerator is of the form $y_i y_j$, which goes to zero faster than y_i and y_j. Property (3.) holds because A is irreducible since all the equations are coupled. Since A is only negative on the diagonal, by the Perron-Frobenius theorem, A^T must have an eigenvector w such that $w_i > 0$ for all i. Property (4.) holds because the function $N(y)$ is negative, so $(w \cdot N(y)) \leq 0$, since $w_i > 0$ for all i. We now need to test property (5.).

Property (5.) If $y \in H$ then $(w \cdot N(y)) = 0$. This implies that

$$w_i \sum_{k\neq i} \frac{T_{jk} \langle I_i S_j \rangle \langle S_j I_k \rangle}{1 - \langle I_j \rangle} = 0$$

and

$$w_i \sum_{k\neq j} \frac{T_{ik} \langle I_k S_i \rangle \langle S_i I_j \rangle}{1 - \langle I_i \rangle} = 0,$$

for all pairs (i, j). If we assume that $y \in H$ and $y \neq 0$, then $y_h \neq 0$ for some h. If we assume that $y_h = \langle S_i I_j \rangle \neq 0$, then we must have $\langle S_i I_k \rangle = 0$, for all $k \in \mathcal{N}_i$. Also, we require $\langle S_j I_k \rangle = 0$ for some k or $\langle I_i S_j \rangle = 0$. We now need to investigate whether such a state can be invariant.

Define $S = \{i : y_i = 0\}$ and $S' = \{i : y_i \neq 0\}$, both of which are non-empty since $y \neq 0$ and $\langle S_i I_j \rangle = 0$ for some pair $(i.j)$ by the above argument. Since A is irreducible, there must exist a pair $k \in S$ and $h \in S'$ such that dy_k/dt depends on y_h.

First assume that $y_h = \langle S_i I_j \rangle$ and $y_k = \langle I_i \rangle$. We have

$$\frac{\mathrm{d}y_k}{\mathrm{d}t} = \sum_{j\neq i} T_{ij} \langle S_i I_j \rangle$$

If this state is invariant, then $dy/dt = 0$, which implies that $dy_k/dt = 0$ for all k. This can only be the case if $\langle S_k I_j \rangle = 0$ for all j. However, we have assumed that $\langle S_i I_j \rangle \neq 0$, so this is not the case and $dy_k/dt \neq 0$.

Now assume $y_k = \langle S_j I_i \rangle$, which gives

$$\frac{\mathrm{d}y_k}{\mathrm{d}t} = \gamma \langle I_i \rangle + \sum_{m \neq j}^{N} T_{im} \langle S_i I_m \rangle - \sum_{m \neq j}^{N} T_{im} \frac{\langle I_j S_i \rangle \langle S_i I_m \rangle}{\langle S_i \rangle}.$$

Since $\langle I_j S_i \rangle / \langle S_i \rangle \leq 1$, the sum of the last two terms cannot be negative. Therefore, if $dy_k/dt = 0$ we have $\langle I_i \rangle = 0$. However, as has been shown by assuming $\langle I_i \rangle = 0$, this case is not possible. Therefore, $dy_k/dt \neq 0$. Therefore, if $\langle S_i I_j \rangle \neq 0$ for some pair (i, j) and $y \in H$, then this state cannot be invariant.

Now assume that $y_h = \langle I_i \rangle \in S'$ for some i, and consider $y_k = \langle S_j I_i \rangle \in S$. Since $\langle S_x I_y \rangle = 0$ for all (x, y), we have

$$\frac{\mathrm{d}y_k}{\mathrm{d}t} = \gamma \langle I_i \rangle.$$

Since $\langle I_i \rangle \in S'$, $dy_k/dt \neq 0$. Therefore, there are no invariant sets in H such that $y \neq 0$, and $y = 0$ is the largest positively invariant set in H.

This shows that properties 1-5 are satisfied for this model. Therefore, there exists a stable endemic steady-state above the epidemic threshold of the standard pair-based SIS model.

Appendix D Population-level individual-based QSD model

The node-level equations give detailed insight into the dynamics of individual nodes in the QSD, however the number of equations scales with N. To build approximations with a reduced number of equations, population-level models can be constructed for undirected networks. The rate of change in the expected number of infected individuals with a given degree, under the conditional distribution, is found by taking the sum over the probability that each node with this degree is infected

$$\sum_{i:k_i=k} \sum_{\alpha:\sigma_{\alpha i}=I} \frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} = \sum_{i:k_i=k} \left(\frac{\sum_j T_{ij} \langle S_i I_j \rangle - \gamma_i \langle I_i \rangle}{1 - P_1} + \frac{\langle I_i \rangle}{(1 - P_1)^2} \sum_j \gamma_j \langle I_j S \rangle \right)$$

The numerator in the first term on the right-hand side is the rate of change that an individual is infected. Taking the sum over all nodes with the same degree, this gives the rate of change in the expected number of infected individuals with that degree, which is given by Equation (B1). Taking the sum of $\langle I_i \rangle$ over all nodes with the same degree gives the expected number of infected nodes with that degree. Therefore, assuming

$$\langle I_i \rangle \approx \frac{[I_{k_i}]}{|C_{k_i}|} \quad , \quad T_{ij} = \bar{T}_{k_i k_j} \quad , \quad \gamma_i = \gamma_{k_i} \qquad (i \in \mathcal{V}, j \in \mathcal{N}_i),$$

where $[A_k]$ is the expected number of individuals with degree k in state A and $\bar{T}_{kl}$ is the rate of transmission from a degree l to a degree k node, we obtain

$$\sum_{i:k_i=k} \sum_{\alpha:\sigma_{\alpha i}=I} \frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} = \frac{\sum_{l \in \mathcal{M}} \bar{T}_{kl} [S_k I_l] - \gamma [I_k]}{1 - P_1} + \frac{[I_k]}{(1 - P_1)^2} \sum_j \gamma \langle I_j S \rangle, \tag{D1}$$

where $[A_k B_l]$ is the expected number of pairs between individuals of degree k and degree l, in states A and B respectively, and k_i is the degree of node i. Above, and throughout, all expected numbers are with respect to the standard probability measure P. Assuming that the states of individuals are independent, (D1) becomes

$$\sum_{i:k_i=k} \sum_{\alpha:\sigma_{\alpha i}=I} \frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} \approx \frac{\sum_{l \in \mathcal{M}} \bar{T}_{kl} |C_{k,l}| \frac{[S_k]}{|C_k|} \frac{[I_l]}{|C_l|} - \gamma [I_k]}{1 - \prod_j \langle S_j \rangle} + \frac{[I_k]}{(1 - \prod_j \langle S_j \rangle)^2} \sum_j \gamma \langle I_j \rangle \prod_{k \neq j} \langle S_k \rangle$$

where $|C_k|$ is the number of degree k nodes in the network and $|C_{k,l}|$ is the number of pairs between degree k and degree l nodes. This equation is not closed, since the final term and the denominators

depend on node-level quantities. However, from (B6) the node-level quantities can be approximated by assuming $\langle S_j \rangle = [S_k]/|C_k|$, where k is the degree of node j. Therefore

$$\prod_i \langle S_i \rangle \approx \prod_{l \in \mathcal{M}} \left(\frac{[S_l]}{|C_l|} \right)^{|C_l|},$$

and

$$\gamma \langle I_j \rangle \prod_{i \neq j} \langle S_i \rangle \approx \frac{[I_k]}{|C_k|} \left(\frac{[S_k]}{|C_k|} \right)^{|C_k|-1} \prod_{l \in \mathcal{M}: l \neq k} \left(\frac{[S_l]}{|C_l|} \right)^{|C_l|}, \tag{D2}$$

where k is the degree of node j. Multiplying Equation (D2) by the number of degree k nodes, $|C_k|$, we obtain the probability of a single degree k node being infected, which we denote $\tilde{P}(I_k = 1)$. Therefore, we obtain

$$\sum_{i:k_i=k} \sum_{\alpha:\sigma_{\alpha i}=I} \frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} \approx \frac{\sum_{l \in \mathcal{M}} \bar{T}_{kl} |C_{k,l}| \frac{[S_k]}{|C_k|} \frac{[I_l]}{|C_l|} - \gamma [I_k]}{(1 - \prod_l (\frac{[S_l]}{|C_l|})^{|C_l|})} + \frac{[I_k]}{(1 - \prod_l (\frac{[S_l]}{|C_l|})^{|C_l|})^2} \sum_{l \in \mathcal{M}} \gamma \tilde{P}(I_l = 1).$$

To find a steady state, we need to find vectors $\langle X \rangle^*$ and $\langle Y \rangle^*$ satisfying

$$0 = \frac{\sum_{l \in \mathcal{M}} \bar{T}_{kl} |C_{k,l}| \frac{[X_k]^*}{|C_k|} \frac{[Y_l]^*}{|C_l|} - \gamma [Y_k]^*}{(1 - \prod_l (\frac{[X_l]^*}{|C_l|})^{|C_l|})} + \frac{[Y_k]^*}{(1 - \prod_l (\frac{[X_l]^*}{|C_l|})^{|C_l|})^2} \sum_{l \in \mathcal{M}} \gamma \tilde{P}(Y_l = 1)^*$$

from which we can approximate the expected number of infected degree k individuals in the QSD by computing $[Y_k]^*/(1 - \prod_l (\frac{[X_l]^*}{|C_l|})^{|C_l|})$. We require $[Y_k]^* \in [0, |C_k|], [X_k]^* = |C_k| - [Y_k]^*$ for all i. Such a solution can be found by defining

$$\begin{aligned}
[\dot{Y_k}] =& \sum_{l \in \mathcal{M}} \bar{T}_{kl} |C_{k,l}| \frac{[X_k]}{|C_k|} \frac{[Y_l]}{|C_l|} - \gamma [Y_k] + \frac{[Y_k] \sum_{l \in \mathcal{M}} \gamma \tilde{P}(Y_l = 1)}{(1 - \prod_l (\frac{[X_l]}{|C_l|})^{|C_l|})} \\
[X_k] =& |C_k| - [Y_k] \\
\tilde{P}(Y_k = 1) =& |C_k| \frac{[Y_k]}{|C_k|} \left(\frac{[X_k]}{|C_k|} \right)^{|C_k|-1} \prod_{l \in \mathcal{M}: l \neq k} \left(\frac{[X_l]}{|C_l|} \right)^{|C_l|},
\end{aligned} \tag{D3}$$

and specifying that $[Y_k(0)] \in [0, |C_k|]$ for all k and calculating the steady-state. Any solution will be a valid solution, since Equation (D3) is bounded such that $[Y_k]^* \in [0, |C_k|]$ for all k (this can be shown using a method similar to Appendix A.2).

Appendix E Node-level pair-based QSD model

If we do not assume independence at the level of individuals, we need to find equations describing pair probabilities in the conditional distribution. We have

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t} \left(\rho(\Sigma_i(t) = I) \right) &= \sum_{\alpha:\sigma_{\alpha i}=I} \frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} = \frac{\sum_j T_{ij} \langle S_i I_j \rangle - \gamma_i \langle I_i \rangle}{1 - P_1} + \frac{\langle I_i \rangle}{(1 - P_1)^2} \sum_j \gamma_j \langle I_j S \rangle, \\
\frac{\mathrm{d}}{\mathrm{d}t} \left(\rho(\Sigma_i(t) = S, \Sigma_j = I) \right) &= \sum_{\substack{\alpha:\sigma_{\alpha i}=S, \\ \sigma_{\alpha j}=I}} \frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} = \frac{\sum_{k \in \mathcal{N}_j \setminus i} T_{jk} \langle S_i S_j I_k \rangle}{1 - P_1} - \frac{\sum_{k \in \mathcal{N}_i \setminus j} T_{ik} \langle I_k S_i I_j \rangle}{1 - P_1} \\
&\quad - \frac{(T_{ij} + \gamma_j) \langle S_i I_j \rangle}{1 - P_1} + \frac{\gamma_i \langle I_i I_j \rangle}{1 - P_1} + \frac{\langle S_i I_j \rangle}{(1 - P_1)^2} \sum_j \gamma_j \langle I_j S \rangle,
\end{aligned}$$

where $\langle A_i \rangle$ is shorthand for the marginal probability $P(\Sigma_i(t) = A)$ with $A \in \{S, I\}$, $\langle A_i B_j \rangle$ is shorthand for $P(\Sigma_i(t) = A, \Sigma_j(t) = B)$ with $A, B \in \{S, I\}$, $\langle A_i B_j C_k \rangle$ is shorthand for $P(\Sigma_i(t) = A, \Sigma_j(t) = B, \Sigma_k(t) = C)$ with $A, B, C \in \{S, I\}$, and $\langle I_j S \rangle$ is shorthand for $P(\Sigma_j = I, \Sigma_k = S$ for all $k \neq j)$. We can simplify this system by assuming statistical independence at the level of pairs.

As described in Appendix B, we approximate the triples in terms of pairs and individuals by assuming

$$\langle A_i B_j C_k\rangle \approx \frac{\langle A_i B_j\rangle\langle B_j C_k\rangle}{\langle B_j\rangle}.$$

Under this assumption, Equation (E1) becomes

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=I)\right) &= \frac{\sum_j T_{ij}\langle S_i I_j\rangle - \gamma_i\langle I_i\rangle}{1-P_1} + \frac{\langle I_i\rangle}{(1-P_1)^2}\sum_j \gamma_j\langle I_j S\rangle,\\
\frac{\mathrm{d}}{\mathrm{d}t}\left(\rho(\Sigma_i(t)=S,\Sigma_j=I)\right) &= \frac{\sum_{k\in\mathcal{N}_j\setminus i} T_{jk}\frac{\langle S_i S_j\rangle\langle S_j I_k\rangle}{\langle S_j\rangle}}{1-P_1} - \frac{\sum_{k\in\mathcal{N}_i\setminus j} T_{ik}\frac{\langle I_k S_i\rangle\langle S_i I_j\rangle}{\langle S_i\rangle}}{1-P_1}\\
&\quad -\frac{(T_{ij}+\gamma_j)\langle S_i I_j\rangle}{1-P_1} + \frac{\gamma_i\langle I_i I_j\rangle}{1-P_1} + \frac{\langle S_i I_j\rangle}{(1-P_1)^2}\sum_j \gamma_j\langle I_j S\rangle,
\end{aligned}$$

Note that $\langle S_i\rangle = 1-\langle I_i\rangle$, $\langle I_i I_j\rangle = \langle I_j\rangle - \langle S_i I_j\rangle$ and $\langle S_i S_j\rangle = \langle S_i\rangle - \langle S_i I_j\rangle$. Both $\langle I_j S\rangle$ and the ground state probability, P_1, are full system size, and therefore, following [12,45], a natural pair approximation for these are

$$\langle I_j S\rangle \approx \langle \widetilde{I_j S}\rangle = \frac{\prod_{x\in\mathcal{N}_j}\langle I_j S_x\rangle \prod_{y\neq j}\prod_{x\in\mathcal{N}_y: x<y, x\neq j}\langle S_y S_x\rangle}{\prod_{x\neq j}\langle S_x\rangle^{k_x-1}\langle Y_j\rangle^{k_j-1}}$$

and

$$P_1 \approx \langle\sigma_1\rangle = \prod_y \prod_{x\in\mathcal{N}_y: x<y}\frac{\langle S_y S_x\rangle}{\langle S_y\rangle^{n_y-1}}.$$

In the QSD, both the pair level and individual level conditional probabilities are in a steady-state, so both equations in Equation (E1) are equal to zero. Therefore, to find the approximation to the QSD under the pair level independence assumption, we need to find vectors $\langle X^*\rangle$, $\langle Y^*\rangle$, and matrices $\langle XX^*\rangle$,$\langle XY^*\rangle$, and $\langle YY^*\rangle$ satisfying,

$$\begin{aligned}
0 &= \frac{\sum_j T_{ij}\langle X_i Y_j\rangle^* - \gamma_i\langle Y_i\rangle^*}{1-\langle\sigma_1\rangle} + \frac{\langle Y_i\rangle^*}{(1-\langle\sigma_1\rangle)^2}\sum_j \gamma_j\langle\widetilde{Y_j X}\rangle^*,\\
0 &= \frac{\sum_{k\in\mathcal{N}_j\setminus i} T_{jk}\frac{\langle X_i X_j\rangle^*\langle X_j Y_k\rangle^*}{\langle X_j\rangle^*}}{1-\langle\sigma_1\rangle} - \frac{\sum_{k\in\mathcal{N}_i\setminus j} T_{ik}\frac{\langle Y_k X_i\rangle^*\langle X_i Y_j\rangle^*}{\langle X_i\rangle^*}}{1-\langle\sigma_1\rangle}\\
&\quad -\frac{(T_{ij}+\gamma_j)\langle X_i Y_j\rangle^*}{1-\langle\sigma_1\rangle} + \frac{\gamma_i\langle Y_i Y_j\rangle^*}{1-\langle\sigma_1\rangle} + \frac{\langle X_i Y_j\rangle^*}{(1-\langle\sigma_1\rangle)^2}\sum_j \gamma_j\langle\widetilde{Y_j X}\rangle^*,
\end{aligned}$$

which, once solved, can be used to find the probability that i is infected in the QSD by computing $\langle Y_i\rangle^*/(1-\langle\sigma_1\rangle^*)$. However, we require solutions $\langle Y_i\rangle^*$ and $\langle X_i Y_j\rangle^* \in [0,1]$ which satisfy $\langle X_i\rangle^* = 1-\langle Y_i\rangle^*$ for all i and $\langle X_i X_j\rangle = \langle X_i\rangle - \langle X_i Y_j\rangle$, and $\langle Y_i Y_j\rangle = \langle Y_j\rangle - \langle X_i Y_j\rangle$ for all i,j in order to be valid solutions to our original problem.

By calculating the steady-state of the system,

$$\begin{aligned}
\langle\dot{Y_i}\rangle &= \sum_j T_{ij}\langle X_i Y_j\rangle - \gamma_i\langle Y_i\rangle + \frac{\langle Y_i\rangle\sum_j \gamma_j\langle\widetilde{Y_j X}\rangle}{1-\langle\sigma_1\rangle},\\
\langle\dot{X_i Y_j}\rangle &= \sum_{k\in\mathcal{N}_j\setminus i} T_{jk}\frac{\langle X_i X_j\rangle\langle X_j Y_k\rangle}{\langle X_j\rangle} - \sum_{k\in\mathcal{N}_i\setminus j} T_{ik}\frac{\langle Y_k X_i\rangle\langle X_i Y_j\rangle}{\langle X_i\rangle}\\
&\quad -(T_{ij}+\gamma_j)\langle X_i Y_j\rangle + \gamma_i\langle Y_i Y_j\rangle + \frac{\langle X_i Y_j\rangle\sum_j \gamma_j\langle\widetilde{Y_j X}\rangle}{1-\langle\sigma_1\rangle},\\
\langle X_i\rangle &= 1-\langle Y_i\rangle,\\
\langle X_i X_j\rangle &= \langle X_i\rangle - \langle X_i Y_j\rangle,\\
\langle Y_i Y_j\rangle &= \langle Y_i\rangle - \langle Y_i X_j\rangle,
\end{aligned}$$

where

$$\widetilde{\langle Y_j X\rangle} = \frac{\prod_{x\in\mathcal{N}_j}\langle Y_j X_x\rangle \prod_{y\neq j}\prod_{x\in\mathcal{N}_y:x<y,x\neq j}\langle X_y X_x\rangle}{\prod_{x\neq j}\langle X_x\rangle^{k_x-1}\langle Y_j\rangle^{k_j-1}}$$

and

$$\langle\sigma_1\rangle = \prod_y \prod_{x\in\mathcal{N}_y:x<y} \frac{\langle X_y X_x\rangle}{\langle X_y\rangle^{n_y-1}}.$$

we can approximate the probability that i is infected in the QSD by computing $\lim_{t\to\infty}\langle Y_i(t)\rangle^*/(1-\langle\sigma_0(t)\rangle^*)$.

Appendix F Population-level pair-based QSD model

To obtain a population-level pair-based model, we sum over nodes with the same degree (and pairs of nodes with same pair of degrees); i.e.

$$\begin{aligned}
\sum_{i:k_i=k}\sum_{\alpha:\sigma_{\alpha i}=I}\frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} &= \frac{\tau\sum_{l\in\mathcal{M}}[S_kI_l]-\gamma[I_k]}{1-P_1}+\frac{[I_k]}{(1-P_1)^2}\sum_j\gamma\langle I_jS\rangle\\
\sum_{\substack{i,j:k_i=k,\\ k_j=l}}\sum_{\substack{\alpha:\sigma_{\alpha i}=S,\\ \sigma_{\alpha j}=I}}\frac{\mathrm{d}\rho_\alpha}{\mathrm{d}t} &= \frac{\tau\sum_{m\in\mathcal{M}}[S_kS_lI_m]-\tau\sum_{m\in\mathcal{M}}[I_mS_kI_l]-\tau[S_kI_l]+\gamma[I_kI_l]-\gamma[S_kI_l]}{1-P_1}\\
&\quad+\frac{[S_kI_l]}{(1-P_1)^2}\sum_j\gamma\langle I_jS\rangle,
\end{aligned}$$

where $[A_kB_lC_h]$ is the expected number of triples between degree k, degree l and degree h individuals in states A, B and C respectively.

As described in Appendix B, we can express the triple terms as

$$[A_lS_kI_m]\approx\frac{k-1}{k}\frac{[A_lS_k][S_kI_m]}{[S_k]}, \tag{F2}$$

We can set equations (F1) to zero and use the approximation (F2) to find equations describing the QSD.

A solution to the resulting system can be found by finding an steady-state of

$$\begin{aligned}
\dot{[Y_k]} =& -\gamma[Y_k]+\tau\sum_{l\in\mathcal{M}}[X_kY_l]+\frac{[Y_k]\sum_{l\in\mathcal{M}}\gamma\tilde{P}(Y_l=1)}{1-\langle\sigma_1\rangle}\\
\dot{[X_kY_l]} =& \tau(\sum_{m\in\mathcal{M}}\frac{l-1}{l}\frac{[X_kX_l][X_lY_m]}{[X_l]}-\sum_{m\in\mathcal{M}}\frac{k-1}{k}\frac{[Y_mX_k][X_kY_l]}{[X_k]}\\
&-[X_kY_l])+\gamma([Y_kY_l]-[X_kY_l])+\frac{[X_kY_l]\sum_{l\in\mathcal{M}}\gamma\tilde{P}(Y_l=1)}{1-\langle\sigma_1\rangle}\\
\dot{[Y_kY_l]} =& \tau(\sum_{m\in\mathcal{M}}\frac{k-1}{k}\frac{[Y_mX_k][X_kY_l]}{[X_k]}+\sum_{m\in\mathcal{M}}\frac{l-1}{l}\frac{[Y_kX_l][X_lY_m]}{[X_l]})\\
&+\tau([X_kY_l]+[Y_kX_l]-2\gamma[Y_kY_l]+\frac{[Y_kY_l]\sum_{l\in\mathcal{M}}\gamma\tilde{P}(Y_l=1)}{1-\langle\sigma_1\rangle}\\
[X_k] =& |C_k|-[Y_k]\\
[X_kX_l] =& |C_{k,l}|-[Y_kY_l]-[X_kY_l]-[X_lY_k],
\end{aligned} \tag{F3}$$

where $\tilde{P}(Y_l=1)=|C_l|\langle Y_iX\rangle$ for some i with $k_i=l$. Here

$$\langle Y_iX\rangle=\frac{\prod_{x,y\neq i}G_{xy}\langle X_xX_y\rangle\prod_x G_{ix}\langle Y_iX_x\rangle}{\prod_{x\neq i}\langle X_x\rangle^{k_x-1}\langle Y_i\rangle^{k_i-1}},$$

which requires node-level terms. We can approximate this by population-level quantities using

$$\langle S_i \rangle \approx \frac{[S_{k_i}]}{|C_{k_i}|}, \tag{F4}$$

and

$$\langle S_i S_j \rangle \approx \frac{[S_{k_i} S_{k_j}]}{|C_{k_i,k_j}|}, \tag{F5}$$

based on the discussion in Appendix B. This gives

$$\langle Y_j X \rangle \approx \frac{\prod\limits_{k \neq k_j} \prod\limits_{l \leq k : l \neq k_j} \left(\frac{[X_k X_l]}{|C_{k,l}|} \right)^{|C_{k,l}|} \left(\frac{[Y_{k_j} X_k]}{|C_{k_j,k}|} \right)^{\frac{|C_{k_j,k}|}{|C_{k_j}|}} \left(\frac{[X_{k_j} X_k]}{|C_{k_j,k}|} \right)^{|C_{k_j,k}| - \frac{|C_{k_j,k}|}{|C_{k_j}|}}}{\prod\limits_{k \neq k_j} \left(\frac{[X_k]}{|C_k|} \right)^{|C_k|(k-1)} \left(\frac{[Y_{k_j}]}{|C_{k_j}|} \right)^{(k_j - 1)} \left(\frac{[X_{k_j}]}{|C_{k_j}|} \right)^{(|C_{k_j}|-1)(k_j-1)}}. \tag{F6}$$

To approximate the ground state recall that in the previous section we have shown that a natural approximation to the ground state probability under the assumption of pair level independence is

$$\langle \sigma_1 \rangle \approx \prod_i \prod_{j<i} \frac{G_{ij} \langle X_i X_j \rangle}{\langle X_i \rangle^{n_i - 1}}.$$

Using equations (F4) and (F5) we can approximate this in terms of population level quantities, which yields

$$\langle \sigma_1 \rangle \approx \prod_k \prod_{l \leq k} \left(\frac{\left(\frac{[X_k X_l]}{|C_{k,l}|} \right)^{|C_{k,l}|}}{\left(\frac{[X_k]}{|C_k|} \right)^{|C_k|(k-1)}} \right) \tag{F7}$$

By substituting equations (F7) and (F6) into Equation (F3) we obtain a closed system of equations.

Acknowledgements

CO and KS acknowledge support from EPSRC grant (EP/N014499/1). The authors would like to thank Ian Smith for use of the ARC Condor high throughput computing system at the University of Liverpool `http://condor.liv.ac.uk/`, which significantly sped up simulation of the stochastic models.

Author contributions

CO, KS and RW created the project, performed the analysis and wrote the manuscript. JM created the project and perfomed the analysis. AL created the project.

Competing interests

The authors have no competing interests to declare.

Data and materials

Matlab code for solving the models will be published online with the manuscript. Python code will be added to the Epidemics on Networks package.

References

1. L. J. S. Allen and A. M. Burgin. Comparison of deterministic and stochastic SIS and SIR models in discrete time. Mathematical Biosciences, 163(1):1 – 33, 2000.
2. H. Andersson and T. Britton. Stochastic epidemics in dynamic populations: quasi-stationarity and extinction. Journal of Mathematical Biology, 41(6):559–580, 2000.
3. J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. The maximum number of infected individuals in SIS epidemic models: Computational techniques and quasi-stationary distributions. Journal of Computational and Applied Mathematics, 233(10):2563–2574, 2010.
4. J. R. Artalejo, A. Economou, and M. J. Lopez-Herrero. Stochastic epidemic models with random environment: quasi-stationarity, extinction and final size. Journal of Mathematical Biology, 67(4):799–831, 2013.
5. S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, and D. U. Hwang. Complex networks: Structure and dynamics. Physics Reports, 424(4):175 – 308, 2006.
6. S. Dambrine and M. Moreau. Note on the stochastic theory of a self-catalytic chemical reaction. i. Physica A: Statistical Mechanics and its Applications, 106(3):559–573, 1981.
7. S. Dambrine and M. Moreau. Note on the stochastic theory of a self-catalytic chemical reaction. ii. Physica A: Statistical Mechanics and its Applications, 106(3):574–588, 1981.
8. J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing continuous-time finite markov chains. Journal of Applied Probability, 4(1):192–196, 1967.
9. R. Dickman and R. Vidigal. Quasi-stationary distributions for stochastic processes with an absorbing state. Journal of Physics A: Mathematical and General, 35(5):1147, 2002.
10. K. T. D. Eames and M. J. Keeling. Modeling dynamic and network heterogeneities in the spread of sexually transmitted diseases. Proceedings of the National Academy of Sciences, 99(20):13330–13335, 2002.
11. S. C. Ferreira, C. Castellano, and R. Pastor-Satorras. Epidemic thresholds of the susceptible-infected-susceptible model on networks: A comparison of numerical and theoretical results. Physical Review E, 86(4):041125, 2012.
12. M. Frasca and K. J. Sharkey. Discrete-time moment closure models for epidemic spreading in populations of interacting individuals. Journal of Theoretical Biology, 399:13–21, 2016.
13. C. Hadjichrysanthou and K. J. Sharkey. Epidemic control analysis: designing targeted intervention strategies against epidemics propagated on contact networks. Journal of Theoretical Biology, 365:84–95, 2015.
14. T. J. Hagenaars, C. A. Donnelly, and N. M. Ferguson. Spatial heterogeneity and the persistence of infectious diseases. Journal of Theoretical Biology, 229(3):349–359, 2004.
15. T. E. Harris. Contact interactions on a lattice. The Annals of Probability, pages 969–988, 1974.
16. C. S. Holling. Resilience and stability of ecological systems. Annual Review of Ecology and Systematics, 4(1):1–23, 1973.
17. M. J. Keeling. The effects of local spatial structure on epidemiological invasions. Proceedings of the Royal Society B: Biological Sciences, 266(1421):859–867, 1999.
18. M. J. Keeling and K. T. D. Eames. Networks and epidemic models. Journal of the Royal Society Interface, 2(4):295–307, 2005.
19. J. O. Kephart, S. R. White, and D. M. Chess. Computers and epidemiology. IEEE Spectrum, 30(5):20–26, 1993.
20. I. Z. Kiss, J. C. Miller, and P. L. Simon. Mathematics of epidemics on networks. Cham: Springer, 2017.
21. D. R. Klein. The introduction, increase, and crash of reindeer on St. Matthew island. The Journal of Wildlife Management, pages 350–367, 1968.
22. R. J. Kryscio and C. Lefevre. On the extinction of the SIS stochastic logistic epidemic, pages 213–228. Statistical Methods in Computer Security. 2004.
23. A. Lajmanovich and J. A. Yorke. A deterministic model for gonorrhea in a nonhomogeneous population. Mathematical Biosciences, 28(3-4):221–236, 1976.
24. T. M. Liggett. Interacting particle systems, volume 276. Springer Science & Business Media, 2012.
25. A. S. Mata and S. C. Ferreira. Pair quenched mean-field theory for the susceptible-infected-susceptible model on complex networks. Europhysics Letters, 103(4):48003, 2013.
26. L. D. Mech. The wolves of isle royale. 1966.
27. W. H. Murray. The application of epidemiology to computer viruses. Computers & Security, 7(2):139–145, 1988.
28. I. Nåsell. The quasi-stationary distribution of the closed endemic SIS model. Advances in Applied Probability, 28(3):895–932, 1996.
29. I. Nåsell. On the quasi-stationary distribution of the stochastic logistic epidemic. Mathematical Biosciences, 156(1-2):21–40, 1999.
30. I. Nåsell. On the time to extinction in recurrent epidemics. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 61(2):309–330, 1999.
31. I. Oppenheim, K. E. Shuler, and G. H. Weiss. Stochastic theory of nonlinear rate processes with multiple stationary states. Physica A: Statistical Mechanics and its Applications, 88(2):191–214, 1977.
32. C. E. Overton, M. Broom, C. Hadjichrysanthou, and K. J. Sharkey. Methods for approximating stochastic evolutionary dynamics on graphs. Journal of Theoretical Biology, 468:45–59, 2019.
33. A. G. Pakes. Limit theorems for the population size of a birth and death process allowing catastrophes. Journal of Mathematical Biology, 25(3):307–325, 1987.
34. R. Parshani, S. Carmi, and S. Havlin. Epidemic threshold for the susceptible-infectious-susceptible model on random networks. Physical Review Letters, 104(25):258701, 2010.
35. R. W. Parsons and P. K. Pollett. Quasistationary distributions for autocatalytic reactions. Journal of Statistical Physics, 46(1-2):249–254, 1987.
36. R. Pastor-Satorras, C. Castellano, P. Van Mieghem, and A. Vespignani. Epidemic processes in complex networks. Reviews of Modern Physics, 87(3):925, 2015.
37. R. Pastor-Satorras and A. Vespignani. Epidemic spreading in scale-free networks. Physical Review Letters, 86(14):3200, 2001.
38. P. K. Pollett. On the problem of evaluating quasistationary distributions for open reaction schemes. Journal of Statistical Physics, 53(5-6):1207–1215, 1988.

39. P. K. Pollett. The determination of quasistationary distributions directly from the transition rates of an absorbing markov chain. Mathematical and Computer Modelling, 22(10-12):279–287, 1995.
40. P. K. Pollett and S. Kumar. On the long-term behaviour of a population that is subject to large-scale mortality or emigration. In Proceedings of the 8th National Conference of the Australian Society for Operations Research, volume 196, page 207, 1987.
41. K. Rock, S. Brand, J. Moir, and M. J. Keeling. Dynamics of infectious diseases. Reports on Progress in Physics, 77(2):026602, 2014.
42. T. Rogers. Maximum-entropy moment-closure for stochastic systems on networks. Journal of Statistical Mechanics: Theory and Experiment, 2011(05):P05007, 2011.
43. V. B. Scheffer. The rise and fall of a reindeer herd. The Scientific Monthly, 73(6):356–362, 1951.
44. K. J. Sharkey. Deterministic epidemic models on contact networks: correlations and unbiological terms. Theoretical Population Biology, 79(4):115–129, 2011.
45. K. J. Sharkey, I. Z. Kiss, R. R. Wilkinson, and P. L. Simon. Exact equations for SIR epidemics on tree graphs. Bulletin of Mathematical Biology, 77(4):614–645, 2015.
46. P. Van Mieghem. The N-intertwined SIS epidemic network model. Computing, 93(2-4):147–169, 2011.
47. P. Van Mieghem, J. Omic, and R. Kooij. Virus spread in networks. IEEE/ACM Transactions on Networking, 17(1):1–14, 2009.
48. Y. Wang, D. Chakrabarti, C. Wang, and C. Faloutsos. Epidemic spreading in real networks: An eigenvalue viewpoint. In 22nd International Symposium on Reliable Distributed Systems, 2003. Proceedings., pages 25–34. IEEE, 2003.
49. J. C. Wierman and D. J. Marchette. Modeling computer virus prevalence with a susceptible-infected-susceptible model with reintroduction. Computational Statistics & Data Analysis, 45(1):3–23, 2004.
50. R. R. Wilkinson and K. J. Sharkey. An exact relationship between invasion probability and endemic prevalence for markovian SIS dynamics on networks. PloS One, 8(7):e69028, 2013.
51. W. W. Zachary. An information flow model for conflict and fission in small groups. Journal of Anthropological Research, 33(4):452–473, 1977.