Journal of Theoretical Biology 534 (2022) 110974

Contents lists available at ScienceDirect

Journal of Theoretical Biology

journal homepage: www.elsevier.com/locate/yjtbi

Approximating steady state distributions for household structured epidemic models

Alex Holmes [a,b,*], Mike Tildesley [a], Louise Dyson [a]

[a] The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, School of Life Sciences and Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom

[b] Mathematics for Real World Systems Centre for Doctoral Training, Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom

ARTICLE INFO

Article history:
Received 27 May 2021
Revised 13 November 2021
Accepted 16 November 2021
Available online 28 November 2021

Keywords:
Epidemiology
Stochastic household modelling
Approximation
Fokker–Planck
Master equation

ABSTRACT

Household-structured infectious disease models consider the increased transmission potential between individuals of the same household when compared with two individuals in different households. Accounting for these heterogeneities in transmission enables control measures to be more effectively planned. Ideally, pre-control data may be used to fit such a household-structured model at an endemic steady state, before making dynamic forward-predictions under different proposed strategies. However, this requires the accurate calculation of the steady states for the full dynamic model. We observe that steady state SIS dynamics with household structure cannot necessarily be described by the master equation for a single household, instead requiring consideration of the full system. However, solving the full system of equations becomes increasingly computationally intensive, particularly for higher-dimensional models. We compare two approximations to the full system: the single household master equation; and a proposed alternative method, using the Fokker–Planck equation. Moment closure is another commonly used method, but for more complicated systems, the equations quickly become unwieldy and very difficult to derive. In comparison, using the master equation for a single household is easily implementable, however it can be quite inaccurate. In this paper we compare these methods in terms of accuracy and ease of implementation. We find that there are regions of parameter space in which each method outperforms the other, and that these regions of parameter space can be characterised by the infection prevalence, or by the correlation between household states.

1. Introduction

Transmission of an infectious disease is often seen to be greater within a household than between those from different households (Kinyanjui et al., 2016). Household structured models take these population heterogeneities into account, and can be used to inform different potential control policies. In particular, they explicitly include (at least) two different forms of infection (Black et al., 2013): infection from within the same household as the infected individual and infection from outside the same household as an infected individual. The first of these typically corresponds to a higher rate of infection, but between a smaller number of contacts, while the latter often corresponds to a lower rate of infection, but affecting a much larger number of individuals (Hilton and Keeling, 2019; Ball et al., 1997).

Household-structured infectious disease models have been used extensively in the literature, particularly in the case of pandemic influenza (Fraser et al., 2011; Wu et al., 2006). Benefits of these models include better representation of the population under consideration (Frank and Neal, 2002) and the ability to incorporate different intervention strategies. This is useful as control policies are often targeted at the level of households, for both practical and structural reasons (Pellis et al., 2009). For example, the current eradication strategy for yaws (a neglected tropical disease) includes treating infected individuals and their contacts, which can be easily modelled using a household structure (Holmes et al., 2020). Similarly, there are a large number of studies investigating household-based vaccination strategies for pandemic influenza (House and Keeling, 2009; Black et al., 2013).

In spite of the widespread use of household models, there are some subtleties in their use. A natural way to fit a household model to data is to assume the system is at steady state, which enables

[*] Corresponding author at: The Zeeman Institute for Systems Biology & Infectious Disease Epidemiology Research, School of Life Sciences and Mathematics Institute, University of Warwick, Coventry CV4 7AL, United Kingdom.

E-mail address: alexander.holmes@Warwick.ac.uk (A. Holmes).

https://doi.org/10.1016/j.jtbi.2021.110974
0022-5193/© 2021 The Authors. Published by Elsevier Ltd.

the fitting of a constant force of infection experienced outside the household. However, when subsequently predicting the response to potential changes in strategy, it is necessary to then link this external force of infection to infections in other households. The different approximations involved can lead to discrepancies, however, between the dynamic forward projection and the steady state. We describe in more detail below how these discrepancies arise.

Compartmental stochastic models in epidemiology can typically be expressed as a set of master equations. That is, a set of ordinary differential equations (ODEs) that describe how the probability of being in each state varies with time (Keeling and Ross, 2008). This gives a set of linear ODEs which can be written in the form $\dot{\mathbf{p}} = A\mathbf{p}$, where A is the state transition matrix. This then has solution $\mathbf{p}(t) = \mathbf{p}(0)e^{At}$, or we can calculate the steady state distribution by calculating the eigenvector corresponding to the 0 eigenvalue of A (Keeling and Ross, 2008) or by solving the system of linear equations $A\mathbf{p} = 0$.

For example, consider the stochastic steady state SIA model in Dyson et al. (2017). This is a household-structured model in which each individual in a household can either be susceptible to infection, S, infected and infectious, I, or asymptomatic and not infectious, A. An individual can be infected from someone within the same household at rate β, or they can be infected from someone outside their household with external force of infection, ε. Data were used to parameterise the model at steady state, allowing the effects of other households on the external force of infection to be considered implicitly, by taking ε to be constant. An infectious individual can recover at rate δ, or they can become asymptomatic at rate λ. Finally, asymptomatic individuals can recover at rate δ, or their symptoms can recur causing them to become infectious again (without further exposure to an infectious individual), which occurs at rate ρ. These dynamics are summarised in figure 1. Letting $P^N_{S,I,A}$ denote the probability of a household of size N containing S, I and A susceptible, infectious and asymptomatic individuals respectively, the master equation describing the dynamics of a single household of size N for this system is given in Eq. 1.

$$\frac{dP^N_{S,I,A}}{dt} = -\left[\left(\varepsilon + \frac{\beta I}{N-1}\right)S + \delta(A+I) + \rho A + \lambda I\right]P^N_{S,I,A} + \left(\varepsilon + \frac{\beta(I-1)}{N-1}\right)(S+1)P^N_{S+1,I-1,A}$$
$$+\delta(A+1)P^N_{S-1,I,A+1} + \delta(I+1)P^N_{S-1,I+1,A} + \rho(A+1)P^N_{S,I-1,A+1} + \lambda(I+1)P^N_{S,I+1,A-1}, \quad (1)$$

subject to the following constraints on S, I, A and N:

$$S + I + A = N,$$

$$S \geqslant 0, I \geqslant 0, A \geqslant 0.$$

How can this system be simulated forwards in time? It is necessary to make an assumption that links the rate of external infection to the infection within households. One approach is to assume the external force of infection, ε, takes the form

$$\varepsilon = \alpha \frac{\sum_{i=1}^{M} I(i)}{\sum_{i=1}^{M} N(i)}, \quad (2)$$

where α is the between–household rate of infection that would achieve the same force of infection at steady state, M is the number of households, $I(i)$ denotes the number of infectious individuals in household i, and $N(i)$ denotes the size of household i. However, if we simulate the system forward to steady state using this new between household rate of infection, α, calculated by rearranging Eq. 2 using the steady state prevalence (calculated by solving the master equation with constant ε, Eq. 1), we find that the simulated steady state is lower than expected. This is shown in Fig. 2, which compares: the numerical solution to the system of master equations (Eq. 1) for the number of households in state (S, I, A), but with ε given by Eq. 2 (solid blue line); the expected steady state (using a constant ε) calculated by solving Eq. 1 (red dashed line); and an ensemble average of trajectories simulated using the Gillespie algorithm for the system in Fig. 1 with ε given by Eq. 2 (solid black line). Both the red dashed line and the solid blue line represent solutions to Eq. 1 (with the red-dashed line representing the steady state). However, the red line is calculated using a constant ε, while the blue line used ε given by Eq. 2. As such, this means the external force of infection is time varying (and the corresponding value of α was selected to ensure the steady states match).

This suggests that the approximation taken by the master equations for the number of households in state (S, I, A) breaks down when there is between–household interaction with a finite number of households. In the limit of a large number of households, the approximation is valid (Ross et al., 2010), but not in the case of a

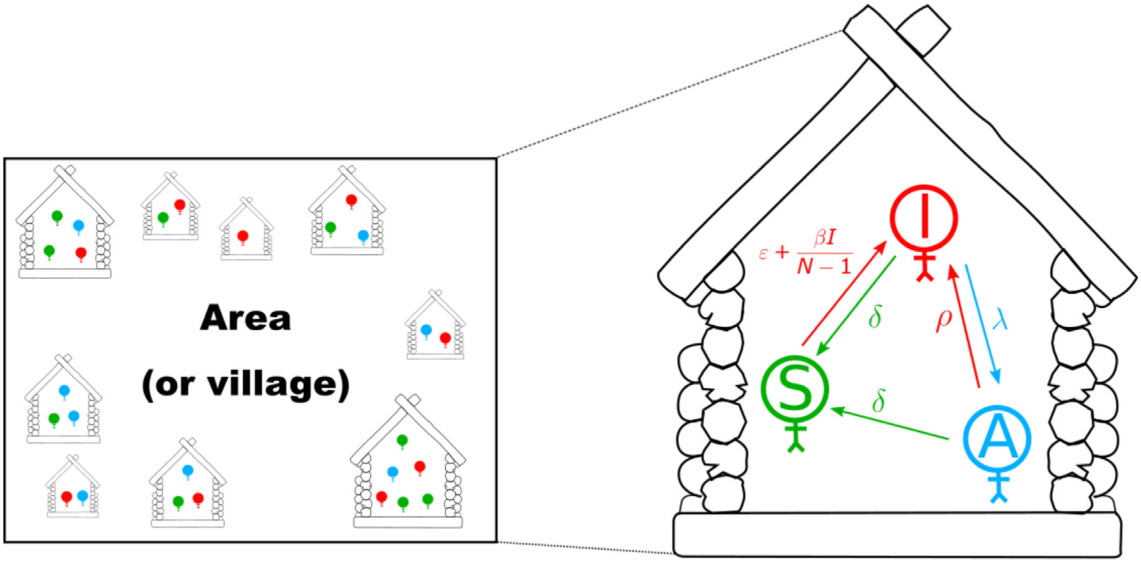


Fig. 1. Visual depiction of the steady-state household-structured SIA model.

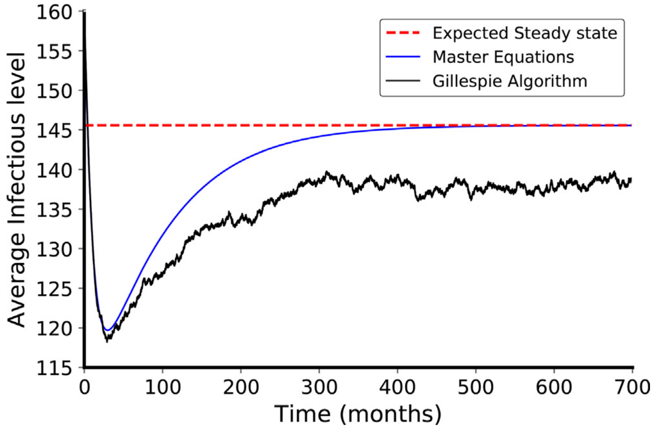


Fig. 2. Plot showing the steady state obtained with a constant ε, the average trajectory obtained by solving the master equations with Eq. 2 and an ensemble average of trajectories simulated using the Gillespie algorithm with ε given by Eq. 2. The population consists of 1500 households with household size distribution detailed in Dyson et al. (2017). Other parameter values used are $(\beta, \delta, \lambda, \rho, \varepsilon, \alpha) = (0.0516, 0.0513, 0.185, 0.0165, 0.004, 0.168)$..

finite population size. Instead, we need a set of master equations that captures the full range of interactions between households. Unfortunately, this significantly increases the number of possible system states, and this quickly becomes impractical unless the number of households is very small (Ball and Lyne, 2001). Previous work has also looked at approximating household models. In Black et al. (2014), the authors investigate infectious disease processes on clumped population structures (e.g. households). They use a branching process approximation to investigate the start of the exponential growth phase, and then apply a diffusion approximation to investigate the variance during the early asymptotic phase of the infectious process. However, the authors do not address the question of how accurate these approximations are at a finite number of clumps (or households), but rather just consider the situation in which the number of clumps tends to infinity.

In this paper we consider different approximations to a similar system (SIS dynamics, starting with households of size 2), to determine the steady state distributions of households in such a population. We investigate which approximations are more accurate in different regions of parameter space. We then extend this to larger household sizes, determine how the accuracy of approximations is affected by the population structure and in particular what happens when we have the same population size partitioned into fewer, but larger, households.

Depending on the parameter values used, the correlation between household states can vary substantially. We investigate the relationship between this correlation and the accuracy of different approximations. Finally, we return to our original question — how accurate is each approximation when used to convert a steady state force of infection into a between household rate of infection?.

2. Methods

2.1. Master equation

We begin by considering the simplest such system — a population of households of size 2 undergoing SIS dynamics. This can be expressed as a set of master equations describing the household states (s, i), where s denotes the number of susceptible individuals and i denotes the number of infectious individuals. For a household of size h, the set of permissible household states under this model can be defined as

Table 1
State transitions for SIS household model.

Event	Symbol	State Transition	Rate
Infection	$a_{m,n,k}$	$(m,n,k) \to (m-1,n+1,k)$	$\frac{\alpha}{N}((n+2k)m)$
Recovery	$b_{m,n,k}$	$(m,n,k) \to (m+1,n-1,k)$	γn
Infection	$c_{m,n,k}$	$(m,n,k) \to (m,n-1,k+1)$	$(\frac{\alpha}{2N}(n+2k)+\beta)n$
Recovery	$d_{m,n,k}$	$(m,n,k) \to (m,n+1,k-1)$	$2\gamma k$

$$\{(s,i) \in \mathbb{N}^2 : s+i=h\}.$$

Similarly, the set of system states can be expressed in terms of these household states, assuming a constant population of N households each of size 2 (so the total population size is $2N$). Let m denote the number of households in household state $(2,0)$ (the number of households with 2 susceptible individuals) and let n denote the number of households in household state $(1,1)$ (the number of households with 1 susceptible individual and 1 infectious individual). Then $N-m-n$ denotes the number of households in household state $(0,2)$ (the number of households consisting of 2 infectious individuals and no susceptible individuals), denoted by k. A system state is then given by the tuple (m,n,k) and the full set of permissible system states is expressed as

$$\{(m,n,k) \in \mathbb{N}^3 : m+n+k=N\}.$$

We consider the stochastic compartmental SIS model incorporating the household structure as described above, with events and corresponding rates summarised in Table 1. In this model, α denotes the rate of between–household transmission, β denotes the rate of within–household transmission, and γ denotes the recovery rate. The possible transitions are summarised in Fig. 3.

Let $p_{m,n,k}(t)$ denote the probability that the system is in a state containing m, n and k households in their respective states. We define the bivariate step operators, $\mathbb{E}^{a,b}$ (Van Kampen, 2007) such that

$$\mathbb{E}^{a,b} f(m,n) = f(m+a, n+b).$$

Substituting in $k = N-m-n$ to eliminate k, we can use the bivariate step operator to write the master equation as

$$\begin{aligned}\frac{\partial p_{m,n}}{\partial t} &= (\mathbb{E}^{1,-1}-1)a_{m,n}p_{m,n} + (\mathbb{E}^{-1,1}-1)b_{m,n}p_{m,n} \\ &\quad + (\mathbb{E}^{0,1}-1)c_{m,n}p_{m,n} + (\mathbb{E}^{0,-1}-1)d_{m,n}p_{m,n},\end{aligned} \tag{3}$$

with the following rates:

$$\begin{aligned}a_{m,n} &= \tfrac{\alpha}{N}(n+2(N-m-n))m \\ b_{m,n} &= \gamma n \\ c_{m,n} &= (\tfrac{\alpha}{2N}(n+2(N-m-n))+\beta)n \\ d_{m,n} &= 2\gamma(N-m-n).\end{aligned}$$

Numerically solving this system of equations at steady state gives us the probability of occupying each system state at steady state. From this, quantities of interest such as the expected number of infectious individuals can be calculated. However, due to the large state space, it is not feasible to solve this system for large population sizes.

2.2. Quasi-steady state

While the above system of equations are difficult to solve for larger populations, it can be solved numerically when dealing with small population sizes. However, due to the absorbing state at $(N, 0, 0)$ (N households in household state $(2, 0)$, so no infection in the population), every system will eventually reach a state of

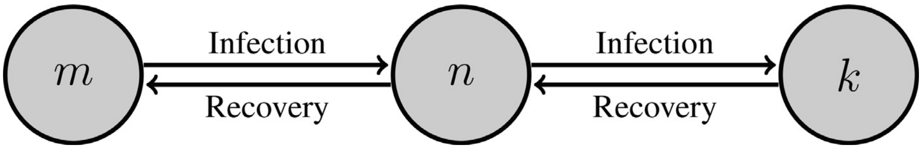


Fig. 3. Figure showing possible transitions between each individual in a population of size N undergoing SIS dynamics.

no infection (Dickman and Vidigal, Jan 2002). Thus, the 0 eigenvalue corresponds to the disease-free eigenvector and we obtain a singular matrix. During this time, a system may spend a long period of time at a quasi-steady state. This quasi-steady state is the steady state of interest. To find the quasi-steady state, the system of master equations are solved conditioned on non-extinction (Mubayi et al., 2019). That is, we consider the sub-matrix formed by removing the rows and columns corresponding to the disease-free states. The eigenvector corresponding to the smallest eigenvalue of this sub-matrix then represents the quasi-steady state distribution for this system. In this way, the system can be studied at low population sizes.

As mentioned before, this system of equations is too large for more than a small number of households to find an exact steady state solution. As such, methods to approximate this system which would allow the steady state distribution to be approximated. One commonly used approximation method that works well for simple systems is moment closure (Keeling, 2000). However, for more complicated systems, the system of moment closure equations can become very difficult to derive (Kuehn, 2016), and so here we restrict ourselves to methods that can be generalised more easily.

Two different methods for approximating the steady state distributions are compared here, which we will refer to as the single household (SHH) approximation and the Fokker–Planck peak (FPP) approximation. To assess which method approximates the true steady state distribution more accurately, we calculate the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951) of each approximation from the true steady state distribution. Letting Q denote the distribution obtained from the approximation and P the distribution obtained from solving the master equation, we denote the KL divergence of Q from P by $D_{KL}(Q||P)$. For the purposes of calculating the KL divergence, we take the steady state distribution to be the proportion of households in each household state.

2.2.1. Single household (SHH) approximation

The first approximation is derived by looking at stochastic SIS dynamics in a household of size 2. This differs from the previous system as these equations only look at transitions between household states, rather than system states. On the other hand, the corresponding fully system of master equations give us a probability of finding each steady state household distribution.

To write down this system, let p_n denote the probability of finding n infectious individuals, and $N - n$ susceptible individuals in a household of size N. There are two events that can then happen: an infectious individual becomes susceptible, or a susceptible individual becomes infectious. However, unlike a standard SIS model, there are two components to the rate of infection. One component comes from inside the household, and one from outside the household. We assume the rate of infection from outside the household is proportional to the fraction of the population that is infectious. A summary of these rates can be seen in Table 2, and the relationship to the master equation of the system can be seen in Fig. 4.

Using these rates, the master equation (a system of $N+1$ differential equations describing the time-evolution of p_n) in the case $n = 2$ can be written down as

$$\begin{aligned}\tfrac{dp_0}{dt} &= -\alpha(p_1 + 2p_2)p_0 + \gamma p_1,\\ \tfrac{dp_1}{dt} &= \alpha(p_1 + 2p_2)p_0 - \left(\tfrac{\alpha}{2}(p_1 + 2p_2) + \beta + \gamma\right)p_1 + 2\gamma p_2,\\ \tfrac{dp_2}{dt} &= \left(\tfrac{\alpha}{2}(p_1 + 2p_2) + \beta\right)p_1 - 2\gamma p_2,\end{aligned} \tag{4}$$

which can be solved analytically to give the SHH approximation to the steady state.

2.2.2. Fokker–Planck peak approximation

We now consider another approximation to this system. Our aim is to find an alternative method that provides a good level of accuracy, without quickly becoming infeasible to derive analytically. To that end, we start by deriving the Fokker–Planck equation for this system (Gardiner, 2004), which can be thought of as a second-order Taylor expansion of the master equation. Doing so, we obtain

$$\begin{aligned}\frac{\partial u}{\partial t} = {} & \frac{\partial}{\partial x}\{(A - B)u(x,y)\} + \frac{\partial}{\partial y}\{(B - A + C - D)u(x,y)\} + \frac{1}{2N}\frac{\partial^2}{\partial x^2}\\ & \times\{(A + B)u(x,y)\} - \frac{1}{N}\frac{\partial^2}{\partial x \partial y}\{(A + B)u(x,y)\} + \frac{1}{2N}\frac{\partial^2}{\partial y^2}\\ & \times\{(A + B + C + D)u(x,y)\},\end{aligned}$$

where

$$x = \frac{m}{N},\ y = \frac{n}{N},\ u(x,y,t) = p_{m,n}(t), A(x,y) = \frac{a_{m,n}}{N},$$

$$B(x,y) = \frac{b_{m,n}}{N}\ C(x,y) = \frac{c_{m,n}}{N}\ D(x,y) = \frac{d_{m,n}}{N}.$$

From the Fokker–Planck equation, we can obtain a deterministic drift vector $\bar{f}$ and a diffusion matrix $\mathbf{D}$. Using these, we define

$$\bar{\alpha}(\bar{x}) = f(\bar{x}) - \frac{1}{2N}\sum_{ij}\frac{\partial D_{ij}}{\partial x_j}\bar{\varepsilon}_i,$$

where $\bar{\varepsilon}_i$ is a unit vector in the direction x_i. The FPP approximation is then obtained by solving the following system of ODES (Mendler et al., 2018):

$$\frac{d\bar{x}}{dt} = \bar{\alpha}(\bar{x}),$$

More details of this derivation can be found in Section A.1.

We note that $\bar{f}(x)$ represents the deterministic drift term for this system, and corresponds to the system of equations obtained for the single household approximation. Similarly, as the number of households $N \to \infty$, the diffusion term tends to 0, and we again obtain the single household approximation. A summary detailing how to obtain each approximation and how they each relate to each other and the master equations can be seen in Fig. 4.

Table 2
State transition rates.

Description	State Transition	Rate
Within household infection	(S,I) → (S-1,I + 1)	$\frac{\beta I}{N-1}$
Between household infection	(S,I) → (S-1,I + 1)	$\alpha\langle I\rangle$
Treatment	(S,I) → (S + 1,I-1)	γ

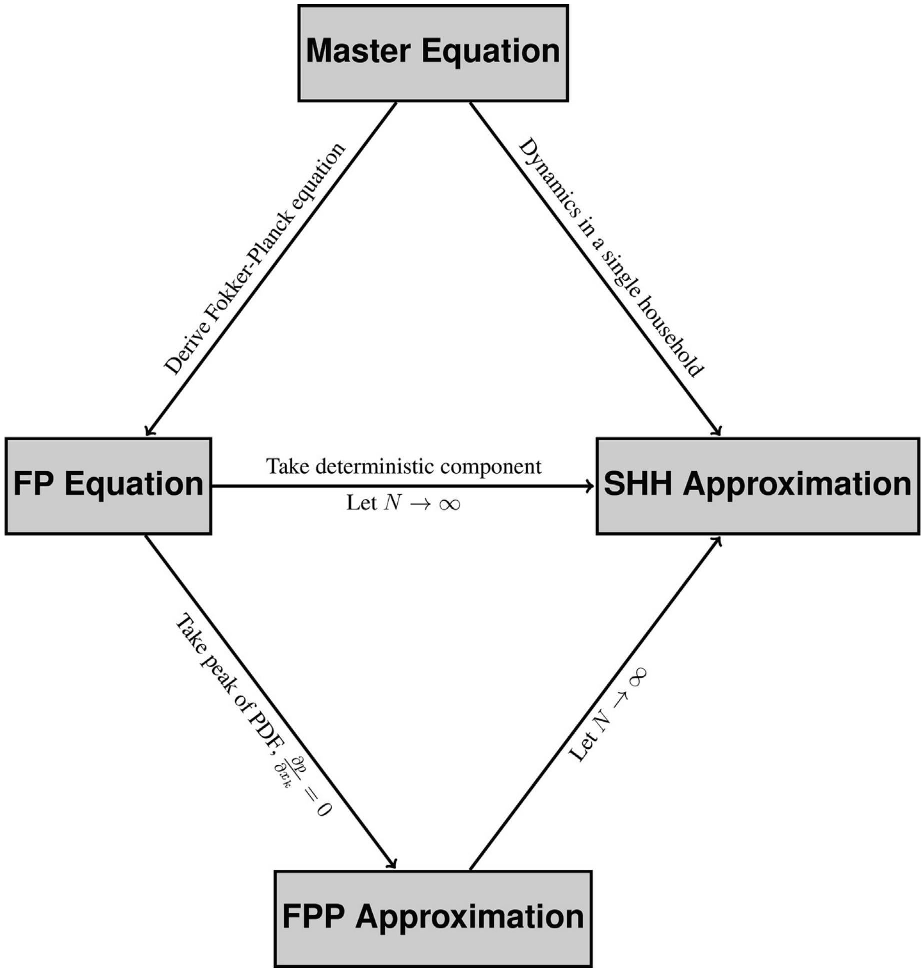

Master Equation
Derive Fokker-Planck equation
Dynamics in a single household
FP Equation
Take deterministic component
Let $N \to \infty$
SHH Approximation
Take peak of PDF, $\frac{\partial p}{\partial x_k} = 0$
Let $N \to \infty$
FPP Approximation

Fig. 4. Figure showing possible transitions between each household state in a population of households of size 2 undergoing SIA dynamics. FPP = Fokker–Planck peak, SHH = single household.

2.3. Phase portraits and correlation

The behaviour around the steady state can vary depending on the parameter values used in the model. More specifically, the correlation between household states $(2,0)$ and $(1,1)$ is greater under certain parameter values. Quantifying the differences in this behaviour could identify another method for distinguishing between parameter regimes in which each approximation is more accurate. To that end, we look at phase portraits of the Fokker–Planck approximated system under different parameter values, as this allows us to examine the steady state behaviour more closely. We plot the nullclines for this system, and the steady state (where the nullclines intersect). We also plot trajectories around the steady state, simulated using the Gillepsie Algorithm and the steady state as determined by this simulated trajectory. Finally, we calculate the steady state distribution from the master equation.

To investigate this further, we generate a set of parameter values using latin hypercube sampling (LHS) (Iman, 2014) to ensure we thoroughly search the parameter space. We then calculate the KL divergence of both the SHH and FPP methods from the true steady state under these parameter values.

The correlation between household states $(2,0)$ and $(1,1)$ is then calculated in two ways. The first is using the Gillespie algorithm to simulate the steady state distribution, from which we can calculate the correlation between household states $(2,0)$ and $(1,1)$. We then plot this correlation against KL divergence.

Secondly we assume the number of households in each household state follows a multinomial distribution, with parameters equal to the steady state solution obtained from the FPP approximation. If our steady state solutions are $(p_m, p_n, 1 - p_m - p_n)$, then

$$\mathrm{Corr}(m,n) = -\frac{\sqrt{p_m p_n}}{\sqrt{(1-p_m)(1-p_n)}}.$$

Similarly, we plot this correlation against KL divergence to ensure we get the same qualitative results when calculating the correlation from the simulated trajectories.

2.4. n-Person households

We extend the work above to now consider households of some arbitrary size n, rather than just size 2 as considered before. In this case, we let m_i denote the number of households in household state $(n-i, i)$. That is, the number of households with $n-i$ susceptible individuals, and i infectious individuals. So m_0 denotes the number of households with n susceptible individuals, while m_n denotes the number of households with n infectious individuals. We consider a fixed population of size N households such that $\sum_{k=0}^{n} m_k = N$. This gives us a population consisting of nN individuals. For a household of size n, we have $n+1$ different possible household states, and $2n$ state transitions. Of these $2n$ state transitions, n are associated with infection events, while the other n are associated with recovery events. This is demonstrated in Fig. 5. Each of these transition rates is summarised in Table 3.

The master equation and Fokker–Planck equation can be derived in the same way as before (see Section A.3) with rates i_k and r_k as given in Table 3. From these, the FPP and SHH approximations can be derived as before.

2.5. Simulating forward steady state master equations

In this section, we look at determining between–household rates of infection from a model that has been fitted at steady state. This can be useful as we may only have steady state data available to us (as in Dyson et al. (2017)), but will still want to simulate the dynamics forward in time after perturbing the system. We can do this by finding the between household rate of infection which, at steady state, corresponds to the distribution of household states we obtain from the steady state model. In particular, we are interested in looking at how well each of the approximations perform.

Suppose we have a household structured model (a population of size $2M$ consisting of M households of size 2) undergoing SIS dynamics that has been parameterised at steady state. That is, we have a constant external force of infection (in a similar manner to the SIA model described previously), ε, rather than explicitly considering the interactions between different households using a between–household rate of infection, α. As the houses are now independent (ε does not depend on the infection level of other households), we can solve this system exactly at steady state using the master equation for a single household. Alternatively, we can use the master equation described initially to obtain a set of ODEs describing the time evolution of the moments of the system. As the equations are linear, this will correspond precisely to the deterministic system we obtain by considering a single household (Hahl and Kremling, 2016).

The master equation can be written as follows

$$\frac{dp_0}{dt} = -2\varepsilon p_0 + \gamma p_1,$$

$$\frac{dp_1}{dt} = 2\varepsilon p_0 - (\varepsilon + \beta + \gamma)p_1 + 2\gamma p_2,$$

$$\frac{dp_2}{dt} = (\varepsilon + \beta)p_1 - 2\gamma p_2.$$

The steady state distribution p^* is given by the eigenvector of this system (when written in matrix form) corresponding to the zero eigenvalue. We then consider the external force of infection to be a function of the infection status of the other households, taking the form given in Eq. 2. Using the SHH approximation, α can be calculated and fed back into the model to obtain a new steady state. Alternatively, the FPP approximation can be used to obtain this value of α by finding the value of α that solves $\alpha_1(\bar{x}) = \alpha_2(\bar{x})$, where $\bar{\alpha}(\bar{x})$ is the system of equations obtained from the FPP approximation.

As before, we are interested in which regions of parameter space one method outperforms the other. To investigate this, we take values of ε, β, and γ, varying each parameter one at a time. We then find the corresponding steady state distribution, and we then find the value of α corresponding to that steady state distribution (as described above). As there is not a monotonic relationship between the KL divergence and α, the existence of an α that corresponds to a value of ε is not guaranteed. Instead, the value of α that minimises the KL divergence of the master equations from the steady state system corresponding to ε is used.

3. Results

3.1. KL divergence from ME

Here we look at the KL divergence of the three methods from the steady state distribution obtained by solving the full system of master equations (Eq. 3) as the parameters vary for a fixed population size of 250 households. A higher KL Divergence indicates the distribution is further from the ‘true’ master equation-derived distribution.

We plot a heat map of the ratio of the two KL divergences, with the colour map centred around 1 (both methods having equal KL divergence). A KL divergence ratio greater than 1 corresponds to the SHH approximation being more accurate (the red areas of the heatmap), while a KL divergence ratio less than one corresponds to the FPP approximation being more accurate (the blue areas of the heatmap). The results of this can be seen in Fig. 6.

Fig. 6 shows that the areas with lower infection and higher recovery parameter values correspond to a greater accuracy of the FPP approximation, while lower recovery and greater infection values correspond to a greater accuracy for the SHH approximation. It can also be useful to consider multiple metrics, to see if any conclusions made are consistent across these different metrics. We provide similar plots using different metrics for assessing accuracy in appendix A.4. Further analysis shows that using a moment closure approximation provides more accurate results than either the SHH and FPP approximations (results not shown). Due to the relative difficulty in deriving systems of moment closure equa-

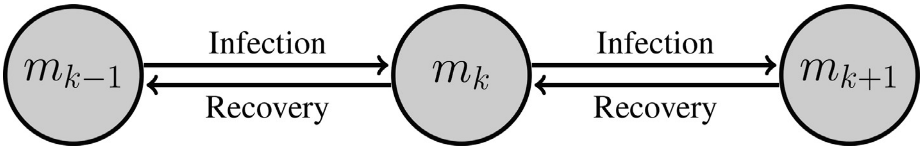
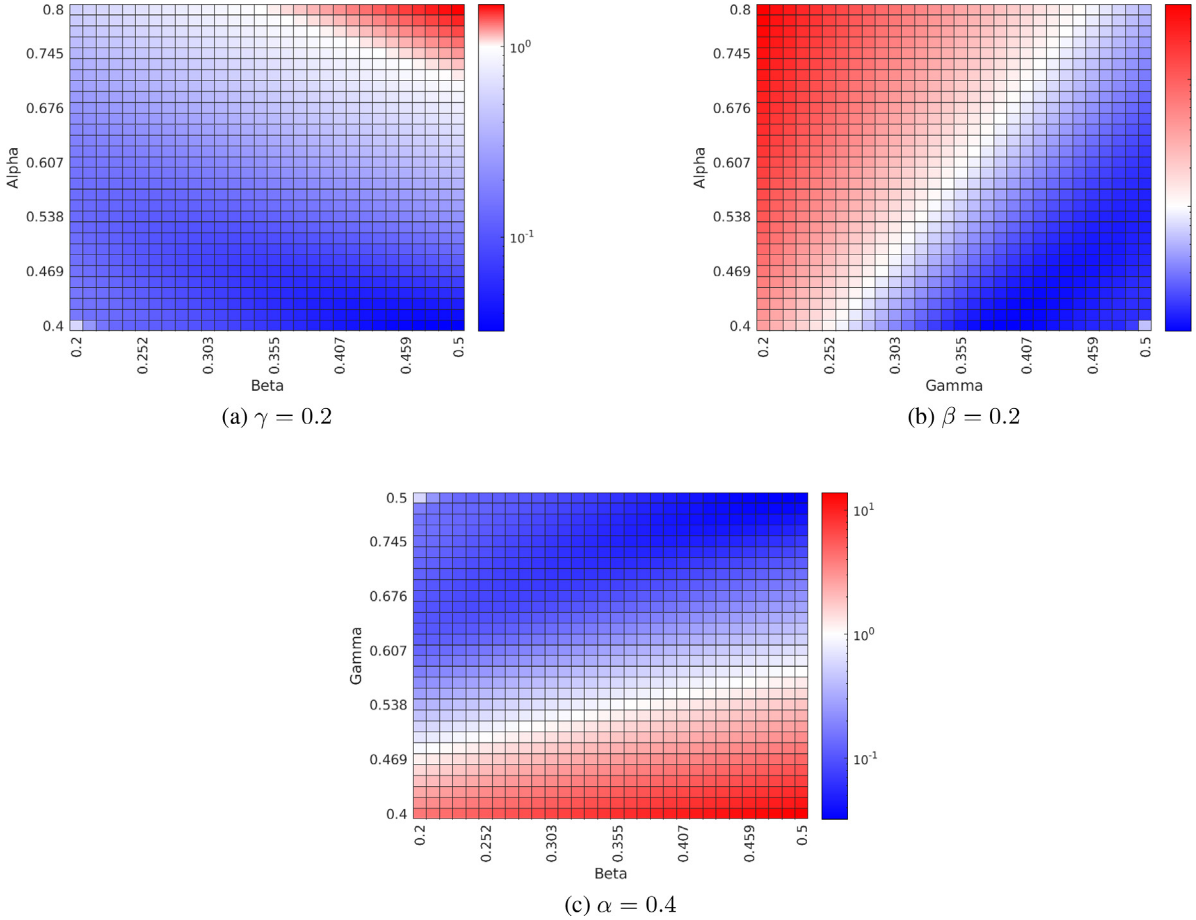


Fig. 5. Figure showing possible transitions between each individual in a population of n-person households undergoing SIS dynamics. m_k denotes the number of households with k infectious individuals.

Table 3
State transitions for n-person SIS household model.

Event	Symbol	State Transition	Rate
Infection	i_k	$(m_k, m_{k+1}) \rightarrow (m_k - 1, m_{k+1} + 1)$	$\left(\frac{\alpha}{nN}\left(\sum_{i=0}^{n} i m_i\right) + \frac{\beta k}{n-1}\right)(n-k)m_k$
Recovery	r_k	$(m_{k-1}, m_k) \rightarrow (m_{k-1} + 1, m_k - 1)$	$k\gamma m_k$



(a) $\gamma = 0.2$

(b) $\beta = 0.2$

(c) $\alpha = 0.4$

Fig. 6. Ratio of KL Divergence of different approximations from the true steady state as parameters vary. Blue areas denote regions of parameter space where the Fokker–Planck approximation outperformed the deterministic approximation, while red areas denote regions of parameter space where the deterministic approximation outperformed the Fokker–Planck approximation. System consisted of 500 households of size 2.

tions, particularly for more complex systems, it is not considered here.

3.2. Phase portraits and correlation

It was observed that the FPP approximation works well in low infection parameter regimes. Now we want to determine whether there are any qualitative differences that explain these differences. To do this, we look at the phase portraits of the system under low-infection and high-infection parameter regimes. Each subplot in Fig. 7 shows three different points plotted, as well as each phase portrait. These are the steady state from the FPP approximation (light blue dot), the true mean steady state obtained by solving the full master equations (red dot), and the steady state obtained by averaging the points on the simulated trajectory from the Gillespie algorithm (black dot). As shown in Fig. 7, the approximation is much closer to the true solution than that obtained from the Gillespie algorithm.

Fig. 7 shows a stark contrast between the behaviour around the steady state under the two different parameter regimes. The two household state (2,0) and (1,1) do not appear to have a strong correlation under the high infection regime (top row), while there is a much stronger correlation between them under a low infection parameter regime (bottom row). Specifically, as the proportion of households in state (2,0) decreases, the proportion of households in state (1,1) increases. Conversely, there is no clear trend when looking at the trajectory around the steady state under the high infection parameter regime.

From this, we hypothesised that households states (2,0) and (1,1) were more correlated under low infection parameter regimes, and would result in the FPP approximation being more accurate. Fig. 8 shows the results of this, with a number of different parameter sets sampled and the KL divergence of each approximation calculated.

As the correlation between the two states increases (from more negative towards 0), the FPP approximation becomes less accurate while the SHH approximation becomes more effective, with equality occurring around a correlation of -0.4 (Fig. 8). The same trend is observed whether we calculate the correlation from simulated trajectories (red and blue dots), or whether we approximate the correlation by assuming a multinomial distribution with parameters given by the steady state obtained using the FPP approximation (black and green dots).

The KL divergences for a population partitioned into households of different sizes can be seen in Table 4, with the steady state distributions for households of sizes 2,3 and 4 shown in Fig. 9. We denote the KL divergence of the distribution obtained under approximation Q from the true steady state distribution P by $D_{KL}(P||Q)$. We let Q1 and Q2 denote the steady state distributions obtained using the SHH and FPP approximations respectively.

Based on visual inspection of Fig. 9 and looking at the KL divergences, the accuracy of the Fokker–Planck approximation relative to the single household approximation appears to decrease as household size increases.

We now return to the example described in the introduction — how can we find appropriate parameters to simulate a system

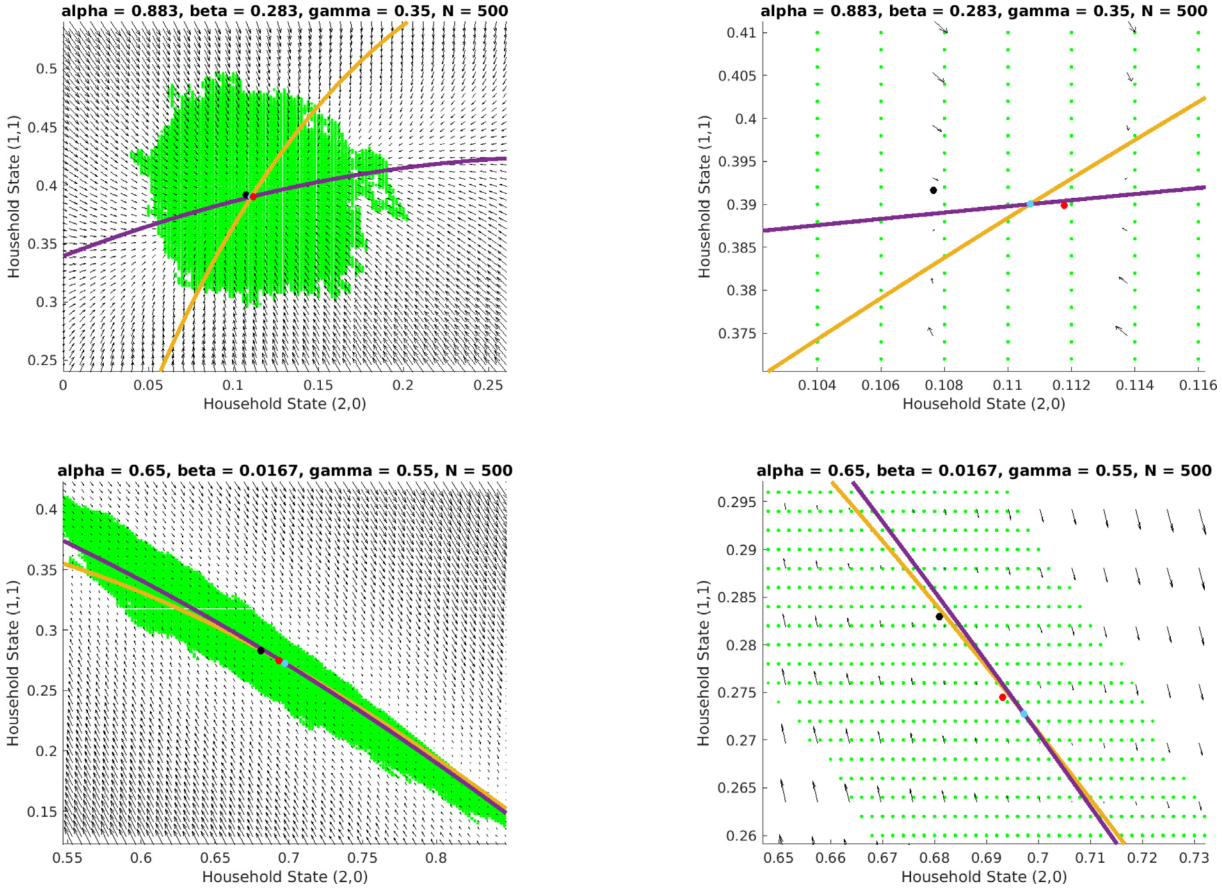


Fig. 7. Phase portraits of SIS system with households of size 2 under different parameter regimes. Figures on the right are expanded copies of those on the left. The light blue dot denotes the distribution obtained from the FPP approximation, the red dot is the true distribution, and the black dot is that obtained from the Gillespie Algorithm.

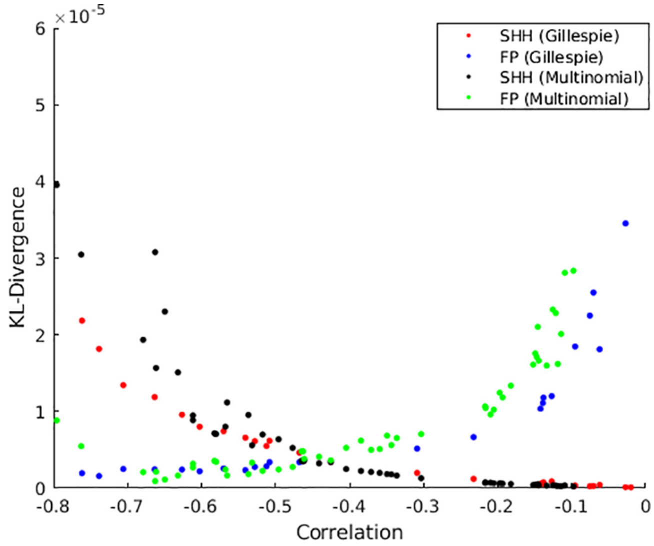


Fig. 8. Plot of the correlation between household states $(2,0)$ and $(1,1)$ (as determined by simulations from the Gillespie algorithm (red and blue) and assuming a multinomial distribution (black and green)), and the KL divergence of the FPP and SHH approximations from the true steady state distribution for $N = 400$.

parameterised at steady state forward in time? Defining ε as in Eq. 2, we take $\varepsilon = 0.05, \beta = 0.1, \gamma = 0.15$, with the aim of determining a between–household rate of infection corresponding to ε with the same steady state distribution as this system.

Table 4
KL divergence for each method under different household sizes using parameters $\alpha = 0.4$, $\beta = 0.3$ and $\gamma = 0.45$.

Household Size	Number of Households	$D_{KL}(P\|\|Q_1)$	$D_{KL}(P\|\|Q_2)$
2	60	0.00381	2.35×10^{-4}
3	40	0.0035	0.00252
4	30	0.00384	0.0193

The steady state distribution of this system is

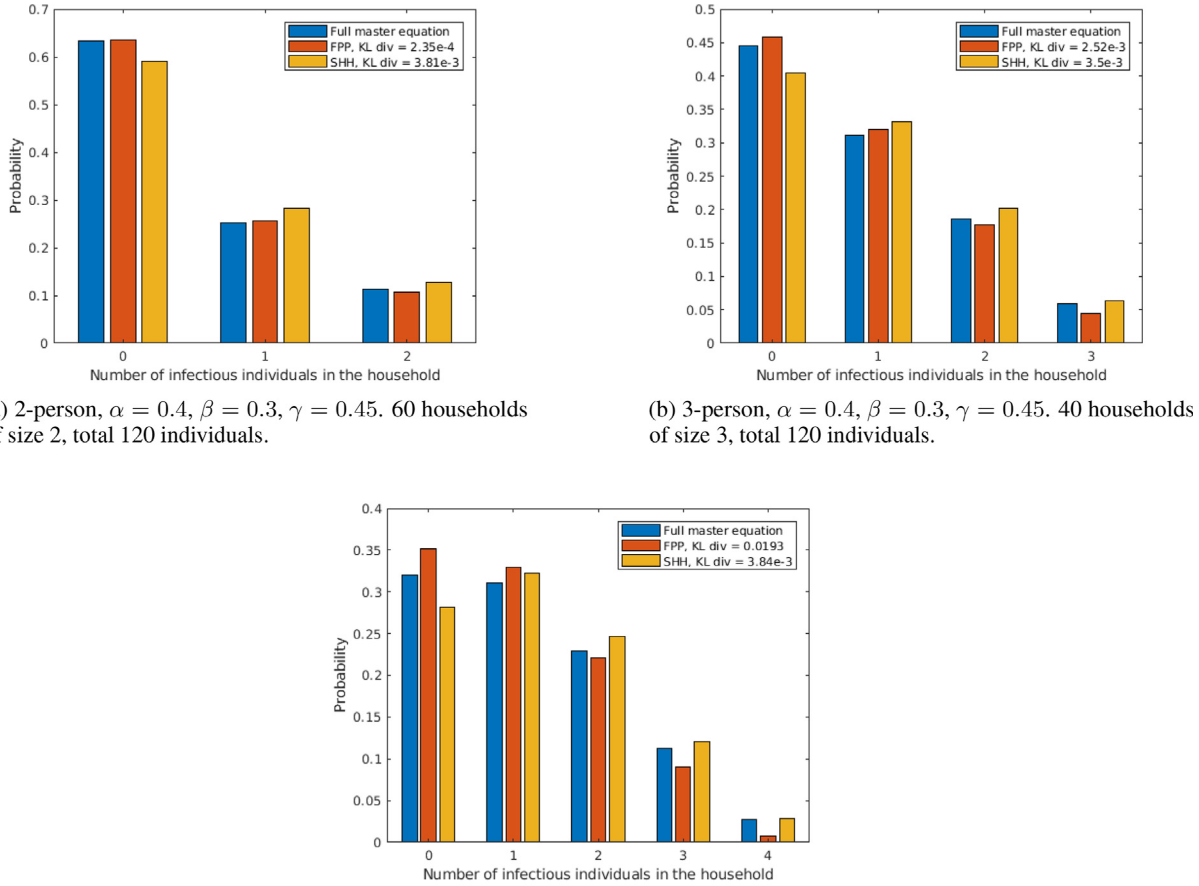
$$\mathbf{p}^* = \begin{pmatrix} p_0 \\ p_1 \\ p_2 \end{pmatrix} = \begin{pmatrix} \frac{1}{2} \\ \frac{1}{3} \\ \frac{1}{6} \end{pmatrix},$$

where p_i denotes the probability of a household having i infectious individuals at steady state. Rearranging Eq. 2, we can calculate the corresponding value of α as being $\alpha = 0.15$.

However, we now put this value of α back into the master equation for the full system (Eq. 3), which gives us a steady state distribution of

$$\mathbf{p}^* = \begin{pmatrix} 0.516 \\ 0.323 \\ 0.161 \end{pmatrix}.$$

The KL divergence of this steady state distribution from the true steady state distribution can be calculated as $D_{KL}(P||Q) = 5.16\times 10^{-4}$, which compares to a KL divergence of 1.12×10^{-5} for the optimal value of α (the value which minimises the KL divergence).



(a) 2-person, $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.45$. 60 households of size 2, total 120 individuals.

(b) 3-person, $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.45$. 40 households of size 3, total 120 individuals.

(c) 4-person, $\alpha = 0.4$, $\beta = 0.3$, $\gamma = 0.45$. 30 households of size 4, total 120 individuals.

Fig. 9. Steady state distributions obtained using the FPP and SHH approximations, with the true steady state as determined by solving the set of master equations for 2, 3 and 4 person households. The KL divergence of each approximation from the true steady state is given in the legend.

If instead the Fokker–Planck peak approximation is used, taking α such that $\alpha_1(x^*, y^*) = \alpha_2(x^*, y^*)$, we get a KL divergence from the true steady state distribution of 3.19×10^{-4}, which is lower than that obtained by using the single household approximation.

The KL divergence for each approximation as we vary parameters can be seen in Fig. 10. We observe the same trends we did previously, with the FPP approximation outperforming the SHH approximation in regions of parameter space corresponding to lower levels of infection.

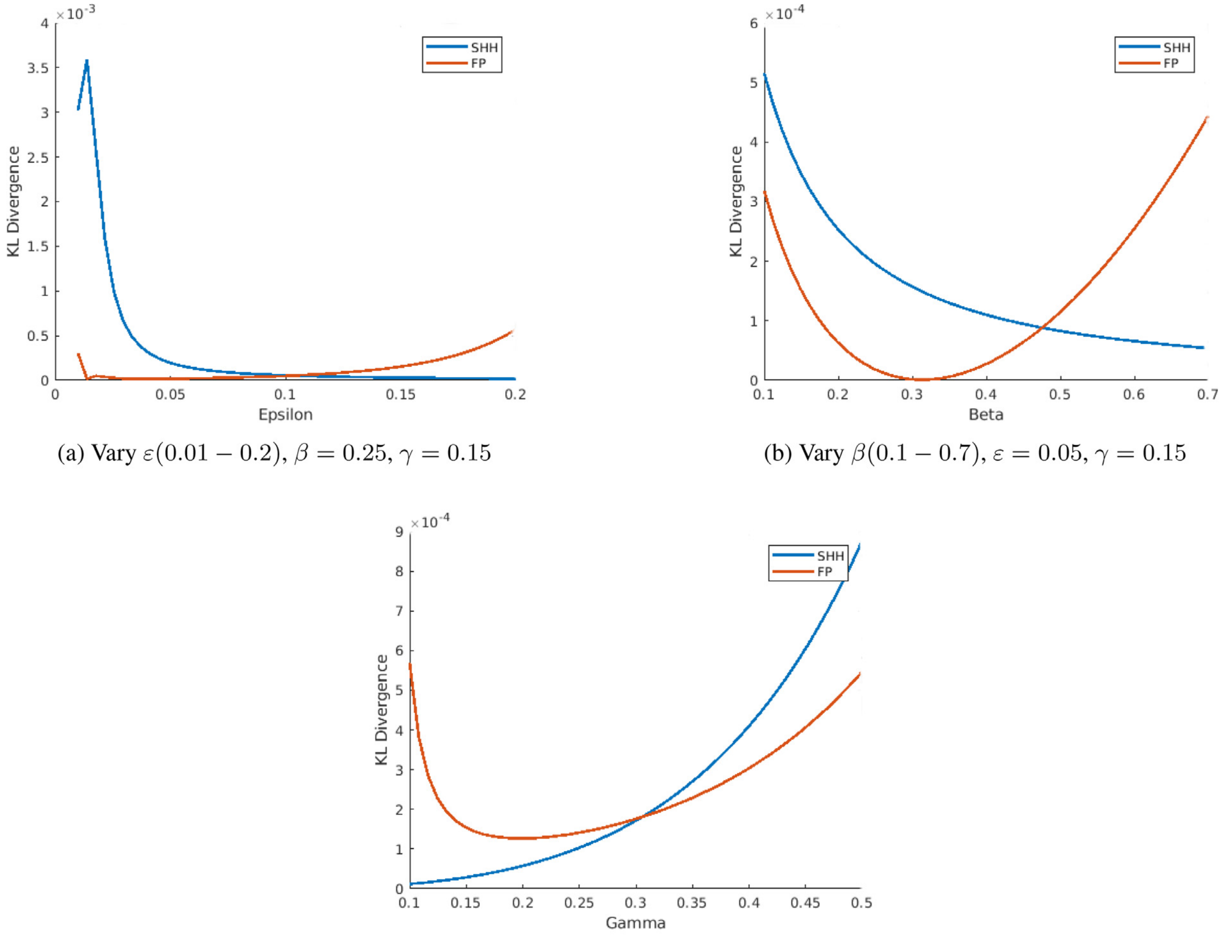
4. Discussion

In this study we investigated the accuracy of different approximations to the steady state distribution of a system consisting of a population of households undergoing stochastic SIS dynamics. It was shown that at steady state, it is insufficient to consider the stochastic dynamics of a single household (or equivalently, the deterministic dynamics for the population of households). Instead, a more refined model is required that considers all reactions between households.

The difficulty with this new model lies in the complexity of the system – it is no longer feasible to solve numerically in most situations. To that end, we looked at the accuracy and ease of implementation of two different approximations, the single household approximation and the Fokker–Planck peak approximation.

In a low infection parameter regime, the FPP approximation outperformed the SHH approximation, meaning the KL divergence between the true solution (obtained numerically) and the solution obtained from the FPP approximation was lower than that from the SHH approximation. However, as we move into a higher infection parameter regime (one with larger infection rates or smaller recovery rates), the SHH outperforms the FPP approximation. This is shown by varying two parameters at a time in Fig. 6. Thus, we found that the accuracy of each approximation was closely related to the prevalence associated with the region of parameter space we are in. While it is an expected result that the SHH approximation will perform better in a high infection parameter regime than in a low infection parameter regime (as the different households would synchronise more quickly), it is less clear as to whether we would expect the SHH approximation to outperform the FPP approximation in this scenario.

The FPP approximation works by taking the peak of the steady state distribution produced using the Fokker–Planck equation. As such, this approximation to the mean value is only valid when the distribution is not skewed. However, we could expect the distribution to be more skewed when pushed into the corner of the domain (corresponding to system state $(0, 0, N)$), as here the bounds of the region are having a larger impact on the steady state



(a) Vary $\varepsilon(0.01 - 0.2)$, $\beta = 0.25$, $\gamma = 0.15$

(b) Vary $\beta(0.1 - 0.7)$, $\varepsilon = 0.05$, $\gamma = 0.15$

(c) Vary $\gamma(0.1 - 0.5)$, $\varepsilon = 0.15$, $\beta = 0.25$

Fig. 10. KL Divergence of the steady state obtained from FPP and SHH approximations when used to simulate a system forwards in time.

distribution. This typically corresponds to a high infection parameter regime (as most households have two infectious individuals, and very few have none). The SHH approximation does not depend on the number of households in the system. As such, we may expect this to perform less well at smaller population sizes.

Looking in closer detail at an example from each parameter regime, it was observed that household states $(2,0)$ and $(1,1)$ in systems in the low infection parameter regime were more strongly correlated than those in a high infection parameter regime. This is due to strong negative correlations (which are present between susceptible and infectious individuals in stochastic models Keeling and Rohani, 2008) primarily occurring along the diagonal boundary of the domain. This suggested that the correlation between household states may be a good statistic to identify which approximation should be used, rather than having to investigate the full 3-dimensional parameter space. While simulating trajectories with which to calculate the correlation can be costly, it can be approximated by assuming the number of households in each state follows a multinomial distribution. Fig. 8 shows that both methods obtain similar KL divergences. It also shows that the FPP approximation is more accurate in high correlation parameter regimes, while the SHH approximation is more accurate in low correlation parameter regimes, with both methods performing similarly at a correlation between -0.3 and -0.4. While there are some differences in the correlations obtained under each method, it is a broad region of correlation values we are interested in, rather than specific values the correlation takes.

This behaviour can typically be explained by the way households move between states. A household cannot move between states $(2,0)$ and $(0,2)$ without first passing through $(1,1)$. In a low infection parameter regime, most of the dynamics will be happening between household states $(2,0)$ and $(1,1)$ causing the correlation between these household states to be stronger. Conversely, in a high infection parameter regime, it is mostly interaction between household states $(1,1)$ and $(0,2)$ that will be occurring, causing the correlation between these two states to be stronger (and the correlation between $(2,0)$ and $(1,1)$ to be weaker). However, we note that multiple household distributions can produce the same prevalence while having different correlations between household states $(2,0)$ and $(1,1)$, and so prevalence and correlation are not completely equivalent. This is particularly noticeable in extreme scenarios whereby between–household infection is largely replaced by a much higher level of within–household infection. We note that correlation between household states does not explain these edge cases any more successfully than prevalence.

After analysing a system of 2-person households, a natural extension was to consider larger household sizes. Due to computational constraints, only households up to size 4 were considered, with the total population size remaining fixed. Fig. 9 shows that for the parameter set considered, the FPP approximation outperformed the SHH approximation for populations of 2-person and 3-person households, but that the reverse was true for 4-person households. It should be noted that the FPP KL divergence was several orders of magnitude lower for the 2-person household system, but the KL divergence for both the FPP and SHH approximations had the same order of magnitude when considering 3-person households. This suggests the relative effectiveness of the FPP approximation to the SHH approximation decreased as the household size increased. Therefore we conjecture that the accuracy of one approximation relative to the other may depend on the complexity of the system (as defined by the number of system states), with the SHH approximation outperforming the FPP approximation in more complex systems.

Our motivation for investigating this problem was for simulating forward systems that have been parameterised at steady state. In particular, we found that just using the SHH approximation to obtain an between–household rate of infection, and then using this between–household rate of infection to simulate the system forward again to steady state didn't produce the same steady state we started at. Using the methodology we present here, there are multiple ways of achieving this. In particular, we wanted to determine whether we could obtain a more accurate solution to this problem using the FPP approximation. The results (displayed in Fig. 10) are broadly consistent with the results obtained previously: the FPP approximation is more accurate in a low infection parameter regime, while the SHH approximation is more accurate than the FPP approximation in a higher infection parameter regime.

There are a number of avenues not considered here that could be considered in future work. Firstly, we only consider stochastic SIS household structured models. Applying this methodology to a wider range of models, such as SIR and SEIR models, would increase its utility. However, difficulties will occur in finding exact solutions to the master equations. One approach is to use a proxy (e.g. simulating realisations using the Gillespie algorithm) for the true solution. However, there will then be some error around any results, and it is possible that this error could be greater than the error in the approximations themselves making it difficult to assess the validity of each approximation. Similarly, we considered households of sizes 2,3 and 4. Further work should look at larger household sizes to be sure that the trend (FPP becomes less accurate, SHH becomes more accurate relative to FPP) continues. Distributions of household sizes should also be investigated, to better match real-world populations.

We considered two approximations in this work, the SHH and FPP approximations. Whilst moment closure was not considered here, it provides a more accurate approximation than the SHH and FPP approximations at the expense of simplicity to derive. Future work should look at other approximations to this system. Finally, we have shown that correlation between households acts as a good metric in determining which approximation is likely to be more accurate. Further work should look into understanding which correlations are most indicative of this when there are more household states to consider.

For values of R0 close to 1, the FPP approximation may not provide valid solutions at all. As such, the FPP approximation should only be used when R0 is sufficiently high. In this paper, we only considered parameters that resulted in R0 > 1.05 to avoid any risk of the FPP approximation not finding a valid approximation. The results provided in Table 4 suggest the FPP approximation may not be useful in more complicated systems (e.g. an SEIR model, or a model with a larger household size) due to the decrease in accuracy as the size of the state space increases.

In conclusion, we have shown that in order to accurately model a population of households of size 2, it is necessary to fully consider all interactions between states, even if just for analysing the steady states of the model. We have shown that under certain parameter regimes, the FPP approximation provides a more accurate approximation than the SHH approximation and that in the 2-person household case, these parameter regimes are well classified by the correlation between two household states. There are many future directions in which this work could be taken, including investigating systems with a higher complexity (e.g. SEIR model) and generalising the use of correlation as a metric to assess the accuracy of each approximation.

CRediT authorship contribution statement

Alex Holmes: Conceptualization, Methodology, Software, Investigation, Writing - original draft, Writing - review & editing, Visualization. **Mike Tildesley:** Methodology, Writing - review & editing, Supervision. **Louise Dyson:** Conceptualization, Methodology, Writing - review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding declaration

This work has been funded by the Engineering and Physical Sciences Research Council through the MathSys CDT [Grant No. EP/S022244/1]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Appendix A. Appendices

A.1. Fokker–Planck derivation

The master equation for a population of N households of size two, all undergoing SIS dynamics, can be written as follows

$$\frac{\partial p_{m,n}}{\partial t} = \left(\mathbb{E}^{1,-1} - 1\right)a_{m,n}p_{m,n} + \left(\mathbb{E}^{-1,1} - 1\right)b_{m,n}p_{m,n} + \left(\mathbb{E}^{0,1} - 1\right)c_{m,n}p_{m,n} + \left(\mathbb{E}^{0,-1} - 1\right)d_{m,n}p_{m,n},$$

with the following rates:

$$a_{m,n} = \frac{\alpha}{N}(n + 2(N - m - n))m$$

$$b_{m,n} = \gamma n$$

$$c_{m,n} = \left(\frac{\alpha}{2N}(n + 2(N - m - n)) + \beta\right)n$$

$$d_{m,n} = 2\gamma(N - m - n),$$

We now want to derive the Fokker–Planck equation, a continuous approximation to the master equation. To this end, we define the following:

$$x = \frac{m}{N},\ y = \frac{n}{N},\ u(x, y, t) = p_{m,n}(t),$$

$$A(x, y) = \frac{a_{m,n}}{N},\ B(x, y) = \frac{b_{m,n}}{N}\ C(x, y) = \frac{c_{m,n}}{N}\ D(x, y) = \frac{d_{m,n}}{N}.$$

Using this, we can write the master equation as follows

$$\frac{\partial u}{\partial t} = N\left(\widehat{\mathbb{E}}^{1,-1} - 1\right)A(x, y)u(x, y, t) + N\left(\widehat{\mathbb{E}}^{-1,1} - 1\right)B(x, y)u(x, y, t) + N\left(\widehat{\mathbb{E}}^{0,1} - 1\right)C(x, y)u(x, y, t) + N\left(\widehat{\mathbb{E}}^{0,-1} - 1\right)D(x, y)u(x, y, t),$$

with

$$\widehat{\mathbb{E}}^{a,b} f(x,y) = f\left(x + \frac{a}{N}, y + \frac{b}{N}\right).$$

We can then calculate the multivariate Taylor expansion of this operator up to second order.

$$\widehat{\mathbb{E}}^{a,b} \approx 1 + \frac{a}{N}\frac{\partial}{\partial x} + \frac{b}{N}\frac{\partial}{\partial y} + \frac{1}{2!}\left[\left(\frac{a}{N}\right)^2 \frac{\partial^2}{\partial x^2} + 2\left(\frac{a}{N}\right)\left(\frac{b}{N}\right)\frac{\partial^2}{\partial x \partial y} + \left(\frac{b}{N}\right)^2 \frac{\partial^2}{\partial y^2}\right].$$

Substituting this into our master equation, we get

$$\begin{aligned}\tfrac{\partial u}{\partial t} &= N\left(\tfrac{1}{N}\tfrac{\partial}{\partial x} - \tfrac{1}{N}\tfrac{\partial}{\partial y} + \tfrac{1}{2N^2}\left\{\tfrac{\partial^2}{\partial x^2} - 2\tfrac{\partial^2}{\partial x \partial y} + \tfrac{\partial^2}{\partial y^2}\right\}\right)A(x,y)u(x,y,t) \\ &+ N\left(\tfrac{-1}{N}\tfrac{\partial}{\partial x} + \tfrac{1}{N}\tfrac{\partial}{\partial y} + \tfrac{1}{2N^2}\left\{\tfrac{\partial^2}{\partial x^2} - 2\tfrac{\partial^2}{\partial x \partial y} + \tfrac{\partial^2}{\partial y^2}\right\}\right)B(x,y)u(x,y,t) \\ &+ N\left(\tfrac{1}{N}\tfrac{\partial}{\partial y} + \tfrac{1}{2N^2}\tfrac{\partial^2}{\partial y^2}\right)C(x,y)u(x,y,t) + N\left(\tfrac{-1}{N}\tfrac{\partial}{\partial y} + \tfrac{1}{2N^2}\tfrac{\partial^2}{\partial y^2}\right)D(x,y)u(x,y,t).\end{aligned}$$

We can then combine terms into the matching derivatives, giving us the desired Fokker–Planck equation

$$\begin{aligned}\frac{\partial u}{\partial t} &= \frac{\partial}{\partial x}\{(A-B)u(x,y)\} + \frac{\partial}{\partial y}\{(B-A+C-D)u(x,y)\} \\ &+ \frac{1}{2N}\frac{\partial^2}{\partial x^2}\{(A+B)u(x,y)\} - \frac{1}{N}\frac{\partial^2}{\partial x \partial y}\{(A+B)u(x,y)\} \\ &+ \frac{1}{2N}\frac{\partial^2}{\partial y^2}\{(A+B+C+D)u(x,y)\}.\end{aligned}$$

A.2. Approximating solution from Fokker–Planck equation

We present the derivation of the FPP approximation from a FP equation for the system, following the work presented in Mendler et al. (2018).

Fokker–Planck equations take the following form

$$\frac{\partial p(\bar{x},t)}{\partial t} = -\sum_i \frac{\partial}{\partial x_i}[f_i(\bar{x})p(\bar{x},t))] + \frac{1}{2N}\sum_{i,j}\frac{\partial^2}{\partial x_i \partial x_j}\left[D_{ij}(\bar{x})p(\bar{x},t)\right]. \quad (5)$$

We can manipulate the equation above in the following way

$$\begin{aligned}\tfrac{\partial p(\bar{x},t)}{\partial t} &= -\sum_i \tfrac{\partial}{\partial x_i}[f_i(barx)p(barx,t))] + \tfrac{1}{2N}\sum_{i,j}\tfrac{\partial^2}{\partial x_i \partial x_j}\left[D_{ij}(barx)p(barx,t)\right] \\ &= -\sum_i \tfrac{\partial}{\partial x_i}\left[f_i(barx)p(barx,t) - \tfrac{1}{2N}\sum_j \tfrac{\partial}{\partial x_j}\left[D_{ij}(barx)p(barx,t)\right]\right] \\ &= -\sum_i \tfrac{\partial}{\partial x_i}\left[f_i(barx)p(barx,t) - \tfrac{1}{2N}\sum_j D_{ij}(\bar{x})\tfrac{\partial p}{\partial x_j} - \tfrac{1}{2N}\sum_j \tfrac{\partial D_{ij}}{\partial x_j}p(\bar{x},t)\right] \\ &= -\sum_i \tfrac{\partial}{\partial x_i}\left[\left(f_i(\bar{x}) - \tfrac{1}{2N}\sum_j \tfrac{\partial D_{ij}}{\partial x_j}\right)p(\bar{x},t) - \tfrac{1}{2N}\sum_j D_{ij}(\bar{x})\tfrac{\partial p}{\partial x_j}\right].\end{aligned}$$

In this form, the FPE takes the form of a continuity equation, $\frac{\partial p}{\partial t} = -\bar{\nabla}\cdot\bar{j}$, where

$$j_i = \left(f_i(\bar{x}) - \frac{1}{2N}\sum_j \frac{\partial D_{ij}}{\partial x_j}\right)p(\bar{x},t) - \frac{1}{2N}\sum_j D_{ij}(\bar{x})\frac{\partial p}{\partial x_j}$$

is a probability current. We then define

$$\bar{\alpha}(\bar{x}) = f(\bar{x}) - \frac{1}{2N}\sum_{ij}\frac{\partial D_{ij}}{\partial x_j}\bar{\varepsilon}_i,$$

where $\bar{\varepsilon}_i$ is a unit vector in the direction x_i. We now consider the following system of ODES:

$$\frac{d\bar{x}}{dt} = \bar{\alpha}(\bar{x}).$$

This system is obtained by setting $\frac{\partial p}{\partial x_j} = 0 \ \forall j$, and so corresponds to the maxima of the stationary probability density. Thus, to approximate the probability density by these maxima, we simply need to calculate steady states of this new system.

Applying this to our specific system, we obtain

$$\bar{f} = \begin{pmatrix} B-A \\ A-B+D-C \end{pmatrix} = \begin{pmatrix} -2\alpha x + 2\alpha x^2 + gy + \alpha xy \\ 2g + (2\alpha - 2g)x - 2\alpha x^2 - (\alpha + \beta + 3g)y + \frac{\alpha}{2}y^2 \end{pmatrix},$$

and the diffusion matrix

$$\mathbf{D} = \begin{pmatrix} A+B & -(A+B) \\ -(A+B) & A+B+C+D \end{pmatrix}$$

$$= \begin{pmatrix} 2\alpha x - 2\alpha x^2 + \gamma y - \alpha xy & -(2\alpha x - 2\alpha x^2 + \gamma y - \alpha xy) \\ -(2\alpha x - 2\alpha x^2 + \gamma y - \alpha xy) & 2\gamma - 2\alpha x^2 + (\alpha + \beta - \gamma)y - \frac{\alpha}{2}y^2 + (2\alpha - 2\gamma)x - 2\alpha xy \end{pmatrix}.$$

The FPP approximation is then obtained by solving the following system of ODES:

$$\frac{d\bar{x}}{dt} = \bar{\alpha}(\bar{x}),$$

A.3. n-Person households

As in the 2-person household case, we can derive the master equation for SIS dynamics in the n-person household case, with rates given in Table 3 as follows:

$$\frac{\partial p_{\mathbf{m}}}{\partial t} = \sum_{k=0}^{n-1}\left(\mathbb{E}_{k,k+1}^{1,-1} - 1\right)i_{k,\mathbf{m}}p_{\mathbf{m}} + \sum_{k=0}^{n-1}\left(\mathbb{E}_{k,k+1}^{-1,1} - 1\right)r_{k,\mathbf{m}}p_{\mathbf{m}},$$

where $\mathbb{E}_{k_1,k_2}^{a,b}$ is a generalisation of the step operator we used before, such that for $k_1 < k_2$,

$$\begin{aligned}&\mathbb{E}_{k_1,k_2}^{a,b} f(m_1, m_2, \ldots, m_{k_1}, \ldots, m_{k_2}, \ldots, m_{n+1}) \\ &\quad = f(m_1, m_2, \ldots, m_{k_1} + a, \ldots, m_{k_2} + b, \ldots, m_{n+1})\end{aligned}$$

As we did in the $n = 2$ case, we can derive the Fokker–Planck equation. Doing this, we can define the deterministic drift vector as the vector f with k_{th} component

$$f_k = -\left[\sum_{i=0}^{n}(I_i - R_i)\delta_{i,k} + \sum_{i=0}^{n}(R_i - I_i)\delta_{i+1,k}\right].$$

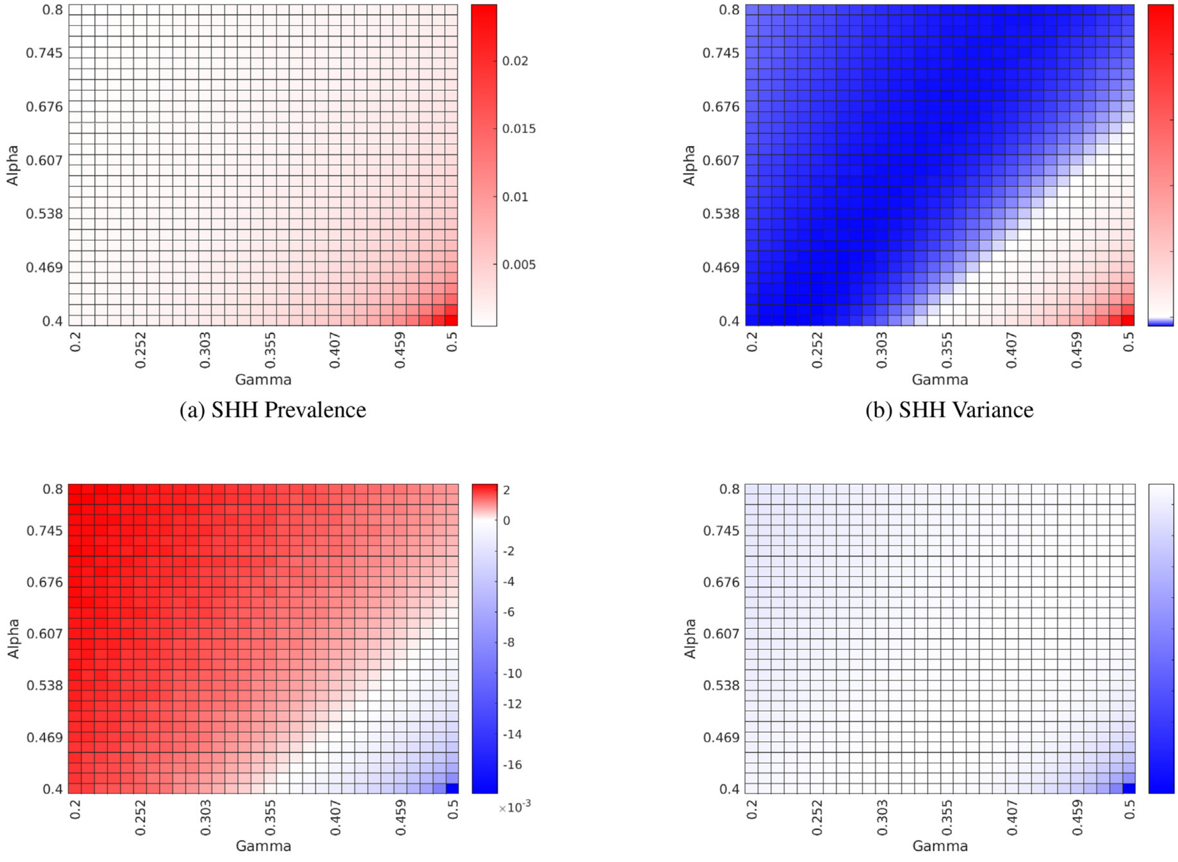
We can then define the diffusion matrix as the matrix D with $(k,l)_{th}$ component

$$\begin{aligned}D_{kl} &= \sum_{i=0}^{n}(I_i + R_i)\delta_{i,k}\delta_{i,l} + \sum_{i=0}^{n}(I_i + R_i)\delta_{i,k}\delta_{i+1,l} + \sum_{i=0}^{n}(I_i + R_i)\delta_{i+1,k}\delta_{i,l} \\ &+ \sum_{i=0}^{n}(I_i + R_i)\delta_{i+1,k}\delta_{i+1,l},\end{aligned}$$

which can then be used to calculate the FPP approximation for this system.

A.4. Alternative metrics for assessing accuracy

We previously looked at how the KL divergence under each approximation varies as we explore the parameter space. However, there are a number of other metrics we could use instead of the KL divergence. Here we consider how the steady state prevalence and the variance of this prevalence changes as parameters vary. In



(a) SHH Prevalence

(b) SHH Variance

(c) FPP Prevalence

(d) FPP Variance

Fig. 11. Heatmaps showing the error in the prevalence (left column) and variance (right column) under the SHH (top row) and FPP (bottom row) approximations. Red areas represent the approximation over-estimating the true value, while blue areas represent the approximation under-estimating the true value. Steady states were obtained for a population of 500 households of size 2. In all four plots we take $\beta = 0.2$.

particular, we are interested in whether any of the approximations systematically under- or over-estimate the true mean and variance. For each approximation, we calculate $\Delta_{var} := \mathrm{Var}_{approx} - \mathrm{Var}_{True}$, and plot this difference as we vary α and β. We also do the same for the error in the prevalence, with results for both the prevalence and variance under each approximation shown in Fig. 11. Unlike the results displayed in Fig. 6, here we instead calculate an absolute error for each approximation, rather than the error in one approximation relative to the other.

In Fig. 11, we see that the SHH approximation consistently over-estimates the true prevalence of the system, while the FPP approximation switches between over- and under-estimating the prevalence. However, we find that the FPP approximation only under-estimates prevalence in regions of parameter space where R_0 is close to one (and so prevalence is low). Conversely, the FPP approximation consistently under-estimates the variance, while the SHH approximation switches between over- and under-estimating the variance. However, we find that the SHH approximation only over-estimates variance in regions of parameter space where R_0 is close to one (and so prevalence is low).

A.5. Model code

All code used in producing these figures can be found athttps://github.com/aholmes95/PhD/tree/master/Approximations%20Paper

References

Ball, Frank, Lyne, Owen, 2001. Stochastic multi-type sir epidemics among a population partitioned into households. Adv. Appl. Prob. 33: 99–123. doi:10.1017/S000186780001065X..

Ball, Frank, Mollison, Denis, Scalia-Tomba, Gianpaolo, 1997. Epidemics with two levels of mixing. Ann. Appl. Probab., 7 (1): 46–89, 02 1997. doi:10.1214/aoap/1034625252. url: https://doi.org/10.1214/aoap/1034625252..

Black, Andrew J., House, Thomas, Keeling, M.J., Ross, J.V., 2013. Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza, J. R. Soc. Interface 10 (81): 20121019, 2013. doi:10.1098/rsif.2012.1019. url: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2012.1019..

Black, Andrew J, Thomas House, Matt J Keeling, Joshua V. Ross, 2014. The effect of clumped population structure on the variability of spreading dynamics, J. Theor. Biol., 359: 45–53. ISSN 0022–5193. doi:10.1016/j.jtbi.2014.05.042. url: https://www.sciencedirect.com/science/article/pii/S0022519314003312..

Dickman, Ronald, Vidigal, Ronaldo, Jan 2002. Quasi-stationary distributions for stochastic processes with an absorbing state. J. Phys. A: Math. Gen. 35 (5), 1147–1166. https://doi.org/10.1088/0305-4470/35/5/303. url:https://doi.org/10.1088.

Dyson, Louise, Michael Marks, Oliver M. Crook, Oliver Sokana, Anthony W. Solomon, Alex Bishop, David C.W. Mabey, T Déirdre Hollingsworth, 2017. Targeted Treatment of Yaws With Household Contact Tracing: How Much Do We Miss? Am. J. Epidemiol. 187(4) (2017) 837–844. ISSN 0002–9262. doi:10.1093/aje/kwx305. url:https://doi.org/10.1093/aje/kwx305..

Ball, Frank, Neal, Peter, 2002. A general model for stochastic sir epidemics with two levels of mixing. Mathematical Biosciences, 180 (1): 73–102. ISSN 0025–5564. doi:10.1016/S0025-5564(02)00125-6. url:http://www.sciencedirect.com/science/article/pii/S0025556402001256..

Fraser, Christophe, Cummings, Derek A.T., Klinkenberg, Don, Burke, Donald S., Ferguson, Neil M., 2011. Influenza Transmission in Households During the 1918 Pandemic. Am. J. Epidemiol. 174 (5): 505–514.. ISSN 0002–9262. doi:10.1093/aje/kwr122. url: https://doi.org/10.1093/aje/kwr122..

Gardiner, C.W., 2004. Handbook of stochastic methods for physics, chemistry and the natural sciences, volume 13 of Springer Series in Synergetics. Springer-Verlag, Berlin, third edition. ISBN 3-540-20882-8..

Hahl, Sayuri K., Kremling, Andreas, 2016. A comparison of deterministic and stochastic modeling approaches for biochemical reaction systems: On fixed points, means, and modes. Frontiers in Genetics, 7: 157. ISSN 1664–8021. doi:10.3389/fgene.2016.00157. url: https://www.frontiersin.org/article/10.3389/fgene.2016.00157..

Hilton, Joe, Keeling, Matt J., 2019. Incorporating household structure and demography into models of endemic disease. J. R. Soc. Interface 16 (157), 20190317. https://doi.org/10.1098/rsif.2019.0317. url:https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2019.0317.

Holmes, A., Tildesley, M.J., Solomon, A.W., Mabey, D.C.W., Sokana, O., Marks, M., Dyson, L., 2020. Modeling treatment strategies to inform yaws eradication. Emerg. Infect Dis. 26 (11), 2685–2693 [PubMed Central:PMC7588528] DOI:10.3201/eid2611.191491.

House, T., Keeling, M.J., 2009. Household structure and infectious disease transmission. Epidemiol. Infection 137 (5), 654–661. https://doi.org/10.1017/S0950268808001411.

Iman, Ronald L., 2014. Latin Hypercube Sampling. American Cancer Society. ISBN 9781118445112. doi:10.1002/9781118445112.stat03803. url:https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118445112.stat03803..

Keeling, Matt J., 2000. Metapopulation moments: coupling, stochasticity and persistence. J. Anim. Ecol. 69 (5), 725–736. https://doi.org/10.1046/j.1365-2656.2000.00430.x. url: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1046/j.1365-2656.2000.00430.x.

Keeling, Matt J., Rohani, Pejman, 2008. Modeling Infectious Diseases in Humans and Animals. Princeton University Press. ISBN 9780691116174, url: http://www.jstor.org/stable/j.ctvcm4gk0.

Keeling, M.J., Ross, J.V., 2008. On methods for studying stochastic disease dynamics. J. R. Soc. Interface 5 (19), 171–181. https://doi.org/10.1098/rsif.2007.1106. url: https://royalsocietypublishing.org/doi/abs/10.1098/rsif.2007.1106.

Kinyanjui, T.M., Pellis, L., House, T., 2016. Information content of household-stratified epidemics. Epidemics, 16: 17–26. ISSN 1755–4365. doi:10.1016/j.epidem.2016.03.002. url: http://www.sciencedirect.com/science/article/pii/S175543651630010X..

Kuehn, Christian, 2016. Moment Closure—A Brief Review, pages 253–271. Springer International Publishing, Cham. ISBN 978-3-319-28028-8. doi:10.1007/978-3-319-28028-8_13. url:https://doi.org/10.1007/978-3-319-28028-8_13..

Kullback, Solomon, Leibler, Richard, 1951. On information and sufficiency. Ann. Math. Stat., 22: 79–86. doi:10.1214/aoms/1177729694..

Mendler, Marc, Johannes Falk, Barbara Drossel, 2018. Analysis of stochastic bifurcations with phase portraits. PLOS ONE, 13 (4): 1–20. doi:10.1371/journal.pone.0196126. url:https://doi.org/10.1371/journal.pone.0196126..

Mubayi, Anuj, Christopher Kribs, Viswanathan Arunachalam, Carlos Castillo-Chavez, 2019. Chapter 5 – studying complexity and risk through stochastic population dynamics: Persistence, resonance, and extinction in ecosystems. In Arni S.R. Srinivasa Rao and C.R. Rao, editors, Integrated Population Biology and Modeling, Part B, volume 40 of Handbook of Statistics, pages 157 – 193. Elsevier, 2019. doi:10.1016/bs.host.2018.11.001. url: http://www.sciencedirect.com/science/article/pii/S0169716118300944..

Pellis, L., Ferguson, N.M., Fraser, C, 2009. Threshold parameters for a model of epidemic spread among households and workplaces. Journal of the Royal Society, Interface, 6 (40): 979–987. ISSN 1742–5689. doi:10.1098/rsif.2008.0493. url: https://europepmc.org/articles/PMC2827443..

Ross, Joshua V., Thomas House, Matt J. Keeling, 2010. Calculation of disease dynamics in a population of households. PLOS ONE, 5 (3): 1–9. doi:10.1371/journal.pone.0009666. url: https://doi.org/10.1371/journal.pone.0009666..

Van Kampen, N.G., 2007. Stochastic processes in physics and chemistry. North Holland.

Wu, Joseph T., Riley, Steven, Fraser, Christophe, Leung, Gabriel M., 2006. Reducing the impact of the next influenza pandemic using household-based public health interventions. PLOS Med., 3 (9): 1–9. doi:10.1371/journal.pmed.0030361. url: https://doi.org/10.1371/journal.pmed.0030361..