Analysis of a stochastic SIR epidemic on a random network incorporating household structure[1]

Frank Ball[2], David Sirl[3], Pieter Trapman[4]

Abstract

This paper is concerned with a stochastic SIR (susceptible → infective → removed) model for the spread of an epidemic amongst a population of individuals, with a random network of social contacts, that is also partitioned into households. The behaviour of the model as the population size tends to infinity in an appropriate fashion is investigated. A threshold parameter which determines whether or not an epidemic with few initial infectives can become established and lead to a major outbreak is obtained, as are the probability that a major outbreak occurs and the expected proportion of the population that are ultimately infected by such an outbreak, together with methods for calculating these quantities. Monte Carlo simulations demonstrate that these asymptotic quantities accurately reflect the behaviour of finite populations, even for only moderately sized finite populations. The model is compared and contrasted with related models previously studied in the literature. The effects of the amount of clustering present in the overall population structure and the infectious period distribution on the outcomes of the model are also explored.

Keywords: SIR epidemic; random social network; households; local and global contacts; threshold behaviour; clustering.

1 Introduction

1.1 Background

There has been considerable recent interest in epidemic models which incorporate departures from the classical assumption that the underlying population is a collection of homogeneous individuals mixing homogeneously. In this paper we analyse and discuss an

[1]Appeared in *Mathematical Biosciences* 224:53–73 (2010); see also erratum ibid. 225:81 (2010). http://www.elsevier.com/wps/find/journaldescription.cws_home/505777/

[2]School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK. Email frank.ball@nottingham.ac.uk

[3]School of Mathematical Sciences, University of Nottingham, University Park, Nottingham NG7 2RD, UK. Email david.sirl@nottingham.ac.uk

[4]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Heidelberglaan 100, 3584 CX Utrecht, Netherlands and Stochastics Section, Faculty of Sciences, Vrije Universitiet Amsterdam, De Boelelaan 1081, 1081 HV Amsterdam, Netherlands. Current Address: Department of Mathematics, Stockholm University, SE-106 91 Stockholm, Sweden. Email p.trapman@uu.nl

epidemic model which relaxes these assumptions by having the population partitioned into households, within which *local* infectious contacts occur, and by using a random graph to model contacts/social structures through which *global* infectious contacts might take place.

The idea of partitioning the population into households, within which infectious contacts are made at a relatively high rate, while maintaining casual, homogeneously-mixing contacts goes back to Bartoszyński [1] (see also Becker and Dietz [2] and Ball *et al.* [3]). Models where an epidemic spreads over a random network with a prescribed degree distribution have also received significant attention (see, for example, Andersson [4], Newman [5], Kenah and Robins [6] and Myers *et al.* [7]). These network models have also been extended by treating the random graph as a local contact structure and introducing casual, homogeneously mixing contacts (Kiss *et al.* [8] and Ball and Neal [9]). The spread of epidemics on a different type of random graph structure, where individuals belong to several groups (workplaces, homes, etc.) and can make infectious contact only with persons in one of the same groups as themselves, is discussed by Britton *et al.* [10].

In a recent paper, Ball *et al.* [11], the authors formulate and analyse a new SIR (susceptible $\rightarrow$ infective $\rightarrow$ removed) epidemic model featuring mixing on two levels: local contacts with individuals in the same household and global contacts with an individual's neighbours in a random network with specified degree distribution. In that paper, rigorous branching process approximations are developed, which lead to a threshold theorem determining whether a major outbreak is possible and the probability of such an outbreak, as well as results which allow one to determine the expected proportion of the population that will be infected by such a major outbreak. Though the analysis is exact only asymptotically as the number of households becomes large, it was demonstrated through extensive simulation that these results provide good approximations for moderately-sized populations.

Though the proofs in [11] are valid for general infectious period distributions, explicit calculations for evaluating the probability of a major outbreak are only done for two special cases of infectious period distribution. In addition, the paper [11] assumes that all households are of the same size, but we conjecture that the results still hold when household size varies under very mild conditions on the household size distribution. The paper [11] focuses on establishing rigorous limit theorems, based on couplings between processes associated with the epidemic and certain branching processes, which show that many quantities of interest in the model can be expressed, asymptotically, in terms of the criticality and extinction probabilities of these branching processes. In this paper we focus on the more applied aspects of analysing this model, using these more theoretical results to derive methods for finding properties of the model.

The remainder of the paper is organised as follows. In Section 1.2 we recap the details of the model—including allowing for variable household size. Then in Sections 2 and 3 we discuss the calculation of the main quantities of interest for the model, with attention given to the practicalities of calculating these quantities of interest as well as the more theoretical results that tell us how to do these calculations. In Section 4 we firstly compare the results of simulations of epidemics on finite populations to our analytical results for large populations, then we compare our model to both the standard households model and

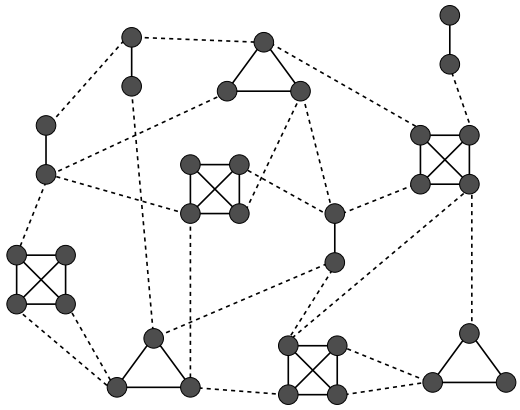

Figure 1: An example of a population of the type we analyse. Individuals are denoted by filled circles, with local (within-household) neighbours connected by solid lines and global neighbours by dashed lines.

the standard network model. We also investigate the amount of clustering present in the population structure in our model and look at the effect this has on the outcome of our model, as well as exploring the effect of the infectious period distribution on our model. Finally, we make some concluding comments in Section 5.

1.2 Model

The model we study consists of a finite, closed population of m households, of which m_n are of size n (for $n = 1, 2, \ldots$). We then construct the network describing the global structure of the population accoring to the 'configuration model' (see [12, 13]). This model works by assigning each individual in the population a number of 'half-edges' (that individual's degree in the global network) according to independent samples from some distribution D with $\mathbb{P}(D = k) = p_k$ $(k = 0, 1, \ldots)$ and then pairing these half-edges uniformly at random to form the edges of the graph describing the global network. An example of a (very small) population of the type we consider is given in Figure 1.

The epidemic evolves as follows. It starts with a single infectious individual chosen uniformly at random from the population. The infectious periods of different infectives are independent and identically distributed according to a non-negative random variable I, which we specify by its Laplace transform $\phi(\theta) = \mathbb{E}[\mathrm{e}^{-\theta I}]$ $(\theta \geq 0)$. Throughout its infectious period each infective makes infectious contacts with any given local neighbour at the points of a Poisson process with rate λ_L and with any given global neighbour at the points of a Poisson process with rate λ_G. If an individual so contacted is susceptible then it becomes infectious, otherwise the contact has no effect. An individual becomes removed at the end of its infectious period and plays no further part in the epidemic. The epidemic terminates when there is no infective remaining in the population. All infectious periods, global degrees and Poisson processes are assumed mutually independent.

For ease of exposition we have assumed that the epidemic begins with a single infective chosen uniformly at random from the population. However, our results are easily modified to accomodate several initial infectives and various ways of chosing them. Although our model does not include a latent period our results, which concern the final outcome of an epidemic, are invariant to very general assumptions concerning a latent period (see, e.g., Pellis *et al.* [14]). To be emphatic, λ_L and λ_G are *per-pair* rates, so that an infectious individual of degree d in a household of size n makes infectious contacts at total rate $d\lambda_G + (n-1)\lambda_L$.

Our results are asymptotic as the number of households $m \to \infty$. In order for the asymptotic analysis to be valid it is necessary to impose conditions on the household size and degree distributions. We assume that the mean μ_D and variance σ_D^2 of D are both finite. We also require that $m \to \infty$ in such a way that

$$m_n/m \to \rho_n \qquad (n = 1, 2, \ldots). \tag{1}$$

Suppose first that there is a maximal household size, $n_{\max}$ (i.e. for all m, $m_n = 0$ for all $n > n_{\max}$). Then $\sum_{n=1}^{n_{\max}} \rho_n = 1$, so $(\rho_1, \rho_2, \ldots)$ is a proper probability distribution, and (1) is sufficient for all the results of the paper to hold. The situation is more delicate when where is no maximal household size. In addition to (1) we assume that $\sum_{n=1}^{\infty} \rho_n = 1$ and $\sum_{n=1}^{\infty} n^2 \rho_n < \infty$. We also need to impose further conditions on the household size distributions for all sufficiently large m. These are met if instead we assume that, for each m, the sizes of the m households are chosen independently from the distribution given by $\mathbb{P}(H = n) = \rho_n$ $(n = 1, 2, \ldots)$. For simplicity, we make this assumption throughout the paper. For the properties we consider, assuming prescribed household sizes or independent and identically distributed household sizes from the appropriate limiting distribution yields identical results. However, this is not true in general; for example, the asymptotic variance of the size of a major outbreak is different under these two assumptions.

The requirement that $\sigma_D^2 < \infty$ ensures that any multiple edges and self-loops amongst individual become sparse in the global network as $m \to \infty$. The condition $\sum_{n=1}^{\infty} n^2 \rho_n < \infty$ ensures that, in the global network, multiple edges between households and household self-loops also become sparse as $m \to \infty$.

2 Early stages

2.1 Informal description of methods

As is prevalent in the literature on epidemic models, we analyse the early stages of an epidemic by way of a branching process approximation, which is exact in the limit as the population size becomes large in an appropriate manner. This leads to a threshold parameter which determines whether a major outbreak is possible; a major outbreak being the event that infinitely many individuals are ultimately infected in the limit as the population size tends to infinity. If such an outbreak is possible we can then further analyse the branching process to determine the probability of such a major outbreak.

The appropriate way to let the population size tend to infinity for our model involves letting m, the number of households, tend to infinity in the manner described in Section 1.2. The branching process we use to analyse the early stages of the epidemic approximates the number of households which become infected in the course of the epidemic. In the limiting process (which has infinitely many households) the infection of infinitely many households is equivalent to the infection of infinitely many individuals; so that non-extinction of the branching process corresponds to a major outbreak of the epidemic.

We now describe the branching process we use to approximate the number of infected households. Note that because we are interested only in the final outcome of the epidemic and not its precise time evolution we can think of the epidemic as evolving in the following way. We first consider the epidemic spreading only within the household containing the initial infective (the local epidemic that it initiates) and then consider the global infectious contacts made by each individual infected by the local epidemic. In the early stages of the epidemic it is highly likely that these global contacts are all with individuals in distinct households, this being critical for the branching approximation. We then let each newly infected household proceed in the same manner of local epidemic followed by global infections. Again in the early stages it is highly likely that these global infectious contacts are with individuals in distinct, previously uninfected households. We can view this as a branching process if we consider the households infected by a local epidemic initiated by a single infective within a typical household to be the children of that household. It is proved in Ball *et al.* [11] that (when the household size is fixed) this approximation becomes exact as $m \to \infty$, in that the size of the epidemic process converges in distribution to the total progeny of the approximating branching process.

To formally describe the (Galton-Watson) branching process we need to specify the offspring distribution. To this end we define a random variable describing the total number of global neighbours infected by the members of a household with a single initial infective. However, note that this offspring distribution is different for the first generation of the branching process than for all subsequent generations. This is because in the second and subsequent generations the initial infective in a household has been infected by one of its global neighbours, so the number of uninfected neighbours of this individual has the same distribution as $\tilde{D} - 1$. Here, the size-biased degree distribution $\tilde{D}$ is defined by $\mathbb{P}(\tilde{D} = k) = \tilde{p}_k = kp_k/\mu_D$ $(k = 1, 2, \ldots)$ and is the degree distribution of an individual from which a half-edge chosen uniformly at random emanates (under this sampling a given individual of degree k is k times as likely to be chosen as a given individual of degree 1). However, in the first generation the initial infective is the initial infective in the whole population, and the degree distribution of this individual is distributed as D, because the initial infective is chosen uniformly at random from the population. We denote by $\tilde{C}$ the random variable describing the number of global neighbours infected by the members of a household with a single initial infective where the initial infective is infected by a global contact and by C the corresponding quantity where the initial infective is chosen uniformly at random from the population. The random variable C thus describes the offspring distribution for the first generation of the approximating branching process and

$\tilde{C}$ that of all subsequent generations.

Whether or not the epidemic can 'take off' is then determined by the criticality of the branching process, i.e. the value of $R_* = \mathbb{E}[\tilde{C}]$. From the discussion in the previous paragraph we see that this is the average number of households infected by a typical infected household. In light of the discussions of Becker and Dietz [15] and Goldstein *et al.* [16] regarding different kinds of reproduction numbers/threshold parameters for epidemic models with household structure, this is but one of several possible threshold parameters for our model. We choose to use this particular threshold parameter due to the relative simplicity of the calculations involved. Calculating R_* is the focus of the next subsection, and the following subsection considers the calculation of the probability generating functions of C and $\tilde{C}$, from which we can obtain the probability of a major outbreak.

2.2 Threshold parameter

To calculate R_* we firstly condition on the size of the household that the globally infected individual is in, i.e.

$$\mathbb{E}[\tilde{C}] = \sum_{n=1}^{\infty} \tilde{\rho}_n \, \mathbb{E}[\tilde{C}^{(n)}], \tag{2}$$

where $\tilde{C}^{(n)}$ is the random variable $\tilde{C}$ conditional on the household being of size n and $\tilde{\rho}_n$ is the size-biased household size distribution, given by $\tilde{\rho}_n = n\rho_n / \sum_{j=1}^{\infty} j\rho_j$. This size-biasing arises because an individual chosen uniformly at random from the population is in a household of size n with probability proportional to $n\rho_n$. Then we decompose $\tilde{C}^{(n)}$ into the number of global infections emanating from each member of the household, i.e.

$$\tilde{C}^{(n)} = C_0 + \sum_{i=1}^{n-1} \chi_i C_i, \tag{3}$$

where we have labelled the individuals in the household $0, 1, \ldots, n-1$, with individual 0 being the globally infected initial infective (the *primary* infective in the household), χ_i is the indicator of the event that individual i is infected by the local epidemic initiated by the primary infective (i.e. $\chi_i = 1$ if i is so infected and 0 otherwise) and C_i is the number of global infections made by individual i. By symmetry, the random variables $(C_1, \chi_1), (C_2, \chi_2), \ldots, (C_{n-1}, \chi_{n-1})$ have the same distribution. Also, for each $i = 1, 2, \ldots, n-1$, whether individual i is infected by the local epidemic is clearly independent of individual i's infectious period if it becomes infective, so χ_i and C_i are independent (although whether or not a given individual is infected does depend on the infectious period of other individuals in its household). Thus, taking expectations of (3),

$$\mathbb{E}[\tilde{C}^{(n)}] = \mathbb{E}[C_0] + \mathbb{E}[T^{(n)}] \, \mathbb{E}[C_1], \tag{4}$$

where $T^{(n)} = \sum_{i=1}^{n-1} \chi_i$ is the final size of the local epidemic amongst the initial susceptibles (the *secondary* individuals) in the household.

The expectation of each C_i can be determined by conditioning on individual i's infectious period (I_i) and the number of uninfected neighbours it has in the random graph (K_i). All infectious periods have the same distribution, and K_1 has the same distribution as D (the specified degree distribution of the random graph); however the distribution of K_0 is the same as $\tilde{D}-1$, where $\tilde{D}$ is the size-biased degree distribution described in Section 1.2 and discussed in Section 2.1 and the '-1' accounts for the fact that one of the primary individual's neighbours is already infected—the one that infected it. Now note that, conditional on I_i and K_i, individual i makes infectious contacts with each of its K_i susceptible neighbours independently at the points of Poisson processes of rate λ_G for a time I_i. Thus, because a Poisson process with rate λ_G has no points before time I_i with probability $1-e^{-\lambda_G I_i}$, $C_i \,|\, I_i, K_i \sim \mathrm{Bin}(K_i, 1-\mathrm{e}^{-\lambda_G I_i})$. (If X is a random variable, n a non-negative integer and $p \in [0,1]$ then $X \sim \mathrm{Bin}(n,p)$ means that X has a binomial distribution with n trials and success probability p, with the convention that X is identically zero if $n=0$.) Therefore, recalling that $\phi(\theta) = \mathbb{E}[\mathrm{e}^{-\theta I}]$,

$$\mathbb{E}[C_i] = \mathbb{E}[\mathbb{E}[C_i \,|\, I_i, K_i]] = \mathbb{E}[K_i(1-\mathrm{e}^{-\lambda_G I_i})] = \mathbb{E}[K_i](1-\phi(\lambda_G)).$$

Thus we immediately have $\mathbb{E}[C_1] = \mu_D(1-\phi(\lambda_G))$. In addition, it follows from the definition of $\tilde{D}$ that $\mathbb{E}[\tilde{D}] = \mathbb{E}[D] + \mathrm{Var}[D]/\,\mathbb{E}[D]$, whence $\mathbb{E}[C_0] = (\mu_D + \sigma_D^2/\mu_D - 1)(1-\phi(\lambda_G))$. Substituting these results into (4) yields, with $\mu_{T^{(n)}} = \mathbb{E}[T^{(n)}]$,

$$\mathbb{E}[\tilde{C}^{(n)}] = \left(\mu_D(\mu_{T^{(n)}}+1) + \frac{\sigma_D^2}{\mu_D} - 1\right)(1-\phi(\lambda_G)),$$

and then substituting into (2) leads to the expression

$$R_* = \mathbb{E}[\tilde{C}] = (1-\phi(\lambda_G))\left(\mu_D(\mu_T+1) + \sigma_D^2/\mu_D - 1\right), \tag{5}$$

where $\mu_T = \sum_{n=1}^{\infty} \tilde{\rho}_n \mu_{T^{(n)}}$ is the (size-biased) mean within-household final size.

Unless the household sizes are all very small we need to evaluate R_* numerically, there being no simple expression for $\mu_{T^{(n)}}$. We formulate an expression for $\mu_{T^{(n)}}$ in terms of Gontcharoff polynomials, first introduced to the study of epidemic models by Daniels [17]. Given a parameter sequence of real numbers $U = (u_i,\, i = 0, 1, \ldots)$, the Gontcharoff polynomials $(G_k(x \,|\, U),\, k = 0, 1, \ldots)$ are defined by $G_0(x \,|\, U) = 1$ and the recurrence

$$G_k(x \,|\, U) = \frac{x^k}{k!} - \sum_{j=0}^{k-1} \frac{u_j^{k-j}}{(k-j)!} G_j(x \,|\, U), \qquad k = 1, 2, \ldots.$$

Note that $G_k(x \,|\, U)$ depends on U only through its first k entries. From Corollary 3.3 of Lefèvre and Picard [18] (cf. Ball [19, equations (2.25) and (2.26)]) one can deduce easily that

$$\mu_{T^{(n)}} = n - 1 - \sum_{i=1}^{n-1} (n-1)_{[i]} q_i^{n-i} G_{i-1}(1 \,|\, V), \tag{6}$$

where $a_{[i]}$ is the falling factorial $a!/(a-i)!$, $q_i = \phi(i\lambda_L)$ and $V = (q_{i+1},\, i = 0, 1, \ldots)$.

2.3 Major outbreak probability

As indicated at the end of Section 2.1, we approximate the probability of a major outbreak by the probability that the above approximating branching process for the early stages of the epidemic spread avoids extinction. In order to calculate this probability we require the probability generating functions (PGFs) of the offspring distributions. Calculating the PGFs $f_C(s) = \mathbb{E}[s^C]$ and $f_{\tilde{C}}(s) = \mathbb{E}[s^{\tilde{C}}]$ is complicated somewhat by the fact that, unless the infectious period is constant, the numbers of successful global contacts emanating from different individuals in the same household are not independent. For example, a large number of contacts emanating from one individual suggests that it had a long infectious period, which in turn increases the chance that any given other individual in the household was infected and thus increases the number of global infectious contacts it might make. However, we can apply the theory of so-called 'final state random variables' developed by Ball and O'Neill [20] to find the PGFs f_C and $f_{\tilde{C}}$. We present the background to this and the proof of the following result in Appendix A.

Theorem 1 *The PGF of C is given by $f_C(s) = \sum_{n=1}^{\infty} \tilde{\rho}_n f_{C^{(n)}}(s)$, where $C^{(n)}$ is the random variable C conditioned on the household size being n, which has PGF*

$$f_{C^{(n)}}(s) = \sum_{j=0}^{n-1} (n-1)_{[j]} \psi_0(s,j) \psi_1(s,j)^{n-1-j} G_j(1 \,|\, U^*), \qquad 0 \le s \le 1.$$

Here, for $l = 0, 1$, $j = 0, 1, \ldots$ and $0 \le s \le 1$,

$$\psi_l(s,j) = \sum_{i=0}^{\infty} \frac{(1-s)^i \phi(j\lambda_L + i\lambda_G)}{i!} f_{K_l}^{(i)}(s), \tag{7}$$

where $f_{K_l}^{(i)}$, $i = 0, 1, \ldots$, is the ith derivative of the PGF f_{K_l} and $G_j(x \,|\, U^*)$, $j = 0, \ldots, n-1$, are Gontcharoff polynomials with parameter sequence $U^* = (u_i^* = \psi_1(s,i),\ i = 0, 1, \ldots)$.

Remarks. 1. In an obvious notation, the same conditioning on household size for the PGF of the random variable $\tilde{C}$ yields $f_{\tilde{C}}(s) = \sum_{n=1}^{\infty} \tilde{\rho}_n f_{\tilde{C}^{(n)}}(s)$; and the PGF of $\tilde{C}^{(n)}$ differs from that of $C^{(n)}$ only in that ψ_0 is different because, as discussed previously, the distribution of the number of susceptible neighbours of the primary infective in the household in the first generation differs from that of subsequent generations. Thus, we always (i.e. for the calculation of both f_C and $f_{\tilde{C}}$) have $K_1 \stackrel{\mathcal{D}}{=} D$ (here and henceforth $A \stackrel{\mathcal{D}}{=} B$ means that the random variables A and B have the same distribution); for the first generation $K_0 \stackrel{\mathcal{D}}{=} D$ (and thus $\psi_0 = \psi_1$) and for subsequent generations $K_0 \stackrel{\mathcal{D}}{=} \tilde{D} - 1$.

2. Some care needs to be taken with the expression (7) when $s = 1$, as $f_{K_l}^{(i)}(1)$ may be infinite for $i > 0$ (i.e. if $\mathbb{E}[K_l^i] = \infty$). If this is the case then we interpret $0 \times \infty$ as 0 and 0^0 as 1, so that only the $i = 0$ term of the sum is non-zero.

We note here that it easily follows from the definition of $\tilde{D}$ that $f_{\tilde{D}-1}(s) = f_D^{(1)}(s)/\mu_D$, so $f_{\tilde{D}-1}^{(i)}(s) = f_D^{(i+1)}(s)/\mu_D$. Also worthy of note is the fact that D and $\tilde{D}-1$ have the same distribution if and only if D has a Poisson distribution.

In practice we can often evaluate the PGF of the random variable $C^{(n)}$ or $\tilde{C}^{(n)}$ numerically using Theorem 1 exactly as stated. This computation is relatively straightforward, the main potential problem being approximating infinite sums in the expression for $\psi_\cdot(\cdot,\cdot)$. If a closed form expression for a derivative of arbitrary order of f_D is available then it is quite efficient to use a finite truncation of (7) to approximate $\psi_l(s,j)$. However, it is possible for numerical problems to arise due to both $f_{K_l}^{(i')}(s)$ and $i'!$ becoming very large and $(1-s)^{i'}$ and $\phi(j\lambda_L + i'\lambda_G)$ becoming very small when i' is large but not large enough that the sum (7) truncated at $i = i'$ is a good approximation of the infinite sum. If such problems arise, or if there is no closed form expression available for $f_{K_l}^{(i)}$, then one needs to approximate $\psi_l(s,j)$ using a form along the lines of (20) of Appendix A, taking care to avoid numerical problems such as those just mentioned. Using these methods we can then approximate σ, the smallest root of $f_{\tilde{C}}(s) - s$ in $[0,1]$ and then the probability of a major outbreak $p_{\text{maj}} = 1 - f_C(\sigma)$ numerically.

To be emphatic, f_C is obtained from Theorem 1 by putting K_0 as D (or possibly some other distribution if we make different assumptions about how the epidemic is initiated) and $f_{\tilde{C}}$ by putting K_0 as $\tilde{D}-1$, whilst K_1 is D in both cases.

3 Final outcome

3.1 Informal description of methods

We now consider the final outcome of an epidemic by examining the expected relative final size of a major outbreak. By relative final size we mean the proportion of the population ultimately infected, thus we ask the question 'when a positive proportion of individuals are ultimately infected, what, on average, is that proportion?' Again our analysis is of the $m \to \infty$ limiting epidemic process, for which we find the probability that an individual chosen uniformly at random from the initial susceptibles is ultimately infected in the event of a major outbreak. This probability is equal to the asymptotic mean proportion of the population (individuals, not households) that are ultimately infected by a major outbreak. This quantity serves as our approximation of the expected proportion infected in a major outbreak in a finite population. (Although not proved in Ball *et al.* [11], the numerical results presented in Section 4.1 and similar results for other epidemic models, e.g. Andersson and Britton [21, Theorem 4.2] and Ball *et al.* [3, Section 4.2], suggest that the variance of the relative final size of a major outbreak converges to 0 as $m \to \infty$.) We determine the probability that a given individual is infected by considering (the size of) its *susceptibility set*. This concept has proved a fruitful framework within which to study the final outcome of epidemics where individuals interact with each other in more than one way (see, for example, Ball and Lyne [22] and Ball and Neal [23]).

The idea behind susceptibility sets is that for each individual in the population we

can, by sampling from the infectious period distribution and then the relevant Poisson processes, make a (random) list of other individuals it would make infectious contact with if it was to be infected itself. We then construct a digraph (directed graph) based on these lists, where the nodes represent individuals and we put a directed arc from i to j when i would make infectious contact with j were i to become infected, i.e. if j is in i's list. The susceptibility set of individual i consists of those individuals from which there exists a path to i in the digraph (including i itself). An individual i will become infected if and only if the initial infective is in i's susceptibility set. The probability of this occuring is related to the size of i's susceptibility set, which for our purposes is the number of *households* it intersects. We also require the notion of an individual's *local* susceptibility set, the portion of an individual's susceptibility set that arises by considering only local (within-household) infectious contacts. When we discuss the size of a local susceptibility set, however, we mean the number of *individuals* it contains.

Again we can make progress in the asymptotic case as $m \to \infty$, and we can construct the susceptibility set of an individual in such a way that its size can be coupled with an approximating branching process. The first generation of this branching process consists of the households containing an individual that makes infectious contact with a member of individual i's local susceptibility set (we call these individuals the *primary* individual in their household, similarly to the primary individuals in the forward process). Subsequent generations then consist of those households with individuals who make infectious contact with the local susceptibility set of any of the primary individuals of the previous generation. (One can think of this process growing through the population by following the arcs backwards in the random digraph described above; whereas the process we looked at to analyse the early stages of an outbreak follows these arcs forwards.) It follows from the methods used to construct the random graph of global contacts that, in the early stages of this 'growth', the individuals that join the susceptibility set are highly likely to be in households that are not already in it, enabling a branching process approximation to be found for the number of households in an individual's susceptibility set.

The offspring distribution for our branching process approximating the size of the susceptibility set of an individual is therefore the number of global neighbours of a household that, were they infected, would make infectious contact with the members of the local susceptibility set of the primary individual in that household. In the first generation the primary individual is an initial susceptible chosen uniformly at random and in subsequent generations the primary individuals are individuals who have joined the susceptibility set by means of making a global contact with an individual who is already in the susceptibility set—this affects the degree of the primary individual in the same way as in the previous section. Thus it seems plausible that we can calculate (as the complement of the extinction probability of a branching process) the probability that an initial susceptible has an asymptotically infinite susceptibility set as $m \to \infty$.

The connection between asymptotically infinite susceptibility sets and ultimate infection in the event of a major outbreak is quite complicated, so we do not attempt to describe it here. Very loosely, we find that, as $m \to \infty$, in the event of a major outbreak the

initial susceptible in question is ultimately infected if and only if its susceptibility set is asymptotically infinite. More detail of this connection, as well as a summary of these ideas concerning susceptibility sets, their approximation with branching processes and their relation to ultimate infection in the event of a major outbreak with more technical detail can be found in Section 3.2 of Ball *et al.* [11]. We next look in detail at the analysis of the approximating branching process described here and thus the calculation of the expected proportion of initial susceptiblesinfected by a major outbreak.

3.2 Expected relative final size of a major outbreak

We have seen that in order to calculate the (asymptotic) expected relative final size of a major outbreak we must analyse the offspring distribution of the branching process outlined in the first paragraphs of the previous subsection. The 'offspring' of a household in which individual i is the primary individual—either the individual whose susceptibility set we are considering or one that has joined the susceptibility set of interest by way of a global contact—in this process are those households with a member who globally infects a member of i's local susceptibility set. In order to find the extinction probability of the branching process we must determine the PGF of the distribution of the number of these offspring. To this end, we denote by B this number of individuals directly leading to (in graph-theory parlance) the *local* susceptibility set of a given individual (say i) from outside i's household when i is the individual whose susceptibility set we are considering. Denote by $\tilde{B}$ the corresponding quantity when i is an individual that has joined the susceptibility set by making a global contact.

Theorem 2 *For $s \in [0,1]$, the PGFs of B and $\tilde{B}$ are given, respectively, by*

$$f_B(s) = \sum_{n=1}^{\infty} \tilde{\rho}_n f_D(1 - p_G + sp_G) f_{M^{(n)}}(f_D(1 - p_G + sp_G)) \tag{8}$$

and

$$f_{\tilde{B}}(s) = \sum_{n=1}^{\infty} \tilde{\rho}_n f_{\tilde{D}-1}(1 - p_G + sp_G) f_{M^{(n)}}(f_D(1 - p_G + sp_G)), \tag{9}$$

where the random variable $M^{(n)}$ is the size of the local susceptibility set of a typical individual residing in a household of size n, not counting that individual.

Proof. For the purposes of the proof let B be a random variable which could be either B or $\tilde{B}$—the differences in the calculation are only slight and are pointed out when they arise. The first step is to condition on the size of the household individual i is in, so

$$f_B(s) = \sum_{n=1}^{\infty} \tilde{\rho}_n f_{B^{(n)}}(s), \tag{10}$$

where $B^{(n)}$ is the quantity B conditioned on the household size of individual i being n. We then decompose $B^{(n)}$ into the number of global contacts made with each member of i's local susceptibility set, i.e.

$$B^{(n)} = B_0 + \sum_{j=1}^{M^{(n)}} B_j, \tag{11}$$

where B_j is the number of contacts made with individual j (again labelling the individuals within the household $0, 1, \ldots, n-1$, with 0 corresponding to the primary individual) and $M^{(n)}$ is the size of i's local susceptibility set, not counting i itself. (If $M^{(n)} = 0$ then i's local susceptibility set consists only of i itself and the sum in (11) is empty and equal to 0.) Now $B_j \mid K_j \sim \mathrm{Bin}(K_j, p_G)$, where K_j is the number of global neighbours of j not already in the susceptibility set and $p_G = 1 - \phi(\lambda_G)$ is the probability that a given global contact is made. We do not need to condition on the infectious period of individual j because the contacts we are considering come from other (distinct) individuals; the independence of the infectious periods of these individuals implies that the events that each of these individuals contacts j are also independent. In the first generation $K_j \overset{\mathcal{D}}{=} D$ for all $j = 0, 1, \ldots, n-1$, however in subsequent generations this holds only for $j = 1, 2, \ldots, n-1$ and $K_0 \overset{\mathcal{D}}{=} \tilde{D} - 1$. We first note that, by independence, $f_{B^{(n)}}(s) = \mathbb{E}[s^{B^{(n)}}] = \mathbb{E}[s^{B_0}]\,\mathbb{E}[s^{\sum_{j=1}^{M^{(n)}} B_j}]$. Now,

$$\begin{aligned}
\mathbb{E}[s^{B_0}] &= \mathbb{E}\left[\mathbb{E}\left[s^{B_0} \mid K_0\right]\right] \\
&= \mathbb{E}[(1 - p_G + sp_G)^{K_0}] \\
&= f_{K_0}(1 - p_G + sp_G),
\end{aligned}$$

where f_{K_0} is either f_D or $f_{\tilde{D}-1}$, as above. Similarly we have

$$\begin{aligned}
\mathbb{E}\left[s^{\sum_{j=1}^{M^{(n)}} B_j}\right] &= \mathbb{E}\left[\mathbb{E}\left[s^{\sum_{j=1}^{M^{(n)}} B_j} \mid M^{(n)}, K_1, K_2, \ldots, K_n\right]\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{M^{(n)}} (1 - p_G + sp_G)^{K_i}\right] \\
&= \mathbb{E}[(f_D(1 - p_G + sp_G))^{M^{(n)}}] \\
&= f_{M^{(n)}}(f_D(1 - p_G + sp_G)).
\end{aligned}$$

Thus,

$$\mathbb{E}[s^{B^{(n)}}] = f_{K_0}(1 - p_G + sp_G) f_{M^{(n)}}(f_D(1 - p_G + sp_G)), \tag{12}$$

so, now denoting the offspring distribution random variable for the first generation by B and for subsequent generations by $\tilde{B}$ and substituting (12) into (10), we get (8) and (9). □

In order to determine $f_{M^{(n)}}$ we use a result of Ball [24, Lemma 3.1] (see also Ball and Neal [23, Lemma 3.1], which gives the same result but not in terms of Gontcharoff

polynomials), where it is shown that the mass function of $M^{(n)}$ is

$$\mathbb{P}(M^{(n)} = k) = (n-1)_{[k]} q_{k+1}^{n-1-k} G_k(1 \mid V), \qquad k = 0, 1, \ldots, n-1,$$

where $G_k(V)$ is a Gontcharoff polynomial with V as in (6). Unfortunately this distribution appears not to admit a simple form for its PGF (unless $\mathbb{P}(I = c) = 1$ for some c, in which case $M^{(n)}$ has the same distribution as $T^{(n)}$, the final size of the corresponding single-household epidemic), but because $M^{(n)}$ has finite support its PGF is easily evaluated numerically from the mass function.

The formulae (8) and (9) enable us to calculate $f_B(s)$ or $f_{\tilde{B}}(s)$ for given s and thus it is a relatively simple matter to (numerically) find the probability of extinction $f_B(\xi)$ of this branching process, where ξ is the smallest solution of $f_{\tilde{B}}(s) = s$ in $[0, 1]$. The proportion of individuals infected by a major outbreak in a finite population is then approximately $z = 1 - f_B(\xi)$. Evaluating f_B and $f_{\tilde{B}}$ only requires evaluating f_D and $f_{\tilde{D}-1} = f_D'/\mu_D$ (i.e. no higher order derivatives of f_D), which is usually straightforward and free of potential numerical problems.

4 Model behaviour

In this section we investigate several aspects of our model. First we compare our asymptotic analytical results to those of simulations for finite populations. Then we compare our model with both the standard households model and the standard network model. Next we investigate some of the clustering properties of the population structure and look at how this affects the behaviour of the epidemic model and finally we look at the effect of the infectious period distribution on the predictions of our model. Throughout this section we use the notation H for a random variable representing the household size distribution. When the distribution of H has finite support we write $H \sim (\rho_1, \rho_2, \ldots, \rho_n)$, where $\rho_i = \mathbb{P}(H = i)$. In addition, we sometimes write $H \sim \text{Poi}^+(\mu)$ to denote the household size distribution being Poisson with parameter $\mu > 0$ conditioned to be strictly positive. If $H \sim \text{Poi}^+(\mu)$, it is easy to show that $\mu_H = \mu/(1 - \mathrm{e}^{-\mu})$ and, for $h = 0, 1, \ldots$, $\mathbb{P}(\tilde{H} - 1 = h) = \mathrm{e}^{-\mu}\mu^h/h!$ ($\tilde{H} - 1$ being the number of local neighbours of an individual chosen uniformly at random from the population). Also note that letting $\mu \downarrow 0$ results in both the household size and the size-biased household size distributions having all their mass concentrated at 1, thus allowing us to recover the standard network model with all households of size 1. It is therefore consistent if we define $\text{Poi}^+(0)$ to be the distribution having unit mass concentrated at 1. We will also use the notation $\text{Gam}(\alpha, \theta)$ to denote a distribution with density function $x^{\alpha-1}\mathrm{e}^{-x/\theta}/\Gamma(\alpha)\theta^\alpha$ ($x \geq 0$), i.e. a gamma distribution with shape parameter α and scale parameter θ (which has mean $\alpha\theta$ and variance $\alpha\theta^2$).

4.1 Comparison with finite populations

Firstly we investigate whether the asymptotic values of the quantities of interest we have calculated give good approximations to these quantites in finite populations. In order to

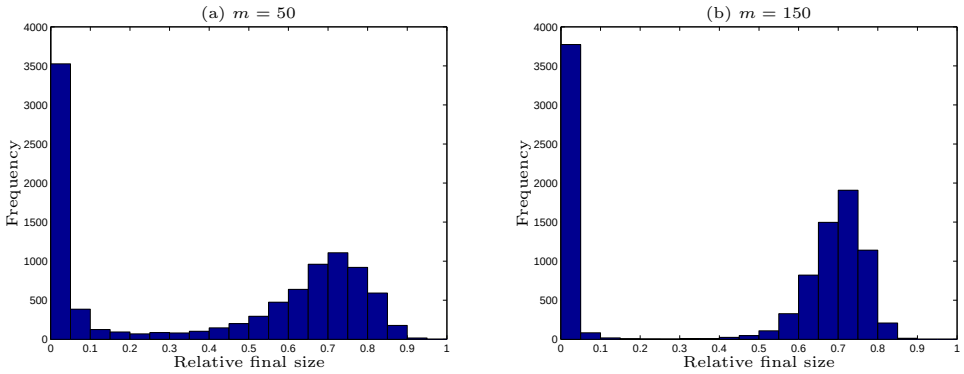

Figure 2: Histograms of relative final sizes from 10,000 simulations of our model showing the appearance of the distinction between minor and major outbreaks. The systems simulated are of 50 and 150 households respectively, with other parameters $H \sim (0.2, 0.25, 0.25, 0.25, 0.04, 0.01)$, $\lambda_L = 1$, $\lambda_G = 1/10$, $D \sim \text{Poi}(8)$ and $I \sim \text{Gam}(3, 1/3)$.

do this we run 10,000 simulations of the epidemic on finite populations and estimate the quantities of interest empirically (for increasing numbers of households m), then compare these estimates to the results of the asymptotic ($m = \infty$) calculations. Each simulation of our model involves generating a random population with the desired structure (both the household sizes, which are independent and identically distributed, and the network of possible global contacts) and running one epidemic on it; we do not simulate all of the epidemic processes on a single randomly generated network. Note that in small populations the determination of a cut-off for whether a particular final size constitutes a major or minor outbreak can be difficult—the population has to be moderately large for the distinction to be clear. We determine this cut-off by inspecting histograms of the relative final size for our simulations and we find that for m larger than about 100 a cut-off of 0.15 of the population size is appropriate for the parameter values we use. Figure 2 shows two such histograms which illustrate the 'overlap' between minor and major outbreaks for smaller population sizes and the increasingly obvious distinction as m becomes larger. If the parameters were chosen so that the major outbreaks are smaller then a lower cut-off would be necessary and a larger number of households would be needed for the distinction between minor and major outbreaks to be clear.

Figure 3 shows how these simulation-based estimates of the probability, p_{maj}, and the expected relative final size, z, of a major outbreak for finite popuations compare with the asymptotic value for two different degree distributions; one is Poisson and the other is a distribution with a power law tail that has mass function

$$p_k \propto \begin{cases} k_*^{-a}, & \text{for } k = 1, 2, \ldots, k_*, \\ k^{-a}, & \text{for } k = k_* + 1, k_* + 2, \ldots, \end{cases}$$

which we denote by $\text{Pow}(k_*, a)$. All parameters are chosen so that a major outbreak occurs

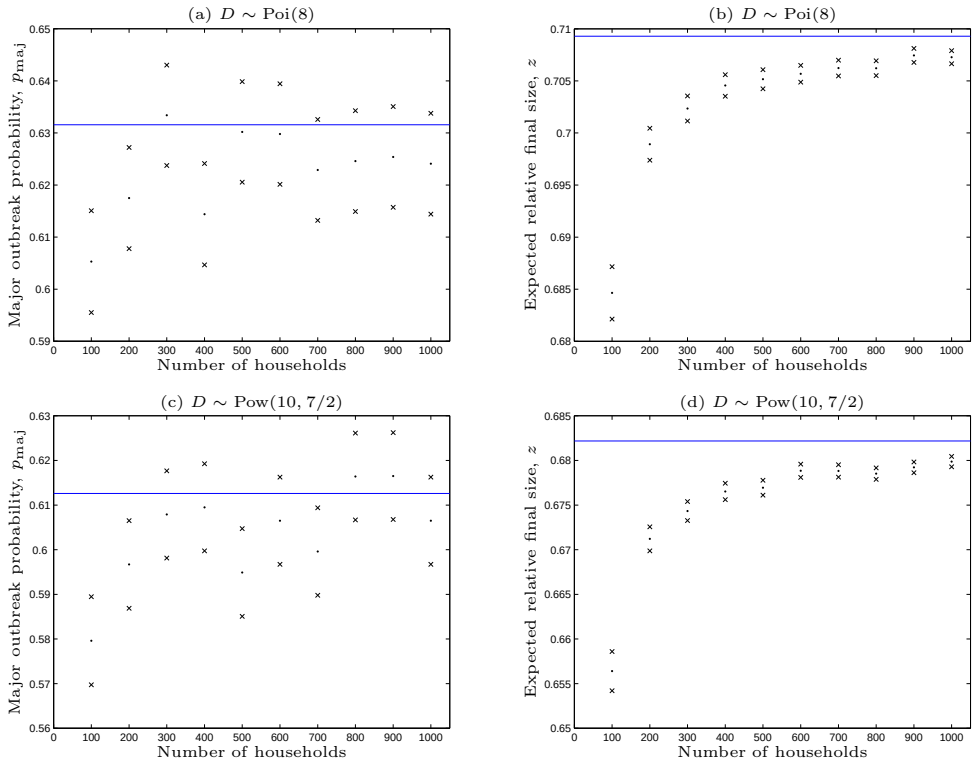

Figure 3: Comparison of simulation-based estimates of major outbreak probability and relative final size of a major outbreak for finite populations with asymptotic results (horizontal lines). The first two plots ((a) and (b)) have $D \sim \text{Poi}(8)$ (so $\mu_D = \sigma_D^2 = 8$ and $R_* \approx 2$) and the second two ((c) and (d)) have $D \sim \text{Pow}(10, 7/2)$ (so $\mu_D \approx 8$, $\sigma_D^2 \approx 96$ and $R_* \approx 3$). Other parameters are $H \sim (0.2, 0.25, 0.25, 0.25, 0.04, 0.01)$, $\lambda_L = 1$, $\lambda_G = 1/10$ and $I \sim \text{Gam}(3, 1/3)$.

with positive probability. The plots show point estimates of the quantities of interest and error bounds which are ± 2 standard errors (SEs) of the estimator. For the probability of a major outbreak, estimated as $\hat{p}_{\text{maj}}$, $\text{SE} = \sqrt{\hat{p}_{\text{maj}}(1 - \hat{p}_{\text{maj}})/n_0}$, where n_0 is the number of simulations. For the mean relative final size $\text{SE} = \hat{\sigma}_{\text{RFS}}/\sqrt{n_1}$, where $\hat{\sigma}^2_{\text{RFS}}$ is the sample variance of the relative final sizes of the n_1 major outbreaks. Note that the standard errors are appreciably smaller for p_{maj} than for z. This is because each simulation only gives one piece of information regarding whether a major outbreak occured, but each simulation that does yield a major outbreak gives one piece of information for each initial susceptible (i.e. whether or not it was ultimately infected). The latter are highly correlated but overall they contain more information than that available from one realisation of the forward process. We see from these plots that the asymptotic values are quite good approximations for the corresponding quantities in finite populations, even for moderately sized populations. We note also that the convergence of these quantities to their asymptotic limits is somewhat

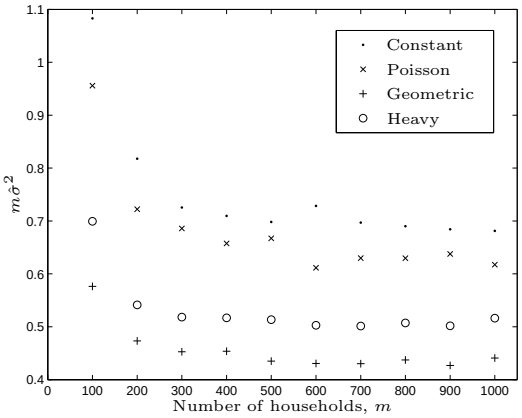

Figure 4: Plot showing apparent convergence of the variance of the relative final size of a major outbreak, scaled by m, to a constant for different degree distributions D. All of the degree distributions have mean 8, the heavy-tailed distribution being $\mathrm{Pow}(10, 7/2)$. The other parameters of the model are $H \sim (0.2, 0.25, 0.25, 0.25, 0.04, 0.01)$, $\lambda_L = 1$, $\lambda_G = 1/10$ and $I \sim \mathrm{Gam}(3, 1/3)$.

slower when the degree distribution D is heavy-tailed. Other simulations (not shown) suggest that changing the household size distribution H realistically (i.e. no heavy tails) has little effect on this convergence.

In connection with the parenthetical observations in the first paragraph of Section 3.1 regarding the possible convergence in distribution of the relative final size of a major outbreak, we also investigate the behaviour of the sample variance $\hat{\sigma}^2_{\mathrm{RFS}}$ of the relative final size of major outbreaks as a function of m. From Figure 4 it seems plausible that $m\hat{\sigma}^2_{\mathrm{RFS}}(m)$ $(= \hat{\sigma}^2_{\mathrm{FS}}(m)/m$, where $\hat{\sigma}^2_{\mathrm{FS}}(m)$ is the sample variance of the absolute, rather than relative final size) converges to a constant value as $m \to \infty$, lending credence to the suggestion made in Section 7 of Ball *et al.* [11] that the relative final size might satisfy a central limit theorem of the form $\sqrt{m}(RFS - \mathbb{E}\, RFS) \to \mathrm{N}(0, \sigma^2)$ with the variance σ^2 depending only on H, D, λ_G, λ_L and I.

4.2 Comparison to standard household model

An interesting question is whether or not our model necessarily behaves differently to the standard households model, where global contacts are homogeneously mixing rather than through a random network. We first determine conditions under which our networked household model (NHM) does give the same outcomes as the standard household model (SHM), then examine whether the outcomes of a natural limit of our model (as $\mu_D \to \infty$) converge to those of the SHM.

4.2.1 SHM and NHM with the same outcomes

A first point we need to be careful of is exactly what we mean by the NHM and SHM having the same outcomes. The analyses we can perform are not on the models directly, but on the branching process approximations of them (valid for large populations). We say that the models are effectively the same if the corresponding approximating branching processes (both forward and backward) have the same distribution in both models (though these distributions may be different for the forward and backward processes), which is equivalent to their offspring distributions having the same distribution. To simplify the presentation we assume that the household size distributions are the same, so it is sufficient to consider the case where the household size n is fixed. We denote the parameters of the NHM by I, D, λ_G and λ_L and those of the SHM by I', λ'_G and λ'_L. Note that although λ_L and λ'_L have the same interpretation, λ'_G is a *total* contact rate whilst λ_G is a per-pair contact rate. For the sake of simplicity we look first at the backward processes.

We observe first that the decomposition (11) of the random variable for the offspring in the backward branching processes in the NHM is $B^{(n)} = B_0 + \sum_{j=1}^{M^{(n)}} B_j$, where $B_0, B_1, \ldots, B_{n-1}$ are independent, with $B_1, B_2, \ldots, B_{n-1} \sim \mathrm{Bin}(D, p_G)$ and $B_0 \sim \mathrm{Bin}(D, p_G)$ or $B_0 \sim \mathrm{Bin}(\tilde{D} - 1, p_G)$ according as we are looking at the first or subsequent generations. It follows from the homogeneously mixing nature of the global contacts in the SHM that the corresponding decomposition, $B^{(n)\prime} = B'_0 + \sum_{j=1}^{M^{(n)\prime}} B'_j$, has $B'_0, B'_1, \ldots, B'_{n-1}$ as independent and identically distributed $\mathrm{Poi}(\lambda'_G \mu_{I'})$ random variables. By considering the PGFs of $B^{(1)}$ and $B^{(1)\prime}$, and in particular their factorial moments, the distributions of $B^{(1)}$ and $B^{(1)\prime}$ are different unless $\tilde{D} - 1 \sim \mathrm{Poi}(\mu_D)$ (for some μ_D), which implies that $D \sim \mathrm{Poi}(\mu_D)$, and $\mu_D p_G = \lambda'_G \mu_{I'}$. Note that if $D \sim \mathrm{Poi}(\mu_D)$ then $B_i \sim \mathrm{Poi}(\mu_D p_G)$. If we take $I \stackrel{\mathcal{D}}{=} I'$ and $\lambda_L = \lambda'_L$, then $M^{(n)} \stackrel{\mathcal{D}}{=} M^{(n)\prime}$ and the backward processes agree if and only if

$$\lambda'_G \mu_I = \mu_D (1 - \phi(\lambda_G)),$$

where $\mu_I = \mathbb{E}[I]$. Note that if this is the case then the expected relative final size of a major outbreak z is the same for the two models, as is the threshold parameter R_*.

To examine what happens in the forward process (still assuming $D \sim \mathrm{Poi}(\mu_D)$), we first consider the case $n = 1$, so our NHM reduces to the standard network model and the SHM reduces to the basic homogeneously mixing model. Here the random variables that describe the number of offspring in the forward processes for the NHM and SHM are, respectively, $C \sim \mathrm{Poi}(\mu_D(1 - \mathrm{e}^{-\lambda_G I}))$ and $C' \sim \mathrm{Poi}(\lambda'_G I')$. For these distributions to coincide it is necessary and sufficient that $\mu_D(1 - \mathrm{e}^{-\lambda_G I}) \stackrel{\mathcal{D}}{=} \lambda'_G I'$. Furthermore, it is simple to verify that this implies that the backward processes also coincide. If we also consider households of size 2, then we can derive a contradiction unless the infectious periods are fixed. The argument is quite involved, so we present it in Appendix B.

Since the forward and backward processes for both models coincide when their infectious periods are constant, we have thus seen that, given a NHM, it is possible to construct a SHM which coincides in the sense that both the backward and forward branching process approximations of the two models have the same distribution if and only if D is Poisson and

I is fixed. Moreover, under these conditions, the forward and backward processes coincide if and only if $\lambda_L c = \lambda'_L c'$ and $\mu_D(1-\mathrm{e}^{-\lambda_G c}) = \lambda'_G c'$ (where c and c' are the constant infectious periods for the NHM and SHM, respectively). Similarly, given a SHM there exists a NHM that coincides in the above sense if and only if I is almost surely constant. In that case the degree distribution of the NHM is necessarily Poisson and the parameters must satisfy the same relations given above.

4.2.2 SHM as a limit of the NHM

Another, perhaps more natural, way that we might be able to recover the SHM from our NHM is by letting $\mu_D \to \infty$ and $\lambda_G \downarrow 0$ in such a way that $\lambda_G \mu_D \to \lambda'_G \in (0, \infty)$. Here we show that under certain circumstances we do recover the properties of the standard household model and give some indication of when and why this can fail to hold. In the calculations which follow in the rest of this section we either put $\lambda_G = \lambda'_G/\mu_D$ and let $\mu_D \to \infty$ or we put $\mu_D = \lambda'_G/\lambda_G$ and let $\lambda_G \downarrow 0$, according to which one yields the simplest arguments; we thus have $\mu_D \to \infty$ and $\lambda_G \downarrow 0$ with $\lambda_G \mu_D$ being equal to rather than tending towards λ'_G. For ease of presentation, any unspecified limit henceforth refers to this limit. Firstly we determine whether or not the threshold parameter R_* of the NHM converges to that of the SHM.

Recall that $\mu_I = \mathbb{E}[I]$ is the mean of the infectious period distribution. Application of l'Hôpital's rule shows that $\lim_{\lambda \downarrow 0} \lambda^{-1}(1 - \mathbb{E}[\mathrm{e}^{-\lambda I}]) = \mu_I$; note this holds even if $\mu_I = \infty$. Suppose that $\kappa = \lim_{\mu_D \to \infty} \sigma_D^2(\mu_D)/\mu_D^2$ exists. Then taking the appropriate limit in (5) shows that $R_* \to \lambda'_G \mu_I(\mu_T + 1 + \kappa)$. The corresponding threshold parameter of the SHM is $R_* = \lambda'_G \mu_I(\mu_T + 1)$ (see equation (3.31) of Ball *et al.* [3]). Thus, if $\mu_I < \infty$, the threshold parameter R_* for the limiting NHM coincides with that of the SHM if and only if $\kappa = 0$, i.e. $\sigma_D^2 = o(\mu_D^2)$. This holds if, for example, D is Poisson or constant, but not if D has a tail which is geometric or heavier.

Now we examine the possibility of the actual branching process approximations of the NHM converging to those of the SHM, i.e. the offspring distributions converging, rather than just the threshold parameter. Firstly we consider the forms of the PGFs describing the offspring distributions of the SHM that we hope to recover from letting $\mu_D \to \infty$ and $\lambda_G \downarrow 0$ in our model. The key feature of the SHM that we need to recover is that, in the forward process, the 'contribution' to C from each infected individual (i.e. the number of infectious global contacts it makes) is a Poisson random variable with (random) mean $\lambda'_G I$. Similarly for the backward process, in the SHM each individual in the susceptibility set is globally contacted by a Poisson number of individuals with mean $\lambda'_G \mu_I$.

We consider the backward process first as it is simpler. With reference to equation (11), we have $B_j \,|\, K_j \sim \mathrm{Bin}(K_j, 1 - \mathbb{E}[\mathrm{e}^{-\lambda_G I}])$ and we need B_j to tend to a Poisson random variable with mean $\lambda'_G \mu_I$, which is equivalent to requiring $\mathbb{E}[B_{j[i]}] \to (\lambda'_G \mu_I)^i$ for all $i = 0, 1, \ldots$ (see, for example, [25, Section 2.3.e]). Now, since $f_{B_j|K_j}(s) = (\mathbb{E}[\mathrm{e}^{-\lambda_G I}] + (1 - \mathbb{E}[\mathrm{e}^{-\lambda_G I}])s)^{K_j}$

and $\mathbb{E}[B_{j\,[i]} \,|\, K_i] = f^{(i)}_{B_j|K_j}(s)|_{s=1}$, we have (dropping the subscript j for cleanliness)

$$\begin{aligned}\mathbb{E}[B_{[i]}] &= \mathbb{E}[\mathbb{E}[B_{[i]} \,|\, K]] \\ &= \mathbb{E}[K_{[i]}(1 - \mathbb{E}[\mathrm{e}^{-\lambda_G I}])^i] \\ &= \frac{\mathbb{E}[K_{[i]}]}{\mu_K^i}\left(\frac{1 - \mathbb{E}[\mathrm{e}^{-\lambda'_G I/\mu_K}]}{1/\mu_K}\right)^i,\end{aligned}$$

which (assuming $\mu_I < \infty$) converges to $(\lambda'_G \mu_I)^i$ if and only if $\mathbb{E}[K_{[i]}]/\mu_K^i \to 1$, (which we write as $\mathbb{E}[K_{[i]}] \sim \mu_K^i$) or, equivalently, if and only if $\mathbb{E}[K^i] \sim (\mu_K)^i$. Note also that $\mathbb{E}[D^n] \sim \mu_D^n$ implies $\mathbb{E}[(\tilde{D}-1)^{n-1}] \sim \mu_{\tilde{D}-1}^{n-1}$ $(n = 1, 2, \ldots)$. Now observe that the random variables $B_0, B_1, \ldots, B_{M^{(n)}}$ and $M^{(n)}$ in (11) are mutually independent. It follows that

$$\mathbb{E}[D^i] \sim \mu_D^i \qquad (i = 1, 2, \ldots) \tag{13}$$

is a necessary and sufficient condition for the backward branching process of the NHM to converge to that of the SHM.

Turning now to the forward process, with reference to (3), we have $C_j \,|\, I_j, K_j \sim \mathrm{Bin}(K_j, 1 - \mathrm{e}^{-\lambda_G I_j})$ and we need $C_j \,|\, I_j$ to tend to a Poisson random variable with mean $\lambda'_G I_j$, which is equivalent to requiring $\mathbb{E}[C_{j\,[i]} \,|\, I_j] \to (\lambda'_G I_j)^i$ for all $i = 0, 1, \ldots$. Therefore, since $f_{C_j|I_j,K_j}(s) = (\mathrm{e}^{-\lambda_G I_j} + (1 - \mathrm{e}^{-\lambda_G I_j})s)^{K_j}$ and $\mathbb{E}[C_{j\,[i]} \,|\, I_j, K_j] = f^{(i)}_{C_j|I_j,K_j}(s)|_{s=1}$, we obtain that (again dropping the subscript j)

$$\begin{aligned}\mathbb{E}[C_{[i]} \,|\, I] &= \mathbb{E}[\mathbb{E}[C_{[i]} \,|\, I, K]] \\ &= \mathbb{E}[K_{[i]}(1 - \mathrm{e}^{-\lambda_G I})^i] \\ &= \frac{\mathbb{E}[K_{[i]}]}{\mu_K^i}\left(\frac{1 - \mathrm{e}^{-\lambda'_G I/\mu_K}}{1/\mu_K}\right)^i,\end{aligned}$$

which converges to $(\lambda'_G I)^i$ if and only if $\mathbb{E}[K_{[i]}] \sim \mu_K^i$. Recall from above that this condition is equivalent to $\mathbb{E}[K^i] \sim (\mu_K)^i$ and that $\mathbb{E}[D^n] \sim \mu_D^n$ implies $\mathbb{E}[(\tilde{D}-1)^{n-1}] \sim \mu_{\tilde{D}-1}^{n-1}$ $(n = 1, 2, \ldots)$. Let $\boldsymbol{I} = (I_0, I_1, \ldots, I_{n-1})$, recall (3) and note that, given $\boldsymbol{I}$, $(C_0, C_1, \ldots, C_{n-1})$ and $(\chi_0, \chi_1, \ldots, \chi_{n-1})$ are independent and $C_i \,|\, \boldsymbol{I} = C_i \,|\, I_i$ $(i = 0, 1, \ldots, n-1)$. It then follows from (3) that (13) is sufficient for the forward branching process approximation of the NHM to converge to that of the SHM.

It is straightforward to show that (13) holds in the cases where $D \equiv \mu_D$ and $D \sim \mathrm{Poi}(\mu_D)$, but not when $D \sim \mathrm{Geom}(1/(1+\mu_D))$ (here $D \sim \mathrm{Geom}(p)$ means that, for $k = 0, 1, \ldots$, $\mathbb{P}(D = k) = p(1-p)^k$). Investigating the latter case further, it can be shown that rather than each infected individual asymptotically making infectious contact with a Poisson distributed number of global neighbours with mean cI, a secondary individual in a household makes infectious contact with a geometrically distributed number of global neighbours with parameter $(1 + cI)^{-1}$ (mean cI), whilst a primary individual makes infectious contact with a number of global neighbours that has a negative binomial distribution

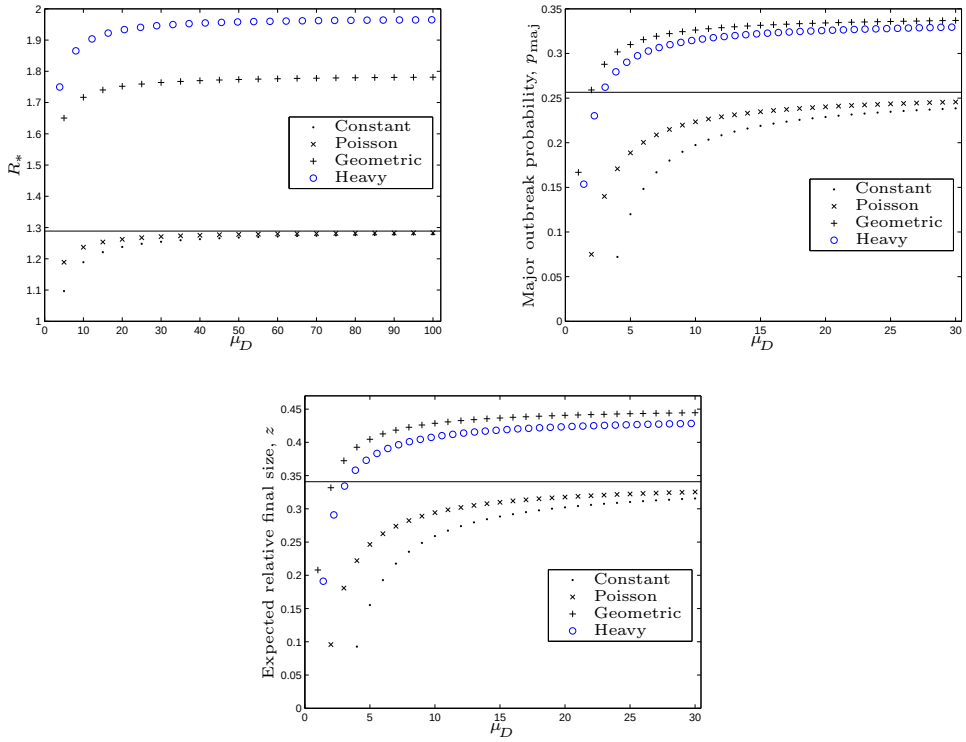

Figure 5: Comparison of asymptotic properties of our model with different degree distribution (and global contact rate) with the corresponding properties of the standard households model (solid horizontal lines). Other model parameters are $H \sim (0.2, 0.25, 0.25, 0.25, 0.04, 0.01)$, $\lambda_L = 2$, $I \sim \mathrm{Gam}(3/2, 1/3)$ and $\lambda_G = 1/\mu_D$. The heavy-tailed distribution is $\mathrm{Pow}(k, 7/2)$, for $k = 5, 10, \ldots, 100$ and $k = 1, 2, \ldots, 29$.

with parameters 2 and $(1 + cI)^{-1}$ (which has mean $2cI$). The same comments apply regarding the backwards process, with 'make contact with' replaced by 'contacted by' and I replaced by μ_I.

This has potentially important implications as just 'people having, on average, lots of contacts/friends', i.e. μ_D being large, in most circumstances does not imply that the standard household model is a good approximation for our network-household model. Figure 5 demonstrates numerically some of the conclusions of this section, that the SHM is a good approximation for our model when D is concentrated around its large mean μ_D, λ_G is small and $\mu_D \lambda_G = \lambda'_G$ is of moderate order. Note that, for large μ_D, the degree distributions that give a larger R_* generally give larger p_{maj} and z, except for the heavy-tailed distribution. Although not what one might initially expect, this is an artefact of the particular distributions we have used. For the heavy-tailed distribution to have a mean of moderate order it must have appreciable mass near zero, so there is a relatively large chance that the first few infectives will have few global neighbours and thus not spread infection.

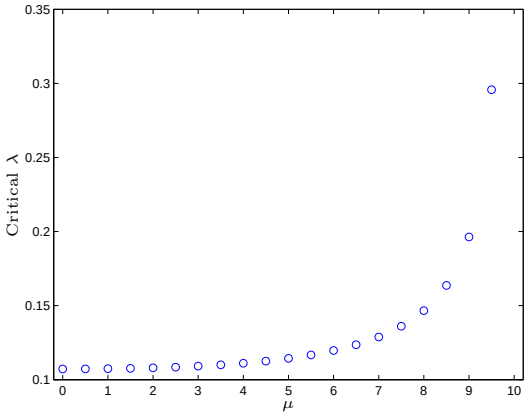

Figure 6: Plot of critical infection rate $\lambda = \lambda_L = \lambda_G$ for models with $H \sim \text{Poi}^+(\mu)$ and $D \sim \text{Poi}(10 - \mu)$ (so Q is always Poi(10)) for $\mu = 0, \frac{1}{2}, \ldots, 9\frac{1}{2}$ ('μ=0' corresponding to the standard network model). The other model parameter is $I \sim \text{Gam}(3, 1/3)$.

4.3 Comparison to standard network model

Another interesting comparison to make is that between our model and the standard network model (SNM) studied extensively by Newman [5] and others (notably Kenah and Robins [6]). This is the special case of our model where the households are all of size 1. A reasonable comparison to make here to gauge the effect of including households is to specify a total degree distribution Q and compare the critical value of λ in the SNM with degree distribution Q to the critical value of $\lambda = \lambda_L = \lambda_G$ in our model with the same overall degree distribution but with some household structure, i.e. with some household size and global degree distributions H and D such that $\tilde{H} - 1 + D \overset{\mathcal{D}}{=} Q$. Note that Q must be chosen carefully so that this is possible. An example of this is given in Figure 6, where we see that increasing the (average) household size at first makes little difference, but then as larger households become more prevalent a much higher infection rate is required to bring the epidemic above threshold in order to overcome the increase in the proportion of infectious contacts that are with individuals who have already been infected. This behaviour is somewhat more satisfactorily explained by looking to the amount of clustering present in the network of possible contacts, which is the subject of the next section.

4.4 Clustering

The clustering of a network measures whether the neighbours of an individual tend to also be neighbours of each other or, expressed in terms of friendships, whether a typical individual's friends are also friends with each other. We consider just one of the several measures of clustering in the literature. In this section we present our results only with heuristic arguments suggesting their truth; rigorous justifications are presented in Appendix C.

Firstly we note some simple facts about the population structure in our model. For large m the fraction of the vertices that are in a household of size h and have global degree d is well approximated by $\tilde{\rho}_h p_d$, by the law of large numbers. Such a vertex has degree $d+(h-1)$. Let $f_D(s)$ and $f_H(s)$ denote the PGFs of the global degree distribution and the household size distribution, respectively. The PGF of the size-biased household size distribution $\tilde{H}$ is given by $f_{\tilde{H}}(s) = \sum_{h=0}^{\infty} \tilde{\rho}_h s^h = \sum_{h=0}^{\infty} h\rho_h s^h/\mu_H = sf'_H(s)/\mu_H$, where we have assumed that $\mu_H = \mathbb{E}[H] < \infty$. The PGF of the *total* degree distribution Q is therefore, for large m, well approximated by $f_Q(s) = \sum_{j=0}^{\infty} q_j s^j = s^{-1} f_{\tilde{H}}(s) f_D(s) = f'_H(s) f_D(s)/\mu_H$.

A natural measure of clustering is the total number of ordered triangles divided by the total number of ordered triples (a triple being three vertices of which the second is a neighbour of both the first and third) in a network of m households [26, 27], say $\mathcal{C}^{(m)}$. We use the fact that the number of triangles not entirely in the same household is small. If m is large then the law of large numbers implies that the number of ordered triples that are not entirely in the same household per household is well approximated by $\mathbb{E}[H(2D(H-1)+D(D-1)]$, while the total number of triples per household is well approximated by $\mathbb{E}[H(D+H-1)(D+H-2)]$. We therefore have

$$\mathcal{C}^{(m)} \approx 1 - \frac{\mathbb{E}[H(2D(H-1)+D(D-1)]}{\mathbb{E}[H(D+H-1)(D+H-2)]}. \tag{14}$$

This formula becomes exact as $m \to \infty$ (see Appendix C), so we define the clustering coefficient $\mathcal{C}$ to be the right hand side of (14). Note that, because H and D are independent, the expectations in (14) are easily evaluated and

$$\mathcal{C} = 1 - \frac{f_H^{(1)}(1) f_D^{(2)}(1) + 2 f_H^{(2)}(1) f_D^{(1)}(1)}{f_H^{(1)}(1) f_D^{(2)}(1) + 2 f_H^{(2)}(1) f_D^{(1)}(1) + f_H^{(3)}(1)}. \tag{15}$$

Note that if H has infinite third moment and finite second moment, while D has finite second moment, then $\mathcal{C} = 1$. In this situation the total degree Q has infinite variance and in the SNM this implies $R_* = \infty$ (see, for example, [5]). However, in the NHM, equation (5) implies that, under these conditions on H and D, $R_* < \infty$. This apparent discrepancy is a consequence of the clustering in the overall contact network in the NHM. A similar phenomenon is observed (in different models to ours) in [26] and [27].

To investigate further the effect of varying clustering on our model we examine a situation where it is simple to vary the household size and global degree distributions whilst keeping the total degree distribution fixed. We do this by taking $H \sim \text{Poi}^+(\mu)$ and $D \sim \text{Poi}(\mu_D)$ with $\mu + \mu_D$ fixed, so that $\mu_H = \mu/(1-\mathrm{e}^{-\mu})$ and $\tilde{H} - 1 \sim \text{Poi}(\mu)$. In this situation it follows easily from (15) that $\mathcal{C} = \mu^2/(\mu+\mu_D)^2$. In Figure 7 we compare the properties of these different models in both the situation where $\lambda_L = \lambda_G$, so the changes in the quantities of interest reflect only the changes in the structure of the network of possible contacts (plots (a) and (b)) and also in the more realistic case $\lambda_L > \lambda_G$, where an increase (decrease) in clustering is accompanied by an increase (decrease) in the overall rate at which a typical infected individual makes infectious contacts (plots (c) and (d)). Plot (b) tells much the

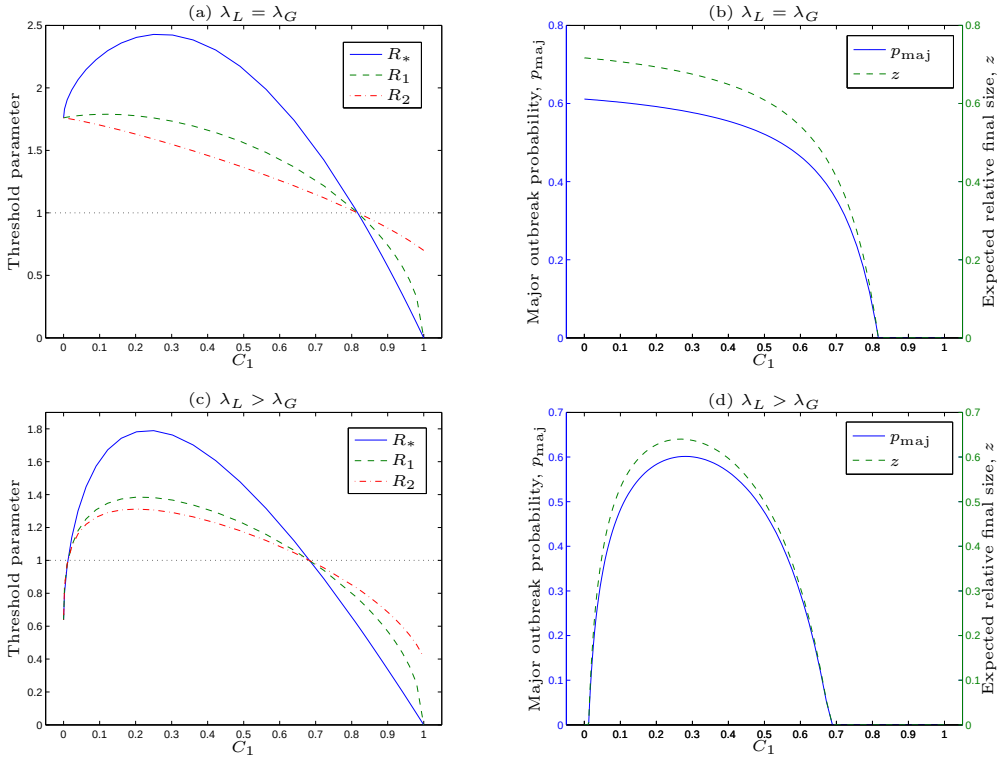

Figure 7: Plots of threshold parameters and the probability and expected relative final size of a major outbreak for networks with varying clustering by having $H \sim \text{Poi}^+(\mu)$ and $D \sim \text{Poi}(10 - \mu)$ (so Q is always Poi(10)), for $\mu_H \in [0, 10]$. Other parameters are $I \sim \text{Gam}(3, 1/3)$ and $\lambda_L = \lambda_G = 1/5$ for (a) and (b) and $\lambda_L = 1$ and $\lambda_G = 1/15$ for (c) and (d).

same story as Figure 6, that more clustering (i.e. larger households) results in more repeat contacts with individuals who are already infected, thus reducing both p_{maj} and z. Although the graph of R_* in plot (a) might appear to contradict this, we must remember that R_* is a household-to-household reproduction number and that the distribution of household sizes changes as $\mathcal{C}$ changes. A somewhat better comparison can be obtained by using a reproduction number which reflects the proliferation of infected individuals rather than households, as is the case with R_* (cf. Section 2.1, final paragraph). Two reproduction numbers which attempt to do this, R_1 and R_2, are therefore included in Figure 7. Details concerning their definition and interpretation are provided in Appendix D. The sometimes substantial difference between R_*, R_1 and R_2 serves as a reminder that simply comparing the reproduction numbers of different models can be misleading; though note that of course all of the reproduction numbers cross their critical value, 1, at the same time.

Plots (c) and (d) of Figure 7, however, demonstrate that the monotonic relationships of the probability and expected relative final size of a major outbreak to $\mathcal{C}$ do not necessarily obtain in the more realistic part of the parameter space where $\lambda_L > \lambda_G$. In this situation, whilst increased clustering results in each infectious contact having a reduced chance of being with a susceptible there are many more such contacts and so some increase in clustering can actually enhance the spread of the infection. However, when the clustering is already high this is outweighed by the fact that increasing the clustering further reduces the number of possible global contacts dramatically.

4.5 The effect of the infectious period distribution

Another aspect of our model that we investigate is its dependence on the choice of the infectious period distribution (IPD). A number of different IPDs are used in the literature, including the exponential distribution (in the so-called general stochastic epidemic and implicitly in most deterministic models), the gamma distribution and an almost surely constant infectious period. It has long been recognised that the exponential distribution, though mathematically convenient as it means the epidemic process is Markov, does not provide a realistic model of the infectiousness of real diseases (see, for example, Grossman [28] and Keeling and Grenfell [29]). A fixed infectious period also offers some mathematical advantages and is usually more realistic than an exponentially distributed one, but it still eliminates a potentially important source of randomness in an epidemic model.

To study the effect of the IPD we use the results of Kuulasmaa [30], in particular Theorem 2.1 of that paper. Trapman [26, Section 3] has applied Kuulasmaa's work to obtain similar results in a context where there is only one kind of contact rather than the two types—local and global—that we consider here. In Appendix E we prove the following theorem.

Theorem 3 *Let* $\mathrm{NHM}(H, D, \phi, \lambda_L, \lambda_G)$ *denote our epidemic model with household size distribution* H*, degree distribution* D*, infectious period distribution given by the Laplace transform* $\phi(\cdot)$ *and infection rates* λ_L *and* λ_G*. Let the epidemics* $\mathcal{E}$ *and* $\mathcal{E}'$ *be given by*

$\mathrm{NHM}(H, D, \phi, \lambda_L, \lambda_G)$ *and* $\mathrm{NHM}(H, D, \phi', \lambda'_L, \lambda'_G)$*, respectively, and suppose that*

$$\phi(i\lambda_L + j\lambda_G) \geq \phi'(i\lambda'_L + j\lambda'_G) \tag{16}$$

for all $i, j = 0, 1, \ldots$*. Then* $R_*(\mathcal{E}) \leq R_*(\mathcal{E}')$*,* $p_{\mathrm{maj}}(\mathcal{E}) \leq p_{\mathrm{maj}}(\mathcal{E}')$ *and* $z(\mathcal{E}) \leq z(\mathcal{E}')$*.*

Remarks. 1. When the conditions of this theorem obtain we say, somewhat loosely, that the process $\mathcal{E}'$ is an upper bound for $\mathcal{E}$ (or $\mathcal{E}$ a lower bound for $\mathcal{E}'$), in the sense that the stated measures of severity are all higher (or at least not smaller). Note that it is not necessarily the case that the final size of $\mathcal{E}'$ is stochastically larger than that of $\mathcal{E}$.

2. As outlined in Appendix E, we can also use Kuulasmaa's result to show that the probability that all individuals in a given subset of the initially susceptible individuals avoid infection in $\mathcal{E}$ is greater than the correponding probability in $\mathcal{E}'$. Thus, in particular, the probability that a given initially susceptible individual avoids infection is greater in $\mathcal{E}$ than in $\mathcal{E}'$. Summing over all such initially susceptible indivuduals then shows that the unconditional expected final size of $\mathcal{E}$ is smaller than that of $\mathcal{E}'$.

One way of using this result is to compare the properties of epidemic models with a relatively complicated infectious period distribution to those with an infectious period distribution that admits simpler computation, in order to obtain bounds on quantities of interest for the model with the more complicated IPD (cf. [11, Section 4.1]). For example, Trapman [26] uses Kuulasmaa's result to give upper and lower bounds on certain properties of an epidemic model with only one type of contact by comparing the model with an arbitrary infectious period to a similar model where the infectious period is either fixed or can only take the values 0 or ∞. Our theorem might also be used to compare the predicitions of two models which have different IPDs but are otherwise the same. For example, if one has in mind a value for the mean infectious period, how would the predictions of a modeller who specified an exponential IPD differ from those of a modeller who specified a fixed infectious period?

The upper bounding processes we derive have, as in Kuulasmaa's and Trapman's work, a constant infectious period, say c'. With this in mind, we want to show that

$$\phi(i\lambda_L + j\lambda_G) \geq \phi'(i\lambda'_L + j\lambda'_G) = \mathrm{e}^{-c'(i\lambda'_L + j\lambda'_G)} = \phi'(\lambda_L)^i \phi'(\lambda_G)^j, \tag{17}$$

for suitable constants c', λ'_L and λ'_G. However, since $\mathbb{E}[\mathrm{e}^{-(a+b)X}] \geq \mathbb{E}[\mathrm{e}^{-aX}]\,\mathbb{E}[\mathrm{e}^{-bX}]$ for any non-negative random variable X and constants $a, b > 0$ (see Kuulasmaa [30, Lemma 4.2]), we have (also using Jensen's inequality)

$$\begin{aligned}
\phi(i\lambda_L + j\lambda_G) &= \mathbb{E}[\mathrm{e}^{-I(i\lambda_L + j\lambda_G)}] \\
&\geq \mathbb{E}[\mathrm{e}^{-i\lambda_L I}]\,\mathbb{E}[\mathrm{e}^{-j\lambda_G I}] \\
&\geq (\mathbb{E}[\mathrm{e}^{-\lambda_L I}])^i (\mathbb{E}[\mathrm{e}^{-\lambda_G I}])^j \\
&= \phi(\lambda_L)^i \phi(\lambda_G)^j.
\end{aligned}$$

It thus follows that $\phi(\lambda_L) \geq \phi'(\lambda'_L) = \mathrm{e}^{-c'\lambda'_L}$ and $\phi(\lambda_G) \geq \phi'(\lambda'_G) = \mathrm{e}^{-c'\lambda'_G}$ (equivalently $p_L \leq p'_L$ and $p_G \leq p'_G$, where $p_k = 1 - \phi(\lambda_k)$ ($k \in \{L, G\}$) is the marginal probability that an infective individual infects a given type k neighbour) are together sufficient for (17).

A naive but reasonable guess at approximating a given epidemic process with a similar epidemic process with a constant infectious period is to simply replace the aribtrary infectious period I with a constant infectious period $c' = \mathbb{E}[I]$. It follows from Jensen's inequality that $\phi(\lambda_i) = \mathbb{E}[\mathrm{e}^{-\lambda_i I}] \geq \mathrm{e}^{-\lambda_i \mathbb{E}[I]} = \mathrm{e}^{-\lambda_i c'}$, so this does result in an upper-bounding epidemic process. Note that in terms of the marginal probabilities of infection, this results in $p_k \leq p'_k$, usually with strict inequality for both $k \in \{L, G\}$.

In order to get tighter bounds on the quantities of interest, we might try choosing the constant infectious period c' as small as possible so that $\phi(\lambda_k) \geq \mathrm{e}^{-\lambda_k c'}$ (equivalently, $p_k \leq p'_k$) for both $k \in \{L, G\}$, which amounts to chosing $c' = \max_k\{-\log\phi(\lambda_k)/\lambda_k\}$. Supposing that $c' = -\log\phi(\lambda_G)/\lambda_G$, we firstly have $\mathrm{e}^{-c'\lambda_G} = \exp(-(-\log\phi(\lambda_G)/\lambda_G)\lambda_G) = \phi(\lambda_G)$. Now, noting that $-\log\phi(\lambda_L)/\lambda_L \geq -\log\phi(\lambda_G)/\lambda_G$, we have

$$\mathrm{e}^{-c'\lambda_L} = \exp(-(-\log\phi(\lambda_G)/\lambda_G)\lambda_L) \leq \exp(-(-\log\phi(\lambda_L)/\lambda_L)\lambda_L) = \phi(\lambda_L).$$

Of course, if instead we have $c' = -\log\phi(\lambda_L)/\lambda_L$ the same argument holds with Ls and Gs interchanged and either way we have $\mathrm{e}^{-c'\lambda_L} \leq \phi(\lambda_L)$ and $\mathrm{e}^{-c'\lambda_G} \leq \phi(\lambda_G)$, so this process also gives an upper bound. In terms of the marginal probabilities of infection we now have $p_k \leq p'_k$, with equality for one $k \in \{L, G\}$ and strict inequality for the other (unless $\lambda_L = \lambda_G$), so one would expect this to be a better upper bound than the first upper bound discussed above where both inequalities are in general strict.

Finally, we can do better again by changing the local and global contact rates too, in such a way that $p_k = p'_k$, for both $k \in \{L, G\}$. We can do this by setting $c' = 1$ (any constant will do) and $\lambda'_k = -\log\phi(\lambda_k)$. It then follows that $\mathrm{e}^{-c'\lambda_k} = \mathrm{e}^{\log\phi(\lambda_k)} = \phi(\lambda_k)$ for both k.

Turning to the question of finding a lower bounding epidemic process, we are hindered somewhat by the fact that when we take the infectious period distribution to be zero or infinity, i.e. $\mathbb{P}(I = \infty) = 1 - \mathbb{P}(I = 0) = \pi$, the infection rates λ_L and λ_G have no influence on the epidemic from the viewpoint of the final size. To ensure that p'_L and p'_G are as large as possible but still respect the inequalities $\phi(\lambda_k) \leq \phi'(\lambda_k) = 1 - \pi'$ (equivalently $p'_k \leq p_k$), we must take $\pi' = \min_k\{p_k\} = 1 - \phi(\min_k\{\lambda_k\})$. It then easily follows that, for $i + j > 0$, $\phi(i\lambda_L + j\lambda_G) \leq 1 - \pi' = \phi'(i\lambda_L + j\lambda_G)$ and so this process with a zero or infinite infectious period gives the desired lower bound.

In the special case where all households are of size 1, i.e. the standard network model, we find that R_* and the expected relative final size of a major outbreak are the same for the best upper bounding process and the lower bounding process we have described. This is because these quantities depend on ϕ and λ_G only through $\phi(\lambda_G)$ and this is the quantity that we have matched up in order to ensure that the bounds are as good as possible. (Note λ_L is irrelevant here.) Similarly, if we consider households of size 2 then the threshold parameter R_* and the expected relative final size of a major outbreak are the same in any 'original' process with arbitrary IPD and the last upper bound described above, as these

quantities depend on ϕ, λ_L and λ_G only through $\phi(\lambda_L)$ and $\phi(\lambda_G)$ and both of these are matched in the approximating process. For any household size, however, the probability of a major outbreak is different from that of the original process in both of the bounding processes, since it is clear from equation (7) that p_{maj} depends on $\phi(i\lambda_L + j\lambda_G)$ for all $i, j = 0, 1, \ldots$ and in constructing the bounding processes we only match these quantities for one or both of $(i, j) = (0, 1), (1, 0)$. Note also that appreciable simplification obtains when $\lambda_L = \lambda_G$, for then the two marginal infection probabilities become one and so the second and third upper bounding processes become equivalent and also the lower bounding process achieves $p'_k = p_k$ for all links (since there is only one type of link.)

Examples of these bounds are plotted in Figure 8, where we have taken as an 'original' model one with an exponential IPD and then calculated the properties of it and of the bounding models we have just described; the upper bounds being numbered in the order that they are presented above. In all cases we see that the upper bounds appear in the plots in the order we expect from the observations concerning the marginal infection probabilities p_k associated with the bounding processes, the better bounds obtaining when the differences $p'_k - p_k$ are smaller. We see from the difference between columns (a) and (b) in the figure that as household sizes become larger the bounds get worse. This is in part because there are more values (i, j) for which the quantity $\phi(i\lambda_L + j\lambda_G)$ is not being matched up, so the approximation is worse. This also explains the fact that the approximations are much worse for p_{maj} than for z.

5 Discussion

In this paper we have shown how to analyse the potential spread of an SIR epidemic in a structured population incorporating household structure and a random graph model with a specified degree distribution to model potential global contacts. We have extended the results of Ball *et al.* [11] to incorporate variable household sizes and calculate the probability of a major outbreak for a general infectious period distribution. In addition we have discussed the numerical implementation of the methods we describe for calculating a threshold parameter and the probability and expected relative final size of a major outbreak in the limit as the number of households becomes large. We have seen that these asymptotic results give good approximations to the behaviour of our model in modestly sized finite populations and given numerical results suggesting that the relative final size in the event of a major outbreak satisfies a central limit theorem as $m \to \infty$, as conjectured in Section 7 of Ball *et al.* [11]. Further, we have compared our model with the standard households model and the standard network model for SIR-type infections and seen that some of the differences between the standard network model and ours can be viewed as a result of the different amount of clustering in the population structure. Moreover, when $\lambda_L = \lambda_G$ (in which case our model provides one way of introducing clustering into the standard network model), we have demonstrated that such clustering can appreciably decrease the spread of disease. We have also shown that the choice of infectious period distribution, for given marginal local and global contact probabilities, can have a very significant impact on both

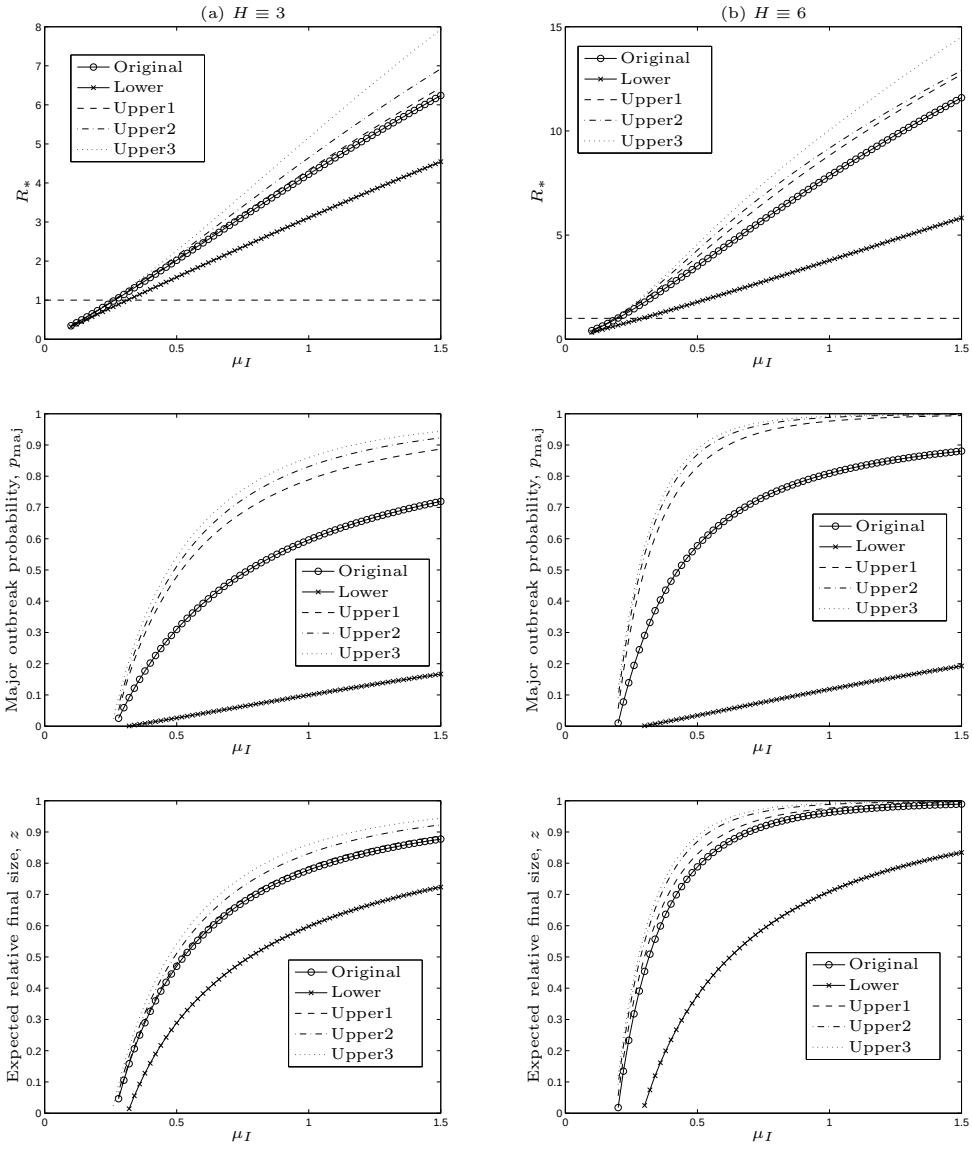

Figure 8: Plot of quantities of interest and their bounds for 'original' models with exponential infectious period with mean μ_I ranging from 0.1 to 1.5 and fixed household sizes (a) $H \equiv 3$ and (b) $H \equiv 6$. Other model parameters are $D \sim \mathrm{Geom}(8/9)$, $\lambda_L = 1$ and $\lambda_G = 1/5$. (In the lower left plot there is a difference between 'original' and 'upper1' but it is very small.)

the probability and size of a major outbreak.

One of the more obvious next steps to take is to analyse the effect of vaccination on the behaviour of our model, including the effect of different models for the action of the vaccine and different vaccine allocation methods—this work is currently in progress. In addition, our model can be generalised to allow one or more of the global degree distribution D and the contact rates λ_L and λ_G to depend on household size. These changes are in principle quite simple to incorporate because as soon as we take what is invariably the first step in our calculations—conditioning on the household size—these quantities are fixed and the calculations then proceed in exactly the same way as in this paper. It follows that the numerical computation of all of the quantities of interest will be only slightly more complicated but potentially much more time-consuming than in the situation where D and the contact rates are independent of household size. Another possible generalisation of our model which should still be amenable to analysis but will be much more complicated to implement numerically is to allow for correlations between the global degrees of individuals in the same household; again the nature of these correlations might depend on household size. Furthermore, the results of Ball and O'Neill [20] on final state random variables are stated in terms of multitype epidemics, so our results could be extended to this situation (e.g. for modelling differences between adults and children).

Another aspect of our model that could be generalised without making the analysis significantly more complex is the infectiousness profile of individuals over time. The model we present in this paper assumes a constant rate of infectiousness for a random time (the infectious period) but all that is important is the total infectiousness over this time period—the area under the (random) function $J_k(t)$, the rate at which type k contacts (local or global) are made t time units after an individual becomes infected. Currently $J_k(t) = \lambda_k \mathbb{1}(t \leq I)$, so $\int_0^\infty J_k(t)dt = I\lambda_k$, but in principle $J_k(\cdot)$ could be a random function drawn from say all integrable functions taking non-negative values on $[0, \infty)$. (This also emphasises the point that our results are insensitive to the inclusion of a latent period, as $\int_0^\infty J_k(t)dt = \int_0^\infty J_k(t-L)dt$, assuming that $J_k(t) = 0$ for $t < 0$.) If we write $\varphi(\boldsymbol{\theta}) = \mathbb{E}[\exp(-\sum_{k\in\{L,G\}} \theta_k \int_0^\infty J_k(t)dt)]$, where $\boldsymbol{\theta} = (\theta_L, \theta_G)$, for the joint Laplace-Stieltjes transform of the local and global infectious pressure exerted by an individual on each of its neighbours, we could change our model specification to include $\varphi(\boldsymbol{\theta})$ instead of $\phi(\theta)$, λ_L and λ_G. The arguments used in this paper are easily modified to show that our results would then hold with all occurences of $\phi(i\lambda_L + j\lambda_G)$ in our formulae replaced with $\varphi(i, j)$, this being the probability that an infectious individual fails to infect all of a given set of i of its local and j of its global neighbours.

Acknowledgment This research was supported by the UK Engineering and Physical Sciences Research Council, under research grant EP/E038670/1.

Appendix A Final state random variables

We first describe the setup for dealing with final state random variables in multiple group epidemics, then introduce multivariate Gontcharoff polynomials and the main results of Ball and O'Neill [20] before applying them to our model. We take the opportunity to present these results in terms of probability generating functions rather than Laplace transforms because the former are most often applied to discrete random variables (as in our situation). To simplify the presentation of Ball and O'Neill's results we omit from their setup the possibility of individuals moving between groups.

We introduce some notation we require throughout this appendix. For suitable vectors $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_m)$, we define $\boldsymbol{x}! = \prod_{i=1}^m x_i!$ and $\boldsymbol{x}^{\boldsymbol{y}} = \prod_{i=1}^m x_i^{y_i}$. We say that $\boldsymbol{x} \leq \boldsymbol{y}$ if the inequality holds componentwise, and that $\boldsymbol{x} < \boldsymbol{y}$ if in addition at least one of the componentwise inequalities is strict. We also adopt the convention that any summation over vector indices is rectangular, i.e. $\sum_{\boldsymbol{i}=\boldsymbol{j}}^{\boldsymbol{k}} = \sum_{i_1=j_1}^{k_1} \cdots \sum_{i_m=j_m}^{k_m}$.

A.1 Multiple group framework and main result

The framework we use is that of an SIR epidemic amongst a population of individuals who each belong to one of m groups, labelled $1, 2, \ldots, m$. For $i = 1, 2, \ldots, m$, there are initially a_i infectives and n_i susceptibles in group i, and we write $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)$ and $\boldsymbol{n} = (n_1, n_2, \ldots, n_m)$. The infectious periods of infectives in group i are each distributed according to a random variable $I^{(i)}$. For $i, j = 1, 2, \ldots, m$, the individual to individual infection rate from a given group i infective to a given group j susceptible is λ_{ij}. As usual, such infections are governed by Poisson processes, and all Poisson processes and infectious periods are mutually independent.

To each infective we also attach a random real-valued attribute describing some quantity of interest. This attribute may depend on the individual's infectious period and in our setting it will be the number of global neighbours that the individual infects. The attributes of different infectives are mutually independent and, for $i = 1, 2, \ldots, m$, the attributes of infectives in group i are distributed according to the random variable $A^{(i)}$. For $i = 1, 2, \ldots, m$, let T_i be the number of susceptibles that are ultimately infected in group i and let A_i be the sum of the attributes over all $a_i + T_i$ infectives in group i. Let $\boldsymbol{T} = (T_1, T_2, \ldots, T_m)$ and $\boldsymbol{A} = (A_1, A_2, \ldots, A_m)$. The components of $\boldsymbol{A}$ are called *final state random variables*. Let

$$\Phi(\boldsymbol{x}, \boldsymbol{s}) = \mathbb{E}[\boldsymbol{x}^{\boldsymbol{n}-\boldsymbol{T}} \boldsymbol{s}^{\boldsymbol{A}}], \tag{18}$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_m)$ and $\boldsymbol{s} = (s_1, s_2, \ldots, s_m)$. In order to give an expression for $\Phi(\boldsymbol{x}, \boldsymbol{s})$ it is convenient to define the multivariate Gontcharoff polynomials first studied by Lefèvre and Picard [18]. If $\boldsymbol{U} = (\boldsymbol{u_j} \in \mathbb{R}^m, \boldsymbol{j} \in \mathbb{Z}_+^m)$ is a collection of real numbers, then the Gontcharoff polynomials (associated with $\boldsymbol{U}$) are defined by $G_{\boldsymbol{0}}(\boldsymbol{x} \,|\, \boldsymbol{U}) = 1$ and, for $\boldsymbol{k} \in \mathbb{Z}_+^m \setminus \{\boldsymbol{0}\}$,

$$G_{\boldsymbol{k}}(\boldsymbol{x} \,|\, \boldsymbol{U}) = \frac{\boldsymbol{x}^{\boldsymbol{k}}}{\boldsymbol{k}!} - \sum_{\boldsymbol{0} \leq \boldsymbol{j} < \boldsymbol{k}} \frac{\boldsymbol{u}_{\boldsymbol{j}}^{\boldsymbol{k}-\boldsymbol{j}}}{(\boldsymbol{k}-\boldsymbol{j})!} G_{\boldsymbol{j}}(\boldsymbol{x} \,|\, \boldsymbol{U}).$$

Note that $G_{\boldsymbol{k}}(\boldsymbol{x} \,|\, \boldsymbol{U})$ is a polynomial of degree $k_1, k_2, \ldots, k_m$ in the variables $x_1, x_2, \ldots, x_m$, respectively, depending only on those parameters $\boldsymbol{u_j} \in \mathbb{R}^m$ for which $\boldsymbol{j} < \boldsymbol{k}$. The following result is an immediate application of Ball and O'Neill [20, Theorem 5.1] to the special case of their model with no movement between groups, noting that $\prod_{i=1}^m s_i^{A_i} = \exp(\sum_{i=1}^m A_i \log s_i)$.

Theorem 4 (Ball and O'Neill [20]) *The joint PGF defined by* (18) *is given by*

$$\Phi(\boldsymbol{x}, \boldsymbol{s}) = \sum_{\boldsymbol{j}=\boldsymbol{0}}^{\boldsymbol{n}} \frac{\boldsymbol{n}!}{(\boldsymbol{n}-\boldsymbol{j})!} (\boldsymbol{\psi}(\boldsymbol{s}, \boldsymbol{j}))^{\boldsymbol{n}+\boldsymbol{a}-\boldsymbol{j}} G_{\boldsymbol{j}}(\boldsymbol{x} \,|\, \boldsymbol{U}^*), \tag{19}$$

where $\boldsymbol{\psi}(\boldsymbol{s}, \boldsymbol{j}) = (\psi_1(s_1, \boldsymbol{j}), \psi_2(s_2, \boldsymbol{j}), \ldots, \psi_m(s_m, \boldsymbol{j}))$ *with*

$$\psi_i(s_i, \boldsymbol{j}) = \mathbb{E}\left[\exp\left(-I^{(i)} \sum_{k=1}^m \lambda_{ik} j_k\right) s_i^{A^{(i)}}\right] \qquad (i = 1, 2, \ldots, m),$$

and $\boldsymbol{U}^* = (\boldsymbol{u_j},\, \boldsymbol{j} \in \mathbb{Z}_+^m)$ *has components* $\boldsymbol{u_j} = \boldsymbol{\psi}(\boldsymbol{s}, \boldsymbol{j})$.

A.2 Application to our model

We now apply these results to the analysis of the number of global infections emanating from a household with a single initial infective in our model. Because the degree distribution of the primary and secondary individuals are different the attributes (number of global infections made) associated with primary and secondary have different distributions and so these are the groups. We see considerable simplification arising from the fact that there are initial susceptibles in only one group and also because we are interested not in the individual final state random variables A_1 and A_2 but their sum.

Consider the problem of determining the distribution of the number of global neighbours infected by the members of a household of size n in our model with a single initial infective. We let the initially infective individual (the primary infective) comprise group 1, and the remaining (secondary) individuals be in group 2; so that $\boldsymbol{a} = (1, 0)$ and $\boldsymbol{n} = (0, n-1)$. We also have $I^{(i)} \stackrel{\mathcal{D}}{=} I$ and $\lambda_{ij} = \lambda_L$ for all $i, j = 1, 2$. The attribute $A^{(i)}$ we associate with each infected individual in group i is the number of its global neighbours it infects. Denoting by T the final size as defined in Section 2.2, this leads to

$$A^{(1)} + \sum_{j=0}^{T} A_j^{(2)} \stackrel{\mathcal{D}}{=} C_0 + \sum_{j=1}^{T} C_j = C^{(n)},$$

where the $A_j^{(2)}$ are independent copies of $A^{(2)}$. Because we are not interested in $A^{(1)}$ and $\sum_{j=0}^T A_j^{(2)}$ individually but rather their sum, we set $\boldsymbol{s} = (s, s)$ in (19). Similarly, the fact that there are no initial susceptibles in group 1 means that $\boldsymbol{T} = (0, T)$ is essentially a scalar and thus the first component of $\boldsymbol{x}$ is irrelevant and we may set $\boldsymbol{x} = (1, x)$. With these choices of $\boldsymbol{x}$ and $\boldsymbol{s}$, $\Phi(\boldsymbol{x}, \boldsymbol{s})$ is the joint PGF of $n - 1 - T$ and $C^{(n)}$.

Further simplification arises from the fact that the index $\boldsymbol{j}$ of the summation in the expression (19) ranges from $\mathbf{0} = (0,0)$ to $\boldsymbol{n} = (0, n-1)$, so we only need to consider $\boldsymbol{j}$ of the form $(0, j)$. It follows easily from the definition of the multivariate Gontcharoff polynomials that when the index has one component which is always 0 we can ignore the variable corresponding to this component and thus reduce an m-variable polynomial to an $(m-1)$-variable polynomial, which in our case means we need only deal with single variable Gontcharoff polynomials.

In order to calculate $\psi_l(s,j) = \mathbb{E}\left[e^{-j\lambda_L I} s^{C_{l-1}}\right]$ $(l = 1, 2)$, we first condition on the number of global neighbours K_l (excluding its infector if $l = 1$) and infectious period I of a group l individual. The number C_{l-1} of global neighbours that it infects is then binomially distributed with parameters K_l and $1 - \mathrm{e}^{-\lambda_G I}$. Thus

$$\begin{aligned}
\psi_l(s,j) &= \mathbb{E}_{K_l,I}\left[e^{-j\lambda_L I}\,\mathbb{E}_{C_{l-1}}\left[s^{C_{l-1}} \mid K_l, I\right]\right] \\
&= \mathbb{E}_{K_l,I}\left[e^{-j\lambda_L I}\left(e^{-\lambda_G I} + \left(1 - e^{-\lambda_G I}\right) s\right)^{K_l}\right] \\
&= \mathbb{E}_{K_l,I}\left[e^{-j\lambda_L I}\left(s + (1-s)\, e^{-\lambda_G I}\right)^{K_l}\right].
\end{aligned}$$

Using the binomial theorem, evaluating the expectations and interchanging the order of the sums, we find that, with $p_k^{(l)} = \mathbb{P}(K_l = k)$ $(k = 0, 1, \ldots)$,

$$\begin{aligned}
\psi_l(s,j) &= \mathbb{E}_I\left[e^{-j\lambda_L I} \sum_{k=0}^{\infty} p_k^{(l)} \sum_{i=0}^{k} \binom{k}{i} s^{k-i}(1-s)^i e^{-i\lambda_G I}\right] \\
&= \sum_{k=0}^{\infty} p_k^{(l)} \sum_{i=0}^{k} \binom{k}{i} s^{k-i}(1-s)^i \phi(j\lambda_L + i\lambda_G) \qquad (20) \\
&= \sum_{i=0}^{\infty} \frac{(1-s)^i \phi(j\lambda_L + i\lambda_G)}{i!} \sum_{k=i}^{\infty} \frac{k!}{(k-i)!} s^{k-i} p_k^{(l)} \\
&= \sum_{i=0}^{\infty} \frac{(1-s)^i \phi(j\lambda_L + i\lambda_G)}{i!} f_{K_l}^{(i)}(s),
\end{aligned}$$

where, for $i = 0, 1, \ldots$, $f_{K_l}^{(i)}$ is the ith derivative of the PGF f_{K_l}.

The PGF of $C^{(n)}$ follows using Theorem 4, since $f_{C^{(n)}}(s) = \Phi((1,1),(s,s)$. Note that in the statement of Theorem 1 we have re-labelled ψ_1 and ψ_2 as ψ_0 and ψ_1, as they are quantities associated with primary and secondary individuals, respectively, which we have labelled 0 and $1, 2, \ldots$. To complete the proof of Theorem 1, note that the decomposition $f_C(s) = \sum_{n=1}^{\infty} \tilde{\rho}_n f_{C^{(n)}}(s)$ follows from a simple conditioning on household size.

Appendix B Matching the NHM and SHM with households of size 2

Here we present the proof that when we allow households of size 2, it is not possible to match up the forward processes of the NHM and the SHM unless the infectious period distributions are both almost surely constant. Recall that we have shown in Section 4.2 that for households of size 1 the forward processes coincide if and only if $D \sim \text{Poi}(\mu_D)$ and $\lambda'_G I' \stackrel{\mathcal{D}}{=} \mu_D(1 - \mathrm{e}^{-\lambda_G I})$, and that this implies that the backward processes coincide. Now, assuming that these conditions hold, we investigate what happens if we also allow households of size 2.

In the SHM, we have $f_{C^{(2)\prime}}(s) = \mathbb{E}[\mathbb{E}[s^{C^{(2)\prime}} \mid I'_0]]$, where I'_0 is the infectious period of the primary case, and

$$C^{(2)\prime} \mid I'_0 \sim \begin{cases} Z'_0 & \text{with probability } \mathrm{e}^{-\lambda'_L I'_0}, \\ Z'_0 + Z'_1 & \text{with probability } 1 - \mathrm{e}^{-\lambda'_L I'_0}, \end{cases}$$

where Z'_0 and Z'_1 are independent, $Z'_0 \sim \text{Poi}(\lambda'_G I'_0)$ and Z'_1 is Poisson with random mean $\lambda'_G I'_1$. (Here I'_1 is the infectious period of the secondary case if one occurs.) Thus,

$$f_{C^{(2)\prime}}(s) = \mathbb{E}[\mathrm{e}^{-\lambda'_L I'_0}\, \mathrm{e}^{-\lambda'_G I'_0(1-s)} + (1 - \mathrm{e}^{-\lambda'_L I'_0})\, \mathrm{e}^{-\lambda'_G I'_0(1-s)} f_1(s)],$$

where $f_1(s) = \mathbb{E}[\mathrm{e}^{-\lambda'_G I'(1-s)}]$. Hence,

$$f_{C^{(2)\prime}}(s) = \mathbb{E}[\mathrm{e}^{-(\lambda'_L I'_0 + \lambda'_G I'_0(1-s))}](1 - f_1(s)) + f_1(s)^2.$$

In the NHM, on the other hand, we have $f_{C^{(2)}}(s) = \mathbb{E}[\mathbb{E}[s^{C^{(2)}} \mid I_0]]$, with

$$C^{(2)} \mid I_0 \sim \begin{cases} Z_0 & \text{with probability } \mathrm{e}^{-\lambda_L I_0}, \\ Z_0 + Z_1 & \text{with probability } 1 - \mathrm{e}^{-\lambda_L I_0}, \end{cases}$$

where Z_0 and Z_1 are independent, $Z_0 \sim \text{Poi}(\mu_D(1-\mathrm{e}^{-\lambda_G I_0}))$ and Z_1 is Poisson with random mean $\mu_D(1 - \mathrm{e}^{-\lambda_G I_1})$. It follows in the same way as above that

$$f_{C^{(2)}}(s) = \mathbb{E}[\mathrm{e}^{-(\lambda_L I_0 + \mu_D(1-\mathrm{e}^{-\lambda_G I_0})(1-s))}](1 - f_2(s)) + f_2(s)^2,$$

where $f_2(s) = \mathbb{E}[\mathrm{e}^{-\mu_D(1-\mathrm{e}^{-\lambda_G I_0})(1-s)}]$.

Now

$$\lambda'_G I' \stackrel{\mathcal{D}}{=} \mu_D(1 - \mathrm{e}^{-\lambda_G I}) \iff f_1(s) = f_2(s),\ s \in [0,1],$$

so $f_{C^{(2)\prime}}(s) = f_{C^{(2)}}(s)$ if and only if

$$\mathbb{E}[\mathrm{e}^{-(\lambda'_L I'_0 + \lambda'_G I'_0(1-s))}] = \mathbb{E}[\mathrm{e}^{-(\lambda_L I_0 + \mu_D(1-\mathrm{e}^{-\lambda_G I_0})(1-s))}]$$

for all $s \in [0,1]$. Assume without loss of generality that $\lambda_G = 1$. Then $\lambda'_G I' \stackrel{\mathcal{D}}{=} \mu_D(1-\mathrm{e}^{-I})$, so we require

$$\mathbb{E}[\mathrm{e}^{-\kappa Y}\mathrm{e}^{-(1-s)\mu_D Y}] = \mathbb{E}[(1-Y)^{\lambda_L}\mathrm{e}^{-(1-s)\mu_D Y}] \qquad (0 \le s \le 1),$$

where $\kappa = \lambda'_L \mu_D / \lambda'_G$ and $Y = 1 - \mathrm{e}^{-I}$, so Y takes values in $[0, 1]$. Putting $\theta = (1 - s)\mu_D$, we therefore have that $f_{C'}(s) \equiv f_C(s)$ if and only if

$$\mathbb{E}[\mathrm{e}^{-\theta Y} \mathrm{e}^{-\kappa Y}] = \mathbb{E}[\mathrm{e}^{-\theta Y}(1 - Y)^{\lambda_L}] \qquad (0 \leq \theta \leq \mu_D).$$

It then follows from the uniqueness of Laplace transforms of measures (see, for example, Feller [31, Section XIII.1]) that this is impossible unless $\mathbb{P}(Y = y) = 1$ for some $y \in [0, 1]$.

Appendix C Clustering proofs

In this appendix we provide precise statements concerning the behaviour of the empirical total degree distribution and the asymptotics of the clustering coefficient $\mathcal{C}^{(m)}$ in our model for large numbers of households. We construct a sequence of random graphs (G_m), with m the number of households, such that as $m \to \infty$, the PGF of the empirical degree distribution converges almost surely pointwise to $f'_H(s) f_D(s)/\mu_H$ and $\mathcal{C}^{(m)}$ converges almost surely to $\mathcal{C}$ given in (14). Here we consider the model with independent, identically distributed household sizes. The model with finite support for household sizes can be dealt with in a similar way.

Firstly, let $Q^{(m)} = (q_j^{(m)}, j = 0, 1, \ldots)$ describe the empirical degree distribution of the random network of m households, constructed as described in Section 1.2. That is, the fraction of vertices in the network that have j incident edges (global plus local) is $q_j^{(m)}$. Or, equivalently, $q_j^{(m)}$ is the probability that the total degree of a vertex chosen uniformly at random from a population of m households is j. Note that $Q^{(m)}$ only depends on the first m household sizes and the (global) degrees of the individuals in these first m households. The pairing of the global 'half-edges' is independent of $Q^{(m)}$.

We construct the sequence of random graphs $(G_1, G_2, \ldots)$ as follows. Let $H_1, H_2, \ldots$ be a sequence of independent and identically distributed household sizes with distribution ρ (i.e. distributed as H). Furthermore, let $D_1, D_2, \ldots$ be a sequence of independent global degrees identically distributed as D. For convenience we define $\hat{H}_m = \sum_{i=1}^{m} H_i$.

The graph G_m consists of m households of respective sizes $H_1, H_2, \ldots, H_m$. If $\sum_{j=1}^{\hat{H}_m} D_j$ is even, then the j-th vertex in k-th household has global degree $D_{\hat{H}_{k-1}+j}$ (if that vertex is present in G_m), i.e. this individual has $D_{\hat{H}_{k-1}+j}$ half-edges attached to it. If $\sum_{j=1}^{\hat{H}_m} D_j$ is odd, then for all but the H_m-th vertex in the m-th household the j-th vertex in the k-th household has global degree $D_{\hat{H}_{k-1}+j}$, while the H_m-th vertex in the m-th household has global degree $D_{\hat{H}_m} + 1$. This implies that the sum of global degrees in G_m is even. To complete the construction of G_m the half-edges are paired uniformly at random. The sequence of networks $G_1, G_2, \ldots$ may be constructed by assuming that the pairings of half-edges in $G_1, G_2, \ldots$ are mutually independent, although in our proofs we require only that for every G_m the marginal probability of every possible pairing of the half-edges is equal.

C.1 The empirical degree distribution

By the strong law of large numbers the fraction of the vertices in G_m that are in a household of size h and have global degree d converges almost surely to $\tilde{\rho}_h p_d$ as $m \to \infty$. This implies that for every $c < \infty$ and $s \in [0,1]$, $\sum_{i=0}^{c} q_i^{(m)} s^i \overset{\text{a.s.}}{\to} \mathbb{E}[s^{D+\tilde{H}-1}\mathbb{1}(D+\tilde{H}-1 \le c)]$ as $m \to \infty$, where $\overset{\text{a.s.}}{\to}$ denotes almost sure convergence and, as before, $q_i^{(m)}$ is the fraction of vertices in G_m of total degree i. This in turn implies that the PGF $f_{Q^{(m)}}$ of the empirical degree distribution $Q^{(m)}$ almost surely converges pointwise to $f'_H(s)f_D(s)/\mu_H$ for $s \in [0,1]$.

C.2 An auxilliary graph

Before discussing the clustering we define and find some basic properties of a graph which represents the global connections between households rather than individuals. Define $\bar{T} = (T_i,\, i = 1, 2, \ldots)$ as the sequence of random variables denoting the total global degree of the households:

$$T_m = \sum_{i=\hat{H}_{m-1}+1}^{\hat{H}_m} D_i.$$

Note that by construction $T_1, T_2, \ldots$ are independent and identically distributed. Let T be a random variable distributed as T_i. Since H and D have finite second moments, T also has finite second moment. Let $L_i = \sum_{j=1}^{i} T_j$ be the sum of the degrees of the first i households.

For ease of exposition, we consider a new random graph $\hat{G}_m$ constructed by the configuration model in which vertices have degrees $T_1, T_2, \ldots T_m$. (To be complete, if the total degree $\sum_{i=1}^{m} T_i$ is odd, 1 half-edge is added to the final vertex). We denote the vertices in this graph by $v_1, v_2, \ldots, v_m$. The half-edges assigned to vertex v_j are denoted by $e_{j,1}, e_{j,2}, \ldots, e_{j,T_j}$. We use the notation $[e_{i,a}, e_{j,b}] \in E$ if $e_{i,a}$ and $e_{j,b}$ are paired. We have therefore constructed $\hat{G}_m$ so that it has the same distribution as G_m with vertices in the same household projected to a single vertex.

By arguments similar to those used to prove [32, Theorem 3.1.2], we can prove that the number of triangles (here defined as the number of circuits of length 3) plus the number of self-loops (an edge for which the start and end vertex are the same) and parallel edges (two edges with the same start and end vertex) in $\hat{G}_m$ converges in distribution to a Poisson distribution, with parameters depending only on the first two moments of T. This implies that the number of these imperfections per household converges in probability to 0 as $m \to \infty$.

C.3 Triangles and clustering

We denote the number of (rooted and oriented) triangles in $\hat{G}_m$ by W_m, i.e. if the vertices v_α, v_β and v_γ form a triangle we count 6 distinct triangles corresponding to the 6 ordered triples $(\alpha,\beta,\gamma), (\alpha,\gamma,\beta), \cdots, (\gamma,\beta,\alpha)$. We also let $\hat{W}_m(\alpha,\beta,\gamma)$ be the number of triangles that can be formed from the *ordered* triple $(v_\alpha, v_\beta, v_\gamma)$ of vertices. (Note that $\hat{W}_m(\alpha,\beta,\gamma)$

may be strictly greater than 1 if $\hat{G}_m$ has multiple edges.) Our first goal is to prove that $m^{-1}W_m \stackrel{\text{a.s.}}{\to} 0$. In order to prove this we need the following lemma.

Lemma 1 *For a sequence $\bar{T} = (T_m,\, m = 1, 2, \ldots)$ of independent and identically distributed non-degenerate random variables with finite second moment, we have, as $m \to \infty$, (a) $\mathbb{1}(\max_{1\le i\le m} T_i < \sqrt{m}) \stackrel{\text{a.s.}}{\to} 1$ and (b) $\mathbb{1}(L_m > \mathbb{P}(T > 0)m/2) \stackrel{\text{a.s.}}{\to} 1$.*

Proof. For the convergence of $\mathbb{1}(\max_{1\le i\le m} T_i < \sqrt{m})$, observe that $\sum_{i=1}^{\infty} \mathbb{P}(T_i > \sqrt{i}) = \sum_{i=1}^{\infty} \mathbb{P}((T_i)^2 > i) = \mathbb{E}[T_i^2] < \infty$. By the first Borel-Cantelli lemma this implies that, almost surely, $\{T_i > \sqrt{i}\}$ occurs for at most finitely many i and the first statement of the lemma follows. The convergence of $\mathbb{1}(L_m > \mathbb{P}(T > 0)m/2)$ follows by the strong law of large numbers. □

The next step is to prove that $m^{-1}W_m \mathbb{1}(\max_{1\le i\le m} T_i < \sqrt{m}) \stackrel{\text{a.s.}}{\to} 0$ as $m \to \infty$. To do this we show that, for every $\varepsilon > 0$, $\sum_{i=1}^{\infty} \mathbb{P}(i^{-1}W_i \mathbb{1}(\max_{1\le j\le i} T_j < \sqrt{i}) > \varepsilon) < \infty$.

Lemma 2 *For $\bar{T} = (T_m,\, m = 1, 2, \ldots)$ as in Lemma 1 and W_m as above, there is a constant $C < \infty$ such that $\mathbb{E}[W_m^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})] < C$ for all m.*

Proof. Firstly, observe that

$$\mathbb{E}[W_m^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})] = \mathbb{E}[(6 \sum_{1\le\alpha<\beta<\gamma\le m} \hat{W}_m(\alpha,\beta,\gamma))^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})].$$

The square of the sum contains terms $\hat{W}(\alpha_1,\beta_1,\gamma_1)\hat{W}(\alpha_2,\beta_2,\gamma_2)$ for which the quantity $|\{\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2\}|$ takes the values 3, 4, 5 and 6.

The sum of terms with $|\{\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2\}| = 3$ is bounded above by

$$6m(m-1)(m-2)\,\mathbb{E}((\hat{W}_m(\alpha,\beta,\gamma)^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})).$$

Observe that

$$\begin{aligned}
&\mathbb{E}[(\hat{W}_m(\alpha,\beta,\gamma))^2 \,|\, \bar{T}] \\
&= 2^6 \sum_{1\le a_1<b_1\le T_\alpha} \sum_{1\le a_2<b_2\le T_\beta} \sum_{1\le a_3<b_3\le T_\gamma} \sum_{1\le a_1'<b_1'\le T_\alpha} \sum_{1\le a_2'<b_2'\le T_\beta} \sum_{1\le a_3'<b_3'\le T_\gamma} \\
&\quad \mathbb{P}([e_{\alpha,a_1}, e_{\beta,b_2}], [e_{\alpha,a_1'}, e_{\beta,b_2'}], [e_{\beta,a_2}, e_{\gamma,b_3}], [e_{\beta,a_2'}, e_{\gamma,b_3'}], [e_{\gamma,a_3}, e_{\alpha,b_1}], [e_{\gamma,a_3'}, e_{\alpha,b_1'}] \in E).
\end{aligned}$$

If we write $\binom{T_{\alpha,\beta,\gamma}}{i,j,k} = \binom{T_\alpha}{i}\binom{T_\beta}{j}\binom{T_\gamma}{k}$ and $a^{[i]} = [(a-1)(a-3)\cdots(a-2i+1)]^{-1}$, straightforward algebra shows that

$$\begin{aligned}
\mathbb{E}[(\hat{W}_m(\alpha,\beta,\gamma))^2 \,|\, \bar{T}] = 2^6 \Big\{ &\tbinom{T_{\alpha,\beta,\gamma}}{2,2,2} L_m^{[3]} + \Big(\tbinom{T_{\alpha,\beta,\gamma}}{3,3,2} + \tbinom{T_{\alpha,\beta,\gamma}}{3,2,3} + \tbinom{T_{\alpha,\beta,\gamma}}{2,3,3} \Big) L_m^{[4]} \\
&+ \Big(\tbinom{T_{\alpha,\beta,\gamma}}{3,3,4} + \tbinom{T_{\alpha,\beta,\gamma}}{3,4,3} + \tbinom{T_{\alpha,\beta,\gamma}}{4,3,3} \Big) L_m^{[5]} + \tbinom{T_{\alpha,\beta,\gamma}}{4,4,4} L_m^{[6]} \Big\},
\end{aligned}$$

where the first term arises if the set of half-edges under consideration has cardinality 6 (the summands with three edges), the second term arises if the set of half-edges under consideration has cardinality 8, etc.

It follows that there exists $c_1 > 0$ such that

$$c_1 \,\mathbb{E}[(\hat{W}_m(\alpha,\beta,\gamma))^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m}) \,|\, \bar{T}] \le (T_\alpha T_\beta T_\gamma)^2 (L_m^{[3]} + mL_m^{[4]} + m^2 L_m^{[5]} + m^3 L_m^{[6]}).$$

By the almost sure convergence of $\mathbb{1}(L_m \ge \mathbb{P}(T>0)m/2)$ to 1 (Lemma 1(b)), $\mathbb{E}[T^2] < \infty$ and

$$\mathbb{E}[(\hat{W}(\alpha,\beta,\gamma))^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})] = \mathbb{E}\left[\mathbb{E}[(\hat{W}(\alpha,\beta,\gamma))^2 \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m}) \,|\, \bar{T}]\right],$$

we conclude that terms with $|\{\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2\}| = 3$ add at most a constant to

$$\mathbb{E}[(6 \sum_{1\le\alpha_1<\beta_1<\gamma_1\le m} (6 \sum_{1\le\alpha_2<\beta_2<\gamma_2\le m} \hat{W}_m(\alpha_1,\beta_1,\gamma_1)\hat{W}_m(\alpha_2,\beta_2,\gamma_2))) \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})].$$

In a similar fashion we can deal with summands for which $|\{\alpha_1,\beta_1,\gamma_1,\alpha_2,\beta_2,\gamma_2\}| \neq 3$, and computations are available from the authors. The lemma now follows. □

We are now ready to prove the following theorem.

Theorem 5 *For $\bar{T} = (T_m,\, m = 1, 2, \ldots)$ as in Lemma 1 and W_m as above, $m^{-1}W_m \xrightarrow{\text{a.s.}} 0$ as $m \to \infty$.*

Proof. By Chebychev's inequality we have that, for any $\varepsilon > 0$,

$$\sum_{m=1}^{\infty} \mathbb{P}(m^{-1}W_m \mathbb{1}(\max_{1\le j\le m} T_m < \sqrt{m}) > \varepsilon) \le \sum_{m=1}^{\infty} (m^{-2}\,\mathbb{E}[\mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m})(W_m)^2]/\varepsilon^2,$$

which by Lemma 2 is bounded above by $C\sum_{m=1}^{\infty} m^{-2} < \infty$ for some $C < \infty$. This implies that $\sum_{m=1}^{\infty} \mathbb{P}(m^{-1}W_m \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m}) > \varepsilon) < \infty$. We therefore have $m^{-1}W_m \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m}) \xrightarrow{\text{a.s.}} 0$ as $m \to \infty$. By Lemma 1(a), we also have that, almost surely, $m^{-1}W_m \mathbb{1}(\max_{1\le j\le m} T_j < \sqrt{m}) = m^{-1}W_m$ for all sufficiently large m, whence $m^{-1}W_m \xrightarrow{\text{a.s.}} 0$ as $m \to \infty$. □

Remark. By similar arguments it is possible to prove that the number of self-loops and parallel edges per vertex converges almost surely to 0 as $m \to \infty$; this is stronger than the convergence in probability which follows from [32, Theorem 3.1.2].

Because the pairing of the global half-edges in G_m follows the same rules as the pairing of the half-edges in $\hat{G}_m$, the number of triangles per household, using global edges in G_m converges almost surely to 0 by Theorem 5. The number of triples in G_m does not depend on the pairing of the half-edges and the number of triples per household converges almost surely to $\mathbb{E}[H(H-1+D)(H-2+D)]$ by the strong law of large numbers; while the number of triples that contain global edges per household converges, by the strong law of large numbers, to $\mathbb{E}[H(2D(H-1)+D(D-1))]$. This implies that $\mathcal{C}^{(m)} \xrightarrow{\text{a.s.}} \mathcal{C}_1$ as $m \to \infty$, where $\mathcal{C}$ is given by (15).

Appendix D Alternative threshold parameters

In this appendix we describe two alternative threshold parameters that are derived from looking at the proliferation of infected individuals rather than households. One such parameter, say R_1, is given by the largest eigenvalue of the matrix

$$M_1 = \begin{pmatrix} \mu_{\tilde{D}-1}(1-\phi(\lambda_G)) & \mu_T \\ \mu_D(1-\phi(\lambda_G)) & 0 \end{pmatrix},$$

recalling that $\mu_T = \sum_{n=1}^{\infty} \tilde{\rho}_n \mu_{T^{(n)}}$ is the expected within-household final size amongst secondary individuals (cf. equation (3.36) of Ball *et al.* [3], who give a similar threshold parameter for the SHM). Here the first row describes the mean number of primary and secondary individuals infected by a primary infective and the second row gives the same quantities for a secondary infective. Note, however, that this formulation assigns all secondary infectives in a household to the primary infective in that household, when clearly some secondary infectives will infect further secondary infectives. The threshold parameter derived from this mean matrix is thus, in some sense, part way between an individual and a household reproduction number and in Figure 7(a) we see that R_1 is increasing for small values of the clustering coefficient $\mathcal{C}$. It is rather difficult to calculate the exact means which 'should' be in the second column of this mean matrix to make the interpretation of its leading eigenvalue as an individual reproduction number strictly correct, but the following approximation turns out to be satisfactory.

As in Becker and Dietz' reproduction number based on potential cases [15], we assign all secondary individuals that the primary infective contacts to that primary infective. The mean number of secondary individuals assigned to the primary infective is thus $a = (\mathbb{E}[\tilde{H}]-1)(1-\phi(\lambda_L))$. Suppose that each secondary infective infects on average b further secondary individuals and that this process continues until the within-household epidemic stops. Thus the mean within-household final size satisfies $\mu_T = a(1+b+b^2+\cdots) = a/(1-b)$, whence $b = 1 - a/\mu_T < 1$. The threshold parameter R_2 is then given by the largest eigenvalue of the matrix

$$M_2 = \begin{pmatrix} \mu_{\tilde{D}-1}(1-\phi(\lambda_G)) & a \\ \mu_D(1-\phi(\lambda_G)) & b \end{pmatrix}.$$

In Figure 7(a) we find that R_2 is decreasing in $\mathcal{C}$, confirming that the proliferation of infected individuals is on average reduced by an increase in clustering when $\lambda_L = \lambda_G$. We note here that it is not difficult to show that R_1 and R_2 take values greater than and less than 1 together, so that R_2 is indeed a threshold parameter for our model. Note also that $R_2 > 0$ when $\mathcal{C} = 1$; this is because when $\mathcal{C} = 1$ there is still some proliferation of infected individuals within the initially infected household, though the infection cannot escape the initial household and this is reflected by the fact that $R_2 = b < 1$ when $\mu_D = \mu_{\tilde{D}-1} = 0$ (i.e. $\mathcal{C} = 1$.)

Appendix E Proofs of comparison results

In this appendix we prove Theorem 3, which enables comparisons to be made between network household models having common degree and household size distributions, D and H, but different infectious period distributions and infection rates, so that the models have the same population structure but different disease dynamics. The proofs utilise a theorem of Kuulasmaa [30] concerning a class of percolation models called locally-dependent random graphs, which we describe first. Let $G = (V, E)$ be a directed graph, where V is a set of vertices and E is a set of directed edges. We assume that V is finite and G is simple, in the sense that E does not contain any loops or multiple edges, though we do allow two edges having opposite directions between any pair of vertices. (Kuulasmaa allows V to be countably infinite, however we require only the finite case where the results are slightly simpler to state.) A percolation model on G is defined as follows. Each directed edge in E is coloured black or white. For each $v \in V$, let E_v be the set of directed edges *from* v in G. The colours of edges in E_v are determined by a probability measure P_v and the colours of edges emanating from distinct vertices are mutually independent. Let P be the product probability measure $\prod_{v \in V} P_v$. The pair (G, P) is called a locally-dependent random graph.

A (self-avoiding) path ξ in G is an ordered set $v_0, v_1, \ldots, v_n$ of (at least two) distinct vertices such that the directed edge from v_i to v_{i+1} is in E for each $i = 0, 1, \ldots, n-1$. A path is called black if and only if *all* of its constituent edges are black. For any set Ξ of paths in G, $\mathcal{B}^{\Xi}$ is the event that at least one of the paths in Ξ is black, i.e. $\mathcal{B}^{\Xi} = \cup_{\xi \in \Xi}\{\xi \text{ is black}\}$.

For each $v \in V$, define the *zero-function* p_v by

$$p_v(J) = P_v(\text{every edge in } J \text{ is white}) \qquad (J \subseteq E_v).$$

Note that the zero-functions $(p_v,\ v \in V)$ uniquely determine P (through a Möbius inversion formula such as those given in Martin-Löf [33, Section 1]). The following result is contained in Theorem 2.1 of Kuulasmaa [30].

Theorem 6 *Let (G, P) and (G, P') be two locally-dependent random graphs, defined on the same directed graph G, with zero-functions $(p_v,\ v \in V)$ and $(p'_v,\ v \in V)$, respectively. Suppose that, for each $v \in V$, $p_v(J) \geq p'_v(J)$ for all $J \subseteq E_v$. Then for any set Ξ of paths in G we have $P(\mathcal{B}^{\Xi}) \leq P'(\mathcal{B}^{\Xi})$.*

To connect this result with our epidemic model, suppose that G is the directed graph of possible contacts corresponding to a given realisation of a network of households and global neighbours and, as in Section 3.1, for each individual in the population (vertex v in V) we draw up a list of who v would make infectious contact with if it were to become infected. For each $v \in V$, the directed edges emanating from v are coloured black if the recieving vertex is in v's list and white otherwise. This yields a locally-dependent random graph whose zero-functions are given by

$$p_v(J) = \phi(|J_L|\lambda_L + |J_G|\lambda_G) \qquad (v \in V,\ J \subseteq E_v), \tag{21}$$

where J_L and J_G are, respectively, the sets of local and global edges in J.

We now prove Theorem 3. Let $\mathcal{E}$ and $\mathcal{E}'$ be two NHM epidemics, as in the statement of Theorem 3, which satisfy (16). Turning first to the threshold parameter R_*, consider a household of size n and label its members $0, 1, \ldots, n-1$. Let $D_0, D_1, \ldots, D_{n-1}$ be independent random variables with $D_0 \stackrel{\mathcal{D}}{=} \tilde{D} - 1$ and $D_i \stackrel{\mathcal{D}}{=} D$ $(i = 1, 2, \ldots, n-1)$, and let $D_H = \sum_{i=0}^{n-1} D_i$. Conditional on $\boldsymbol{D} = (D_0, D_1, \ldots, D_{n-1})$, let $G_{\boldsymbol{D}}$ be the directed graph on $n + D_H$ vertices, labelled $0, 1, \ldots, n-1+D_H$, in which there is a pair of directed local edges between every pair of members of the household (individuals $0, 1, \ldots, n-1$) and, for $i = 0, 1, \ldots, n-1$, a pair of directed global edges between i and each of its D_i global neighbours (labelled $k + \sum_{j=0}^{i-1} D_j$, for $k = 1, 2, \ldots, D_i$). Let $(G_{\boldsymbol{D}}, P_{\boldsymbol{D}})$ be the locally-dependent random graph with zero-functions given by (21) and let $(G_{\boldsymbol{D}}, P'_{\boldsymbol{D}})$ be the locally-dependent random graph with zero-functions given by (21) but with $(\phi, \lambda_L, \lambda_G)$ replaced by $(\phi', \lambda'_L, \lambda'_G)$.

For $i = n, n+1, \ldots, n-1+D_H$, let $\Xi_i^{\boldsymbol{D}}$ be the set of all paths in $G_{\boldsymbol{D}}$ from vertex 0 to vertex i. Given that (16) holds, we have, recalling (3) and Theorem 6, that

$$\mathbb{E}[\tilde{C}^{(n)} \,|\, \boldsymbol{D}] = \sum_{i=n}^{n-1+D_H} P_{\boldsymbol{D}}(\mathcal{B}^{\Xi_i^D}) \leq \sum_{i=n}^{n-1+D_H} P'_{\boldsymbol{D}}(\mathcal{B}^{\Xi_i^D}) = \mathbb{E}[\tilde{C}^{(n)\prime} \,|\, \boldsymbol{D}],$$

say. Taking expectations with respect to $\boldsymbol{D}$ yields $\mathbb{E}[\tilde{C}^{(n)}] \leq \mathbb{E}[\tilde{C}^{(n)\prime}]$ and $R_*(\mathcal{E}) \leq R_*(\mathcal{E}')$ then follows using (2).

Consider next the probability of a major outbreak $p_{\mathrm{maj}}(\mathcal{E})$. Let k be a strictly positive integer and construct a (random) tree of households $\mathcal{T}_k$ as follows. The initial (root) household has size, H_0 say, distributed according to ρ (defined in Section 1.2). Suppose that $H_0 = n$ and let $D_0, D_1, \ldots, D_{n-1}$ be independent and identically distributed to D. These give the number of global neighbours of the members of the root household. For each such global neighbour, its household size is found by sampling independently from the size-biased household size distribution $\tilde{\rho}$. These households comprise the first generation of $\mathcal{T}_k$. Subsequent generations are defined in a similar fashion, except that $D_0 \stackrel{\mathcal{D}}{=} \tilde{D} - 1$ rather than D. Of course the degrees of distinct individuals are mutually independent, as are the sizes of distinct households. The construction is continued up to and including generation k, yielding $\mathcal{T}_k$. (It is possible that the network of households dies out before generation k, say at generation $k' < k$, in which case $\mathcal{T}_k$ is the tree given by these first k' generations.) Let $G_{\mathcal{T}_k}$ be the corresponding directed graph in which there is a pair of directed local edges between any two individuals who are in the same household and a pair of directed global edges between any two individuals who are global neighbours in the construction of $\mathcal{T}_k$. Let $(G_{\mathcal{T}_k}, P_{\mathcal{T}_k})$ and $(G_{\mathcal{T}_k}, P'_{\mathcal{T}_k})$ be locally-dependent random graphs with zero-functions given, respectively, by (21) and (21) with $(\phi, \lambda_L, \lambda_G)$ replaced by $(\phi', \lambda'_L, \lambda'_G)$.

Let $Y = (Y_0, Y_1, \ldots)$ denote the approximating branching process introduced in Section 2.1. Thus $Y_0 = 1$ and for $j = 1, 2, \ldots$, Y_j is the number of infectious households in generation j. Specify an individual in the root household in $\mathcal{T}_k$ to be the initial infective and label this individual 0. Observe that $(G_{\mathcal{T}_k}, P_{\mathcal{T}_k})$ can be used to construct a realisation of $Y_0, Y_1, \ldots, Y_k \,|\, \mathcal{T}_k$ in an obvious fashion. In particular, if $\Xi_{\mathcal{T}_k}$ is the set of all paths in

$G_{\mathcal{T}_k}$ from individual 0 to any individual in generation k of $\mathcal{T}_k$, then $Y_k \neq 0$ if and only if the event $\mathcal{B}^{\Xi_{\mathcal{T}_k}}$ occurs. (Note that $\Xi_{\mathcal{T}_k}$ is empty if the construction of $\mathcal{T}_k$ dies out before generation k.) Thus Theorem 6 implies that

$$\mathbb{P}(Y_k = 0 \,|\, \mathcal{T}_k) = 1 - P_{\mathcal{T}_k}(\mathcal{B}^{\Xi_{\mathcal{T}_k}}) \geq 1 - P_{\mathcal{T}_k}(\mathcal{B}^{\Xi_{\mathcal{T}_k}}) = \mathbb{P}(Y'_k = 0 \,|\, \mathcal{T}_k), \tag{22}$$

where $Y' = (Y'_0, Y'_1, \ldots)$ denotes the approximating branching process for $\mathcal{E}'$. Taking expectations of (22) with respect to $\mathcal{T}_k$ yields $\mathbb{P}(Y_k = 0) \geq \mathbb{P}(Y'_k = 0)$ and $p_{\mathrm{maj}}(\mathcal{E}) \leq p_{\mathrm{maj}}(\mathcal{E}')$ follows by letting $k \to \infty$.

The corresponding result for the expected relative final size of a major outbreak is proved similarly, except that we now take $\Xi_{\mathcal{T}_k}$ to be the set of all paths in $G_{\mathcal{T}_k}$ from individuals in generation k of $\mathcal{T}_k$ to individual 0.

To prove the assertion at the beginning of Remark 2 following Theorem 3, we construct a realisation, $\mathcal{N}$ say, of the network and then define the obvious locally-dependent random graphs $(G_{\mathcal{N}}, P_{\mathcal{N}})$ and $(G_{\mathcal{N}}, P'_{\mathcal{N}})$. Then the assertion, conditional on the network configuration $\mathcal{N}$, comes from considering $\mathcal{B}^{\Xi}$, where Ξ is the set of paths leading from the initial infective to the given set of initial susceptibles. The unconditional result then follows from taking expectations with respect to the network configuration $\mathcal{N}$.

References

[1] R. Bartoszyński, On a certain model of an epidemic, Zastos. Mat. 13 (1972/73) 139–151.

[2] N. G. Becker, K. Dietz, The effect of household distribution on transmission and control of highly infectious diseases, Math. Biosci. 127 (1995) 207–219.

[3] F. G. Ball, D. Mollison, G. Scalia-Tomba, Epidemics with two levels of mixing, Ann. Appl. Probab. 7 (1) (1997) 46–89.

[4] H. Andersson, Epidemic models and social networks, Math. Sci. 24 (2) (1999) 128–147.

[5] M. E. J. Newman, Spread of epidemic disease on networks, Phys. Rev. E 66 (1) (2002) 016128.

[6] E. Kenah, J. M. Robins, Second look at the spread of epidemics on networks, Phys. Rev. E 76 (3) (2007) 036113.

[7] L. A. Meyers, B. Pourbohloul, M. E. J. Newman, D. M. Skowronski, R. C. Brunham, Network theory and SARS: predicting outbreak diversity, J. Theoret. Biol. 232 (1) (2005) 71–81.

[8] I. Z. Kiss, D. M. Green, R. R. Kao, The effect of contact heterogeneity and multiple routes of transmission on final epidemic size, Math. Biosci. 203 (1) (2006) 124–136.

[9] F. G. Ball, P. J. Neal, Network epidemic models with two levels of mixing, Math. Biosci. 212 (1) (2008) 69–87.

[10] T. Britton, M. Deijfen, A. N. Lagerås, M. Lindholm, Epidemics on random graphs with tunable clustering, J. Appl. Probab. 45 (3) (2008) 743–756.

[11] F. G. Ball, D. J. Sirl, P. Trapman, An SIR epidemic model on a random network with household structure, To appear in Adv. Appl. Probab. 41 (3).

[12] E. A. Bender, E. R. Canfield, The asymptotic number of labeled graphs with given degree sequences, J. Combinatorial Theory (A) 24 (3) (1978) 296–307.

[13] M. Molloy, B. Reed, A critical point for random graphs with a given degree sequence, Rand. Struct. Alg. 6 (2-3) (1995) 161–179.

[14] L. Pellis, N. M. Ferguson, C. Fraser, The relationship between real-time and discrete-generation models of epidemic spread, Math. Biosci. 216 (1) (2008) 63–70.

[15] N. G. Becker, K. Dietz, Reproduction numbers and critical immunity levels for epidemics in a community of households, in: Athens Conference on Applied Probability and Time Series Analysis, Vol. I (1995), Vol. 114 of Lecture Notes in Statistics, Springer, New York, 1996, pp. 267–276.

[16] E. Goldstein, K. Paur, C. Fraser, E. Kenah, J. Wallinga, M. Lipsitch, Reproductive numbers, epidemic spread and control in a community of households, Math. Biosci. In Press, Corrected Proof.

[17] H. E. Daniels, The distribution of the total size of an epidemic, in: Proc. 5th Berkeley Sympos. Math. Statist. and Prob., Vol. IV, Univ. California Press, Berkeley, Calif., 1967, pp. 281–293.

[18] C. Lefèvre, P. Picard, A nonstandard family of polynomials and the final size distribution of Reed-Frost epidemic processes, Adv. in Appl. Probab. 22 (1) (1990) 25–48.

[19] F. G. Ball, A unified approach to the distribution of total size and total area under the trajectory of infectives in epidemic models, Adv. in Appl. Probab. 18 (2) (1986) 289–310.

[20] F. G. Ball, P. D. O'Neill, The distribution of general final state random variables for stochastic epidemic models, J. Appl. Probab. 36 (2) (1999) 473–491.

[21] H. Andersson, T. Britton, Stochastic epidemic models and their statistical analysis, Vol. 151 of Lecture Notes in Statistics, Springer-Verlag, New York, 2000.

[22] F. G. Ball, O. D. Lyne, Stochastic multitype SIR epidemics among a population partitioned into households, Adv. in Appl. Probab. 33 (1) (2001) 99–123.

[23] F. G. Ball, P. J. Neal, A general model for stochastic SIR epidemics with two levels of mixing, Math. Biosci. 180 (2002) 73–102.

[24] F. G. Ball, Susceptibility sets and the final outcome of stochastic SIR epidemic models, Research Report 00-09, Division of Statistics, School of Mathematical Sciences, University of Nottingham (2000).

[25] R. Durrett, Probability: theory and examples, 2nd Edition, Brooks-Cole, Belmont, CA, 2004.

[26] P. Trapman, On analytical approaches to epidemics on networks, Theor. Pop. Biol. 71 (2007) 160–173.

[27] J. P. Gleeson, S. Melnik, Analytical results for bond percolation and k-core sizes on clustered networks, http://arXiv.org/abs/0811.4511 (2008).

[28] Z. Grossman, Oscillatory phenomena in a model of infectious diseases, Theoret. Population Biol. 18 (2) (1980) 204–243.

[29] M. J. Keeling, B. T. Grenfell, Effect of variability in infection period on the persistence and spatial spread of infectious diseases, Math. Biosci. 147 (2) (1998) 207–226.

[30] K. Kuulasmaa, The spatial general epidemic and locally dependent random graphs, J. Appl. Probab. 19 (4) (1982) 745–758.

[31] W. Feller, An introduction to probability theory and its applications. Vol. II., Second edition, John Wiley & Sons Inc., New York, 1971.

[32] R. Durrett, Random graph dynamics, Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, Cambridge, 2006.

[33] A. Martin-Löf, Symmetric sampling procedures, general epidemic processes and their threshold limit theorems, J. Appl. Probab. 23 (2) (1986) 265–282.