Threshold behaviour and final outcome of an epidemic on a random network with household structure*

Frank Ball† David Sirl† Pieter Trapman‡

12th May 2010

Abstract

This paper considers a stochastic SIR (susceptible→infective→removed) epidemic model in which individuals may make infectious contacts in two ways, both within 'households' (which for ease of exposition are assumed to have equal size) and along the edges of a random graph describing additional social contacts. Heuristically-motivated branching process approximations are described, which lead to a threshold parameter for the model and methods for calculating the probability of a major outbreak, given few initial infectives, and the expected proportion of the population who are ultimately infected by such a major outbreak. These approximate results are shown to be exact as the number of households tends to infinity by proving associated limit theorems. Moreover, simulation studies indicate that these asymptotic results provide good approximations for modestly-sized finite populations. The extension to unequal sized households is discussed briefly.

Keywords Branching processes; coupling; epidemic processes; final outcome; households; local and global contacts; random graphs; susceptibility set; threshold theorem.

AMS MSC (2000) Primary: 92D30, 60K35; Secondary: 05C80, 60J80.

*First published in *Advances in Applied Probability* 41:765–796 (2009); available at `http://projecteuclid.org/euclid.aap/1253281063`.

†University of Nottingham.

‡University Medical Center Utrecht and Vrije Universiteit Amsterdam.

1 Introduction

Epidemic models which include some element of realistic population structure have been the subject of a considerable amount of recent study in recognition of the fact that the classical homogeneously-mixing models are quite unrealistic for all but the smallest of populations.

One approach to this has been to allow *local* contacts of some kind, modelling contacts which occur on a regular basis in addition to maintaining the 'well-mixed' *global* contacts to model chance interactions with random members of the population. A common form for these local contacts to take arises by partitioning the population into *households*, where these local contacts can occur only between individuals who are in the same household (see, for example, Becker and Dietz (1995) and Ball *et al.* (1997)). This can be extended to the overlapping groups model where the population may be partitioned in more than one way (for example, by household and by workplace), with local interactions taking place at (possibly) different rates within groups of the different partitions, see Ball and Neal (2002). Another mode of local interactions is described by the so-called great circle model (Ball *et al.*, 1997; Ball and Neal, 2002, 2003), where the population is spread around a circle and individuals have local contact with only their nearest neighbours. This model is closely related to 'small-world' models (Watts and Strogatz, 1998), which have received considerable attention, particularly in the physics literature.

Another way of accounting for the inhomogeneous nature of interactions is by using random graphs to model social networks (see, for example, Andersson (1997, 1998, 1999), Newman (2002), Durrett (2006), Kenah and Robins (2007) and Britton *et al.* (2008)). Perhaps the most important aspect of these random graph models is that they incorporate a specified degree distribution, the degree of a node in the graph corresponding to the number of other members in the population an individual can possibly make infectious contact with. These models have been extended to also incorporate 'casual contacts' by way of the classical homogeneous mixing effects, see Kiss *et al.* (2006) and Ball and Neal (2008).

In this paper we investigate a model for an SIR (susceptible→infective→removed) epidemic in a closed finite population, which draws together the main aspects of the generalisations of the standard homogeneously mixing model described above. We consider a population grouped into households, with infectious contacts at a given per-pair rate, where individuals also make global contacts along the edges of a random graph over the whole population. We use branching process approximations to derive (i) a threshold parameter, which determines whether a disease with just a few initial infectives can become

established and infect a non-negligible proportion of the population (an event we call a *major outbreak*); (ii) the probability that a major outbreak occurs; and (iii) the expected proportion of the population that is infected by a major outbreak. These results are approximations that become exact in the limit as the size of the population becomes large in an appropriate way.

A feature of our model is that there is clustering present in the network of possible contacts, roughly meaning that there are significant numbers of triangles (and other short cycles) present in the network. This is an important aspect as the presence of triangles captures the phenomenon of people having mutual friends. The effect of such clustering in random networks in an epidemiological setting has been considered, in different models, by Trapman (2007) and Britton *et al.* (2008).

In the remainder of the paper we firstly describe, in Section 2, the full detail of our model. Then in Section 3 we give the ideas behind the above-mentioned branching process approximations. In Section 4 we derive explicit formulae which allow us to calculate the quantities of interest for two important special cases, then give some brief numerical examples in Section 5, including demonstrating that our asymptotic results give good approximations for even moderately-sized finite populations. In Section 6 we rigorously establish the branching process approximations by proving related limit theorems as the population size tends to infinity. The paper concludes with a brief discussion in Section 7.

2 Model

We consider a closed population of m households, each of n individuals, and construct the network of possible global contacts using the 'configuration model' (as in Durrett (2006, Chapter 3)) as follows. Firstly assign to each individual a number of *half-edges*, these numbers being independent realisations of a random variable D (the degree distribution) with $\mathbb{P}(D = k) = p_k$, $k = 0, 1, \ldots$. Conditional on the total number of half-edges being even we then pair these half-edges with each other uniformly at random, whence each such pair of half-edges forms an edge in the (random) graph describing the possible global contacts. We denote by μ_D and σ_D^2 the mean and variance of the distribution D and assume that both of these quantities are finite. We also note for later reference that if we follow an edge from one vertex to another then the degree distribution of the second vertex is the *size-biased* distribution $\tilde{D}$, where $\mathbb{P}(\tilde{D} = k) = kp_k/\mu_D$, $k = 1, 2, \ldots$. This is because in the construction of the graph the half-edges are paired uniformly at random, so it is k times more likely that following an edge leads

one to a vertex of degree k than to a vertex of degree 1. By the degree of an individual we mean the number of individuals adjacent to it in the network of global contacts, not counting those in its own household.

Note that there may be some imperfections in the graph, in the form of parallel edges and self-loops. However, our assumption that $\sigma_D^2 < \infty$ ensures that as $m \to \infty$, the number of these imperfections in the network of global contacts converges in distribution to a Poisson random variable whose mean is a function of (μ_D, σ_D^2) (Durrett, 2006, Theorem 3.1.2). By treating the households as macro-individuals, with degree distribution given by the sum of n independent copies of D, it follows that the numbers of parallel edges between households and household self-loops also converge in distribution to Poisson random variables as $m \to \infty$. Thus the probability that these imperfections are absent in the graph is bounded away from zero as $m \to \infty$, and consequently (cf. Janson (2009)) our asymptotic results continue to hold if the graph is conditioned on having no such imperfections.

When an infective individual makes infectious contact with a susceptible individual, the susceptible becomes infective and remains so for a random period of time distributed according to a non-negative real-valued random variable I, which we specify by its Laplace transform $\phi(\theta) = \mathbb{E}[\mathrm{e}^{-\theta I}]$, $\theta \geq 0$, and call the infectious period. An infective individual makes infectious contact with each other member of his/her household at the points of a Poisson process with rate λ_L and similarly with each individual he is adjacent to in the network of global contacts at rate λ_G. To be emphatic, both λ_L and λ_G are *per-pair* rates, so an infectious individual of degree k makes infectious contacts at overall rate $\lambda_L(n-1) + \lambda_G k$. As usual, all Poisson processes and infectious periods are assumed to be mutually independent.

For ease of presentation, we assume that an epidemic is initiated by a single infective individual within the population, either a given specific individual or an individual chosen uniformly at random from the population. Our assumption that all households are of the same size is also made for ease of presentation although, as indicated in Section 7, our results generalise easily to incorporate unequal household sizes.

3 Heuristics and description of main results

We now give informal descriptions of the branching process approximations we use, firstly to approximate the early stages of an outbreak, leading to a threshold parameter and a method of calculating the probability of a major outbreak and, secondly, to approximate the expected relative final size of (i.e. the proportion of the population infected by) a major outbreak. These

approximations become exact in the limit as the number of households $m \to \infty$, with the household size n held fixed.

3.1 Forward processes

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. 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 (see, for example, Pellis *et al.* (2008)). We first consider the epidemic spreading only within the household containing the initial infective (the local epidemic that it initiates) and then consider the number of individuals infected via global infectious contacts made by those infected by the local epidemic. Because of the way the network is constructed, in the early stages of the epidemic it is highly likely that these globally contacted individuals are all in distinct households (this being critical for the branching process approximation). We then consider each newly infected household in the same manner: local epidemic followed by global infections. Again in the early stages it is highly likely that those infected by such global infectious contacts are in distinct households and furthermore that they are in 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 (offspring) of that household.

Note that the offspring distribution of the above branching process is different for the initial (i.e. zeroth) generation than for subsequent generations, since in subsequent generations the initial infective in a household has been infected by one of its global neighbours, so the number of uninfected global neighbours of this individual is equal in distribution to $\tilde{D} - 1$, whilst in the zeroth generation the initial infective is the initial infective in the whole population, and the degree distribution of this individual is either distributed as D or is a fixed constant, according as the initial infective is chosen (uniformly) at random or a specific individual in the population is chosen to be the initial infective. We therefore define the random variable C to be the number of global neighbours infected by members of the initial infective's household and $\tilde{C}$ to be the number infected by the household of a single infective that was infected by a global neighbour. Our branching process approximation is then defined by it having a single ancestor (since the epidemic starts with one initial infective) and offspring distribution C in the initial generation and $\tilde{C}$ in subsequent generations. Throughout the paper, we denote a branching process of this type by $\mathrm{BP}(1, C, \tilde{C})$, or by $\mathrm{BP}(1, \boldsymbol{c}, \tilde{\boldsymbol{c}})$,

where $\boldsymbol{c} = (c_0, c_1, \ldots)$ and $\tilde{\boldsymbol{c}} = (\tilde{c}_0, \tilde{c}_1, \ldots)$ are the mass functions of C and $\tilde{C}$, respectively.

The above branching process approximation of the epidemic is made fully rigorous in Section 6.4.1, where it is shown that, as $m \to \infty$, the total number of households infected by the epidemic converges in distribution to the total progeny of the branching process (see Theorem 1). Thus, whether or not the epidemic can 'take off' and lead to a major outbreak is determined by whether or not the branching process is supercritical (i.e. whether or not $R_* = \mathbb{E}[\tilde{C}] > 1$). Further, by standard branching process theory, the probability of such a major outbreak is given by $1 - f_C(\sigma)$, where σ is the smallest solution of $f_{\tilde{C}}(s) = s$ in $[0, 1]$, and $f_C(s) = \mathbb{E}[s^C]$ and $f_{\tilde{C}}(s) = \mathbb{E}[s^{\tilde{C}}]$ (for $s \in [0, 1]$) denote the probability generating functions (PGFs) of C and $\tilde{C}$, respectively. (Here and henceforth we denote by $f_X(\cdot)$ the PGF of the random variable X.) Calculation of R_*, $f_C(s)$ and $f_{\tilde{C}}(s)$ is considered in Section 4.1.

3.2 Backward processes

We now consider the expected final size of a major outbreak. Again our analysis is of the $m \to \infty$ limiting epidemic process, for which we find the probability that a given individual is infected in the event of a major outbreak. By an exchangeability argument 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. We determine the probability that a given individual is infected by considering its *susceptibility set* (cf. Ball and Lyne (2001) and Ball and Neal (2002)).

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 infect were it to be infected itself. We then construct a digraph (directed graph) based on these lists, in which the vertices represent individuals in the population and we put a directed arc from i to j when, were i to become infected, it would make infectious contact with j, 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). Note that an individual will become infected by an epidemic if and only if the initial infective is in its susceptibility set. We also need the concept of a *local* susceptibility set, constructed in the same way but considering only local (within-household) infectious contacts.

We approximate the size of the susceptibility set of an individual chosen uniformly at random from the population by the total progeny of an appropriate branching process. To construct this branching process we break up the susceptibility set into 'generations' in much the same way as we look at the spread of infection in the early stages of the epidemic. Starting with an individual i, consider those individuals j, not in i's household, who are in i's susceptibility set by virtue of an arc leading from j to an individual in i's local susceptibility set. These individuals are all in different households with high probability as $m \to \infty$ and the households they are in comprise the first 'generation' of the susceptibility set. Repeating this process for each of these individuals j (i.e. looking at the individuals who make infectious global contact with a member of j's local susceptibility set) gives the second 'generation'; and by continuing this process we can construct the whole of i's susceptibility set. Because each individual j that joins the susceptibility set by virtue of a global contact is in a household not previously associated with the susceptibility set with high probability, the number of households in each generation is approximated well by the branching process $\mathrm{BP}(1, B, \tilde{B})$, where B and $\tilde{B}$ denote the offspring random variables for the initial and subsequent generations, which again are typically different.

We show in Section 6.5.2 that, as $m \to \infty$, the conditional probability that a typical initial susceptible (i say) is infected, given that a major outbreak occurs, is given by the probability that the branching process $\mathrm{BP}(1, B, \tilde{B})$ avoids extinction (see Theorem 2). An intuitive explanation of this result is as follows. As $m \to \infty$, (i) the number of households in i's susceptibility set converges in distribution to the total progeny of $\mathrm{BP}(1, B, \tilde{B})$; and (ii) a major outbreak necessarily infects at least $\log m$ households (cf. Lemma 6). Thus, as $m \to \infty$, the probability that i's susceptibility set intersects one of these $\log m$ households is 0 if $\mathrm{BP}(1, B, \tilde{B})$ goes extinct and 1 otherwise. The latter result follows because if $\mathrm{BP}(1, B, \tilde{B})$ does not go extinct then the size of i's susceptibility set is of exact order m as $m \to \infty$.

The above claim and standard branching process theory imply that the expected relative final size of a major outbreak in a large finite population is approximately $1 - f_B(\xi)$, where ξ is the smallest solution of $f_{\tilde{B}}(s) = s$ in $[0, 1]$. Calculation of $f_B(s)$ and $f_{\tilde{B}}(s)$ is considered in Section 4.2.

4 Calculations

4.1 Forward process

Consider first the threshold parameter $R_* = \mathbb{E}[\tilde{C}]$. Label the individuals in a household $0, 1, \ldots, n-1$, with individual 0 the initial infective, and define χ_i to be the indicator of the event that individual i is infected in the local (i.e. single-household) epidemic and C_i to be the number of global neighbours with which i makes infectious contact, if i were to become infected. Then

$$\tilde{C} = C_0 + \sum_{i=1}^{n-1} \chi_i C_i \tag{4.1}$$

and it follows, since C_1 and χ_1 are independent and $(C_1, \chi_1), (C_2, \chi_2), \ldots, (C_{n-1}, \chi_{n-1})$ are identically distributed, that

$$R_* = \mathbb{E}[C_0] + \mathbb{E}[T]\,\mathbb{E}[C_1], \tag{4.2}$$

where T is the final size of the within-household epidemic (not counting the initial infective). Denote by I_i and K_i the infectious period and number of global neighbours, not including its infector, of individual i (this only affects the initial infective within the household). Now, since infectious contacts between different pairs of individuals are independent, $C_i \,|\, K_i, I_i \sim \mathrm{Bin}(K_i, 1 - \mathrm{e}^{-\lambda_G I_i})$. Thus $\mathbb{E}[C_i \,|\, K_i, I_i] = K_i(1 - \mathrm{e}^{-\lambda_G I_i})$, whence, by the independence of K_i and I_i,

$$\mathbb{E}[C_i] = \mathbb{E}[K_i](1 - \phi(\lambda_G)). \tag{4.3}$$

Now, for $i = 1, 2, \ldots, n-1$, K_i has the same distribution as D, the prescribed degree distribution, so $\mathbb{E}[K_i] = \mu_D$ for such i. However, for the reasons noted in the first paragraph of Section 2, since the initial infective in the household was infected by a global infection its degree has the size-biased distribution $\tilde{D}$, and because one of these neighbours (the one that infected it) has already been infected, K_0 has the same distribution as $\tilde{D} - 1$, so $\mathbb{E}[K_0] = \mathbb{E}[\tilde{D}] - 1$. It follows from the definition of $\tilde{D}$ that $\mathbb{E}[\tilde{D}] = \mathbb{E}[D] + \mathrm{Var}\, D / \mathbb{E}[D]$. Substituting these into (4.3) and then (4.2), and letting $\mu_T = \mathbb{E}[T]$, yields

$$R_* = \left(\mu_D \left(\mu_T + 1 \right) + \frac{\sigma_D^2}{\mu_D} - 1 \right) (1 - \phi(\lambda_G)). \tag{4.4}$$

The mean μ_T may be evaluated (typically numerically) by using equations (2.25) and (2.26) of Ball (1986), thus enabling R_* to be calculated.

Calculation of the PGFs $f_C(s)$ and $f_{\tilde{C}}(s)$ is more difficult because the number of global infections caused by a particular individual is dependent on that individual's infectious period, which also influences whether or not other individuals in the household become infective and thus the number of global contacts they might make. It is possible to use the notion of 'final state random variables' introduced by Ball and O'Neill (1999) to find f_C and $f_{\tilde{C}}$, but it is not straightforward, so we do not present it here. This methodology will be discussed in a forthcoming paper concentrating on the more applied aspects of our model. However, there are two special cases where the above dependencies do not exist and the analysis is much simpler. These are when the infectious period is fixed (i.e. almost surely equal to a given constant) and when the infectious period can be only zero or infinity.

Trapman (2007) describes (using results of Kuulasmaa (1982)) how these special cases lead to bounds on quantities of interest for a very general class of epidemic models. Trapman's arguments hold for any epidemic model where there is only one 'kind' of infectious contact rather than the two (local and global) that we are concerned with, but the methods can be easily adapted. In addition, a fixed infectious period is often a reasonable assumption to make in practice and it is commonly used because it leads to simplifications of the kind shown shortly (see, for example, Britton *et al.* (2007) and Britton *et al.* (2008)). We therefore proceed to calculate the PGFs f_C and $f_{\tilde{C}}$ in these two special cases as they can be used to calculate the above-mentioned bounds and they also may give insight into the importance of and interplay between the parameters of our model. The role of the infectious period distribution will be discussed in the above-mentioned applied paper.

4.1.1 Zero or infinite infectious period.

Suppose that $\mathbb{P}(I = \infty) = 1 - \mathbb{P}(I = 0) = p$ for some $p \in [0, 1]$. For the moment we ignore the differences between the initial and subsequent generations and denote the generic offspring random variable by unadorned C. Here we have

$$C = \begin{cases} 0, & \text{with probability } 1-p, \\ C_0 + \sum_{i=1}^{n-1} C_i, & \text{with probability } p, \end{cases}$$

where C_i is the number of global neighbours infected by an infectious individual i. Thus $C_0 \stackrel{\mathcal{D}}{=} K_0$ (where $\stackrel{\mathcal{D}}{=}$ denotes equality in distribution) and $C_1, C_2, \ldots, C_{n-1}$ are independent and identically distributed with

$$C_i = \begin{cases} 0, & \text{with probability } 1-p, \\ K_i, & \text{with probability } p. \end{cases}$$

Also note that the number, N say, of the $n-1$ C_i's which take the value K_i (i.e. the number of initially susceptible individuals in the household with $I=\infty$) is binomially distributed, with parameters $n-1$ and p. We therefore have

$$\begin{aligned} f_C(s) = \mathbb{E}[s^C] &= (1-p)s^0 + p\,\mathbb{E}[s^{C_0+\sum_{i=1}^{n-1} K_i}] \\ &= 1-p+p\,\mathbb{E}[s^{C_0}]\,\mathbb{E}[s^{\sum_{i=1}^{N} K_i}] \\ &= 1-p+pf_{K_0}(s)f_N(f_D(s)) \\ &= 1-p+pf_{K_0}(s)(1-p+pf_D(s))^{n-1}, \end{aligned}$$

where K_0 is D or d in the initial generation and $\tilde{D}-1$ in subsequent generations (in which case the PGF is $f_{\tilde{C}}$ rather than f_C).

4.1.2 Fixed infectious period.

Now suppose that $\mathbb{P}(I=c)=1$ for some $c>0$. Again we temporarily ignore the differences between the initial and subsequent generations, label the individuals $0,1,\ldots,n-1$ and denote by C_i the number of global neighbours infected by an infectious individual i. Then, letting T denote the final size of the within-household epidemic, we have $C=C_0+\sum_{i=1}^{T} C_i$ and, conditional on the final size, $C_1, C_2, \ldots, C_T$ are mutually independent. Now $C_i \,|\, K_i \sim \mathrm{Bin}(K_i, 1-\mathrm{e}^{-c\lambda_G})$, so $f_{C_i}(s) = f_{K_i}(1-p_G+sp_G)$, where $p_G = 1-\mathrm{e}^{-c\lambda_G}$. Thus, by the usual formula for the PGF of a random sum,

$$f_C(s) = f_{C_0}(s)f_T(f_{C_1}(s)) = f_{K_0}(1-p_G+sp_G)f_T(f_D(1-p_G+sp_G)), \quad (4.5)$$

where again K_0 is D or d in the initial generation and $\tilde{D}-1$ in subsequent generations. The PGF f_T is easily calculated using Theorem 2.6 of Ball (1986).

4.2 Backward process

Now consider the branching process approximation of the growth (as described in Section 3.2) of the susceptibility set of an individual, i_* say, chosen uniformly at random from the population. The offspring distribution of this process has the same distribution as the number of individuals that make global contact with the local susceptibility set of a single individual, say individual i. Again we have a distinction between the initial and subsequent generations but we ignore this for now and denote the random variable of interest by B. Firstly we write

$$B = B_0 + \sum_{j=1}^{M} B_j,$$

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 i) and M is the size of i's local susceptibility set, not counting i itself. (If $M = 0$ then i's local susceptibility set consists of only i itself and the sum is empty.) Now $B_j \,|\, K_j \sim \mathrm{Bin}(K_j, p_G)$, where K_j is the number of global neighbours of j excluding, in the case of the initial individual, the individual it made contact with in order to join the susceptibility set and $p_G = 1 - \phi(\lambda_G)$ is the probability that an infective individual makes infectious contact with a given global neighbour. We do not need to condition on the infectious period of individual j because the contacts we are considering come from other individuals; the independence of the infectious periods of these individuals implies that they make contacts with j independently of each other. For a similar reason, $B_0, B_1, \ldots, B_M$ are independent. Arguing as in the the derivation of (4.5) yields that

$$f_B(s) = f_{K_0}(1 - p_G + sp_G) f_M(f_D(1 - p_G + sp_G)), \tag{4.6}$$

where now K_0 is D in the initial generation (because of how i_* was chosen) and $\tilde{D} - 1$ in subsequent generations.

In order to determine f_M we use equation (3.5) of Ball and Neal (2002), which gives a triangular system of linear equations whose solution is the mass function of M, from which one can easily calculate the PGF. Note that (4.6) holds for any choice of infectious period distribution. It is easily verified that in the fixed infectious period case $T \stackrel{\mathcal{D}}{=} M$, so $f_{\tilde{B}}(s) = f_{\tilde{C}}(s)$ and, if the initial infective is chosen uniformly at random from the population, $f_B(s) = f_C(s)$; and in the zero or infinite infectious period case $M \sim \mathrm{Bin}(n-1, p)$, where $p = \mathbb{P}(I = \infty)$, whence $f_M(s) = (1 - p + ps)^{n-1}$.

5 Numerical results

We now explore, numerically, some of the features of our model and investigate how they depend on some of its parameters. As a way of examining how the household size n affects the model, Figure 1 shows the critical values of the per-pair global contact rate λ_G and the per-individual local contact rate $\lambda_L(n-1)$, above which the epidemic is supercritical, for several household sizes, with the degree distribution and infectious period distribution fixed. Note that the expected total rate of global contacts per individual remains constant over these plots since D is held fixed. Note also that if $\lambda_L = 0$ then n is immaterial, as is λ_L when $n = 1$. In these situations there is no local contact, so we recover the standard network model and the critical value of λ_G is at the point the plotted lines converge to as $\lambda_L \to 0$. The plot reflects

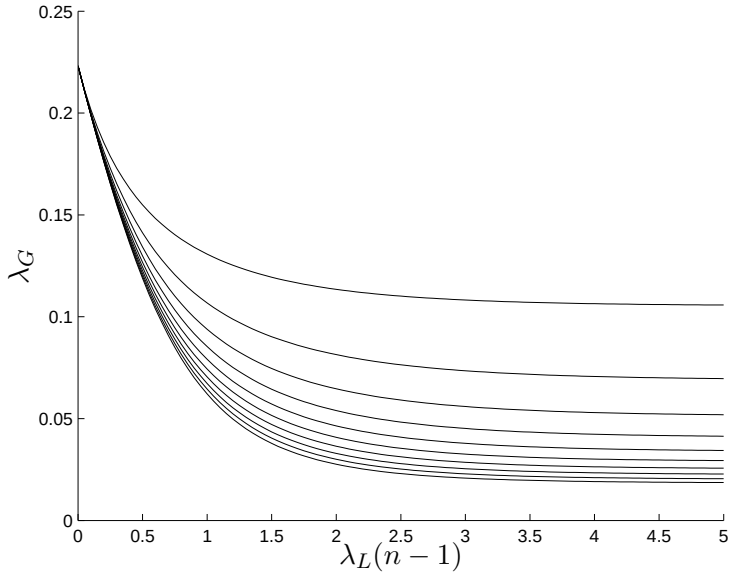

Figure 1: Critical values of λ_G and $\lambda_L(n-1)$ above which the epidemic is supercritical, for $n = 2, 3, \ldots, 10$ (top to bottom in the plot). Other parameters are $I \equiv 1$ and $D \sim \text{Poi}(5)$ (i.e. Poisson with mean 5).

the fact that, even as the per-individual total contact rate remains constant, increasing the household size spreads the potential infectious contacts over a larger number of neighbours, thus avoiding repeated contacts with the same individual and increasing the spread of the disease. We also observe that, fixing D and letting $\lambda_L \to \infty$, the critical value of λ_G tends to that for the standard network model with the same infectious period distribution and degree distribution $\sum_{i=1}^{n} D_i$, where the D_i are independent copies of D. This is because, in this limit, once an individual is infected the whole household that it is in necessarily becomes infected, and is easily verified using (4.4).

Perhaps the most interesting aspect of this model to explore is the dependence of its behaviour on the distribution of D, the number of global neighbours of a typical individual. Considerable research, conjecture and discussion has gone into trying to determine distributions which capture the features of many real life contact networks—Section III.C of Newman (2003) has an extensive list of references. In Figure 2 we investigate the probability of a major outbreak in our epidemic model for various distributions D with different properties, in particular different tail behaviours. We use the standard Poisson and geometric (with support including 0) distributions, as

well as an almost surely constant degree and two variants of heavy-tailed distributions. The first has mass function

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

and the second, with mass function $p_k \propto k^{-a}\mathrm{e}^{-k/\kappa}$ $(k = 1, 2, \ldots)$, is a power law with exponential cut-off which has gained much attention in recent physics literature. We denote these distributions by $\mathrm{Pow}(k_*, a)$ and $\mathrm{PowC}(\kappa, a)$, respectively. The behaviour of these plots for relatively small

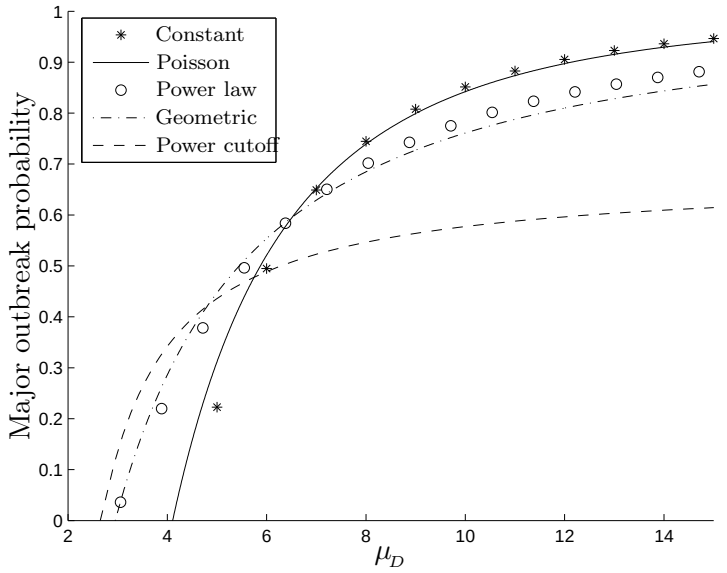

Figure 2: The probability of a major outbreak versus μ_D for different classes of degree distribution D. The distribution labelled 'Power law' is $\mathrm{Pow}(k_*, 7/2)$, for $k_* = 5, 6, \ldots, 18$ and the distribution labelled 'Power cutoff' is $\mathrm{PowC}(\kappa, 3/2)$, for $\kappa \in [10, 485]$ (smaller values of k_* or κ yield subcritical epidemics). The other parameters of the model are $n = 3$, $I \equiv 1$, $\lambda_L = 1$ and $\lambda_G = 1/10$.

values of μ_D (where the model is close to critical) is largely determined by the probability of D taking very large values, i.e. the tail of the distribution, as this dictates what opportunity the disease might have to really 'take hold'; however when μ_D is large the behaviour of D at small values is more important, as the epidemic can usually move quite freely and this determines the chance that it might be contained by the network structure.

We also briefly investigate whether our asymptotic methods give reasonable approximations to the quantities of interest in finite populations. We estimate the probability and expected relative final size of a major outbreak in finite populations from simulations and compare these to the results we get from our asymptotic analysis. Each simulation consists of generating a random network and running *one* epidemic on it. Figure 3 shows estimates of these quantities of interest for increasing numbers m of households together with the theoretical ($m = \infty$) values for two choices of degree distribution. The estimates of the major outbreak probability are based on 10,000 simulations for each parameter combination and those that result in a major outbreak are then used to estimate the expected relative final size. We have plotted point estimates of the quantities of interest, together with error bounds based on ± 2 standard errors (SE) of the estimator. For the probability of a major outbreak, estimated as $\hat{p}$, $\mathrm{SE} = [\hat{p}(1-\hat{p})/n_0]^{1/2}$, where $n_0 = 10,000$ is the number of simulations. For the relative final size, $\mathrm{SE} = \hat{\sigma} n_1^{-1/2}$, where $\hat{\sigma}^2$ is the sample variance of the relative final sizes and n_1 is the number of simulations that resulted in a major outbreak.

Note that in small finite populations the determination of a cutoff for whether a particular final size constitutes a major outbreak is practically impossible; only once the population size is sufficiently large (for m larger than about 100 in our simulations) does the distinction become clear. In our calculations we have used a cutoff of 0.15 of the population size, this being determined by inspecting histograms of the relative final size of the simulations. Also note that the vertical scale of plots (a) and (c) is different from that of plots (b) and (d). Figure 3 shows that our asymptotic results give good approximations for these quantities of interest for populations of only a few hundred households. Though the asymptotic values of both the major outbreak probability and the expected relative final size seem to consistently overestimate these values for the finite populations (as one would expect since the approximating branching process treats each global infection as an infection of a previously uninfected household, thus overestimating disease spread), even for populations of only 100 households the relative error is much less than 5%. It also seems that having a heavy-tailed degree distribution may make the convergence to the asymptotic value a little slower (compare plots (b) and (d) at around 200–500 households), but the effect seems to be only very slight. Another interesting observation is that the relative final size seems to be appreciably more efficiently estimated by our simulation methods than the probability of a major outbreak. This is owing (at least in part) to the fact that from each simulation we simply observe the occurrence or otherwise of a major outbreak—one observation of the forward process—whereas when a major outbreak does occur, the proportion infected

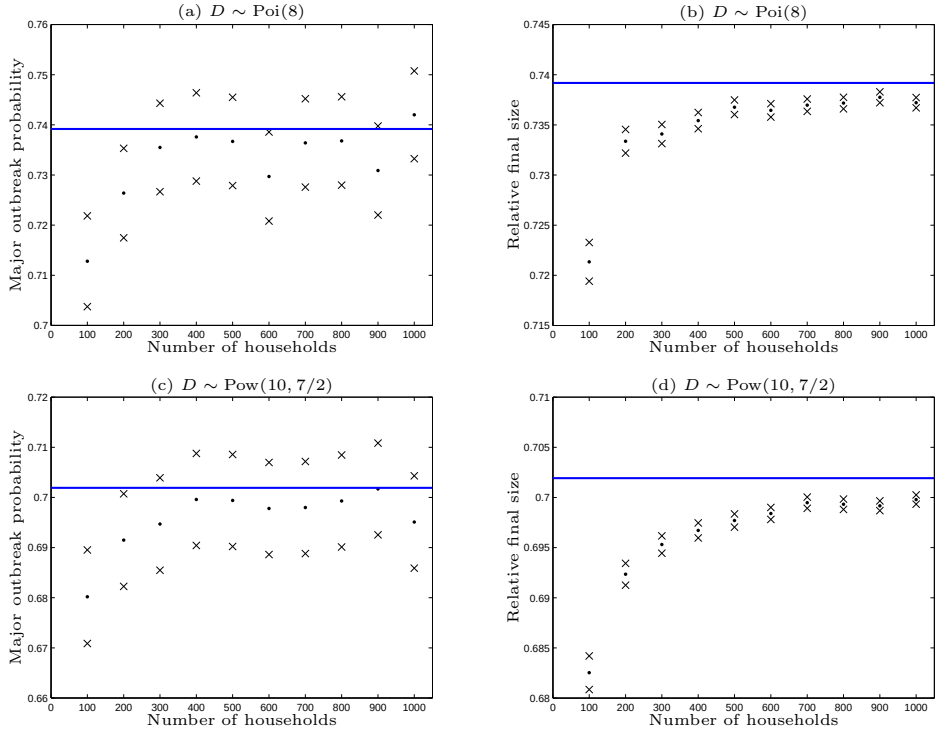

Figure 3: Comparison of simulation estimates of major outbreak probability and expected relative final size for finite populations with asymptotic results. The Poisson degree distribution (plots (a) and (b)) has $\mu_D = \sigma_D^2 = 8$ and the power law distribution (plots (c) and (d)) has $\mu_D \approx 8.04$ and $\sigma_D^2 \approx 96$. Other parameters are $n = 3$, $I \equiv 1$, $\lambda_L = 1$ and $\lambda_G = 1/10$.

has information about the susceptibility set of every initial susceptible in the population—many (highly correlated) observations of the backward process.

6 Proofs

6.1 Overview

In this section we provide a fully rigorous justification of the results discussed in Section 3 concerning the threshold behaviour of the epidemic model and its final outcome in the event of a major outbreak. This subsection gives a brief outline of our methods of proof. The starting point is a sequence $\boldsymbol{D} = (D_1, D_2, \ldots)$ of independent copies of D. For $m = 1, 2, \ldots$, $(D_1, D_2, \ldots, D_{mn})$ is used to give the degrees of the mn individuals in a population of m households. We then define a realisation of the epidemic, $E^{(m)}$ say, viewed on a generation basis, and a realisation of an approximating branching process, say $Y^{(m)} = (Y_k^{(m)}, k = 0, 1, \ldots)$ (see Section 6.2). In $E^{(m)}$ the network is formed, i.e. the half-edges are paired up, as the epidemic progresses. The branching process $Y^{(m)}$ is similar to the branching process, Y say, described in Section 3.1, except the empirical distribution of the degrees $D_1, D_2, \ldots, D_{mn}$ is used in place of the degree distribution D. The epidemic $E^{(m)}$ and approximating branching process $Y^{(m)}$ are coupled so that they coincide until a random number, $\tau^{(m)} + 1$, of households have been infected in $E^{(m)}$. It is shown that $\mathbb{P}(\tau^{(m)} > k) \to 1$ as $m \to \infty$ for all $k \in \mathbb{Z}_+$, so $\hat{Z}^{(m)}$, the number of households infected in $E^{(m)}$, and $\hat{Y}^{(m)}$, the total progeny of $Y^{(m)}$, have the same limiting distribution as $m \to \infty$. (We use $\mathbb{Z}_+$ to denote the positive integers including 0 and $\mathbb{N}$ to denote the strictly positive integers.) Now, $Y^{(m)}$ converges in distribution to Y as $m \to \infty$, so $\hat{Z}^{(m)}$ is asymptotically distributed as $\hat{Y}$, the total progeny of Y (see Theorem 1), thus providing a formal justification of the threshold behaviour described in Section 3.1.

Suppose now that $R_* > 1$, so that major outbreaks are possible. Let $t_m = \lfloor 2 \log\log m / \log R_* \rfloor$, where, for $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the greatest integer $\leq x$. We show (cf. Lemma 7) that there exists $\beta > 1$ such that $\lim_{m\to\infty} \mathbb{P}(\log m < Z_{t_m}^{(m)} < (\log m)^\beta) = \mathbb{P}(\hat{Y} = \infty)$, where $Z_{t_m}^{(m)}$ is the number of infectious households in generation t_m of $E^{(m)}$. It follows that, with probability tending to 1 as $m \to \infty$, a major outbreak has at least $\log m$ and at most $(\log m)^\beta$ infectious households in generation t_m.

We next consider the probability that a typical individual, i^* say, that is susceptible at time t_m in $E^{(m)}$ ultimately becomes infected. We do this by stopping the construction of $E^{(m)}$ at time t_m, leaving the $Z_{t_m}^{(m)}$ infectious

('live') half-edges unconnected, and constructing the susceptibility set, $\mathcal{S}^{(m)}$ say, of i^* in 'generations' as described in Section 3.2, pairing up the half-edges as we construct the susceptibility set. If at any point in the construction of $\mathcal{S}^{(m)}$ a half-edge is paired up with one of the $Z_{t_m}^{(m)}$ live half-edges from the epidemic then i^* is ultimately infected, otherwise i^* is not infected by the epidemic. Note that for any individual, i say, in $\mathcal{S}^{(m)}$ we need to explore all of i's global neighbours (and not just those that join $\mathcal{S}^{(m)}$), since if *any* half-edge emanating from i is paired with one of the $Z_{t_m}^{(m)}$ live half-edges then i^* is ultimately infected. Thus we need to construct simultaneously $\mathcal{A}^{(m)}$, the set of global neighbours of $\mathcal{S}^{(m)}$, also on a generation basis.

Let $(S^{(m)}, A^{(m)}) = ((S_k^{(m)}, A_k^{(m)}), k = 0, 1, \ldots)$ describe the number of households in successive generations of $(\mathcal{S}^{(m)}, \mathcal{A}^{(m)})$. In Section 6.2, we construct realisations of $(S^{(m)}, A^{(m)})$ and an approximating two-type branching process $(X^{(m)}, X_A^{(m)}) = ((X_k^{(m)}, X_{Ak}^{(m)}), k = 0, 1, \ldots)$. The process $X^{(m)}$ is a single-type branching process that is similar to the branching process, X say, described in Section 3.2, except, as with $Y^{(m)}$, the empirical distribution of $D_1, D_2, \ldots, D_{mn}$ is used instead of the degree distribution D. The process $X_A^{(m)}$ corresponds to global neighbours of $\mathcal{S}^{(m)}$ who are not in $\mathcal{S}^{(m)}$; individuals in $X_A^{(m)}$ have no offspring. The processes $(S^{(m)}, A^{(m)})$ and $(X^{(m)}, X_A^{(m)})$ are coupled so that they coincide until $\bar{\tau}^{(m)}+1$ households have joined $\mathcal{S}^{(m)} \cup \mathcal{A}^{(m)}$, where $\mathbb{P}(\bar{\tau}^{(m)} > k) \to 1$ as $m \to \infty$ for all $k \in \mathbb{Z}_+$. Let $\hat{W}^{(m)}$ and $\hat{W}_A^{(m)}$ denote the number of households in $\mathcal{S}^{(m)}$ and $\mathcal{A}^{(m)}$, respectively, and let $\hat{X}^{(m)}$ and $\hat{X}$ denote the total progenies of $X^{(m)}$ and X, respectively. As with $E^{(m)}$, $\hat{W}^{(m)}$ and $\hat{X}^{(m)}$ have the same limiting distribution as $m \to \infty$, which, since $X^{(m)}$ converges in distribution to X as $m \to \infty$, is given by the distribution of $\hat{X}$. Now, for any $k \in \mathbb{N}$, if $\hat{W}^{(m)} + \hat{W}_A^{(m)} \leq k$ then the probability that $\mathcal{S}^{(m)}$ intersects with one of the $Z_{t_m}^{(m)}$ live half-edges tends to 0 as $m \to \infty$ (since a major outbreak has at most $(\log m)^\beta$ infectious households at generation t_m of the forward process), so the limiting (as $m \to \infty$) probability that i^* is ultimately infected by a major outbreak is at most $\mathbb{P}(\hat{X} = \infty)$. (Note that $(X^{(m)}, X_A^{(m)})$ goes extinct if and only if $X^{(m)}$ goes extinct.)

We also construct, for all sufficiently small $\varepsilon \in (0, 1)$, a branching process ${}_\varepsilon X^{(m)}$, which is a lower bound for $S^{(m)}$ as long as $\hat{W}^{(m)} \leq \varepsilon m$; whence $\mathbb{P}(\hat{W}^{(m)} > \varepsilon m) \geq \mathbb{P}({}_\varepsilon \hat{X}^{(m)} = \infty)$, where ${}_\varepsilon \hat{X}^{(m)}$ denotes the total progeny of ${}_\varepsilon X^{(m)}$. As $m \to \infty$, ${}_\varepsilon \hat{X}^{(m)}$ converges in distribution to ${}_\varepsilon \hat{X}$, the total progeny of a branching process ${}_\varepsilon X$ say. Moreover, for any $\varepsilon > 0$, if $\hat{W}^{(m)} > \varepsilon m$ the probability that $\mathcal{S}^{(m)}$ intersects one of the $Z_{t_m}^{(m)}$ live half-edges tends to 1 as $m \to \infty$ (since a major outbreak has at least $\log m$ infectious households at generation t_m of the forward process), so the limiting probability that i^* is ultimately infected is at least $\mathbb{P}({}_\varepsilon \hat{X} = \infty)$. Furthermore, $\mathbb{P}({}_\varepsilon \hat{X} = \infty) \to$

$\mathbb{P}(\hat{X} = \infty)$ as $\varepsilon \downarrow 0$, which, combined with the result described at the end of the previous paragraph, shows that the probability that i^* is ultimately infected by a major outbreak tends to $\mathbb{P}(\hat{X} = \infty)$ as $m \to \infty$ (see Theorem 2). It follows that the expected proportion of the population that are infected by a major outbreak also tends to $\mathbb{P}(\hat{X} = \infty)$ as $m \to \infty$ (see Corollary 2).

Our results are proved by conditioning on the degree sequence $\boldsymbol{D}$ and showing that they hold for $\mathbb{P}$-almost all $\boldsymbol{D}$. The unconditional results then follow using the dominated convergence theorem. As remarked above, the network of global contacts is now constructed as the epidemic/susceptibility set evolves, not a priori as in our model description in Section 2. This implicitly means that, rather than conditioning on the total number of half-edges $\sum_{i=1}^{mn} D_i$ being even, we simply ignore the single left-over half-edge in the event of $\sum_{i=1}^{mn} D_i$ being odd. This small change does not affect the asymptotic results as $m \to \infty$ (cf. van der Hofstad *et al.* (2007, Section 1.1)).

The remainder of this section is organised as follows. The main constructions are described in Section 6.2, with the epidemics $E^{(m)}$ and their approximating branching processes being described in Section 6.2.1 and the susceptibility set processes and their approximating branching processes being described in Section 6.2.2. Some notation concerning the offspring distributions of various branching processes is given in Section 6.2.3. Section 6.3 contains some preliminary results, and the main results are given in Sections 6.4 and 6.5, which analyse the epidemics $E^{(m)}$ and the susceptibility set processes $(S^{(m)}, A^{(m)})$, respectively.

6.2 Construction of approximating branching processes

Let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ be a probability space, on which is defined a sequence $\boldsymbol{D} = (D_1, D_2, \ldots)$ of independent random variables, each distributed according to the degree distribution D. Also let $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$ be a probability space, on which are defined the following mutually independent random quantities:

(i) for every $(\boldsymbol{d}, j) = ((d_1, d_2, \ldots, d_n), j) \in \mathbb{Z}_+^n \times \{1, 2, \ldots, n\}$, a sequence of random variables $\Phi_1^{(\boldsymbol{d},j)}, \Phi_2^{(\boldsymbol{d},j)}, \ldots$, which are independent copies of the random variable $\Phi^{(\boldsymbol{d},j)}$ defined below.

(ii) for every $(\boldsymbol{d}, j) \in \mathbb{Z}_+^n \times \{1, 2, \ldots, n\}$, a sequence of random variables $(\Psi_1^{(\boldsymbol{d},j)}, \Psi_{A1}^{(\boldsymbol{d},j)})$, $(\Psi_2^{(\boldsymbol{d},j)}, \Psi_{A2}^{(\boldsymbol{d},j)}), \ldots$, which are independent copies of the random variable $(\Psi^{(\boldsymbol{d},j)}, \Psi_A^{(\boldsymbol{d},j)})$ also defined below.

We also require other random variables defined on $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, but these are described only informally because the detail is unnecessary for our proofs.

The random variable $\Phi^{(\boldsymbol{d},j)}$ describes the number of global neighbours with which infectious contact is made by members of a household of individuals with degrees given by $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)$ in which individual j is initially infected and is defined as follows. Let G be the random directed graph on the vertices $V = \{1, 2, \ldots, n\}$ obtained as follows. For each vertex i we take an independent realisation, I_i say, of the infectious period distribution I and then put an arc from i to each other vertex in V independently with probability $1 - \mathrm{e}^{-\lambda_L I_i}$. Given G, let $C_1, C_2, \ldots, C_n$ be independent random variables with $C_i \,|\, I_1, I_2, \ldots, I_n \sim \mathrm{Bin}(d'_i, 1 - \mathrm{e}^{\lambda_G I_i})$, where $d'_i = d_i$ if $i \neq j$ and $d'_j = d_j - 1$. Then $\Phi^{(\boldsymbol{d},j)} = \sum_{i=1}^{n} \mathbb{1}_{\{j \rightsquigarrow i\}} C_i$, where $j \rightsquigarrow i$ denotes the event that there is a path from vertex j to vertex i in G (with the convention that $i \rightsquigarrow i$).

In a similar manner, the two components of the random variable $(\Psi^{(\boldsymbol{d},j)}, \Psi_A^{(\boldsymbol{d},j)})$ describe the number of global neighbours of the local susceptibility set of individual j in a household of individuals with degrees given by $\boldsymbol{d} = (d_1, d_2, \ldots, d_n)$ that do and do not make global infectious contact with their neighbour in that susceptibility set. To this end, let G be the random graph described above and, conditional on G, let $B_1, B_2, \ldots, B_n$ be independent random variables with $B_i \sim \mathrm{Bin}(d'_i, p_G)$, where $d'_1, d'_2, \ldots, d'_n$ are as above and $p_G = 1 - \phi(\lambda_G)$. We then have $(\Psi^{(\boldsymbol{d},j)}, \Psi_A^{(\boldsymbol{d},j)}) = \sum_{i=1}^{n} \mathbb{1}_{\{i \rightsquigarrow j\}} (B_i, d'_i - B_i)$.

We now introduce some further notation. For $k = 1, 2, \ldots$, let $\boldsymbol{D}_k = (D_{k1}, D_{k2}, \ldots, D_{kn})$, where, for $i = 1, 2, \ldots, n$, $D_{ki} = D_{(k-1)n+i}$ is the degree of the ith individual in the kth household. Let $H_k = \sum_{i=1}^{n} D_{ki}$ denote the total degree of the kth household. Lastly, denote by $\mu_H^{(m)} = \frac{1}{m} \sum_{i=1}^{m} H_i$ the (empirical) mean number of edges emanating from each of the first m households.

The epidemic, susceptibility sets and approximating branching processes are defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1, \mathcal{F}_1, \mathbb{P}_1) \times (\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$. Our construction and most of our calculations will henceforth be conditional on the degree sequence $\boldsymbol{D}$. To this end, we denote $\mathbb{P}(\cdot \,|\, \boldsymbol{D})$ by $\mathbb{P}_{\boldsymbol{D}}(\cdot)$ and similarly $\mathbb{E}[\cdot \,|\, \boldsymbol{D}]$ by $\mathbb{E}_{\boldsymbol{D}}[\cdot]$. Conditional on this degree sequence and for every $m = 1, 2, \ldots$, we now describe the construction of a branching process, $Y^{(m)}$, which approximates the early stages of the spread of the epidemic amongst households $1, 2, \ldots, m$; then another (two-type) branching process, $(X^{(m)}, X_A^{(m)})$, which approximates the 'early growth' of the susceptibility set (and its global neighbours) of a typical initially susceptible individual in that population.

6.2.1 The forward processes.

We first describe the branching process $Y^{(m)}$. Set $Y_0^{(m)} = 1$ and choose an individual uniformly at random from $1, 2, \ldots, mn$. Suppose it is individual $\iota \in \{1, 2, \ldots, n\}$ of household $\Delta_0 \in \{1, 2, \ldots, m\}$. Then $Y_1^{(m)} = \Phi_1^{(\boldsymbol{D}_{\Delta_0} + e_\iota, \iota)}$, where e_i is the unit n-vector with a 1 in the ith position. For subsequent generations $k \geq 2$, we continue the construction as follows. For each $j = 1, 2, \ldots, Y_{k-1}^{(m)}$, sample a half-edge uniformly at random from the $m\mu_H^{(m)}$ half-edges in the population and, supposing it emanates from individual ι of household Δ, set $Y_{kj}^{(m)} = \Phi_{\nu(\Delta,\iota)+1}^{(\boldsymbol{D}_\Delta, \iota)}$, where $\nu(\Delta, \iota)$ is the number of times we have sampled previously from the sequence $\Phi_1^{(\boldsymbol{D}_\Delta, \iota)}, \Phi_2^{(\boldsymbol{D}_\Delta, \iota)}, \ldots$. Lastly, set

$$Y_k^{(m)} = \sum_{j=1}^{Y_{k-1}^{(m)}} Y_{kj}^{(m)}.$$

The branching process $Y^{(m)}$ and the epidemic process $E^{(m)}$ can be coupled by using the same $\boldsymbol{D}$, Φ's and uniformly random samples. However, the coupling breaks down as soon as a half-edge is sampled that emanates from a household that either has been used previously in the epidemic or is a neighbour of such a previously used household. If a previously used half-edge is sampled then in $E^{(m)}$ another half-edge needs to be sampled. If an unused half-edge that emanates from a previously used household is sampled then in $E^{(m)}$ the spread of the epidemic within that household is different from in $Y^{(m)}$ since there are fewer susceptibles. Finally, if a half-edge emanating from a household neighbouring a household previously used in $E^{(m)}$ is sampled then the spread of the epidemic from that household is in general different from that in $Y^{(m)}$, since the (effective) degree distribution of individuals in that household may be different from that assumed in $Y^{(m)}$. (When constructing $E^{(m)}$ one needs also to pair up non-infectious half-edges from infectious individuals.) In all of these cases the construction of $E^{(m)}$ can be continued appropriately but the detail is not important for our purposes. However, we do need a bound on the size of, and number of half-edges that emanate from, the 'bad set' of households that must be avoided in order that $Y^{(m)}$ and $E^{(m)}$ remain coupled. To that end we describe another branching process $T^{(m)} = (T_k^{(m)}, k = 1, 2, \ldots)$, which provides such a bound.

Let $T_0^{(m)}$ be the total degree of the initial household in $Y^{(m)}$, so $T_0^{(m)} = H_{\Delta_0}$, where Δ_0 is as above. For $k = 1, 2, \ldots$, $T_k^{(m)}$ is determined as follows. For each $j = 1, 2, \ldots, T_{k-1}^{(m)}$, a half-edge is sampled uniformly at random from the $m\mu_H^{(m)}$ half-edges in the population, say this half-edge emanates from household Δ_j, and then put $T_{kj}^{(m)} = H_{\Delta_j} - 1$. Finally, set $T_k^{(m)} = \sum_{j=1}^{T_{k-1}^{(m)}} T_{kj}^{(m)}$.

The processes $Y^{(m)}$, $E^{(m)}$ and $T^{(m)}$ can be coupled in an obvious fashion so that their sampled half-edges correspond. Let $\hat{T}_k^{(m)} = \sum_{l=0}^{k} T_l^{(m)}$ $(k = 0, 1, \ldots)$ be the total progeny of $T^{(m)}$ up to generation k. Then $2\hat{T}_{k+1}^{(m)}$ provides an upper bound for the number of half-edges that emanate from (and hence also for the size) of the bad set of households in generation k of $E^{(m)}$. The index $k+1$ arises because the bad set consists of not just all households infected up to generation k of $E^{(m)}$ but also their neighbouring households. The factor 2 arises because $T^{(m)}$ does not count the receiving half-edge when the half-edges are paired up.

The above construction of $Y^{(m)}$ (and implicitly $E^{(m)}$) is continued for a fixed number of generations, t_m, and $T^{(m)}$ is continued for t_m+1 generations. (Of course, some or all of these processes may die out beforehand.)

6.2.2 The backward processes.

The two-type branching process $(X^{(m)}, X_A^{(m)})$ is defined analogously to $Y^{(m)}$ except the random variables $(\Psi_i^{(\boldsymbol{d},j)}, \Psi_{Ai}^{(\boldsymbol{d},j)})$ are used instead of $\Phi_i^{(\boldsymbol{d},j)}$ (recall that there are no offspring in $X_A^{(m)}$). The process $X^{(m)}$ approximates the growth, described by generations as in Section 3.2, of the susceptibility set of an individual chosen uniformly at random from all susceptible individuals at time t_m in the epidemic process $E^{(m)}$ and $X_A^{(m)}$ approximates the number of global neighbours of this susceptibility set, also on a generation basis. The processes $(X^{(m)}, X_A^{(m)})$ and $(S^{(m)}, A^{(m)})$ can be coupled in a similar fashion to that used for $Y^{(m)}$ and $E^{(m)}$, though note that now the coupling breaks down if a sampled half-edge emanates from either (i) a household previously used in the susceptibility set, (ii) a household neighbouring such a household, or (iii) a household or neighbour of a household used in the forward process up to time t_m. Also note that this coupling may break down at generation 0 (if the initial individual is in a household that is either infected in $E^{(m)}$ or a neighbour of a household infected in $E^{(m)}$). As with the epidemic process $E^{(m)}$, the construction of $(S^{(m)}, A^{(m)})$ can be continued appropriately after the coupling breaks down but we do not require such detail.

6.2.3 Further notation and limiting processes.

For $m = 1, 2, \ldots$, let $\boldsymbol{c}^{(m)} = (c_0^{(m)}, c_1^{(m)}, \ldots)$ and $\tilde{\boldsymbol{c}}^{(m)} = (\tilde{c}_0^{(m)}, \tilde{c}_1^{(m)}, \ldots)$ denote the offspring distributions of the initial individual and all subsequent individuals, respectively, in $Y^{(m)}$. For $\boldsymbol{d} = (d_1, d_2, \ldots, d_n) \in \mathbb{Z}_+^n$, let

$$p_{\boldsymbol{d}}^{(m)} = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}_{\{\boldsymbol{D}_i = \boldsymbol{d}\}} \qquad \text{and} \qquad \tilde{p}_{\boldsymbol{d}}^{(m)} = \frac{|\boldsymbol{d}|}{m\mu_H^{(m)}} \sum_{i=1}^{m} \mathbb{1}_{\{\boldsymbol{D}_i = \boldsymbol{d}\}},$$

where $|\boldsymbol{d}| = \sum_{j=1}^{n} d_j$. Then the 'household type' (i.e. the degrees of individuals within the household) of the initial individual in $Y^{(m)}$ is distributed according to $p_{\boldsymbol{d}}^{(m)}$ ($\boldsymbol{d} \in \mathbb{Z}_+^n$) and the household type of any subsequent individual is distributed according to $\tilde{p}_{\boldsymbol{d}}^{(m)}$ ($\boldsymbol{d} \in \mathbb{Z}_+^n$). It follows that, for $k = 0, 1, \ldots$,

$$c_k^{(m)} = \sum_{\boldsymbol{d} \in \mathbb{Z}_+^n} p_{\boldsymbol{d}}^{(m)} \, \mathbb{P}(\Phi_{\boldsymbol{d}} = k) \qquad \text{and} \qquad \tilde{c}_k^{(m)} = \sum_{\boldsymbol{d} \in \mathbb{Z}_+^n} \tilde{p}_{\boldsymbol{d}}^{(m)} \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{d}} = k), \quad (6.1)$$

where $\Phi_{\boldsymbol{d}}$ and $\tilde{\Phi}_{\boldsymbol{d}}$ are random variables with distributions given by

$$\mathbb{P}(\Phi_{\boldsymbol{d}} = k) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{P}(\Phi^{(\boldsymbol{d}+e_i,i)} = k) \qquad (k = 0, 1, \ldots, |\boldsymbol{d}|), \quad (6.2)$$

and

$$\mathbb{P}(\tilde{\Phi}_{\boldsymbol{d}} = k) = \sum_{i=1}^{n} \frac{d_i}{|\boldsymbol{d}|} \, \mathbb{P}(\Phi^{(\boldsymbol{d},i)} = k) \qquad (k = 0, 1, \ldots, |\boldsymbol{d}| - 1). \quad (6.3)$$

For $m = 1, 2, \ldots$, the offspring distributions of the initial and subsequent individuals in $X^{(m)}$, $\boldsymbol{b}^{(m)}$ and $\tilde{\boldsymbol{b}}^{(m)}$, are defined analagously to $\boldsymbol{c}^{(m)}$ and $\tilde{\boldsymbol{c}}^{(m)}$, using (6.1)–(6.3) with Φ replaced by Ψ and $\tilde{\Phi}$ by $\tilde{\Psi}$ throughout. Replacing Φ by (Ψ, Ψ_A) and $\tilde{\Phi}$ by $(\tilde{\Psi}, \tilde{\Psi}_A)$ throughout gives the offspring distributions associated with the two-type process $(X^{(m)}, X_A^{(m)})$.

Further, for $m = 1, 2, \ldots$, let $\boldsymbol{r}^{(m)} = (r_0^{(m)}, r_1^{(m)}, \ldots)$ denote the distribution of the number of initial ancestors and $\tilde{\boldsymbol{r}}^{(m)} = (\tilde{r}_0^{(m)}, \tilde{r}_1^{(m)}, \ldots)$ denote the offspring distribution of both the ancestors and any subsequent individuals in $T^{(m)}$. Then, for $k = 0, 1, \ldots$,

$$r_k^{(m)} = \sum_{\{\boldsymbol{d} \in \mathbb{Z}_+^n \,:\, |\boldsymbol{d}| = k\}} p_{\boldsymbol{d}}^{(m)} \qquad \text{and} \qquad \tilde{r}_k^{(m)} = \sum_{\{\boldsymbol{d} \in \mathbb{Z}_+^n \,:\, |\boldsymbol{d}| = k+1\}} \tilde{p}_{\boldsymbol{d}}^{(m)}.$$

For $m = 1, 2, \ldots$, let $\mu_c^{(m)} = \sum_{k=1}^{\infty} k c_k^{(m)}$ be the mean of the empirical distribution $\boldsymbol{c}^{(m)}$, and define $\tilde{\mu}_c^{(m)}, \mu_b^{(m)}, \tilde{\mu}_b^{(m)}, \mu_r^{(m)}$ and $\tilde{\mu}_r^{(m)}$ analogously.

In Section 6.3 we prove that the offspring distributions of $Y^{(m)}$, $X^{(m)}$ and $T^{(m)}$ and the distribution of the number of ancestors in $T^{(m)}$ converge almost surely as $m \to \infty$ to those of branching processes we denote by Y, X and T respectively. To that end, for $\boldsymbol{d} \in \mathbb{Z}_+^n$, let $p_{\boldsymbol{d}} = \prod_{i=1}^{n} p_{d_i}$ and $\tilde{p}_{\boldsymbol{d}} = p_{\boldsymbol{d}} \, |\boldsymbol{d}| \, / n\mu_D$. (Recall that $p_k = \mathbb{P}(D = k)$ $(k = 0, 1, \ldots)$ and $\mu_D = \sum_{k=1}^{\infty} k p_k$.) Also, for $k = 0, 1, \ldots$, let

$$p_H(k) = \sum_{\{\boldsymbol{d} \in \mathbb{Z}_+^n \,:\, |\boldsymbol{d}| = k\}} p_{\boldsymbol{d}} = \mathbb{P}(D_1 + D_2 + \cdots + D_n = k) = \mathbb{P}(H_1 = k)$$

and, for $k = 1, 2, \ldots$, let $\tilde{p}_H(k) = kp_H(k)/n\mu_D$. Now, for $k = 0, 1, \ldots$, let c_k be defined analogously to $c_k^{(m)}$ but with $p_{\boldsymbol{d}}^{(m)}$ replaced by $p_{\boldsymbol{d}}$, and define $\tilde{c}_k, b_k$ and $\tilde{b}_k$ similarly. Also, for $k = 0, 1, \ldots$, let $r_k = p_H(k)$ and $\tilde{r}_k = \tilde{p}_H(k+1)$. Let $\boldsymbol{c} = (c_0, c_1, \ldots)$ and define $\tilde{\boldsymbol{c}}$, $\boldsymbol{b}$, $\tilde{\boldsymbol{b}}$, $\boldsymbol{r}$ and $\tilde{\boldsymbol{r}}$ similarly. Let $\mu_c = \sum_{k=1}^{\infty} kc_k$ and define $\tilde{\mu}_c$, μ_b, $\tilde{\mu}_b$, μ_r and $\tilde{\mu}_r$ in the obvious fashion.

Let $Y = (Y_0, Y_1, \ldots)$, $X = (X_0, X_1, \ldots)$ and $T = (T_0, T_1, \ldots)$ be the branching processes $\mathrm{BP}(1, \boldsymbol{c}, \tilde{\boldsymbol{c}})$, $\mathrm{BP}(1, \boldsymbol{b}, \tilde{\boldsymbol{b}})$ and, in an obvious notation, $\mathrm{BP}(\boldsymbol{r}, \tilde{\boldsymbol{r}}, \tilde{\boldsymbol{r}})$, respectively. Note that the branching processes Y and X are those described in Sections 3.1 and 3.2, respectively. Note especially that, in the notation of Section 3.1, this implies that $\tilde{\mu}_c = R_*$. We also require a two-type branching process (X, A), defined analagously to $(X^{(m)}, X_A^{(m)})$ but again using $p_{\boldsymbol{d}}$ and $\tilde{p}_{\boldsymbol{d}}$ in defining the offspring distribution instead of the empirical versions $p_{\boldsymbol{d}}^{(m)}$ and $\tilde{p}_{\boldsymbol{d}}^{(m)}$.

6.3 Preliminary results

In this section we collect some results required in the analysis of the forward and backward processes. Recall that we have made the assumption that $\sigma_D^2 = \operatorname{Var} D$ is finite (although some results only require $\mu_D < \infty$).

Lemma 1. *There exists $A_1 \in \mathcal{F}_1$, with $\mathbb{P}_1(A_1) = 1$, such that, for all $\omega_1 \in A_1$,*

(i) $\lim_{m\to\infty} \mu_H^{(m)}(\omega_1) = n\mu_D$;

(ii) $\lim_{m\to\infty} p_{\boldsymbol{d}}^{(m)}(\omega_1) = p_{\boldsymbol{d}}$ *and* $\lim_{m\to\infty} \tilde{p}_{\boldsymbol{d}}^{(m)}(\omega_1) = \tilde{p}_{\boldsymbol{d}}$ *for each* $\boldsymbol{d} \in \mathbb{Z}_+^n$;

(iii) $\lim_{m\to\infty} \boldsymbol{c}^{(m)}(\omega_1) = \boldsymbol{c}$, $\lim_{m\to\infty} \tilde{\boldsymbol{c}}^{(m)}(\omega_1) = \tilde{\boldsymbol{c}}$, $\lim_{m\to\infty} \boldsymbol{b}^{(m)}(\omega_1) = \boldsymbol{b}$, $\lim_{m\to\infty} \tilde{\boldsymbol{b}}^{(m)}(\omega_1) = \tilde{\boldsymbol{b}}$, $\lim_{m\to\infty} \boldsymbol{r}^{(m)}(\omega_1) = \boldsymbol{r}$ *and* $\lim_{m\to\infty} \tilde{\boldsymbol{r}}^{(m)}(\omega_1) = \tilde{\boldsymbol{r}}$;

(iv) $\lim_{m\to\infty} \mu_c^{(m)}(\omega_1) = \mu_c$, $\lim_{m\to\infty} \tilde{\mu}_c^{(m)}(\omega_1) = \tilde{\mu}_c$, $\lim_{m\to\infty} \mu_b^{(m)}(\omega_1) = \mu_b$, $\lim_{m\to\infty} \tilde{\mu}_b^{(m)}(\omega_1) = \tilde{\mu}_b$, $\lim_{m\to\infty} \mu_r^{(m)}(\omega_1) = \mu_r$ *and* $\lim_{m\to\infty} \tilde{\mu}_r^{(m)}(\omega_1) = \tilde{\mu}_r$.

Here and henceforth, convergence of a sequence of sequences is interpreted elementwise, so, for example, $\lim_{m\to\infty} \boldsymbol{c}^{(m)} = \boldsymbol{c}$ *means that* $\lim_{m\to\infty} c_k^{(m)} = c_k$ *for each* $k = 0, 1, \ldots$.

Proof. By the strong law of large numbers, there exists $A_2 \in \mathcal{F}_1$ with $\mathbb{P}_1(A_2) = 1$ such that $\lim_{m\to\infty} \mu_H^{(m)}(\omega_1) = n\mu_D$ $(\omega_1 \in A_2)$ and, for each $\boldsymbol{d} \in \mathbb{Z}_+^n$, there exists $A_{\boldsymbol{d}} \in \mathcal{F}_1$ with $\mathbb{P}_1(A_{\boldsymbol{d}}) = 1$ such that $\lim_{m\to\infty} p_{\boldsymbol{d}}^{(m)}(\omega_1) = p_{\boldsymbol{d}}$ $(\omega_1 \in A_{\boldsymbol{d}})$. Let

$A_3 = A_2 \cap \bigcap_{\boldsymbol{d} \in \mathbb{Z}_+^n} A_{\boldsymbol{d}}$. Then $\mathbb{P}_1(A_3) = 1$ and it is easily verified that (i) and (ii) hold for all $\omega_1 \in A_3$, whence (iii) also holds for all $\omega_1 \in A_3$ by Scheffé's theorem (see, for example, Billingsley (1968, p. 224)). Next, consider

$$\begin{aligned}
\tilde{\mu}_c^{(m)} = \sum_{k=1}^{\infty} k \tilde{c}_k^{(m)} &= \sum_{k=1}^{\infty} k \sum_{\boldsymbol{d} \in \mathbb{Z}_+^n} \tilde{p}_{\boldsymbol{d}}^{(m)} \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{d}} = k) \\
&= \sum_{k=1}^{\infty} k \sum_{\boldsymbol{d} \in \mathbb{Z}_+^n} \frac{|\boldsymbol{d}|}{m \mu_H^{(m)}} \sum_{i=1}^{m} \mathbb{1}_{\{\boldsymbol{D}_i = \boldsymbol{d}\}} \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{d}} = k) \\
&= \frac{1}{\mu_H^{(m)}} \cdot \frac{1}{m} \sum_{i=1}^{m} |\boldsymbol{D}_i| \sum_{k=1}^{\infty} k \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{D}_i} = k).
\end{aligned}$$

Now, $\mathbb{P}(\tilde{\Phi}_{\boldsymbol{D}_i} = k) = 0$ for $k \geq |\boldsymbol{D}_i| - 1$, so

$$\mathbb{E}\left[|\boldsymbol{D}_1| \sum_{k=1}^{\infty} k \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{D}_1} = k) \right] \leq \mathbb{E}\left[|\boldsymbol{D}_1| \left(|\boldsymbol{D}_1| - 1 \right) \right] < \infty,$$

as $\sigma_D^2 < \infty$. Thus, by the strong law of large numbers, there exists $A_4 \in \mathcal{F}_1$ with $\mathbb{P}_1(A_4) = 1$ such that, for all $\omega_1 \in A_4$,

$$\begin{aligned}
\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} |\boldsymbol{D}_i(\omega_1)| \sum_{k=1}^{\infty} k \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{D}_i(\omega_1)} = k) &= \mathbb{E}\left[|\boldsymbol{D}_1| \sum_{k=1}^{\infty} k \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{D}_1} = k) \right] \\
&= \sum_{\boldsymbol{d} \in \mathbb{Z}_+^n} p_{\boldsymbol{d}} \, |\boldsymbol{d}| \sum_{k=1}^{\infty} k \, \mathbb{P}(\tilde{\Phi}_{\boldsymbol{d}} = k).
\end{aligned}$$

It follows that $\lim_{m \to \infty} \tilde{\mu}_c^{(m)}(\omega_1) = \tilde{\mu}_c$ for all $\omega_1 \in A_2 \cap A_4$. Similar arguments hold for the other means in (iv) and the lemma is thus proved. □

Remark. Throughout the remainder of the paper, A_1 refers to a set that satisfies the statement of Lemma 1.

The following result concerns the convergence of certain quantities associated with a sequence of branching processes when their offspring distributions converge in distribution.

Lemma 2. *Suppose that $\boldsymbol{a}$, $\tilde{\boldsymbol{a}}$, $\boldsymbol{a}^{(m)}$ and $\tilde{\boldsymbol{a}}^{(m)}$ $(m = 1, 2, \ldots)$ are probability distributions satisfying $\boldsymbol{a}^{(m)} \to \boldsymbol{a}$ and $\tilde{\boldsymbol{a}}^{(m)} \to \tilde{\boldsymbol{a}}$ as $m \to \infty$. Let $Y^{(m)} \sim \mathrm{BP}(1, \boldsymbol{a}^{(m)}, \tilde{\boldsymbol{a}}^{(m)})$ $(m = 1, 2, \ldots)$ and $Y \sim \mathrm{BP}(1, \boldsymbol{a}, \tilde{\boldsymbol{a}})$. Then, denoting by $\hat{Y}^{(m)}$ (respectively $\hat{Y}$) the total progeny of $Y^{(m)}$ (respectively Y),*

(i) $\lim_{m\to\infty} \mathbb{P}(\hat{Y}^{(m)} = k) = \mathbb{P}(\hat{Y} = k)$ *(*$k = 1, 2, \ldots$*);*

(ii) $\lim_{m\to\infty} \mathbb{P}(\hat{Y}^{(m)} = \infty) = \mathbb{P}(\hat{Y} = \infty)$*, provided* $\tilde{a}_1 \neq 1$.

Proof. Part (i) follows immediately by considering the sum of the probabilities of the finite number of sample paths of $Y^{(m)}$ with $\hat{Y}^{(m)} = k$. Part (ii) is a simple extension of Lemma 4.1 of Britton *et al.* (2007). □

Remarks.

1. The condition in part (ii) of the lemma is in practice only a technical condition which will always hold true. As pointed out by Britton *et al.* (2007), although the case $\tilde{a}_1 = 1$ really can be an exception (for example if $\tilde{a}_0^{(m)} = 1 - \tilde{a}_1^{(m)} = 1/m$), such a scenario is, from an applied viewpoint, decidedly pathological.

2. We sometimes use a slight variant of Lemma 2, where the branching processes are indexed by $\varepsilon \in (0, 1)$ and their offspring distributions converge as $\varepsilon \downarrow 0$. Of course, the analogous results hold, and the proof is exactly the same.

Lastly, we have a result concerning the probability of picking a 'bad' half-edge in our constructions of the forward and backward processes.

Lemma 3. *Suppose that, for each* $m = 1, 2, \ldots$*, we draw elements uniformly at random, with replacement, from the set* $\mathcal{J}^{(m)} = \{1, 2, \ldots, m\mu_H^{(m)}\}$*. Suppose also that, for each* m*, there is an increasing sequence of (random) sets* $\mathcal{J}_1^{(m)} \subset \mathcal{J}_2^{(m)} \subset \cdots \subset \mathcal{J}^{(m)}$ *and at the* i*th pick we wish to avoid picking a member of* $\mathcal{J}_i^{(m)}$*. Denote the* i*th pick by* $\chi_i^{(m)}$ *and let* $\tau^{(m)} = \min\{i : \chi_i^{(m)} \in \mathcal{J}_i^{(m)}\} - 1$ *be the number of picks we make before making a pick from a set we wish to avoid. Suppose further that there exist strictly positive integers* $g(m)$ *and* $h(m)$ *(*$m = 1, 2, \ldots$*) satisfying* $\lim_{m\to\infty} g(m)h(m)m^{-1} = 0$ *and*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(J_{g(m)}^{(m)} \leq h(m)) = 1 \tag{6.4}$$

for all $\omega_1 \in A_1$*, where* $J_i^{(m)} = |\mathcal{J}_i^{(m)}|$*. Then, for all* $\omega_1 \in A_1$*,*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\tau^{(m)} > g(m)) = 1. \tag{6.5}$$

Proof. In view of (6.4), for $\omega_1 \in A_1$,

$$\begin{aligned}
&\liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\tau^{(m)} > g(m)) \\
&\quad = \liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}\left(\tau^{(m)} > g(m) \,\middle|\, J^{(m)}_{g(m)} \le h(m)\right) \mathbb{P}_{\boldsymbol{D}(\omega_1)}\left(J^{(m)}_{g(m)} \le h(m)\right) \\
&\quad \ge \liminf_{m\to\infty} \left(1 - \frac{h(m)}{m\mu^{(m)}_H(\omega_1)}\right)^{g(m)} \mathbb{P}_{\boldsymbol{D}(\omega_1)}\left(J^{(m)}_{g(m)} \le h(m)\right) \\
&\quad \ge \liminf_{m\to\infty} \left(1 - \frac{g(m)h(m)}{m\mu^{(m)}_H(\omega_1)}\right) \mathbb{P}_{\boldsymbol{D}(\omega_1)}\left(J^{(m)}_{g(m)} \le h(m)\right) \\
&\quad = 1,
\end{aligned}$$

using (6.4), Lemma 1(i) and the fact that $g(m)h(m)m^{-1} \to 0$ as $m \to \infty$. The assertion (6.5) then follows. □

6.4 Analysis of forward process

6.4.1 Threshold theorem for the epidemic $E^{(m)}$.

In order to prove a threshold theorem for the epidemic we first establish a bound for the size of the bad set of half-edges after k generations of the epidemic $E^{(m)}$. Recall (from the discussion at the end of Section 6.2.1) that the number of half-edges in this set is bounded by $2\hat{T}^{(m)}_{k+1}$.

Lemma 4. *For all $\omega_1 \in A_1$,*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{T}^{(m)}_k > \log m) = 0 \qquad (k = 1, 2, \ldots).$$

Proof. Fix $\omega_1 \in A_1$. Then note that $\mathbb{E}_{\boldsymbol{D}(\omega_1)}[\hat{T}^{(m)}_k] = \mu^{(m)}_r(\omega_1)\{1 + \tilde{\mu}^{(m)}_r(\omega_1) + (\tilde{\mu}^{(m)}_r(\omega_1))^2 + \cdots + (\tilde{\mu}^{(m)}_r(\omega_1))^k\}$ and also that $\mu^{(m)}_r(\omega_1) \le \tilde{\mu}^{(m)}_r(\omega_1) + 1$. Thus $\mathbb{E}_{\boldsymbol{D}(\omega_1)}[\hat{T}^{(m)}_k] \le (k+1)(\tilde{\mu}^{(m)}_r(\omega_1) + 1)^{k+1}$ and, by Markov's inequality,

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{T}^{(m)}_k(\omega_1) > \log m) \le \frac{k+1}{\log m}(\tilde{\mu}^{(m)}_r(\omega_1) + 1)^{k+1}.$$

The lemma now follows, since $\tilde{\mu}^{(m)}_r(\omega_1) \to \tilde{\mu}_r$ as $m \to \infty$ for all $\omega_1 \in A_1$, by Lemma 1(iv). □

For $m = 1, 2, \ldots$, let $\hat{Z}^{(m)}$ denote the total number of households infected in the epidemic $E^{(m)}$, including the initial household, and let $\hat{Y}^{(m)}$ and $\hat{Y}$ be the total progeny, including the initial individual, of the branching processes $Y^{(m)}$ and Y, respectively. We now show that the total number of households infected in $E^{(m)}$ converges in distribution to the total progeny of Y.

Theorem 1. *For $k = 1, 2, \ldots$,*

(i) for all $\omega_1 \in A_1$, $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k) = \mathbb{P}(\hat{Y} = k)$;

(ii) $\lim_{m\to\infty} \mathbb{P}(\hat{Z}^{(m)} = k) = \mathbb{P}(\hat{Y} = k)$.

Proof. Fix $\omega_1 \in A_1$ and let $\tau^{(m)}$ be the number of households infected by $E^{(m)}$ before a bad half-edge is chosen. Fix $k \in \mathbb{N}$. Then

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k) = \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k\,,\, \tau^{(m)} \le k) + \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k\,,\, \tau^{(m)} > k). \tag{6.6}$$

Let $\mathcal{J}_l^{(m)}$ $(l = 1, 2, \ldots)$ be the set of half-edges we wish to avoid when choosing the lth household to spread the epidemic to. Then $J_k^{(m)} = |\mathcal{J}_k^{(m)}| \le 2\hat{T}_k^{(m)}$, so

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(J_k^{(m)} \le 2\log m) \ge \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{T}_k^{(m)} \le \log m) \to 1$$

as $m \to \infty$, by Lemma 4. Thus, using Lemma 3 with $g(m) = k$ and $h(m) = 2\log m$, $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\tau^{(m)} > k) = 1$. Therefore, $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k\,,\, \tau^{(m)} \le k) = 0$ and, recalling (6.6),

$$\begin{aligned}
\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k) &= \lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} = k\,,\, \tau^{(m)} > k) \\
&= \lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Y}^{(m)} = k\,,\, \tau^{(m)} > k) \\
&= \lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Y}^{(m)} = k) \\
&= \mathbb{P}(\hat{Y} = k),
\end{aligned}$$

using Lemmas 1(iii) and 2(i), proving assertion (i). Further,

$$\lim_{m\to\infty} \mathbb{P}(\hat{Z}^{(m)} = k) = \lim_{m\to\infty} \mathbb{E}\left[\mathbb{P}_{\boldsymbol{D}}(\hat{Z}^{(m)} = k)\right] = \mathbb{P}(\hat{Y} = k),$$

using the dominated convergence theorem, proving assertion (ii). □

6.4.2 Early behaviour of major outbreaks.

Theorem 1 shows that the total number of households infected in $E^{(m)}$ converges in distribution as $m \to \infty$ to the total progeny of Y, so if $\tilde{\mu}_c \le 1$ only minor outbreaks can occur in the limit as $m \to \infty$ (recall that $R_* = \tilde{\mu}_c$). We now assume that $\tilde{\mu}_c > 1$ and study the early behaviour of $E^{(m)}$ when a major outbreak occurs. For $m = 1, 2, \ldots$, let

$$t_m = \lfloor 2\log\log m / \log\tilde{\mu}_c \rfloor\,.$$

We obtain a bound on the size of the 'bad set' of half-edges at time t_m and show that, with probability tending to 1 as $m \to \infty$, in a major outbreak there are at least $\log m$ infected households after t_m generations of the epidemic process.

Lemma 5. *There exists $\beta \in (1, \infty)$ such that, for all $\omega_1 \in A_1$,*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{T}^{(m)}_{t_m+1} \geq (\log m)^\beta) = 0.$$

Proof. Fix $\omega_1 \in A_1$ and note that, for all sufficiently large m,

$$\begin{aligned}\mathbb{E}_{\boldsymbol{D}(\omega_1)}[\hat{T}^{(m)}_{t_m+1}] &= \mu^{(m)}_r(\omega_1)(1 + \tilde{\mu}^{(m)}_r(\omega_1) + \cdots + (\tilde{\mu}^{(m)}_r(\omega_1))^{t_m+1}) \\ &\leq \mu^{(m)}_r(\omega_1)\frac{(\tilde{\mu}^{(m)}_r(\omega_1))^{t_m+2}}{\tilde{\mu}^{(m)}_r(\omega_1) - 1}.\end{aligned}$$

Thus, by Markov's inequality, for such m,

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{T}^{(m)}_{t_m+1} \geq (\log m)^\beta) \leq \frac{(\tilde{\mu}^{(m)}_r(\omega_1))^{t_m+2}}{(\log m)^\beta}\,\frac{\mu^{(m)}_r(\omega_1)}{\tilde{\mu}^{(m)}_r(\omega_1) - 1}. \tag{6.7}$$

It is readily shown, by considering its logarithm and using Lemma 1(iv), that, for all sufficiently large β, the right hand side of (6.7) tends to 0 as $m \to \infty$, and the lemma follows. □

Lemma 6. *For all $\omega_1 \in A_1$,*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Y^{(m)}_{t_m} > \log m) = \mathbb{P}(\hat{Y} = \infty).$$

Proof. Note that either (i) $\boldsymbol{c}$ and $\tilde{\boldsymbol{c}}$ both have infinite support, or (ii) $\boldsymbol{c}$ and $\tilde{\boldsymbol{c}}$ are supported on $\{0, 1, \ldots, nd_{\max}\}$ and $\{0, 1, \ldots, nd_{\max} - 1\}$ (or subsets thereof) respectively, where $d_{\max} = \max\{k : p_k > 0\}$.

Consider (i) first. For sufficiently small $\varepsilon > 0$, let $k_0 = \min\{k : \sum_{i=k+1}^\infty c_i < \varepsilon\}$, $\varepsilon' = \sum_{i=k_0+1}^\infty c_i$, $\tilde{k}_0 = \min\{k : \sum_{i=k+1}^\infty \tilde{c}_i < \varepsilon\}$ and $\tilde{\varepsilon}' = \sum_{i=\tilde{k}_0+1}^\infty \tilde{c}_i$, so $\varepsilon' < \varepsilon$ and $\tilde{\varepsilon}' < \varepsilon$. Note that k_0 and $\tilde{k}_0$ are well-defined whenever $\varepsilon < 1 - (c_0 \vee \tilde{c}_0)$ (where $a \vee b = \max(a, b)$), and also that both k_0 and $\tilde{k}_0$ tend to ∞ as $\varepsilon \downarrow 0$. Now let $Y^\varepsilon = (Y^\varepsilon_k,\, k = 0, 1, \ldots) \sim \mathrm{BP}(1, \boldsymbol{c}^\varepsilon, \tilde{\boldsymbol{c}}^\varepsilon)$, where $\boldsymbol{c}^\varepsilon$ has elements $c^\varepsilon_i = c_i + \frac{\varepsilon'}{k_0+1}$ for $i = 0, 1, \ldots, k_0$ and $c^\varepsilon_i = 0$ for $i > k_0$, and $\tilde{\boldsymbol{c}}^\varepsilon = (\tilde{c}^\varepsilon_i,\, i = 0, 1, \ldots)$ is defined similarly but with c_i, ε' and k_0 replaced by $\tilde{c}_i, \tilde{\varepsilon}'$ and $\tilde{k}_0$, respectively. Also let $\mu_\varepsilon = \sum_{k=1}^\infty k c^\varepsilon_k$ and $\tilde{\mu}_\varepsilon = \sum_{k=1}^\infty k \tilde{c}^\varepsilon_k$.

Now, note that $\sum_{i=0}^k c_i < \sum_{i=0}^k c^\varepsilon_i$ $(k = 0, 1, \ldots)$, so $\mu_\varepsilon < \mu_c$. We also have $\mu_\varepsilon = \sum_{i=1}^{k_0} i c^\varepsilon_i \geq \sum_{i=1}^{k_0} i c_i \to \mu_c$ as $\varepsilon \downarrow 0$, so $\mu_\varepsilon \to \mu_c$ as $\varepsilon \downarrow 0$. Similarly,

$\tilde{\mu}_\varepsilon \to \tilde{\mu}_c$ as $\varepsilon \downarrow 0$. Now fix $\omega_1 \in A_1$. Then, by Lemma 1(iii), $\boldsymbol{c}^{(m)}(\omega_1) \to \boldsymbol{c}$ and $\tilde{\boldsymbol{c}}^{(m)}(\omega_1) \to \tilde{\boldsymbol{c}}$ as $m \to \infty$, so there exists $M(\varepsilon, \omega_1)$ such that, for all $m \geq M(\varepsilon, \omega_1)$, $c_i^{(m)}(\omega_1) < c_i^\varepsilon$, for $i = 1, 2, \ldots, k_0$, and $\tilde{c}_i^{(m)}(\omega_1) < \tilde{c}_i^\varepsilon$, for $i = 1, 2, \ldots, \tilde{k}_0$. Thus, for $m \geq M(\varepsilon, \omega_1)$ and $k = 0, 1, \ldots$, $\sum_{i=0}^k c_i^{(m)}(\omega_1) < \sum_{i=0}^k c_i^\varepsilon$ and $\sum_{i=0}^k \tilde{c}_i^{(m)}(\omega_1) < \sum_{i=0}^k \tilde{c}_i^\varepsilon$ (note, for example, that $\sum_{i=0}^k c_i^\varepsilon = 1$ if $k \geq k_0$), whence $Y^{(m)}(\omega_1) \overset{\text{st}}{\geq} Y^\varepsilon$, where $\overset{\text{st}}{\geq}$ denotes stochastic ordering. Therefore, for $\omega_1 \in A_1$ and $m \geq M(\varepsilon, \omega_1)$,

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(Y_{t_m}^{(m)} > \log m) \geq \mathbb{P}(Y_{t_m}^\varepsilon > \log m) = \mathbb{P}\left(\frac{Y_{t_m}^\varepsilon}{\mu_\varepsilon \tilde{\mu}_\varepsilon^{t_m - 1}} > \frac{\log m}{\mu_\varepsilon \tilde{\mu}_\varepsilon^{t_m - 1}}\right). \quad (6.8)$$

Now, note that $\sum_{i=1}^\infty \tilde{c}_i^\varepsilon i \log i < \infty$ (the summand is 0 for $i > \tilde{k}_0$), so by the well known result concerning the exponential growth of branching processes (see, for example, Haccou *et al.* (2005, Theorem 6.1)), there exists a random variable $\mathcal{W}^\varepsilon$, which takes the value 0 if and only if Y^ε goes extinct (i.e. if and only if $\hat{Y}^\varepsilon = \sum_{i=0}^\infty Y_i^\varepsilon < \infty$), such that

$$\frac{Y_{t_m}^\varepsilon}{\mu_\varepsilon \tilde{\mu}_\varepsilon^{t_m - 1}} \overset{\text{a.s.}}{\to} \mathcal{W}^\varepsilon \qquad \text{as } m \to \infty.$$

Next, since $t_m = \lfloor 2 \log\log m / \log \tilde{\mu}_c \rfloor$, observe that, for suitable $\theta_m \in [0, 1)$,

$$\log \frac{\log m}{\mu_\varepsilon \tilde{\mu}_\varepsilon^{t_m - 1}} = \left(1 - 2\frac{\log \tilde{\mu}_\varepsilon}{\log \tilde{\mu}_c}\right) \log\log m + \log \frac{\tilde{\mu}_\varepsilon}{\mu_\varepsilon} + \theta_m \log \tilde{\mu}_\varepsilon.$$

Recalling that $\tilde{\mu}_\varepsilon \to \tilde{\mu}_c$ as $\varepsilon \to 0$ we see that, for sufficiently small ε, $\log \tilde{\mu}_\varepsilon / \log \tilde{\mu}_c > 1/2$ and thus $\log m / (\mu_\varepsilon \tilde{\mu}_\varepsilon^{t_m - 1}) \to 0$ as $m \to \infty$. It then follows from (6.8) that, for such ε,

$$\liminf_{m \to \infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Y_{t_m}^{(m)} > \log m) \geq \mathbb{P}(\mathcal{W}^\varepsilon > 0) = \mathbb{P}(\hat{Y}^\varepsilon = \infty). \quad (6.9)$$

Now, $\boldsymbol{c}^\varepsilon \to \boldsymbol{c}$ and $\tilde{\boldsymbol{c}}^\varepsilon \to \tilde{\boldsymbol{c}}$ as $\varepsilon \downarrow 0$, so letting $\varepsilon \downarrow 0$ in (6.9) and using Lemma 2(ii) yields

$$\liminf_{m \to \infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Y_{t_m}^{(m)} > \log m) \geq \mathbb{P}(\hat{Y} = \infty). \quad (6.10)$$

Now, for $k = 1, 2, \ldots$,

$$\begin{aligned} \limsup_{m \to \infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Y_{t_m}^{(m)} > \log m) &\leq \limsup_{m \to \infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Y}^{(m)} > \log m) \\ &\leq \limsup_{m \to \infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Y}^{(m)} > k) \\ &= \mathbb{P}(\hat{Y} > k), \end{aligned}$$

using Lemmas 1(iii) and 2(i). Letting $k \to \infty$ then yields

$$\limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Y_{t_m}^{(m)} > \log m) \le \mathbb{P}(\hat{Y} = \infty),$$

which, together with (6.10), establishes the lemma.

In case (ii), a suitable lower bounding branching process is obtained by setting, for $\varepsilon < c_{nd_{\max}} \wedge \tilde{c}_{nd_{\max}-1}$ (where $a \wedge b = \min(a, b)$), $c_i^\varepsilon = c_i + \varepsilon/(nd_{\max})$ $(i = 0, 1, \ldots, nd_{\max} - 1)$, $c_{nd_{\max}}^\varepsilon = c_{d_{\max}} - \varepsilon$, $\tilde{c}_i^\varepsilon = \tilde{c}_i + \varepsilon/(nd_{\max} - 1)$ $(i = 0, 1, \ldots, nd_{\max} - 2)$, $\tilde{c}_{nd_{\max}-1}^\varepsilon = \tilde{c}_{nd_{\max}-1} - \varepsilon$, and (6.10) follows as above. □

For $m = 1, 2, \ldots$ and $k = 0, 1, \ldots$, let $Z_k^{(m)}$ denote the number of infectious households in generation k of $E^{(m)}$.

Lemma 7. *Let β be as in Lemma 5. Then, for all $\omega_1 \in A_1$,*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)} \left(Z_{t_m}^{(m)} > \log m \,,\, \hat{T}_{t_m+1}^{(m)} < (\log m)^\beta \right) = \mathbb{P}(\hat{Y} = \infty). \qquad (6.11)$$

Proof. Lemmas 5 and 6 show that (6.11) holds with $Z_{t_m}^{(m)}$ replaced by $Y_{t_m}^{(m)}$. Application of Lemma 3, with $g(m) = (\log m)^\beta$ and $h(m) = 2(\log m)^\beta$ then shows that $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Z_{t_m}^{(m)} = Y_{t_m}^{(m)}) = 1$ and the assertion follows. □

Corollary 1. *(i) For all $\omega_1 \in A_1$, $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} > \log m) = \mathbb{P}(\hat{Y} = \infty)$;*

(ii) $\lim_{m\to\infty} \mathbb{P}(\hat{Z}^{(m)} > \log m) = \mathbb{P}(\hat{Y} = \infty)$.

Proof. Fix $\omega_1 \in A_1$. For $k = 1, 2, \ldots$,

$$\begin{aligned} \limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} > \log m) &\le \limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} > k) \\ &= \mathbb{P}(\hat{Y} > k) \qquad \text{(using Theorem 1(i))}, \end{aligned}$$

and letting $k \to \infty$ yields

$$\limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} > \log m) \le \mathbb{P}(\hat{Y} = \infty).$$

Also,

$$\begin{aligned} \liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} > \log m) &\ge \liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(Z_{t_m}^{(m)} > \log m) \\ &= \mathbb{P}(\hat{Y} = \infty) \qquad \text{(using Lemmas 5 and 7)}, \end{aligned}$$

and assertion (i) follows. Assertion (ii) then follows using the dominated convergence theorem, as in the proof of Theorem 1(ii). □

Note that Theorem 1 and Corollary 1 imply that if (h_m) is any sequence of real numbers satisfying $h_m \to \infty$ as $m \to \infty$ and $h_m < \log m$ for all m, then $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{Z}^{(m)} \in [h_m, \log m)) = 0$ for all $\omega_1 \in A_1$ and $\lim_{m\to\infty} \mathbb{P}(\hat{Z}^{(m)} \in [h_m, \log m)) = 0$. Thus, for $m = 1, 2, \ldots$, it is natural to define a major outbreak as one which infects at least $\log m$ households, i.e. as one in which the event $\bar{G}^{(m)} = \{\omega \in \Omega : \hat{Z}^{(m)}(\omega) > \log m\}$ occurs. Let $G^{(m)} = \{\omega \in \Omega : Z^{(m)}_{t_m} > \log m\,,\, \hat{T}^{(m)}_{t_m+1} < (\log m)^\beta\}$, where β is as in Lemma 5. Clearly $G^{(m)} \subseteq \bar{G}^{(m)}$, and Lemma 5 and Corollary 1 imply $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\bar{G}^{(m)} \setminus G^{(m)}) = 0$ for all $\omega_1 \in A_1$ and $\lim_{m\to\infty} \mathbb{P}(\bar{G}^{(m)} \setminus G^{(m)}) = 0$, so we can take $G^{(m)}$ as our working definition of a major outbreak.

6.5 Analysis of backward process

6.5.1 Lower bounding branching processes.

We now analyse the 'backward' process, which describes the generation-wise growth of the susceptibility set (and its neighbours) of a typical individual that is susceptible at time t_m in the forward process, in order to find the asymptotic probability that such an individual is ultimately infected, given that a major outbreak occurs (i.e. $Z^{(m)}_{t_m} > \log m$ and $\hat{T}^{(m)}_{t_m+1} < (\log m)^\beta$, where β is as in Lemma 5). To this end, it is fruitful to have, for all sufficiently small $\varepsilon > 0$, a branching process ${}_\varepsilon X^{(m)}$ which asymptotically bounds $S^{(m)}$ from below until the susceptibility set covers a proportion ε of the households in the population (cf. Whittle (1955)). In order to do this, we need an almost sure bound, $\bar{\eta}(\varepsilon)$, for the proportion of households that are neighbours of the susceptibility set when the size (in terms of households) of the susceptibility set is at most εm, which we now obtain.

Suppose that D has infinite support. Recall the definitions of $p_H(\cdot)$ and $\tilde{p}_H(\cdot)$ from Section 6.2.3. Let $k_1 = \min\{k : p_H(k) > 0\}$ and $\varepsilon_0 = 1 - p_H(k_1) - p_H(k_1+1)$. Then, for $\varepsilon \in (0, \varepsilon_0)$, let $\kappa(\varepsilon) = \max\{k : \sum_{i=k_1}^{k} p_H(i) \in (0, 1-\varepsilon)\}$, $\kappa^*(\varepsilon) = \max\{k < \kappa(\varepsilon) : p_H(k) > 0\}$ and $\eta(\varepsilon) = \sum_{i=\kappa^*(\varepsilon)}^{\infty} \tilde{p}_H(i)$. (The definition of $\kappa^*(\varepsilon)$ requires $\kappa(\varepsilon) > k_1$, which in turn requires $\varepsilon < \varepsilon_0$.) Note that $\eta(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$. Let $\bar{\eta}(\varepsilon) = 2n\mu_D\eta(\varepsilon)$ and, for $m = 1, 2, \ldots$, let $H^{(m)}_{(1)}, H^{(m)}_{(2)}, \ldots, H^{(m)}_{(m)}$ be the order statistics of the household degrees $H_1, H_2, \ldots, H_m$.

Lemma 8. *For any $\omega_1 \in A_1$ and $\varepsilon \in (0, \varepsilon_0)$,*

$$\frac{1}{m\mu_H^{(m)}(\omega_1)} \sum_{k=m-[\varepsilon m]+1}^{m} H^{(m)}_{(k)}(\omega_1) \leq \eta(\varepsilon) \tag{6.12}$$

and

$$\frac{1}{m}\sum_{k=m-[\varepsilon m]+1}^{m} H_{(k)}^{(m)}(\omega_1) \le \bar{\eta}(\varepsilon) \tag{6.13}$$

for all sufficiently large m.

Proof. Fix $\omega_1 \in A_1$ and note that, for $k = 0, 1, \ldots$,

$$\begin{aligned}
\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{H_i(\omega_1)=k\}} &= \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \sum_{\{\boldsymbol{d}\colon |\boldsymbol{d}|=k\}} \mathbb{1}_{\{\boldsymbol{D}_i(\omega_1)=\boldsymbol{d}\}} \\
&= \lim_{m\to\infty} \sum_{\{\boldsymbol{d}\colon |\boldsymbol{d}|=k\}} p_{\boldsymbol{d}}^{(m)}(\omega_1) \\
&= \sum_{\{\boldsymbol{d}\colon |\boldsymbol{d}|=k\}} p_{\boldsymbol{d}} \qquad \text{(using Lemma 1(ii))} \\
&= p_H(k), \qquad (6.14)
\end{aligned}$$

whence

$$\begin{aligned}
\lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m} \mathbb{1}_{\{H_i(\omega_1)\ge\kappa(\varepsilon)+1\}} &= 1 - \lim_{m\to\infty} \frac{1}{m}\sum_{i=1}^{m}\sum_{j=0}^{\kappa(\varepsilon)} \mathbb{1}_{\{H_i(\omega_1)=j\}} \\
&= \sum_{j=\kappa(\varepsilon)+1}^{\infty} p_H(j).
\end{aligned}$$

Thus, since $\sum_{j=\kappa(\varepsilon)+1}^{\infty} p_H(j) > \varepsilon$ (by the definition of $\kappa(\varepsilon)$), we have, for all sufficiently large m, say $m \ge N_0(\varepsilon, \omega_1)$, that $m^{-1}\sum_{i=1}^{m} \mathbb{1}_{\{H_i(\omega_1)\ge\kappa(\varepsilon)+1\}} > \varepsilon$, whence $H_{(m-[\varepsilon m]+1)}^{(m)}(\omega_1) > \kappa(\varepsilon)$. Hence, for $m \ge N_0(\varepsilon, \omega_1)$,

$$\begin{aligned}
\frac{1}{m\mu_H^{(m)}(\omega_1)} \sum_{k=m-[\varepsilon m]+1}^{m} H_{(k)}^{(m)}(\omega_1) &\le \frac{1}{m\mu_H^{(m)}(\omega_1)} \sum_{i=1}^{m}\sum_{k=\kappa(\varepsilon)+1}^{\infty} k\mathbb{1}_{\{H_i(\omega_1)=k\}} \\
&= \frac{m\mu_H^{(m)}(\omega_1) - \sum_{i=1}^{m}\sum_{k=0}^{\kappa(\varepsilon)} k\mathbb{1}_{\{H_i(\omega_1)=k\}}}{m\mu_H^{(m)}(\omega_1)} \\
&= 1 - \frac{1}{\mu_H^{(m)}(\omega_1)} \sum_{k=0}^{\kappa(\varepsilon)} \frac{k}{m} \sum_{i=1}^{m} \mathbb{1}_{\{H_i(\omega_1)=k\}} \\
&\to 1 - \sum_{k=1}^{\kappa(\varepsilon)} \tilde{p}_H(k) = \sum_{k=\kappa(\varepsilon)+1}^{\infty} \tilde{p}_H(k)
\end{aligned}$$

as $m \to \infty$, using (6.14) and Lemma 1(i). Assertion (6.12) follows upon recalling the definition of $\kappa^*(\varepsilon)$. The second assertion (6.13) follows from the first assertion after applying Lemma 1(i) and recalling the definition of $\bar{\eta}(\varepsilon)$. □

Remarks.

1. If $d_{\max} < \infty$ (i.e. D has finite support) then it is readily seen that Lemma 8 holds with $\eta(\varepsilon) = 2d_{\max}\varepsilon/\mu_D$ and $\bar{\eta}(\varepsilon) = nd_{\max}\varepsilon$.

2. For $\omega_1 \in A_1$, Lemma 8 provides, for all sufficiently large m, a bound for the number of half-edges that emanate from households in a susceptibility set (and hence also for the number of households neighbouring a susceptibility set), if the susceptibility set contains no more than εm households. The number of such half-edges, $H^{(m)}(\varepsilon)$ say, is given by the sum of the degrees of the households in the susceptibility set, which is bounded by the sum of the degrees of the $[\varepsilon m]$ households of highest degree. Thus, by (6.13), $H^{(m)}(\varepsilon) \leq m\bar{\eta}(\varepsilon)$ for all sufficiently large m.

Recall from Section 6.2.2 that the coupling of the susceptibility set process $S^{(m)}$ and its approximating branching process $X^{(m)}$ breaks down when a half-edge is sampled that emanates from an appropriate 'bad' set of households. This can happen in two fundamentally different ways. First, a half-edge through which we try to extend the susceptibility set may be paired up with another half-edge through which we want to extend the susceptibility set in the same generation. Note that in this case, neither of the two half-edges concerned actually extends the susceptibility set. Second, the half-edge may be paired with a bad half-edge which is not one through which we wish to extend the susceptibility set in the current generation, in which case the susceptibility set may still be extended, though the offspring distribution is different to that in the branching process. We treat these two cases sequentially.

For $m = 1, 2, \ldots$ and $k = 0, 1, \ldots$, let $\hat{X}_k^{(m)} = \sum_{i=0}^{k} X_i^{(m)}$ be the total number of individuals that have lived in the approximating branching process $X^{(m)}$ by time k and let $\hat{W}_k^{(m)} = \sum_{i=0}^{k} S_i^{(m)}$ be the total number of households in the susceptibility set process $S^{(m)}$ up to and including generation k. Further, let $\hat{X}^{(m)} = \sum_{i=0}^{\infty} X_i^{(m)}$ and $\hat{W}^{(m)} = \sum_{i=0}^{\infty} S_i^{(m)}$. Suppose that $\omega_1 \in A_1$. Then, for all sufficiently large m, while $\hat{W}_k^{(m)} \leq \varepsilon m$, the probability that a half-edge is paired with another half-edge through which we want to extend the susceptibility set in the same generation is no more than $\eta(\varepsilon)$. For such m, suppose that at some generation k there are $X_{k-1}^{(m)} = i$ 'live' half-edges through which we attempt to extend the susceptibility set. Denote by

Y_L the number of these half-edges that do not pair up with another of these i live half-edges and let $\check{Y}_L \sim \mathrm{Bin}(i, 1-\sqrt{\eta(\varepsilon)})$. We now show that $Y_L \overset{\mathrm{st}}{\geq} \check{Y}_L$.

First, define another random variable $\hat{Y}_L$ as follows. Take a live half-edge, then with probability $\eta(\varepsilon)$ pair it up with another live half-edge, otherwise it 'survives' to be connected with a non-live half-edge. Repeat this process until all live half-edges have been either paired up or designated to survive. Note that if there is a single live half-edge left at the end of this procedure, it must survive. Let $\hat{Y}_L$ be the number of surviving half-edges under this regime. Since the proportion of half-edges that are actually live is less than $\eta(\varepsilon)$, $Y_L \overset{\mathrm{st}}{\geq} \hat{Y}_L$. We now show that $\hat{Y}_L \overset{\mathrm{st}}{\geq} \check{Y}_L$ by describing these two random variables as the number of renewals of a discrete-time renewal process by time i and showing that the corresponding lifetime distributions, $\hat{T}$ and $\check{T}$ say, satisfy $\hat{T} \overset{\mathrm{st}}{\leq} \check{T}$. This we achieve by taking a lifetime in the renewal process as being the number of half-edges examined to find a surviving half-edge. It is immediate that $\mathbb{P}(\check{T}=k) = (1-\eta(\varepsilon)^{\frac{1}{2}})\eta(\varepsilon)^{\frac{k-1}{2}}$, $k=1,2,\ldots$. Now, since pairing one live half-edge with another obviously uses up two half-edges, $\hat{T}$ cannot take even values and $\mathbb{P}(\hat{T}=2k+1) = (1-\eta(\varepsilon))\eta(\varepsilon)^k$, $k=0,1,\ldots$. Elementary calculation shows that $\mathbb{P}(\hat{T}\geq k) \leq \mathbb{P}(\check{T}\geq k)$, $k=1,2,\ldots$, so $\check{T} \overset{\mathrm{st}}{\geq} \hat{T}$, whence $\hat{Y}_L \overset{\mathrm{st}}{\geq} \check{Y}_L$.

The above argument shows that, in a given generation, the number of half-edges that survive to be paired with non-live half-edges is stochastically larger than if they survive independently with probability $1-\sqrt{\eta(\varepsilon)}$. Now consider a live half-edge that survives this first stage and thus is paired with a half-edge chosen uniformly at random from all the non-live half-edges. The probability that it avoids being paired with a half-edge from the bad set is therefore larger than if it were paired with a half-edge chosen uniformly at random from all of the half-edges. Recall that $\tilde{p}^{(m)}_{\boldsymbol{d}}$ is the probability that a half-edge chosen at random from all $m\mu^{(m)}_H$ half-edges in the population emanates from a household of type $\boldsymbol{d}$. Further, for $\omega_1 \in A_1$ and m sufficiently large, conditional on choosing a household of type $\boldsymbol{d}$, if $\hat{W}^{(m)}_k \leq \varepsilon m$ and $Y^{(m)}_{t_m} < (\log m)^\beta$ then the probability of choosing a bad household is bounded above by

$$\tilde{\gamma}^{(m)}_{\boldsymbol{d}}(\varepsilon) = \frac{(\log m)^\beta + \varepsilon m + \bar{\eta}(\varepsilon)m}{m\tilde{p}^{(m)}_{\boldsymbol{d}}} \wedge 1.$$

This bound is obtained by noting that, under the stated conditions, there are fewer than $(\log m)^\beta$ bad households from the forward process, fewer than εm households in the susceptibility set and fewer than $\bar{\eta}(\varepsilon)m$ households that are neighbours of the susceptibility set; and then assuming that all of these

bad households are of type $\boldsymbol{d}$.

It follows from this discussion that, for m sufficiently large and if $Y_{t_m}^{(m)} < (\log m)^\beta$, then while $\hat{W}_k^{(m)} \leq \varepsilon m$ the susceptibility set process $S^{(m)}$ is stochastically larger than a branching process, ${}_\varepsilon X^{(m)}$ say, in which each potential birth (live half-edge) is aborted independently with probability $\sqrt{\eta(\varepsilon)}$ and the potential offspring (live half-edges) of an unaborted birth are obtained by first sampling $\boldsymbol{d}$ according to $\tilde{p}_{\boldsymbol{d}}^{(m)}$, then with probability $\tilde{\gamma}_{\boldsymbol{d}}^{(m)}(\varepsilon)$ this unaborted birth is aborted at this stage and otherwise its potential offspring is distributed according to the random variable $\tilde{\Psi}_{\boldsymbol{d}}$ defined at the end of the paragraph following (6.3).

The number, $\bar{X}_1^{(m)}$ say, of potential births that emanate from the initial individual in the susceptibility set may be found as follows. First a household is chosen uniformly at random from the households not infected by time t_m in the forward process. Suppose that this household is of type $\boldsymbol{d}$. Then, if this household is not a neighbour of a household in the forward process, $\bar{X}_1^{(m)}$ is distributed according to the random variable $\Psi_{\boldsymbol{d}}$, also defined in the paragraph immediately following (6.3). If the sampled household is a neighbour of a household in the forward process then $\bar{X}_1^{(m)}$ has a different distribution. Suppose that $\hat{T}_{t_m+1}^{(m)} < (\log m)^\beta$. Then the number of households that are neighbours of the forward process is less than $2(\log m)^\beta$ and it follows that $\bar{X}_1^{(m)}$ is stochastically larger than a random variable, $\bar{\bar{X}}_1^{(m)}$ say, obtained by first sampling $\boldsymbol{d}$ according to $p_{\boldsymbol{d}}^{(m)}$ and then setting $\bar{\bar{X}}_1^{(m)} = 0$ with probability $\gamma_{\boldsymbol{d}}^{(m)} = \frac{2(\log m)^\beta}{m p_{\boldsymbol{d}}^{(m)}} \wedge 1$, otherwise $\bar{\bar{X}}_1^{(m)}$ is distributed according to $\Psi_{\boldsymbol{d}}$.

Assume that there is a single ancestor in the branching process ${}_\varepsilon X^{(m)}$, which has a number of potential offspring distributed as $\bar{\bar{X}}_1^{(m)}$. We now have a complete description of how ${}_\varepsilon X^{(m)}$ evolves. Let ${}_\varepsilon \hat{X}^{(m)}$ and ${}_\varepsilon \hat{W}^{(m)}$ be, respectively, the total number of potential and unaborted births in ${}_\varepsilon X^{(m)}$. Recall the event $G^{(m)}$ defined at the end of Section 6.4.2, giving our working definition of a major outbreak. The above arguments show that

$$\begin{aligned} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{W}^{(m)} > [\varepsilon m] \,|\, G^{(m)}) &\geq \mathbb{P}_{\boldsymbol{D}(\omega_1)}({}_\varepsilon \hat{W}^{(m)} \geq [\varepsilon m]) \\ &\geq \mathbb{P}_{\boldsymbol{D}(\omega_1)}({}_\varepsilon \hat{W}^{(m)} = \infty) \\ &= \mathbb{P}_{\boldsymbol{D}(\omega_1)}({}_\varepsilon \hat{X}^{(m)} = \infty). \end{aligned} \tag{6.15}$$

For the branching process ${}_\varepsilon X^{(m)}$, let ${}_\varepsilon \boldsymbol{b}^{(m)} = ({}_\varepsilon b_0^{(m)}, {}_\varepsilon b_1^{(m)}, \ldots)$ denote the distribution of the number of potential offspring of the initial individual and let ${}_\varepsilon \tilde{\boldsymbol{b}}^{(m)} = ({}_\varepsilon \tilde{b}_0^{(m)}, {}_\varepsilon \tilde{b}_1^{(m)}, \ldots)$ denote the distribution of the number of potential

offspring of a typical potential birth. Then

$$
\begin{aligned}
{}_\varepsilon b_0^{(m)} &= \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} p_{\boldsymbol{d}}^{(m)} \left(\gamma_{\boldsymbol{d}}^{(m)} + (1-\gamma_{\boldsymbol{d}}^{(m)})\,\mathbb{P}(\Psi_{\boldsymbol{d}}=0)\right),\\
{}_\varepsilon b_k^{(m)} &= \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} p_{\boldsymbol{d}}^{(m)} (1-\gamma_{\boldsymbol{d}}^{(m)})\,\mathbb{P}(\Psi_{\boldsymbol{d}}=k) \qquad (k=1,2,\ldots),\\
{}_\varepsilon \tilde{b}_0^{(m)} &= \sqrt{\eta(\varepsilon)} + (1-\sqrt{\eta(\varepsilon)}) \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} \tilde{p}_{\boldsymbol{d}}^{(m)} \left(\tilde{\gamma}_{\boldsymbol{d}}^{(m)}(\varepsilon) + (1-\tilde{\gamma}_{\boldsymbol{d}}^{(m)}(\varepsilon))\,\mathbb{P}(\tilde{\Psi}_{\boldsymbol{d}}=0)\right)
\end{aligned}
$$

and

$$
{}_\varepsilon \tilde{b}_k^{(m)} = (1-\sqrt{\eta(\varepsilon)}) \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} \tilde{p}_{\boldsymbol{d}}^{(m)} (1-\tilde{\gamma}_{\boldsymbol{d}}^{(m)}(\varepsilon))\,\mathbb{P}(\tilde{\Psi}_{\boldsymbol{d}}=k) \qquad (k=1,2,\ldots).
$$

Note that ${}_\varepsilon\boldsymbol{b}^{(m)}$ does not depend on ε, however it is distinct from $\boldsymbol{b}^{(m)}$ and we retain the notation ${}_\varepsilon\boldsymbol{b}^{(m)}$ to indicate that it is associated with the branching process ${}_\varepsilon X^{(m)}$.

The following lemma is useful for determining the limits of the distributions ${}_\varepsilon\boldsymbol{b}^{(m)}$ and ${}_\varepsilon\tilde{\boldsymbol{b}}^{(m)}$ as $m\to\infty$. Its proof is standard and is hence omitted.

Lemma 9. *Suppose that, for all $\boldsymbol{d}\in\mathbb{Z}_+^n$ and $m=1,2,\ldots$, the real numbers*

(i) $p_{\boldsymbol{d}}^{(m)}$ and $p_{\boldsymbol{d}}$ are non-negative and satisfy $p_{\boldsymbol{d}}^{(m)}\to p_{\boldsymbol{d}}$ as $m\to\infty$ and $\sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} p_{\boldsymbol{d}}^{(m)} = \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} p_{\boldsymbol{d}} = 1$;

(ii) $\alpha_{\boldsymbol{d}}^{(m)}$ and $\alpha_{\boldsymbol{d}}$ belong to $[0,1]$ and satisfy $\alpha_{\boldsymbol{d}}^{(m)}\to\alpha_{\boldsymbol{d}}$ as $m\to\infty$;

(iii) $c_{\boldsymbol{d}}$ belong to $[0,1]$.

Then, as $m\to\infty$,

$$
\sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} p_{\boldsymbol{d}}^{(m)}\alpha_{\boldsymbol{d}}^{(m)}c_{\boldsymbol{d}} \to \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} p_{\boldsymbol{d}}\alpha_{\boldsymbol{d}}c_{\boldsymbol{d}}.
$$

For $\boldsymbol{d}\in\mathbb{Z}_+^n$ and $\varepsilon\in(0,\varepsilon_0)$, let

$$
\tilde{\gamma}_{\boldsymbol{d}}(\varepsilon) = \begin{cases} \left(\frac{\varepsilon+\bar{\eta}(\varepsilon)}{\tilde{p}_{\boldsymbol{d}}}\right)\wedge 1 & \text{if } \tilde{p}_{\boldsymbol{d}}>0,\\ 0 & \text{if } \tilde{p}_{\boldsymbol{d}}=0. \end{cases}
$$

Lemma 10. *For all* $\omega_1 \in A_1$, $\lim_{m\to\infty} {}_\varepsilon\boldsymbol{b}^{(m)} = \boldsymbol{b}$ *and* $\lim_{m\to\infty} {}_\varepsilon\tilde{\boldsymbol{b}}^{(m)} = {}_\varepsilon\tilde{\boldsymbol{b}}$, *where* $\boldsymbol{b} = (b_0, b_1, \ldots)$ *is as in Section 6.2.3, and* ${}_\varepsilon\tilde{\boldsymbol{b}} = ({}_\varepsilon\tilde{b}_0, {}_\varepsilon\tilde{b}_1, \ldots)$ *is given by*

$$ {}_\varepsilon\tilde{b}_0 = \sqrt{\eta(\varepsilon)} + (1 - \sqrt{\eta(\varepsilon)}) \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} \tilde{p}_{\boldsymbol{d}} \left(\tilde{\gamma}_{\boldsymbol{d}}(\varepsilon) + (1 - \tilde{\gamma}_{\boldsymbol{d}}(\varepsilon))\, \mathbb{P}(\tilde{\Psi}_{\boldsymbol{d}} = 0) \right) $$

and

$$ {}_\varepsilon\tilde{b}_k = (1 - \sqrt{\eta(\varepsilon)}) \sum_{\boldsymbol{d}\in\mathbb{Z}_+^n} \tilde{p}_{\boldsymbol{d}} (1 - \tilde{\gamma}_{\boldsymbol{d}}(\varepsilon))\, \mathbb{P}(\tilde{\Psi}_{\boldsymbol{d}} = k) \qquad (k = 1, 2, \ldots). $$

Proof. Note that, for $\omega_1 \in A_1$, $\gamma_{\boldsymbol{d}}^{(m)}(\omega_1) \to 0$ and $\tilde{\gamma}_{\boldsymbol{d}}^{(m)}(\varepsilon, \omega_1) \to \gamma_{\boldsymbol{d}}(\varepsilon)$ as $m \to \infty$ (for all $\boldsymbol{d}$ with $p_{\boldsymbol{d}} > 0$). The required assertions then follow using Lemma 9. □

Remark. It is easily verified that $\sum_{k=0}^{\infty} {}_\varepsilon\tilde{b}_k = 1$, i.e. that ${}_\varepsilon\tilde{\boldsymbol{b}}$ is a proper probability distribution.

Recall the definition of ε_0 in the paragraph preceding Lemma 8 and, for $\varepsilon \in (0, \varepsilon_0)$, let ${}_\varepsilon X = ({}_\varepsilon X_k,\, k = 0, 1, \ldots) \sim \mathrm{BP}(1, \boldsymbol{b}, {}_\varepsilon\tilde{\boldsymbol{b}})$. Let ${}_\varepsilon\hat{X}$ denote the total progeny of ${}_\varepsilon X$, excluding the ancestor. Let $(\hat{X}, \hat{X}_A)$ denote the total progeny of the branching process (X, X_A) (defined at the end of Section 6.2.3), including the ancestor. Also let $\hat{X}_A^{(m)} = \sum_{i=0}^{\infty} X_{Ai}^{(m)}$, so $(\hat{X}^{(m)}, \hat{X}_A^{(m)})$ is the total progeny of $(X^{(m)}, X_A^{(m)})$.

Lemma 11. *(i) For all* $\omega_1 \in A_1$,

(a) $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{X}^{(m)} + \hat{X}_A^{(m)} = k) = \mathbb{P}(\hat{X} + \hat{X}_A = k) \qquad (k = 1, 2, \ldots)$;

(b) $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{X}^{(m)} + \hat{X}_A^{(m)} = \infty) = \mathbb{P}(\hat{X} = \infty)$.

(ii) For all $\omega_1 \in A_1$ *and* $\varepsilon \in (0, \varepsilon_0)$,

(a) $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}({}_\varepsilon\hat{X}^{(m)} = k) = \mathbb{P}({}_\varepsilon\hat{X} = k) \qquad (k = 1, 2, \ldots)$;

(b) $\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}({}_\varepsilon\hat{X}^{(m)} = \infty) = \mathbb{P}({}_\varepsilon\hat{X} = \infty)$.

Proof. For all $\omega_1 \in A_1$ and $\boldsymbol{d} \in \mathbb{Z}_+^n$, $p_{\boldsymbol{d}}^{(m)}(\omega_1) \to p_{\boldsymbol{d}}$ and $\tilde{p}_{\boldsymbol{d}}^{(m)}(\omega_1) \to \tilde{p}_{\boldsymbol{d}}$ as $m \to \infty$, so, using Scheffé's theorem, $\boldsymbol{b}^{(m)}(\omega_1) \to \boldsymbol{b}$ and $\tilde{\boldsymbol{b}}^{(m)}(\omega_1) \to \tilde{\boldsymbol{b}}$ as $m \to \infty$. Part (ii)(b) then follows using Lemma 2(ii) and noting that, almost surely, $\hat{X}^{(m)} + \hat{X}_A^{(m)} = \infty$ if and only if $\hat{X}^{(m)} = \infty$. A similar argument

shows that, for all $\omega_1 \in A_1$, the offspring laws of $(X^{(m)}, X_A^{(m)})$ converge to those of (X, X_A) as $m \to \infty$. Part (i)(a) then follows from the extension of Lemma 2(i) to two-type branching processes. Part (ii) of the lemma follows immediately from Lemmas 10 and 2. □

6.5.2 Relative final size of a major outbreak.

For $m = 1, 2, \ldots$, let $B^{(m)}$ be the event that an individual chosen uniformly at random from all individuals that are susceptible at time t_m in the forward process is ultimately infected by the epidemic $E^{(m)}$. Thus, if $\mathcal{A}^{(m)}$ denotes the set of global neighbours of $\mathcal{S}^{(m)}$ then $B^{(m)}$ occurs if and only if one of the $Z_{t_m}^{(m)}$ 'live' half-edges from the forward process is paired in the construction of $\mathcal{S}^{(m)} \cup \mathcal{A}^{(m)}$. Recall the working definition of a major outbreak, viz. $G^{(m)} = \{Z_{t_m}^{(m)} > \log m\,,\, \hat{T}_{t_m+1}^{(m)} < (\log m)^\beta\}$, where β is as in Lemma 5.

Theorem 2. *For all $\omega_1 \in A_1$,*

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) = \mathbb{P}(\hat{X} = \infty).$$

Proof. For $m = 1, 2, \ldots$, let $T_P^{(m)}$ be the number of half-edge pairings made in the construction of $\mathcal{S}^{(m)} \cup \mathcal{A}^{(m)}$ until one of the $Z_{t_m}^{(m)}$ live half-edges from the forward process is chosen. In determining $T_P^{(m)}$ it is assumed that, if necessary, the pairings continue after $\mathcal{S}^{(m)} \cup \mathcal{A}^{(m)}$ goes extinct and that $T_P^{(m)}$ includes the pairing when the first live half-edge is chosen.

Fix $\omega_1 \in A_1$. First we obtain an upper bound for $\mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)})$. For all fixed $k \in \mathbb{N}$,

$$1 - \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) \geq \mathbb{P}_{\boldsymbol{D}(\omega_1)}(T_P^{(m)} > k\,,\, \hat{X}^{(m)} + \hat{X}_A^{(m)} \leq k\,,\, \bar{\tau}^{(m)} > k \,|\, G^{(m)}), \tag{6.16}$$

where $\bar{\tau}^{(m)}$ is the number of households in the construction of $\mathcal{S}^{(m)} \cup \mathcal{A}^{(m)}$ when the first bad half-edge is chosen. Note that $\bar{\tau}^{(m)} = 1$ if the initial individual in $(X^{(m)}, X_A^{(m)})$ belongs to the set of bad households at time t_m in the forward process. Given $G^{(m)}$, the number of such bad households is less than $(\log m)^\beta$, so $\mathbb{P}_{\boldsymbol{D}(\omega_1)}(\bar{\tau}^{(m)} = 1 \,|\, G^{(m)}) \to 0$ as $m \to \infty$. Arguing as in the proof of Theorem 1 then shows that, for all $k \in \mathbb{N}$,

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\bar{\tau}^{(m)} > k \,|\, G^{(m)}) = 1. \tag{6.17}$$

Let $Q^{(m)}$ denote the number of half-edges used up to time t_m in the forward process. Now, for all $k \in \mathbb{N}$,

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(T_P^{(m)} > k \,|\, G^{(m)}\,,\, Q^{(m)}\,,\, Z_{t_m}^{(m)}) = \prod_{i=1}^{k} \left(\frac{m\mu_h^{(m)}(\omega_1) - Q^{(m)} - 2(i-1) - Z_{t_m}^{(m)}}{m\mu_h^{(m)}(\omega_1) - Q^{(m)} - 2(i-1)} \right) \tag{6.18}$$

and, since we have conditioned on $G^{(m)}$, $Q^{(m)} < 2(\log m)^\beta$ and $Z_{t_m}^{(m)} < 2(\log m)^\beta$. It then follows from (6.18) that

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(T_P^{(m)} > k \,|\, G^{(m)}) = 1 \tag{6.19}$$

for all $k \in \mathbb{N}$. Letting $m \to \infty$ in (6.16), using (6.17) and (6.19), and noting that $\hat{X}^{(m)} + \hat{X}_A^{(m)}$ and $G^{(m)}$ are conditionally independent given $\boldsymbol{D}(\omega_1)$, yields, for all $k \in \mathbb{N}$,

$$\begin{aligned} \limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) &\leq \limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{X}^{(m)} + \hat{X}_A^{(m)} > k) \\ &= \mathbb{P}(\hat{X} + \hat{X}_A > k), \end{aligned}$$

using Lemma 11(i)(a). Letting $k \to \infty$ then yields

$$\limsup_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) \leq \mathbb{P}(\hat{X} = \infty). \tag{6.20}$$

Now we obtain a lower bound for $\mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)})$. First note, using (6.18), that for any $\varepsilon \in (0,1)$, we have

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(T_P^{(m)} > [\varepsilon m] \,|\, G^{(m)}) \leq \left(1 - \frac{\log m}{m\mu_H^{(m)}(\omega_1)}\right)^{[\varepsilon m]} \leq \exp\left(\frac{-[\varepsilon m]\log m}{m\mu_H^{(m)}(\omega_1)}\right).$$

Now, $\mu_H^{(m)}(\omega_1) \to n\mu_D$ as $m \to \infty$ (since $\omega_1 \in A_1$), so $[\varepsilon m]\log m / m\mu_H^{(m)}(\omega_1) \to \infty$ as $m \to \infty$, whence

$$\lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(T_P^{(m)} \leq [\varepsilon m] \,|\, G^{(m)}) = 1. \tag{6.21}$$

Also note that, since $\mathcal{S}^{(m)}$ is obviously contained in $\mathcal{S}^{(m)} \cup \mathcal{A}^{(m)}$,

$$\mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) \geq \mathbb{P}_{\boldsymbol{D}(\omega_1)}(T_P^{(m)} \leq [\varepsilon m]\,,\, \hat{W}^{(m)} > [\varepsilon m] \,|\, G^{(m)}), \tag{6.22}$$

for any $\varepsilon \in (0,1)$. Thus, using (6.22) and (6.21), then (6.15) and Lemma 11(ii)(b), for any $\varepsilon \in (0, \varepsilon_0)$,

$$\begin{aligned} \liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) &\geq \liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(\hat{W}^{(m)} > [\varepsilon m] \,|\, G^{(m)}) \\ &\geq \liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}({}_\varepsilon\hat{X}^{(m)} = \infty) \\ &= \mathbb{P}({}_\varepsilon\hat{X} = \infty). \end{aligned} \tag{6.23}$$

It is easily verified, using the dominated convergence theorem, that $({}_\varepsilon\boldsymbol{b}, {}_\varepsilon\tilde{\boldsymbol{b}}) \to (\boldsymbol{b}, \tilde{\boldsymbol{b}})$ as $\varepsilon \downarrow 0$, so letting $\varepsilon \downarrow 0$ in (6.23) and using Lemma 2(ii) yields

$$\liminf_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)}) \geq \mathbb{P}(\hat{X} = \infty),$$

which together with (6.20) establishes the assertion of the theorem. □

For $m = 1, 2, \ldots$, let $\bar{Z}_k^{(m)}$ be the total number of individuals infected by time k in the forward epidemic process $E^{(m)}$ $(k = 0, 1, \ldots)$ and let $\bar{Z}^{(m)}$ denote the total number of individuals who are ultimately infected in $E^{(m)}$.

Corollary 2. *(i) For all* $\omega_1 \in A_1$, $\lim_{m\to\infty} \frac{1}{mn} \mathbb{E}_{\boldsymbol{D}(\omega_1)}[\bar{Z}^{(m)} \,|\, G^{(m)}] = \mathbb{P}(\hat{X} = \infty)$*;*

(ii) $\lim_{m\to\infty} \frac{1}{mn} \mathbb{E}[\bar{Z}^{(m)} \,|\, G^{(m)}] = \mathbb{P}(\hat{X} = \infty)$.

Proof. Fix $\omega_1 \in A_1$. For $m = 1, 2, \ldots$, let $\bar{X}_{t_m}$ denote the number of susceptible individuals at time t_m in the forward process, and label these individuals $1, 2, \ldots, \bar{X}_{t_m}$. Then

$$\bar{Z}^{(m)} = \bar{Z}_{t_m}^{(m)} + \sum_{i=1}^{\bar{X}_{t_m}} \mathbb{1}_{\{i \text{ ultimately infected}\}}.$$

Given the occurrence of $G^{(m)}$, $\bar{Z}_{t_m}^{(m)} < 2n(\log m)^\beta$ and $\bar{X}_{t_m} > nm - 2n(\log m)^\beta$. Thus

$$\lim_{m\to\infty} \frac{1}{mn} \mathbb{E}_{\boldsymbol{D}(\omega_1)}[\bar{Z}^{(m)} \,|\, G^{(m)}] = \lim_{m\to\infty} \mathbb{P}_{\boldsymbol{D}(\omega_1)}(B^{(m)} \,|\, G^{(m)})$$

and assertion (i) follows using Theorem 2. Assertion (ii) then follows by the dominated convergence theorem. □

Finally, note from the discussion at the end of Section 6.4.2 that Corollary 2 holds with $G^{(m)}$ replaced by $\bar{G}^{(m)}$, where $\bar{G}^{(m)}$ is the event that the epidemic $E^{(m)}$ infects at least $\log m$ households.

7 Concluding comments

We have analysed the spread of an SIR epidemic within a population structure that features some significant departures from traditional homogeneous mixing; specifying both a local household structure and using random networks with an arbitrary degree distribution (with finite variance) to model potential 'global' contacts. Rigorous limit theorems were obtained, valid as the number of households $m \to \infty$, from which one can determine the probability of a major outbreak and the expected relative final size of such an outbreak. The potential usefulness of these results was verified by showing, numerically, that these asymptotic results provide good approximations for the behaviour of moderately sized finite populations.

As stated in Section 2, our results easily generalise to allow for unequal household sizes. For example, we can decompose R_* in a variable household size framework as $R_* = \sum_{n=1}^{\infty} \tilde{\rho}_n R_*^{(n)}$, where $\tilde{\rho}_n$ is the size-biased proportion of households of size n and $R_*^{(n)}$ is the threshold parameter R_* in the case of a fixed household size n. (The size-bias of $\tilde{\rho}_n$ arises because if a proportion ρ_n of households are of size n then an individual chosen uniformly at random is in a household of size n with probability proportional to $n\rho_n$; thus we require $\sum_{n=1}^{\infty} n\rho_n < \infty$.) Full details of this generalisation will appear in a forthcoming paper, which will discuss our model from a more applied viewpoint.

Another condition that we have required is that the variance, σ_D^2, of the degree distribution is finite. Whilst this is necessary for all of our proofs, the PGFs of C, $\tilde{C}$, B and $\tilde{B}$ are all well-defined so long as $\mu_D < \infty$ and numerical studies (along the lines of those encompassed by Figure 3) indicate that our methods at least give good approximations when $\sigma_D^2 = \infty$. This is particularly relevant in light of several of the studies cited by Newman (2003, Section III.C), which suggest that degree distributions which asymptotically follow some power law are appropriate models in some real-world situations. We note, however, that when $\sigma_D^2 = \infty$ it is not known (to our knowledge) whether self-loops and parallel edges remain sufficiently sparse in the network, so the argument that our results continue to hold if we condition on there being no such imperfections (second paragraph of Section 2) may not be valid.

Of course there are other features of our model that in many circumstances will be unrealistic. In particular, the method of construction of the random graph—pairing the half-edges uniformly at random—ensures not only that there are (asymptotically) very few 1-cycles (self-loops) and 2-cycles (parallel edges) in the resulting multigraph, but also that there are very few 3-cycles (triangles). Thus, in the asymptotic model that we analyse, individuals have no mutual acquaintances outside their household, which is unrealistic. Similarly the random graph model has very few edges which join individuals in the same pair of households, i.e. the acquaintances of two individuals are, with probability close to 1, all in distinct households. That this is the case stems from the construction of the random graph: although there is heterogeneity amongst the individuals (through differing degrees), the uniformly at random pairing of half-edges means that the mixing is still homogeneous—this being critical for the branching process approximations. In this sense it seems fair to say that our model incorporates some heterogeneity of both the individuals in the population (via the differing degrees of individuals and varying household sizes) and their mixing (having both local

and global infection).

Nevertheless, our model does capture some important heterogeneities which are present in real populations and which doubtless have a significant effect on the spread of disease through these populations. Some additional features, such as having the degree distribution D or the infection rates λ_L and λ_G depend on household size or incorporating correlation between the degrees of individuals within the same household can in principle be included in our model relatively simply, though the calculations quickly become very cumbersome.

The usual approach for obtaining fully rigorous results concerning the final size of a major epidemic on a random network is via the existence and uniqueness of a giant component in an associated bond percolation model (see e.g. Britton *et al.* (2007) and the discussion in Section 4 of Britton *et al.* (2008)). This requires that the infectious period is constant (though see Kenah and Robins (2007)) and fully rigorous results concerning the component structure of the percolation model, which may not be easy to prove. We have developed a different approach, which does not require a constant infectious period. Although not the focus of the paper, it seems plausible that our methods can be used to prove existence and uniqueness of a giant component for our random network (and indeed for other network models) and that they might also be applicable to epidemics on other random graph models, such as the random intersection graph considered by Britton *et al.* (2008).

Further study of this model will include an analysis of the effect of vaccination on epidemic spread (work ongoing) and it seems likely that a central limit theorem for the final size of a major outbreak might be derived using methods similar to those of Ball and Neal (2008).

Acknowledgements This research was supported by the UK Engineering and Physical Sciences Research Council, under research grant number EP/E038670/1 (Frank Ball and David Sirl) and by the Netherlands Organisation for Scientific Research (NWO) through a VICI grant awarded to Ronald Meester (Pieter Trapman).

References

Andersson, H. (1997). Epidemics in a population with social structures. *Math. Biosci.* **140,** 79–84.

Andersson, H. (1998). Limit theorems for a random graph epidemic model. *Ann. Appl. Probab.* **8,** 1331–1349.

Andersson, H. (1999). Epidemic models and social networks. *Math. Sci.* **24,** 128–147.

Ball, F. G. (1986). 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,** 289–310.

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

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

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

Ball, F. G. and Neal, P. J. (2003). The great circle epidemic model. *Stochastic Process. Appl.* **107,** 233–268.

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

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

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

Billingsley, P. (1968). *Convergence of probability measures.* John Wiley & Sons Inc., New York.

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

Britton, T., Janson, S. and Martin-Löf, A. (2007). Graphs with specified degree distributions, simple epidemics, and local vaccination strategies. *Adv. in Appl. Probab.* **39,** 922–948.

Durrett, R. (2006). *Random graph dynamics.* Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge.

Haccou, P., Jagers, P. and Vatutin, V. (2005). *Branching processes: Variation, growth, and extinction of populations.* Cambridge University Press, Cambridge.

van der Hofstad, R., Hooghiemstra, G. and Znamenski, D. (2007). Distances in random graphs with finite mean and infinite variance degrees. *Electron. J. Probab.* **12,** 703–766.

Janson, S. (2009). The probability that a random multigraph is simple. *Combin. Probab. Comput.* **18,** 205–225.

Kenah, E. and Robins, J. M. (2007). Second look at the spread of epidemics on networks. *Phys. Rev. E* **76,** 036113.

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

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

Newman, M. E. J. (2002). Spread of epidemic disease on networks. *Phys. Rev. E* **66,** 016128.

Newman, M. E. J. (2003). The structure and function of complex networks. *SIAM Rev.* **45,** 167–256 (electronic).

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

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

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442.

Whittle, P. (1955). The outcome of a stochastic epidemic—a note on Bailey's paper. *Biometrika* **42,** 116–122.