

Household epidemic models revisited

Frank Ball, Tom Britton and Peter Neal

May 26, 2025

Abstract

We analyse a generalized stochastic household epidemic model defined by a bivariate random variable (X_G, X_L) , representing the number of global and local infectious contacts that an infectious individual makes during their infectious period. Each global contact is selected uniformly among all individuals and each local contact is selected uniformly among all other household members. The main focus is when all households have the same size $h \geq 2$, and the number of households is large. Large population properties of the model are derived including a central limit theorem for the final size of a major epidemic, the proof of which utilises an enhanced embedding argument. A modification of the epidemic model is considered where local contacts are replaced by global contacts independently with probability p . We then prove monotonicity results for the probability of the major outbreak and the limiting final fraction infected z (conditioned on a major outbreak). a) The probability of a major outbreak is shown to be increasing in both h and p for any distribution of X_L . b) The final size z increases monotonically with both h and p if the probability generating function (pgf) of X_L is log-convex, which is satisfied by traditional household epidemic models where X_L has a mixed-Poisson distribution. Additionally, we provide counter examples to b) when the pgf of X_L is not log-convex.

Keywords: Household epidemic model; SIR epidemic; Final size; Large population limits; Branching process; Central limit theorem; Coupling; Monotonicity.

1 Introduction

The spreading of an infectious disease in a community is highly affected by how individuals mix. The earliest epidemic models assumed homogeneous mixing between all individuals, but this has later been generalized to allow for e.g. a multitype community (where mixing rates depend on the types of the two involved individuals involved), social networks (where individuals are connected through some underlying social structure, often characterized by a degree distribution for how many acquaintances individuals have), household epidemic models (where the population is divided into small units with higher mixing within the units) and spatial models (where the rate at which people mix depend on their spatial distance from each other), or a combination of these heterogeneous mixing features.

Stochastic household epidemic models were first considered by McKendrick [15], with increased interest in the past 30 years starting with Becker and Dietz [11] and Ball *et al.* [5]. A common definition of such a model uses the SIR or SEIR model concept to define an infectious period I , possibly having a random duration, during which an infectious individual has infectious contacts on the global scale and within the household according to independent Poisson processes having rates β_G and β_L (global and local). Each global contact is with a uniformly selected individual from the entire population (including also household members for convenience) and each local contact is with a uniformly selected individual of the same household. This implies that, conditional on the duration of the infectious period of an infective $I = x$, the random number of (not necessarily unique) global and local contacts are Poisson distributed with means $\beta_G x$ and $\beta_L x$, respectively. The corresponding unconditional numbers of global and local contacts (X_G, X_L) are hence mixed-Poisson distributed: $X_G \sim \text{MixPo}(\beta_G I)$ and $X_L \sim \text{MixPo}(\beta_L I)$, where both random variables depend on the same random variable I (the infectious period). If the population size is large it is unlikely that an individual makes multiple global contacts with the same individual. However, within small households multiple infectious contacts with the same individual are common.

The time dynamics of this household epidemic depends on the infectious period I and its possible preceding latent period L , but the final size describing who eventually gets infected is independent of L and only depends on I through (X_G, X_L) . In the present paper we study a more general model where (X_G, X_L) may follow an arbitrary but specified distribution on $\mathbb{Z}_+^2$. Hence, X_G and X_L need not be mixed-Poisson, nor do they have to be positively correlated as in the traditional model. In fact, it is quite possible that someone who becomes ill soon after infection makes fewer global infectious contacts but on the other hand more local (household) infectious contacts, thus making X_G and X_L negatively correlated. We consider also a modification of the model where each local contact is replaced by a global contact with probability p , in order to analyse what happens with the epidemic as more of the contacts become global, i.e. p increases.

The focus of this paper is the distribution of the final size of the epidemic as the population size, N , tends to infinity in the case where all households have the same size $h \geq 2$. We extend limiting results for the traditional epidemic models to our more general household epidemic model: a branching process approximation of the initial stages of the epidemic, an expression for the basic reproduction number R_* , and a law of large number and central limit theorem for the final size, conditional on the epidemic taking off. The central limit theorem (Theorem 2.1) uses an embedding argument based on the approach introduced in Scalia-Tomba [18] and successfully applied to household epidemics in [5]. However, given that the number of global infectious contacts made by infectives is not necessarily mixed-Poisson an additional layer of embedding is required to allow for a general distribution for global contacts, necessitating a novel proof. An explicit, and relatively easy to compute, expression for the variance of the central limit theorem is given in (2.4) with details on numerical computation given in Appendix A. In the absence of local infection, $X_L \equiv 0$, the model behaves as a homogeneously mixing epidemic and Theorem 2.1 holds for extensions of the Reed-Frost model considered in Martin-Löf [14] and Picard and Lefèvre [16], with Theorem 2.1 corresponding to [14], Theorem 1. Provided that $\mu_G = \mathbb{E}[X_G] < \infty$, in the limit as $N \rightarrow \infty$, we obtain the same asymptotic final size distribution whether the global contacts made by an individual are *with* or *without* replacement. The standard household

model assumes the local contacts are made with replacement and such a model is the main focus of this paper. However, the central limit theorem given in Theorem 2.1 holds if instead we assume that local contacts are made without replacement through minor modifications to the arguments.

We provide novel insight into how the household size, h , and the probability, p , that a local contact is replaced by a global contact, affect the probability that the epidemic takes off (a major epidemic occurs) and the (asymptotic) final size of a major epidemic. For larger households (increasing h) and a greater proportion of global contacts (increasing p), the epidemic more closely resembles a homogeneously mixing epidemic with fewer multiple contacts made by an infective with the same individual. Hence, intuitively the probability that the epidemic takes off and the final size of a major epidemic are increasing in both h and p . In Theorem 2.2, we show that this is the case for the probability a major epidemic outbreak regardless of the choice of (X_G, X_L) . The effect of h and p on the final size of the epidemic depends on the distribution of X_L , with X_G only affecting the final size through its mean μ_G . Specifically, in Theorem 2.3, we show that the final size of the epidemic is increasing in both h and p if the logarithm of the probability generating function (pgf) of X_L is convex. This is the case if X_L follows a mixed-Poisson distribution, so the monotonicity results hold for the standard construction of the SIR household model. However, for more general X_L the situation is more complex, with scenarios where the counter-intuitive result holds of smaller household sizes and increased local infectious contacts (with repeated contacts) leading to a larger final size.

The remainder of the paper is structured as follows. In Section 2, we present the general stochastic household epidemic model and state the main results of the paper, a central limit theorem for the final size of the epidemic (Theorem 2.1) and sufficient conditions for the probability of a major outbreak (Theorem 2.2) and the final size of the epidemic (Theorem 2.3) to be increasing in h and p . In Section 3, we present numerical illustrations of the main results, demonstrating the usefulness of the central limit theorem for finite N and providing examples where the final size of the epidemic is not increasing in h and/or p . The proofs of the central limit theorem and of the effects of h and p on the probability and final size of a major outbreak are given in Sections 4 and 5, respectively. In Section 6, we discuss the findings of the paper and possible extensions. Finally, in the appendices we present details of how to compute key quantities such as the probability of a major outbreak, the final proportion infected and the variance of the final size (Appendix A), along with Appendices B and C, which provide the proofs of Theorems 2.4 and 2.5 concerning how the final size of a major epidemic behaves near $p = 1$ (almost all local contacts replaced by global contacts) and as $h \rightarrow \infty$, respectively.

2 Model and main results

2.1 The general household epidemic model

The main ingredient for our epidemic model is the bivariate random variable (X_G, X_L) with distribution on $\mathbb{Z}_+^2$. X_G and X_L , respectively, denote the number of global and local contacts that a randomly selected individual makes.

Consider a population consisting of n households, all having size h . We investigate the

limiting situation where the population size $N = nh$ tends to infinity in such a way that the household size h remains fixed and the number of household $n \rightarrow \infty$.

An individual who gets infected draws their random pair (X_G, X_L) . Each of the X_G global infectious contacts is with a uniformly selected individual from the entire population. Each of the X_L local infectious contact is selected uniformly among the other $h - 1$ household members. All contact selections are made independently, and a susceptible individual who is contacted gets infected (and repeats the procedure), whereas contacts with previously infected individuals have no effect. It is worth pointing out that the contacts of an individual may not all be to unique individuals. In particular, the X_L local contacts may very well include multiple contacts to some individual(s). Such multiple contacts have no effect on the epidemic - it is the number of unique contacts that determines the propagation of the epidemic.

We consider also a modification of the model containing an additional parameter p . In this model, each local contact is, independently of everything else, replaced by a global contact with probability p . Thus, as p increases, there are fewer local and more global contacts.

The epidemic is initiated by a number of individuals, chosen uniformly at random from the population, being infected and all other individuals being uninfected and susceptible. The epidemic continues until it eventually stops by no new individuals getting infected. The final number infected is denoted Z , or $Z_{n,h}$ if we want to emphasize its dependence on the number and size of households. Clearly $1 \leq Z \leq N(= nh)$.

We denote the original model by $\mathcal{E}_{n,h}(X_G, X_L)$ and the model with swapping of local contacts to global contacts by $\mathcal{E}_{n,h}(X_G, X_L, p)$, so $\mathcal{E}_{n,h}(X_G, X_L)$ is identical to $\mathcal{E}_{n,h}(X_G, X_L, 0)$. Note that $\mathcal{E}_{n,h}(X_G, X_L, 1)$ is a homogeneously mixing epidemic.

2.2 Relation to traditional household epidemic models

As described in the introduction, traditional household epidemic models are often defined by infectious individuals having a random infectious period I , during which the infective has global contacts at rate β_G and household contacts at rate β_L (or $(h-1)\beta_L$, so β_L to each household member, but we choose the former parametrisation). In that case, the numbers of global and local contacts have distribution $(X_G, X_L) = (\text{MixPo}(\beta_G I), \text{MixPo}(\beta_L I))$, where we note that the two random variables are dependent having parameter containing the same random variable I . The final size of the epidemic depends only on the distribution of (X_G, X_L) , so the traditional model can be viewed as a subclass of $\mathcal{E}_{n,h}(X_G, X_L)$.

2.3 The $\mathcal{E}_{n,h}(X_G, X_L, p)$ model described as an $\mathcal{E}_{n,h}(X'_G, X'_L)$ model

It is worth mentioning that $\mathcal{E}_{n,h}(X_G, X_L, p)$ can, for a fixed value p , be described by $\mathcal{E}_{n,h}(X'_G, X'_L)$, i.e. the model without swapping, where the new random vector (X'_G, X'_L) is different from the original vector (X_G, X_L) . More precisely, the new vector is simply the (random) number of global and local contacts that occur *after* the swapping has happened. Suppose that $X_L = k$ and let $Y_L \sim \text{Bin}(k, p)$ denote how many contacts are swapped, then $X'_G = X_G + Y_L$ and $X'_L = X_L - Y_L$. Unconditionally, and showing the

dependence on p , we hence have

$$(X_G^{(p)}, X_L^{(p)}) = (X_G + Y_L^{(p)}, X_L - Y_L^{(p)}), \text{ where } Y_L^{(p)} \sim \text{MixBin}(X_L, p).$$

Note that, in the expressions above, $Y_L^{(p)}$ depends on X_L which is evident from the mixed-binomial distribution but hidden when writing the random vector $(X_G + Y_L^{(p)}, X_L - Y_L^{(p)})$.

2.4 Main results for the general household epidemic model

We now state our main results, firstly for the $\mathcal{E}_{n,h}(X_G, X_L)$ model and then for the $\mathcal{E}_{n,h}(X_G, X_L, p)$ model. These results are asymptotic results as $n \rightarrow \infty$ and for fixed h , we consider a sequence of epidemics, indexed by the number of households n . The epidemic $\mathcal{E}_{n,h}(X_G, X_L)$ is initiated by m_n individuals, chosen uniformly at random from the population, being infected, with the remaining $nh - m_n$ individuals being susceptible. Let $\bar{Z}_{n,h} = (nh)^{-1}Z_{n,h}$ denote the proportion of the population infected in $\mathcal{E}_{n,h}(X_G, X_L)$ and let $V_{n,h}$ denote the number of households where at least one individual is infected. Let $\mathcal{G}^{n,h} = \{V_{n,h} \geq \lfloor \log n \rfloor\}$, the event that the epidemic infects at least $k_n = \lfloor \log n \rfloor$ households. We say that a major epidemic has occurred if $\mathcal{G}^{n,h}$ occurs. The choice of $k_n = \lfloor \log n \rfloor$ households being infected to define a major epidemic is somewhat arbitrary and the results in this paper hold for any sequence k_n such that $k_n \rightarrow \infty$ and $k_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$.

Before stating Theorem 2.1, which extends known results for the traditional household epidemic models (where $(X_G, X_L) = (\text{MixPo}(\beta_G I), \text{MixPo}(\beta_L I))$) to a general random vector (X_G, X_L) , we require some extra notation.

Consider a household of size h , with initially 1 infective and $h - 1$ susceptibles. Let $\mathcal{E}_h^H(X_G, X_L)$ denote the ensuing *within-household* epidemic in which infected individuals make global and local infections according to the random pair (X_G, X_L) . Let C denote the number of global contacts that emanate from $\mathcal{E}_h^H(X_G, X_L)$. Let S denote the size of the susceptibility set of a typical individual in the household, where the susceptibility set of a given individual is the set of individuals, including themselves, who if infected globally will lead to the chosen individual being infected locally. A formal definition is given in Section 4.4. Note that S has support $\{1, 2, \dots, h\}$. Let

$$f_S(s) = \sum_{k=1}^h \mathbb{P}(S = k) s^k \quad (0 \leq s \leq 1) \quad (2.1)$$

and

$$f_C(s) = \sum_{k=0}^{\infty} \mathbb{P}(C = k) s^k \quad (0 \leq s \leq 1)$$

denote the pgfs of S and C , respectively. Note that the distributions of S and C depend on h but for notational convenience we suppress explicitly mentioning the dependence on h unless it is the focus of our study. Let $R_* = \mathbb{E}[C]$, the mean number of global contacts emanating from a household epidemic. Then, letting $\mu_G = \mathbb{E}[X_G]$, it is straightforward (see the appendix of Ball *et al.* [5]) to show that

$$R_* = \mathbb{E}[C] = \mu_G \mathbb{E}[S].$$

We now consider the household exposed to global infection. For $\pi \in [0, 1]$, let $\tilde{\mathcal{E}}_h^H(X_G, X_L, \pi)$ denote the following epidemic. Initially the whole household is susceptible. During the course of the epidemic, individuals avoid external infection independently with probability π . Infected individuals make global and local infections according to the random pair (X_G, X_L) . For $t \geq 0$, let $R(t)$ and $G(t)$ be respectively the total number infected in the household and the total number of global contacts emanating from the household in $\tilde{\mathcal{E}}_h^H(X_G, X_L, e^{-t})$.

For $t \geq 0$, let

$$\nu_R(t) \left(= \frac{1}{h} \mathbb{E}[R(t)] \right) = 1 - f_S(e^{-t}). \quad (2.2)$$

Suppose that $R_* > 1$ and define z to be the solution in $(0, 1]$ of

$$z = 1 - f_S(e^{-z\mu_G}) = \nu_R(\mu_G z). \quad (2.3)$$

(It is seen easily that z exists and is unique, since $\nu_R(\cdot)$ is concave, $\mu_G \nu_R'(0) = R_*$ and $\nu_R(\infty) = 1$.) Let

$$\begin{aligned} \sigma^2 = \frac{1}{h} \big[& (1 + b(\tau)\mu_G)^2 \text{var}(R(\tau)) + b(\tau)^2 h \nu_R(\tau) (\sigma_G^2 - \mu_G) \\ & + 2b(\tau)(1 + b(\tau)\mu_G) (\text{cov}(R(\tau), G(\tau)) - \mu_G \text{var}(R(\tau))) \big], \end{aligned} \quad (2.4)$$

where $\sigma_G^2 = \text{var}(X_G)$, $\tau = \mu_G z$ and $b(t) = \nu_R'(t)/[1 - \mu_G \nu_R'(t)]$.

Theorem 2.1. *Suppose that $R_* > 1$, and that there exists $m \geq 1$ such that $m_n = m$ for all sufficiently large n , and $a > 0$ such that $\mathbb{E}[X_G^{2+a}] < \infty$. Let $z > 0$ be given by (2.3) and ρ be the unique solution in $[0, 1)$ of*

$$\rho = f_C(\rho). \quad (2.5)$$

Then

$$\bar{Z}_{n,h} \xrightarrow{D} Z \quad \text{as } n \rightarrow \infty,$$

where the random variable Z has probability mass function

$$\mathbb{P}(Z = 0) = 1 - \mathbb{P}(Z = z) = \rho^m. \quad (2.6)$$

Furthermore, there exists $0 < \sigma^2 < \infty$ given by (2.4), such that

$$\sqrt{nh} (\bar{Z}_{n,h} - z) \Big| \mathcal{G}^{n,h} \xrightarrow{D} \mathcal{N}(0, \sigma^2) \quad \text{as } n \rightarrow \infty. \quad (2.7)$$

Theorem 2.1 holds if instead (X_G, X_L) are the numbers of unique individuals contacted by an infective in the population and their household, respectively. In this case X_L has support $\{0, 1, \dots, h-1\}$ and corresponds to sampling local infectious contacts *without* replacement from the other members of the household. Sampling without replacement affects the distributions of C and S but does not otherwise affect the derivation of the central limit theorem. We discuss this in more detail in Section 4.9.

In Section 4.8 we give two alternative but equivalent expressions for σ^2 . Note that z depends on the distribution of X_G only through its mean μ_G . In Appendix A, we give

expressions for $E[R(t)]$, $\text{var}(R(t))$, $\text{cov}(R(t), G(t))$ and $f_C(s)$ in terms of Gontcharoff polynomials, which enable ρ , z and σ^2 to be computed.

We now turn our attention to the $\mathcal{E}_{n,h}(X_G, X_L, p)$ model. Theorems 2.2 and 2.3 analyse $\pi^{(h,p)}$, the limiting probability of a major outbreak assuming a single initial infective, and $z^{(h,p)}$, the limiting final fraction getting infected in the event of a major outbreak, in particular their dependence on h and p for a given vector (X_G, X_L) . (To connect with Theorem 2.1, note that in an obvious notation, $\pi^{(h,p)} = 1 - \rho^{(h,p)}$.)

Theorem 2.2. *The limiting probability of a major outbreak $\pi^{(h,p)}$ is monotonically increasing in h and p for any random vector (X_G, X_L) .*

This hence means that the probability of a major outbreak increases if households are larger and/or local contacts are replaced by global contacts, both features making the epidemic model becoming closer to homogeneously mixing.

The second theorem concerns the final outbreak size $z^{(h,p)}$ assuming a major outbreak has occurred. Here the result depends on the *distribution* of X_L and in particular how much randomness there is. To this end we define the pgf of X_L : $f_{X_L}(s) = \sum_{k=1}^{\infty} s^k P(X_L = k)$.

Theorem 2.3. *Assume that $\log(f_{X_L}(s))$ is convex on $0 \leq s \leq 1$. Then the limiting final size $z^{(h,p)}$ is monotonically increasing in h and p for any X_G (dependent or independent of X_L).*

The mixed-Poisson distribution has a log-convex pgf, so Theorem 2.3 holds for the traditional household epidemic model. Log-convexity of the pgf of X_L implies that $\sigma_L^2 \geq \mu_L$ where $\sigma_L^2 = \text{var}(X_L)$ and $\mu_L = E[X_L]$. In Section 2.5, we present counter examples to Theorem 2.3 in the case where $\sigma_L^2 < \mu_L$. (See Theorem 2.4 (a) below.)

The following theorem is proved in Appendix B. We define $z^{(h,p)}$ to be strictly increasing (decreasing) in p near 1 if there exists $p_*^{(h)} \in [0, 1)$ such that $z^{(h,p)}$ is strictly increasing (decreasing) in p for $p \in [p_*^{(h)}, 1]$. For $\sigma_L^2 < \mu_L$, let $z^*(\mu_L, \sigma_L^2) = 1 - \frac{\mu_L - \sigma_L^2}{3\mu_L^2}$ and note that $z^*(\mu_L, \sigma_L^2) \in (0, 1)$ since X_L takes values in $\mathbb{Z}_+$. Let $\alpha = \mu_L + \mu_G$ and, for $\alpha > 1$, let $z_{\text{hom}}(\alpha)$ be the unique solution of $1 - z = e^{-\alpha z}$ in $(0, 1)$. Note that $z_{\text{hom}}(\alpha)$ is the proportion infected by a major outbreak in a homogeneously mixing epidemic, where each individual makes on average α infectious contacts.

Theorem 2.4. *Suppose that $h \geq 2$ and $\alpha = \mu_G + \mu_L > 1$, so $\mathcal{E}_{n,h}(X_G, X_L, 1)$ is supercritical.*

- (a) *If $\sigma_L^2 \geq \mu_L$, then $z^{(h,p)}$ is strictly increasing in p near 1.*
- (b) *Suppose that $\sigma_L^2 < \mu_L$. Then $z^{(h,p)}$ is strictly increasing in p near 1 if $z_{\text{hom}}(\alpha) < z^*(\mu_L, \sigma_L^2)$ and strictly decreasing in p near 1 if $z_{\text{hom}}(\alpha) > z^*(\mu_L, \sigma_L^2)$.*

Finally, we consider the final size $z^{(h,p)}$ in the limit as household size $h \rightarrow \infty$, with the proof given in Appendix C.

Theorem 2.5. *Suppose that $\alpha = \mu_G + \mu_L > 1$. Then for any $0 \leq p \leq 1$, $z^{(h,p)} \rightarrow z_{\text{hom}}(\alpha)$ as $h \rightarrow \infty$.*

2.5 Counter examples to Theorem 2.3 when $\sigma_L^2 < \mu_L$

In this section we provide simple counter examples showing that our main results are not necessarily true when X_L has too little randomness.

2.5.1 An example where final size decreases with household size

Consider the simple case where $X_L \equiv 1$, meaning that all infected individuals have exactly one household contact, uniformly selected among all household neighbours, and some fixed μ_G . Note that $\log(f_{X_L}(s)) = \log(s)$, so the pgf of X_L is a concave function. From (2.3), we know that the final size z is given by the solution in $(0, 1)$ of the equation $1 - z = f_S(e^{-\mu_G z})$.

We start with the case $h = 2$. The susceptibility set is then identical to 2, since the other household member must contact the index locally. So $S \equiv 2$, and the right-hand side of the final size equation equals $e^{-2\mu_G z}$.

When $h = 3$ the susceptibility set of an individual can in fact take only the values 1 or 3. The former if both housemates contact each other locally, and the latter otherwise. Consequently, we have $P(S_3 = 1) = 0.25$ and $P(S_3 = 3) = 0.75$. The right-hand side of the final size equation then equals $0.25e^{-\mu_G z} + 0.75e^{-3\mu_G z}$.

If we choose $\mu_G = 2$ the final size equation for $h = 2$ becomes $1 - z = e^{-4z}$, with solution $z_2 = 0.980$. When $h = 3$ the final size equation is $1 - z = 0.25e^{-2z} + 0.75e^{-6z}$ with solution $z_3 = 0.961$, thus showing that $h = 2$ gives a larger major outbreak than $h = 3$.

2.5.2 An example where moving local to global contacts lead to smaller final size

For an example such that the final size decreases as local contacts are swapped to global contacts we continue the example from the previous subsection with $X_L \equiv 1$, $\mu_G = 2$ and $h = 2$. When $p = 0$ we have the final size equation considered above, leading to final size $z_2 = 0.980$. If we swap *all* local contacts to global contacts (so $p = 1$) we simply have a homogeneous community where all individuals have $\mu_G = 3$ global contacts. The final size equation is then $1 - z = e^{-3z}$, with solution $z = 0.941$. So, if *all* local contacts are swapped to global contacts we get a *smaller* outbreak, implying that the final size cannot increase monotonically with p (in fact it decreases monotonically).

3 Numerical illustrations

3.1 Accuracy of asymptotic approximations

Figure 1 shows histograms of the fraction of the population infected, $\bar{Z}_{n,h}$, in the epidemic $\mathcal{E}_{n,h}(X_G, X_L)$ when $h = 2$, $X_G \sim \text{Po}(1)$ and $X_L \sim \text{Po}(1)$ independently, and $n = 125, 250, 500$ and $1,000$ (so the total population size $N = 250, 500, 1,000$ and $2,000$). Each epidemic is initiated by a single infective and each histogram is based on 100,000 simulations. Superimposed on each histogram is the density $\pi^{(h,0)} f_N(x)$, where $f_N(x)$ is the probability density function of the normal distribution $N(z, \frac{\sigma^2}{N})$, which approximates

the distribution of $\bar{Z}_{n,h}$ for a major outbreak by Theorem 2.1. For $N = 1,000$ and $2,000$, there is a clear distinction between major and minor outbreaks. The distinction is fairly clear for $N = 500$ but not when $N = 250$, where the choice of a cutoff to separate minor and major outbreaks is far from clear. Figure 2 shows histograms of 100,000 simulated major epidemics, using the same parameters as in Figure 1 and a cutoff of $z = 0.2$, with the $N(z, \frac{\sigma^2}{N})$ probability density function superimposed. Also shown are estimates of the skewness β_1 and kurtosis β_2 of the distribution of $\bar{Z}_{n,h}$ conditional upon a major outbreak. (Note that $\beta_1 = 0$ and $\beta_2 = 3$ for a normal distribution.) The asymptotic normal distribution gives a good approximation for $N \geq 500$. The true distribution of $\bar{Z}_{n,h}$ is skewed slightly to the left, with the degree of skewness decreasing as N increases, and slightly more peaked than the asymptotic normal distribution. Note that Theorem 2.1 implies that, for any $z_* \in (0, z)$, the probability a major outbreak infects at least a fraction z_* of the population tends to one as $n \rightarrow \infty$. In the numerical study below, following inspection of histograms, we define a major outbreak to be one with $\bar{Z}_{n,h} \geq 0.2$. Of course, the choice of cutoff depends on the parameters of an epidemic.

For a population of size N consisting of households of size $h = 2$, let π_N be the major outbreak probability, and z_N and σ_N be the mean and scaled standard deviation of the fraction infected by a major outbreak. (Thus $\sigma_N^2 = N \text{var}(\bar{Z}_{n,h} | \bar{Z}_{n,h} \geq 0.2)$, cf. Theorem 2.1.) Table 1 shows estimates of π_N , z_N and σ_N for the epidemic $\mathcal{E}_{n,h}(X_G, X_L)$ with household size $h = 2$ and various choices for the population size $N = nh$ and distribution for (X_G, X_L) . For each choice of N and distribution for (X_G, X_L) , $n_{\text{sim}} = 100,000$ epidemics were simulated and π_N was estimated by $\hat{\pi}_N$, the fraction of simulations with $\bar{Z}_{n,2} > 0.2$, with an approximate 95% confidence interval for π_N given by $\hat{\pi}_N \pm 1.96 \sqrt{\hat{\pi}_N(1 - \hat{\pi}_N)/n_{\text{sim}}}$. The simulations with $\bar{Z}_{n,h} \leq 0.2$ were then discarded and further simulations made until there were n_{sim} simulations with $\bar{Z}_{n,h} > 0.2$, which were used to estimate z_N and the scaled standard deviation σ_N . Let $\hat{z}_N$ and $\hat{\sigma}_N^2$ be the sample mean and variance of these n_{sim} simulations of $\bar{Z}_{n,h}$. Then z_N was estimated by $\hat{z}_N$, with an approximate 95% confidence interval given by $\hat{z}_N \pm 1.96 \hat{\sigma}_N / \sqrt{n_{\text{sim}}}$ and σ_N was estimated by $\hat{\sigma}_N = \sqrt{N} \tilde{\sigma}_N$, with an approximate 95% confidence interval given by $\left[\hat{\sigma}_N \sqrt{(n_{\text{sim}} - 1)/q_2}, \hat{\sigma}_N \sqrt{(n_{\text{sim}} - 1)/q_1} \right]$, where q_1 and q_2 are respectively the 2.5% and 97.5% quantiles of the $\chi_{n_{\text{sim}}-1}^2$ distribution. The $N = \infty$ entries in Table 1 give the asymptotic values π , z and σ given by Theorem 2.1.

The distributions of (X_G, X_L) in Table 1 all have $E[X_G] = E[X_L] = 1$ and are defined as follows. Constant: $(X_G, X_L) \equiv (1, 1)$. Binomial: $X_G \sim \text{Bin}(2, \frac{1}{2})$ and $X_L \sim \text{Bin}(2, \frac{1}{2})$ independently. Poisson: $X_G \sim \text{Po}(1)$ and $X_L \sim \text{Po}(1)$ independently. Mixed-Poisson I : $X_G|I \sim \text{Po}(I)$ and $X_L|I \sim \text{Po}(I)$ independently, where I is a single realisation of the given distribution. Note that in Table 1, the distributions are listed in increasing order of $\text{var}(X_G)$ and $\text{var}(X_L)$. There are no entries under $\hat{\pi}_N$ when (X_G, X_L) has the Constant distribution since, then $\pi = 1$ and for the values of N considered, π_N is extremely close to one.

It can be seen from Table 1 that $\hat{\pi}_N$ generally increases with N and π is an overestimate of π_N for finite N , as one would expect on intuitive grounds. Further, the convergence of π_N to its asymptotic value π is faster when X_G and X_L have a smaller variance. A similar comment holds for the fraction infected by a major outbreak z , though convergence of z_N to z is generally faster than that of π_N to π . Note that the confidence intervals for z_N are smaller than those for π_N . The simulations suggest that σ is an underestimate of

σ_N and that the scaled standard deviation of the size of a major outbreak converges to its asymptotic value more slowly than the mean. Caution is required when interpreting results for small N , since then the distinction between major and minor outbreaks is less clear, particularly for distributions with larger $\text{var}(X_G)$ and $\text{var}(X_L)$.

The accuracy of the asymptotic normal distribution as an approximation for the size of a major epidemic in a finite population is explored further in Table 2, which is based on $n_{\text{sim}} = 100,000$ simulations for each choice of distribution for (X_G, X_L) , population size n and household size h . For each such choice, the table shows the value of the Kolmogorov-Smirnov one-sample test statistic $D_{n_{\text{sim}}} = \sup_x |F_{n_{\text{sim}}}(x) - F(x)|$, where $F_{n_{\text{sim}}}$ is the empirical distribution function of the n_{sim} simulated fractions infected by a major outbreak and F is the distribution function of the approximating $N(z, \frac{\sigma^2}{N})$ distribution obtained using Theorem 2.1. Note that the corresponding tests all reject the null hypothesis that the fraction infected by a major outbreak follows a $N(z, \frac{\sigma^2}{N})$ distribution, with a very low p -value, since the true distribution is not $N(z, \frac{\sigma^2}{N})$ and the sample size n_{sim} is very large. Nevertheless, the values of $D_{n_{\text{sim}}}$ give a measure of the accuracy of the normal approximation. The values of $D_{n_{\text{sim}}}$ clearly decrease with N , consistent with the convergence in Theorem 2.1. They also generally decrease with increasing household size h , though that is less clear for the Constant and Binomial cases. Among the Poisson and mixed-Poisson choices for the distribution of (X_G, X_L) , the accuracy of the approximation generally decreases with increasing variance. Overall, Table 2 confirms the usefulness of the asymptotic normal approximation for finite population sizes.

3.2 Exploring model behaviour

In this section, we illustrate numerically the dependence of $\pi^{(h,p)}$, $z^{(h,p)}$ and $\sigma^{(h,p)}$ on h , p and the distribution of (X_G, X_L) . (Recall that h is the household size, p is the probability that a local contact is replaced by a global contact, $\pi^{(h,p)}$ is the asymptotic probability of a major outbreak, given one initial infective, and $z^{(h,p)}$ and $\sigma^{(h,p)}$ are the asymptotic mean and scaled standard deviation of the fraction of the population infected by a major outbreak.) Unless specified otherwise, the naming of the distributions follows exactly that used in Table 1. Figures 3 and 4 show the dependence of $z^{(h,p)}$ and $\sigma^{(h,p)}$ on h and p when (X_G, X_L) is (a) Constant, (b) Binomial, (c) Poisson and (d) Mixed-Poisson with $I \sim \text{Exp}(1)$. Note that in both the Poisson and Mixed-Poisson cases, $z^{(h,p)}$ is increasing in both h and p , as predicted by Theorem 2.3 since for both of these distributions $\log(f_{X_L}(s))$ is convex. The same holds for this Binomial case, even though then $\log(f_{X_L}(s))$ is not convex, so the condition that $\log(f_{X_L}(s))$ is convex is not necessary for the conclusions of Theorem 2.3 to hold. Observe that in this Constant case, where $(X_G, X_L) \equiv (1, 1)$, $z^{(h,p)}$ is decreasing with p when $h = 3, 4, 5, 6$, while $z^{(2,p)}$ first increases and then decreases with p , and $z^{(2,0)} = z^{(2,1)}$. The final observation has a simple explanation. When $p = 0$, an infected individual necessarily contacts their housemate, so the epidemic can be viewed as a homogeneously mixing one of fully infected households in which each infected household makes precisely two global contacts. When $p = 1$, the epidemic is homogeneously mixing with each individual making two (global) contacts. Hence, $z^{(2,0)} = z^{(2,1)}$. For $h = 3, 4, 5, 6$, $z^{(h,p)}$ is decreasing with h but the comparison with $h = 2$ depends on the value of p .

Turning to the scaled standard deviation, note that in the Poisson and Mixed-Poisson

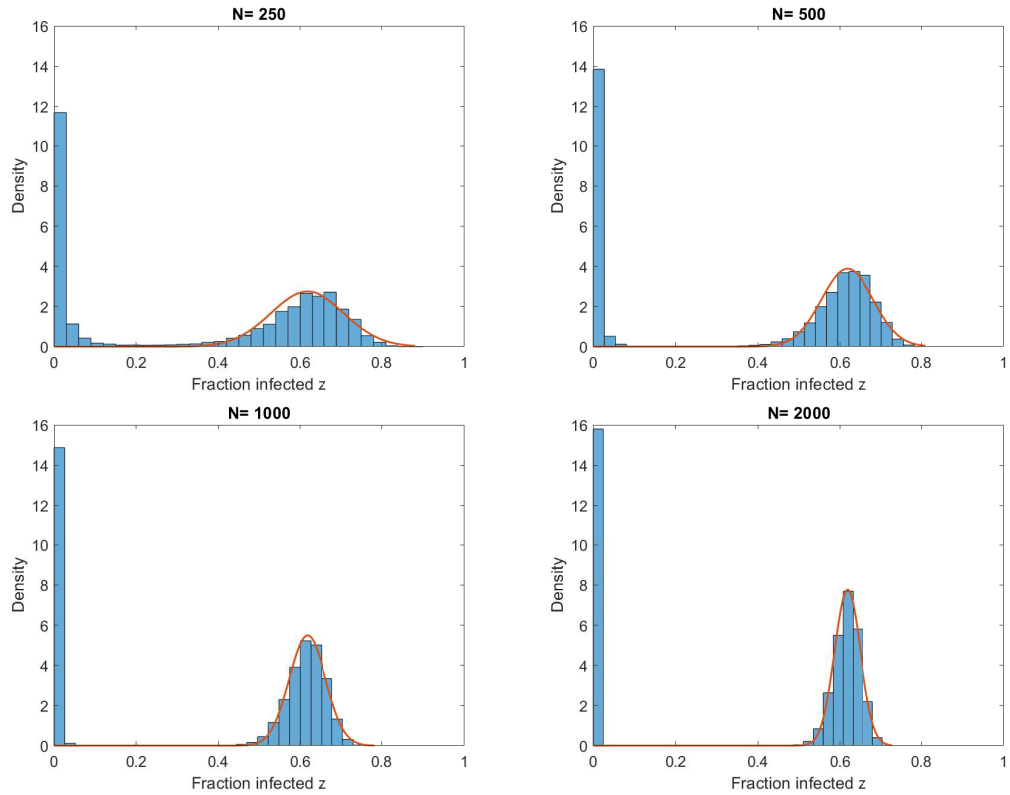


Figure 1: Histograms of 100,000 simulations of the fraction of the population infected in $\mathcal{E}_{n,2}(X_G, X_L)$ when $X_G \sim \text{Po}(1)$ and $X_L \sim \text{Po}(1)$ independently, for population sizes $N = nh = 250, 500, 1,000$ and $2,000$, with a normal approximation superimposed; see text for details.

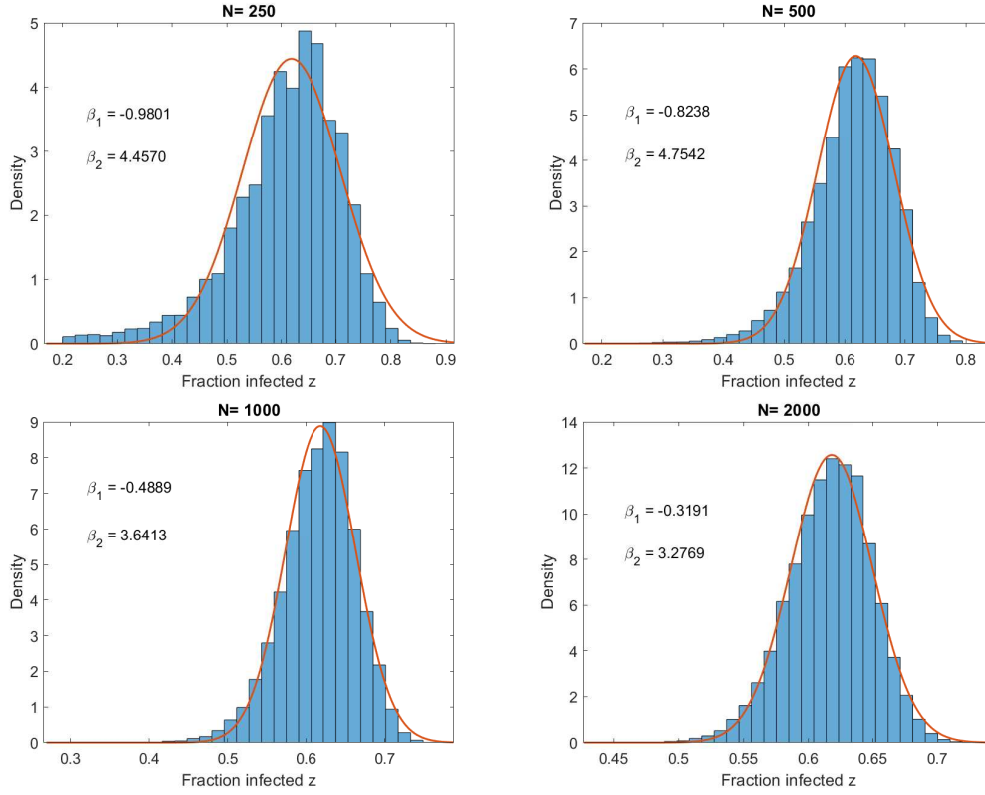


Figure 2: Histograms of 100,000 simulations of the fraction of the population infected in a major outbreak in $\mathcal{E}_{n,2}(X_G, X_L)$ when $X_G \sim \text{Po}(1)$ and $X_L \sim \text{Po}(1)$ independently, for population sizes $N = nh = 250, 500, 1,000$ and $2,000$, with a normal approximation superimposed; see text for details.

(X_G, X_L)	N	$\hat{\pi}_N$		$\hat{z}_N$		$\hat{\sigma}_N$	
Constant	250			0.7960	(0.7957, 0.7962)	0.7471	(0.7438, 0.7504)
	500			0.7966	(0.7964, 0.7968)	0.7402	(0.7369, 0.7434)
	1,000			0.7967	(0.7965, 0.7968)	0.7421	(0.7389, 0.7454)
	2,000			0.7967	(0.7966, 0.7968)	0.7394	(0.7362, 0.7427)
	5,000			0.7967	(0.7967, 0.7968)	0.7384	(0.7352, 0.7417)
	10,000			0.7968	(0.7967, 0.7968)	0.7367	(0.7335, 0.7399)
	∞			0.7968		0.7386	
Binomial	250	0.8103	(0.8078, 0.8127)	0.6762	(0.6757, 0.6766)	1.1642	(1.1591, 1.1693)
	500	0.8181	(0.8157, 0.8205)	0.6794	(0.6791, 0.6797)	1.1138	(1.1089, 1.1187)
	1,000	0.8199	(0.8175, 0.8223)	0.6805	(0.6803, 0.6808)	1.0963	(1.0915, 1.1011)
	2,000	0.8230	(0.8207, 0.8254)	0.6812	(0.6811, 0.6814)	1.0902	(1.0854, 1.0950)
	5,000	0.8233	(0.8210, 0.8257)	0.6814	(0.6813, 0.6815)	1.0928	(1.0881, 1.0976)
	10,000	0.8238	(0.8215, 0.8262)	0.6816	(0.6815, 0.6817)	1.0852	(1.0805, 1.0900)
	∞	0.8238		0.6817		1.0854	
Poisson	250	0.5916	(0.5885, 0.5946)	0.6084	(0.6078, 0.6091)	1.5814	(1.5745, 1.5884)
	500	0.6053	(0.6023, 0.6083)	0.6135	(0.6131, 0.6139)	1.5249	(1.5182, 1.5316)
	1,000	0.6126	(0.6096, 0.6157)	0.6159	(0.6156, 0.6162)	1.4670	(1.4606, 1.4735)
	2,000	0.6169	(0.6139, 0.6199)	0.6170	(0.6168, 0.6172)	1.4359	(1.4297, 1.4423)
	5,000	0.6153	(0.6123, 0.6183)	0.6178	(0.6177, 0.6179)	1.4270	(1.4208, 1.4333)
	10,000	0.6179	(0.6149, 0.6209)	0.6179	(0.6178, 0.6180)	1.4196	(1.4134, 1.4259)
	∞	0.6181		0.6181		1.4201	
Mixed-Poisson $I \sim \text{Gamma}(2, 2)$	250	0.3992	(0.3962, 0.4023)	0.5640	(0.5633, 0.5648)	1.9193	(1.9110, 1.9278)
	500	0.4127	(0.4097, 0.4158)	0.5661	(0.5655, 0.5666)	2.0076	(1.9989, 2.0165)
	1,000	0.4252	(0.4221, 0.4283)	0.5687	(0.5683, 0.5691)	1.9666	(1.9580, 1.9752)
	2,000	0.4284	(0.4253, 0.4314)	0.5708	(0.5705, 0.5710)	1.8831	(1.8749, 1.8914)
	5,000	0.4316	(0.4285, 0.4346)	0.5718	(0.5716, 0.5719)	1.8508	(1.8428, 1.8590)
	10,000	0.4350	(0.4319, 0.4381)	0.5722	(0.5721, 0.5723)	1.8461	(1.8381, 1.8542)
	∞	0.4391		0.5725		1.8378	
Mixed-Poisson $I \sim \text{Exp}(1)$	250	0.2933	(0.2905, 0.2961)	0.5357	(0.5348, 0.5365)	2.0870	(2.0779, 2.0962)
	500	0.3024	(0.2995, 0.3052)	0.5320	(0.5313, 0.5326)	2.3291	(2.3190, 2.3394)
	1,000	0.3150	(0.3122, 0.3179)	0.5326	(0.5322, 0.5331)	2.4141	(2.4035, 2.4247)
	2,000	0.3224	(0.3195, 0.3253)	0.5346	(0.5343, 0.5349)	2.3315	(2.3213, 2.3417)
	5,000	0.3254	(0.3225, 0.3283)	0.5359	(0.5357, 0.5361)	2.2697	(2.2598, 2.2797)
	10,000	0.3274	(0.3245, 0.3303)	0.5363	(0.5362, 0.5365)	2.2453	(2.2355, 2.2552)
	∞	0.3247		0.5368		2.2347	
Mixed-Poisson $I \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$	250	0.1892	(0.1868, 0.1917)	0.5013	(0.5004, 0.5021)	2.2397	(2.2299, 2.2496)
	500	0.1838	(0.1814, 0.1862)	0.4900	(0.4892, 0.4907)	2.6346	(2.6231, 2.6462)
	1,000	0.1900	(0.1876, 0.1924)	0.4831	(0.4825, 0.4837)	2.9566	(2.9437, 2.9696)
	2,000	0.1984	(0.1959, 0.2009)	0.4810	(0.4806, 0.4815)	3.1439	(3.1302, 3.1578)
	5,000	0.2011	(0.1986, 0.2036)	0.4816	(0.4813, 0.4819)	3.0990	(3.0854, 3.1126)
	10,000	0.2044	(0.2019, 0.2069)	0.4819	(0.4818, 0.4821)	3.0495	(3.0362, 3.0629)
	∞	0.2060		0.4829		2.9959	

Table 1: Simulation results against theoretical (asymptotic) calculations for epidemics with $h = 2$. See text for details.

(X_G, X_L)	N	$h = 2$	$h = 3$	$h = 4$	$h = 5$
Constant	250	0.0477	0.0391	0.0455	0.0469
	500	0.0350	0.0293	0.0314	0.0332
	1,000	0.0228	0.0196	0.0266	0.0226
	2,000	0.0168	0.0135	0.0174	0.0185
	5,000	0.0092	0.0098	0.0149	0.0099
	10,000	0.0083	0.0081	0.0085	0.0070
Binomial	250	0.0303	0.0272	0.0320	0.0349
	500	0.0215	0.0204	0.0224	0.0233
	1,000	0.0154	0.0127	0.0200	0.0197
	2,000	0.0107	0.0100	0.0107	0.0116
	5,000	0.0085	0.0072	0.0071	0.0100
	10,000	0.0048	0.0072	0.0057	0.0090
Poisson	250	0.0414	0.0357	0.0314	0.0342
	500	0.0285	0.0244	0.0225	0.0224
	1,000	0.0193	0.0173	0.0169	0.0162
	2,000	0.0154	0.0115	0.0123	0.0123
	5,000	0.0100	0.0081	0.0077	0.0110
	10,000	0.0076	0.0081	0.0062	0.0072
Mixed-Poisson $I \sim \text{Gamma}(2, 2)$	250	0.0469	0.0479	0.0431	0.0371
	500	0.0363	0.0322	0.0270	0.0267
	1,000	0.0276	0.0212	0.0196	0.0190
	2,000	0.0176	0.0149	0.0152	0.0149
	5,000	0.0115	0.0103	0.0098	0.0100
	10,000	0.0102	0.0078	0.0067	0.0070
Mixed-Poisson $I \sim \text{Exp}(1)$	250	0.0643	0.0542	0.0517	0.0493
	500	0.0481	0.0387	0.0342	0.0324
	1,000	0.0371	0.0267	0.0236	0.0224
	2,000	0.0240	0.0187	0.0174	0.0162
	5,000	0.0152	0.0121	0.0118	0.0098
	10,000	0.0110	0.0082	0.0079	0.0074
Mixed-Poisson $I \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$	250	0.0923	0.0589	0.0597	0.0623
	500	0.0797	0.0509	0.0461	0.0438
	1,000	0.0616	0.0357	0.0297	0.0283
	2,000	0.0434	0.0256	0.0202	0.0217
	5,000	0.0213	0.0164	0.0148	0.0122
	10,000	0.0134	0.0118	0.0100	0.0094

Table 2: Kolmogorov-Smirnov one-sample test statistics $D_{n_{\text{sim}}}$ for testing the goodness-of-fit of the approximating $N(z, \frac{\sigma^2}{N})$ distribution, obtained using Theorem 2.1, to a random sample of $n_{\text{sim}} = 100,000$ simulated major outbreaks for each parameter combination. See text for details.

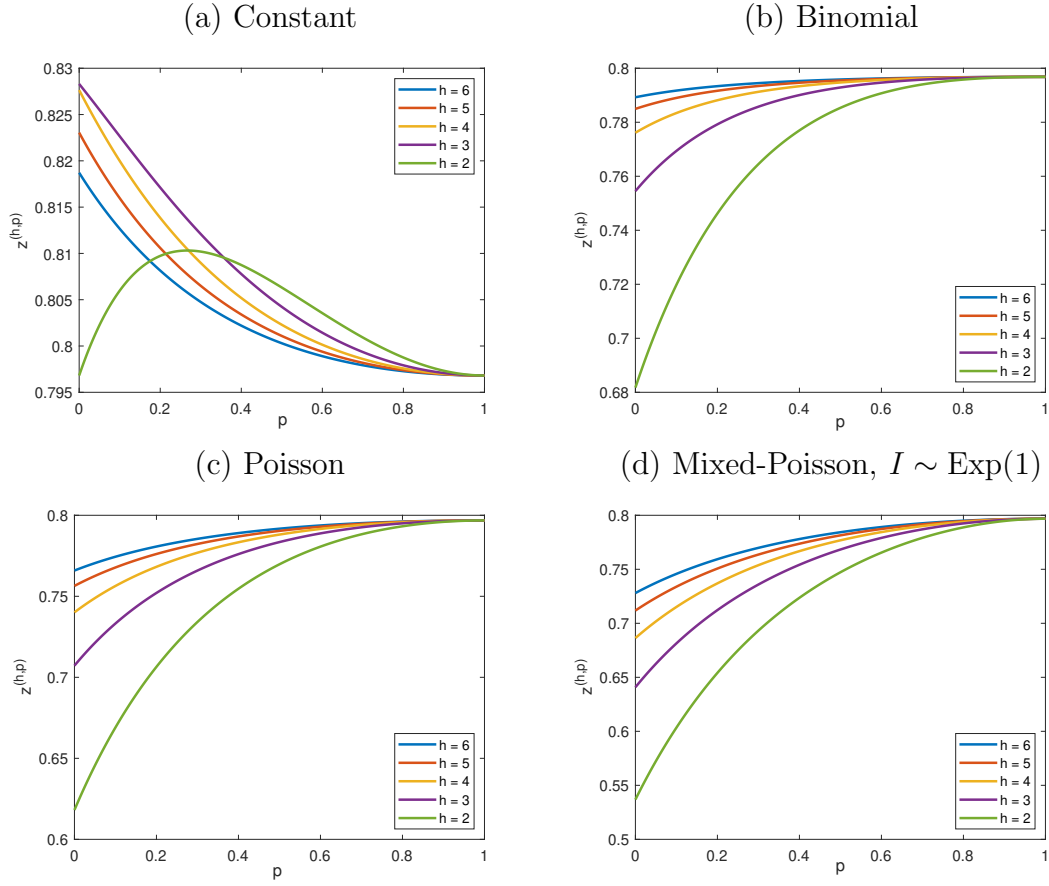


Figure 3: Graphs of the fraction of the population infected by a major outbreak, $z^{(h,p)}$, against p for different choices of household size h and distribution of (X_G, X_L) .

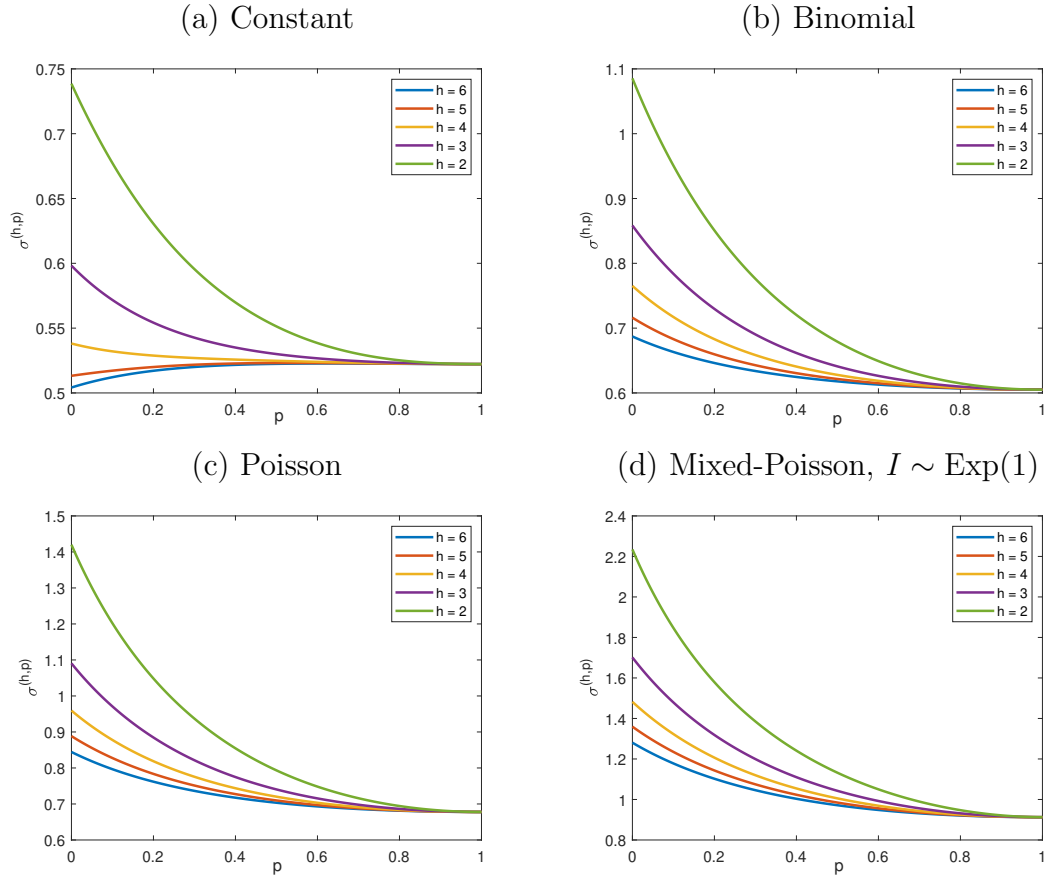


Figure 4: Graphs of the scaled variance, $\sigma^{(h,p)}$, of the fraction of the population infected by a major outbreak against p for different choices of household size h and distribution of (X_G, X_L) .

cases, $\sigma^{(h,p)}$ is decreasing in both h and p . The same observation holds for all of the cases we have considered in which $\log(f_{X_L}(s))$ is convex. A possible intuitive explanation is that increasing h and increasing p both have the effect of making the epidemic more homogeneous. The observation also holds for this Binomial case but, as we illustrate below, it and the above observation concerning $z^{(h,p)}$, do not hold generally when X_G and X_L follow independent Binomial distributions. In this Constant case, $\sigma^{(h,p)}$ is decreasing with h , however it is decreasing with p for $h = 2, 3, 4$ and increasing with p for $h = 5, 6$. Note that for the distributions considered, $z^{(h,p)}$ decreases and $\sigma^{(h,p)}$ increases as the variances of X_L and X_G increase.

Figure 5 shows plots of $z^{(h,p)}$ and $\sigma^{(h,p)}$ when $X_G \sim \text{Bin}(2, \frac{3}{4})$ and $X_L \sim \text{Bin}(2, \frac{3}{4})$ independently. Note that these plots are broadly similar to the corresponding plots in the above Constant case, except here $z^{(h,p)}$ is also non-monotonic with p when $h = 3$.

Finally, Figure 6 shows plots of the probability of a major outbreak, $\pi^{(h,p)}$, for various choices of distribution for (X_G, X_L) . Note that in all cases, $\pi^{(h,p)}$ is increasing in both h and p , as predicted by Theorem 2.2. For fixed (h, p) , $\pi^{(h,p)}$ decreases as the variances of X_G and X_L increase. Note that in the Poisson case, $\pi^{(h,p)} = z^{(h,p)}$, while in the other cases in which $\log(f_{X_L}(s))$ is convex, $\pi^{(h,p)} < z^{(h,p)}$ (see also Table 1 when $(h, p) = (2, 0)$). This

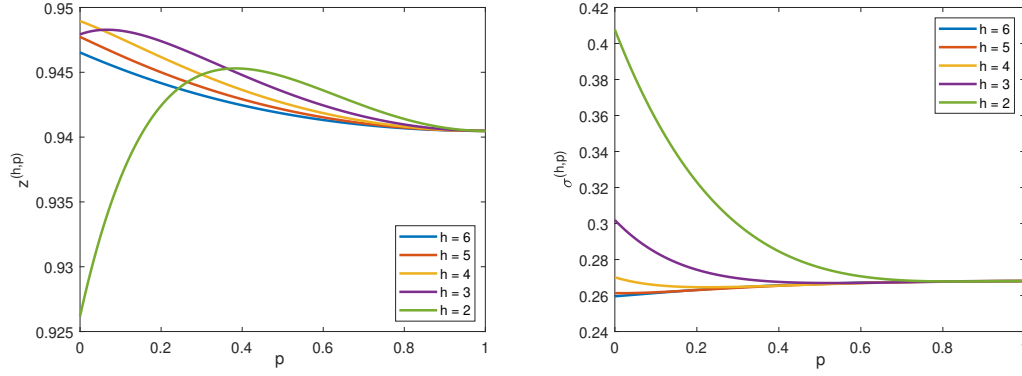


Figure 5: Graphs of $z^{(h,p)}$ (left panel) and $\sigma^{(h,p)}$ (right panel) when $X_G \sim \text{Bin}(2, \frac{3}{4})$ and $X_L \sim \text{Bin}(2, \frac{3}{4})$ independently.

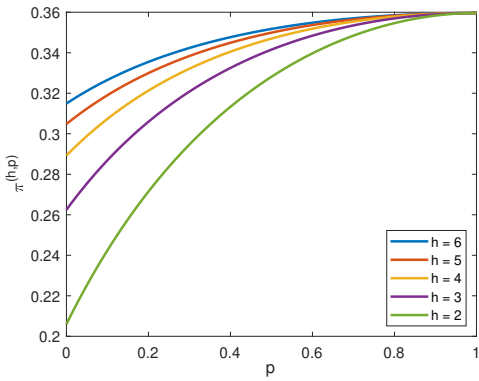
is usually the case for epidemic models. However, in the Binomial case, $\pi^{(h,p)} > z^{(h,p)}$.

4 Central limit theorem proof

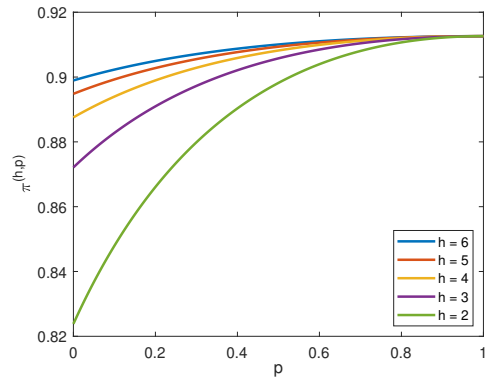
4.1 Introduction

In this section we prove Theorem 2.1. We begin in Section 4.2 by defining a sequence of $\mathcal{E}_{n,h}(X_G, X_L)$ epidemics, $\tilde{\mathcal{E}}_n$, indexed by n the number of households. In Section 4.3, we give a branching process approximation for the early stages of the epidemic and show that the probability of a minor outbreak (which infects at most $\lfloor \log n \rfloor$ households) converges to ρ^m as $n \rightarrow \infty$, where ρ satisfies (2.5). In Section 4.4 we define the embedding process which is utilised for the central limit theorem. The embedding process is based on a Sellke construction, see Sellke [17], of the epidemic with an extra level of embedding. We define a sequence of epidemics $\mathcal{E}_n$ based on the embedded construction and show that $\tilde{\mathcal{E}}_n$ and $\mathcal{E}_n$ can be coupled to give the same epidemic final size, albeit with potentially different global infectors of individuals. This enables us to focus on the embedded construction in the remainder of the section. In Section 4.5, we prove a law of large numbers result and show that $\bar{Z}_{n,h} \xrightarrow{D} Z$ as $n \rightarrow \infty$, where the probability mass function of Z satisfies (2.6). In Section 4.6 we prove Theorem 2.1 by exploiting an upper and lower bound for the proportion infected in the event of a major epidemic and showing that both these bounds have the same limit. A key component in the proof is Theorem 4.1 whose proof is postponed to Section 4.7. In Section 4.8, we discuss σ^2 and give two equivalent expressions for σ^2 in (4.30) and (4.31). The first expression, (4.30), arises naturally in the proof of Theorem 2.1, whilst the second expression, (4.31), is often simpler to work with in terms of computing σ^2 numerically. The proof that the expressions in (4.30) and (4.31) are equal, and equivalent to that given by (2.4) in Section 2.4 are deferred to Appendix D. Finally, in Section 4.9 we discuss the minor modifications to the central limit theorem for the case where the contacts (X_G, X_L) are sampled *without* replacement from the population and household, respectively.

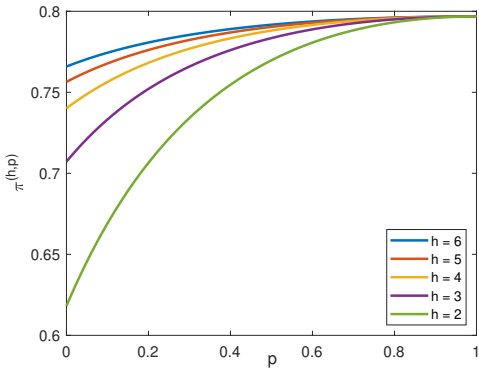
(a) Mixed-Poisson, $I \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2})$



(b) Binomial



(c) Poisson



(d) Mixed-Poisson, $I \sim \text{Exp}(1)$

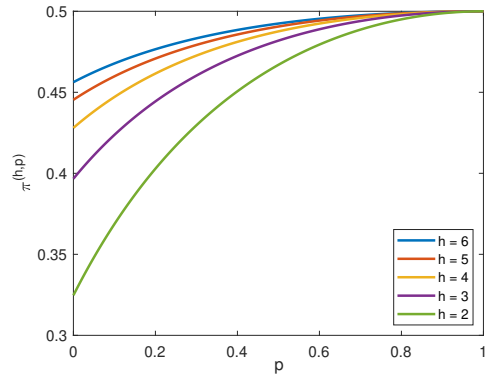


Figure 6: Graphs of the probability of a major outbreak, $\pi^{(h,p)}$, against p for different choices of household size h and distribution of (X_G, X_L) .

4.2 Model description

For $i = 1, 2, \dots$ and $j = 1, 2, \dots, h$, let $\mathbf{X}_{ij}$ be i.i.d. copies of $\mathbf{X} = (X_G, X_L)$ with $\mathbf{X}_{ij}$ determining the number of global and local infectious contacts made by the j^{th} individual in household i . We construct the epidemic $\tilde{\mathcal{E}}_n$ using $\{\mathbf{X}_{ij} = (X_{G,(i,j)}, X_{L,(i,j)}); i = 1, 2, \dots, n, j = 1, 2, \dots, h\}$ as follows. We assign to each individual a list of household contacts $\mathbf{H}_{ij} = (H_{ij1}, H_{ij2}, \dots)$, where H_{ijk} is the individual within the household contacted by the k^{th} household infectious contact made by individual j in household i . (Note that the $\{H_{ijk}\}$ s are independent and uniformly distributed on $\{1, 2, \dots, h\} \setminus \{j\}$.) The individual (i, j) makes a household infectious contact with individual (i, l) if $l \in \{H_{ij1}, H_{ij2}, \dots, H_{ijX_{L,(i,j)}}\}$. In addition, for each n , we let $U_1^n, U_2^n, \dots$ be i.i.d. copies of U^n , where

$$P(U^n = (i, j)) = \frac{1}{nh} \quad (i = 1, 2, \dots, n; j = 1, 2, \dots, h).$$

Therefore, U^n can be used to choose an individual uniformly at random from the population underlying the epidemic $\tilde{\mathcal{E}}_n$.

The epidemic $\tilde{\mathcal{E}}_n$ starts with m_n initial infectives, and we assume that there exists $m \geq 1$ such that $m_n = m$ for all sufficiently large n . The m_n initial infectives are given by the first m_n unique U^n . For $n = 1, 2, \dots$ and $k = 1, 2, \dots$, let $\mathcal{J}_k^n = \cup_{i=1}^k \{U_i^n\}$. Then $\mathcal{J}_{K_n}^n$ denotes the set of initial infectives where K_n satisfies

$$K_n = \min \{k : |\mathcal{J}_k^n| = m_n\}.$$

The epidemic is then constructed by considering infectives one at a time. Suppose that prior to considering individual (i_0, j_0) , say, there has been a total of M global infectious contacts. The local infectious contacts made by individual (i_0, j_0) are governed by $X_{L,(i_0,j_0)}$ and $\mathbf{H}_{i_0j_0}$. The global infectious contacts made by individual (i_0, j_0) are with, if $X_{G,(i_0,j_0)} > 0$, individuals $U_{M+1}^n, U_{M+2}^n, \dots, U_{M+X_{G,(i_0,j_0)}}^n$. The process continues until there are no more infectives in the population.

4.3 Branching process approximation

For the epidemic $\tilde{\mathcal{E}}_n$ we have defined a major epidemic as one that infects at least $k_n = \lfloor \log n \rfloor$ households. Therefore we define a minor epidemic as one that infects fewer than $\lfloor \log n \rfloor$ households, that is, if $V_{n,h} < \lfloor \log n \rfloor$ and in this section we show that

$$P(V_{n,h} < \lfloor \log n \rfloor) \rightarrow \rho^m \quad \text{as } n \rightarrow \infty, \quad (4.1)$$

where ρ satisfies (2.5).

In order to prove (4.1), we couple the sequence of epidemics $\tilde{\mathcal{E}}_n$ to a Galton-Watson branching process $\mathcal{B}$. Specifically, the branching process $\mathcal{B}$ has m ancestors and the number of offspring from individuals are i.i.d. copies of C , defined just before (2.1) in Section 2.4. Hence, ρ denotes the extinction probability of the branching process $\mathcal{B}$. Let V denote the total size, including initial ancestors, of the branching process $\mathcal{B}$.

Lemma 4.1. *For any $k = 1, 2, \dots$,*

$$P(V_{n,h} \leq k) \rightarrow P(V \leq k) \quad \text{as } n \rightarrow \infty.$$

Proof. We prove the lemma by constructing $\tilde{\mathcal{E}}_n$ and $\mathcal{B}$ on a common probability space. For $i = 1, 2, \dots$ and $j = 1, 2, \dots, h$, let $\bar{\mathbf{X}}_{ij}$ be i.i.d. copies of $\mathbf{X}$ and let $\bar{\mathbf{H}}_{ij}$ be independent with $\bar{\mathbf{H}}_{ij} \stackrel{D}{=} \mathbf{H}_{1j}$. For $i = 1, 2, \dots$, let C_i denote the number of global contacts emanating from the i^{th} household epidemic constructed using $\{\bar{\mathbf{X}}_{ij}, \bar{\mathbf{H}}_{ij}; j = 1, 2, \dots, h\}$, where the individual $(i, 1)$ is the initial infective in the household. Let C_i denote the number of offspring of the i^{th} individual in the branching process $\mathcal{B}$ with $C_1, C_2, \dots, C_m$ denoting the offspring of the m ancestors.

Let $\tilde{U}_1^n, \tilde{U}_2^n, \dots$ be i.i.d. copies of $\tilde{U}^n$, where $\tilde{U}^n$ is a discrete uniform distribution on $\{1, 2, \dots, n\}$. We construct a realisation of $\tilde{\mathcal{E}}_n$ by assigning the i^{th} global contact in $\tilde{\mathcal{E}}_n$ to household $\tilde{U}_i^n$. Given that household $\tilde{U}_i^n$ has not previously been infected we assign infectious histories $\{\bar{\mathbf{X}}_{ij}, \bar{\mathbf{H}}_{ij}; j = 1, 2, \dots, h\}$ to the individuals in household $\tilde{U}_i^n$ and assume that the individual contacted globally is individual $(i, 1)$. Therefore the number of global contacts emanating from the first household epidemic in household $\tilde{U}_i^n$ is C_i .

Let $M_n = \min \left\{ k > 1 : \tilde{U}_k^n \in \left\{ \tilde{U}_1^n, \tilde{U}_2^n, \dots, \tilde{U}_{k-1}^n \right\} \right\}$, the number of global contacts that occur until the first attempted infection of a previously infected household. This is the well known *Birthday Problem*, see for example Ball and Donnelly [3], and

$$\mathbb{P}(M_n \leq k) \leq \frac{k(k-1)}{2n} \quad (k = 2, 3, \dots, n). \quad (4.2)$$

Therefore, for any $k = 1, 2, \dots$,

$$\begin{aligned} \mathbb{P}(V_{n,h} \leq k) &= \mathbb{P}(V_{n,h} \leq k | M_n > k) \mathbb{P}(M_n > k) + \mathbb{P}(V_{n,h} \leq k | M_n \leq k) \mathbb{P}(M_n \leq k) \\ &= \mathbb{P}(V \leq k | M_n > k) \mathbb{P}(M_n > k) + \mathbb{P}(V_{n,h} \leq k | M_n \leq k) \mathbb{P}(M_n \leq k) \\ &\rightarrow \mathbb{P}(V \leq k) \quad \text{as } n \rightarrow \infty, \end{aligned} \quad (4.3)$$

as required. $\square$

Given that (4.2) implies $\mathbb{P}(M_n > \lfloor \log n \rfloor) \rightarrow 1$ as $n \rightarrow \infty$, it is straightforward to show that

$$|\mathbb{P}(V_{n,h} \leq \log n) - \mathbb{P}(V \leq \log n)| \rightarrow 0 \quad \text{as } n \rightarrow \infty.$$

Since $\mathbb{P}(k < V < \infty) \rightarrow 0$ as $k \rightarrow \infty$ and $\mathbb{P}(V < \infty) = \rho^m$, it follows by the triangle inequality that

$$\begin{aligned} |\mathbb{P}(V_{n,h} \leq \log n) - \rho^m| &\leq |\mathbb{P}(V_{n,h} \leq \log n) - \mathbb{P}(V \leq \lfloor \log n \rfloor)| + \mathbb{P}(\lfloor \log n \rfloor < V < \infty) \\ &\rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

4.4 Embedding

In order to obtain a central limit theorem for the final size, we use an embedding argument similar to [18], [5] and Ball and Neal [7], utilising a Sellke ([17]) construction of the epidemic. This involves taking an alternative approach to modelling global infection but we show that the final size of the epidemic is unchanged. Specifically, we assume that any given individual encounters global infections at the points of a homogeneous unit rate Poisson point process as the amount of global infectious pressure they are exposed to increases. In [5] and [7], an infectious individual with infectious period I contributes

$\lambda_G I/N$ units of global infectious to each individual in the population with the number of new global infectious encounters arising following a Poisson distribution with mean $\lambda_G I$. In our setting, each infective makes a given number of global contacts distributed according to X_G . This means that we cannot directly apply the embedding arguments used in the earlier referenced works but require an additional layer of embedding which links the total number of global contacts in the epidemic process to the independent Poisson point processes of global contacts attached to individuals.

Before defining a sequence of embedded epidemics, $\mathcal{E}_n$, indexed by n the number of households and showing that $\mathcal{E}_n$ and $\tilde{\mathcal{E}}_n$ give the same final size, we require some additional notation. This includes the formal definition of a susceptibility set whose pgf plays a key role in obtaining, z , the mean final proportion infected in a major outbreak given by (2.3).

For $i = 1, 2, \dots$ and $j, l = 1, 2, \dots, h$, let $(i, j) \rightsquigarrow (i, l)$ denote that there is a path of household infection from individual (i, j) to individual (i, l) with the convention that $(i, j) \rightsquigarrow (i, j)$. Note that $(i, j) \rightsquigarrow (i, l)$ is determined by $\{(X_{L,(i,k)}, \mathbf{H}_{ik}); k = 1, 2, \dots, h\}$. For $i = 1, 2, \dots$ and $j = 1, 2, \dots, h$, let $\mathcal{S}^{ij}$ denote the susceptibility set of individual (i, j) which is defined to be

$$\mathcal{S}^{ij} = \{l \in \{1, 2, \dots, h\} : (i, l) \rightsquigarrow (i, j)\}.$$

That is, $\mathcal{S}^{ij}$ is the set of individuals whom if infected by a global infection will infect individual (i, j) , if susceptible, via a chain of local infections within the household. Let $S_{ij} = |\mathcal{S}^{ij}|$ denote the size of the susceptibility set of individual (i, j) . Note that for all (i, j) , $S_{ij} \stackrel{D}{=} S_{11}$ and for $k \neq i$, S_{ij} and S_{kl} are independent with the pgf of S_{11} given by $f_S(s)$, cf. (2.1).

Finally, before introducing the embedded epidemic process we attach to each individual (i, j) an independent, homogeneous Poisson point process, η_{ij} , with rate 1. For $t \geq 0$, let $\zeta_{ij}(t)$ denote the number of points of η_{ij} in $[0, t]$. Thus $\zeta_{ij}(t) \sim \text{Po}(t)$.

Suppose that global contacts occur with an individual at the points of a homogeneous Poisson point process with rate 1. Specifically, we assume that individual (i, j) receives global contacts at the points of η_{ij} as the individual is exposed to increasing amounts of global infection. We assume that when an individual is infected globally the local household epidemic from that individual occurs instantaneously. Let $\chi_{ij}(t) = 1 - \prod_{l \in \mathcal{S}^{ij}} 1_{\{\zeta_{il}(t)=0\}}$. Then $\chi_{ij}(t)$ is an indicator random variable for whether or not individual (i, j) is infected when all members of the population are exposed to t units of global infectious pressure, since an individual is infected once somebody in their susceptibility set receives a global infectious contact.

For $i = 1, 2, \dots$ and $t \geq 0$, let $(R_i(t), G_i(t), Y_i(t))$ be a trivariate random variable determining the state of household when each individual is exposed to t units of global infection. Let $R_i(t) = \sum_{j=1}^h \chi_{ij}(t)$ denote the number of individuals infected in the household, let $G_i(t) = \sum_{j=1}^h X_{G,(i,j)} \chi_{ij}(t)$ denote the number of global contacts made by those infected in the household and let $Y_i(t) = \sum_{j=1}^h \zeta_{ij}(t) [= \sum_{j=1}^h \chi_{ij}(t) \zeta_{ij}(t)]$ denote the number of global contacts made into the household. By construction the $\{(R_i(t), G_i(t))\}$ s are i.i.d. copies of $(R(t), G(t))$, defined in Section 2.

For $t \geq 0$, let $\nu_R(t) = \mathbb{E}[R_1(t)]/h = \mathbb{E}[\chi_{11}(t)] = 1 - f_S(e^{-t})$, cf. (2.2). Since, for all $t \geq 0$,

$X_{G,(1,1)}$ and $\chi_{11}(t)$ are independent, we have that

$$\nu_G(t) = \frac{1}{h} \mathbb{E}[G_1(t)] = \mu_G[1 - f_S(e^{-t})].$$

Finally, $\nu_Y(t) = \mathbb{E}[Y(t)]/h = t$.

We are now in position to describe the construction of the embedded epidemic process $\mathcal{E}_n$ and obtain an expression for the proportion, $\bar{Z}_{n,h}$, of the population infected.

The embedded epidemic process considers each individual, and hence, household being exposed to infection at a constant rate. If each member of the population is exposed to t units of global infection, the total number of global infectious contacts is random and distributed according to $\text{Po}(nht)$, the number of points in $[0, t]$ of the Poisson point process η^n , where η^n is defined to be the superposition of the Poisson processes $\{\eta_{ij}; i = 1, 2, \dots, n, j = 1, 2, \dots, h\}$. To study the original epidemic process using the embedded epidemic process, we reverse this procedure and for a given $x \in \mathbb{R}^+$, we find the random time $S_n(x)$ such that the number of global contacts in the population on the interval $[0, S_n(x)]$ is equal to $\lfloor xnh \rfloor$. More specifically, for $n = 1, 2, \dots$ and $x \geq 0$, let

$$S_n(x) = \min \left\{ t \geq 0 : \sum_{i=1}^n Y_i(t) = \lfloor xnh \rfloor \right\}. \quad (4.4)$$

Let T_0^n denote the number of global infections required to generate m_n infectives to initiate the epidemic, and remember that $m_n = m$ for all sufficiently large n . Therefore $T_0^n \xrightarrow{p} m$ as $n \rightarrow \infty$. Let $\bar{T}_0^n = T_0^n/(nh)$. Then

$$S_n(\bar{T}_0^n) = \min \left\{ t \geq 0 : \sum_{i=1}^n Y_i(t) = T_0^n \right\}$$

is the initial amount of global infection in the epidemic process $\mathcal{E}_n$ to generate m_n infectives (T_0^n global infectious contacts). We say that the set of individuals whose susceptibility set contains an initial infective form generation 0 of $\mathcal{E}_n$. (Therefore generation 0 of $\mathcal{E}_n$ is obtained by running the local epidemics from the initial infectives.) Generation 0 will generate $\sum_{i=1}^n G_i(S_n(\bar{T}_0^n))$ global infectious contacts. Thus

$$T_1^n (= nh\bar{T}_1^n) = T_0^n + \sum_{i=1}^n G_i(S_n(\bar{T}_0^n)),$$

is the number of global infections, including those required for the initial infectives, after the global infections emanating from generation 0 have been considered. Following [5], Section 4.2.2, we can define $T_0^n, T_1^n, \dots$, with $\bar{T}_k^n = T_k^n/(nh)$, to satisfy, for $k = 0, 1, \dots$,

$$T_{k+1}^n (= nh\bar{T}_{k+1}^n) = T_0^n + \sum_{i=1}^n G_i(S_n(\bar{T}_k^n)).$$

For $k = 1, 2, \dots$, we say an individual belongs to the k^{th} generation of infectives if the first time a member of their susceptibility set is infected globally is by a member of generation $k - 1$. Using the embedding process an individual (i, j) belongs to generation k if

$$\chi_{ij}(S_n(\bar{T}_{k-1}^n)) = 0 \quad \text{and} \quad \chi_{ij}(S_n(\bar{T}_k^n)) = 1,$$

and T_{k+1}^n is the total number of global infections, including those required for the initial infectives, from the first k generations of infectives. The process continues until there are no additional global infections created in a generation. That is, $T_{k+1}^n = T_k^n$, and consequently we can define $T_\infty^n = nh\bar{T}_\infty^n$ to satisfy

$$\bar{T}_\infty^n = \min \left\{ x \geq 0 : T_0^n + \sum_{i=1}^n G_i(S_n(x)) = \lfloor xnh \rfloor \left(= \sum_{i=1}^n Y_i(S_n(x)) \right) \right\}. \quad (4.5)$$

Hence,

$$\bar{T}_\infty^n = \bar{T}_0^n + \frac{1}{nh} \sum_{i=1}^n G_i(S_n(\bar{T}_\infty^n)) \left(= \frac{1}{nh} \sum_{i=1}^n Y_i(S_n(\bar{T}_\infty^n)) \right).$$

Therefore, $\bar{Z}_{n,h}$, the proportion of the population infected by the epidemic $\mathcal{E}_n$, satisfies

$$\bar{Z}_{n,h} = \frac{1}{nh} \sum_{i=1}^n R_i(S_n(\bar{T}_\infty^n)) = \frac{1}{nh} \sum_{i=1}^n \sum_{j=1}^h \chi_{ij}(S_n(\bar{T}_\infty^n)).$$

We show how the epidemic processes $\mathcal{E}_n$ and $\tilde{\mathcal{E}}_n$ can be coupled to give the same final size. We construct $\mathcal{E}_n$ using $\{\mathbf{X}_{ij} = (X_{G,(i,j)}, X_{L,(i,j)}), \mathbf{H}_{ij}, \eta_{ij}; i = 1, 2, \dots, n, j = 1, 2, \dots, h\}$. To construct $\tilde{\mathcal{E}}_n$ from $\mathcal{E}_n$, we use $\{\mathbf{X}_{ij} = (X_{G,(i,j)}, X_{L,(i,j)}), \mathbf{H}_{ij}; i = 1, 2, \dots, n, j = 1, 2, \dots, h\}$ so local epidemics are unchanged and the number of global contacts made by a given individual are the same in both processes. Using $\{\eta_{ij}; i = 1, 2, \dots, n, j = 1, 2, \dots, h\}$, we construct $U_1^n, U_2^n, \dots$. For $k = 1, 2, \dots$, we set $U_k^n = (i', j')$ if the k^{th} point of η^n comes from $\eta_{i'j'}$. This construction means that the initial m_n infectives in $\tilde{\mathcal{E}}_n$ are $\mathcal{I}_{K_n}^n$ and that the individual contacted by the k^{th} global contact is the same in both epidemics although the assignment of the infector might be different. Consequently, those individuals whose susceptibility sets have been globally infected, and thus are guaranteed to be infected, by the first t global infections in $\tilde{\mathcal{E}}_n$, is precisely the set of individuals for whom $\chi_{ij}(S_n(t/nh)) = 1$ ($i = 1, 2, \dots, n; j = 1, 2, \dots, h$). Therefore, T_∞^n is the total number of global infectious contacts in both $\mathcal{E}_n$ and $\tilde{\mathcal{E}}_n$, with $\bar{Z}_{n,h}$ denoting the proportion of individuals infected.

4.5 Law of large numbers

In this section we prove that the proportion of the population infected, $\bar{Z}_{n,h}$, converges to a random variable Z whose probability mass function is defined in (2.6).

Lemma 4.2. *Suppose that there exists $m \in \mathbb{N}$ such that $m_n = m$ for all sufficiently large n . For $R_* > 1$, there exists $0 < \tau < \infty$ which solves $\tau = \nu_G(\tau)$ with*

$$\min \{ |S_n(\bar{T}_\infty^n)|, |S_n(\bar{T}_\infty^n) - \tau| \} \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \rightarrow \infty. \quad (4.6)$$

Proof. Firstly, $m_n = m$ for all sufficiently large n , implies that $\bar{T}_0^n \xrightarrow{\text{a.s.}} 0$ as $n \rightarrow \infty$. By the strong law of large numbers, $(nh)^{-1} \sum_{i=1}^n G_i(t) \xrightarrow{\text{a.s.}} \nu_G(t)$ as $n \rightarrow \infty$, for all $t \geq 0$.

Also, $\nu_G(\infty) = \mu_G < \infty$. A similar, but simpler, argument to the proof of [7], Lemma 3.8, yields

$$\sup_{t \geq 0} \left| \frac{1}{nh} \sum_{i=1}^n G_i(t) - \nu_G(t) \right| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \rightarrow \infty. \quad (4.7)$$

By a similar argument, for any $T > 0$, we have that

$$\sup_{0 \leq t \leq 2T} \left| \frac{1}{nh} \sum_{i=1}^n Y_i(t) - t \right| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \rightarrow \infty. \quad (4.8)$$

For any $x \geq 0$, using (4.4), we have that

$$\begin{aligned} |S_n(x) - x| &= \left| S_n(x) - \frac{1}{nh} \sum_{i=1}^n Y_i(S_n(x)) + \frac{\lfloor xnh \rfloor}{nh} - x \right| \\ &\leq \left| S_n(x) - \frac{1}{nh} \sum_{i=1}^n Y_i(S_n(x)) \right| + \left| \frac{\lfloor xnh \rfloor - xnh}{nh} \right|. \end{aligned}$$

Since $S_n(x)$ is increasing in x , it follows that for $S_n(T) \leq 2T$,

$$\begin{aligned} 0 \leq \sup_{0 \leq x \leq T} |S_n(x) - x| &\leq \sup_{0 \leq x \leq T} \left(\left| S_n(x) - \frac{1}{nh} \sum_{i=1}^n Y_i(S_n(x)) \right| + \left| \frac{\lfloor xnh \rfloor - xnh}{nh} \right| \right) \\ &\leq \sup_{0 \leq t \leq 2T} \left| \frac{1}{nh} \sum_{i=1}^n Y_i(t) - t \right| + \frac{1}{nh}. \end{aligned} \quad (4.9)$$

Given $(nh)^{-1} \sum_{i=1}^n Y_i(2T) \geq T$ implies that $S_n(T) \leq 2T$ and $(nh)^{-1} \sum_{i=1}^n Y_i(2T) \xrightarrow{\text{a.s.}} 2T$ as $n \rightarrow \infty$, it follows from (4.9) and (4.8) that

$$\sup_{0 \leq x \leq T} |S_n(x) - x| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \rightarrow \infty.$$

Let $\mathcal{K} = \{t \in [0, \infty] : t = \nu_G(t)\}$. Since $\nu_G(\cdot)$ is a strictly concave function of t , it follows that $\mathcal{K} = \{0, \tau\}$ for $R_* > 1$. Also $\nu'_G(\tau) \neq 1$ for all $\tau \in \mathcal{K}$. Let $(\Omega, \mathcal{F}, \mathcal{P})$ denote the probability space on which the random vectors $(R_1(t), G_1(t), Y_1(t)), (R_2(t), G_2(t), Y_2(t)), \dots$ are defined. Fix $T > \tau$ and let

$$\begin{aligned} F_1 &= \left\{ \omega \in \Omega : \sup_{t \geq 0} \left| \frac{1}{nh} \sum_{i=1}^n G_i(t, \omega) - \nu_G(t) \right| \rightarrow 0 \text{ as } n \rightarrow \infty \right\} \\ F_2 &= \left\{ \omega \in \Omega : \sup_{0 \leq x \leq T} \left| \frac{1}{nh} \sum_{i=1}^n S_n(x, \omega) - x \right| \rightarrow 0 \text{ as } n \rightarrow \infty \right\} \end{aligned}$$

and

$$F_3 = \left\{ \omega \in \Omega : \lim_{n \rightarrow \infty} \bar{T}_0^n(\omega) = 0 \right\}.$$

Then

$$\min \{ |S_n(\bar{T}_\infty^n, \omega) - \tau| : \tau \in \mathcal{K} \} \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

for all $\omega \in F_1 \cap F_2 \cap F_3$. The lemma follows since $\mathcal{P}(F_1 \cap F_2 \cap F_3) = 1$. $\square$

A corollary of Lemma 4.2 concerns the proportion infected in the epidemic. For $R_* > 1$, let $z = \tau/\mu_G$. Note that $z = \nu_R(\mu_G z)$, so z coincides with the definition at (2.3).

Corollary 4.1. *Suppose that there exists $m \in \mathbb{N}$ such that $m_n = m$ for all sufficiently large n . For $R_* > 1$, we have that*

$$\min \{|\bar{Z}_{n,h}|, |\bar{Z}_{n,h} - z|\} \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \rightarrow \infty. \quad (4.10)$$

Proof. An identical line of argument to the derivation of (4.7) gives

$$\sup_{t \geq 0} \left| \frac{1}{nh} \sum_{i=1}^n R_i(t) - \nu_R(t) \right| \xrightarrow{\text{a.s.}} 0 \quad \text{as } n \rightarrow \infty, \quad (4.11)$$

Then using Lemma 4.2, (4.6) and (4.11) it is straightforward to prove (4.10) along similar lines to the proof of Lemma 4.2. $\square$

The final step to prove that $\bar{Z}_{n,h} \xrightarrow{D} Z$, where Z has probability mass function given by (2.6), is to show that for any $0 < \epsilon < z$, $P(\bar{Z}_{n,h} < \epsilon) \rightarrow \rho^m$ as $n \rightarrow \infty$. Let $\bar{V}_{n,h} = V_{n,h}/n$. By construction we have that $\bar{V}_{n,h}/h \leq \bar{Z}_{n,h} \leq \bar{V}_{n,h}$ and therefore it suffices to show that there exists $\epsilon' > 0$,

$$P(\bar{V}_{n,h} \leq \epsilon') \rightarrow \rho^m \quad \text{as } n \rightarrow \infty. \quad (4.12)$$

It is straightforward using a lower bound branching process, cf. Whittle [20], [5], Ball and Lyne [4], to show that (4.12) holds by following a similar line of argument to the proof of Ball and Neal [10], Theorem 3.2. An outline of the argument is as follows. We couple $\tilde{\mathcal{E}}_n$ and $\mathcal{B}$ until $k_n = \lfloor \log n \rfloor$ households have been infected. The first k_n household epidemics will generate approximately $R_* k_n$ global infections. More precisely, we can show that for any $0 < \delta < R_* - 1$, the first $\lfloor \log n \rfloor$ household epidemics create at least a further $\lfloor \delta \log n \rfloor$ local epidemics in distinct households with probability tending to 1 as $n \rightarrow \infty$. For any $0 < \epsilon' < [R_* - 1]/R_*$, we consider a super-critical lower bound branching process approximation to the epidemic starting from $\lfloor \delta \log n \rfloor$ individuals where each birth in the branching process is aborted independently with probability ϵ' . Since the lower bound branching process is super-critical and $E[C^2] < \infty$, we have that the extinction probability, $\rho(\epsilon')$, from a single ancestor is bounded away from 1 by Ball and Neal [9], Lemma A3, with $\rho(\epsilon')^{\lfloor \delta \log n \rfloor} \rightarrow 0$ as $n \rightarrow \infty$. Whilst the proportion of households infected is less than ϵ' , the probability that a global contact is with a previously infected household is at most ϵ' . We can then use the lower bound branching process to show that

$$P(\bar{V}_{n,h} \leq \epsilon' | V_{n,h} > \lfloor \log n \rfloor) \rightarrow 0 \quad \text{as } n \rightarrow \infty,$$

which combined with (4.1) yields (4.12).

4.6 Proof of Theorem 2.1

We are now in position to prove (2.7) in Theorem 2.1. By conditioning on the event $\mathcal{G}^{n,h}$, that at least $\log n$ households are infected in the epidemic, it follows from Lemma 4.2, Corollary 4.1 and the discussion after Corollary 4.1 that

$$S_n(\bar{T}_\infty^n) | \mathcal{G}^{n,h} \xrightarrow{p} \tau \quad \text{and} \quad \bar{Z}_{n,h} | \mathcal{G}^{n,h} \xrightarrow{p} z (= \nu_R(\tau)) \quad \text{as } n \rightarrow \infty.$$

Throughout the remainder of the proof we implicitly condition on $\mathcal{G}^{n,h}$.

Considering $\bar{Z}_{n,h}|\mathcal{G}^{n,h}$ directly is not straightforward. However, we note that conditional upon $\mathcal{G}^{n,h}$ at least $k_n = \lfloor \log n \rfloor$ households are infected. This allows us to construct lower, $\bar{Z}_{n,h}^L$, and upper, $\bar{Z}_{n,h}^U$, bounds for the proportion infected in the event of a global epidemic by considering who becomes infected in the first k_n households and *restarting* the epidemic with k_n infected households and $n - k_n$ initially susceptible households. We show that $\bar{Z}_{n,h}^L \stackrel{\text{st}}{\leq} \bar{Z}_{n,h}|\mathcal{G}^{n,h} \stackrel{\text{st}}{\leq} \bar{Z}_{n,h}^U$ with $\sqrt{nh}(\bar{Z}_{n,h}^L - z) \xrightarrow{D} N(0, \sigma^2)$ and $\sqrt{nh}(\bar{Z}_{n,h}^U - z) \xrightarrow{D} N(0, \sigma^2)$ from which (2.7), and hence Theorem 2.1, follow.

In order to obtain suitable $\bar{Z}_{n,h}^L$ and $\bar{Z}_{n,h}^U$, we first define a sequence of epidemic processes $\hat{\mathcal{E}}_n(D_n)$, indexed by the number of households n , where $D_n \in \mathbb{N}$ denotes the number of global contacts from outside the population to initiate the epidemic. Suppose that in $\hat{\mathcal{E}}_n(D_n)$ there are initially $n - k_n$ totally susceptible households, with the remaining k_n households consisting entirely of removed individuals. We label the initially susceptible households $1, 2, \dots, n - k_n$ and the initially removed households $n + 1 - k_n, n + 2 - k_n, \dots, n$. The epidemic is constructed in a similar manner to $\mathcal{E}_n$ using $\{\mathbf{X}_{ij}, \mathbf{H}_{ij}, \eta_{ij}; i = 1, 2, \dots, n, j = 1, 2, \dots, h\}$ with D_n initial global contacts to determine the initial infectives within the population. However, throughout the epidemic global contacts with households $n + 1 - k_n, n + 2 - k_n, \dots, n$ have no effect as they are with removed individuals. Thus for the initially removed households only the η_{ij} s are required. The epidemic only effectively takes place between $n - k_n$ households with the k_n initially removed households included to absorb unsuccessful global infections when we relate $\hat{\mathcal{E}}_n(D_n)$ to $\mathcal{E}_n$. Due to the construction of $\hat{\mathcal{E}}_n(D_n)$, we can use the trivariate random vectors $(R_i(t), G_i(t), Y_i(t))$ for the embedding process. Let $\hat{T}_\infty^n(D_n)$ satisfy

$$\hat{T}_\infty^n(D_n) = \min \left\{ x \geq 0 : D_n + \sum_{i=1}^{n-k_n} G_i(S_n(x)) = \lfloor xnh \rfloor \left(= \sum_{i=1}^n Y_i(S_n(x)) \right) \right\}. \quad (4.13)$$

We note that the difference between (4.13) and (4.5) for $\bar{T}_\infty^n$ is the number of global infections to initiate the epidemic and that in $\hat{\mathcal{E}}_n(D_n)$ only $n - k_n$ households contribute to the generation of new global infections. By construction if $D_n^1 < D_n^2$ then $\hat{T}_\infty^n(D_n^1) \leq \hat{T}_\infty^n(D_n^2)$.

We have the following central limit theorem for the proportion infected in the epidemic $\hat{\mathcal{E}}_n(D_n)$ with the proof deferred to Section 4.7. Theorem 4.1 is central to proving (2.7). (We show in Section 4.8 that the expression for σ^2 given in (4.14) below is equivalent to that given in (2.4) in Section 2.4.)

Theorem 4.1. *Let D_n be a sequence of positive integers such that $D_n \rightarrow \infty$ and $D_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$. Let*

$$\hat{Z}_{n,h}(D_n) = \frac{1}{nh} \sum_{i=1}^{n-k_n} R_i(S_n(\hat{T}_\infty^n(D_n))),$$

the proportion of individuals who are infected during the epidemic $\hat{\mathcal{E}}_n(D_n)$. Then

$$\sqrt{nh} \left(\hat{Z}_{n,h}(D_n) - z \right) \xrightarrow{D} N(0, \sigma^2) \quad \text{as } n \rightarrow \infty,$$

where

$$\sigma^2 = \text{var} \left(R_1(\tau) + b(\tau) [G_1(\tau) - Y_1(\tau)] \right). \quad (4.14)$$

Conditional on $\mathcal{G}^{n,h}$, we can consider the first k_n households infected. Let $\tilde{D}_n^I$ denote the total number of global infectious contacts emanating from the first k_n local household epidemics plus the initial T_0^n global infectious contacts required to create the m_n initial infectives. Then $\tilde{D}_n^I/k_n \xrightarrow{p} R_*$ as $n \rightarrow \infty$, where $R_* = E[C]$ is the mean number of global contacts emanating from a local epidemic initiated by a single infective in an otherwise susceptible household. Let $\tilde{D}_n^B$ denote the number of global infections required to infect k_n distinct households and note that using the birthday problem $P(\tilde{D}_n^B = k_n) \rightarrow 1$ as $n \rightarrow \infty$, cf. 4.2. Let $\tilde{D}_n^A = \tilde{D}_n^I - \tilde{D}_n^B$ the number of excess global contacts between the number of global contacts required to infect the first k_n households and the number of global infectious contacts generated by these first k_n household epidemics. Then $\tilde{D}_n^A/k_n \xrightarrow{p} R_* - 1 > 0$ as $n \rightarrow \infty$. Let $\tilde{D}_n^C$ denote the sum of all the global contacts from individuals in the first k_n infected households whether or not they are infected in the initial local epidemic in the household plus the initial T_0^n global infectious contacts. Note that $\tilde{D}_n^C \stackrel{D}{=} T_0^n + \sum_{i=1}^{k_n} \sum_{j=1}^h X_{G,(i,j)}$ with $E[\tilde{D}_n^C] = E[T_0^n] + hk_n\mu_G$. Let $\tilde{D}_n^D = \tilde{D}_n^C - \tilde{D}_n^B$, the number of excess global contacts between the number of global contacts required to infect the first k_n households and the total number of potential global infectious contacts generated by these first k_n households should everybody become infected.

We create a lower bounding epidemic process $\bar{\mathcal{E}}_n^L$ by using the same construction as $\mathcal{E}_n$ except that in the first k_n households to be infected only the first global contact is successful. All subsequent global infectious contacts with these k_n households, which we denote by $\mathcal{F}_n$, are unsuccessful. For households in $\mathcal{F}_n^C$, the epidemic progresses as in $\mathcal{E}_n$. Let $\bar{Z}_{n,h}^L$ denote the proportion of the population infected in $\bar{\mathcal{E}}_n^L$, then $\bar{Z}_{n,h}^L \leq \bar{Z}_{n,h}$. Similarly we create an upper bounding epidemic process $\bar{\mathcal{E}}_n^U$ by using the same construction as $\mathcal{E}_n$ except that in the first k_n households all individuals are made infectious. All subsequent global infectious contacts with these k_n households have no effect, as the individual contacted has already been infected. For households in $\mathcal{F}_n^C$, the epidemic again progresses as in $\mathcal{E}_n$. Let $\bar{Z}_{n,h}^U$ denote the proportion of the population infected in $\bar{\mathcal{E}}_n^U$, then $\bar{Z}_{n,h}^U \geq \bar{Z}_{n,h}$. Let $\bar{Z}_{n,h}^L = \bar{Z}_{n,h}^{L,0} + \bar{Z}_{n,h}^{L,1}$, where $\bar{Z}_{n,h}^{L,0}$ is the proportion of the population who both belong to $\mathcal{F}_n$ and are infected in $\bar{\mathcal{E}}_n^L$ and $\bar{Z}_{n,h}^{L,1}$ is the proportion of the population who both belong to $\mathcal{F}_n^C$ and are infected in $\bar{\mathcal{E}}_n^L$. Define $\bar{Z}_{n,h}^{U,0}$ and $\bar{Z}_{n,h}^{U,1}$ similarly, with $\bar{Z}_{n,h}^U = \bar{Z}_{n,h}^{U,0} + \bar{Z}_{n,h}^{U,1}$.

By construction, the lower bounding and upper bounding epidemic processes behave as if the households in $\mathcal{F}_n$ are removed after considering the first k_n households and $\tilde{D}_n^B$ global infections. Given that Poisson processes have independent increments and $\mathcal{G}^{n,h}$ with $\tilde{D}_n^A = D_n^1$, we can couple the construction of $\bar{\mathcal{E}}_n^L$ to a realisation of $\hat{\mathcal{E}}_n(D_n^1)$ such that $\{\bar{Z}_{n,h}^{L,1} | \mathcal{G}^{n,h}, \tilde{D}_n^A = D_n^1\} = \hat{Z}_{n,h}(D_n^1)$. Similarly, given that $\tilde{D}_n^D = D_n^2$, we can couple the construction of $\bar{\mathcal{E}}_n^U$ to a realisation of $\hat{\mathcal{E}}_n(D_n^2)$ such that $\{\bar{Z}_{n,h}^{U,1} | \mathcal{G}^{n,h}, \tilde{D}_n^D = D_n^2\} = \hat{Z}_{n,h}(D_n^2)$.

Let $D_n^L = \lfloor (R_* - 1)k_n/2 \rfloor$ and $D_n^U = \lfloor 2hk_n\mu_G \rfloor$. Then $P(D_n^L \leq \tilde{D}_n^A) \rightarrow 1$ and $P(D_n^U \geq \tilde{D}_n^D) \rightarrow 1$ as $n \rightarrow \infty$. Also since $D_n^L, D_n^U \rightarrow \infty$ and $D_n^L/\sqrt{n}, D_n^U/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, it follows from Theorem 4.1 that both $\sqrt{nh} \left(\hat{Z}_{n,h}(D_n^L) - z \right)$ and $\sqrt{nh} \left(\hat{Z}_{n,h}(D_n^U) - z \right)$ converge in distribution to $N(0, \sigma^2)$ as $n \rightarrow \infty$.

Let $\bar{Z}_{n,h}^0$ ($\bar{Z}_{n,h}^1$) denote the proportion of the population who both belong to $\mathcal{F}_n$ ($\mathcal{F}_n^C$) and are infected in $\mathcal{E}_n$. We have that if $D_n^L \leq \tilde{D}_n^A$ and $D_n^U \geq \tilde{D}_n^D$,

$$\hat{Z}_{n,h}(D_n^L) \leq \bar{Z}_{n,h}^{L,1} | \mathcal{G}^{n,h} \leq \bar{Z}_{n,h}^1 | \mathcal{G}^{n,h} \leq \bar{Z}_{n,h}^{U,1} | \mathcal{G}^{n,h} \leq \hat{Z}_{n,h}(D_n^U).$$

Given that $P(\{D_n^L \leq \tilde{D}_n^A\} \cup \{D_n^U \geq \tilde{D}_n^D\}) \rightarrow 1$ as $n \rightarrow \infty$, it follows that $\sqrt{nh}(\bar{Z}_{n,h}^1 - z) \xrightarrow{D} N(0, \sigma^2)$ as $n \rightarrow \infty$. Finally, (2.7) follows using Slutsky's theorem (see, for example, Billingsley [12], Theorem 3.1), since $\sqrt{nh}\bar{Z}_{n,h}^0 \xrightarrow{p} 0$ as $n \rightarrow \infty$.

4.7 Proof of Theorem 4.1

In order to prove Theorem 4.1, we show that $\sqrt{nh}(\hat{Z}_{nh}(D_n) - z)$ has the same limiting distribution, as $n \rightarrow \infty$, as the normalised sum of a certain linear combination of $\{(R_i(\tau), G_i(\tau), Y_i(\tau)); i = 1, 2, \dots, n\}$. This requires first defining for $T > 0$ a sequence of stochastic processes $\mathbf{W}_{[n,T]}$ and showing in Lemma 4.3 that the limiting stochastic process is a zero-mean trivariate Gaussian process.

For $J = R, G, Y$ and $t \geq 0$, let

$$W_n^J(t) = \frac{1}{\sqrt{nh}} \sum_{i=1}^{n_J} [J_i(t) - h\nu_J(t)], \quad (4.15)$$

where $\nu_R(t)$ is defined in (2.2), $\nu_G(t) = \mu_G \nu_R(t)$, $\nu_Y(t) = t$, $n_R = n_G = \hat{n} = n - k_n$ and $n_Y = n$. That is, for R and G we sum over the $n - k_n$ initially susceptible households and for Y we sum over all n households, since global contacts with the initially susceptible households are important. Let $\mathbf{W}_n(t) = (W_n^R(t), W_n^G(t), W_n^Y(t))$ and let, for $T > 0$,

$$\mathbf{W}_{[n,T]} = \{\mathbf{W}_n(t) : 0 \leq t \leq T\}. \quad (4.16)$$

Also for $T > 0$, let $\mathbf{W}_{[*T]} = (W^R, W^G, W^Y)$ be a zero-mean trivariate Gaussian process with, for $J, L \in \{R, G, Y\}$ and $0 \leq s, t \leq T$,

$$\text{cov}(W^J(s), W^L(t)) = \frac{1}{h} \text{cov}(J_1(s), L_1(t)).$$

Lemma 4.3. *For any $T \geq 0$,*

$$\mathbf{W}_{[n,T]} \xrightarrow{w} \mathbf{W}_{[*T]} \quad \text{as } n \rightarrow \infty$$

where $\xrightarrow{w}$ denotes weak convergence in the space of bounded functions from $[0, T]$ to $\mathbb{R}^3$ endowed with the supremum metric (see, van der Vaart and Wellner [19], page 34).

Proof. Fix $T > 0$. The lemma follows using [19], Theorem 1.5.4, by showing that the finite-dimensional distributions of $\mathbf{W}_{[n,T]}$ converge to those of $\mathbf{W}_{[*T]}$ and that the sequence $\mathbf{W}_{[n,T]}$ ($n = 1, 2, \dots$) is asymptotically tight.

For any $m \in \mathbb{N}$, $\mathbf{t} \in [0, T]^m$ and $\alpha_{Jk} \in \mathbb{R}$ ($J = R, G, Y; k = 1, 2, \dots, m$),

$$\begin{aligned} & \sum_{k=1}^m \{\alpha_{Rk} W_n^R(t_k) + \alpha_{Gk} W_n^G(t_k) + \alpha_{Yk} W_n^Y(t_k)\} \\ &= \frac{1}{\sqrt{nh}} \sum_{i=1}^{\hat{n}} Q_i(\boldsymbol{\alpha}, \mathbf{t}) + \sum_{k=1}^m \frac{1}{\sqrt{nh}} \sum_{i=\hat{n}+1}^n \alpha_{Yk} [Y_i(t_k) - h\nu_Y(t_k)], \end{aligned} \quad (4.17)$$

where, for $i = 1, 2, \dots$,

$$Q_i(\boldsymbol{\alpha}, \mathbf{t}) = \sum_{k=1}^m \{ \alpha_{Rk} [R_i(t_k) - h\nu_R(t_k)] + \alpha_{Gk} [G_i(t_k) - h\nu_G(t_k)] + \alpha_{Yk} [Y_i(t_k) - h\nu_Y(t_k)] \}.$$

The $\{Q_i(\boldsymbol{\alpha}, \mathbf{t})\}$ s are i.i.d. with $E[Q_1(\boldsymbol{\alpha}, \mathbf{t})] = 0$. Since, for any $t \geq 0$, $R_1(t) \leq h$, $G_1(t) \leq \sum_{j=1}^h C_{1j}^G$ and $Y_1(t) \sim \text{Po}(ht)$, it is straightforward to show that

$$E[Q_1(\boldsymbol{\alpha}, \mathbf{t})^2] < \infty.$$

Therefore, since $k_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, the central limit theorem yields

$$\frac{1}{\sqrt{nh}} \sum_{i=1}^{\hat{n}} Q_i(\boldsymbol{\alpha}, \mathbf{t}) \xrightarrow{D} N\left(0, \frac{1}{h} \text{var}(Q_1(\boldsymbol{\alpha}, \mathbf{t}))\right) \quad \text{as } n \rightarrow \infty.$$

It is straightforward to show that the final term on the right-hand side of (4.17) converges in probability to 0 as $n \rightarrow \infty$, so using Slutsky's theorem,

$$\sum_{k=1}^m \{ \alpha_{Rk} W_n^R(t_k) + \alpha_{Gk} W_n^G(t_k) + \alpha_{Yk} W_n^Y(t_k) \} \xrightarrow{D} N\left(0, \frac{1}{h} \text{var}(Q_1(\boldsymbol{\alpha}, \mathbf{t}))\right) \quad \text{as } n \rightarrow \infty.$$

By considering linear combinations of $\mathbf{W}_n(t_k)$ and using the Cramér-Wold device, it follows that the finite-dimensional distributions of $\mathbf{W}_{[n,T]}$ converge to those of $\mathbf{W}_{[*],T}$.

By [19], Lemma 1.4.3, the sequence $\mathbf{W}_{[n,T]}$ ($n = 1, 2, \dots$) is asymptotically tight if and only if each of the sequences $W_{[n,T]}^J$ ($J = R, G, Y; n = 1, 2, \dots$) is asymptotically tight. We start by showing that the sequence $W_{[n,T]}^G$ ($n = 1, 2, \dots$) is asymptotically tight.

For $t \geq 0$, let $\bar{G}_1(t) = G_1(t) - h\nu_G(t)$, with $W_n^G(t) = (nh)^{-1/2} \sum_{i=1}^{n_G} \bar{G}_i(t)$. Since $\bar{G}_1(\cdot), \bar{G}_2(\cdot), \dots$ are i.i.d., the 3 conditions which are given for asymptotic tightness of $W_{[n,T]}^G$ ($n = 1, 2, \dots$) in [19], Theorem 2.11.9 simplify to showing that as $n \rightarrow \infty$:

(i) For every $\xi > 0$,

$$n_G E \left[\left\| \frac{\bar{G}_1}{\sqrt{nh}} \right\|_T 1_{\{\|\bar{G}_1/\sqrt{nh}\|_T > \xi\}} \right] \rightarrow 0,$$

where $\|f\|_T = \sup_{0 \leq t \leq T} |f(t)|$.

(ii) For every $\delta_n \downarrow 0$,

$$\sup_{|s-t| < \delta_n} \frac{n_G}{nh} E \left[(\bar{G}_1(s) - \bar{G}_1(t))^2 \right] \rightarrow 0.$$

(iii) For every $\delta_n \downarrow 0$,

$$\int_0^{\delta_n} \sqrt{\log N_{[]}^n(\epsilon, T)} d\epsilon \rightarrow 0,$$

where for $\epsilon > 0$, the bracketing number $N_{[]}^n(\epsilon, T)$ is defined to be the minimum number of sets N_ϵ in a partition $[0, T] = \cup_{j=1}^{N_\epsilon} \mathcal{A}_{\epsilon j}^n$ such that, for each $\mathcal{A}_{\epsilon j}^n$, we have

$$\frac{n_G}{nh} E \left[\sup_{s, t \in \mathcal{A}_{\epsilon j}^n} (\bar{G}_1(t) - \bar{G}_1(s))^2 \right] \leq \epsilon^2. \quad (4.18)$$

Given that, for all $t \geq 0$, $|\bar{G}_1(t)| \leq \sum_{j=1}^h X_{G,(1,j)} + h\mu_G = Q^G$, say, it follows that, for any $\xi > 0$,

$$n_G \mathbb{E} \left[\left\| \frac{\bar{G}_1}{\sqrt{nh}} \right\|_T 1_{\{\|\bar{G}_1/\sqrt{nh}\|_T > \xi\}} \right] \leq \sqrt{\frac{n}{h}} \mathbb{E} \left[Q^G 1_{\{Q^G > \sqrt{nh}\xi\}} \right]. \quad (4.19)$$

The same argument as a proof of Markov's inequality yields, for $a > 0$,

$$\mathbb{E} \left[Q^G 1_{\{Q^G > \sqrt{nh}\xi\}} \right] = (\sqrt{nh}\xi)^{-(1+a)} \mathbb{E} \left[Q^G 1_{\{Q^G > \sqrt{nh}\xi\}} (\sqrt{nh}\xi)^{(1+a)} \right] \leq \frac{\mathbb{E} [(Q^G)^{2+a}]}{(\xi\sqrt{nh})^{1+a}}. \quad (4.20)$$

Since $\mathbb{E}[X_G^{2+a}] < \infty$ implies that $\mathbb{E}[(Q^G)^{2+a}] < \infty$, it is straightforward to show condition (i) holds using (4.19) and (4.20).

Given that $G_1(\cdot)$ and $\nu_G(\cdot)$ are non-decreasing in t , it is straightforward to show that for any $u \leq s \leq t \leq v$,

$$[\bar{G}_1(t) - \bar{G}_1(s)]^2 \leq [G_1(t) - G_1(s)]^2 + h^2 [\nu_G(t) - \nu_G(s)]^2 \quad (4.21)$$

$$\leq [G_1(v) - G_1(u)]^2 + h^2 [\nu_G(v) - \nu_G(u)]^2. \quad (4.22)$$

Also, jumps in $G_1(\cdot)$ only occur when a global infectious contacts are made with the household, so, for all $0 \leq s < t$,

$$|G_1(t) - G_1(s)| \leq \left(\sum_{j=1}^h X_{G,(1,j)} \right) 1_{\{Y_i(t) \neq Y_i(s)\}} \quad (4.23)$$

with $\sum_{j=1}^h X_{G,(1,j)}$ independent of $1_{\{Y_i(t) \neq Y_i(s)\}}$. It then follows from (4.21), (4.23) and

$$h^2(\nu_G(t) - \nu_G(s))^2 = h^2 \left(\frac{1}{h} \mathbb{E}[G_1(t) - G_1(s)] \right)^2 \leq \mathbb{E}[(G_1(t) - G_1(s))^2],$$

that for all $0 \leq s < t$,

$$\begin{aligned} \mathbb{E} \left[(\bar{G}_1(t) - \bar{G}_1(s))^2 \right] &\leq 2\mathbb{E} [(G_1(t) - G_1(s))^2] \\ &\leq 2\mathbb{E} \left[\left(\sum_{j=1}^h X_{G,(1,j)} \right)^2 \right] \mathbb{E} [1_{\{Y_1(t) \neq Y_1(s)\}}] \\ &\leq 2\mathbb{E} \left[\left(\sum_{j=1}^h X_{G,(1,j)} \right)^2 \right] \mathbb{E} [Y_1(t) - Y_1(s)] \\ &\leq 2h^2 \mathbb{E} [X_G^2] h(t - s). \end{aligned} \quad (4.24)$$

Condition (ii) follows since for all $s \geq 0$, the right hand side of (4.24) converges to 0 as $t \downarrow s$.

Fix $\epsilon > 0$ and $\mathcal{A} = [u, v]$, where $0 \leq u < v$ such that $|u - v| \leq \epsilon^2 / (4h^2 \mathbb{E}[X_G^2])$. It follows from (4.22) and (4.24) that

$$\begin{aligned} \frac{n_G}{nh} \mathbb{E} \left[\sup_{s, t \in \mathcal{A}} (\bar{G}_1(t) - \bar{G}_1(s))^2 \right] &\leq \frac{1}{h} \mathbb{E} [(G_1(v) - G_1(u))^2] + h [\nu_G(v) - \nu_G(u)]^2 \\ &\leq \frac{2}{h} \mathbb{E} [(G_1(v) - G_1(u))^2] \\ &\leq \frac{2}{h} \times 2h^3 \mathbb{E}[X_G^2] \times \frac{\epsilon^2}{4h^2 \mathbb{E}[X_G^2]} = \epsilon^2. \end{aligned}$$

Therefore a partition of $[0, T]$ into intervals $\mathcal{A}_{ej}^n$ of length $L_\epsilon = \epsilon^2 / (4h^2 \mathbb{E}[X_G^2])$ exists such that (4.18) holds. Hence, $N_{\square}^n(\epsilon, T) \leq c/\epsilon^2$, where $c = 1 + 4Th^2 \mathbb{E}[X_G^2]$. Then,

$$\begin{aligned} \int_0^{\delta_n} \sqrt{\log N_{\square}^n(\epsilon, T)} d\epsilon &\leq \int_0^{\delta_n} \sqrt{\log \left(\frac{c}{\epsilon^2} \right)} d\epsilon \\ &= \frac{\sqrt{c}}{2} \int_{\log(c/\delta_n^2)}^{\infty} \sqrt{u} \exp\left(-\frac{u}{2}\right) du \rightarrow 0 \quad \text{as } n \rightarrow \infty. \end{aligned}$$

Hence condition (iii) is satisfied, concluding the proof of asymptotic tightness of $W_{[n, T]}^G$ ($n = 1, 2, \dots$).

The asymptotic tightness of $W_{[n, T]}^R$ ($n = 1, 2, \dots$) follows by an identical argument with $X_G \equiv 1$. Finally, using properties of Poisson processes, it is straightforward to show that conditions (i)-(iii) hold with $\bar{G}_1$ replaced by $\bar{Y}_1$, where $\bar{Y}_1(t) = Y_1(t) - h\nu_Y(t)$ and $n_G = n - k_n$ replaced by $n_Y = n$. Therefore, the sequence $\mathbf{W}_{[n, T]}$ ($n = 1, 2, \dots$) is asymptotically tight and the lemma follows. $\square$

Proof of Theorem 4.1. Using similar arguments to Section 4.5, we have that if $D_n \rightarrow \infty$ and $D_n/\sqrt{n} \rightarrow 0$ as $n \rightarrow \infty$, then

$$S_n(\hat{T}_\infty^n(D_n)) \xrightarrow{p} \tau, \quad \text{as } n \rightarrow \infty. \quad (4.25)$$

This is because the probability that the epidemic fails to take-off from D_n initial global contacts, of which $\text{Bin}(D_n, (n - k_n)/n)$ are with initially susceptible households, tends to 0 as $n \rightarrow \infty$. Therefore, using similar arguments to Corollary 4.1, we have that $\hat{Z}_{n, h}(D_n) \xrightarrow{p} z$ as $n \rightarrow \infty$.

Using the mean value theorem, we have that

$$\begin{aligned} \sqrt{nh} \left(\hat{Z}_{n, h}(D_n) - z \right) &= \sqrt{nh} \left[\frac{1}{nh} \sum_{i=1}^{\hat{n}} R_i(S_n(\hat{T}_\infty^n(D_n))) - \nu_R(\tau) \right] \\ &= \sqrt{nh} \left[\frac{1}{nh} \sum_{i=1}^{\hat{n}} R_i(S_n(\hat{T}_\infty^n(D_n))) - \nu_R(S_n(\hat{T}_\infty^n(D_n))) + \nu_R(S_n(\hat{T}_\infty^n(D_n))) - \nu_R(\tau) \right] \\ &= W_n^R(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n + \nu_R'(a_{n1}) \sqrt{nh} [S_n(\hat{T}_\infty^n(D_n)) - \hat{T}_\infty^n(D_n) + \hat{T}_\infty^n(D_n) - \tau], \end{aligned} \quad (4.26)$$

where a_{n1} lies between $S_n(\hat{T}_\infty^n(D_n))$ and τ and $\tilde{a}_n = \sqrt{nh}[\hat{n} - n]\nu_R(S_n(\hat{T}_\infty^n(D_n)))/n \xrightarrow{p} 0$ as $n \rightarrow \infty$. By definition, see (4.13),

$$\frac{1}{nh} \sum_{i=1}^n Y_i(S_n(\hat{T}_\infty^n(D_n))) = \hat{T}_\infty^n(D_n) \quad \text{and} \quad \nu_Y(S_n(\hat{T}_\infty^n(D_n))) = S_n(\hat{T}_\infty^n(D_n)).$$

Therefore, we can rewrite (4.26) as

$$\begin{aligned} \sqrt{nh} \left(\hat{Z}_{n,h}(D_n) - z \right) &= W_n^R(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n - \nu'_R(a_{n1})W_n^Y(S_n(\hat{T}_\infty^n(D_n))) \\ &\quad + \nu'_R(a_{n1})\sqrt{nh}[\hat{T}_\infty^n(D_n) - \tau]. \end{aligned} \quad (4.27)$$

Hence, we need to consider the distribution of $\sqrt{nh}[\hat{T}_\infty^n(D_n) - \tau]$.

Let $\hat{D}_n = D_n/(nh)$ and note that

$$\begin{aligned} \sqrt{nh}[\hat{T}_\infty^n(D_n) - \tau] &= \sqrt{nh} \left[\hat{D}_n + \frac{1}{nh} \sum_{i=1}^{\hat{n}} G_i(S_n(\hat{T}_\infty^n(D_n))) - \nu_G(\tau) \right] \\ &= \sqrt{nh}\hat{D}_n + W_n^G(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n\mu_G \\ &\quad + \sqrt{nh}[\nu_G(S_n(\hat{T}_\infty^n(D_n))) - \nu_G(\hat{T}_\infty^n(D_n)) + \nu_G(\hat{T}_\infty^n(D_n)) - \nu_G(\tau)]. \end{aligned}$$

By the mean value theorem, there exists a_{n2} lying between $S_n(\hat{T}_\infty^n(D_n))$ and τ , such that

$$\begin{aligned} \sqrt{nh}[\hat{T}_\infty^n(D_n) - \tau] &= \sqrt{nh}\hat{D}_n + W_n^G(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n\mu_G \\ &\quad + \sqrt{nh}\nu'_G(a_{n2})[S_n(\hat{T}_\infty^n(D_n)) - \hat{T}_\infty^n(D_n) + \hat{T}_\infty^n(D_n) - \tau] \\ &= \sqrt{nh}\hat{D}_n + W_n^G(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n\mu_G - \nu'_G(a_{n2})W_n^Y(S_n(\hat{T}_\infty^n(D_n))) \\ &\quad + \nu'_G(a_{n2})\sqrt{nh}[\hat{T}_\infty^n(D_n) - \tau], \end{aligned}$$

using (4.13). Hence,

$$\begin{aligned} \sqrt{nh}[\hat{T}_\infty^n(D_n) - \tau] &= \frac{1}{1 - \nu'_G(a_{n2})} \left\{ \sqrt{nh}\hat{D}_n + W_n^G(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n\mu_G - \nu'_G(a_{n2})W_n^Y(S_n(\hat{T}_\infty^n(D_n))) \right\}. \end{aligned} \quad (4.28)$$

Inserting (4.28) into (4.27), we obtain that

$$\begin{aligned} \sqrt{nh} \left(\hat{Z}_{n,h}(D_n) - z \right) &= W_n^R(S_n(\hat{T}_\infty^n(D_n))) + \tilde{a}_n + \frac{\sqrt{nh}\hat{D}_n + \tilde{a}_n\mu_G}{[1 - \nu'_G(a_{n2})]} \\ &\quad + \left[\frac{\nu'_R(a_{n1})}{1 - \nu'_G(a_{n2})} \right] \left[W_n^G(S_n(\hat{T}_\infty^n(D_n))) - W_n^Y(S_n(\hat{T}_\infty^n(D_n))) \right]. \end{aligned} \quad (4.29)$$

Using (4.25), we have by the sandwich theorem that $a_{n1}, a_{n2} \xrightarrow{p} \tau$ as $n \rightarrow \infty$. Therefore, $\nu'_G(a_{n2}) \xrightarrow{p} \nu'_G(\tau) < 1$ as $n \rightarrow \infty$. Given that $\sqrt{nh}\hat{D}_n \rightarrow 0$ and $\tilde{a}_n \xrightarrow{p} 0$ as $n \rightarrow \infty$, the

second and third terms on the right-hand side of (4.29) converge in probability to 0 as $n \rightarrow \infty$. Also, we have that

$$\frac{\nu'_R(a_{n1})}{1 - \nu'_G(a_{n2})} \xrightarrow{p} \frac{\nu'_R(\tau)}{1 - \nu'_G(\tau)} \quad \text{as } n \rightarrow \infty.$$

It follows by Slutsky's theorem that $\sqrt{nh} \left(\hat{Z}_{n,h}(D_n) - z \right)$ and

$$W_n^R(S_n(\hat{T}_\infty^n(D_n))) + b(\tau) \left[W_n^G(S_n(\hat{T}_\infty^n(D_n))) - W_n^Y(S_n(\hat{T}_\infty^n(D_n))) \right]$$

have the same limiting distribution, should one exist, as $n \rightarrow \infty$. By Slutsky's lemma and the continuous mapping theorem, [19], Example 1.4.7 and Theorem 1.3.6, respectively, it follows from Lemma 4.3 and (4.25) that

$$\mathbf{W}_n(S_n(\hat{T}_\infty^n(D_n))) \xrightarrow{D} \mathbf{W}(\tau) \quad \text{as } n \rightarrow \infty.$$

Hence,

$$\begin{aligned} & W_n^R(S_n(\hat{T}_\infty^n(D_n))) + b(\tau) \left[W_n^G(S_n(\hat{T}_\infty^n(D_n))) - W_n^Y(S_n(\hat{T}_\infty^n(D_n))) \right] \\ & \xrightarrow{D} W^R(\tau) + b(\tau) [W^G(\tau) - W^Y(\tau)] \quad \text{as } n \rightarrow \infty, \end{aligned}$$

and the theorem follows since

$$\sigma^2 = \text{var} \left(W^R(\tau) + b(\tau) [W^G(\tau) - W^Y(\tau)] \right) = \frac{1}{h} \text{var} \left(R_1(\tau) + b(\tau) [G_1(\tau) - Y_1(\tau)] \right).$$

□

4.8 Variance calculations

In this section we discuss σ^2 and present an alternative representation of the variance. The variance σ^2 satisfies

$$\sigma^2 = \frac{1}{h} \text{var} \left(R_1(\tau) + b(\tau) [G_1(\tau) - Y_1(\tau)] \right) \quad (4.30)$$

$$\begin{aligned} & = (1 + b(\tau)\mu_G)^2 \nu_R(\tau) [1 - \nu_R(\tau)] + (h - 1)(1 + b(\tau)\mu_G)^2 \text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) \\ & \quad + b(\tau)^2 \nu_R(\tau) [\sigma_G^2 - \mu_G] + 2(h - 1)b(\tau)(1 + \mu_G b(\tau)) \text{cov}(\chi_{11}(\tau), X_{G,(1,2)}), \end{aligned} \quad (4.31)$$

where $b(t) = \nu'_R(t)/[1 - \mu_G \nu'_R(t)]$. Since $\nu^G(\cdot)$ is a concave function, we have that $\nu^G(\tau) < 1$ giving $b(\tau) < \infty$.

We make the following observations regarding σ^2 and defer showing that σ^2 satisfies (4.31) to Appendix D.

1. The expression for σ^2 in (4.31) involves simply the relationship by individuals (1, 1) and (1, 2). Note that if the number of global and local contacts made by individuals are independent then $\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) = 0$.

2. In the case $X_L \equiv 0$ (no local infection) we obtain a homogeneously mixing model with $P(S_{11} = 1) = 1$, giving $\nu_R(t) = 1 - e^{-t}$ and $\nu_R(t) + \nu'_R(t) - 1 = 0$. Therefore for $X_L \equiv 0$, letting $\xi = \exp(-\tau)[= \nu'_R(\tau)]$, we have that $\nu_R(\tau) = 1 - \xi$, $b(\tau) = \xi/[1 - \mu_G\xi]$, $\tau = \mu_G(1 - \xi)$ and $1 + b(\tau)\mu_G = [1 - \mu_G\xi]^{-1}$, so

$$\begin{aligned}\sigma^2 &= (1 + b(\tau)\mu_G)^2 \nu_R(\tau)[1 - \nu_R(\tau)] + b(\tau)^2 \nu_R(\tau)[\sigma_G^2 - \mu_G] \\ &= \frac{[1 - \xi]\xi}{(1 - \mu_G\xi)^2} + \frac{\xi^2}{(1 - \mu_G\xi)^2} [1 - \xi][\sigma_G^2 - \mu_G] \\ &= \frac{\xi(1 - \xi) + \xi^2(1 - \xi)[\sigma_G^2 - \mu_G]}{(1 - \mu_G\xi)^2}.\end{aligned}\tag{4.32}$$

The expression for σ^2 given in (4.32) agrees with the variance term given in [14], Theorem 1, for a constant number of initial infectives m . The model considered in [14] is the generalised Reed-Frost model, where infectious individuals make X_G contacts with distinct members of the population. As we note in Section 4.9 below the difference between sampling global contacts with and without replacement vanishes as $n \rightarrow \infty$.

3. The expression for σ^2 given at (2.4) in Section 2.4 is of course equivalent to (4.30) or (4.31) above, as is shown at the end of the proof of (4.31) in Appendix D.

4.9 Global and local contacts sampled without replacement

In this section, we briefly describe the minor modifications required for the central limit theorem to hold when the global and local contacts made by an infective are *without replacement* from the remainder of the population and household, respectively. As $n \rightarrow \infty$, the probability an infective makes either a global self-contact or multiple global contacts with a given individual converges to 0 provided that $\mu_G = E[X_G] < \infty$. Moreover, it is straightforward to show that the total number of global self-contacts and multiple global contacts made by individuals with the same individual, V_n say, satisfies

$$V_n \xrightarrow{D} V \sim \text{Po}\left(\frac{E[X_G(X_G + 1)]}{2}\right) \quad \text{as } n \rightarrow \infty,$$

provided that $E[X_G^2] < \infty$, which is the case as under the assumptions of Theorem 2.1, there exists $a > 0$ such that $E[X_G^{2+a}] < \infty$. The effect of V_n additional global contacts to replace global self- and multiple contacts is negligible and does not affect the law of large numbers and central limit theorem for final proportion infected by a major epidemic. A similar result holds if we preclude the possibility of an individual making global contacts with their own household.

Turning to local (household) infectious contacts, if X_L denotes the total number of distinct household contacts then X_L has support on $\{0, 1, \dots, h-1\}$. Consequently, $\mathbf{H}_{ij}$, the successive individuals contacted locally by individual (i, j) , is a random vector of length $h-1$, whose entries are a random permutation of $\{1, 2, \dots, h\} \setminus j$, with individual (i, j) making a household infectious contact with individual (i, l) if $l \in \{H_{ij1}, H_{ij2}, \dots, H_{ijX_{L,(i,j)}}\}$. The susceptibility set of individuals can then be constructed from $\{(X_{L,(i,k)}, \mathbf{H}_{ik}); k = 1, 2, \dots, h\}$ in a similar manner to Section 4.4 with the proof of the central limit theorem

continuing unchanged. The only change is the values taken by z and σ^2 , which change owing to the different distribution of S , the size of a susceptibility set (see Remark A.2 in Appendix A). Note that the probability of a minor outbreak, ρ , also changes, since C has a different distribution.

5 Proofs of Theorems 2.2 and 2.3

5.1 Proof of Theorem 2.2

In this section we prove that the probability of a major outbreak, $\pi^{(h,p)}$, is increasing in h and p for any random vector (X_G, X_L) . This is proved in two separate lemmas where we vary h (keeping p fixed) in Lemma 5.1 and vary p (keeping h fixed) in Lemma 5.2. Lemma 5.2 is proved under weaker assumptions on $\mathbf{X} = (X_G, X_L)$ and the independent replacement of local contacts by global contacts.

We show first that $\pi^{(h,p)}$ is increasing in h . We assume without loss of generality that $p = 0$. Recall the single-household epidemic model from $\mathcal{E}_h^H(X_G, X_L)$ from Section 2.4. Let $R^{(h)}$ be the size of that epidemic, including the initial infective, and $C^{(h)}$ be the number of global contacts that emanate from infectives in that epidemic. Thus $C^{(h)}$ is the offspring random variable for the branching process, $\mathcal{B}^{(h)}$, which approximates the of the epidemic $\mathcal{E}_{n,h}(X_G, X_L)$. Let $\rho^{(h)}$ denote the extinction probability of $\mathcal{B}^{(h)}$.

Lemma 5.1. *For a given contacts random vector $\mathbf{X} = (X_G, X_L)$, $\rho^{(h)}$ is strictly decreasing in h .*

Proof. It is immediate that $C^{(1)} \stackrel{\text{st}}{\leq} C^{(2)}$ and hence that $\rho^{(1)} \geq \rho^{(2)}$. Let $(X_{G,k}, X_{L,k})$ ($k = 1, 2, \dots$) be i.i.d. copies of (X_G, X_L) and U_k ($k = 1, 2, \dots$) be an independent sequence of independent $U(0, 1)$ random variables. We use these random variables to construct a realisation of $C^{(h)}$ for each $h = 2, 3, \dots$, as follows.

Fix $h \geq 2$. We determine $(R^{(h)}, C^{(h)})$ by considering the infectives in $\mathcal{E}_h^H(X_G, X_L)$ one at a time. We use $X_{L,1}$ to determine the number of distinct local contacts, $Z_1^{(h)}$, made by the initial infective. Precise details are given below. If $Z_1^{(h)} = 0$ the epidemic stops and $(R^{(h)}, C^{(h)}) = (1, X_{G,1})$. Otherwise, we take one of $Z_1^{(h)}$ newly infected individuals and use $X_{L,2}$ to determine the number of distinct contacts it makes with the remaining $h - 1 - Z_1^{(h)}$ susceptibles. We continue the process in the obvious fashion, stopping when we have run out of infectives to consider. Let $W_0^{(h)} = 1$ and $W_k^{(h)} = 1 + Z_1^{(h)} + Z_2^{(h)} + \dots + Z_k^{(h)}$ ($k = 1, 2, \dots$). Then $R^{(h)} = \min\{k \geq 1 : W_k^{(h)} - k = 0\}$ and $C^{(h)} = \sum_{i=1}^{R^{(h)}} X_{G,i}$. For completeness we define $W_k^{(h)} = W_{R^{(h)}}^{(h)}$ for $k > R^{(h)}$.

To determine whether local contacts are with susceptibles, we treat the local contacts one at a time. Suppose that just prior to the l^{th} local contact a total of $Y_l^{(h)}$ individuals have been infected (including the initial infective). Then that local contact is with a susceptible if and only if $U_l \leq (h - Y_l^{(h)})/(h - 1)$; otherwise the contact is with a non-susceptible individual and does not result in a new infective. Since $(h - y)/(h - 1) \leq (h + 1 - y)/h$ for $y = 1, 2, \dots$, it follows immediately from the construction that $W_k^{(h+1)} \geq W_k^{(h)}$ for $k = 0, 1, \dots$, whence $R^{(h+1)} \geq R^{(h)}$ and $C^{(h+1)} \geq C^{(h)}$. Thus, $C^{(h)} \stackrel{\text{st}}{\leq} C^{(h+1)}$ and $\rho^{(h)} \geq \rho^{(h+1)}$.

The inequality $(h-y)/(h-1) \leq (h+1-y)/h$ is strict for $y > 1$, so $P(C^{(h+1)} > C^{(h)}) = 1$, whence $\rho^{(h)} > \rho^{(h+1)}$. $\square$

Consider a random vector $\mathbf{X} = (X_G, X_L)$ for the number of global and household contacts made by a typical individual. For $0 \leq p \leq 1$, let $\mathcal{B}(p)$ denote the branching process where a proportion p of household contacts are converted to global contacts. Throughout, the branching process approximation is based on the assumption that each global infectious contact (birth) is with a previously uninfected household. Therefore, each infected household is infected globally once. The local epidemic (within the household) is determined by the number of local contacts, distributed independently according to X_L , and p , the proportion of household contacts that are converted to global contacts. We allow for a general rule for the transferring of household to global contacts. Let $Y_T^{(p)}$ denote the number of household contacts transferred to global contacts, so that in $\mathcal{B}(p)$ the number of global and household contacts made by a typical infective are distributed according to $(X_G + Y_T^{(p)}, X_L - Y_T^{(p)})$. Note that $Y_T^{(0)} = 0$ and $Y_T^{(1)} = X_L$, and if $Y_T^{(p)} = Y_L^{(p)} \sim \text{MixBin}(X_L, p)$ we are in the scenario described in Section 2.3. For $0 \leq p < q \leq 1$, we assume that $Y_T^{(p)}|X_L \stackrel{\text{st}}{\leq} Y_T^{(q)}|X_L$, that is, a coupling exists such that at least as many household contacts are transferred for an individual in $\mathcal{B}(q)$ as for the corresponding individual in $\mathcal{B}(p)$.

Lemma 5.2. *For a given household size h and contacts random vector $\mathbf{X} = (X_G, X_L)$, the extinction probability, ρ_p , of the branching process $\mathcal{B}(p)$ is monotonically decreasing in p .*

Proof. Fix $0 \leq p < q \leq 1$. We prove the lemma by showing that $\rho_q \leq \rho_p$.

Construct $\mathcal{B}(p)$ and $\mathcal{B}(q)$ on a common probability space such that the i^{th} individual in both process makes $X_{G,i} + X_{L,i}$ attempted births and $Y_{T,i}^{(p)} \leq Y_{T,i}^{(q)}$. We construct a lower bound branching process $\hat{\mathcal{B}}(p, q)$ in which the i^{th} individual makes $X_{G,i} + Y_{T,i}^{(p)}$ global and $X_{L,i} - Y_{T,i}^{(q)}$ local contacts. Note that in $\hat{\mathcal{B}}(p, q)$ the i^{th} individual has $Y_{T,i}^{(q)} - Y_{T,i}^{(p)}$ fewer contacts than its counterparts in $\mathcal{B}(p)$ and $\mathcal{B}(q)$ and we term these missing contacts, *ghost* contacts. Let ρ_p, ρ_q and $\hat{\rho}_{p,q}$ denote the extinction probabilities in the branching processes $\mathcal{B}(p), \mathcal{B}(q)$ and $\hat{\mathcal{B}}(p, q)$, respectively. Let $(\hat{V}_{p,q}, \hat{W}_{p,q})$ denote the number of global and ghost contacts emanating from a typical infectious individual in $\hat{\mathcal{B}}(p, q)$. Then

$$\hat{\rho}_{p,q} = \mathbb{E} \left[\hat{\rho}_{p,q}^{\hat{V}_{p,q}} \right].$$

We define

$$\hat{f}_{p,q}(\theta, s) = \mathbb{E} \left[\theta^{\hat{V}_{p,q}} s^{\hat{W}_{p,q}} \right],$$

the joint pgf of $(\hat{V}_{p,q}, \hat{W}_{p,q})$, so $\hat{\rho}_{p,q}$ solves

$$\theta = \hat{f}_{p,q}(\theta, 1).$$

Also we have that ρ_q solves

$$\theta = \hat{f}_{p,q}(\theta, \theta) = \mathbb{E} \left[\theta^{\hat{V}_{p,q} + \hat{W}_{p,q}} \right],$$

since all ghost contacts in $\hat{\mathcal{B}}(p, q)$ correspond to global contacts in $\mathcal{B}(q)$.

The ghost contacts in $\hat{\mathcal{B}}(p, q)$ correspond to local contacts within the household in $\mathcal{B}(p)$. The additional $\hat{W}_{p,q}$ local contacts in $\mathcal{B}(p)$ will result in $\tilde{W}_{p,q} \leq \hat{W}_{p,q}$ additional infectives from whom to grow the epidemic. It is likely that $\tilde{W}_{p,q} < \hat{W}_{p,q}$ as some contacts could be with individuals who are already members of the local household epidemic and/or repeat contacts with a new individual. Let $P(\hat{V}_{p,q}, \hat{W}_{p,q})$ denote the probability that the branching process goes extinct from those individuals infected by the additional $\hat{W}_{p,q}$ local contacts. Thus, ρ_p solves

$$\theta = \mathbb{E} \left[\theta^{\hat{V}_{p,q}} P(\hat{V}_{p,q}, \hat{W}_{p,q}) \right].$$

The $\tilde{W}_{p,q}$ individuals will initiate a local epidemic in a household with at least one removed individual (the initial infective). The number of global infections emanating from the local epidemic from the $\tilde{W}_{p,q}$ is stochastically smaller than $\sum_{i=1}^{\tilde{W}_{p,q}} \tilde{V}_{p,i}$, where the $\tilde{V}_{p,i}$ s are i.i.d. copies of $\tilde{V}_p$, the number of global contacts emanating from a household where individuals have i.i.d. contacts according to $(X_{G,i} + Y_{T,i}^{(p)}, X_{L,i} - Y_{T,i}^{(p)})$ and households initially have 1 infective, 1 removed and $h - 2$ susceptibles. Let $\tilde{\rho}_p$ solve

$$\theta = \mathbb{E} \left[\theta^{\tilde{V}_p} \right],$$

the extinction probability of a branching process where the offspring distribution is $\tilde{V}_p$. Then $\tilde{V}_p \stackrel{\text{st}}{\leq} V_p$, where V_p is the number of global contacts emanating from a household where individuals have i.i.d contacts according to $(X_{G,i} + Y_{T,i}^{(p)}, X_{L,i} - Y_{T,i}^{(p)})$ and households initially have 1 infective and $h - 1$ susceptibles. Thus, $\tilde{\rho}_p \geq \rho_p$ and for $0 \leq \theta \leq 1$,

$$\mathbb{E} \left[\theta^{\hat{V}_{p,q}} P(\hat{V}_{p,q}, \hat{W}_{p,q}) \right] \geq \mathbb{E} \left[\theta^{\hat{V}_{p,q}} \tilde{\rho}_p^{\hat{W}_{p,q}} \right] = \hat{f}_{p,q}(\theta, \tilde{\rho}_p). \quad (5.1)$$

Let ρ_* solve

$$\rho_* = \hat{f}_{p,q}(\rho_*, \tilde{\rho}_p).$$

Then by (5.1) it follows that $\rho_p \geq \rho_*$.

Since $\tilde{\rho}_p \geq \rho_*$, it follows that

$$\rho_* \geq \hat{f}_{p,q}(\rho_*, \rho_*). \quad (5.2)$$

Given that ρ_q is the smallest solution in $[0, 1]$ of $\theta = \hat{f}_{p,q}(\theta, \theta)$, an immediate consequence of (5.2) is that $\rho_* \geq \rho_q$, whence $\rho_p \geq \rho_q$, as required. $\square$

We observe that Lemma 5.2 holds if we assume instead that the, X_L , local contacts made by an individual are without replacement, with each local contact made by an individual being equally likely to be with anybody in their household they have not previously contacted.

5.2 Proof of Theorem 2.3

In this section we prove that the final size of a major outbreak, $z^{(h,p)}$, is increasing in h and p for any random vector (X_G, X_L) , for which the pgf of X_L is log-convex. As in Section 5.1 we prove the result in two separate lemmas where we vary h (keeping p fixed) in Lemma 5.3. and vary p (keeping h fixed) in Lemma 5.4.

We show first that $z^{(h,p)}$ is increasing in h . We assume without loss of generality that $p = 0$ and for ease of notation write $z^{(h,0)}$ as $z^{(h)}$.

Lemma 5.3. *For a given contact random vector $\mathbf{X} = (X_G, X_L)$, with $\log(f_{X_L}(s))$ being convex, the final size of a major outbreak, $z^{(h)}$, is strictly increasing in h .*

Proof. In the proof, we use the following way of sampling a $\text{Bin}(n, 1 - q)$ random variable. First sample $Z \sim \text{Po}(\lambda)$, where $\lambda = -n \log q$. Then place Z balls independently and uniformly at random into n boxes and let Y be the number of boxes that contain at least one ball. Then $Y \sim \text{Bin}(n, 1 - q)$. (The numbers of balls in the n boxes are independent $\text{Po}(-\log q)$ random variables.) Note this implies that $Y \stackrel{\text{st}}{\leq} Z$.

The susceptibility set $\mathcal{S}^{(h)}$ of a typical individual a in a household of size h can be constructed as follows. We first look to see which individuals make contact with a ; there are $X_1^{(h)} \sim \text{Bin}(h - 1, 1 - q_0^{(h)})$ such individuals, where

$$q_0^{(h)} = \mathbb{E} \left[\left(\frac{h-2}{h-1} \right)^{X_L} \right] = f_{X_L} \left(1 - \frac{1}{h-1} \right).$$

If $X_1^{(h)} = 0$, the process stops and $S^{(h)} = 1$. Otherwise, we take one of the $X_1^{(h)}$ individuals that have been added to the susceptibility set, individual b say, and look to see which of the remaining $h - 1 - X_1^{(h)}$ individuals make contact with b . Each of these individuals have failed to make contact with a , so the probability they make contact with b is $1 - q_1^{(h)}$, where

$$q_1^{(h)} = \frac{\mathbb{E} \left[\left(\frac{h-3}{h-1} \right)^{X_L} \right]}{\mathbb{E} \left[\left(\frac{h-2}{h-1} \right)^{X_L} \right]} = \frac{f_{X_L} \left(1 - \frac{2}{h-1} \right)}{f_{X_L} \left(1 - \frac{1}{h-1} \right)}.$$

The process is then continued in the obvious fashion. Specifically, for $k = 2, 3, \dots, h - 1$,

$$X_k^{(h)} | X_1^{(h)}, X_2^{(h)}, \dots, X_{k-1}^{(h)} \sim \text{Bin}(h - 1 - X_1^{(h)} - X_2^{(h)} - \dots - X_{k-1}^{(h)}, 1 - q_{k-1}^{(h)}),$$

where

$$q_k^{(h)} = \frac{\mathbb{E} \left[\left(\frac{h-(k+2)}{h-1} \right)^{X_L} \right]}{\mathbb{E} \left[\left(\frac{h-(k+1)}{h-1} \right)^{X_L} \right]} = \frac{f_{X_L} \left(1 - \frac{k+1}{h-1} \right)}{f_{X_L} \left(1 - \frac{k}{h-1} \right)}. \quad (5.3)$$

Let $Y_0^{(h)} = 1$ and $Y_k^{(h)} = 1 + X_1^{(h)} + X_2^{(h)} + \dots + X_k^{(h)}$ ($k = 1, 2, \dots, h$), where $X_h^{(h)} = 0$. Then $S^{(h)} \stackrel{D}{=} \min\{k \geq 1 : Y_k^{(h)} - k = 0\}$.

For $k = 0, 1, \dots, h - 2$, let $\lambda_k^{(h)} = -(h - 1) \log q_k^{(h)}$. Note that (5.3) holds also for $k = 0$, since $f_{X_L}(1) = 1$. Hence, for $k = 0, 1, \dots, h - 2$,

$$\lambda_k^{(h)} = -(h - 1) \log q_k^{(h)} = (h - 1) g_{X_L} \left(\frac{k+1}{h-1} \right) - (h - 1) g_{X_L} \left(\frac{k}{h-1} \right).$$

where $g_{X_L}(x) = -\log f_{X_L}(1-x)$ ($0 \leq x \leq 1$). The function g_{X_L} is concave, increasing and differentiable on $[0, 1]$ (recall that $\log f_{X_L}$ is convex). For $k = 0, 1, \dots, h-2$,

$$\lambda_k^{(h)} = (h-1) \int_{k/(h-1)}^{(k+1)/(h-1)} g'_{X_L}(y) dy = \int_k^{k+1} g'_{X_L}\left(\frac{u}{h-1}\right) du, \quad (5.4)$$

where we have made the substitution $u = (h-1)y$. Now g'_{X_L} is decreasing on $(0, 1)$, since g_{X_L} is concave, so it follows from (5.4) that $\lambda_k^{(h)} \geq \lambda_k^{(h')}$ if $h > h'$.

It is immediate that $S^{(h)} \stackrel{\text{st}}{\geq} S^{(1)}$, for $h > 1$, so suppose that $h > h' \geq 2$. We construct coupled realisations of $S^{(h)}$ and $S^{(h')}$, satisfying $S^{(h)} \geq S^{(h')}$, as follows. Let $\tilde{Z}_k^{(h)} \sim \text{Po}(\lambda_k^{(h)})$ ($k = 0, 1, \dots, h-2$) be independent random variables and define $\tilde{Z}_k^{(h')}$ ($k = 0, 1, \dots, h'-2$) similarly. Since, $\lambda_k^{(h)} \geq \lambda_k^{(h')}$ ($k = 0, 1, \dots, h'-2$), $\tilde{Z}_k^{(h)}$ and $\tilde{Z}_k^{(h')}$ can be coupled so that $\tilde{Z}_k^{(h)} \geq \tilde{Z}_k^{(h')}$ ($k = 0, 1, \dots, h'-2$). We show by induction that the processes $Y_k^{(h)}$ ($k \geq 0$) and $Y_k^{(h')}$ ($k \geq 0$) can be coupled so that $Y_k^{(h)} \geq Y_k^{(h')}$ for all $k = 0, 1, \dots, h'$, whence $S^{(h)} \stackrel{\text{st}}{\geq} S^{(h')}$. (Note that $S^{(h')}$ is necessarily $\leq h'$.)

Now $Y_0^{(h)} = Y_0^{(h')} = 1$. Suppose that $Y_i^{(h)} \geq Y_i^{(h')}$ for $i = 0, 1, \dots, k$, where $k \leq h' - 1$. Let $y = Y_k^{(h)}$ and $y' = Y_k^{(h')}$, so $y \geq y'$. We use the above balls-in-boxes approach to obtain a realisation of $X_{k+1}^{(h)}$. We place $\tilde{Z}_k^{(h)}$ balls uniformly at random in $h-1$ boxes, labelled $1, 2, \dots, h-1$. Then $X_{k+1}^{(h)}$ is given by the number of boxes with label $\geq y$ which contain at least one ball. A realisation of $X_{k+1}^{(h')}$ can be obtained similarly, using $\tilde{Z}_k^{(h')}$. Let $X_{k+1}'^{(h')}$ be the number of boxes with label $\geq y$ that contain at least one ball in the realisation of $X_{k+1}^{(h')}$. Now $\tilde{Z}_k^{(h)} \geq \tilde{Z}_k^{(h')}$ and $(h-y)/(h-1) > (h'-y)/(h'-1)$, so using a sequence of independent $U(0, 1)$ random variables as in the proof of Lemma 5.1, it is straightforward to couple $X_{k+1}^{(h)}$ and $X_{k+1}'^{(h')}$ so that $X_{k+1}^{(h)} \geq X_{k+1}'^{(h')}$, whence $Y_k^{(h)} \geq Y_k^{(h')}$, as required.

It follows immediately from the above argument that $f_{S^{(h)}}(s) \leq f_{S^{(h')}}(s)$ ($0 \leq s \leq 1$) if $h > h'$. Moreover, it easily seen that this inequality is strict for $s \in [0, 1)$. Hence, $z^{(h)} > z^{(h')}$ if $h > h'$. $\square$

The proof of Lemma 5.4 is similar to that of Lemma 5.2. In Lemma 5.2 we select a random typical individual and study the *forward* epidemic process of who is infected from the resulting epidemic. We couple this to a *forward* branching process and compute the probability of extinction of the branching process. In Lemma 5.4 we select a random typical individual and study the *backward* epidemic process of who, if infected, will infect our selected individual. That is, we identify the susceptibility set of the individual and couple this to a *backward* branching process and compute its probability of extinction. Dependencies in the backward branching process mean that we require conditions on (X_G, X_L) , namely, that the pgf of X_L is log-convex and $Y_L^{(p)} \sim \text{MixBin}(X_L, p)$, each local contact is independently with probability p replaced by a global contact.

For $0 \leq p \leq 1$, let $\mathcal{S}(p)$ denote the susceptibility set of a randomly chosen individual in a household of size h , where individuals have household contacts distributed according to X_L and each local contact is replaced by a global contact independently with probability p . Let $S(p) = |\mathcal{S}(p)|$, the size of the susceptibility set. Let $\mathcal{B}^B(p)$ denote the backward branching process where individuals (household susceptibility sets) have sizes independently distributed according to $S(p)$ and an individual, with a susceptibility set of size

$S(p)$ has $\text{Po}([\mu_G + p\mu_L]S(p))$ offspring. The offspring of a household susceptibility set in $\mathcal{B}^B(p)$ correspond to the set of individuals, who if infected, will infect the household susceptibility set via a global infection. Let ρ_p^B denote the extinction probability of $\mathcal{B}^B(p)$. Note that ρ_p^B satisfies

$$\mathbb{E} \left[e^{-(\mu_G + p\mu_L)S(p)(1-\rho_p^B)} \right] = \rho_p^B,$$

so $z^{(h,p)} = 1 - \rho_p^B$; cf. (2.3).

Lemma 5.4. *For a given household size h and contacts random vector $\mathbf{X} = (X_G, X_L)$, with $\log(f_{X_L}(s))$ being convex, if $Y_L^{(p)} \sim \text{MixBin}(X_L, p)$, then ρ_p^B is monotonically decreasing in p .*

Proof. Fix $0 \leq p < q \leq 1$. We prove the lemma by showing that $\rho_q^B \leq \rho_p^B$.

Construct $\mathcal{B}_B(p)$ and $\mathcal{B}_B(q)$ on a common probability space as follows. Attach to each individual a local contact random variable X_L to be used to construct susceptibility sets in the household. For each (potential) local contact assign an independent $U \sim U(0, 1)$ random variable and if $U \leq p$ ($U \leq q$) convert the local contact to a global contact in $\mathcal{B}_B(p)$ ($\mathcal{B}_B(q)$). Thus in $\mathcal{B}_B(q)$ each individual makes the same number or fewer local contacts than the corresponding individual in $\mathcal{B}_B(p)$. Each individual has *backward* global contacts to grow the branching process beyond the current household. Attach to each individual a random variable $X_B \sim \text{Po}(\mu_G + \mu_L)$ of potential global contacts into the individual. To each (potential) global contact assign an independent $\tilde{U} \sim U(0, 1)$ random variable and if $\tilde{U} \leq [\mu_G + p\mu_L]/[\mu_G + \mu_L]$ ($\tilde{U} \leq [\mu_G + q\mu_L]/[\mu_G + \mu_L]$) the global contact is kept in $\mathcal{B}_B(p)$ ($\mathcal{B}_B(q)$). Thus in $\mathcal{B}_B(q)$ each individual has the same number or more global contacts in than the corresponding individual in $\mathcal{B}_B(p)$.

As in Lemma 5.2, we construct a lower bound branching process $\hat{\mathcal{B}}_B(p, q)$ in which the i^{th} individual has the same number of global contacts in as the i^{th} individual in $\mathcal{B}_B(p)$ and the same number of local contacts out as the i^{th} individual in $\mathcal{B}_B(q)$. Let $\hat{S}(p, q)$ denote the susceptibility set of a randomly chosen individual in the branching process $\hat{\mathcal{B}}_B(p, q)$ with $\hat{S}(p, q) = |\hat{S}(p, q)|$. Then $\hat{S}(p, q) \stackrel{D}{=} S(q)$. More explicitly, by selecting a typical individual in a typical household we can construct realisations of $\mathcal{S}(p)$, $\mathcal{S}(q)$ and $\hat{S}(p, q)$ such that $\hat{S}(p, q) = \mathcal{S}(q) \subseteq \mathcal{S}(p)$, with $\hat{S}(p, q) = S(q) \leq S(p)$.

Let $\hat{W}^{(p,q)}$ be the number of potential global infectious contacts made with individuals in $\hat{S}(p, q)$ in $\hat{\mathcal{B}}_B(p, q)$. Then

$$\hat{W}^{(p,q)} | \hat{S}(p, q) \sim \text{Po} \left(\lambda_p \hat{S}(p, q) \right),$$

where $\lambda_p = \mu_G + p\mu_L$. Let $\hat{\rho}_{p,q}^B$ denote the extinction probability of the branching process $\hat{\mathcal{B}}_B(p, q)$. Then

$$\hat{\rho}_{p,q}^B = \mathbb{E} \left[\exp \left(-\lambda_p [1 - \hat{\rho}_{p,q}^B] \hat{S}(p, q) \right) \right].$$

For $0 \leq \theta \leq 1$ and $s = 1, 2, \dots$, let

$$f_p(\theta; s) = \exp(-\lambda_p[1 - \theta]s).$$

Hence, $\hat{\rho}_{p,q}^B$ solves

$$\hat{\rho}_{p,q}^B = \mathbb{E} \left[f_p(\hat{\rho}_{p,q}^B; \hat{S}(p, q)) \right] = \mathbb{E} \left[f_p(\hat{\rho}_{p,q}^B; S(q)) \right].$$

Similarly,

$$\rho_q^B = \mathbb{E} \left[f_q(\rho_q^B; S(q)) \right] = \mathbb{E} \left[f_p(\rho_q^B; S(q)) \exp \left(-(q-p)\mu_L[1-\rho_q^B]S(q) \right) \right]. \quad (5.5)$$

Let $\mathcal{V}^C$ denote the individuals in $\mathcal{S}^c(q)$ that make contact with members of $\mathcal{S}(q)$ in the construction of $\mathcal{S}(p)$. Let $V^C = |\mathcal{V}^C|$. For these V^C individuals we can construct the restricted susceptibility set, $\bar{\mathcal{S}}_R$, from members of $\mathcal{S}^c(q)$. In other words, the restricted susceptibility set precludes individuals in $\mathcal{S}(q)$. Let $\bar{S}_R = |\bar{\mathcal{S}}_R|$. (Note that if $V^C = 0$ then $\bar{\mathcal{S}}_R = \emptyset$.) Then

$$\bar{S}_R | V^C, S_0 \leq^{\text{st}} \sum_{i=1}^{V^C} S_i(p), \quad (5.6)$$

where $S_1(p), S_2(p), \dots$ are i.i.d. according to $S(p)$. The justification for (5.6) is as follows. Recall the definition of $q_k^{(h)}$ at (5.3) and note that (5.4) implies $q_k^{(h)}$ is nondecreasing in k . Let $\bar{X}_L$ denote the local infectious contact distribution of a member of $\mathcal{S}^c(q) \setminus \mathcal{V}^C$ who does not make any household contacts with $\mathcal{S}(q)$. Then $\bar{X}_L \leq^{\text{st}} X_L$, since $q_k^{(h)}$ is nondecreasing in k . Further, since $|\mathcal{S}^c(q)| \leq h-1$, we can couple the construction of the susceptibility set of one member of $\mathcal{V}^C$ with the construction of the susceptibility set of an individual in a new household of size h where all individuals have local contact distributions according to X_L , so that the size of the susceptibility set in the latter case is no smaller than the former case. If $V^C > 1$ we can repeat the process in turn for each member of $\mathcal{V}^C$ considering only those individuals in $\mathcal{S}^c(q)$ who have not previously been added to $\bar{\mathcal{S}}_R$.

Let $P(V^C, S(q))$ denote the probability of extinction of a branching process with an atypical initial individual, whose susceptibility set is formed of $\bar{\mathcal{S}}_R$, and subsequent individuals have susceptibility sets of size i.i.d. according to $S(p)$, and each member of the susceptibility set has $\text{Po}(\lambda_p)$ offspring. Then it follows from (5.6) that

$$P(V^C, S(q)) \geq [\rho_p^B]^{V^C}. \quad (5.7)$$

Also we have that ρ_p^B solves

$$\rho_p^B = \mathbb{E} \left[f_p(\rho_p^B; S(q)) P(V^C, S(q)) \right]. \quad (5.8)$$

However, from (5.7), we have that

$$\begin{aligned} \mathbb{E} \left[f_p(\rho_p^B; S(q)) P(V^C, S(q)) \right] &\geq \mathbb{E} \left[f_p(\rho_p^B; S(q)) (\rho_p^B)^{V^C} \right] \\ &= \mathbb{E} \left[f_p(\rho_p^B; S(q)) \mathbb{E} \left[(\rho_p^B)^{V^C} | S(q) \right] \right]. \end{aligned} \quad (5.9)$$

Therefore if ρ_* is the smallest solution in $[0, 1]$ of

$$\theta = \mathbb{E} \left[f_p(\theta; S(q)) \theta^{V^C} \right] = \mathbb{E} \left[f_p(\theta; S(q)) \mathbb{E} \left[\theta^{V^C} | S(q) \right] \right],$$

it follows from (5.8) and (5.9) that $\rho_p^B \geq \rho_*$.

We complete the proof of the lemma by showing, for $0 \leq \theta \leq 1$, that

$$\mathbb{E}[\theta^{V^C} | S(q)] \geq \exp(-[1 - \theta](q - p)\mu_L S(q)). \quad (5.10)$$

Since then it follows that

$$\begin{aligned} \mathbb{E}[f_p(\theta_p; S(q))\theta^{V^C}] &= \mathbb{E}[f_p(\theta; S(q))\mathbb{E}[\theta^{V^C} | S(q)]] \\ &\geq \mathbb{E}[f_p(\theta; S(q)) \exp(-[1 - \theta](q - p)\mu_L S(q))] = \mathbb{E}[f_q(\theta; S(q))], \end{aligned} \quad (5.11)$$

and together with (5.5), (5.11) implies that $\rho_* \geq \rho_q^B$, whence $\rho_p^B \geq \rho_q^B$, as required.

For $0 \leq s \leq 1$, let $f_{X_L, p}(s)$ be the pgf of $\text{MixBin}(X_L, 1 - p)$, so

$$\begin{aligned} f_{X_L, p}(s) &= \mathbb{E}[\mathbb{E}[s^{X_L, p} | X_L]] \\ &= \mathbb{E}[(p + (1 - p)s)^{X_L}] = f_{X_L}(p + [1 - p]s). \end{aligned}$$

Note that

$$f_{X_L, p}(s) = f_{X_L}(p + [1 - p][1 - s]) = f_{X_L}(1 - [1 - p]s)$$

Now

$$V^C | S(q) \sim \text{Bin}(h - S(q), 1 - r_{S(q)})$$

where for $k = 1, 2, \dots, h - 1$,

$$r_k = \frac{f_{X_L, p}(1 - k/[h - 1])}{f_{X_L, q}(1 - k/[h - 1])} = \frac{f_{X_L}(1 - (1 - p)k/[h - 1])}{f_{X_L}(1 - (1 - q)k/[h - 1])}$$

is the probability that an individual fails to infect locally a given set of k individuals, when the probability of a local contact being transferred to a global contact is p , given the individual fails to infect locally a given set of k individuals, when the probability of a local contact being transferred to a global contact is q . Hence, using $\text{Bin}(n, 1 - r) \stackrel{\text{st}}{\leq} \text{Po}(-n \log r)$ and $g_{X_L}(s) = -\log f_{X_L}(1 - s)$ ($0 \leq s \leq 1$), we have that

$$\begin{aligned} \mathbb{E}[\theta^{V^C} | S(q)] &\geq \exp\left([h - S(q)] \log \left\{ \frac{f_{X_L}(1 - (1 - p)S(q)/[h - 1])}{f_{X_L}(1 - (1 - q)S(q)/[h - 1])} \right\} [1 - \theta]\right) \\ &= \exp\left(-[h - S(q)][1 - \theta] \left\{ g_{X_L}\left(\frac{(1 - p)S(q)}{h - 1}\right) - g_{X_L}\left(\frac{(1 - q)S(q)}{h - 1}\right) \right\}\right) \\ &= \exp\left(-\frac{h - S(q)}{h - 1}[1 - \theta]\{q - p\}S(q)g'_{X_L}(\xi)\right), \end{aligned}$$

where $(1 - q)S(q)/(h - 1) < \xi < (1 - p)S(q)/(h - 1)$. Now $g'_{X_L}(\theta)$ is decreasing as f_{X_L} is log-convex, so for $0 \leq \theta \leq 1$,

$$g'_{X_L}(\theta) = \frac{f'_{X_L}(1 - \theta)}{f_{X_L}(1 - \theta)} \leq \frac{f'_{X_L}(1)}{f_{X_L}(1)} = \mu_L.$$

Hence,

$$\begin{aligned} \mathbb{E}[\theta^{V^C} | S(q)] &\geq \exp\left(-\frac{h - S(q)}{h - 1}[1 - \theta](q - p)\mu_L S(q)\right) \\ &\geq \exp(-[1 - \theta](q - p)\mu_L S(q)), \end{aligned}$$

proving (5.10) and completing the proof of the lemma. $\square$

6 Discussion

In the paper we analysed a stochastic household epidemic model characterized by the random vector (X_G, X_L) describing the number of global and local (=household) contacts individuals have, all global contacts being uniform in the entire community and all local contacts uniform in the household. Large population properties of the epidemic model were derived for the probability and size of a major outbreak. Then it was shown that the outbreak probability increases the larger household are considered, and the more of the local contacts are transferred to global contacts. The corresponding monotonicity results for the limiting relative final size z were shown to require conditions on the distribution of X_L with counter examples provided when these conditions were not satisfied.

For ease and clarity of presentation we have assumed that all households are of the same size. It is trivial to extend the central limit theorem to the case of unequal sized households provided that there exists $h_{max} < \infty$ such that all households are of size at most h_{max} . Additional conditions on the household size distribution will be required to extend the central limit theorem to the case where there is no maximum household size, see for example [4] Section 5. The monotonicity results with increasing household size are conjectured to hold if we replace increasing household size by a stochastically increasing household distribution. That is, if we have epidemics in two populations with the same (X_G, X_L) and household size distributions H_1 and H_2 , in populations 1 and 2, respectively, such that H_1 is stochastically smaller than H_2 then $\pi_1 < \pi_2$ and, provided that X_L has a log-convex pgf, $z_1 < z_2$, where π_k and z_k ($k = 1, 2$) are the probability of, and proportion infected in, a major outbreak in populations 1 and 2, respectively.

The somewhat surprising counter examples to the monotonicity result: bigger epidemics with larger households or when swapping local to global contacts, occurred when the number of local contacts X_L had low or no randomness. For example, in a household of size 3 and $X_L \equiv 1$ this would mean that an individual who gets infected would certainly infect one but not both of its household members. From an applied point of view this seems like an exceptional case, so we believe the monotonicity results are valid in most real world situations.

In Ball *et al.* [2] we analysed an epidemic model with two types of subgroups where each individual belongs to precisely one subgroup of each type. Therefore each type of subgroup forms a partition of the population and it was assumed all subgroups of a given type have a common size. We allowed for the possibility of overlap between subgroups, that is, the possibility of two or more individuals belonging to the intersection of a subgroup of type 1 and a subgroup of type 2. The model was defined by contact rates during the infectious periods (rather than arbitrary random vector as in the current paper), leading to mixed-Poisson distributed contacts of different categories. A branching process approximation for the initial stages of the epidemic and a law of large number approximation for the final proportion infected were derived for the model. Numerical investigations suggested that the final size is increasing in the size of both group structures, and also increases as the amount of overlap between the two group structures decreases. These results served as inspiration for the current paper, but here we simplified to having only one group structure. A relevant question is of course if the monotonicity can be proven also when there are two (possibly overlapping) group structures. It follows immediately that in the case that there is no overlap between the two subgroups, our results in the present paper

carry over to the case with two (and more) group structures: the final size increases if either (or both) of the two subgroup sizes increases. This follows, since as noted in Ball and Neal [6], the construction of the susceptibility set, which now extends across multiple, and in the limit possibly infinitely many groups, alternates between the two types of subgroups, so the distribution of the size of a susceptibility set of a typical individual is stochastically increasing as the size of subgroups increases. For the situation where the two group structures are partly overlapping it remains an open problem, as is the numerically motivated conjecture that the final size increases as the amount of overlap between the two group structures decreases.

The embedding argument employed in the proof of the central limit theorem in Section 4 can be utilised to study a wide range of epidemic models. As has been noted above, the central limit theorem can be applied to extensions of the Reed-Frost epidemic model where individuals are assumed to make infectious contacts with members of the population without replacement, see [14] and [16]. In a household context the key elements of the proof evolve around deriving the joint distribution of the number infected in a household, $R(t)$, and the number of global infections out of the household, $G(t)$, given that there has been a specific number of infectious contacts into the household, $Y(t)$. Therefore the approach is applicable to a wider class of models including, for example, assuming that not every individual is infected the first time they are contacted by an infectious individual but instead assuming there is a distribution on the number of infectious contacts required to infect an individual. Beyond the household model, the embedding argument could be employed to central limit theorems for the final size of epidemics in other two-level mixing population structures such as the great circle epidemic model, [7], and network epidemic models, Ball and Neal [8], allowing progress beyond the mixed-Poisson distributions of global and local contacts in these earlier works.

An important assumption in the current model is that the random number of (uniformly chosen) local contacts X_L is independent of household size. Some household epidemic models are defined by assuming the contact rate to *each* household member equals some constant β_H . The overall rate to infect household members if in a household of size h then equals $(h-1)\beta_H$ (in our model the contact rate, or equivalently total number of contacts, is assumed to be the same irrespective of household size). In such a situation, the epidemic is easily shown to increase the larger the household size is. Most network epidemic models makes a similar assumption: the rate or probability of contacting a given neighbour is fixed and independent of the number of neighbours. For a network epidemic model to more closely mimic the current model a fixed overall rate of infecting neighbours would be required, which is then distributed uniformly among the neighbours. The effect would be that highly connected individuals are no longer necessarily super-spreaders to the same extent. Would such an epidemic increase if the mean degree increased? Would the final size increase if the degree distribution has a heavier tail? These are some interesting open questions.

A Calculation of asymptotic properties of $\mathcal{E}_{n,h}(X_G, X_L)$

In this appendix we outline calculation of the major outbreak probability, $1 - \rho$, and the asymptotic mean, z , and scaled variance, σ^2 of the fraction infected by a major out-

break; see Theorem 2.1. We make extensive use of Gontcharoff polynomials (see Lefèvre and Picard [13]). Let $U = u_0, u_1, \dots$ be a sequence of real numbers. The Gontcharoff polynomials associated with U , i.e. $G_i(x|U)$ ($i = 0, 1, \dots$) are defined by

$$\sum_{i=0}^n n_{[i]} u_i^{n-i} G_i(x|U) = x^n \quad (n = 0, 1, \dots), \quad (\text{A.1})$$

where $n_{[i]} = n(n-1)\dots(n-i+1)$ denotes a falling factorial, with the convention that $n_{[0]} = 1$. Note that $G_0(x) = 1$ ($x \in \mathbb{R}$) and that $G_i(x|U)$ ($i = 1, 2, \dots$) can be computed recursively for fixed x . Further, $G_i(x|U)$ is a polynomial of degree i and (Lefèvre and Picard [13], Property 2.4) for $0 \leq j \leq i$,

$$G_i^{(j)}(x|U) = G_{i-j}(x|E^j U), \quad (\text{A.2})$$

where $G_i^{(j)}(x|U)$ denotes the j^{th} derivative of $G_i(x|U)$ and $E^j U$ is the sequence $u_j, u_{j+1}, \dots$.

For $h = 1, 2, \dots$ and $\pi \in [0, 1]$, recall the epidemic model $\tilde{\mathcal{E}}_h^H(X_G, X_L, \pi)$ defined in Section 2.4. Let $\tilde{R}^{(h)}$ be the number of individuals infected in $\tilde{\mathcal{E}}_h^H(X_G, X_L, \pi)$ and $\tilde{G}^{(h)}$ be the total number of global contacts that emanate from those infectives. Further, let $\tilde{S}^{(h)} = h - \tilde{R}^{(h)}$ be the number of susceptibles at the end of the epidemic. Note that if $\pi = e^{-t}$ then $(\tilde{R}^{(h)}, \tilde{G}^{(h)}) \stackrel{D}{=} (R(t), G(t))$, defined in Section 2.4. We derive expressions for $E[\tilde{S}_{[i]}^{(h)}]$ ($i = 1, 2$) and $E[\tilde{S}^{(h)}\tilde{G}^{(h)}]$, from which $\nu_R(t) = h^{-1}E[R(t)]$, $\text{var}(R(t))$ and $\text{cov}(R(t), G(t))$ follow easily.

For $k = 1, 2, \dots, h-1$, let $A_k^{(h)}$ be the event that an infective in $\tilde{\mathcal{E}}_h^H(X_G, X_L, \pi)$ fails to contact anyone in a given set of k susceptibles in the household. For $s \in [0, 1]$, let $q_0(s) = E[s^{X_G}]$ and $q_k(s) = E[s^{X_G} 1_{A_k^{(h)}}]$ ($k = 1, 2, \dots, h-1$). Then,

$$q_k(s) = f_{X_G, X_L} \left(s, 1 - \frac{k}{h-1} \right) \quad (k = 0, 1, \dots, h-1), \quad (\text{A.3})$$

where f_{X_G, X_L} is the joint pgf of (X_G, X_L) . Let $\tilde{f}_h(s_1, s_2) = E[s_1^{\tilde{S}^{(h)}} s_2^{\tilde{G}^{(h)}}]$ ($s_1, s_2 \in [0, 1]$). Then it follows using Ball [1], Theorem 3.3, that

$$\tilde{f}_h(s_1, s_2) = \sum_{i=0}^h h_{[i]} (q_i(s_2))^{h-i} \pi^i G_i(s_1|U(s_2)) \quad (s_1, s_2 \in [0, 1]), \quad (\text{A.4})$$

where the sequence $U(s_2)$ satisfies $u_i(s_2) = q_i(s_2)$ ($i = 0, 1, \dots, h-1$). (Note from (A.1) that, for $i = 1, 2, \dots$, $G_i(x|U)$ is determined by $u_0, u_1, \dots, u_{i-1}$.)

Remark A.1. Observe from (A.3) that $q_k(s)$, and hence also $U(s_2)$ and $G_i(s_1|U(s_2))$, depends on h . We have suppressed this dependence for ease of presentation but note that in the sequel all Gontcharoff polynomials need to be recalculated if the household size is changed.

Remark A.2. Note that if the local contacts made by an infective are without replacement then, for $k = 0, 1, \dots, h-1$,

$$P(A_k^{(h)} | X_L) = \frac{\binom{h-1-k}{X_L}}{\binom{h-1}{X_L}} = \frac{(h-1-X_L)_{[k]}}{(h-1)_{[k]}},$$

so (A.3) is replaced by

$$q_k(s) = \mathbb{E} \left[s^{X_G} \frac{(h-1-X_L)_{[k]}}{(h-1)_{[k]}} \right] \quad (k = 0, 1, \dots, h-1).$$

Differentiating (A.4) partially k times with respect to s_1 , using (A.2), yields (see [1], Proposition 3.1)

$$\mathbb{E}[\tilde{S}_{[k]}^{(h)}] = \sum_{i=k}^h h_{[i]} q_i^{h-i} \pi^i G_{i-k}(1|E^k U) \quad (k = 1, 2, \dots, h),$$

where $q_i = q_i(1) = f_{X_L}(1 - \frac{i}{h-1})$ ($i = 0, 1, \dots, h-1$) and U satisfies $u_i = q_i$ ($i = 0, 1, \dots, h-1$).

Remark A.3. Note from [1], Lemma 3.1, that $P(S^{(h)} = i) = (h-1)_{[i-1]} G_{i-1}(1|EU) q_i^{h-i}$ ($i = 1, 2, \dots, h$), where $S^{(h)}$ is the size of the susceptibility set of a typical individual in a household of size h . Setting $\pi = e^{-t}$, so $R(t) \stackrel{D}{=} h - \tilde{S}^{(h)}$, yields (2.2), and also enables $\nu'_R(t)$ to be computed easily.

For a function $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ and $(k_1, k_2) \in \mathbb{Z}_+^2$, let $f^{(k_1, k_2)}(s_1, s_2)$ denote the partial derivative of f of order k_1 in s_1 and k_2 in s_2 . Then, $\mathbb{E}[\tilde{S}^{(h)} \tilde{G}^{(h)}] = \tilde{f}_h^{(1,1)}(1, 1)$. For $i = 0, 1, \dots, h-1$ and $(s_1, s_2) \in [0, 1]^2$, let $\alpha_i(s_1, s_2) = G_i(s_1|EU(s_2))$. Differentiating (A.4) with respect to s_1 , using (A.2) and noting that $G_0^{(1)}(s_1|U(s_2)) = 0$, yields

$$\tilde{f}_h^{(1,0)}(s_1, s_2) = \sum_{i=1}^h h_{[i]} (q_i(s_2))^{h-i} \pi^i \alpha_{i-1}(s_1, s_2). \quad (\text{A.5})$$

Recalling that $q_i = q_i(1)$, differentiating (A.5) with respect to s_2 yields

$$\mathbb{E}[\tilde{S}^{(h)} \tilde{G}^{(h)}] = \sum_{i=1}^h h_{[i]} q_i^{h-i} \pi^i \alpha_{i-1}^{(0,1)}(1, 1) + \sum_{i=1}^{h-1} h_{[i+1]} q_i^{(1)}(1) q_i^{h-i-1} \pi^i \alpha_{i-1}(1, 1).$$

Using (A.1),

$$\sum_{i=0}^n n_{[i]} (q_{i+1}(s_2))^{n-i} \alpha_i(s_1, s_2) = s_1^n \quad (n = 0, 1, \dots, h-1), \quad (\text{A.6})$$

whence, differentiating (A.6) partially with respect to s_2 ,

$$\sum_{i=0}^{n-1} n_{[i+1]} q_{i+1}^{(1)}(1) q_{i+1}^{n-i-1} \alpha_i(1, 1) + \sum_{i=1}^n n_{[i]} q_{i+1}^{n-i} \alpha_i^{(0,1)}(1, 1) = 0 \quad (n = 0, 1, \dots, h-1). \quad (\text{A.7})$$

Now $\alpha_i(1, 1)$ ($i = 0, 1, \dots, h-1$) can be computed using (A.6), $\alpha_0^{(0,1)} = 0$ and $\alpha_i^{(0,1)}$ ($i = 1, 2, \dots, h-1$) can be computed using (A.7), thus enabling $\mathbb{E}[\tilde{S}^{(h)} \tilde{G}^{(h)}]$ to be computed.

For $h = 1, 2, \dots$, let $C^{(h)}$ be the total number of global contacts that emanate from $\mathcal{E}_h^H(X_G, X_L)$ defined in Section 2.4. Then, using [1], Theorem 3.3,

$$f_{C^{(h)}}(s) = \sum_{i=0}^{h-1} (h-1)_{[i]} (q_i(s))^{h-i} G_i(1|U(s)) \quad (s \in [0, 1]),$$

thus enabling $f_{C^{(h)}}(s)$, and hence ρ , to be computed.

The above enables the asymptotic properties of $\mathcal{E}_{n,h}(X_G, X_L)$ to be computed. To compute the asymptotic properties of $\mathcal{E}_{n,h}(X_G, X_L, p)$, for $p \neq 0$, note that elementary calculation yields

$$\begin{aligned} f_{X_G^{(p)}, X_L^{(p)}}(s_1, s_2) &= f_{X_G, X_L}(s_1, ps_1 + (1-p)s_2) \quad ((s_1, s_2) \in [0, 1]^2), \\ \mathbb{E}[X_G^{(p)}] &= \mathbb{E}[X_G] + p\mathbb{E}[X_L], \\ \text{var}(X_G^{(p)}) &= \text{var}(X_G) + 2p\text{cov}(X_G, X_L) + p^2\text{var}(X_L) + p(1-p)\mathbb{E}[X_L]. \end{aligned}$$

B Proof of Theorem 2.4

For $h = 1, 2, \dots$ and $p \in [0, 1]$, let $S^{(h,p)}$ denote the size of the (household) susceptibility set of a typical individual in $\mathcal{E}_{n,h}(X_G, X_L, p)$. The mean number of global contacts made by a typical individuals in $\mathcal{E}_{n,h}(X_G, X_L, p)$ is $\mu_G + p\mu_L$, so it follows from (2.3) that $z^{(h,p)}$ is given by the largest solution in $[0, 1]$ of

$$1 - z = f_{S^{(h,p)}}(e^{-z(\mu_G + p\mu_L)}).$$

Suppose that $\alpha = \mu_G + \mu_L > 1$, so the (homogeneously mixing) epidemic when $p = 1$ is supercritical, and let z_1 be the unique solution of $1 - z = e^{-\alpha z}$ in $(0, 1)$. (Note that $z_1 = z_{\text{hom}}(\alpha)$, defined just before the statement of Theorem 2.4 in Section 2.4.) For $h = 1, 2, \dots$ let

$$g_h(p) = f_{S^{(h,p)}}(e^{-z_1(\mu_G + p\mu_L)}) \quad (p \in [0, 1]). \quad (\text{B.1})$$

The behaviour of $z^{(h,p)}$ near $p = 1$ is determined by the derivatives of g_h at $p = 1$.

It follows from Remark A.3, (A.1) and a little algebra that

$$\sum_{i=0}^n \binom{h-1-i}{n-i} v_i^{n+1-h} \mathbb{P}(S^{(h)} = i+1) = \binom{h-1}{n} \quad (n = 0, 1, \dots, h-1), \quad (\text{B.2})$$

where

$$v_i = q_{i+1} = f_{X_L} \left(1 - \frac{i+1}{h-1} \right) \quad (i = 0, 1, \dots, h-2).$$

Now $f_{X_L^{(p)}}(s) = f_{X_L}(p + (1-p)s)$, so let

$$v_i(p) = f_{X_L^{(p)}} \left(1 - \frac{i+1}{h-1} \right) = f_{X_L} \left(1 - \frac{(1-p)(i+1)}{h-1} \right) \quad (p \in [0, 1]).$$

(Note that $v_i(p)$ depends also on h but we suppress that dependence for ease of notation.)

Fix $h \geq 2$ and let $f_i(p) = \mathbb{P}(S^{(h,p)} = i+1)$ ($i = 0, 1, \dots, h-1$). Then, using (B.2),

$$\sum_{i=0}^n \binom{h-1-i}{n-i} (v_i(p))^{n+1-h} f_i(p) = \binom{h-1}{n} \quad (n = 0, 1, \dots, h-1). \quad (\text{B.3})$$

Noting that $v_i(1) = 1$ ($i = 0, 1, \dots, h-2$), it follows from (B.3) that

$$f_0(1) = 1 \quad \text{and} \quad f_i(1) = 0 \quad (i = 1, 2, \dots, h-1). \quad (\text{B.4})$$

Now,

$$\frac{d}{dp} (v_i(p)^{n+1-h})|_{p=1} = \frac{(n+1-h)(i+1)\mu_L}{h-1},$$

so differentiating (B.3) and using (B.4) yields

$$\sum_{i=0}^n \binom{h-1-i}{n-i} f_i^{(1)}(1) = \binom{h-1}{n} \left(\frac{h-1-n}{h-1} \right) \mu_L \quad (n = 0, 1, \dots, h-1). \quad (\text{B.5})$$

Successively setting $n = 0, 1, 2, \dots, h-1$ in (B.5) yields, after a little algebra,

$$f_0^{(1)}(1) = \mu_L, \quad f_1^{(1)}(1) = -\mu_L \quad \text{and} \quad f_i^{(1)}(1) = 0 \quad (i = 2, 3, \dots, h-1). \quad (\text{B.6})$$

Now

$$\frac{d^2}{dp^2} (v_0(p)^{n+1-h})|_{p=1} = \frac{(h-1-n)(h-n)\mu_L^2}{(h-1)^2} - \frac{(h-1-n)\mu_{L,[2]}}{(h-1)^2},$$

where $\mu_{L,[2]} = E[X_L(X_L - 1)]$. Differentiating (B.3) twice yields, after using (B.4) and (B.6),

$$\begin{aligned} \sum_{i=0}^n \binom{h-1-i}{n-i} f_i^{(2)}(1) &= 2 \frac{(h-1-n)}{(h-1)} \left[\binom{h-1}{n} - 2 \binom{h-2}{n-1} 1_{\{n \geq 1\}} \right] \mu_L^2 \\ &+ \binom{h-1}{n} \frac{(h-1-n)}{(h-1)^2} [\mu_{L,[2]} - (h-n)\mu_L^2] \quad (n = 0, 1, \dots, h-1). \end{aligned} \quad (\text{B.7})$$

Successively setting $n = 0, 1, 2, \dots, h-1$ in (B.7) yields, after a little algebra,

$$\begin{aligned} f_0^{(2)}(1) &= \frac{1}{h-1} [\mu_{L,[2]} + (h-2)\mu_L^2], \quad f_1^{(2)}(1) = -\frac{1}{h-1} [\mu_{L,[2]} + 4(h-2)\mu_L^2], \\ f_2^{(2)}(1) &= 3 \frac{h-2}{h-1} \mu_L^2 \quad \text{and} \quad f_i^{(2)}(1) = 0 \quad (i = 3, 4, \dots, h-1). \end{aligned} \quad (\text{B.8})$$

Returning to g_h , note from (B.1) that

$$g_h(p) = \sum_{k=1}^h f_{k-1}(p) e^{-kz_1(\mu_G + p\mu_L)} \quad (p \in [0, 1]). \quad (\text{B.9})$$

Differentiating (B.9) yields, after using (B.4) and (B.6),

$$g_h^{(1)}(1) = \mu_L e^{-\alpha z_1} (1 - z_1 - e^{-\alpha z_1}) = 0, \quad (\text{B.10})$$

since $1 - z_1 = e^{-\alpha z_1}$. Differentiating (B.9) twice, and using (B.4), (B.6) and (B.8), yields after a little algebra

$$g_h^{(2)}(1) = \frac{z_1(1-z_1)}{h-1} [\mu_{L,[2]} + (2-3z_1)\mu_L^2] = \frac{z_1(1-z_1)}{h-1} [\sigma_L^2 - \mu_L + 3(1-z_1)\mu_L^2]. \quad (\text{B.11})$$

Recall that $h \geq 2$ is fixed and let $z(p) = z^{(h,p)}$ ($p \in [0, 1]$). Then (B.10) and (B.11) imply that $z^{(1)}(1) = 0$ and $z^{(2)}(1) > 0 (< 0)$ if $\sigma_L^2 - \mu_L + 3(1-z_1)\mu_L^2 > 0 (< 0)$, from which Theorem 2.4 follows easily.

C Proof of Theorem 2.5

We prove the theorem in the case $p = 0$, with a similar proof holding for $0 < p < 1$. Note that the case $p = 1$ is trivial as the epidemic is a homogeneously mixing epidemic with mean number of contacts made by each individual being $\mu_G + \mu_L$.

For $h = 1, 2, \dots$, let $S^{(h)}$ denote the size of the susceptibility set of a typical individual in a household of size h . The probability that an individual with $X_L = x_L$ contacts the same individual twice in the household converges to 0 as the household size $h \rightarrow \infty$. Therefore for large h , the probability an individual contacts a given individual in their household via a local infection is approximately $\mu_L/(h-1)$. It is then straightforward to couple the construction of $S^{(h)}$ to a branching process with offspring distribution $V_h \sim \text{Bin}(h-1, \mu_L/(h-1))$, with $V_h \xrightarrow{D} \tilde{V} \sim \text{Po}(\mathbb{E}[X_L])$ as $h \rightarrow \infty$.

Let $\tilde{\mathcal{B}}$ denote the branching process with offspring distribution $\tilde{V}$ and let $\tilde{S}$ denote the total size of the branching process $\tilde{\mathcal{B}}$. Then for $0 \leq s \leq 1$, the probability generating function of $\tilde{S}$ satisfies

$$\begin{aligned} \mathbb{E}[s^{\tilde{S}}] &= f_{\tilde{S}}(s) = s \mathbb{E}[f_{\tilde{S}}(s)^{\tilde{V}}] \\ &= s \exp(-\mu_L [1 - f_{\tilde{S}}(s)]) . \end{aligned} \quad (\text{C.1})$$

It follows that $z^{(h,0)} \rightarrow \tilde{z}$ as $h \rightarrow \infty$, where $\tilde{z}$ satisfies

$$1 - \tilde{z} = f_{\tilde{S}}(\exp(-\mu_G \tilde{z})), \quad (\text{C.2})$$

and it remains to show that $\tilde{z} = z_{\text{hom}}(\alpha)$, where $\alpha = \mu_G + \mu_L$.

We set $s = \exp(-\mu_G \tilde{z})$ in (C.1), and then using (C.2), we have that

$$\begin{aligned} f_{\tilde{S}}(e^{-\mu_G \tilde{z}}) &= e^{-\mu_G \tilde{z}} \exp(-\mu_L [1 - f_{\tilde{S}}(e^{-\mu_G \tilde{z}})]) \\ &= e^{-\mu_G \tilde{z}} \exp(-\mu_L \tilde{z}) = \exp(-[\mu_G + \mu_L] \tilde{z}) . \end{aligned}$$

Therefore $\tilde{z}$ solves

$$\tilde{z} = 1 - f_{\tilde{S}}(e^{-\mu_G \tilde{z}}) = 1 - \exp(-[\mu_G + \mu_L] \tilde{z}) = 1 - \exp(-\alpha \tilde{z}),$$

which is the defining equation for $z_{\text{hom}}(\alpha)$. Therefore, $\tilde{z} = z_{\text{hom}}(\alpha)$, as required.

D Comparison of variance expressions

In this appendix we prove that the expressions for σ^2 in (4.30) and (4.31) are equivalent. The first step, using (4.30), is to note that,

$$\begin{aligned} \sigma^2 &= \frac{1}{h} [\text{var}(R_1(\tau)) + b(\tau)^2 \text{var}(G_1(\tau)) + b(\tau)^2 \text{var}(Y_1(\tau)) \\ &\quad + 2b(\tau) \text{cov}(R_1(\tau), G_1(\tau)) - 2b(\tau) \text{cov}(R_1(\tau), Y_1(\tau)) - 2b(\tau)^2 \text{cov}(G_1(\tau), Y_1(\tau))] . \end{aligned} \quad (\text{D.1})$$

We consider separately each of the variance and covariance terms on the right-hand side of (D.1).

Using exchangeability of individuals,

$$\begin{aligned}\text{var}(R_1(\tau)) &= h\text{var}(\chi_{11}(\tau)) + h(h-1)\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) \\ &= h\nu_R(\tau)[1 - \nu_R(\tau)] + h(h-1)\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)).\end{aligned}\quad (\text{D.2})$$

Similarly,

$$\text{var}(G_1(\tau)) = h\text{var}(\chi_{11}(\tau)X_{G,(1,1)}) + h(h-1)\text{cov}(\chi_{11}(\tau)X_{G,(1,1)}, \chi_{12}(\tau)X_{G,(1,2)}). \quad (\text{D.3})$$

Since $\chi_{11}(\tau)^2 = \chi_{11}(\tau)$,

$$\begin{aligned}\text{var}(\chi_{11}(\tau)X_{G,(1,1)}) &= \text{E}[\chi_{11}(\tau)X_{G,(1,1)}^2] - \nu_R(\tau)^2\mu_G^2 \\ &= \nu_R(\tau)[\sigma_G^2 + \mu_G^2] - \nu_R(\tau)^2\mu_G^2 \\ &= \nu_R(\tau)[1 - \nu_R(\tau)]\mu_G^2 + \nu_R(\tau)\sigma_G^2.\end{aligned}\quad (\text{D.4})$$

An observation similar to that made in [7], Section 4, is that, conditional upon $\chi_{11}(\tau) = 0$ (individual (1, 1)'s susceptibility set is not contacted when each members of the population is exposed to τ units of global infectious pressure), $(X_{G,(1,1)}, X_{L,(1,1)})$ and $\chi_{12}(\tau)$ are independent, so

$$\text{E}[(1 - \chi_{11}(\tau))(1 - \chi_{12}(\tau))X_{G,(1,1)}X_{G,(1,2)}] = \mu_G^2\text{E}[(1 - \chi_{11}(\tau))(1 - \chi_{12}(\tau))].$$

Since also

$$\begin{aligned}&\text{E}[(1 - \chi_{11}(\tau))(1 - \chi_{12}(\tau))X_{G,(1,1)}X_{G,(1,2)}] \\ &= \text{E}[X_{G,(1,1)}X_{G,(1,2)}] - 2\text{E}[\chi_{11}(\tau)X_{G,(1,1)}X_{G,(1,2)}] + \text{E}[\chi_{11}(\tau)X_{G,(1,1)}\chi_{12}(\tau)X_{G,(1,2)}] \\ &= \mu_G^2 - 2\mu_G\text{E}[\chi_{11}(\tau)X_{G,(1,2)}] + \text{E}[\chi_{11}(\tau)X_{G,(1,1)}\chi_{12}(\tau)X_{G,(1,2)}],\end{aligned}$$

it follows that

$$\begin{aligned}&\text{E}[\chi_{11}(\tau)X_{G,(1,1)}\chi_{12}(\tau)X_{G,(1,2)}] \\ &= \mu_G^2\{\text{E}[(1 - \chi_{11}(\tau))(1 - \chi_{12}(\tau))] - 1\} + 2\mu_G\text{E}[\chi_{11}(\tau)X_{G,(1,2)}] \\ &= \mu_G^2\text{E}[\chi_{11}(\tau)\chi_{12}(\tau)] + 2\mu_G\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}).\end{aligned}\quad (\text{D.5})$$

Thus,

$$\begin{aligned}\text{cov}(\chi_{11}(\tau)X_{G,(1,1)}, \chi_{12}(\tau)X_{G,(1,2)}) &= \text{E}[\chi_{11}(\tau)X_{G,(1,1)}\chi_{12}(\tau)X_{G,(1,2)}] - \nu_R(\tau)^2\mu_G^2 \\ &= \mu_G^2\text{E}[\chi_{11}(\tau)\chi_{12}(\tau)] + 2\mu_G\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) - \nu_R(\tau)^2\mu_G^2 \\ &= \mu_G^2\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) + 2\mu_G\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}).\end{aligned}\quad (\text{D.6})$$

Hence, substituting (D.4) and (D.6) into (D.3),

$$\begin{aligned}\text{var}(G_1(\tau)) &= h\nu_R(\tau)[1 - \nu_R(\tau)]\mu_G^2 + h\nu_R(\tau)\sigma_G^2 \\ &\quad + h(h-1)\mu_G^2\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) + 2h(h-1)\mu_G\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) \\ &= \mu_G^2\text{var}(R_1(\tau)) + h\nu_R(\tau)\sigma_G^2 + 2h(h-1)\mu_G\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}).\end{aligned}\quad (\text{D.7})$$

Since $Y_1(\tau) \sim \text{Po}(h\tau)$, we have that $\text{var}(Y_1(\tau)) = h\tau$.

Turning to the covariance terms, since $\chi_{11}(\tau)$ and $X_{G,(1,1)}$ are independent,

$$\begin{aligned}\text{cov}(R_1(\tau), G_1(\tau)) &= h\text{cov}(\chi_{11}(\tau), \chi_{11}(\tau)X_{G,(1,1)}) + h(h-1)\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)X_{G,(1,2)}) \\ &= h\mu_G\nu_R(\tau)[1 - \nu_R(\tau)] + h(h-1)\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)X_{G,(1,2)}).\end{aligned}$$

A similar argument to the derivation of (D.5) yields

$$\mathbb{E}[\chi_{11}(\tau)\chi_{12}(\tau)X_{G,(1,2)}] = \mu_G\mathbb{E}[\chi_{11}(\tau)\chi_{12}(\tau)] + \text{cov}(\chi_{11}(\tau), X_{G,(1,2)}),$$

so

$$\begin{aligned}\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)X_{G,(1,2)}) &= \mathbb{E}[\chi_{11}(\tau)\chi_{12}(\tau)X_{G,(1,2)}] - \mu_G\nu_R(\tau)^2 \\ &= \mu_G\mathbb{E}[\chi_{11}(\tau)\chi_{12}(\tau)] + \text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) - \mu_G\nu_R(\tau)^2 \\ &= \mu_G\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) + \text{cov}(\chi_{11}(\tau), X_{G,(1,2)}).\end{aligned}$$

Hence,

$$\begin{aligned}\text{cov}(R_1(\tau), G_1(\tau)) &= h\mu_G\nu_R(\tau)[1 - \nu_R(\tau)] + h(h-1)\mu_G\text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) \\ &\quad + h(h-1)\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) \\ &= \mu_G\text{var}(R_1(\tau)) + h(h-1)\text{cov}(\chi_{11}(\tau), X_{G,(1,2)}).\end{aligned}\tag{D.8}$$

Next, we have that

$$\text{cov}(R_1(\tau), Y_1(\tau)) = h\text{cov}(\chi_{11}(\tau), \zeta_{11}(\tau)) + h(h-1)\text{cov}(\chi_{11}(\tau), \zeta_{12}(\tau)).\tag{D.9}$$

Since $\chi_{11}(\tau) = 1$ if $\zeta_{11}(\tau) > 0$ and $\zeta_{11}(\tau) \sim \text{Po}(\tau)$, we have that

$$\begin{aligned}\text{cov}(\chi_{11}(\tau), \zeta_{11}(\tau)) &= \mathbb{E}[\chi_{11}(\tau)\zeta_{11}(\tau)] - \tau\nu_R(\tau) = \mathbb{E}[\zeta_{11}(\tau)] - \tau\nu_R(\tau) \\ &= \tau[1 - \nu_R(\tau)].\end{aligned}\tag{D.10}$$

Note that $\text{P}(\chi_{11}(\tau) = 1 \mid \zeta_{12}(\tau) = k) = \text{P}(\chi_{11}(\tau) = 1 \mid \zeta_{12}(\tau) = 1)$ ($k = 1, 2, \dots$), so

$$\begin{aligned}\mathbb{E}[\chi_{11}(\tau)\zeta_{12}(\tau)] &= \sum_{k=0}^{\infty} k\text{P}(\chi_{11}(\tau) = 1 \mid \zeta_{12}(\tau) = k)\text{P}(\zeta_{12}(\tau) = k) \\ &= \tau\text{P}(\chi_{11}(\tau) = 1 \mid \zeta_{12}(\tau) = 1),\end{aligned}$$

since $\zeta_{12}(\tau) \sim \text{Po}(\tau)$. Also, $\text{P}((1, 2) \notin \mathcal{S}_{1,1} \mid S_{11} = i) = \frac{h-i}{h-1}$ ($i = 1, 2, \dots, h$), so

$$\text{P}(\chi_{11}(\tau) = 0 \mid \zeta_{12}(\tau) = 1) = \sum_{i=1}^h \text{P}(S_{11} = i) \left[\frac{h-i}{h-1} \right] e^{-i\tau}.$$

Noting that $\nu_R(\tau) = 1 - f_S(e^{-\tau}) = 1 - \sum_{i=1}^h \text{P}(S_{11} = i)e^{-i\tau}$, a short calculation yields

$$\mathbb{E}[\chi_{11}(\tau)\zeta_{12}(\tau)] = \tau \left[\nu_R(\tau) + \frac{\nu_R(\tau) + \nu'_R(\tau) - 1}{h-1} \right],$$

whence

$$\begin{aligned}\text{cov}(\chi_{11}(\tau), \zeta_{12}(\tau)) &= \mathbb{E}[\chi_{11}(\tau)\zeta_{12}(\tau)] - \tau\nu_R(\tau) \\ &= \frac{\tau[\nu_R(\tau) + \nu'_R(\tau) - 1]}{h - 1}.\end{aligned}\quad (\text{D.11})$$

Substituting (D.10) and (D.11) into (D.9) yields

$$\begin{aligned}\text{cov}(R_1(\tau), Y_1(\tau)) &= h\tau[1 - \nu_R(\tau)] + h\tau[\nu_R(\tau) + \nu'_R(\tau) - 1] \\ &= h\tau\nu'_R(\tau).\end{aligned}\quad (\text{D.12})$$

Also

$$\text{cov}(G_1(\tau), Y_1(\tau)) = \mu_G \text{cov}(R_1(\tau), Y_1(\tau)) = h\mu_G \tau \nu'_R(\tau). \quad (\text{D.13})$$

Substituting (D.2), (D.7), $\text{var}(Y_1(\tau)) = h\tau$, (D.8), (D.12) and (D.13) into (D.1) yields

$$\begin{aligned}\sigma^2 &= \frac{1}{h} [\text{var}(R_1(\tau)) + b(\tau)^2 \mu_G^2 \text{var}(R_1(\tau)) + hb(\tau)^2 \nu_R(\tau) \sigma_G^2 \\ &\quad + 2h(h-1)b(\tau)^2 \mu_G \text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) + b(\tau)^2 h\tau + 2b(\tau) \mu_G \text{var}(R_1(\tau)) \\ &\quad + 2b(\tau)h(h-1) \text{cov}(\chi_{11}(\tau), X_{G,(1,2)}) - 2b(\tau)[1 + \mu_G b(\tau)]h\tau\nu'_R(\tau)].\end{aligned}$$

Now,

$$1 + \mu_G b(\tau) = 1 + \frac{\mu_G \nu'_R(\tau)}{1 - \mu_G \nu'_R(\tau)} = \frac{1}{1 - \mu_G \nu'_R(\tau)},$$

so

$$\{1 + \mu_G b(\tau)\} \nu'_R(\tau) = \frac{\nu'_R(\tau)}{1 - \mu_G \nu'_R(\tau)} = b(\tau).$$

Hence, using $\tau = \nu_G(\tau) = \mu_G \nu_R(\tau)$,

$$\begin{aligned}\sigma^2 &= \frac{1}{h} [\{1 + \mu_G b(\tau)\}^2 \text{var}(R_1(\tau)) + b(\tau)^2 \{h\nu_R(\tau) \sigma_G^2 + h\tau - 2h\tau\} \\ &\quad + 2h(h-1)b(\tau)[1 + \mu_G b(\tau)] \text{cov}(\chi_{11}(\tau), X_{G,(1,2)})] \\ &= (1 + b(\tau)\mu_G)^2 \nu_R(\tau)[1 - \nu_R(\tau)] + b(\tau)^2 \nu_R(\tau)[\sigma_G^2 - \mu_G] \\ &\quad + (h-1) [(1 + b(\tau)\mu_G)^2 \text{cov}(\chi_{11}(\tau), \chi_{12}(\tau)) + 2b(\tau)(1 + \mu_G b(\tau)) \text{cov}(\chi_{11}(\tau), X_{G,(1,2)})].\end{aligned}\quad (\text{D.14})$$

The right-hand side of (D.14) agrees with (4.31) completing the proof.

The expression for σ^2 given by (2.4) follows after a little algebra by substituting (D.12), (D.13) and $\text{var}(Y_1(\tau)) = h\tau$ into (D.1) and noting that (D.7) and (D.8) imply

$$\text{var}(G_1(\tau)) = 2\mu_G \text{cov}(R_1(\tau), G_1(\tau)) - \mu_G^2 \text{var}(R_1(\tau)) + h\nu_R(\tau) \sigma_G^2.$$

Funding: T.B. was supported in part by the Swedish Research Council (grant 2020-0474).

References

- [1] Ball, F. (2019) Susceptibility Sets and the Final Outcome of Collective Reed-Frost Epidemics. *Meth. Comp. Appl. Prob.* **21**, 401–421.
- [2] Ball F, Britton T and Neal P. (2024). An epidemic model on a network having two group structures with tunable overlap. *Submitted*. arXiv:2410.06696
- [3] Ball, F. and Donnelly, P. (1995) Strong approximations for epidemic models. *Stoc. Proc. Appl.* **55**, 1–21.
- [4] Ball, F. and Lyne, O.D. (2001). Stochastic multitype SIR epidemics among a population partitioned into households. *Adv. Appl. Prob.* **33**, 99–123.
- [5] Ball, F., Mollison, D. and Scalia-Tomba, G. (1997). Epidemics with two levels of mixing. *Ann. Appl. Prob.* **7**, 46–89.
- [6] Ball, F. and Neal, P. (2002). A general model for stochastic SIR epidemics with two levels of mixing. *Math. Biosci.* **180**, 73–102.
- [7] Ball, F. and Neal, P. (2003). The great circle epidemic model. *Stoc. Proc. Appl.* **107**, 233–268.
- [8] Ball, F. and Neal, P. (2008). Network epidemic models with two levels of mixing. *Math. Biosci.* **212**, 69–87.
- [9] Ball, F. and Neal, P. (2017). The asymptotic variance of the giant component of configuration model random graphs. *Ann. Appl. Prob.* **27**, 1057–1092.
- [10] Ball, F. and Neal, P. (2024). Strong convergence of an epidemic model with mixing groups. *Adv. Appl. Prob.* **56**, 430–463.
- [11] 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. [https://doi.org/10.1016/0025-5564\(94\)00055-5](https://doi.org/10.1016/0025-5564(94)00055-5)
- [12] Billingsley, P. (1999). *Convergence of Probability Measures*. Second Edition. Wiley, New York.
- [13] Lefèvre, C. and Picard, P. (1990) A non-standard family of polynomials and the final size distribution of Reed-Frost epidemic processes. *Adv. Appl. Prob.* **22**, 25–48.
- [14] Martin-Löf, A. (1986). Symmetric sampling procedures, general epidemic processes and their threshold limit theorems. *J. Appl. Prob.* **23**, 265–282.
- [15] McKendrick, A.G. (1926). Applications of mathematics to medical problems. *Proc. Edin. Math. Soc.* **44**, 98–130.
- [16] Picard, P. and Lefèvre, C. (1990). A unified analysis of the final size and severity distribution in collective Reed-Frost epidemic processes. *Adv. Appl. Prob.* **22**, 269–294.

- [17] Sellke, T. (1983) On the asymptotic distribution of the size of a stochastic epidemic. *J. Appl. Prob.*, **20**, 390–394.
- [18] Scalia-Tomba, G. (1985). Asymptotic final-size distribution for some chain-binomial processes. *Adv. Appl. Prob.* **17**, 477–495.
- [19] van der Vaart, A.W. and Wellner, J.A. (1996) *Weak Convergence and Empirical Processes*. Springer.
- [20] Whittle, P. (1955) The outcome of a stochastic epidemic - a note on Bailey's paper. *Biometrika*, **42**, 116–122.