The effect of clumped population structure on the variability of spreading dynamics

Andrew J. Black[a,*], Thomas House[b,c], M. J. Keeling[b,c,d], J. V. Ross[a]

[a]*School of Mathematical Sciences, The University of Adelaide, Adelaide SA 5005, Australia.*
[b]*Mathematics Institute, Zeeman Building, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.*
[c]*Warwick Infectious Disease Epidemiology Research (WIDER) Centre, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.*
[d]*School of Life Sciences, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK.*

Abstract

Processes that spread through local contact, including outbreaks of infectious diseases, are inherently noisy, and are frequently observed to be far noisier than predicted by standard stochastic models that assume homogeneous mixing. One way to reproduce the observed levels of noise is to introduce significant individual-level heterogeneity with respect to infection processes, such that some individuals are expected to generate more secondary cases than others. Here we consider a population where individuals can be naturally aggregated into clumps (subpopulations) with stronger interaction within clumps than between them. This clumped structure induces significant increases in the noisiness of a spreading process, such as the transmission of infection, despite complete homogeneity at the individual level. Given the ubiquity of such clumped aggregations (such as homes, schools and workplaces for humans or farms for livestock) we suggest this as a plausible explanation for noisiness of many epidemic time series.

Keywords: epidemics, continuous-time Markov chain, offspring distribution, diffusion approximation

1. Introduction

Processes that spread between individuals in a population have, under various guises, been extensively studied and applied to many biological and physical problems (Boccaletti et al., 2006; Danon et al., 2011). It has been noticed for some time that many models of these processes do not capture the noisy behaviour of empirical data. For example many epidemic outbreaks are far noisier than predicted by simple models (Watts et al., 2005), suggesting that some element (or elements) are missing from these.

This enhanced variability has several important implications for the dynamics, especially during the early stages of invasion. Most notably, early extinctions will be far more common than anticipated from a birth-death process, and we may expect to observe several small outbreaks before the number of cases becomes large. This latter phenomenon has been observed in several real epidemics, as for example in the case of SARS in 2002–2003 (Anderson et al., 2004). Another possible explanation for such *stuttering* chains is due to a change in transmissibility due to evolutionary factors, such that the basic reproduction number, R_0, transitions from below one to above one (Lloyd-Smith et al., 2009). This phenomenon has been identified as an evident gap in modelling and a natural question arises of whether, in a given outbreak, stuttering is a consequence of evolutionary factors or simply a result of natural variability (Lloyd-Smith et al., 2009). Obviously, once many individuals are infected, the relative impact of the noise is reduced. Still, fully accounting for the variability is important in order to better inform decision making, for example via uncertainty analyses which capture the extent of probabilistic uncertainty surrounding the effectiveness of interventions (Black et al., 2013; Gilbert et al., 2014). Therefore understanding and capturing this larger-than-expected noise is of fundamental public-health importance in terms of understanding invasion, persistence and eradication of infection. Incorrect assumptions could substantially bias our parameter inferences (and hence predictions) from early outbreak data.

One plausible modification to simple stochastic epidemic models is to add individual-level heterogeneity to the population structure, such that individuals respond differently to the spreading process. In the context of infectious diseases this heterogeneity is often incorporated in terms of risk-structured populations where different risk groups have different rates of transmission (Andersson and Britton, 2000; Keeling and Rohani, 2007); at a more local scale, similar heterogeneity can also be derived within network-based models through varying degree distributions, with the presence of super-spreaders being a key example (Lloyd-Smith et al., 2005). For such network models, the early expected infection prevalence then grows at a rate proportional to the second moment of the degree distribution, while the early variance about this expected growth scales with the third moment of the degree distribution (Graham and House, 2013). It is therefore possible to generate enhanced variability while maintaining a given growth rate through judicious choice of the degree distribution. One objection to this method for generating increased variability is

*Corresponding author: School of Mathematical Sciences, The University of Adelaide, Adelaide SA 5005, Australia. Phone: +61883134177
Email address: andrew.black@adelaide.edu.au (Andrew J. Black)

that power-law like degree distributions that allow large third moments but modest second moments, are considered by some to be unrealistic representations of real social transmission networks (Clauset et al., 2009). Another objection is that this approach would necessitate a *fine tuning* of moments that would need to hold for *all* networks exhibiting significant noise in spreading dynamics.

A question of both theoretical and applied interest is whether greater variability can be replicated with a simpler and more robustly tuneable model. Here we rely on the natural aggregation (clumping) of individual hosts and consider a model with two-levels of mixing – within-clumps and between-clumps – but where all individuals are homogeneous; that is, all individuals are identical, but there are differential rates of transmission, high within a clump and low between clumps. Specifically, we consider a population made up of $N = m \times n$ individuals, in a total of m clumps each of size n, which we index with $i = 1, \ldots, m$. We consider a stochastic epidemic process on this population: clump i has S_i susceptibles and I_i infectives, such that $S_i + I_i \leq n$. The rates of this process are:

$$(S_i, I_i) \to (S_i - 1, I_i + 1), \text{ at rate } S_i \left(\frac{\beta I_i}{(n-1)} + \frac{\alpha}{N} \sum_{j=1}^{m} I_j \right);$$

$$(S_i, I_i) \to (S_i, I_i - 1)\ , \text{at rate } \gamma I_i, \qquad (1)$$

where β is the effective within-clump transmission rate parameter, α is the effective between-clump transmission rate parameter and γ is the per-infective recovery rate. Note that *frequency-dependent* transmission is used in Eq. (1). Another common choice is density-dependent transmission where β is not scaled by the factor $1/(n-1)$. We do not present results for this form of transmission as it allows for unrealistically large within-clump transmission, but we do discuss how our results are changed by this.

We wish to consider the limit as the number of patches, m, becomes large but where the patch size, n, remains finite and often relatively small. Note that for such a population structure, individuals are homogeneous, and the network topology (defined by within and between clump contact rates) also exhibits the desirable 'small worlds' property of high clustering and low shortest path lengths (Travers and Milgram, 1969; Watts and Strogatz, 1998). Many real-world populations exhibit this clumped population structure, where strong links exist within a clump and weaker links exist between clumps. High-profile examples include the transmission of avian influenza in poultry aggregated into sheds (Savill et al., 2006), spread of Foot-and-Mouth disease in livestock aggregated into farms (Tildesley et al., 2006), transmission of measles by children aggregated into schools (Riley et al., 1978), and the spread of a number of infections such as pandemic influenza through human populations that are aggregated into households (Black et al., 2013). For this type of clumped population, expressions for the probability of a major outbreak (successful invasion) and the final size of an epidemic (and hence results for percolation as a special case) have been derived (Ball et al., 1997).

Our aim is to assess the variability of spreading in this clumped population during the early phase of an epidemic before there is a significant depletion of susceptible clumps in the overall population. Within this early phase we can identify two dynamical regimes (see Figure 1). *(i) Initial behaviour:* Here the epidemic dynamics are well approximated by a branching process between clumps. *(ii) Early asymptotic behaviour:* After a significant number of between-household transmission events, the proportion of infected individuals becomes significant and the mean prevalence of infection, $\langle I(t) \rangle$, grows exponentially with fixed, *early growth* rate r,

$$\langle I(t) \rangle \propto \mathrm{e}^{rt}\ . \qquad (2)$$

Figure 1 shows stochastic simulations of this process to illustrate the dynamics of the two regimes. Figure 1(a) shows typical dynamics from a random-mixing model ($n = 1$) while Figure 1(b) shows dynamics from the clumped model ($n = 20$).

During the initial phase *(i)* clumping of the population mainly affects the variability in the timing of the start of the exponential growth phase, often giving rise to stuttering dynamics. Using the branching process approximation, we calculate a number of quantities which allows us to understand how the variability changes across clump sizes and within-clump transmission rates. To investigate the variability in stage *(ii)* we derive an analytic diffusion approximation to calculate the variance in the infectious process (Kurtz, 1970, 1971).

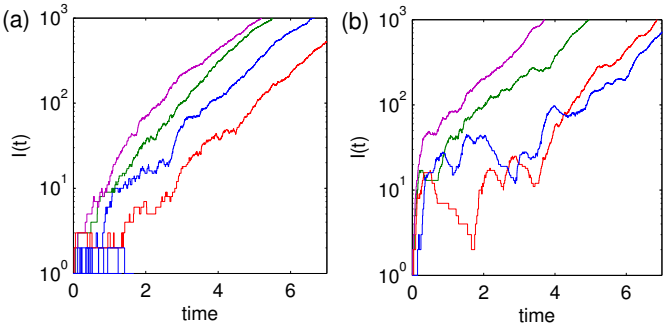

Figure 1: Stochastic simulations of the process Eq. (1) illustrating the early time behaviour and the two dynamical regimes with $n = 1$ (a) and $n = 20$ (b). The variability in the first phase mainly affects the timing of the second, exponential phase. Parameters: $m = 10^4, \beta = 20, r = 1$.

To maintain a fair comparison across clump sizes, n, and within-clump transmission rates, β, we calculate all quantities for the same early growth rate, which we fix arbitrarily as $r = 1$. For given values of β and n we scale the between-clump transmission rate, α, to achieve this. There is another quantity which we could instead fix: the clump reproductive ratio R_*, which is the expected number of secondary clumps infected by a primary clump (Ball et al., 1997; Ross et al., 2010). We note that these two quantities are linked (Svensson, 2007; Wallinga and Lipsitch, 2007) and we chose to fix r instead

of R_* because the early growth rate is most easily estimated from data. Throughout this paper we also fix the recovery rate at $\gamma = 1$, which can be done without loss of generality by rescaling time.

2. Initial behaviour

During this phase the epidemic dynamics are well approximated by a branching process between clumps (Ball et al., 1997; Ross et al., 2010). The within-clump dynamics are modelled as a continuous-time Markov chain, $X(t)$, with transition rate matrix $Q = (q_{ij}; i, j \in S)$. The transition rates are as in Eq. (1) with $\alpha = 0$ and $q_{ii} = -\sum_{j \in S} q_{ij}$. The state space is $S = A \cup C$, where C is an irreducible set of transient states (all possible states of a clump with $I > 0$ and $S + I \leq n$) and A is the set of absorbing states (corresponding to $I = 0$) within the clump. The function $\mathcal{I}(X(t))$ then gives the number of infected individuals within a clump at time t. New clumps are then infected according to a Poisson process with time-dependent rate $\alpha\mathcal{I}(X(t))$.

The growth rate of the process, r, is defined by the equation (Ball et al., 1997),

$$E\left[\int_0^\infty \alpha\mathcal{I}(X(t))e^{-rt}dt\right] = 1. \tag{3}$$

The left-hand-side of Eq. (3) can be efficiently evaluated numerically using exponential discounting (Ross et al., 2010). Combined with a basic root finding algorithm, this allows us to easily compute r. All results are derived by finding the value of α which gives $r = 1$ for a given clump size and within-clump transmission rate β. Figure 2 shows contours of constant r for different sized clumps. This extends to larger values of β than α as it is typically assumed that the within-clump transmission rate will be larger than the between-clump transmission rate.

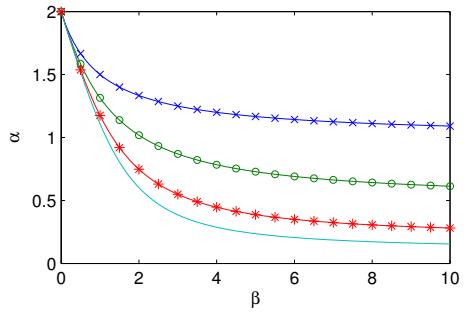

Figure 2: Values of α and β which give the same early growth rate, $r = 1$. Clump sizes are: 2, 4, 10 and 20 (crosses, circles, stars and plain line respectively).

By fixing the mean growth rate for all clump sizes and parameters, the mean number of infected individuals in the early stages of the epidemic is also fixed. The spreading dynamics will be an interplay of the within-clump and the between-clump dynamics and hence, although the mean growth rate remains fixed, there can be a large change in the variability of the process with different parameters. This variability in the initial stages of the branching process is crucial for understanding the variability in the numbers of patches infected and hence the probability of a major outbreak (or, disease extinction) and the timing of the start of the exponential growth phase. Intuitively, more variability will lead to a greater probability of extinction and a larger variability in the timing of the start of the exponential phase.

To understand how the variability of the spreading process changes with clump size and other parameters we calculate three main quantities. The first is the offspring distribution, denoted by $h(m)$. This is the probability distribution of the number of secondary clumps infected by a primary clump. This will be Poisson with a random mean,

$$h(m) = \int_0^\infty \rho(b)\frac{e^{-b}b^m}{m!}db, \tag{4}$$

where $\rho(b)$ is the probability density function corresponding to the random variable $\beta = \int \alpha\mathcal{I}(X(t))dt$, which is the total force of external infections created by a clump. Previous work has detailed how the offspring distribution can be efficiently calculated (Ross et al., 2010). The mean of the offspring distribution is then just R_*, the expected number of secondary clumps infected by a primary. Alternatively, R_* can be calculated more efficiently using path integral techniques, (Pollett and Stefanov, 2002; Ross et al., 2010)

$$R_* = E\left[\int_0^\infty \alpha\mathcal{I}(X(t))dt\right]. \tag{5}$$

The second quantity we calculate is the time to first infection of a secondary clump, conditioned on there being such an infection, which we denote by τ and henceforth refer to as the *conditional first infection time*. In Appendix A we detail how this can be calculated efficiently.

The last quantity we consider is the final size distribution within a clump. This is the probability distribution that a given proportion of the clump will have been infected over the course of the within-clump epidemic. This can be calculated via a number of methods (Ball, 1986; House et al., 2013).

2.1. Results

Figure 3 shows the mean and variance of the offspring distribution as a function of β and n; the mean is equal to R_*. This clearly displays non-monotonic behaviour indicating that the distribution changes in complex ways as the parameters are varied. Figure 4 shows the mean and variance of the conditional first infection time as a function of β and n. Once again, non-monotonic behaviour can be seen, but the mean time and in particular the variance do not show nearly as significant decrease with increasing β for fixed n; an increase in the clump size n, in particular across very small clump sizes, significantly increases both the mean and variance of

conditional first infection time. Finally, the mean and variance of the final size distribution, as a function of β and n, is shown in Figure 5. Here the mean increases monotonically with both n and β, whereas the variance is non-monotonic. Taken together, Figures 3, 4 and 5 show that there is a complex interplay between the spread of the disease between the clumps and the within-clump dynamics which changes significantly as the clump size and transmission rates are varied.

In deciphering the dynamics of this branching process it helps to consider the two different types of scaling that are possible. The most natural, and of primary interest for this paper, is keeping β fixed while increasing the clump size n. A basic consequence of this is that for larger n, less clumps contribute to the overall prevalence, hence if the within clump dynamics are noisy, so will the overall dynamics. Another scaling is to hold n fixed and vary β. This is a less natural situation as the extremes are somewhat unphysical, but it can aid our understanding of the dynamics.

As it is independent of α, the easiest aspect to understand is the within-clump dynamics and hence the final size distribution. The mean and variance are shown in Figure 5 along with the actual distributions for three values of β. Firstly, given that $\gamma = 1$, it is required that $\beta > 1$ for a large epidemic within the clump to occur. This sets a threshold for an epidemic, although we can see that even for $\beta < 1$ there is some probability of multiple infections within the patch. As β is increased, large epidemics become possible and the final size distribution becomes strongly bimodal. This can be seen clearly in the variance which is non-monotonic. Importantly, the time-scale of the within-clump epidemic decreases as β is increased. At the extreme, as $\beta \to \infty$ then the clump increasingly acts as one unit, such that after the initial infection the whole clump becomes infected very quickly (this fact is used to derive an approximation in the next section of the paper.) The variance then decreases as the probability of no further infection decreases.

Understanding the changes in the offspring and conditional first infection time distribution follow on from the within-clump dynamics. As larger within-clump epidemics become possible, then R_* also increases as a clump can potentially give rise to more external infections. The bi-modality in the final size distribution then accounts for the rise in the variance in the offspring distribution. For a large enough value of β, increasing n leads to a linear increase in the proportion of the clump that becomes infected. An important point here is, that because we hold r fixed, an increase in R_* must be offset by an increase in the time between clump infections, hence why Figure 4 shows an increase in the mean time to first infection. When $\beta \to \infty$ the mean time a patch is infectious for is controlled almost completely by the recovery process as the whole clump becomes infectious straight after initial infection.

Multiple components of the behaviour illustrated in Figures 3-5 drive the early stochastic dynamics. When β and n are

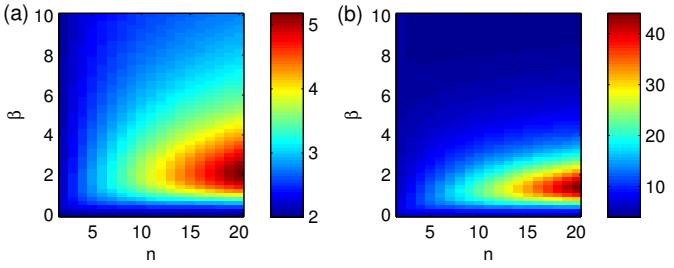

Figure 3: The mean (a) and variance (b) of the offspring distribution. The mean is equal to R_*.

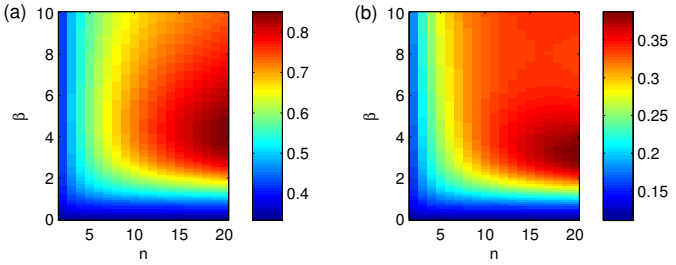

Figure 4: The mean (a) and variance (b) of τ, the conditional time to first infection.

large (and hence α is small to maintain $r = 1$, Figure 2), the within-clump dynamics are characterised by rapid spread that generally infects the vast majority of the clump within a short period of time (Figure 5). Yet the mean time to infect another clump is relatively long with high variance (Figure 4). Therefore we are likely to see saw-tooth aggregate dynamics, where the level of infection in one clump starts to decline before other clumps are infected. A secondary effect predominates at intermediate β (and high n), due to the high variance in within-clump final size (Figure 5) – which in turn generates high variance in both the offspring distribution and conditional time to first infection (Figures 3 and 4). These high variances combine to give long (and variable) periods between the seeding of infection and the onset of sustained exponential growth once a significant number of clumps are infected. Together these two elements of highly variable dynamics and rapid within-clump dynamics compared to between-clump transmission, explain the early behaviour observed in Figure 1(b).

3. Early asymptotic behaviour

After transmission has become established, but before there has been an appreciable depletion of susceptible clumps, the mean prevalence of infection grows exponentially. The branching process discussed in the previous section is still valid and would allow us to calculate the variance in the overall number of infected but the method is problematic to implement (Nerman, 1981; Ball and Donnelly, 1995). Instead, we investigate this phase of the epidemic by deriving a diffusion approximation for the process in the limit of the number of clumps $m \to \infty$

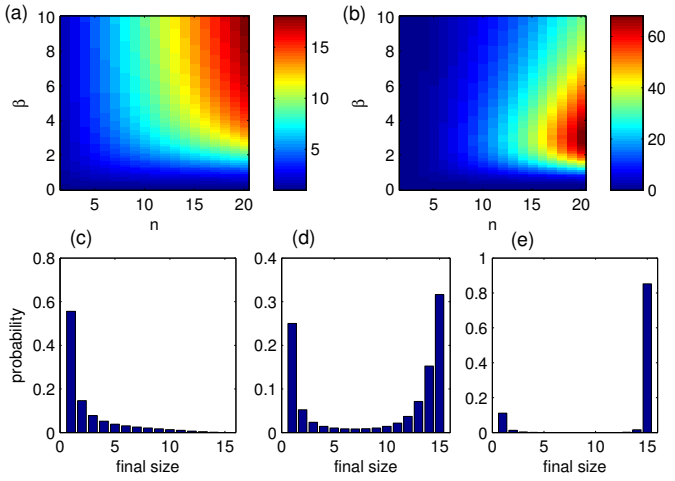

Figure 5: The mean (a) and variance (b) of the final size distribution for the within-clump epidemic. Parts (c), (d) and (e) show the final size distribution for a clump size of 15 with $\beta = 0.8$, 3 and 8 respectively. The variance tends to decrease as we increase β past this threshold as large outbreaks become more probable. For a large value of β increasing the clump size just increases the proportion who become infected.

(Kurtz, 1970, 1971). This is a perturbative expansion in m^{-1}, the inverse number of clumps, which allows us to approximate the full stochastic dynamics with a deterministic part, describing the mean behaviour, plus a stochastic correction (van Kampen, 1992; Black and McKane, 2011). From this we can calculate the variance in the total prevalence of infection. The details of this calculation are given in Appendix B, but are summarised here. Asymptotically, we find that

$$Var(I(t)) \to v e^{2rt}\,, \tag{6}$$

where v is independent of time. As we fix $r = 1$ for all calculations, the variation induced by the clumping comes entirely from the multiplicative factor v. This gives the size of the envelope around the mean in which typical stochastic realisations lie.

Of particular interest is how the variance (Eq. (6)) increases with clump size, n, for fixed values of within-clump transmission, β. In the limit $\beta \to \infty$ we can derive an expression for the variance by assuming that each between clump transmission leads to rapid (instantaneous) infection of the entire clump; this is therefore equivalent of a standard SIR model where each transmission event causes a change of magnitude n in the global susceptible and infected population sizes:

$$(S, I) \to (S - n, I + n)\,. \tag{7}$$

In this limiting case, and using the same steps as for the full model, we find that the factor v increases linearly with the clump size n,

$$v = \frac{n(r+\gamma)+\gamma}{r}. \tag{8}$$

Again, full details of this calculation are given in Appendix C.

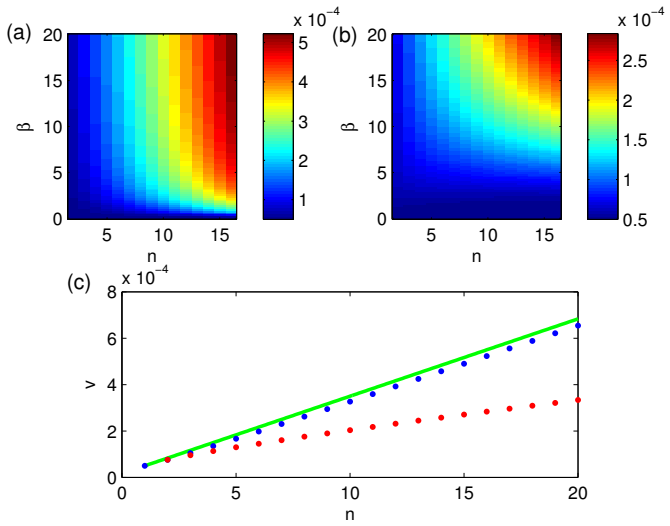

Figure 6: The asymptotic early value of $v = \langle I^2 \rangle e^{-2rt}$ as a function of the clump size, n, and the within-clump transmission parameter, β, assuming density-dependent transmission (a) and frequency-dependent transmission (b). Density dependent transmission assumes that the transmission rate β is *not* scaled by the factor $1/(n-1)$. Panel (c) shows v as a function of n for $\beta = 20$. The green line is the asymptotic result from Eq. (8). The blue and red points assume density- and frequency-dependent transmission respectively.

Figure 6 shows numerical results for v as a function of the clump size, n, and within-clump transmission rate. As we fix $r = 1$ throughout, then the mean overall level of infection grows at the same rate for all parameters. Here we consider both frequency-dependent transmission as in the previous section, and also density-dependent transmission, where β is not scaled by the factor $1/(n-1)$. Density-dependent transmission allows for much larger transmission rates within a clump and so is a better comparison with the limiting case, (8), where $\beta \to \infty$. Figure 6 shows that, as in the very early stages, changes in clump size alone can induce large differences in the variability of the spreading process. In contrast to the early stages, the variance increases monotonically with both clump size and within-clump transmission rate. Figure 6(c) shows how v changes with clump size for a fixed value of transmission parameter β, and also includes the theoretical limit from Eq. (8), which is a linear increase with n. For finite values of β the increase in variance is always sub-linear.

4. Discussion and Conclusion

We have shown that increased variability, in comparison to a homogeneous-mixing model, can be generated with a simple model of spreading, having two levels of mixing – within-clump and between-clump – but with all individuals homogeneous. In both the initial phase (*i*) and the early growth phase (*ii*), increasing the clump size, while keeping all other factors constant, increases the variance in the process. The enhanced variability arises in the early dynamics through the variability in the lag before the early growth phase begins, and hence gives rise to stuttering behaviour as seen in actual

epidemics (Anderson et al., 2004; Watts et al., 2005). Combined with partial detection of cases, this could be mistaken as stochastic fade-out followed by re-introductions of infection.

In Section 2 we presented results using what is know as frequency-dependent transmission where the transmission rate parameter, β, is scaled by $1/(n-1)$. Another common choice is density-dependent transmission where β is not scaled. In terms of our model, frequency-dependent transmission is the most realistic scenario, as within-clump transmission rates cannot become too large. Using density-dependent transmission means that large clumps have much stronger rates of infection, leading to larger / faster outbreaks, due to their increased number of susceptible individuals. In terms of our results in Section 2, the more extreme within-clump dynamics which results from assuming density dependent transmission means the window where we observe interesting population level dynamics is much reduced. In Section 3 we do present results using density-dependent transmission. This is primarily to compare with the analytic $\beta \to \infty$ limit, which is obviously easier to achieve when β is not scaled by $1/(n-1)$.

All of the methods used to analyse the dynamics in stage (i) can be naturally extended to more realistic models such as SE_2I_2R – that is the model considered here extended to have Erlang-2 distributed exposed and infectious periods, more clearly reflecting the shape of the true distributions – which has been used for pandemic influenza studies (Black et al., 2013). This involves using larger stochastic matrices, so there is a computational cost, but because most of our methods only rely on solving linear sets of equations, they scale efficiently.

Our model is clearly related to meta-population models, that have been considered previously as models of infection spread in aggregated populations (Lloyd and May, 1996; Riley and others, 2003; Rozhnova et al., 2012). The novelty in our work is that we consider a different population level limit – meta-populations often are considered as a small number of large clumps (for example representing cities in a country), where as we consider a large number of small clumps (more reminiscent of households within a country). Taking this limit of a large number of clumps means we can analyse the variability in the early growth phase (ii) using a diffusion approximation, but currently this does have limitations, most obviously in the range of clump sizes for which we can explicitly calculate the variance. This is due to numerical errors in calculating the eigenvectors of the Jacobian matrix, which in turn are used to expand a matrix exponential, Eq. (B.13). In practice, clump size $n = 16$ was the largest value we could use over the entire range of the other parameters. It is possible that a different approach to evaluating the coefficients of this expansion would allow us to go to higher values of n. It is possible to extend the diffusion approximation result to heterogeneous populations, but the usefulness of this is questionable. For example, the results will depend on the proportions of each size of clump and this has to be incorporated via careful choice of initial conditions.

To conclude, we have shown that a simple model, which involves a very generic structure and relatively little parameter tuning can reproduce real-world features as has only been demonstrated with highly heterogeneous approaches to date. It has also uncovered a number of interesting features which warrant further investigation.

Acknowledgements

This research was supported under the Australian Research Council's Discovery Projects funding scheme (project number DP110102893) (AJB & JVR) and the UK Engineering and Physical Sciences Research Council (TH & MJK). MJK and JVR also received support from the Royal Society (International Exchanges Scheme). We would like to thank the two anonymous referees who's comments have substantially improved this paper.

Appendix A. Initial behaviour methodology

Offspring distribution

The offspring distribution, denoted by $h_i(m)$, is the probability mass function of the number of secondary clumps infected by a primary clump conditional on starting in state i. This will be Poisson with a random mean,

$$h_i(m) = \int_0^\infty \rho(b) \frac{e^{-b} b^m}{m!} db, \tag{A.1}$$

where $\rho(b)$ is the probability density function corresponding to the random variable $\beta = \int \alpha \mathcal{I}(X(t)|X(0) = i)dt$, which is the total force of external infection created by a clump. We evaluate the offspring distribution as detailed in (Ross et al., 2010).

Conditional first infection time

Next we wish to calculate the mean and variance of τ, the conditional first time to infection. This uses basic theory of Markov processes (Waugh, 1958; Norris, 1997).

First, we augment the original Markov chain, representing the within-clump dynamics, with a third variable, a, which counts the number of external infections caused by the clump. The transition rates for this new chain are then,

$$\begin{aligned} (S, I, a) &\to (S-1, I+1, a), \quad \text{at rate} \quad \frac{\beta S I}{(n-1)} \\ (S, I, a) &\to (S, I, a+1), \quad \text{at rate} \quad \alpha I \\ (S, I, a) &\to (S, I-1, a), \quad \text{at rate} \quad \gamma I \end{aligned} \tag{A.2}$$

Next we condition the Q matrix on observing at least 1 external infection, i.e. $a > 0$. This involves modifying the elements of the transition matrix Q (Waugh, 1958):

$$\bar{q}_{ij} = \left(\frac{u_j}{u_i}\right) q_{ij}, \tag{A.3}$$

for all states i from which $u_i > 0$ and $\bar{q}_{ij} = q_{ij}$ otherwise, where u_i is the probability of at least 1 external infection given that

the process starts from state i. These probabilities correspond to the complementary probability of no further infections, available from the offspring distribution, as detailed earlier, starting from each state i; that is, $u_i = 1 - h_i(0)$. Finally, we make the states of the system corresponding to $a = 1$ absorbing by setting the relevant elements of $\bar{q}_{ij}$ equal to zero. The mean of the conditional first infection time starting from state j, $\langle\tau\rangle_j$, is then found by solving a set of linear equations (Norris, 1997),

$$\sum_{j\in C} \bar{q}_{ij}\langle\tau\rangle_j = -1. \tag{A.4}$$

The second moment can be found from,

$$\sum_{j\in C} \bar{q}_{ij}\langle\tau^2\rangle_j = -2\langle\tau\rangle_i, \tag{A.5}$$

which allows us to calculate the variance.

Appendix B. Diffusion approximation for exponential growth phase

Before we go into the details of this calculation, we give a brief outline of the various steps. The first step is to show that the stochastic process is of the correct form, so that we can derive a diffusion approximation in the limit that the number of clumps, $m \to \infty$. Once this is done we then apply the approximation to our system, giving equations for the mean and variance of the *proportions* of the clumps of each type. Finally, we describe how these equations can be approximately solved when we consider just the early time dynamics. This is valid in the period of time between when the exponential growth starts and before the peak in the epidemic.

Diffusion approximation

We define $H_{x,y}(t)$ as the number of clumps with x susceptible and y infected at time t. The state of the system is then defined as $\mathbf{H}(t) = \{H_{x,y}(t)|x + y \leq n\}$ where n is the size of the clumps. This is equivalent to the original process, but reduces the number of variables in the problem from $2m$ to $(n + 1)(n + 2)/2$. In this representation an infection or recovery event results in a clump in the current configuration being replaced by a clump with the updated configuration. The total number of infected individuals across all clumps is given by $Y(t) = \mathbf{y} \cdot \mathbf{H}(t)$, where $\mathbf{y}$ gives the number of infected in each clump configuration. The transition rates for this new process are then,

$$\begin{aligned}(H_{x,y}, H_{x,y-1}) &\to (H_{x,y} - 1, H_{x,y-1} + 1)\\ &\text{at rate } \gamma y H_{x,y}\ ,\\ (H_{x,y}, H_{x-1,y+1}) &\to (H_{x,y} - 1, H_{x-1,y+1} + 1)\\ &\text{at rate } xH_{x,y}\left(\beta y + \alpha Y(t)\right)\ .\end{aligned} \tag{B.1}$$

We next define the proportion of clumps in each configuration, $\phi_{x,y}(t) = m^{-1}H_{x,y}(t)$, and the overall proportion infected, $I(t) = (nm)^{-1}Y(t)$. From herein we drop the dependence on time of the $H_{x,y}$ and $\phi_{x,y}$ for clarity. The two transmission rates can then be written as

$$\begin{aligned}\gamma y H_{x,y} &= my\gamma\phi_{x,y}\ ,\\ xH_{x,y}\left(\beta y + \alpha Y(t)\right) &= mx\phi_{x,y}\left(\beta y + \alpha I(t)\right)\ .\end{aligned} \tag{B.2}$$

These are of the correct form for a density-dependent Markov chain, and the results of Kurtz give a simple procedure for deriving the diffusion approximation (Kurtz, 1970, 1971). The same results can be obtained from a linear-noise approximation or a WKB approximation (van Kampen, 1992; Black and McKane, 2011).

To leading order in the diffusion approximation, the proportions of clumps in each configuration is given by a set of ODEs,

$$\dot{\boldsymbol{\phi}} = \mathbf{F} = \sum_{\mu=1,2} l_\mu \mathbf{w}_\mu. \tag{B.3}$$

where $(\mathbf{w}_1)_i = \gamma y_i\phi_i$ and $(\mathbf{w}_2)_i = (\alpha I(t) + \beta y_i)x_i\phi_i$ The two matrices l_μ encode all the recovery ($\mu = 1$) and infection ($\mu = 2$) events respectively. The elements of these are

$$\begin{aligned}(l_1)_{ij} &= -\delta_{y_i,y_j} + \delta_{x_i,x_j}\delta_{y_i-1,y_j}, \quad y_{i,j} > 0,\\ (l_2)_{ij} &= -\delta_{x_i,x_j} + \delta_{x_i-1,x_j}\delta_{y_i+1,y_j}, \quad x_{i,j} > 0.\end{aligned} \tag{B.4}$$

Each column of the two matrices l_1 and l_2 represents the changes in the number of households of each type caused by a given event. Some columns will be all zero where no events can take place. Writing Eq. (B.3) in terms of its components we find,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\phi_{x,y} = \gamma&\left(-y\phi_{x,y} + (y+1)\phi_{x,y+1}\right)\\ &+ \beta\left(-xy\phi_{x,y} + (x+1)(y-1)\phi_{x+1,y-1}\right)\\ &+ \alpha I(t)\left(-x\phi_{x,y} + (x+1)\phi_{x+1,y-1}\right)\ .\end{aligned} \tag{B.5}$$

This rigorously establishes a number of results from so called 'self-consistent' methods (Ghoshal et al., 2004; House and Keeling, 2008).

The next order in the diffusion approximation quantifies the stochastic fluctuations around the deterministic trajectory given by Eq. (B.3). Firstly, the Jacobian of the system evaluated about the time-dependent trajectory given by (B.3) is,

$$B(t) = \boldsymbol{\nabla}\mathbf{F}(t). \tag{B.6}$$

The time-varying covariance matrix of this process can then be expressed as (Kurtz, 1970, 1971)

$$\Sigma^2(t) = M(t)\int_0^t M^{-1}(s)G(s)(M^{-1}(s))^{\mathrm{T}}ds\,(M(t))^{\mathrm{T}}, \tag{B.7}$$

where

$$M(t) = \exp\left(\int_0^t B(s)ds\right), \tag{B.8}$$

and

$$G(t) = \sum_{\mu=1,2} l_\mu\,\mathrm{diag}(\mathbf{w}_\mu)\,l_\mu^{\mathrm{T}} \tag{B.9}$$

is the local covariance matrix, where $\mathrm{diag}(\mathbf{w}_\mu)$ is a matrix with the elements of $\mathbf{w}_\mu$ along the diagonal.

Early time approximation

In the early growth phase we can make a number of approximations to make the equations above tractable, yielding insight into the dynamics of the system. As with simpler epidemic models with no clump structure, Eq. (B.3) has an unstable fixed point, $\boldsymbol{\phi}^* = \delta_{n,x}\delta_{0,y}$, which corresponds to all clumps being susceptible. Linearising Eq. (B.3) about this fixed point gives an approximate solution

$$\boldsymbol{\phi}(t) = \delta_{n,x}\delta_{0,y} + \hat{\boldsymbol{\phi}}\epsilon e^{rt}, \tag{B.10}$$

where $\hat{\boldsymbol{\phi}}$ is the eigenvector of the dominant eigenvalue, r, of the Jacobian, evaluated at the unstable fixed point and $\epsilon \ll 1$ is a small perturbation away from this. This approximation is accurate while the proportion of susceptible clumps remains high, i.e. before the peak in the epidemic. Substituting this solution into the equations for the Jacobian and the local covariance matrix, we find that the Jacobian is constant, i.e. $B(t) = \hat{B} + O(\epsilon)$ while

$$G(t) = \epsilon\hat{G}e^{rt} + O(\epsilon^2). \tag{B.11}$$

The matrix $\hat{G}$ is found by substituting the initial condition, $\boldsymbol{\phi}(0) = \delta_{n,x}\delta_{0,y} + \hat{\boldsymbol{\phi}}\epsilon$, into equation Eq. (B.9) and retaining only terms of order ϵ, i.e.,

$$\hat{G} = \left.\frac{dG(0)}{d\epsilon}\right|_{\epsilon=0}. \tag{B.12}$$

As the Jacobian is constant then Eq. (B.8) becomes $M(t) = e^{\hat{B}t}$. The eigenvalues of $\hat{B}$ are distinct, apart from $n+1$ repeated zeros. However, the zero eigenvalues can be made to have distinct eigenvectors so we can do a spectral decomposition and write the matrix exponential as

$$M(t) = e^{\hat{B}t} = \sum W_i e^{\kappa_i t}, \tag{B.13}$$

where the matricies W_i need to be determined and i runs over the eigenvalues of the problem. The inverse, which is needed for the integral in Eq. (B.7), is simply

$$M^{-1}(t) = e^{-\hat{B}t} = \sum W_i e^{-\kappa_i t}. \tag{B.14}$$

The matrices, W_i, are constructed from the eigenvectors of $\hat{B}$. Specifically, $W_i = v_i y_i^T$, where v_j is the right eigenvector corresponding to the jth eigenvalue and y_j^T is the jth row of V^{-1} where V is the matrix whose columns are the right eigenvectors of $\hat{B}$ (Moler and Van Loan, 2003). It then follows from the orthogonality of the eigenvectors that $W_i W_j = 0$ for $i \neq j$. This allows us to simplify Eq. (B.7) considerably, as many of the terms are zero. We find,

$$\Sigma^2(t) = \epsilon \sum_{i,j} W_i^2 \hat{G}(W_j^2)^T e^{(\kappa_i+\kappa_j)t} \int_0^t e^{-(\kappa_i+\kappa_j-r)s} ds. \tag{B.15}$$

This is the most general expression we can give, as for some i and j the eigenvalues in the integrand sum to zero and for $n > 2$ the eigenvalues can also be complex. As r is the only positive eigenvalue then it is easy to find the fastest growing part, which gives the asymptotic solution,

$$\Sigma_F^2(t) = W_F^2 G(W_F^2)^T e^{2rt}/r \tag{B.16}$$

where W_F is the matrix in expansion (B.13) corresponding to the dominant eigenvalue.

Defining the matrix $\Pi = \mathbf{y}\cdot\mathbf{y}^{\mathrm{T}}/n^2$, the variance in the overall number of infectives is $Var(I(t)) = \sum_{i,j}\Pi_{ij}\Sigma_{ij}^2(t)$. Asymptotically

$$Var(I(t)) \to v\mathrm{e}^{2rt}, \tag{B.17}$$

where v is independent of time. As we fix $r = 1$ for all calculations, the variation induced by the clumping comes entirely from the multiplicative factor v.

Appendix C. Large β limit

In the limit $\beta \to \infty$ we can derive an equation for the variance via a modified stochastic process. In this limit the entire clump becomes infected straight after the initial infection, so we can approximate this process by a simple SIR model with transitions,

$$\begin{aligned}(S,I) &\to (S-n, I+n), \text{ at rate } \frac{\alpha S I}{m};\\ (S,I) &\to (S, I-1), \text{ at rate } \gamma I,\end{aligned} \tag{C.1}$$

where n the clump size. The deterministic approximation ($m \to \infty$) of this process is

$$\begin{aligned}\dot{x} &= -n\alpha xy,\\ \dot{y} &= n\alpha xy - \gamma y.\end{aligned} \tag{C.2}$$

where $x = S/m$ and $y = I/m$. Linearising about the unstable fixed point $(x, y) = (1, 0)$ we find $\dot{y} = (n\alpha - \gamma)y$, thus the early growth rate is $r = n\alpha - \gamma$. The Jacobian evaluated at the fixed point is,

$$B = \begin{pmatrix} 0 & n\alpha \\ 0 & n\alpha - \gamma \end{pmatrix}. \tag{C.3}$$

The matrix exponential is then straightforward to calculate,

$$e^{Bt} = \begin{pmatrix} 1 & n\alpha \\ 0 & 0 \end{pmatrix} + \begin{pmatrix} 0 & -n\alpha \\ 0 & 1 \end{pmatrix} e^{rt} = W_0 + W_1 e^{rt}. \tag{C.4}$$

The local covariance matrix is,

$$G(x(t), y(t)) = \begin{pmatrix} n^2\alpha x(t)y(t) & -n^2\alpha x(t)y(t) \\ -n^2\alpha x(t)y(t) & n^2\alpha x(t)y(t) + \gamma y(t) \end{pmatrix}. \tag{C.5}$$

Substituting in the early time solutions we can write this as

$$G(t) = \begin{pmatrix} n^2\alpha & -n^2\alpha \\ -n^2\alpha & n^2\alpha + \gamma \end{pmatrix} I(0)e^{rt} = \hat{G}I(0)e^{rt} \tag{C.6}$$

where $I(0)$ is the initial proportion infected. Using Eq. (B.16), from the full household model, the fastest growing part of the covariance matrix is given by

$$\Sigma_F^2(t) = W_1^2 \hat{G}(W_1^2)^T I(0)e^{2rt}/r. \tag{C.7}$$

Carrying out the matrix multiplications we find the part corresponding to the variance in the number infected,

$$\Sigma_I^2(t) = \frac{n^2\alpha + \gamma}{n\alpha - \gamma} I(0) e^{2(n\alpha-\gamma)t}. \quad \text{(C.8)}$$

Finally we make the substitution $\alpha = \alpha'/n$ where α' is the transmission rate when $n = 1$, which is fixed for a given r and γ, i.e. $\alpha' = r + \gamma$. This gives,

$$\Sigma_I^2(t) = \frac{n(r+\gamma) + \gamma}{r} I(0) e^{2rt}. \quad \text{(C.9)}$$

Thus the factor multiplying the exponential is linear in n. Setting $n = 1$, we recover the basic SIR result in Dangerfield et al. (2009).

References

Anderson, R. M., Fraser, C., Ghani, A. C., Donnelly, C. A., Riley, S., Ferguson, N. M., Leung, G. M., Lam, T., Hedley, A. J., 2004. Epidemiology, transmission dynamics and control of SARS: the 2002-2003 epidemic. Phil. Trans. R. Soc. Lond. B 359, 1091–1105.

Andersson, H., Britton, T., 2000. Stochastic Epidemic Models and Their Statistical Analysis. Vol. 151 of Springer Lectures Notes in Statistics. Springer, Berlin.

Ball, F., 1986. A unified approach to the distribtuion of total size and total area under the trajectory of infectives in epidemic models. Adv. App. Prob. 18, 289–310.

Ball, F., Donnelly, P., 1995. Strong approximations for epidemic models. Stoch. Proc. Appl 55, 1–21.

Ball, F., Mollison, D., Scalia-Tomba, G., 1997. Epidemics with two levels of mixing. Ann. App. Prob. 7 (1), 46–89.

Black, A. J., House, T., Keeling, M. J., Ross, J. V., 2013. Epidemiological consequences of household-based antiviral prophylaxis for pandemic influenza. J. R. Soc. Interface 10, 20121019.

Black, A. J., McKane, A. J., 2011. WKB calculation of an epidemic outbreak distribution. J. Stat. Mech. 12, P12006.

Boccaletti, S., Latora, V., Moreno, Y., Chavez, M., Hwang, D., 2006. Complex networks: Structure and dynamics. Physics Reports 424 (4-5), 175–308.

Clauset, A., Shalizi, C. R., Newman, M. E. J., 2009. Power-law distributions in empirical data. SIAM Review 51 (4), 661–703.

Dangerfield, C. E., Ross, J. V., Keeling, M. J., 2009. Integrating stochasticity and network structure into an epidemic model. J. R. Soc. Interface 6, 761–774.

Danon, L., Ford, A. P., House, T., Jewell, C. P., Keeling, M. J., Roberts, G. O., Ross, J. V., Vernon, M. C., 2011. Networks and the epidemiology of infectious disease. Interdisciplinary Perspectives on Infectious Diseases 2011, 1–28.

Ghoshal, G., Sander, L. M., Sokolov, I. M., 2004. SIS epidemics with household structure: the self-consistent field method. Math. Biosci. 190, 71–85.

Gilbert, J. A., Meyers, L. A., Galvani, A. P., Townsend, J. P., 2014. Probabilistic uncertainty analysis of epidemiological modeling to guide public health intervention policy. Epidemics, in publication.

Graham, M., House, T., 2013. Dynamics of stochastic epidemics on heterogeneous networks. J. Math. Bio., to appear.

House, T., Keeling, M. J., 2008. Deterministic epidemic models with explicit household structure. Math. Biosci. 213, 29–39.

House, T., Ross, J. V., Sirl, D., 2013. How big is an outbreak likely to be? methods for epidemic final size calculation. Proc. R. Soc. A 469, 20120436.

Keeling, M. J., Rohani, P., 2007. Modeling Infectious Diseases in Humans and Animals. Princeton University Press, New Jersey.

Kurtz, T., 1970. Solutions of ordinary differential equations as limits of pure jump Markov processes. J. Appl. Probab. 7 (1), 49–58.

Kurtz, T., 1971. Limit theorems for sequences of jump Markov processes approximating ordinary differential processes. J. Appl. Probab. 8 (2), 344–356.

Lloyd, A. L., May, R. M., 1996. Spatial heterogeneity in epidemic models. J. Theor. Biol. 179, 1–11.

Lloyd-Smith, J. O., George, D., Pepin, K. M., Pitzer, V. E., Pulliam, J. R., Dobson, A. P., Hudson, P. J., Grenfell, B. T., 2009. Epidemic dynamics at the human-animal interface. Science 326, 1362–1367.

Lloyd-Smith, J. O., Schreiber, S. J., Kopp, P. E., Getz, W. M., 2005. Superspreading and the effect of individual variation on disease emergence. Nature 438, 255–259.

Moler, C., Van Loan, C., 2003. Nineteen dubious ways to compute the exponential of a matrix, twenty-five years later. SIAM Review 20, 801–836.

Nerman, O., 1981. On the convergence of supercritical general (C-M-J) branching processes. Z. Wahrscheinlichkeitsch 57, 365–395.

Norris, J. R., 1997. Markov chains. Cambridge University Press, Cambridge.

Pollett, P., Stefanov, V., 2002. Path integrals for continuous-time Markov chains. J. Appl. Probab. 39, 901–904.

Riley, E. C., Murphy, G., Riley, R. L., 1978. Airborne spread of measles in a suburban elementray school. Am. J. Epidemiol. 107, 421–432.

Riley, S., et al., 2003. Transmission dynamics of the etiological agent of SARS in Hong Kong: impact of public health interventions. Science 300, 1961–1966.

Ross, J. V., House, T., Keeling, M. J., 2010. Calculation of disease dynamics in a population of households. PLoS ONE 5, e9666.

Rozhnova, G., Nunes, A., J., M. A., 2012. Phase lag in epidemics on a network of cities. Phys. Rev. E 85, 051912.

Savill, N. J., St Rose, S. G., Keeling, M. J., Woolhouse, M. E. J., 2006. Silent spread of H5N1 in vaccinated poultry. Nature 442, 757.

Svensson, A., 2007. A note on generation times in epidemic models. Mathematical Biosciences 208, 300–311.

Tildesley, M. J., Savill, N. J., Shaw, D. J., Deardon, R., Brooks, S. P., Woolhouse, M. E. J., Grenfell, B. T., Keeling, M. J., 2006. Optimal reactive vaccination strategies for an outbreak of foot-and-mouth disease in great britain. Nature 440, 83–86.

Travers, J., Milgram, S., 1969. An experimental study of the small world problem. Sociometry 32 (4), 425–443.

van Kampen, N. G., 1992. Stochastic processes in physics and chemistry. Elsevier, Amsterdam.

Wallinga, J., Lipsitch, M., 2007. How generation intervals shape the relationship between growth rates and reproductive numbers. Proc. R. Soc. B 274, 599–604.

Watts, D. J., Muhamad, R., Medina, D. C., Dodds, P. S., 2005. Multiscale, resurgent epidemics in a hierarchical metapopulation model. Proc. Natl. Acad. Sci. USA 102 (32), 11157–11162.

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

Waugh, W. A. O., 1958. Conditioned Markov processes. Biometrika 45, 241–249.