JOURNAL OF THE ROYAL SOCIETY Interface

J. R. Soc. Interface (2009) **6**, 1121 –1134
doi:10.1098/rsif.2008.0524
Published online 4 March 2009

Spread of infectious disease through clustered populations

Joel C. Miller[1,2,*]

[1]*University of British Columbia Centre for Disease Control, Vancouver, British Columbia, V5Z 4R4, Canada*
[2]*Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA*

Networks of person-to-person contacts form the substrate along which infectious diseases spread. Most network-based studies of this spread focus on the impact of variations in degree (the number of contacts an individual has). However, other effects such as clustering, variations in infectiousness or susceptibility, or variations in closeness of contacts may play a significant role. We develop analytic techniques to predict how these effects alter the growth rate, probability and size of epidemics, and validate the predictions with a realistic social network. We find that (for a given degree distribution and average transmissibility) clustering is the dominant factor controlling the growth rate, heterogeneity in infectiousness is the dominant factor controlling the probability of an epidemic and heterogeneity in susceptibility is the dominant factor controlling the size of an epidemic. Edge weights (measuring closeness or duration of contacts) have impact only if correlations exist between different edges. Combined, these effects can play a minor role in reinforcing one another, with the impact of clustering the largest when the population is maximally heterogeneous or if the closer contacts are also strongly clustered. Our most significant contribution is a systematic way to address clustering in infectious disease models, and our results have a number of implications for the design of interventions.

Keywords: epidemic; clustering; reproductive ratio; epidemic probability; attack rate

1. INTRODUCTION

Recently, H5N1 avian influenza and SARS have raised the profile of emerging infectious diseases. Both can infect humans, but have a primary animal host. Typically, such zoonotic diseases emerge periodically into the human population and disappear (e.g. Ebola, hantavirus, rabies), but sometimes (e.g. HIV) the disease achieves sustained person-to-person spread. With the advent of modern transportation networks, diseases that formerly emerged in isolated villages and died out without further spread may now spread worldwide.

A number of interventions are available to control emerging diseases, each with distinct costs and benefits. To design optimal policies, we must address several related, but nevertheless distinct, questions. How fast would an epidemic spread? How likely is a single introduced infection to result in an epidemic? How many people would an epidemic infect? We quantify these using $\mathcal{R}_0$, the *basic reproductive ratio*, which measures the average number of new cases each infection causes early in the outbreak; $\mathcal{P}$, the probability that a single infection sparks an epidemic; and $\mathcal{A}$, the *attack rate* or fraction of the population infected in an epidemic. Understanding these different quantities and what affects them helps us to select policies with maximal impact for given cost.

Many different models are used to study disease spread. Perhaps the most important decision in developing a model is how the interactions of the population are represented. Owing to the complexity of the population, it is invariably necessary to make simplifying assumptions. The errors (and therefore the conclusions) resulting from many of these approximations are not well quantified. In this paper, we will focus on quantifying the impact of clustering (the tendency to interact in small groups) and individual-scale heterogeneity on the spread of an epidemic.

Based on how they handle clustering, models for population structure fit into a hierarchy of three classes (which in turn may be subdivided). At the simplest level, the population is assumed to mix without any clustering. Most existing models fall into this category. At the most complex level, agent-based models are used: the movements of each individual are tracked and people who are in the same location are able to infect one another. These models typically require significant resources to develop, and the clustering is explicitly included. An intermediate level of complexity attempts to introduce the clustering as a parameter (or several parameters). Usually these models consider clustering

*Address for correspondence: Harvard School of Public Health, 677 Huntington Avenue, Boston, MA 02115, USA (joel.c.miller.research@gmail.com).

Electronic supplementary material is available at http://dx.doi.org/10.1098/rsif.2008.0524 or via http://rsif.royalsocietypublishing.org.

Received 9 December 2008
Accepted 30 December 2008

only in terms of the number of triangles in a network, but as we shall see, other structures may play a role.

Before introducing the details of our model, we review some previous work. All the models we consider are susceptible–infected–recovered (SIR) epidemic models (Anderson & May 1991), in which individuals begin susceptible, become infected by contacting infected individuals and finally recover with immunity.

For unclustered populations, ordinary differential equation (ODE) models were among the earliest models used (Kermack & McKendrick 1927) and remain the most common. They are deterministic, and so cannot directly calculate $\mathcal{P}$, but they give insights into the factors controlling $\mathcal{R}_0$ and $\mathcal{A}$. Because they assume mass-action mixing, it is difficult to incorporate individual heterogeneity in the number of contacts. More recently, some network-based models have been introduced for unclustered populations (Andersson 1998; Newman 2002; Meyers *et al.* 2005, 2006; Kenah & Robins 2007; Miller 2007; Meyers 2007). These models represent the population as nodes with edges between nodes representing contacts, along which disease spreads stochastically. Heterogeneity in the number of contacts is introduced by modifying the degree (number of edges) of each node. By neglecting clustering, these studies are able to make analytic predictions through branching process arguments. A recent sociological study (Mossong *et al.* 2008) has used surveys with participants recording the length and nature of their contacts. These data are valuable for providing the contact distribution needed for the above network models, and allow us to apply network results to real populations. However, these data do not directly tell us anything about the clustering of the population resulting from family/work/other groups. Other recent work by Kenah & Robins (2007) and Miller (2007) analytically addresses the impact of heterogeneity in infectiousness and susceptibility in unclustered networks.

Using agent-based simulations (Eubank *et al.* 2004; Barrett *et al.* 2005; Ferguson *et al.* 2005; Del Valle *et al.* 2006; Germann *et al.* 2006; Ajelli & Merler 2008) allows us to directly incorporate clustering. In these simulations, the population is a collection of individuals who move and contact one another. The modeller has complete control over the parameters governing interactions and how the disease spreads. This allows us to study many effects, but introduces many parameters. It is difficult to test the accuracy of the assumptions used to generate these models and to extract which parameters are essential to the disease dynamics. The expense of developing these simulations is frequently prohibitive.

In this paper, we introduce a systematic approach for calculating the impact of clustering and quantifying the error. Because our model investigates disease spread in clustered networks, we provide a more detailed review of previous work on clustering and disease. A few investigations have been made into the interaction of clustering with disease spread using network models. The attempts that have been made (Keeling 1999; Newman 2003*a*; Serrano & Boguñá 2006*a*,*b*; Britton *et al.* 2007; Eames 2008) typically use approximations whose errors are not quantified, resulting in apparently contradictory results. A few papers (Kuulasmaa 1982; Trapman 2007; Miller 2008) have considered clustering and heterogeneities, rigorously showing that increased heterogeneity tends to decrease $\mathcal{P}$ and $\mathcal{A}$, but without quantitative predictions. Recently, Eames (2008) has considered the spread of epidemics in a class of random networks for which the number of triangles could be controlled. It may be inferred from his fig. 3 that clustering decreases the growth rate and that sufficient clustering can increase the epidemic threshold. However, at small and moderate levels, clustering appears not to alter the final size of epidemics significantly. Similar observations have been made by Bansal (2008). At first glance, this contradicts the observations of Serrano & Boguñá (2006*a*,*b*) that clustering significantly reduces the size of epidemics, but that sufficiently strong clustering reduces the epidemic threshold (see also Newman 2003*a*), allowing epidemics at lower transmissibility. The discrepancy in epidemic size may be resolved by noting that the networks in Serrano & Boguñá (2006*a*,*b*) have low average degree. We will see that clustering affects the size only if the typical degree is small or clustering is very high. The apparent discrepancy in epidemic threshold with strong clustering may be resolved by noting that the form of strong clustering considered by Serrano & Boguñá (2006*a*,*b*) forces preferential contacts between high-degree nodes. The reduction in epidemic threshold is perhaps better understood in terms of degree–degree correlations than in terms of clustering.

In this paper, we develop techniques to incorporate general small-scale structure (beyond triangles) into the calculations of $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$. To calculate $\mathcal{R}_0$, we develop a systematic series expansion that allows us to interpolate between unclustered and clustered results by including more terms. To calculate $\mathcal{P}$ and $\mathcal{A}$, we use a similar approach, but give only the estimates on the size of correction terms. Our methods give us a rigorous means to understand how the unclustered results relate to more realistic populations, and our results resolve the apparent discrepancies mentioned above. Our theory accurately predicts epidemic behaviour in a more realistic contact network derived from an agent-based simulation of Portland, Oregon, by EpiSimS (Del Valle *et al.* 2006). We expand this to investigate the interplay of clustering, heterogeneities in individual infectiousness or susceptibility, and variations in edge weights in their effects on $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$.

The paper is organized as follows: §2 describes our model and networks and summarizes earlier work on unclustered networks. These results will be the leading-order terms for our expansions for clustered networks in the remainder of the paper. Section 3 considers how epidemics spread in a clustered network assuming homogeneous transmission. We derive the corrections to $\mathcal{R}_0$ and show that the corrections to $\mathcal{P}$ and $\mathcal{A}$ are insignificant unless the typical degree is small or clustering very high. Section 4 considers epidemics in clustered networks with heterogeneous infectiousness or susceptibility, building on §3. Section 5 extends this further to consider epidemics spreading on clustered networks with weighted edges. Edges with large weights tend to occur in family or work groups, which

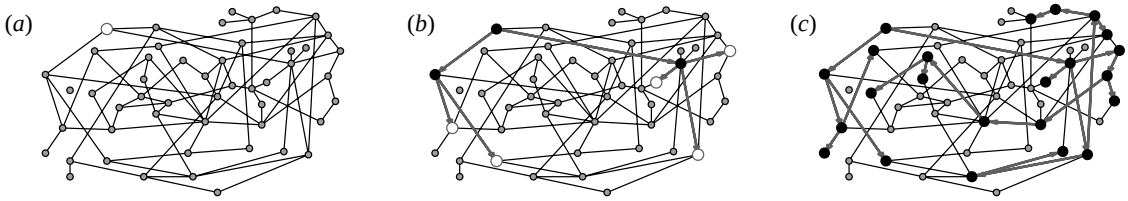

Figure 1. A sample network and several stages of an outbreak. Nodes begin susceptible (small circles), become infected (large open circles), possibly infecting others along edges, and then recover (large filled circles). The outbreak finishes when no infected nodes remain.

magnifies the impact of clustering. Finally, §6 discusses the implications of our results, particularly for designing interventions. We conclude that, in general, heterogeneity significantly impacts $\mathcal{P}$ and $\mathcal{A}$, but not $\mathcal{R}_0$, while clustering impacts $\mathcal{R}_0$ significantly, but not $\mathcal{P}$ and $\mathcal{A}$. Heterogeneity or edge weights may enhance the impact of clustering.

2. FORMULATION

2.1. The disease model

We consider the spread of a disease using a discrete SIR model on a static network G. Nodes of G represent individuals and edges represent (potentially infectious) contacts. The contact structure of the network is fixed during the course of the outbreak. The *degree* k of a node u is the number of edges containing u. Figure 1 shows a sample outbreak. For a single infection, the *index case* is chosen uniformly from the population to begin an *outbreak*. Infection spreads along an edge from an infected node u to a susceptible node v with probability T_{uv}, the *transmissibility*. The time it takes for infection and recovery to occur may vary but does not affect our results. Once u recovers it cannot be reinfected. Typically, for a large random network with a population of $N{=}|G|$ nodes, the final size of outbreaks is either large, with $\mathcal{O}(N)$ cumulative infections, or small, with $\mathcal{O}(\log N)$ infections (Bollobás 2001). Large outbreaks are *epidemics* and small outbreaks are *non-epidemics*.

2.1.1. Transmissibility. A number of factors influence the transmissibility from u to v such as the viral load and duration of infection of u, the vaccination history and general health of v, the duration and nature of the contact between u and v and characteristics of the disease.

For each node u, we denote its ability to infect others by $\mathcal{I}_u$ and its ability to be infected by $\mathcal{S}_u$. Each edge has a weight w_{uv}. The parameter α measures disease-specific quantities. In most of our calculations, we assume that these are scalars and follow Del Valle *et al.* (2007) and Miller (2007), setting

$$T_{uv} = T(\mathcal{I}_u, \mathcal{S}_v, w_{uv}) = 1 - \mathrm{e}^{-\alpha \mathcal{I}_u \mathcal{S}_v w_{uv}}. \tag{2.1}$$

If all contacts are identical, w_{uv} may be absorbed into α

$$T_{uv} = T(\mathcal{I}_u, \mathcal{S}_v) = 1 - \mathrm{e}^{-\alpha \mathcal{I}_u \mathcal{S}_v}. \tag{2.2}$$

Note that T_{uv} is a number assigned to an edge, while $T(\mathcal{I}_u, \mathcal{S}_v)$ is a function that states what the transmissibility between two nodes would be if they shared an edge.

With mild abuse of notation, we denote the probability density functions (pdfs) of $\mathcal{I}$, $\mathcal{S}$ and w by $P(\mathcal{I})$, $P(\mathcal{S})$ and $P(w)$, respectively. We assign $\mathcal{I}$ and $\mathcal{S}$ independently, but allow w to be assigned either independently or based on observed contacts (i.e. by observing contacts in a population, we may create a static network with edge weights assigned based on the observed contact). If w is assigned independently, then it is possible to eliminate edge weights from the analysis by marginalizing over the distribution of weights. However, if weights are not independent (for example work or family contacts tend to have correlated weights), then the details of the distribution and the correlations are important.

Given the infectiousness $\mathcal{I}_u$ of node u, we follow Miller (2007, 2008) and define its *out-transmissibility*

$$T_{\mathrm{out}}(u) = \iint T(\mathcal{I}_u, \mathcal{S}, w) P(\mathcal{S}) P(w) \mathrm{d}\mathcal{S}\, \mathrm{d}w. \tag{2.3}$$

This is the marginalized probability that u infects a randomly chosen neighbour given $\mathcal{I}_u$. From the definition of T_{out} and the pdf $P(\mathcal{I})$, we can calculate the pdf $Q_{\mathrm{out}}(T_{\mathrm{out}})$. We symmetrically define the *in-transmissibility* T_{in} and its pdf $Q_{\mathrm{in}}(T_{\mathrm{in}})$.

We denote the average of a quantity by $\langle\cdot\rangle$. The average transmissibility $\langle T\rangle$ is

$$\langle T\rangle = \iiint T(\mathcal{I}, \mathcal{S}, w) P(\mathcal{I}) P(\mathcal{S}) P(w) \mathrm{d}\mathcal{I}\, \mathrm{d}\mathcal{S}\, \mathrm{d}w. \tag{2.4}$$

2.1.2. Epidemic percolation networks. Rather than studying outbreaks as dynamic processes on networks, we may consider them in the context of epidemic percolation networks (EPNs; Kenah & Robins 2007*a*,*b*; Miller 2008). The EPN framework allows us to study epidemics as static objects and is useful for quickly estimating $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$. In this section, we summarize properties of EPNs; more details are provided in Kenah & Robins (2007), Miller (2007, 2008) and in §A of the electronic supplementary material.

Once the properties of the nodes and edges are assigned, an EPN $\mathcal{E}$ is created as follows: we place each node of G into $\mathcal{E}$. For each edge $\{u,v\}$ in G, we place directed edges (u,v) and (v,u) into $\mathcal{E}$ independently with probability T_{uv} and T_{vu}, respectively. The nodes infected in an outbreak correspond exactly to those nodes that may be reached from the index case following the edges of $\mathcal{E}$. More specifically, the distribution of out-components of a node u in different EPN realizations matches the distribution of outbreaks resulting from different epidemic realizations in the original model with u as the index case. It may be shown that the

distributions of out- and in-component sizes give us information about the probability of nodes to start an epidemic or become infected in an epidemic. We will see that in a large population the structure of a single EPN can be used to accurately estimate $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$.

Once we create an EPN and choose the index case, we define the *rank* of node v as the length of the shortest directed path from the index case to v.[1] If no such path exists, v is never infected.

Interchanging all arrow directions interchanges $\mathcal{P}$ and $\mathcal{A}$. This means that if we can calculate $\mathcal{P}$, then $\mathcal{A}$ may be calculated by the same technique, but with the direction of infection reversed. Owing to this, we focus our attention on calculating $\mathcal{P}$ and apply the same methodology to calculate $\mathcal{A}$. An important consequence is that if T is constant, then $\mathcal{P}=\mathcal{A}$ (Newman 2002; Miller 2007).

2.1.3. The basic reproductive ratio. We expect that epidemics are possible if and only if the *basic reproductive ratio* $\mathcal{R}_0$ is greater than 1. That is, if an average infection causes more than one new case, an epidemic may occur, but otherwise the outbreak dies out quickly. However, this use of $\mathcal{R}_0$ is not consistent with the typical definition: the average number of new infections caused by a single infected individual introduced into a fully susceptible population, which gives $\mathcal{R}_0=\langle T\rangle\langle k\rangle$. A more appropriate definition is the average number of new infections caused by infected individuals early in outbreaks. The distinction is subtle, but results from the fact that whether an outbreak can grow depends on whether the people of low rank infect more than one person each (Diekmann *et al.* 1990). Low-rank individuals may be different from the average individual. Most obviously, they have more contacts (Feld 1991; Newman 2002); but with clustering, they also have a disproportionately large fraction of neighbours infected or recovered.

In order to quantify $\mathcal{R}_0$ more rigorously, we first define N_r to be the number of people of rank r for a given outbreak simulation. We then define the *rank reproductive ratio*

$$\mathcal{R}_{0,r}=\frac{\mathbb{E}[N_{r+1}]}{\mathbb{E}[N_r]} \tag{2.5}$$

to be the expected number of new cases caused by a rank r node (averaged over all possible outbreak realizations). $\mathcal{R}_{0,0}=\langle T\rangle\langle k\rangle$ corresponds to the usual definition of $\mathcal{R}_0$. In practice, we find that $\mathcal{R}_{0,r}$ reaches a plateau quickly as r increases before eventually decreasing as the finite size of the population becomes important. Consequently, an improved definition of $\mathcal{R}_0$ is the limit of $\mathcal{R}_{0,r}$ as r grows, subject to the assumption that $\mathcal{R}_{0,r}$ is unaffected by the finite size of G. This gives (cf. Trapman 2007)

$$\mathcal{R}_0=\lim_{r\to\infty}\lim_{|G|\to\infty}\mathcal{R}_{0,r} \tag{2.6}$$

[1]We follow Ludwig (1975) in using the term rank rather than generation which has been used elsewhere, but is potentially ambiguous. The rank is the smallest number of infectious contacts between the index case and a node. It is possible that a different path takes less time. The path infection actually follows is the path that is shorter in time, rather than number of links.

and generalizes the definition given by Diekmann *et al.* (1990) for ODE models. Under this definition, epidemics are possible if $\mathcal{R}_0>1$, but not if $\mathcal{R}_0<1$. We discuss this further in §B of the electronic supplementary material. In a large population, considering multiple index cases with a single EPN gives a good estimate of $\mathbb{E}[N_r]$ and hence $\mathcal{R}_{0,r}$.

2.2. Configuration model networks

We consider two different types of networks. The first is a class of (unclustered) random networks for which we can derive analytic results based only on the degree distribution. These analytic results will form the leading-order term of our perturbation expansions. The second is a more complicated network resulting from an agent-based simulation, which we will use to demonstrate the accuracy of our perturbation expansions.

Our random networks are created by an algorithm that has been discovered independently a number of times (e.g. Molloy & Reed 1995). These have come to be called configuration model are (CM; Newman 2003*b*) networks. These networks maximally random given the degree distribution. As the number of nodes in a CM network grows, the frequency of short cycles becomes negligible. The resulting lack of clustering allows us to calculate analytic results for epidemics. We briefly discuss these results assuming T is constant. More details are in Andersson (1998), Newman (2002), Meyers *et al.* (2006), Kenah & Robins (2007), Marder (2007), Miller (2007) and Noël *et al.* (2009) and §C of the electronic supplementary material (which also addresses edge weights).

In the early stages of an outbreak in a CM network, the probability that a newly infected (non-index case) node has degree k is $kP(k)/\langle k\rangle$. Clustering is unimportant and so the node will have $k-1$ susceptible neighbours, regardless of its rank. Thus, the expected number of infections caused by a newly infected node is

$$\mathcal{R}_0=T\frac{\langle k^2-k\rangle}{\langle k\rangle}. \tag{2.7}$$

To calculate the probability $\mathcal{P}$ that infection of a randomly chosen index case results in an epidemic, we instead calculate the probability $f=1-\mathcal{P}$ that it does not. Then f is the probability that each neighbour of the index case is either not infected, or infected but does not start an epidemic. Defining h to be the probability that a secondary case does not start an epidemic,

$$f=\sum_k P(k)[1-T+Th]^k. \tag{2.8}$$

We find a similar relationship for h, except that the probability for a secondary case to have degree k is $kP(k)/\langle k\rangle$ and only $k-1$ neighbours are susceptible

$$h=\frac{1}{\langle k\rangle}\sum_k kP(k)[1-T+Th]^{k-1}. \tag{2.9}$$

We solve this recurrence relationship for h numerically, and use the result to find f. $\mathcal{P}$ follows immediately. Because T is constant, this also gives $\mathcal{A}$ (Newman 2002; Miller 2007).

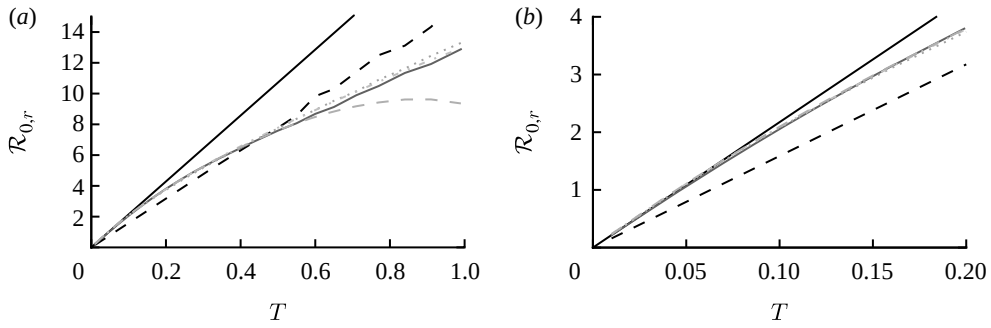

Figure 2. (a,b) Simulated values of the rank reproductive ratio $\mathcal{R}_{0,r} = \mathbb{E}[N_{r+1}]/\mathbb{E}[N_r]$ for $r=0, \ldots, 4$ using an EPN from the (fixed) EpiSimS network with a homogeneous population, compared with the unclustered prediction. (b) At small T, $\mathcal{R}_{0,1}$–$\mathcal{R}_{0,4}$ match the unclustered prediction (black solid curve, unclustered $\mathcal{R}_0$ prediction; black dashed curve, $\mathcal{R}_{0,0}$; grey solid curve, $\mathcal{R}_{0,1}$; dotted curve, $\mathcal{R}_{0,2}$; dot-dashed curve, $\mathcal{R}_{0,3}$; grey dashed curve, $\mathcal{R}_{0,4}$). Each data point for $\langle T\rangle \leq 0.5$ is for 10^5 index cases in a single EPN, while each data point for $T>0.5$ is for 10^3 index cases. Noise becomes less significant at larger r.

If T is not constant, the calculation becomes more difficult, and is discussed further in §C of the electronic supplementary material and Kenah & Robins (2007) and Miller (2007). In general, if T can vary for CM networks, $\mathcal{R}_0 = \langle T\rangle\langle k^2-k\rangle/\langle k\rangle$, while the values calculated assuming constant T give upper bounds for $\mathcal{P}$ and $\mathcal{A}$.

2.3. The EpiSimS network

We are interested in understanding the impact of clustering on disease spread. The term *clustering* is rather vague, and is usually measured by the number of triangles in a network (Watts & Strogatz 1998). However, any sufficiently short cycles impact the spread of an infectious disease. For our purposes, we think of a clustered network as a network with enough short cycles to impact disease dynamics.

It is relatively simple to measure the degree distribution of a population using survey methods. We can easily calculate $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$ for a CM network with the same degree distribution, but the errors between these values and the values for the original clustered network are unknown. Our goal in this paper is to develop analytical techniques to quantify these errors.

To test our predictions, we turn to an agent-based network derived from a single EpiSimS (Eubank *et al.* 2004; Barrett *et al.* 2005; Del Valle *et al.* 2006) simulation of Portland, Oregon. The simulation includes roads, buildings and a statistically accurate (based on census data) population of approximately 1.6 million people who perform daily tasks based on population surveys. This gives a highly detailed knowledge of the interactions in the synthetic population. The degree distribution and contact structure emerge from the simulation. The resulting network has significant clustering and average degree of approximately 16. More details are in §D of the electronic supplementary material.

3. CLUSTERED NETWORKS WITH HOMOGENEOUS NODES

In this section, we assume that the population is homogeneous and all contacts are equally weighted. Consequently, transmissibility is constant: $T_{uv} = T$ for all edges. It follows that $\mathcal{P}=\mathcal{A}$ (Newman 2002; Miller 2007). We develop a predictive theory for $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$ and test the theory with simulations on the EpiSimS network. We begin with $\mathcal{R}_0$.

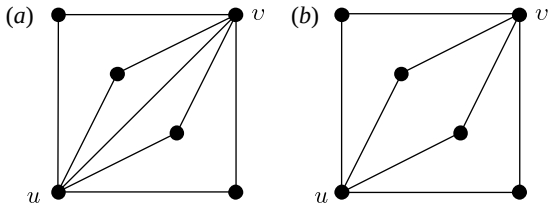

Figure 3. Different options for paths of length 2 between nodes u and v: (a) $n_{uv}=4$, $x_{uv}=1$; (b) $n_{uv}=4$, $x_{uv}=0$.

3.1. The basic reproductive ratio

The simulated rank reproductive ratio $\mathcal{R}_{0,r}$ is shown in figure 2 for $0\leq r\leq 4$. At all values of T, $\mathcal{R}_{0,0} = T\langle k\rangle$ is clearly distinct from $\mathcal{R}_{0,r}$, $r>0$ (which are close together). For $r>0$, $\mathcal{R}_{0,r}$ is asymptotic to the unclustered approximation $T\langle k^2-k\rangle/\langle k\rangle$ as $T\to 0$. This is because at small T the disease only rarely follows all edges of short cycles and so clustering has no impact. As T increases, these curves lie significantly below the unclustered approximation, because clustering reduces the number of available susceptibles. $\mathcal{R}_{0,4}$ peels away from $\mathcal{R}_{0,1}$, $\mathcal{R}_{0,2}$ and $\mathcal{R}_{0,3}$ for larger T because the population is finite, and so the number of susceptibles available to infect after rank 4 is reduced. In larger populations, $\mathcal{R}_{0,4}$ would not deviate.

We conclude that $\mathcal{R}_{0,r}$ converges quickly, and that $\mathcal{R}_{0,1}$ is a good approximation to $\mathcal{R}_0$, but $\mathcal{R}_{0,0}$ is not. This implies that the network has an important structure contained in the paths of length 2, but not in the paths of length 3. This fortunate observation allows us to approximate $\mathcal{R}_0$ by $\mathcal{R}_{0,1}$, which we may analytically calculate with relative ease ($\mathcal{R}_{0,r}$ becomes combinatorially hard as r grows). To find $\mathcal{R}_{0,1} = \mathbb{E}[N_2]/\mathbb{E}[N_1]$, we first note that $\mathbb{E}[N_1] = T\langle k\rangle$. Calculating $\mathbb{E}[N_2]$ is more difficult: consider all pairs of nodes u and v with at least one path of length 2 between them. Let n_{uv} be the number of paths of length 2 between u and v and χ_{uv} be an indicator function: $\chi_{uv}=1$ if $\{u,v\}$ is an edge and $\chi_{uv}=0$ if it is not (figure 3). The probability that an infection of u results in infection of v in exactly two

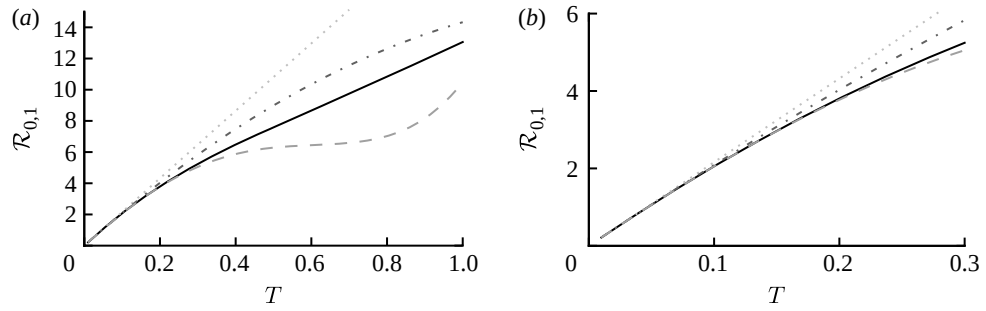

Figure 4. (a,b) Comparison of first three asymptotic approximations for $\mathcal{R}_{0,1}$ from equation (3.1) with the exact value for the EpiSimS network. (b) The comparison at small T is shown (solid curve, exact $\mathcal{R}_{0,1}$; dotted curve, first approximation; dot-dashed curve, second approximation; dashed curve, third approximation).

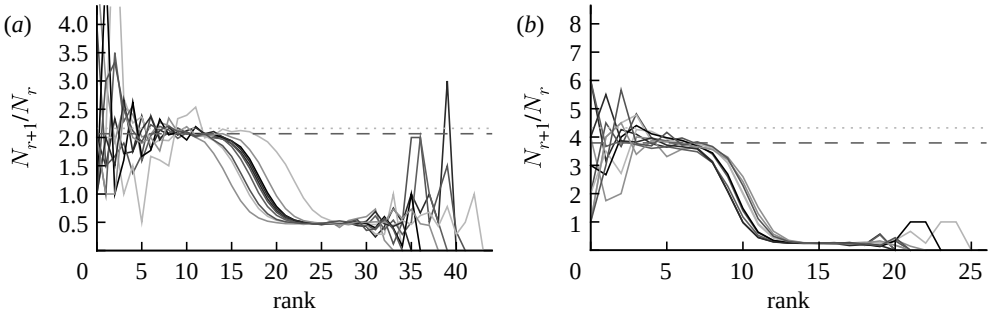

Figure 5. The progression of 10 simulated epidemics for (a) $T=0.1$ and (b) $T=0.2$ in the EpiSimS network. (a) N_{r+1}/N_r against rank and (b) the cumulative fraction of the population infected are shown (dotted curve, unclustered $\mathcal{R}_0$ prediction; dashed curve, $\mathcal{R}_{0,1}$).

steps is $[1-(1-T^2)^{n_{uv}}][1-T]^{\chi_{uv}}$. Summing this over all pairs yields (where N is the size of the population and each pair u and v appears twice)

$$\mathbb{E}[N_2] = \frac{1}{N}\sum_u \sum_{v\neq u} [1-(1-T^2)^{n_{uv}}][1-T]^{\chi_{uv}},$$

which allows us to calculate $\mathcal{R}_{0,1}$ exactly. This sum is straightforward to calculate, but we can increase our understanding with a small T expansion. We approximate $\mathbb{E}[N_2]$ for $T \ll 1$ by

$$\begin{aligned}\mathbb{E}[N_2] &= \frac{1}{N}\sum_u \sum_{v\neq u} T^2 n_{uv}(1-T)^{\chi_{uv}} \\ &\quad - \binom{n_{uv}}{2} T^4 + \mathcal{O}(T^5) \\ &= T^2\langle k^2 - k\rangle - 2T^3\langle n_\triangle\rangle - T^4\langle n_\square\rangle + \mathcal{O}(T^5),\end{aligned}$$

where $\langle n_\triangle\rangle = 1/N\sum_u\sum_{v\neq u} n_{uv}\chi_{uv}$ is the average number of triangles each node is in, and $\langle n_\square\rangle = 1/N\sum_u\sum_{v\neq u}\binom{n_{uv}}{2}$ is the average number of squares each node is in (cf. Hastings 2006). Higher order terms involve more complicated shapes. This gives

$$\mathcal{R}_{0,1} = \frac{\langle k^2-k\rangle}{\langle k\rangle} T - \frac{2\langle n_\triangle\rangle}{\langle k\rangle} T^2 - \frac{\langle n_\square\rangle}{\langle k\rangle} T^3 + \mathcal{O}\left(\frac{T^4}{\langle k\rangle}\right). \tag{3.1}$$

At the leading order, we recover the unclustered prediction for $\mathcal{R}_0$, reflecting the fact that at small T the probability the outbreak follows all edges of a cycle is negligible. As T increases, the first corrections are due to triangles, then squares, then pairs of triangles sharing an edge and sequentially larger and larger structures made up of paths of length 2. A comparison of these approximations with the exact value is shown in figure 4.

Although we have defined $\mathcal{R}_0$ for an ensemble of realizations, figure 5 shows that $\mathcal{R}_{0,1}$ accurately predicts the observed ratio N_{r+1}/N_r for individual simulations once the outbreaks are well established. Early in outbreaks, the behaviour is dominated by stochastic effects, and so the ratio of successive rank sizes is noisy. Once the outbreak has grown large enough, random events become unimportant and the ratio settles at $\mathcal{R}_{0,1}$.[2]

3.2. Epidemic probability and size

In order to assess the effect of clustering on $\mathcal{P}$ and $\mathcal{A}$, we compare epidemics on the EpiSimS network with the analytic predictions derived assuming a CM network of the same degree distribution in figure 6. The epidemic threshold is not notably altered, and the values of $\mathcal{P}$ and $\mathcal{A}$ are almost indistinguishable from the predictions made assuming no clustering, despite the large amount of clustering in the network.

Although initially surprising, these results may be understood intuitively as follows: if T is large enough

[2]Early noise controls how quickly outbreaks become epidemics, and so once stochastic effects become small, the curves appear to be translations in time. We note that it is common to consider the temporal average of a number of outbreaks. However, prior to taking an average, the curves should be shifted in time so that they coincide once the stochastic effects are no longer important. Failure to do so underestimates the early growth, peak incidence and late decay, while it overestimates the epidemic duration. This can lead to an incorrect understanding of 'typical' outbreaks.

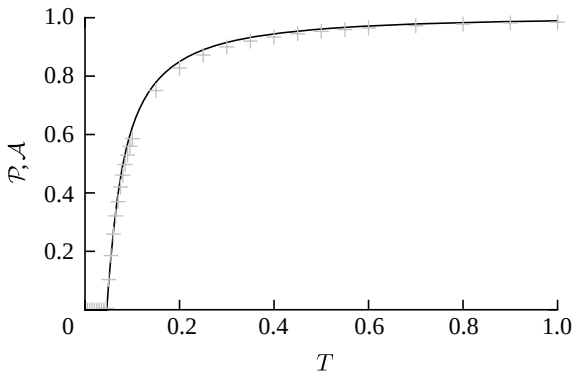

Figure 6. Probability $\mathcal{P}$ and attack rate $\mathcal{A}$ of epidemics for the (clustered) EpiSimS network (pluses) versus T, compared with the prediction derived from the degree distribution assuming no clustering. Each data point is from a single EPN (the variation in $\mathcal{P}$ resulting from different EPNs is negligible).

that the disease follows all edges of a short cycle, then some other edge from a node of that cycle is likely to start an epidemic and the cycle does not prevent an epidemic. On the other hand, if T is smaller so that it does not follow all edges of a cycle, then the disease never sees the existence of the cycle, and the outbreak progresses as if there were no cycle.

To make this more rigorous, we first look at the epidemic threshold. We assume that $\mathcal{R}_0$ is well approximated by $\mathcal{R}_{0,1}$. Let $T_0 = \langle k\rangle/\langle k^2-k\rangle$ be the threshold without clustering and $T_0+\delta T$ be the threshold found by including the correction due to triangles. From equation (3.1), it follows that

$$\frac{\delta T}{T_0} = \frac{2\langle n_\triangle\rangle\langle k\rangle}{\langle k^2-k\rangle^2} + \mathcal{O}\left(\left[\frac{2\langle n_\triangle\rangle\langle k\rangle}{\langle k^2-k\rangle^2}\right]^2\right). \tag{3.2}$$

Because a given node of degree k is contained in at most $(k^2-k)/2$ triangles, we conclude $2\langle n_\triangle\rangle/\langle k^2-k\rangle \leq 1$. So if $\langle k\rangle/\langle k^2-k\rangle$ is small, then the leading-order term of equation (3.2) is small and triangles do not significantly alter the epidemic threshold regardless of the density of triangles. For the EpiSimS network, $\langle k\rangle/\langle k^2-k\rangle$ takes the value 0.046, and so we do not anticipate clustering to play an important role in determining the threshold.

Above threshold, we assume that $\mathcal{P}$ may be expanded much as (3.1)

$$\mathcal{P} = \mathcal{P}_0 + \mathcal{P}_1\langle n_\triangle\rangle + \mathcal{P}_2\langle n_\triangle\rangle^2 + \cdots + Q_1\langle n_\square\rangle + \cdots, \tag{3.3}$$

where $\mathcal{P}_0$ is the epidemic probability in a CM network of the same degree distribution. Although calculating $\mathcal{R}_{0,1}$ only requires information about the nodes of distance at most two from the index case, $\mathcal{P}$ may depend on the effects occurring at larger distance, and so the expansion has many additional terms. In general, we expect that if the average degree is large, then the various coefficients of the correction terms are all small. The larger a structure is, the smaller we expect its corresponding coefficient to be. The coefficient for triangles $\mathcal{P}_1$ may be found by

$$\mathcal{P}_1\langle n_\triangle\rangle = -\frac{1}{N}\sum_{u\in G}\sum_{\triangle\in G}\hat{p}_\triangle(u),$$

where $\hat{p}_\triangle(u)$ is the probability that a given triangle prevents an epidemic if u is the index case (regardless of whether u is part of the triangle). Reversing the order of summation, we get

$$\begin{aligned}\mathcal{P}_1\langle n_\triangle\rangle &= -\frac{N_\triangle}{N}\left\langle\sum_{u\in G}\hat{p}_\triangle(u)\right\rangle_\triangle \\ &= -\frac{1}{3}\langle n_\triangle\rangle\left\langle\sum_{u\in G}\hat{p}_\triangle(u)\right\rangle_\triangle,\end{aligned}$$

where $N_\triangle$ is the number of triangles in G and $\langle\cdot\rangle_\triangle$ is the average of the given quantity taken over all triangles. Thus

$$\mathcal{P}_1 = -\frac{1}{3}\left\langle\sum_{u\in G}\hat{p}_\triangle(u)\right\rangle_\triangle,$$

and we can find $\mathcal{P}_1$ by considering the average effect of a single triangle in an unclustered network.

To calculate the impact of a triangle with nodes u, v and w on $\mathcal{P}$ for a given network, we consider that triangle and a randomly chosen edge $\{x,y\}$ elsewhere in the network. If we replace the edges $\{v,w\}$ and $\{x,y\}$ with $\{v,x\}$ and $\{w,y\}$, then we have a new network without the triangle, but with the same degree distribution. We must estimate the expected change in $\mathcal{P}$ caused by switching the edges.

We begin by assuming that u is the index case. The triangle can affect $\mathcal{P}$ only if the infection tries to cross all three edges, that is if the infection process 'loses' an edge because of clustering. This may happen in three distinct ways. In the first, node u infects both v and w, and then v and/or w tries to infect the other. In the second, u infects v but not w, then v infects w and finally w tries to infect u. The third is symmetric to the second (with u infecting w).

To leading order we can ignore other short cycles, so the probability that an edge leading out of u (not to v or w) will not cause an epidemic is $g = 1-T+Th$, where h (as before) is the probability that a randomly chosen secondary case does not cause an epidemic in an unclustered network and can be calculated using equation (2.9).

We perform a sample calculation with the first case: u infects both v and w. Assume that u has degree k_u, v has degree k_v and w has degree k_w. The probability that u infects both v and w without some other edge leading from u, v or w starting an epidemic is $T^2 g^{k_u+k_v+k_w-6}$. If the $\{v,w\}$ edge were broken and v and w were joined to x and y, respectively (figure 7), then the new probability of u to infect both v and w without an epidemic becomes $T^2 g^{k_u+k_v+k_w-4}$. The difference is $T^2 g^{k_u+k_v+k_w-6}(1-g^2)$, which is the product of three terms, all at most 1. If the sum $k_u+k_v+k_w$ is moderately large, then either $g^{k_u+k_v+k_w-6} \ll 1$ or $1-g^2 \ll 1$ (if g is not close to 1 then the first term is small, otherwise the second term is small). Thus, the triangle has little impact on the epidemic probability in this case.[3] Similar analysis applies to the other two cases where the w to u or v to u infections are lost. Provided the typical sum of degrees of nodes in a triangle is relatively large, the probability

[3]If $\mathcal{P}$ is small, then the relative change may be large, but the absolute change is small.

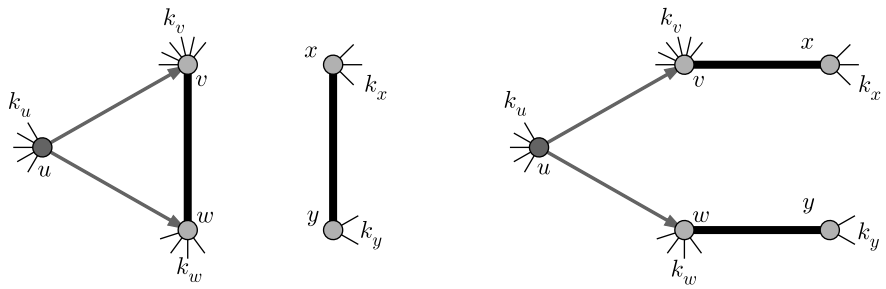

Figure 7. Replacing the edges $\{v,w\}$ and $\{x,y\}$ with $\{v,x\}$ and $\{w,y\}$ breaks the triangle and allows more infections, without affecting the degree distribution.

of an epidemic when the index case is in the triangle is not impacted significantly.

If the index case is not part of the triangle, then the above analysis is modified because we must also consider each node in the path from the index case to the triangle. We must first calculate the probability that infection reaches a node in the triangle while simultaneously no intermediate node sparks an epidemic, and then we calculate the probability as above that the triangle prevents an epidemic. If the index case is u_1 and the path from u_1 to the triangle goes through u_2, …, u_n and then reaches u, then the probability that the triangle prevents an epidemic $\hat{p}(u_1)$ is given by $T^n(g^{-2n+\sum_i k_{u_i}})\hat{p}(u)$. This falls off very quickly, and so nodes not in the triangle are unimportant, unless typical degrees are small.

By contrast, in a network with small average degree and a significant number of triangles, this becomes significant. This explains the observations of Serrano & Boguñá (2006*a*,*b*) who use networks with average degree less than 3 and find that clustering significantly alters $\mathcal{A}$.

It is tempting to generalize our conclusion and state that if the average degree is large, clustering has no impact on $\mathcal{P}$ or $\mathcal{A}$. However, there are a number of counter-examples: consider a network made up of isolated cliques with N_c nodes, then in expansion (3.3), the coefficient for cliques of N_c nodes will not be small. Consequently, care must be taken when using such an expansion to ensure that neglected terms resulting from larger scale structures are in fact negligible. For social networks, we generally anticipate this highly segregated situation to be unimportant.

We conclude that for most reasonable networks, clustering is only important for $\mathcal{P}$ and $\mathcal{A}$ if the typical degrees of nodes are low in which case $\mathcal{R}_0$ is small. A consequence of these results is that if $\mathcal{R}_0$ is moderately large, then $\mathcal{P}$ and $\mathcal{A}$ are effectively unaltered by clustering. If $\mathcal{R}_0$ is small, however, clustering may or may not play a role in determining $\mathcal{P}$ and $\mathcal{A}$, depending on whether $\mathcal{R}_0$ is small because the degrees are small or T is small.

4. CLUSTERED NETWORKS WITH HETEROGENEOUS NODES

When we drop the assumption of constant transmissibility, disease spread becomes more complicated. If $\mathcal{I}$ is heterogeneous and u infects a neighbour, then the *a posteriori* expectation for $T_{\text{out}}(u)$ becomes higher: it is likely to infect more neighbours. This accentuates the effect of short cycles, enhancing the impact of clustering on $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$. A similar argument applies with heterogeneity in $\mathcal{S}$: if v is not infected by one of its neighbours, then the *a posteriori* expectation for $T_{\text{in}}(v)$ becomes lower: it is less likely to be infected by other neighbours, and so has multiple opportunities to prevent an epidemic. Again this accentuates the effect of short cycles.

In this section, we investigate how varying the infectiousness and susceptibility of nodes in the EpiSimS network enables clustering to alter the values of $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$. We will make use of the *ordering assumption* and its consequences from Miller (2008): if u_1 is 'more infectious' than u_2 in a given instance (or v_1 'more susceptible' than v_2), then u_1 is always more infectious than u_2 (or v_1 always more susceptible than v_2). More specifically, the ordering assumption states that $T_{\text{out}}(u_1) > T_{\text{out}}(u_2)$ if and only if $T(\mathcal{I}_{u_1},\mathcal{S}) \geq T(\mathcal{I}_{u_2},\mathcal{S})$ for all $\mathcal{S}$, with inequality for some $\mathcal{S}$, and the corresponding statement for T_{in}. The results of Miller (2008) show that if the ordering assumption holds, heterogeneity tends to reduce $\mathcal{P}$ and $\mathcal{A}$, and the upper bounds on $\mathcal{P}$ and $\mathcal{A}$ correspond to homogeneous populations (constant T).

For simulations in this section, we consider five different illustrative cases, which will be denoted throughout by the symbols given in table 1. In the first four cases, we use equation (2.2), so that $T_{uv} = 1 - \mathrm{e}^{-\alpha \mathcal{I}_u \mathcal{S}_v}$ with the distribution of $\mathcal{I}$ and $\mathcal{S}$ varying for each. We vary α to change the average transmissibility. In the fifth case, the out-transmissibility is maximally heterogeneous: a fraction $\langle T \rangle$ of the population infect all neighbours, while the remaining $1-\langle T \rangle$ infect no neighbours.

The fifth case gives a lower bound on $\mathcal{P}$ for a homogeneously susceptible population (Trapman 2007). It is hypothesized to remain a lower bound on $\mathcal{P}$ if susceptibility is allowed to vary (Miller 2008). We could also consider maximal heterogeneity in susceptibility, but the results for $\mathcal{P}$ and $\mathcal{A}$ merely correspond to interchanging their values for maximal heterogeneity in infectiousness, and so we do not need to consider it explicitly.

4.1. The basic reproductive ratio

We use simulations to calculate the rank reproductive ratio $\mathcal{R}_{0,r}$ for the cases of table 1 and plot the result for

Table 1. For the calculations of §§4 and 5, we determine T_{uv} using equations (2.1) and (2.2) with the distributions of $\mathcal{I}$ and $\mathcal{S}$ given in the first four rows, or by considering a maximally heterogeneous population for which $\langle T\rangle$ of the population infects all neighbours and $1-\langle T\rangle$ infects no neighbours. (The function δ is the Dirac delta function.)

symbol	infectiousness	susceptibility
◇	$P(\mathcal{I})=\delta(\mathcal{I}-1)$	$P(\mathcal{S})=0.5\delta(\mathcal{S}-0.001)+0.5\delta(\mathcal{S}-1)$
■	$P(\mathcal{I})=0.3\delta(\mathcal{I}-0.001)+0.7\delta(\mathcal{I}-1)$	$P(\mathcal{S})=\delta(\mathcal{S}-1)$
×	$P(\mathcal{I})=0.5\delta(\mathcal{I}-0.1)+0.5\delta(\mathcal{I}-1)$	$P(\mathcal{S})=0.2\delta(\mathcal{S}-0.1)+0.8\delta(\mathcal{S}-1)$
●	$P(\mathcal{I})=0.5\delta(\mathcal{I}-0.1)+0.5\delta(\mathcal{I}-1)$	$P(\mathcal{S})=0.8\delta(\mathcal{S}-0.01)+0.2\delta(\mathcal{S}-1)$
⬠	maximally heterogeneous $P(T_{\mathrm{out}})=\langle T\rangle\delta(T_{\mathrm{out}}-1)+(1-\langle T\rangle)\delta(T_{\mathrm{out}})$	homogeneous $T_{\mathrm{in}}=\langle T\rangle$

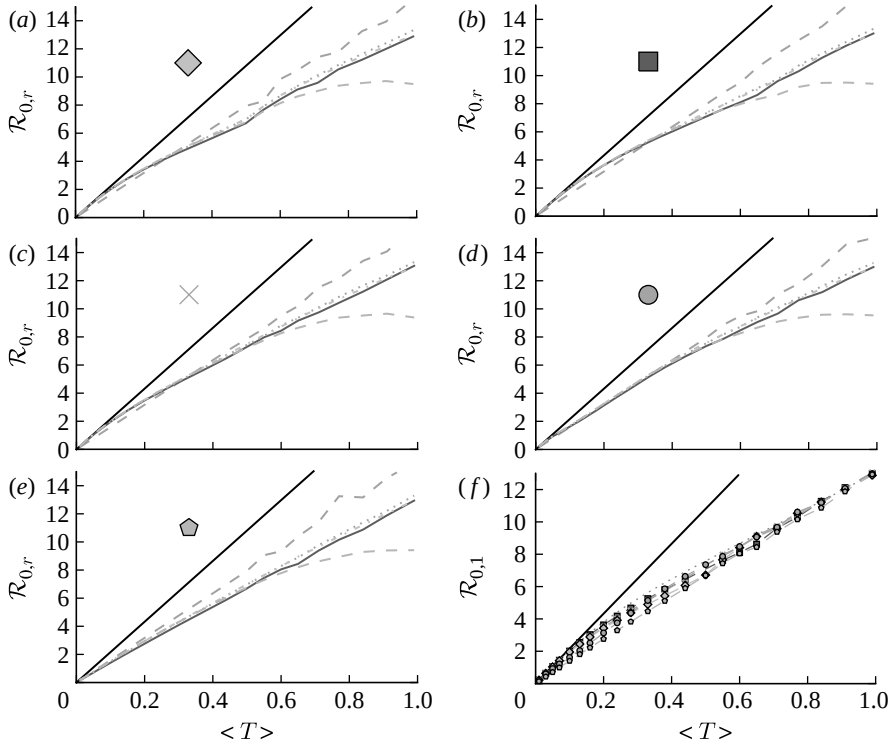

Figure 8. (*a–e*) $\mathcal{R}_{0,r}=\mathbb{E}[N_{r+1}]/\mathbb{E}[N_r]$ calculated from EPNs for the heterogeneous examples of table 1 (black solid curve, unclustered $\mathcal{R}_0$; black dashed curve, $\mathcal{R}_{0,0}$; grey solid curve, $\mathcal{R}_{0,1}$; dotted curve, $\mathcal{R}_{0,2}$; dot-dashed curve, $\mathcal{R}_{0,3}$; grey dashed curve, $\mathcal{R}_{0,4}$). (*f*) $\mathcal{R}_{0,1}$ values for all of the different cases, including both unclustered $\mathcal{R}_0$ (solid curve) and homogeneous $\mathcal{R}_{0,1}$ (dotted curve) are compared.

$0\leq r\leq 4$ in figure 8. Note that $\mathcal{R}_{0,1}$ remains a good approximation to $\mathcal{R}_0$. In the first four cases, $\mathcal{R}_0$ is again asymptotic to the unclustered approximation as $\langle T\rangle\to 0$. There are small kinks for ◇ and ■ at $\langle T\rangle=0.5$ and $\langle T\rangle=0.7$, respectively, resulting from the nature of those distributions. The heterogeneities act to enhance the effect of clustering on $\mathcal{R}_0$, but the effect is relatively small.

In the final, maximally heterogeneous case ⬠, $\mathcal{R}_{0,1}$ remains a good approximation to $\mathcal{R}_0$. At small values of $\langle T\rangle$, the heterogeneity causes clustering to have a larger impact than in a homogeneous population as seen in figure 8*f*, and so this is not asymptotic to the unclustered approximation. At larger values of $\langle T\rangle$, the heterogeneous and homogeneous growth rates are similar.

As before, we can calculate $\mathcal{R}_{0,1}$ analytically, which helps to explain our observations. If the ordering assumption holds, we may use a simplified notation $T(T_{\mathrm{out}},T_{\mathrm{in}})$ to denote the transmissibility from a node with out-transmissibility T_{out} to a node with in-transmissibility T_{in}.[4] We have $\mathbb{E}[N_1]=\langle T\rangle\langle k\rangle$ and

$$\begin{aligned}\mathbb{E}[N_2] &= \frac{1}{N}\sum_u\sum_{v\neq u}\iint[1-(1-T_{\mathrm{out}}T_{\mathrm{in}})^{n_{uv}}]\\ &\quad\times[1-T(T_{\mathrm{out}},T_{\mathrm{in}})]^{\chi_{uv}}Q_{\mathrm{out}}(T_{\mathrm{out}})\\ &\quad\times Q_{\mathrm{in}}(T_{\mathrm{in}})\mathrm{d}T_{\mathrm{out}}\,\mathrm{d}T_{\mathrm{in}}\\ &= \langle k^2-k\rangle\langle T\rangle^2-2\langle n_\triangle\rangle\langle T_{\mathrm{out}}T_{\mathrm{in}}T(T_{\mathrm{out}},T_{\mathrm{in}})\rangle\\ &\quad-\langle n_\square\rangle\langle T_{\mathrm{out}}^2\rangle\langle T_{\mathrm{in}}^2\rangle+\cdots,\end{aligned}$$

[4]We can use this notation because the ordering assumption allows us to uniquely identify I from T_{out} and S from T_{in}. If the ordering assumption fails, similar results hold, but the notation is more cumbersome.

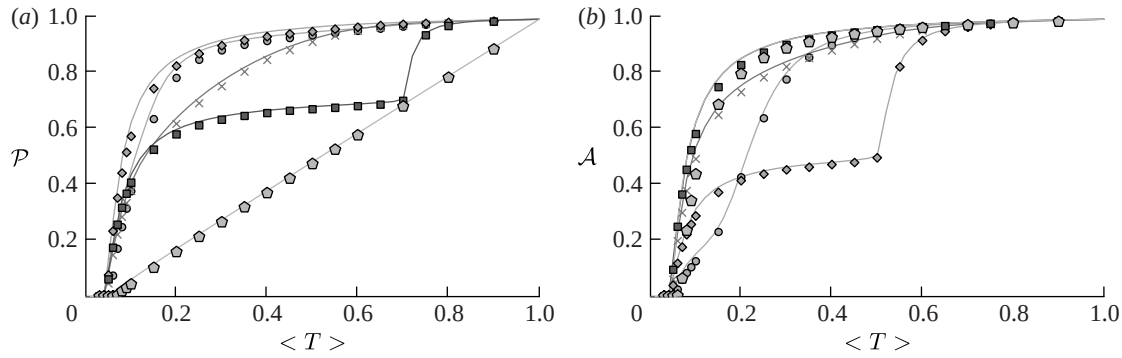

Figure 9. Comparison of (*a*) $\mathcal{P}$ and (*b*) $\mathcal{A}$ observed from EPNs in the clustered EpiSimS network with heterogeneities (symbols) with that predicted by the unclustered theory (curves) using table 1. Each data point is based on a single EPN. For both ■ and ⬟, $T_{\rm in}(v)=\langle T\rangle$ for all nodes, and so the unclustered prediction for $\mathcal{A}$ is the same.

and so we may express the growth rate as a perturbation about the unclustered case $\mathcal{R}_0=\langle T\rangle\langle k^2-k\rangle/\langle k\rangle$ giving

$$\begin{aligned}\mathcal{R}_{0,1}=&\frac{\langle k^2-k\rangle}{\langle k\rangle}\langle T\rangle\\&-\frac{2\langle n_\triangle\rangle}{\langle k\rangle}\frac{\langle T_{\rm out}T_{\rm in}T(T_{\rm out},T_{\rm in})\rangle}{\langle T\rangle}\\&-\frac{\langle n_\square\rangle}{\langle k\rangle}\frac{\langle T_{\rm out}^2\rangle\langle T_{\rm in}^2\rangle}{\langle T\rangle}+\cdots.\end{aligned}\tag{4.1}$$

For the second term, it may be shown that $\langle T\rangle^3\le\langle T_{\rm out}T_{\rm in}T(T_{\rm out},T_{\rm in})\rangle\le\langle T\rangle^2$. The minimum occurs when T is constant, suggesting that the maximum growth rate occurs in a homogeneous population. The maximum $\langle T\rangle^2$ occurs either for ⬟

$$Q_{\rm out}(T_{\rm out})=(1-\langle T\rangle)\delta(T_{\rm out})+\langle T\rangle\delta(T_{\rm out}-1),\tag{4.2}$$

i.e. when the out-transmissibility is maximally heterogeneous, or when the in-transmissibility is maximally heterogeneous

$$Q_{\rm in}(T_{\rm in})=(1-\langle T\rangle)\delta(T_{\rm in})+\langle T\rangle\delta(T_{\rm in}-1).\tag{4.3}$$

Consequently, we expect that for given $\langle T\rangle$, the minimum growth rate occurs with maximally heterogeneous infectiousness or susceptibility. These two minima for $\mathcal{R}_{0,1}$ have previously been hypothesized to give lower bounds on $\mathcal{P}$ and $\mathcal{A}$, respectively (Miller 2008).

We note that in the maximally heterogeneous case, the correction term in (4.1) is significant at the leading order in T. Consequently, if $\langle n_\triangle\rangle$ is comparable with $\langle k^2-k\rangle/2$ (i.e. the clustering coefficient (Watts & Strogatz 1998) is comparable with 1), then the threshold value of $\langle T\rangle$ may be increased by clustering, and $\mathcal{R}_0$ is not asymptotic to the unclustered prediction as $\langle T\rangle\to 0$.

4.2. Probability and size

Figure 9 shows that the unclustered predictions provide a good estimate of $\mathcal{P}$ and $\mathcal{A}$ in the clustered EpiSimS network. We expect that in a network with sufficiently large average degree, the impact of clustering should once again be small.

We use arguments similar to that before, taking a triangle with nodes u, v and w. The reasoning becomes more difficult because knowledge that u infects v increases the expectation that u infects w. Consequently, the lost edges in triangles are more frequently encountered by the outbreak. However, the knowledge that u infects v also increases the expectation that u infects its other neighbours. For a triangle to prevent an epidemic, we need both that no edge outside the triangle leads to an epidemic and that the lost edge would otherwise have caused an epidemic. If the typical degree of the network is not small, then the fact that the lost edge is encountered more frequently may be offset by the fact that when it is encountered, other edges are more likely to spark an epidemic.

For ⬟ where nodes infect all or none of their neighbours, the effect of different triangles that share the index case cannot be separated easily. The probability that the index case directly infects a set of m nodes of interest is $\langle T\rangle$, rather than T^m. Thus, expansions as in (3.3) do not work as well: terms that were previously higher order become significant. Close to the epidemic threshold, this can play an important role. However, well above the epidemic threshold, if the index case infects all of its neighbours, then an epidemic is almost guaranteed and so $\mathcal{P}\approx\langle T\rangle$ regardless of whether the network is clustered. Thus for ⬟, clustering affects $\mathcal{P}$ only close to the epidemic threshold.

In the opposite case where nodes would be infected by any neighbour or else no neighbour, the values of $\mathcal{P}$ and $\mathcal{A}$ are interchanged. Thus, for maximally heterogeneous susceptibility, $\mathcal{P}$ could be significantly altered close to the threshold. The reason for this is as follows: for the first step, the spread is indistinguishable from that of an outbreak with constant T. However, when infections of rank 1 attempt to infect their neighbours, they cannot infect any of the neighbours of the index case. By contrast, in the constant T case, any neighbour not infected by the index case would be susceptible at later steps. Consequently, the impact of triangles becomes much more important (by a factor of $1/\langle T\rangle$) and our earlier argument for neglecting them fails. The interaction of maximal heterogeneity with clustering in this case is larger, but it nevertheless becomes unimportant far from the threshold.

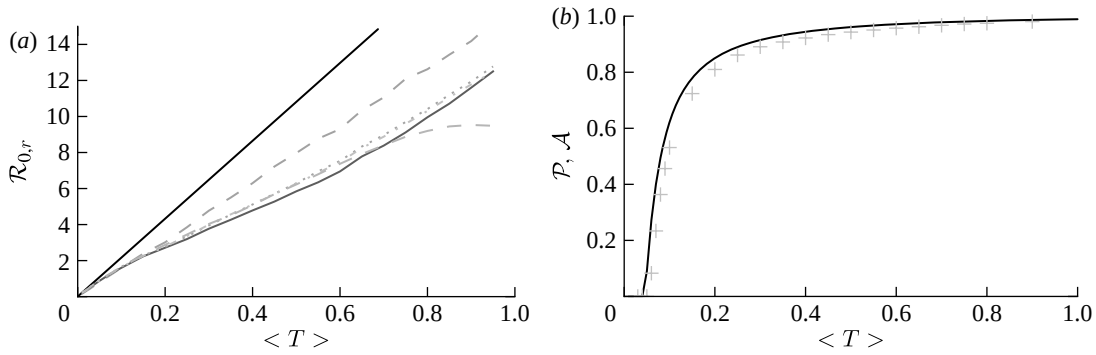

Figure 10. (*a*) $\mathcal{R}_{0,r}$ (black solid curve, unclustered $\mathcal{R}_0$; black dashed curve, $\mathcal{R}_{0,0}$; grey solid curve, $\mathcal{R}_{0,1}$; dotted curve, $\mathcal{R}_{0,2}$; dot-dashed curve, $\mathcal{R}_{0,3}$; grey dashed curve, $\mathcal{R}_{0,4}$) and (*b*) $\mathcal{P}$ and $\mathcal{A}$ for the weighted EpiSimS network with a homogeneous population.

Our prediction that heterogeneity allows clustering to be more significant close to the threshold is borne out for ● where there is relatively strong heterogeneity in susceptibility just above the epidemic threshold. The epidemic threshold for ● is increased compared with the other cases. By contrast, there is much stronger heterogeneity in susceptibility for ◆ at $\langle T\rangle = 0.5$ and in infectiousness for ■ at $\langle T\rangle = 0.7$. This results in a reduction in $\mathcal{A}$ and $\mathcal{P}$, respectively, but because it is far from threshold, there is little deviation from the unclustered predictions.

5. CLUSTERED NETWORKS WITH WEIGHTED EDGES

When we allow edges to be weighted, new complications arise. The weights we use in our simulations are the durations of contacts from the EpiSimS simulation and are discussed in more detail in §D of the electronic supplementary material. If a contact in the original EpiSimS simulation is longer, then a higher weight is assigned. If the weights of different edges were independent, then we could simply take $T_{uv} = \int T(\mathcal{I}_u, \mathcal{S}_v, w)P(w)\mathrm{d}w$. However, edge weights are not independent: clustered connections tend to have larger weights. If brief contacts are negligible, then the disease spreads on a subnetwork of the original network. The new network has a comparable number of short cycles to the original, but lower typical degree. This should enhance the impact of clustering.

For our calculations in this section, we first isolate the impact of weighted edges by taking a homogeneous population ($\mathcal{I} = \mathcal{S} = 1$) and using $T_{uv} = 1 - \mathrm{e}^{-\alpha w_{uv}}$. We vary α in order to set $\langle T\rangle$. We then investigate a heterogeneous population using equation (2.1) with the first four distributions of table 1.

Results for a homogeneous population are shown in figure 10. Because $T_{uv} = T_{vu}$ for all pairs, it follows that $\mathcal{P} = \mathcal{A}$. If different edge weights were uncorrelated, then the value of $\mathcal{R}_0$ would match with figure 2 and $\mathcal{P}$ and $\mathcal{A}$ would match with figure 6. We see, however, that $\mathcal{R}_0$ is significantly reduced from the homogeneous unweighted population (but $\mathcal{R}_{0,1}$ remains a good approximation). $\mathcal{P}$ and $\mathcal{A}$ are mildly reduced close to the threshold. These observations are consistent with our expectation that clustering should be accentuated by incorporating edge weights. Although the predictions for $\mathcal{P}$ and $\mathcal{A}$ are not far off, we expect that they would improve if we adjusted the degree distribution to match that of the effective network on which the disease spreads.

When the population is moderately heterogeneous (figure 11), we still find that $\mathcal{R}_{0,1}$ is a reasonable approximation to the true value of $\mathcal{R}_0$; however, it slightly underestimates $\mathcal{R}_0$ as $\langle T\rangle$ grows. Unfortunately, the analytic calculation of $\mathcal{R}_{0,1}$ is much more difficult, and so it is more appropriate to use simulations to estimate its value. If there were no correlation between weights of different edges, then the calculation would reduce to that of §4.

We consider $\mathcal{P}$ and $\mathcal{A}$ in figure 12. The unclustered predictions are reasonable approximations of the actual values. The error is larger than before because we have combined two effects (edge weights and heterogeneity) that both accentuate the impact of clustering. In spite of this, the predicted values of $\mathcal{P}$ and $\mathcal{A}$ are not far off, and the direction of the error is consistent: the unclustered prediction is always an overestimate.

6. DISCUSSION

We have investigated the interplay of clustering, node heterogeneity and edge weights on the growth rate $\mathcal{R}_0$, probability $\mathcal{P}$ and size of epidemics $\mathcal{A}$ in social networks. For unclustered networks with independently distributed edge weights, it is possible to predict all these quantities analytically. Under weak assumptions, we can accurately estimate $\mathcal{R}_0$, $\mathcal{P}$ and $\mathcal{A}$ for clustered networks.

If the typical degrees are not small, then for a given average transmissibility and degree distribution, the following can be stated.

— The dominant effect controlling the growth rate of epidemics is clustering. Increased clustering reduces $\mathcal{R}_0$.
— The dominant effect controlling the probability of epidemics is heterogeneity in infectiousness. Increased heterogeneity reduces $\mathcal{P}$.
— The dominant effect controlling the size of epidemics is heterogeneity in susceptibility. Increased heterogeneity reduces $\mathcal{A}$.

We are thus able to neglect clustering and still closely estimate $\mathcal{P}$ based only on the degree distribution and the out-transmissibility pdf Q_{out}. The estimate for

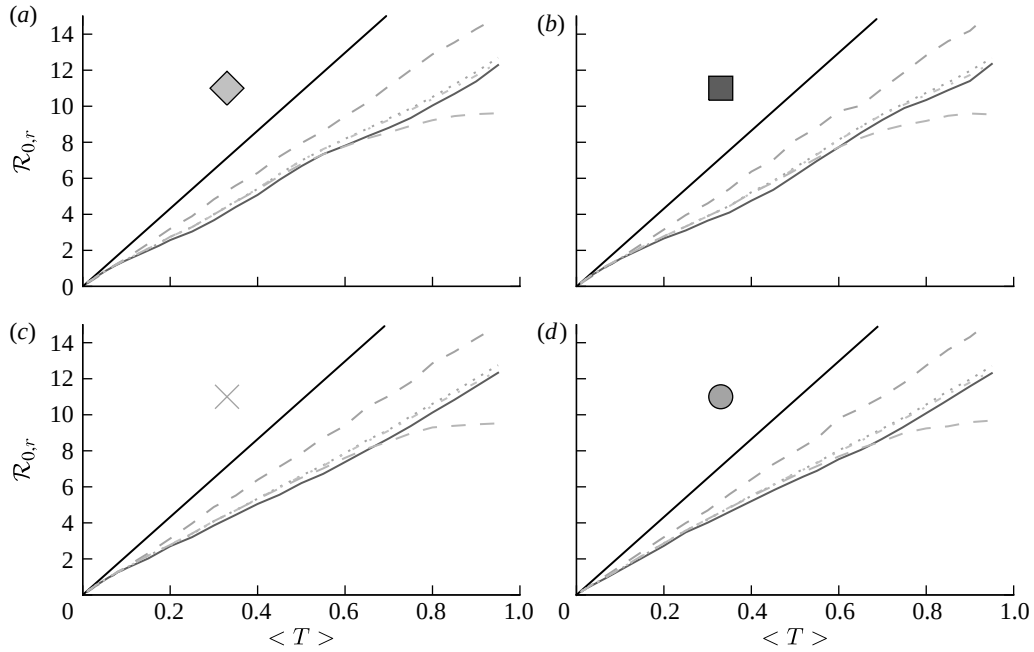

Figure 11. (*a–d*) $\mathcal{R}_{0,r}$ with heterogeneous transmissibility and weighted edges on the EpiSimS network (black solid curve, unclustered $\mathcal{R}_0$; black dashed curve, $\mathcal{R}_{0,0}$; grey solid curve, $\mathcal{R}_{0,1}$; dotted curve, $\mathcal{R}_{0,2}$; dot-dashed curve, $\mathcal{R}_{0,3}$; grey dashed curve, $\mathcal{R}_{0,4}$).

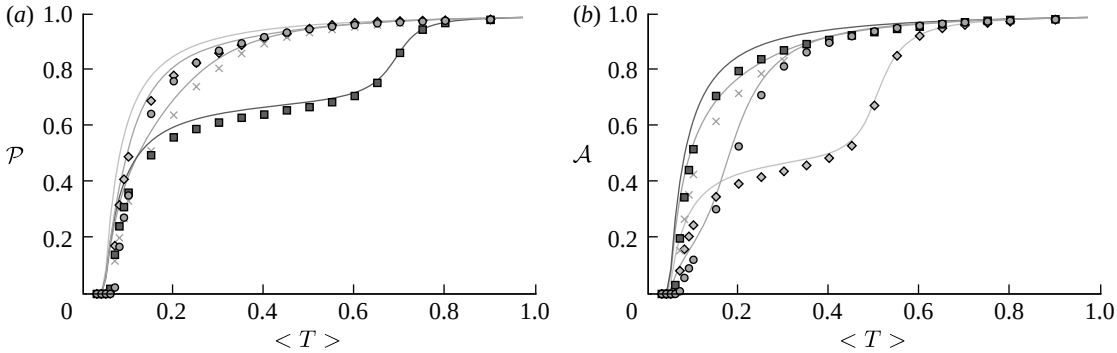

Figure 12. Simulated (*a*) $\mathcal{P}$ and (*b*) $\mathcal{A}$ (symbols) for the weighted EpiSimS network compared with predictions in unclustered networks with the same edge weight distribution (curves).

$\mathcal{A}$ depends only on the degree distribution and the in-transmissibility pdf $Q_{\rm in}$. The impact of clustering is significant in altering $\mathcal{R}_0$, and its impact is mildly enhanced by heterogeneities. This enhancement occurs because the probability of following all edges of a cycle is increased if some of the edges are correlated owing to the heterogeneity. If heterogeneity is large, clustering may play a small role in moving the epidemic threshold, but otherwise its effect on the threshold is negligible. In networks with small typical degree, it has been observed that clustering can modify $\mathcal{P}$ or $\mathcal{A}$ (Serrano & Boguñá 2006*a*,*b*), which is consistent with our estimates.

If edge weights are included, but are independently distributed, then their impact is in modifying $Q_{\rm in}(T_{\rm in})$ and $Q_{\rm out}(T_{\rm out})$. The resulting modification may be calculated explicitly, and edge weights have no further effect. If edge weights are correlated, then they have a more important role in governing the behaviour of epidemics, particularly if higher weight edges tend to be the clustered edges (as frequently occurs in social networks). If this happens, then the impact of clustering is enhanced and the growth rate of epidemics is further reduced.

When we move from predicting $\mathcal{P}$ and $\mathcal{A}$ to predicting $\mathcal{R}_0$, we find that the growth rate is well approximated by $\mathcal{R}_{0,1} = \mathbb{E}[N_2]/\mathbb{E}[N_1]$. This may be calculated analytically in the homogeneous case (constant T). When heterogeneities are included, the calculation becomes harder, and when edge weights are included it becomes largely intractable. However, these are easily estimated through simulation.

These observations show that using $\mathcal{R}_0$ to predict $\mathcal{A}$ will generally be inadequate. In a homogeneous but clustered population, $\mathcal{R}_0$ is reduced but $\mathcal{A}$ is unaffected, and so predictions of $\mathcal{A}$ based on $\mathcal{R}_0$ will be too small. In networks that are not clustered but have heterogeneities in susceptibility, $\mathcal{R}_0$ is unaffected but $\mathcal{A}$ is substantially reduced. Consequently, the value of $\mathcal{A}$ predicted from $\mathcal{R}_0$ will be too large.

Perhaps our most important conclusion about clustering is that it plays an important role in altering the growth of an epidemic, but it plays only a small role

in determining whether an epidemic may occur or how big it would be. If the relevant questions are, 'how likely is an epidemic and how large would it be?', then the modeller may proceed ignoring clustering. If however, the question is 'how fast will an epidemic grow initially?', then clustering must be considered, but only enough to calculate $\mathcal{R}_{0,1}$.

Our results have implications for designing intervention strategies. A number of strategies are available to control epidemic spread, including travel restrictions, quarantines and vaccination. Most of the mathematical theory predicting the effects of these strategies has been developed under the assumption of no clustering. Most immediately, if we measure $\mathcal{R}_0=2$ at the early stages of an epidemic, traditional approaches will suggest that vaccinating just over half of the population will bring the epidemic below threshold. However, if the population is clustered, then the observed $\mathcal{R}_0$ was already affected by the fact that some transmission chains were redundant. Following vaccination, some of these chains will no longer be redundant and the disease may still spread with $\mathcal{R}_0>1$.

Achieving a better understanding of the effect of clustering further helps to guide our intuition when choosing between strategies. For example, let us assume that we have the choice between two strategies: in the first, we stagger work schedules in such a way that a typical person's contacts are reduced by one-third; in the second, we implement population-wide behaviour changes so that the same reduction in number of contacts is achieved, but the work contacts are unaltered. The first reduces clustering while the second increases the relative frequency of clustering. The value of $\mathcal{R}_0$ is smaller in the second case than in the first because of the larger clustering, but $\mathcal{P}$ and $\mathcal{A}$ are reduced by a comparable amount in both cases. Which strategy is best depends on our goals and relative costs.

Strategies that enhance heterogeneity in infectiousness or susceptibility can be important to help reduce $\mathcal{P}$ or $\mathcal{A}$, even when there is little impact on $\mathcal{R}_0$. Depending on which quantity we want to minimize, different choices will be optimal. Consider a choice between vaccinating all individuals with a vaccine that reduces T_{uv} by a factor of $1/2$ for all pairs u and v or a contact tracing strategy that will remove half of all new infections before they have a chance to infect anyone. Both strategies reduce $\langle T\rangle$ by a half. However, the first reduces T_{out} uniformly, while the second increases heterogeneity in T_{out}. Thus, if we have the choice of the two strategies, then contact tracing is more likely to eliminate the disease before an epidemic can happen. If our choice is instead between a global vaccine reducing T_{in} by a factor of $1/2$ for all individuals and a completely effective vaccine that is only available for half of the population, the latter choice will be more effective for reducing $\mathcal{A}$.

This work was supported by the Division of Mathematical Modeling at the UBC CDC under CIHR (grant nos. MOP-81273 and PPR-79231) and the BC Ministry of Health (Pandemic Preparedness Modeling Project), by DOE at LANL under contract DE-AC52-06NA25396 and the DOE Office of ASCR programme in Applied Mathematical Sciences and by the RAPIDD programme of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health. Luís M. A. Bettencourt contributed greatly to the early development of this work. I am grateful to Sara Y del Valle for providing the EpiSimS network data.

REFERENCES

Ajelli, M. & Merler, S. 2008 The impact of the unstructured contacts component in influenza pandemic modeling. *PLoS ONE* **3**, e1519. (doi:10.1371/journal.pone.0001519)

Anderson, R. M. & May, R. M. 1991 *Infectious diseases of humans*. Oxford, UK: Oxford University Press.

Andersson, H. 1998 Limit theorems for a random graph epidemic model. *Ann. Appl. Probab.* **8**, 1331–1349. (doi:10.1214/aoap/1028903384)

Bansal, S. 2008 Ecology of infectious diseases with contact networks and percolation theory. PhD thesis, University of Texas at Austin.

Barrett, C. L., Eubank, S. G. & Smith, J. P. 2005 If smallpox strikes Portland…. *Sci. Am.* **292**, 42–49.

Bollobás, B. 2001 *Random graphs*. Cambridge, UK: Cambridge University Press.

Britton, T., Deijfen, M., Lagerås, A. N. & Lindholm, M. 2007 Epidemics on random graphs with tunable clustering. (http://arXiv:0708.3939)

Del Valle, S. Y., Stroud, P. D., Smith, J. P., Mniszewski, S. M., Riese, J. M., Sydoriak, S. J. & Kubicek, D. A. 2006 EpiSimS: epidemic simulation system. Technical report LAUR-06-6714, Los Alamos National Laboratory.

Del Valle, S. Y., Hyman, J. M., Hethcote, H. W. & Eubank, S. G. 2007 Mixing patterns between age groups in social networks. *Soc. Netw.* **29**, 539–554. (doi:10.1016/j.socnet.2007.04.005)

Diekmann, O., Heesterbeek, J. A. P. & Metz, J. A. J. 1990 On the definition and the computation of the basic reproduction ratio $\mathcal{R}_0$ in models for infectious diseases in heterogeneous populations. *J. Math. Biol.* **28**, 365–382. (doi:10.1007/BF00178324)

Eames, K. T. D. 2008 Modelling disease spread through random and regular contacts in clustered populations. *Theor. Popul. Biol.* **73**, 104–111. (doi:10.1016/j.tpb.2007.09.007)

Eubank, S., Guclu, H., Anil Kumar, V. S., Marathe, M. V., Srinivasan, A., Toroczkai, Z. & Wang, N. 2004 Modelling disease outbreaks in realistic urban social networks. *Nature* **429**, 180–184. (doi:10.1038/nature02541)

Feld, S. L. 1991 Why your friends have more friends than you do. *Am. J. Sociol.* **96**, 1464–1477. (doi:10.1086/229693)

Ferguson, N. M., Cummings, D. A. T., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., Iamsirithaworn, S. & Burke, D. S. 2005 Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature* **437**, 209–214. (doi:10.1038/nature04017)

Germann, T. C., Kadau, K., Longini Jr, I. M. & Macken, C. A. 2006 Mitigation strategies for pandemic influenza in the United States. *Proc. Natl Acad. Sci. USA* **103**, 5935–5940. (doi:10.1073/pnas.0601266103)

Hastings, M. B. 2006 Systematic series expansions for processes on networks. *Phys. Rev. Lett.* **96**, 148701. (doi:10.1103/PhysRevLett.96.148701)

Keeling, M. J. 1999 The effects of local spatial structure on epidemiological invasions. *Proc. R. Soc. B* **266**, 859–867. (doi:10.1098/rspb.1999.0716)

Kenah, E. & Robins, J. M. 2007*a* Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing. *J. Theor. Biol.* **249**, 706–722. (doi:10.1016/j.jtbi.2007.09.011)

Kenah, E. & Robins, J. M. 2007*b* Second look at the spread of epidemics on networks. *Phys. Rev. E* **76**, 36113. (doi:10.1103/PhysRevE.76.036113)

Kermack, W. O. & McKendrick, A. G. 1927 A contribution to the mathematical theory of epidemics. *Proc. R. Soc. A* **115**, 700–721. (doi:10.1098/rspa.1927.0118)

Kuulasmaa, K. 1982 The spatial general epidemic and locally dependent random graphs. *J. Appl. Probab.* **19**, 745–758. (doi:10.2307/3213827)

Ludwig, D. 1975 Final size distributions for epidemics. *Math. Biosci.* **23**, 33–46. (doi:10.1016/0025-5564(75)90119-4)

Marder, M. 2007 Dynamics of epidemics on random networks. *Phys. Rev. E* **75**, 066103. (doi:10.1103/PhysRevE.75.066103)

Meyers, L. A. 2007 Contact network epidemiology: bond percolation applied to infectious disease prediction and control. *Bull. Am. Math. Soc.* **44**, 63–86. (doi:10.1090/S0273-0979-06-01148-7)

Meyers, L. A., Pourbohloul, B., Newman, M. E. J., Skowronski, D. M. & Brunham, R. C. 2005 Network theory and SARS: predicting outbreak diversity. *J. Theor. Biol.* **232**, 71–81. (doi:10.1016/j.jtbi.2004.07.026)

Meyers, L. A., Newman, M. & Pourbohloul, B. 2006 Predicting epidemics on directed contact networks. *J. Theor. Biol.* **240**, 400–418. (doi:10.1016/j.jtbi.2005.10.004)

Miller, J. C. 2007 Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Phys. Rev. E* **76**, 010101. (doi:10.1103/PhysRevE.76.010101)

Miller, J. C. 2008 Bounding the size and probability of epidemics on networks. *J. Appl. Probab.* **45**, 498–512. (doi:10.1239/jap/1214950363)

Molloy, M. & Reed, B. 1995 A critical point for random graphs with a given degree sequence. *Random Struct. Algorithms* **6**, 161–179.

Mossong, J. *et al.* 2008 Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLoS Med.* **5**, 381–391. (doi:10.1371/journal.pmed.0050074)

Newman, M. E. J. 2002 Spread of epidemic disease on networks. *Phys. Rev. E* **66**, 16128. (doi:10.1103/PhysRevE.66.016128)

Newman, M. E. J. 2003*a* Properties of highly clustered networks. *Phys. Rev. E* **68**, 026121. (doi:10.1103/PhysRevE.68.026121)

Newman, M. E. J. 2003*b* The structure and function of complex networks. *SIAM Rev.* **45**, 167–256. (doi:10.1137/S003614450342480)

Noël, P.-A., Davoudi, B., Dubé, L. J., Brunham, R. C. & Pourbohloul, B. 2009 Time evolution of epidemic disease on finite and infinite networks. *Phys. Rev. E.* **79**, 026101. (doi:10.1103/PhysRevE.79.026101)

Serrano, M. Á. & Boguñá, M. 2006*a* Clustering in complex networks. II. Percolation properties. *Phys. Rev. E* **74**, 056115. (doi:10.1103/PhysRevE.74.056115)

Serrano, M. Á. & Boguñá, M. 2006*b* Percolation and epidemic thresholds in clustered networks. *Phys. Rev. Lett.* **97**, 088701. (doi:10.1103/PhysRevLett.97.088701)

Trapman, P. 2007 On analytical approaches to epidemics on networks. *Theor. Popul. Biol.* **71**, 160–173. (doi:10.1016/j.tpb.2006.11.002)

Watts, D. J. & Strogatz, S. H. 1998 Collective dynamics of 'small-world' networks. *Nature* **393**, 409–410. (doi:10.1038/30918)