OPEN ACCESS Freely available online

PLoS COMPUTATIONAL BIOLOGY

Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics

Erik M. Volz[1]*, Joel C. Miller[2,3], Alison Galvani[4], Lauren Ancel Meyers[5,6]

1 Department of Epidemiology, University of Michigan, Ann Arbor, Michigan, United States of America, **2** Department of Epidemiology, Harvard University, Cambridge, Massachusetts, United States of America, **3** Fogarty International Center, National Institutes of Health, Washington, D.C., United States of America, **4** Department of Epidemiology, Yale University, New Haven, Connecticut, United States of America, **5** Santa Fe Institute, Santa Fe, New Mexico, United States of America, **6** Section of Integrative Biology, The University of Texas, Austin, Texas, United States of America

Abstract

The spread of infectious diseases fundamentally depends on the pattern of contacts between individuals. Although studies of contact networks have shown that heterogeneity in the number of contacts and the duration of contacts can have far-reaching epidemiological consequences, models often assume that contacts are chosen at random and thereby ignore the sociological, temporal and/or spatial clustering of contacts. Here we investigate the simultaneous effects of heterogeneous and clustered contact patterns on epidemic dynamics. To model population structure, we generalize the configuration model which has a tunable degree distribution (number of contacts per node) and level of clustering (number of three cliques). To model epidemic dynamics for this class of random graph, we derive a tractable, low-dimensional system of ordinary differential equations that accounts for the effects of network structure on the course of the epidemic. We find that the interaction between clustering and the degree distribution is complex. Clustering always slows an epidemic, but simultaneously increasing clustering and the variance of the degree distribution can increase final epidemic size. We also show that bond percolation-based approximations can be highly biased if one incorrectly assumes that infectious periods are homogeneous, and the magnitude of this bias increases with the amount of clustering in the network. We apply this approach to model the high clustering of contacts within households, using contact parameters estimated from survey data of social interactions, and we identify conditions under which network models that do not account for household structure will be biased.

Citation: Volz EM, Miller JC, Galvani A, Ancel Meyers L (2011) Effects of Heterogeneous and Clustered Contact Patterns on Infectious Disease Dynamics. PLoS Comput Biol 7(6): e1002042. doi:10.1371/journal.pcbi.1002042

Editor: Mark M. Tanaka, University of New South Wales, Australia

Received November 2, 2010; **Accepted** March 23, 2011; **Published** June 2, 2011

Funding: The authors acknowledge financial support from NIH U01 GM087719. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: erikvolz@umich.edu

Introduction

Contacts sufficient for transmission of infectious disease occur repeatedly within stable relationships such as between sex partners or within households and workplaces. Epidemiologists increasingly use random network models that explicitly capture such interactions to study disease dynamics [1]. This work has shown that infectious disease dynamics can be profoundly influenced by two key network properties– the distribution in the number of contacts per individual (the degree distribution) [2] and the transitivity or clustering of contacts, such as within households [3,4]. However, we lack a general framework for studying the combined epidemiological impacts of clustering and degree distribution. For public health, such understanding may be critical to predicting epidemiological events across diverse populations and tailoring control strategies appropriately.

As epidemiological models grow in complexity, we face the question of how much complexity is necessary and useful. For example, which features of network structure significantly influence disease dynamics and which can we ignore without introducing large biases? In some cases, mass action models that assume panmixis may be adequate and thus we can ignore network structure altogether. In others, incorporating realistic degree distributions and/or clustering may be important. A published simulation-based study [5] suggests that clustering affects epidemic dynamics when transmissibility is low and contacts between two individuals are highly autocorrelated. However, there remains a clear need for general, systematic model selection rules.

The impact of the degree distribution on epidemics in the absence of clustering is complex, but has received considerable attention and is relatively well understood [1,2,6,7]. For example, in networks with power law degree distributions (so-called scale free networks), as the variance of the degree distribution diverges to infinity, the reproduction number for a given pathogen also diverges to infinity while the minimum transmissibility necessary for epidemics to occur approaches zero (meaning even diseases with very low infectiousness have the potential to cause epidemics).

In contrast, the effects of clustering on epidemics are still unclear. Some studies suggest that clustering decreases epidemic thresholds, making an epidemic more likely to occur after an initial introduction [8]. Others studies suggest that the relationships between clustering and the epidemic threshold is subtle [9–11], and depends on the nature of clustering in the population. The effects of clustering on the timescale of an epidemic are less ambiguous, with most studies suggesting that clustering decreases

Author Summary

The transmission dynamics of infectious diseases are sensitive to the patterns of interactions among susceptible and infectious individuals. Human social contacts are known to be highly heterogeneous (the number of social contacts ranges from few to very many) and to be highly clustered (the social contacts of a single individual tend also to contact each other). To predict the impacts of these patterns on infectious disease transmission, epidemiologists have begun to use random network models, in which nodes represent susceptible, infectious, or recovered individuals and links represent contacts sufficient for disease transmission. This paper introduces a versatile mathematical model that takes both heterogeneous connectivity and clustering into account and uses it to quantify the relative impact of clustered contacts on epidemics and the prediction biases that can arise when clustering and variability in infectious periods are ignored.

the rate of epidemic propagation. Here, we describe and analyze a versatile model that allows extensive exploration of the interactive impacts of clustering and degree distribution on epidemic dynamics. Although clustering always retards an epidemic, the timescale of the epidemic is more sensitive to the variance of the degree distribution than to clustering.

Following the approach introduced in [12,13], we model the spread of infectious disease through structured host populations using networks that are straightforward generalizations of the configuration model [14]. Our model is designed so that one can easily tune the parameters describing the degree distribution and the number of cliques in the network (a *clique* is a completely connected subgraph), which is closely related to the clustering coefficient. Although these networks are not tree-like locally, they can be analyzed using branching processes and percolation theory, as shown in [12,13], and more recently in [15] and [16].

Our epidemic model generalizes the approaches recently introduced in [17,18] for modeling the dynamics of epidemics in networks. These models exactly predict epidemic spread in a class of random networks. The resulting model consists of a low dimensional system of ordinary differential equations that describes the prevalence of infection over time. Recently, an alternative system of approximate ODEs was independently developed [19] which describes epidemics in networks with arbitrary degree distributions and clustering coefficients. This heuristic approach is intended to be fairly generic, and it is not clear if there are clustered networks for which this model is exact. Our complimentary approach allows straightforward analytical solutions (using percolation theory and branching process methods) for a simple class of random networks. In some cases, our model agrees closely with the one presented in [19], but it can differ substantially around epidemic thresholds. This result suggests that the clustering coefficient (a single value for the entire network) alone is not always sufficient to determine the full epidemiological impact of clustering.

We also revisit one of the early, pioneering approaches to modeling disease transmission through complex contact networks: approximating the final size of an epidemic (the giant component of the network) using bond percolation [12,13]. A recent paper introduces a method that correctly accounts for variation in infectious periods when making such calculations [16]. In contrast to what is found in unclustered networks, in which such variation does not significantly impact epidemic sizes [20–23], we find that in highly clustered networks ignoring variation in infectious periods can introduce considerable bias.

In addition, we model a realistic population by estimating network parameters from a large diary-based survey of social interactions [24]. We quantify the amount of network clustering that occurs within households and show that ignoring household clustering can lead to significant prediction errors including overestimation of both prevalence and, somewhat counter-intuitively, the epidemiological significance of households.

Materials and Methods

We consider a basic susceptible-infected-recovered model. Infectious nodes transmit to neighbors at a constant rate β and transition to the immune recovered state at a constant rate γ. Once recovered, the node cannot be re-infected, and can no longer transmit to neighbors. Key parameters and variables are defined in table 1.

Our solutions are based on the class of undirected random graphs originally described in [12,13], which are refinements of bipartite configuration models [8,25,26]. A node can be a member of multiple cliques of various size. A two-clique is a pair of nodes with an edge between them, and we will call these *lines*. A three-clique is three nodes with all three possible edges, which we call *triangles*. Each node is a member of a random number of lines and triangles. The probability that a node is a member of l lines and t triangles is described by the probability mass function $p_{l,t}$. Our model captures network structure using the probability generating function (PGF):

$$g(x,y)=\sum_{l,t}p_{l,t}x^l y^t.$$

The degree distribution, which describes the probability that a node is a member of k edges, is generated by the following univariate PGF:

$$G(x)=g(x,x^2).$$

Finite-size realizations of these random networks can be easily generated as described in the next section. Most of this section concerns the derivation of equations that describe epidemic dynamics; these solutions are asymptotically exact in the limit of large population size, and as discussed below, compare well to large random networks.

Clustering is often characterized using the clustering coefficient, C, which is the ratio of $3\times$ the number of triangles [12,13], denoted N_Δ, to the number of 2-paths in the network, denoted N_3. C can be interpreted as the probability that two random edges that share a common node are joined by a third edge to form a triangle. Thus we have

$$\begin{aligned} C &= 3N_\Delta/N_3 \\ &= \frac{g^{(y)}(1,1)}{\frac{1}{2}G''(1)}. \end{aligned} \tag{1}$$

When differentiating the PGF, we will use superscripts so that, for example, $g^{(x)}$ would indicate the first derivative with respect to x and $g^{(x,x)}$ would indicate the second derivative with respect to x. The PGF can be used to calculate many useful properties of the graph; for example, the expected number of lines and triangles to which a random node belongs is

Table 1. Definitions for key parameters and variables.

Parameter	Definition
β	Transmission rate
β_k	Transmission rate within a clique of size k
γ	Recovery rate
C	Clustering coefficient
N	The number of nodes in the network
S,I,R	The fraction of the population susceptible, infectious, and recovered respectively
$p_{l,t}$	The frequency of nodes in the network that is a member of l lines and t triangles
$g(x,y)$	Probability generation function for the numbers of lines and triangles of which a node is a member
$\theta_2(t)$	A survivor function for remaining susceptible given that a node is a member of a single line
$\theta_3(t)$	A survivor function for remaining susceptible given that a node is a member of a single triangle
ϕ_S,ϕ_I,ϕ_R	The probabilities that a neighbor of a susceptible node along a line is susceptible, infectious or recovered
ϕ_{XY}	The probabilities that the two neighbors of a susceptible in a triangle are in states X and Y
$n_{ij}N$	The number of 3-cliques with i susceptible and j infectious members
$M_{SI}N$	The number of lines with one susceptible and one infectious member

doi:10.1371/journal.pcbi.1002042.t001

$$M=\sum_{l,t} lp_{l,t}=g^{(x)}(1,1) \tag{2}$$

$$\hat{M}=\sum_{l,t} tp_{l,t}=g^{(y)}(1,1). \tag{3}$$

Generating random clustered networks

Random graphs [12,13] can be algorithmically generated by assigning a random number of lines and triangles to a set of N nodes from the distribution $p_{l,t}$. Edges can then be created by

1. generating a set of half-lines or "stubs", such that the number of times a node appears in the set is equal to the number of lines to which it belongs,
2. generating a set of "corners", such that the number of times a node appears in the set is equal to the number of triangles to which it belongs,
3. ensuring that the number of stubs is divisible by two and the number of corners is divisible by three, for example by randomly deleting any remainder,
4. repeatedly constructing an edge between two stubs drawn at random and without replacement,
5. and, repeatedly constructing edges between three corners drawn at random and without replacement.

This algorithm may produce loops and double-edges, but the frequency of such edges will be negligibly small for large graphs [27], and we simply delete them if they do occur.

Disease transmission through clustered random networks

The ODEs that describe epidemic dynamics in clustered networks can be expressed in several equivalent forms and derived from at least two different perspectives. Below, we present two systems of equations that respectively describe the change in the number of cliques with i susceptible and j infectious nodes and the probability that a susceptible node is connected to such a clique. Both of these systems can also describe the dynamics of the number of infected and susceptible individuals in the population as a function of time. First we present the system of equations based on the probabilities ϕ_X that a random node u is connected by a line to a node in state X and the probabilities ϕ_{XY} that u is connected in a triangle to two nodes in states X and Y. Below we present an alternative derivation based on the numbers of cliques with different configurations. The derivation of this system is very similar to what was presented in [28], but is less mathematically parsimonious than the system of equations in this section, which requires only 7 ODEs. And, below we show how this system can be extended to networks with generalized distributions of clique sizes, that is, networks that include cliques larger than size three.

We follow the recently introduced edge-based compartmental modeling approach of [18]. This approach is based on the consideration of the fate of a single randomly chosen node u in the network. The probability this node is susceptible is equal to the proportion of nodes that are susceptible, and the probability it is infected or recovered is similarly the proportion of nodes that are infected or recovered. If we know the probability the node is susceptible as a function of time, then we can calculate its probability of being infected or recovered, so we focus our attention on calculating $S(t)$, the probability the randomly chosen test node is susceptible. Following [18] we modify the test node so that it does not transmit infection once infected. This does not alter the probability it is susceptible, but eliminates some conditional probability arguments we would have to consider otherwise.

Assume u is a member of l lines and t triangles. Then the probability it is susceptible is $\theta_2^l\theta_3^t$ where θ_2 is the probability that a random line has not transmitted to the test node and θ_3 is the probability that neither of the other nodes in a triangle has transmitted to the test node. So assuming we can calculate θ_2 and θ_3 as functions of time, we have S as a function of time. From this we use $I=1-S-R$ and $\dot{R}=\gamma I$ to find I and R.

Let us first consider θ_2. We divide θ_2 into ϕ_S, ϕ_I, and ϕ_R, the probabilities that a neighbor along a line has not transmitted infection to u and is either susceptible, infected, or recovered respectively. The probability the neighbor has not transmitted is

$$\theta_2 = \phi_S + \phi_I + \phi_R, \tag{4}$$

and $1-\theta_2$ is the probability that it has transmitted. We create compartments for these states and display the flux between them in Figure 1.

The fluxes from ϕ_I to ϕ_R and $1-\theta_2$ are proportional to each other, and each begins as zero, so we can show that $\phi_R = \frac{\gamma}{\beta}(1-\theta_2)$. We find ϕ_S by a different approach, similar to the calculation of S. A neighbor found along a randomly chosen line will tend to have more lines than a node chosen uniformly at random. The random number of such lines is described by the *excess degree distribution* [29], and we calculate the generating function for this distribution as follows. Denote $q_{l,t} \propto l p_{l,t}$ to be the probability that there are l lines and t triangles connected to a susceptible node that we reach by following a line from an infectious to a susceptible node not counting the line by which we arrived. Similarly, $r_{l,t} \propto t p_{l,t}$ is the probability that if we follow a triangle to a susceptible node, there are l lines and t triangles connected to that node, not counting the one by which we arrived. Then we have the generating functions

$$g_q(x,y) = \sum_{l,t} q_{l,t} x^l y^t = g^{(x)}(\theta_2 x, \theta_3 y)/g^{(x)}(\theta_2,\theta_3) \tag{5}$$

$$g_r(x,y) = \sum_{l,t} r_{l,t} x^l y^t = g^{(y)}(\theta_2 x, \theta_3 y)/g^{(y)}(\theta_2,\theta_3). \tag{6}$$

Equations 5 and 6 generate the excess degree distributions for lines and triangles.

A neighbor reached by following a line connected to u is susceptible with probability $\theta_2^{l-1}\theta_3^t$ (recall that u does not cause infection) where l is a realization of the excess degree distribution. Summing over values of l, we find $\phi_S = \sum_{l,t} l p_{l,t} \theta_2^{l-1}\theta_3^t / g^{(x)}(1,1) = g^x(\theta_2,\theta_3)/g^{(x)}(1,1)$. Now we rearrange equation 4 which gives $\phi_I = \theta_2 - \gamma(1-\theta_2)/\beta - g^{(x)}(\theta_2,\theta_3)/g^{(x)}(1,1)$.

We can finally calculate θ_2 by noting that Figure 1 shows $\dot{\theta}_2 = -\beta\phi_I$. We find

$$\dot{\theta}_2 = -\beta\theta_2 + \beta\frac{g^{(x)}(\theta_2,\theta_3)}{g^{(x)}(1,1)} + \gamma(1-\theta_2). \tag{7}$$

To complete the system, we need a corresponding equation for θ_3. Here the system is more complicated. For the line case, if the neighbor had not transmitted, there were just three states to consider. But when considering triangles, if neither neighbor has transmitted, there are $\binom{3}{2} = 6$ states to consider. We define ϕ_{SS} to be the probability both neighbors are susceptible, ϕ_{SI} to be the probability one neighbor is susceptible, while the other is infected but has not transmitted to u, ϕ_{II} to be the probability both are infected but neither has transmitted to u, and similarly define ϕ_{SR}, ϕ_{IR}, and ϕ_{RR}. Figure 1 shows the compartments and flux between them.

We do not have a simple relation for ϕ_{RR} and θ_3, so our derivation changes mildly. The starting point will be $\dot{\theta}_3$, which satisfies

$$\dot{\theta}_3 = -\beta\phi_{SI} - 2\beta\phi_{II} - \beta\phi_{IR}.$$

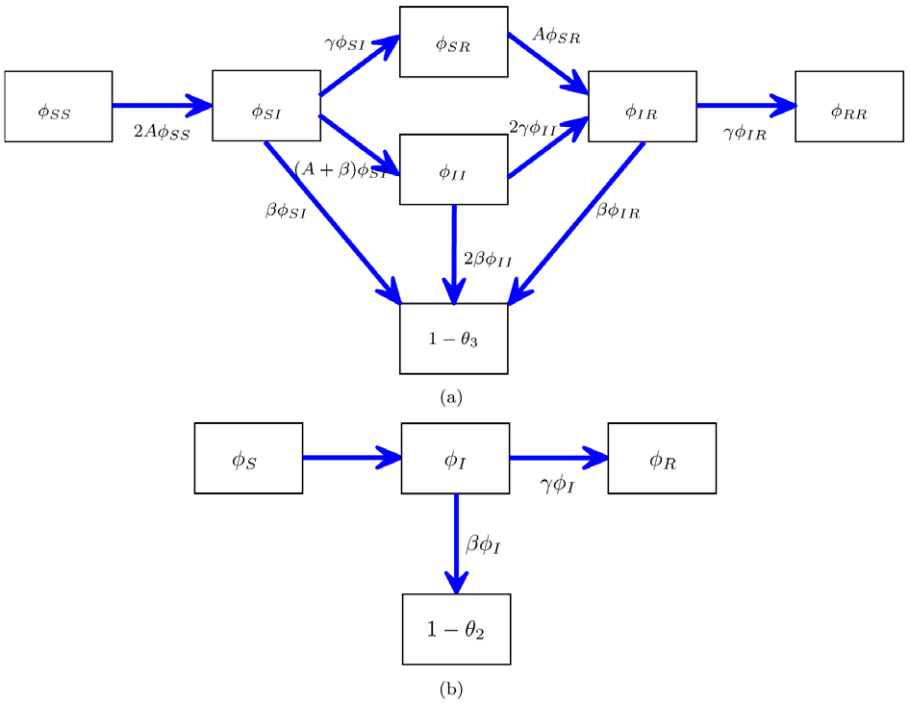

Figure 1. A schematic of the system of equations 7–8. A: The flux between the probabilities that a node u is connected to a triangle with all possible configurations as well as the probability that a node $v \neq u$ in the triangle has transmitted to u. B: The flux between the probabilities that a node u is connected by a line to a node v that is susceptible, infectious, recovered, and the probability that v has transmitted to u.
doi:10.1371/journal.pcbi.1002042.g001

To calculate the right hand side, we first find ϕ_{SS}, the probability that both neighbors in a triangle are still susceptible. Under the assumption that transmissions have not happened in the triangle, the probability that one neighbor is still susceptible is $\sum_{l,t} tp_{l,t}\theta_2^l\theta_3^{t-1}/g^{(y)}(1,1) = g^{(y)}(\theta_2,\theta_3)/g^{(y)}(1,1)$. Since we require both be susceptible,

$$\phi_{SS} = \left(\frac{g^{(y)}(\theta_2,\theta_3)}{g^{(y)}(1,1)}\right)^2.$$

We take A to be the rate that a neighbor in a triangle is infected from outside the triangle. Then $A = -\dot{\phi}_{SS}/2\phi_{SS}$. After some simplification, we find

$$A = -\frac{g^{(x,y)}(\theta_2,\theta_3)\dot{\theta}_2 + g^{(y,y)}(\theta_2,\theta_3)\dot{\theta}_3}{g^{(y)}(1,1)}.$$

We are now ready to find equations for ϕ_{SI}, ϕ_{II} and ϕ_{IR}. We will also need to find ϕ_{SR} to complete the system, but we will not need ϕ_{RR}. We find

$$\begin{aligned}
\dot{\phi}_{SI} &= 2A\phi_{SS} - (\beta+\gamma+2\beta)\phi_{SI} \\
\dot{\phi}_{SR} &= \gamma\phi_{SI} - A\phi_{SR} \\
\dot{\phi}_{II} &= (A+\beta)\phi_{SI} - 2(\beta+\gamma)\phi_{II} \\
\dot{\phi}_{IR} &= A\phi_{SR} + 2\gamma\phi_I I - (\beta+\gamma)\phi_{IR}
\end{aligned} \tag{8}$$

This completes our system of equations. We are able to calculate θ_2 and θ_3 as functions of time, which in turn leads to S, from which we can find I and R as well:

$$\dot{R} = \gamma I, \quad S(t) = g(\theta_2,\theta_3), \quad I = 1 - S - R. \tag{9}$$

Alternative derivation of epidemic dynamics. This model is based on the idea that the number of transmissions events in the network per unit time is a linear function of several time dependent variables:

1. $M_{SI}(t) \propto$ the number of lines that begin at a susceptible node and terminate at an infectious node,
2. $n_{21}(t) \propto$ the number of triangles with two susceptible nodes and one infectious node,
3. $n_{12}(t) \propto$ the number of triangles with one susceptible and two infectious nodes, and
4. $n_{11}(t) \propto$ the number of triangles with one susceptible node, one infectious node, and one recovered node.

The variables M_{XY} are dimensionless quantities that do not depend on N. For comparison to simulations, the number of half-lines M_{XY} would be NM_{XY}. The constant of proportionality depends on the variable under consideration. Given a graph size N, the total number of lines and triangles in the graph are respectively

$$\frac{N}{2}M = \frac{N}{2}\sum_{l,t} lp_{l,t} = \frac{N}{2}g^{(x)}(1,1) \tag{10}$$

$$\frac{N}{3}\hat{M} = \frac{N}{3}\sum_{l,t} tp_{l,t} = \frac{N}{3}g^{(y)}(1,1), \tag{11}$$

since there are 2 nodes per line and 3 per triangle. For the variables n_{ij} defined above, the total number of triangles is Nn_{ij}. And the total number of lines between susceptibles and infected is NM_{SI}. However, below we also use the variable M_{SS} which is proportional to the number of lines connecting two susceptibles. In this case, the total number of such lines is $\frac{N}{2}M_{SS}$ since this variable counts lines twice (once for each susceptible node in the clique).

We will assume that the number of transmissions per unit time over a line or triangle are proportional to

$$\begin{aligned}
T_2 &= \beta M_{SI}, \\
T_3 &= \beta(2n_{21} + 2n_{12} + n_{11}).
\end{aligned}$$

To model epidemic spread, we construct a set of ODEs in terms of the M and n variables as well as two survivor functions for susceptible nodes [17,30]:

1. $\theta_2(t)$: the probability that a neighbor in a "line" has not transmitted infection prior to time t, and
2. $\theta_3(t)$: the probability that both neighbors in a "triangle" have not transmitted infection prior to time t.

The probability that a node with l' lines and t' triangles remains susceptible is $\theta_2^{l'}\theta_3^{t'}$ (see [17,30] for a justification). Consequently the fraction of the population, S, that remains susceptible at any time is

$$S = \sum_{l,t} p_{l,t}\theta_2^l\theta_3^t = g(\theta_2,\theta_3).$$

The probability that an edge beginning at a susceptible node will terminate at an infectious node is M_{SI}/M_S, where M_S is proportional to the number of half-lines or *stubs* connected to susceptible nodes. Similarly, the probability that a susceptible node is connected to a triangle with i susceptible nodes and j infectious nodes is $i \times n_{i,j}/\hat{M}_S$.These two variables can be expressed in terms of the PGF:

$$\begin{aligned}
M_S &= \sum_{l,t} l \times p_{l,t}\theta_2^l\theta_3^t = \theta_2 g^{(x)}(\theta_2,\theta_3), \\
\hat{M}_S &= \sum_{l,t} t \times p_{l,t}\theta_2^l\theta_3^t = \theta_3 g^{(y)}(\theta_2,\theta_3).
\end{aligned}$$

The system of ODEs relies on several more variables derived from the generating function. When a transmission event occurs, lines and triangles that were formally counted among M_{SS} or n_{21} may instead be counted among M_{SI} or n_{12}. Quantifying the magnitude of these changes requires that we calculate the average degree of a newly infected node. This is accomplished with the excess degree distribution and its corresponding generating function [29] (equations 5,6). The mean number of lines and triangles in these joint distributions gives us the expected number of lines or triangles of a newly infected node. We denote the means as δ_{ij}, which is the average excess number of type-j links for a

susceptible node selected with probability proportional to the number of type-i links. Using the generating functions, we have

$$\begin{aligned}\delta_{ll} &= \theta_2 g_q^{(x)}(1,1),\\ \delta_{lt} &= \theta_3 g_q^{(y)}(1,1),\\ \delta_{tt} &= \theta_3 g_r^{(y)}(1,1),\\ \delta_{tl} &= \theta_2 g_r^{(y)}(1,1).\end{aligned} \tag{12}$$

The hazard of infection along a single edge is proportional to the probability that the edge terminates at an infectious node (M_{SI}/M_S) and the transmission rate, implying [17]

$$\dot{\theta}_3 = -\theta_3 \frac{T_3}{\hat{M}_S}, \tag{13}$$

$$\dot{\theta}_2 = -\theta_2 \frac{T_2}{M_S}. \tag{14}$$

Dynamics of M_{SI} and M_{SS} require careful consideration of how edges are rearranged following a transmission event. $\dot{M}_{SS}$ describes the time derivative of the normalized number of lines between susceptibles. T_2 transmissions occur per unit time along lines, and the newly infected individual is connected to an average of δ_{ll} lines in addition to the one by which the individual was infected. The probability that such a line is shared with a susceptible node is the ratio of the number of lines between susceptibles to the total number of half-lines connected to susceptibles: M_{SS}/M_S. Note that this probability does not correspond to what we would have in randomly mixing population, which would just be the fraction of susceptible half-lines in the network: M_S/M. The extent to which M_{SS}/M_S differs from M_S/M reflects the extent to which the state of neighbors in the network is correlated due to the spread of the epidemic. Therefore, M_{SS} will decrease at a rate of $2T_2\delta_{ll}M_{SS}/M_S$.

Furthermore, T_3 transmissions will occur via triangles, and the newly infected node will be connected to an expected number δ_{tl} lines. Each of these will also terminate at a susceptible node with probability M_{SS}/M_S. Then we conclude

$$\dot{M}_{SS} = -2\frac{M_{SS}}{M_S}(T_2\delta_{ll} + T_3\delta_{tl}). \tag{15}$$

The equation for $\dot{M}_{SI}$ can be derived similarly. The edge rearrangement follows a similar pattern as for M_{SS}, but we must account for the increase of M_{SI} when a newly infected node is connected to another susceptible (with probability M_{SS}/M_S) and the decrease of M_{SI} when the new infection has connections to other infecteds (with probability M_{SI}/M_S). Then the new infection has connections to other infecteds (with probability M_{SI}/M_S), yielding terms of the form $(T_2\delta_{ll} + T_3\delta_{tl})M_{XY}/M_X$.

In addition to the edge-rearrangement terms, we must account for changes due to recovery ($-\gamma M_{SI}$) and direct transmission ($-\beta M_{SI}$).

$$\dot{M}_{SI} = -M_{SI}(\gamma+\beta) + (T_2\delta_{ll} + T_3\delta_{tl})\left(\frac{M_{SS}}{M_S} - \frac{M_{SI}}{M_S}\right) \tag{16}$$

Finally, the equations for the number of triangles with i susceptible and j infectious constituents, n_{ij}, is found by considering rearrangements as above, as well as flux between classes that are due to an infectious member of the triangle transmitting to a susceptible member, or recovering. For example, a triangle with one susceptible and two infectious nodes (state (1,2)) will transition to the state (0,3) at the rate 2β, because there are two edges between susceptible and infecteds in this clique. It will also transition to the state (1,1) at the rate 2γ, because there are two infectious nodes in the clique that can recover. To summarize, we find

$$\begin{aligned}\dot{n}_{30} &= -(T_3\delta_{tt} + T_2\delta_{lt})\frac{3n_{30}}{\hat{M}_S},\\ \dot{n}_{21} &= -(2\beta+\gamma)n_{21} + (T_3\delta_{tt} + T_2\delta_{lt})\left(\frac{3n_{30}}{\hat{M}_S} - \frac{2n_{21}}{\hat{M}_S}\right),\\ \dot{n}_{20} &= \gamma n_{21} - (T_3\delta_{tt} + T_2\delta_{lt})\frac{2n_{20}}{\hat{M}_S},\\ \dot{n}_{12} &= 2\beta n_{21} - (2\beta+2\gamma)n_{12} + (T_3\delta_{tt} + T_2\delta_{lt})\left(\frac{2n_{21}}{\hat{M}_S} - \frac{n_{12}}{\hat{M}_S}\right),\\ \dot{n}_{11} &= 2\gamma n_{12} - (\beta+\gamma)n_{11} + (T_3\delta_{tt} + T_2\delta_{lt})\left(\frac{2n_{20}}{\hat{M}_S} - \frac{n_{11}}{\hat{M}_S}\right).\end{aligned} \tag{17}$$

An extra differential equation can be solved for the epidemic prevalence at any time.

$$\begin{aligned}\dot{I} &= -\dot{S} - \gamma I\\ &= \frac{d}{dt}g(\theta_2,\theta_3) - \gamma I\\ &= \dot{\theta}_2 g^{(x)}(\theta_2,\theta_3) + \dot{\theta}_3 g^{(y)}(\theta_2,\theta_3) - \gamma I\end{aligned} \tag{18}$$

This system can also be related to the one in the previous section by the change of variables $\phi_X = M_{SX}/M_S$ and $\phi_{SI} = n_{11}/\hat{M}_S$, etc.

If an initial fraction $\varepsilon \ll 1$ of the population is infected at the beginning of the epidemic and the total number of lines and cliques are respectively proportional to M and $\hat{M}$ (equation 3), we use the initial conditions

$$\begin{aligned}\theta_3(0) &= \varepsilon\\ \theta_2(0) &= \varepsilon\\ M_{SI}(0) &= \varepsilon M\\ M_{SS}(0) &= (1-2\varepsilon)M\\ n_{30}(0) &= (1-\varepsilon)\hat{M}/3\\ n_{21}(0) &= \varepsilon\hat{M}/3,\end{aligned} \tag{19}$$

and the remaining variables would be zero.

Generalization to clique sizes >3

It is straightforward to generalize the derivation for triangles (3-cliques) to larger clique sizes, and to furthermore allow the transmission rate to be a function of clique size. Let $n_{i,j}^k$ denote the number of cliques of size k with i susceptible and j infectious nodes. We will generalize the preceding model to allow transmission rates to vary between cliques of different sizes. The

transmission rate for edges within a clique of size k will be denoted β_k. We consider clique sizes from $k=3$ to a maximum of m. Having multiple clique sizes requires us to introduce additional dummy variables into the generating function. The vector y of dummy variables with elements $y_2, y_3, \cdots, y_m$ correspond to each of the $m-2$ clique sizes and unclustered edges. Note that the element y_2 is the dummy variable corresponding to lines, previously denoted x. Then the following will generate the degree distribution:

$$g(y) = \sum_{t_2,t_3,t_4,\cdots,t_m} p_{t_2,t_3,\cdots,t_m} \prod_{k=2}^{m} y_k^{t_k}.$$

θ will be the vector of survivor functions with elements $\theta_2, \theta_3, \cdots, \theta_m$.

Letting the derivative of g with respect to the dummy variable y_k be denoted $g^{(y_k)}$, the number of cliques of size k in the network is proportional to $\hat{\boldsymbol{M}}^k := g^{(y_k)}(\vec{1})/k$ (because there are k nodes for every k clique). In addition, the number of links from susceptible nodes to cliques of size k is $\hat{\boldsymbol{M}}_S^k = \theta_k g^{(y_k)}(\vec{\theta})$.

We find the dynamics of n^k by tabulating the flux to and from cliques with similar configurations. A $k-$clique with i susceptible and j infectious nodes will have $i \times j$ edges between susceptible and infectious nodes, so that transmissions within cliques will occur at the rate $\beta_k ij$. The rate of transmissions that occur within cliques of size k is

$$T_k = \beta_k \sum_{i=1}^{k-1} \sum_{j=1}^{k-i} ij n_{i,j}^k.$$

The rate of transmissions by unclustered edges will be $T_2 = \beta_2 M_{SI}$, and nodes in cliques of size $k>2$ with i susceptible and j infectious nodes will be infected from outside of the clique (i.e. by an edge with an infectious node not in the clique) at a rate

$$r(i,j,k) := \left(\sum_{l=2}^{m} T_l \delta_{lk} \right) \frac{i n_{ij}^k}{\hat{\boldsymbol{M}}_S^k},$$

where δ_{lk} is the average number of k cliques of a node selected by randomly choosing a susceptible member of a random $l-$clique:

$$\delta_{lk} = \theta_k g^{(y_l,y_k)}(\theta)/g^{(y_l)}(\theta).$$

A clique with j infectious nodes will have recovery events at the rate γj.

Putting these terms together yields the following solution for the dynamics of $n_{i,j}^k$. These equations are defined for all i and j such that $i+j \le k$.

$$\dot{n}_{i,j}^k = \begin{cases} r(i+1,j-1,k) + \beta_k (i+1)(j-1) n_{i+1,j-1}^k + \gamma (j+1) n_{i,j+1}^k \\ -r(i,j,k) - \beta_k ij n_{i,j}^k - \gamma j n_{i,j}^k & \text{if } i < k \text{ and } j > 0, \\ \gamma (j+1) n_{i,j+1}^k - r(i,j,k) & \text{else.} \end{cases} \quad (20)$$

The survivor functions will be determined by the following set of differential equations:

$$\dot{\theta}_k = -T_k \theta_k. \quad (21)$$

The equations for M_{SS} and M_{SI} will be the same as equations 16 and 16, except that indirect transmissions by cliques larger than three must be taken into account.

$$\begin{aligned} \dot{\boldsymbol{M}}_{SS} &= -2 \frac{M_{SS}}{M_S} \left(\sum_{j=2}^{m} T_j \delta_{j2} \right) \\ \dot{\boldsymbol{M}}_{SI} &= -M_{SI}(\gamma+\beta) + \left(\sum_{j=2}^{m} T_j \delta_{j2} \right) \left(\frac{M_{SS}}{M_S} - \frac{M_{SI}}{M_S} \right). \end{aligned} \quad (22)$$

Calculation of the survivor functions only requires cliques such that $i>0$ and $i+j \ge 2$, so it is not necessary to solve for all possible configurations of i susceptible and j infectious nodes. In general, if cliques range in size from 3 to m, this will require $\binom{m+1}{2} - 1$ equations.

Bond percolation approximations for final epidemic size

For an infectious disease spreading in a population in which all individuals have the same susceptibility and the same infectiousness and all transmissions are independent, the epidemic process can be exactly represented through a bond percolation process. Consider an individual u chosen to be the initial infection. Assume the per-contact probability of transmission is τ. If we delete each edge of the network with probability $1-\tau$, then the probability that u is in the same component of the residual network as a given set of nodes is equal to the probability that that set of nodes is infected in the epidemic [20,23,31].

However, if there is variable infection duration or some other cause of heterogeneity in infectiousness, this is no longer the case: those individuals with longer infectious period are more infectious. Assuming the only heterogeneities are due to variable infectiousness, it has been shown [22] that in networks without short cycles the final size of large outbreaks depends only on the average infectiousness in the limit of large networks.

When there are short cycles, the size of epidemics does depend on how infectiousness is distributed. The assumption that all individuals have the average infectiousness only gives an upper bound on epidemic size [23,31]. This bound is often a reasonable approximation [9]. Recently, an alternative percolation technique was developed [16] which accounts for variable infectious periods and can accurately calculate final sizes in some clustered networks.

Taking the transmission rate to be β and the recovery rate to be γ, the average probability of infecting a neighbor is $\bar{\tau} = \beta/(\beta+\gamma)$. First, we investigate how closely the bond percolation approach reproduces epidemics with constant transmission and recovery rates for the clustered networks considered here. Second, we present an alternative simple solution for final size in clustered networks that takes variable infectious periods into account.

The original bond percolation method for clustered networks [12,13] can be used to determine the probability that there would be zero, one or two secondary infections following an initial infection in a triangle. If the transmission probability $\bar{\tau}$ is constant, the probability of having one or two secondary infections in a triangle is (refer to [12,13]):

- one secondary infection: $\bar{\alpha}_1 := 2\bar{\tau}(1-\bar{\tau})^2$,
- two secondary infections: $\bar{\alpha}_2 := \bar{\tau}^2 + 2\bar{\tau}^2(1-\bar{\tau})$.

In fact, these probabilities are functions of the infectious period of the initial case in a triangle, which is itself an exponentially distributed random variable. We can solve for the true probabilities by integrating over the infectious period (in this case

t denotes time). Conditional on the infectious period being t, the probability of transmission by single infected to a single neighbor of that infected is $1-e^{-\beta t}$. When the infectious period is exponentially distributed with rate γ, we have the following:

- One secondary infection:

$$\begin{aligned}\alpha_1 &:= \int_0^\infty \gamma e^{-\gamma t}\left(2(1-e^{-\beta t})e^{-\beta t}(1-\bar{\tau})\right)\mathrm{d}t\\ &= 2(1-\bar{\tau})^2 - 2\frac{\gamma}{2\beta+\gamma}(1-\bar{\tau})\\ &= [2\beta/(\gamma+2\beta)][\gamma/(\beta+\gamma)]^2\end{aligned}$$

- Two secondary infections:

$$\begin{aligned}\alpha_2 &:= \int_0^\infty \gamma e^{-\gamma t}\left((1-e^{-\beta t})^2 + 2(1-e^{-\beta t})e^{-\beta t}\bar{\tau}\right)\mathrm{d}t\\ &= 1+(1-2\bar{\tau})\frac{\gamma}{2\beta+\gamma} + 2(1-\bar{\tau})(\bar{\tau}-1)\end{aligned}$$

This distribution is generally different from the one based on $\bar{\alpha}_1$ and $\bar{\alpha}_2$, and the expected number of secondary infections is strictly less with variable infectious periods. To see this, we denote the averages $R=2\alpha_2+\alpha_1$ and $\bar{R}=2\bar{\alpha}_2+\bar{\alpha}_1$, and note that only second order terms of τ will differ between R and $\bar{R}$. We have $\bar{\tau}^2=\beta^2/(\beta+\gamma)^2$, and

$$\begin{aligned}\langle\tau^2\rangle &= \int_0^\infty \gamma e^{-\gamma t}(1-e^{-\beta t})^2\mathrm{d}t\\ &= \frac{2\beta^2}{(\beta+\gamma)(2\beta+\gamma)}\end{aligned} \tag{23}$$

It is straightforward to see that $\langle\tau^2\rangle>\bar{\tau}^2$. Furthermore, if we collect all terms involving τ^2 in the equation for R, we find a leading factor of $-2\bar{\tau}$. Consequently, these terms will be negative and will have larger magnitude in the expression for R than for $\bar{R}$, so $R<\bar{R}$.

Now we present an asymptotically exact solution for final epidemic size. Let u be a random node. Let q_2 be the probability that a neighbor of u along a line is not infected from another node at the end of the epidemic. Then following the methods described in [12,13], this probability must satisfy

$$q_2 = \frac{g^{(x)}(\theta_2(\infty),\theta_3(\infty))}{g^{(x)}(1,1)}. \tag{24}$$

Similarly, let q_3 be the probability that a neighbor in a triangle never receives an infectious dose from outside that triangle.

$$q_3 = \frac{g^{(y)}(\theta_2(\infty),\theta_3(\infty))}{g^{(y)}(1,1)} \tag{25}$$

We need to calculate θ_2 and θ_3 at ∞ in order to calculate final epidemic size. It suffices to find θ_2 and θ_3 in terms of q_2 and q_3 and then solve the system.

We have θ_2 is the probability that a line does not transmit to u. Clearly this can be calculated by considering the probability the neighbor is never infected plus the probability the neighbor is infected, but does not infect u. This is

$$\theta_2(\infty) = q_2 + (1-q_2)(1-\bar{\tau}) \tag{26}$$

Finding θ_3 is slightly harder. This is the probability that neither neighbor in a 3-clique is infected from outside, or exactly one receives infection from outside, or both receive infection from outside and transmission does not reach u. As above, α_1 is the probability that a node in a triangle will lead to exactly one further transmission within the triangle, and α_0 will be the probability it will cause no transmissions. The probability that an infected neighbor in a triangle recovers prior to transmitting to either of its neighbors is $\alpha_0=\gamma/(\gamma+2\beta)$. Then

$$\theta_3(\infty) = q_3^2 + 2q_3(1-q_3)\left(\frac{\alpha_1}{2}+\alpha_0\right) + (1-q_3)^2(1-\bar{\tau})^2 \tag{27}$$

The first term means neither neighbor is infected. The second term has exactly one neighbor infected (factor of 2 because there are two choices), with the neighbor either infecting the other neighbor, but nothing further or the neighbor infects no one. The third term is both getting infected from outside; we do not need to consider the correlations in this case.

Equations 24–25 can be solved numerically by iteration from small initial values of q_2 and q_3 [12,13]. Given $\theta_2(\infty)$ and $\theta_3(\infty)$, the final size can be calculated:

$$R(\infty) = 1 - g(\theta_2(\infty),\theta_3(\infty)). \tag{28}$$

In the SI, we show how these calculations can be extended to models with generalized distributions of clique sizes.

Comparison to alternative models

To validate the model assumptions, we compare solutions of the system given by equations 13–17 to stochastic simulations in continuous time based on the Gillespie algorithm [32]. Random networks are generated as described above. At time $t=0$, a number of $I(0)$ initial infections are selected uniformly at random within the network. When a susceptible is infected, new transmission and recovery events are queued with exponentially distributed waiting times.

We also compare our model to a similar model consisting of ODEs based on moment-closure [19]. This model was developed for networks with a given degree distribution generated by $G(x)$ and a clustering coefficient C. Unlike our model, this system does not specify a joint distribution for the number of lines and triangles. Rather, this system is based on the concept that potential triangles, of which a degree k node will have $\binom{k}{2}$, will exist with independent probability ϕ. This system also uses PGFs within a low-dimensional system of ODEs, and proposes that $S=G(\theta)$, with $\dot{\theta}=-\theta\beta[SI]/M_S$, where $[SI]$ is the number of half-edges from a susceptible node that terminates at an infectious node. Equations for $[SI]$ are derived in terms of the number of connected triples, or 2-paths, of nodes that pass through a susceptible. This model makes the approximation that the number of 2-paths connecting two susceptibles and an infected is a simple function of the clustering coefficient ϕ:

$$[SSI] \approx [SS][SI]\frac{G''(\theta)}{N(G'(\theta))^2}\left((1-C)+CG'(1)\frac{[SI]}{\theta G'(\theta)M_I}\right).$$

The number of 2-paths connecting a susceptible with two infecteds is

$$[ISI] \approx [SI]^2 \frac{G''(\theta)}{N(G'(\theta))^2}\left((1-C)+CG'(1)N\frac{[II]}{M_I^2}\right).$$

We will subsequently refer to this as the House-Keeling (HK) model.

Results

We used our low-dimensional model to explore the interactions between the variance of the degree distribution and the level of clustering, as they impact epidemic dynamics. To do this, we constructed a negative binomial degree distribution which allows us to hold the mean degree constant while interpolating variances that range from the mean of the distribution to infinity. The negative binomial distribution with parameters p and r is generated by

$$g_{nb}(x;r,p)=\left(\frac{p}{1-(1-p)x}\right)^r. \quad (29)$$

We modified this distribution so that a tuneable fraction p_t of edges are part of a triangle while keeping the mean of the distribution constant. To construct this distribution, we modify the PGF so that all edges occur in pairs; the degree will always be an even integer. The number of *pairs* of edges follows a negative binomial distribution. With probability p_t, a pair of edges is part of a triangle, and with probability $1-p_t$, the pair of edges forms two lines with nodes that are not themselves connected. Because lines always appear in pairs, it is easy to keep the mean of the distribution constant while tuning the amount of clustering with p_t, which can range between zero and one. Then given a random number k 2-tuples generated by equation 29, the number of lines and triangles was generated by $((1-p_t)x^2+p_ty)^k$, where y is the dummy variable for triangles, and x is the dummy variable for lines. Note that the exponent of 2 for x causes all lines to occur in pairs. Using the composition property of PGFs, the degree distribution can be generated by

$$g(x,y)=g_{nb}((1-p_t)x^2+p_ty). \quad (30)$$

We compared solutions of the clustering model to 50 stochastic simulations on random networks with 5,000 nodes and 10 initial infections (Figure 1).

The degree distribution was generated by equation 30, with a mean of 2 and a variance of 3. The fraction of edges that are part of a triangle was $p_t=90\%$. For comparison, we also plot a solution to the clustering model with $p_t=0$, so that there is no clustering. Our results show that clustering slows the epidemic and reduces the final number ultimately infected. The system of equations 13–17 correctly predicts the final size, while the trajectory passes through the central mass of simulated trajectories. The analytical model approximately corresponds to the median time for a stochastic simulation to reach a given prevalence.

We examined the effects of clustering on the final size of the epidemic (Figure 2). The clustering model (equations 13–17) correctly reproduces the final epidemic size observed in simulations. However, the MN percolation solution [12,13] is noticeably biased for non-zero clustering, although it does correctly trend downwards (Figure 3). Over-estimation of the final epidemic size by the MN model is expected because the number of secondary infections within a triangle is overestimated when the infectious period is not constant, as detailed in the methods section.

To calibrate the HK model with our chosen p_t, we used the univariate generating function

$$G(x)=g(x,x^2),$$

as there are two edges for every triangle. The HK clustering model also overestimates final size for this class of random graph, which is not unexpected, because the HK model assumes a different mechanism for generating transitivity in the network. The lack of alignment between the HK model and equations 13–17 indicates that clustering can impact disease dynamics not only through macroscopic effects such as the clustering coefficient, but also through microscopic characteristics. As we show below, the discrepancy between the HK and clustering models is greatest when the variance of the degree distribution is low; and the large discrepancy between the two models in Figure 3 occurs at the lowest variance considered.

When we systematically explored the effects of the variance of the degree distribution and clustering on the estimated final size of an epidemic, we found that the final epidemic size decreases as clustering increases (Supporting Figure 1 in Text S1). Consistent with previous studies, the final size usually decreases as variance increases. This can happen, for example, if the degree distribution has more nodes with degree $=1$ when it is more skewed, which are easily isolated from the giant component. There is an exception, however, when the variance is very small, and clustering is high. In this region, with variance between 1 and 1.5, the final epidemic size can actually increase with larger variance.

We also examined the bias (absolute difference from the true value) of alternative calculations of final size as a function of the variance of the degree distribution and clustering (Supporting Figure 1 in Text S1). The bias of percolation approximations increases with clustering in all cases. However bias is insubstantial when the variance is large, even if clustering is also large. This is a result, at least in part, of the nonlinear relationship between p_t and the clustering coefficient. Given a constant fraction of links to triangles, p_t, the number of triangles in the network is

$$N_\Delta=\sum_{l,t} t\times p_{l,t}/3 = g^{(y)}(1,1)/3,$$

which is constant with respect to the variance of the degree distribution (holding the mean constant). The number of paths with two edges, that is the number of connected triples is

$$N_3=\frac{1}{2}G''(1)=\sum_{k=l+2t} p_k\binom{k}{2},$$

which increases with the second moment of the distribution ($\sum_k p_k k^2$). Thus, increasing the variance of the distribution (holding the mean constant) decreases the ratio of N_Δ to N_3. The clustering coefficient, $\phi=3N_\Delta/N_3$ is more important than the total number of triangles in determining epidemic outcomes. As we increase variance, ϕ converges to zero, and the clustering model converges to the percolation and HK model solutions. Variance and ϕ, rather than N_Δ, are the important quantities for determining final size, because as the variance of the degree distribution increases, the mean excess degree, $G''(1)/G'(1)$, also increases. The number of two paths through a node of degree k is

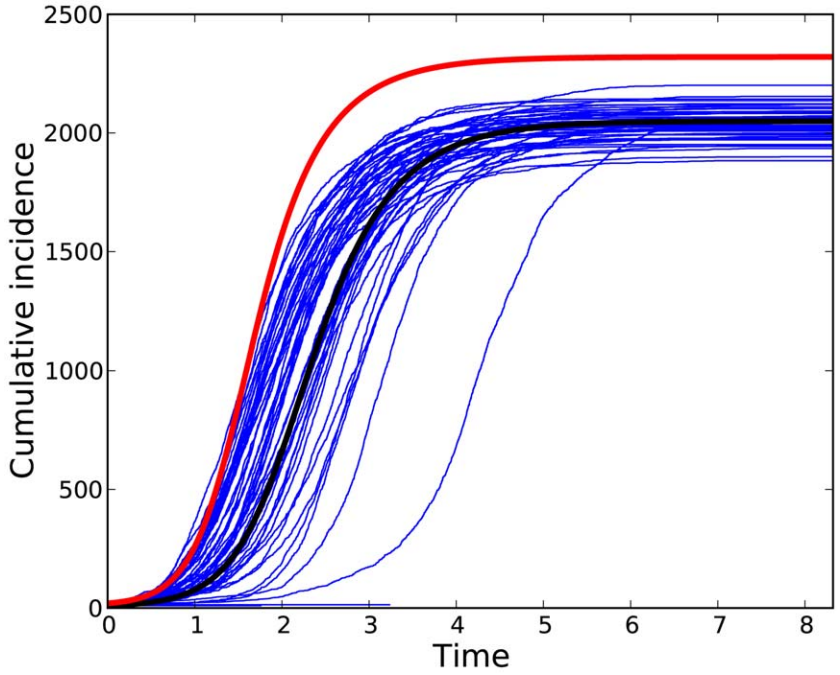

Figure 2. Cumulative number of infections through time. Fifty stochastic simulations (blue dashed lines) are compared to the solution of equations (black line) 13–17. The degree distribution is generated by equation 29 with $p=2/3$ and $r=1/2$. $N=5000, I(0)=10, p_t=0.9, \beta=1.5$, and $\gamma=1$. For comparison, a trajectory with $p_t=0$ is shown in red.
doi:10.1371/journal.pcbi.1002042.g002

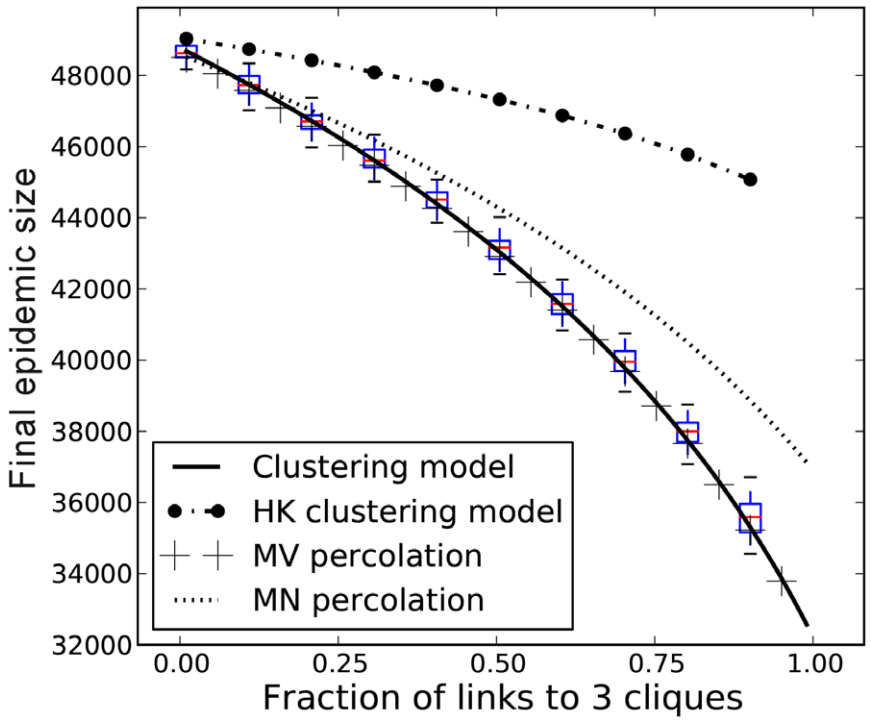

Figure 3. Comparison of clustering models. The degree distribution is Poisson for the number of pairs of edges (mean degree = 2). The black line corresponds to the solution of equations 13–17. The boxplots illustrate the 90% confidence interval from 50 stochastic simulations on networks with 5000 nodes. The remaining trajectories correspond to to the original bond percolation calculations [12,13], our modified bond percolation calculations, and the HK clustering model [19], respectively. $\beta=1.5, \gamma=1$.
doi:10.1371/journal.pcbi.1002042.g003

$\binom{k}{2}$. If we consider a node with mean excess degree $k=G''(1)/G'(1)$, which is the mean degree of a new infected node early in the epidemic, the probability that two neighbors of that node are connected is

$$\frac{p_t}{k-1}=\frac{p_t G'(1)}{G''(1)},$$

which will decrease as variance of the degree distribution decreases.

To measure the timescale of epidemics, we define t_p to be the time to peak incidence, $t_p=\text{argmax}(-\dot{S}(t))$. When we evaluated the influence of degree variance and clustering on the timescale of epidemics, we found that while clustering always slows the epidemic and increases t_p, variance accelerates an epidemic and decreases t_p (Supporting Figure 2 in Text S1). We also found that t_p is much more elastic with respect to variance than p_t (Supporting Figure 2 in Text S1). The HK model is in close agreement with the clustering model (equations 13–17), but can differ by as much as 10% when p_t is large.

The spread of infectious disease through households

Many respiratory diseases such as influenza spread through networks of close-proximity contacts. Transmission can be especially intense within households, where contacts are highly clustered. The clustering of close-proximity contacts that occurs within households is an important factor in the spread of such diseases and such clustering has been the subject of many mathematical models [16,33]. In this section we illustrate how the model in equations 19–21 can be parameterized from real data that includes household contacts. The model developed below is designed for didactic purposes; it does not provide a realistic representation of a specific disease spreading in a specific population. This model excludes a number of complexities, such as age structure, clustering of non-household contacts, and dynamic partnerships. Nonetheless, the model illustrates the conditions under which it is important to include clustering of household contacts. Model misspecification can bias both model predictions and model-based estimates of parameter values.

To parameterize this model, we used data from the POLYMOD study [24], which consists of a sample of 7,290 individuals in eight European countries. These data are diary-based estimates of the number and type of contacts sufficient for transmission of a respiratory pathogen over a 24 hour period. Crucial for our purposes, the data provide a breakdown of contacts made both inside and outside of households. After pooling the data from each country, we find that the number of contacts outside of households was well described by a geometric distribution, which is generated by

$$g_o(x)=\frac{p}{1-(1-p)x}, \tag{31}$$

with $p=0.092$. The geometric distribution was selected by the minimum AIC criterion in comparison with Poisson and negative binomial distributions fit to the data using maximum likelihood. For household sizes, we used the empirical distribution rather than fitting the data to an idealized distribution.To ensure that the system is computationally tractable, we limited the maximum household size at eight, and rounded down any households of larger size; only 2% of households included more than eight individuals. Letting the vector of dummy variables $y=(y_2,\cdots,y_8)$ correspond to household sizes, the following generates the household size distribution:

$$g_h(y)=.08+.13y_2+.18y_3+.21y_4+.14y_5+.09y_6+ .06y_7+.11y_8. \tag{32}$$

The first term in $g_h(y)$ accounts for the probability of living alone. This model assumes that the household size is independent of the number of contacts made outside the household. This approximation is supported by the data, which shows very low correlation between the number of contacts reported within and outside of households (Pearson correlation coefficient $\rho=2.9\%$). Consequently, the generating function for the entire system is the product of marginal PGFs.

$$h(x,y)=g_o(x)g_h(y). \tag{33}$$

For most respiratory diseases, it is reasonable to assume that the transmission rate within households, β_h, is greater than the transmission rate outside of households, β_o [34]. Applying the PGF 32 to the system of equations 19–21 and using the transmission rates β_h and β_o completes the model.

Figure 4 shows the final epidemic size (cumulative number of infections) for the clique model over a range of transmission probabilities both within and outside of households. The transmission probability is the per-edge probability that an infected will transmit prior to recovery, and is $\beta_h/(\beta_h+\gamma)$ within households and $\beta_o/(\beta_o+\gamma)$ outside of households. The final size is much more sensitive to β_o than β_h because the mean number of non-household contacts is much greater than household contacts (10.9 versus 3.3) and the household contacts only occur within cliques.

To determine the epidemiological significance of household clustering, we compared the clique model to a null model that had an identical degree distribution but no clustering. The null model retains household contacts with the transmission rate β_h, but in the null model, such edges do not appear in cliques. In general, the null model without clustering will over-estimate epidemic size. Consequently, null model-based estimates of the epidemiological importance of household contacts will tend to be inflated. The following discussion is oriented around the estimation of epidemic size given epidemic parameters. However, model misspecification will also bias estimates of transmission rates and other parameters made by fitting the network models to empirical epidemic data.

We have identified two sources of bias in the null model without clustering:

1. Clustering by household introduces redundancies relative to the null model that limit transmission, regardless of transmissibilities.
2. When there are two classes of edges (high transmissibility and low transmissibility), the household model aggregates the high transmissibility edges into the redundant parts of the network.

The second factor accounts for most of the bias in this example; clustering alone introduces little bias. For example, comparisons of the null model and clique model with $\beta_h=\beta_o=7.5\%$ and $\gamma=1$ show that true final size is 29%, and the null model is biased by less than 0.36%.

The bias is greatest when transmissibility is high within households, but low outside of households (Figure 3). Outside of this small region, the null model can provide good approximation. Nevertheless, there is good reason to believe that for many real epidemics, the parameters will lie close to the region of high bias. For example, the per-day transmissibility of influenza within

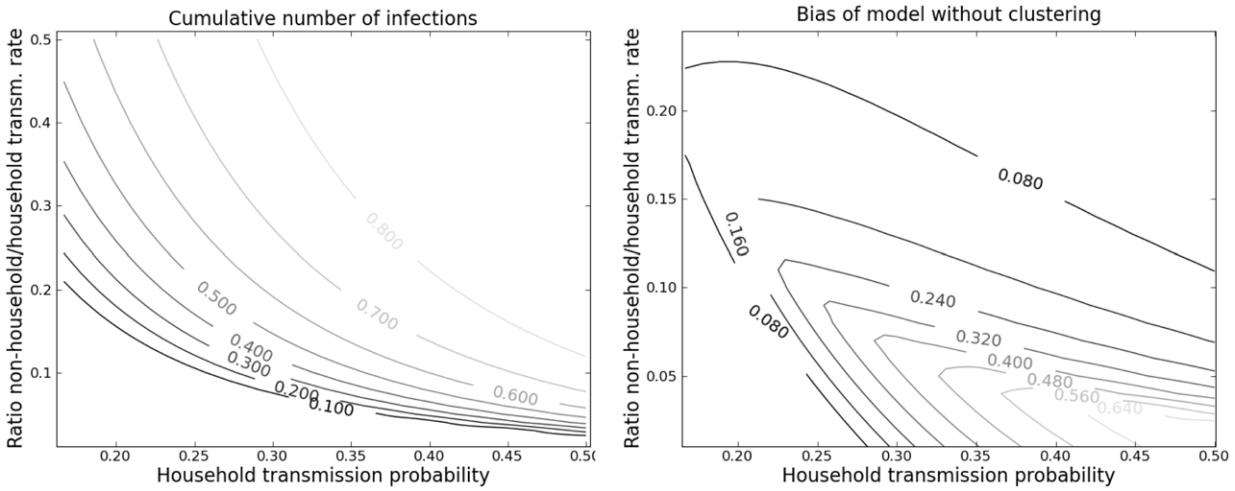

Figure 4. Epidemic size and bias of network models without clustering. Left: The final cumulative number of infections as predicted by the clique model is shown as a function of the transmissibility within households and the transmissibility for contacts outside of households. Right: The difference between final epidemic size in a model without clustering and the final size predicted by the clique model.
doi:10.1371/journal.pcbi.1002042.g004

households has been estimated to be around 5% [34], and based on a 6 day infectious period, this implies a cumulative transmission probability of 20–30%. If transmission rates per edge outside of households are an order of magnitude less than household transmission rates, the network model without clustering may be biased by more than 25%.

Discussion

We have investigated the interactive effects of clustering and the degree distribution of contact networks on the timescale and final size of infectious disease epidemics. For this purpose, we developed a model that generalizes the one presented in [17]. This model has previously been generalized in other dimensions [18], including the incorporation of simultaneous network dynamics, such as edge swapping [30,35], populations with heterogeneous contact rates [36], multiple edge types with distinct transmission rates [28], preferential attachment [28], and growing networks with natural birth and mortality [37]. These extensions can be combined and extended further to model, for example, epidemics in clustered networks that also have dynamically rearranging ties, or networks in which larger clique sizes or other network motifs are prominent [15].

Model selection for epidemic dynamics in networks is a challenging problem; and our work has made two contributions to understanding the biases introduced by model misspecification. We have shown that when infectious periods vary among individuals, models that assume homogeneous transmissibility across all edges in a clustered network can be very biased; and the magnitude of this bias increases with the amount of clustering in the network. In contrast, bond percolation models that neglect variable infectious periods suffer negligible bias in configuration model networks without clustering [21].

The impact of clustering and degree distributions on SIR epidemic dynamics was previously investigated with the HK model [19]. We have compared that model to ours by calibrating the clustering coefficient of the HK model to match the fraction of links to triangles in ours. Our comparison indicates that the models are in close agreement when the variance of the degree distribution is high, but substantial differences in the expected final size and timescale of the epidemic exist when the degree distribution is homogeneous and clustering is extensive. This suggests that epidemic dynamics depend not only on the clustering coefficient, but also on the specific nature of clustering in the network. While the HK model is easy to parameterize when a population has a known clustering coefficient, our model facilitates parameterization using data with well defined cliques, such as human populations with household structure [16,34].

This model allows the number of cliques of different sizes connected to a node to be correlated, but assumes that no two cliques connected to a node share other members. For example, it is not possible for two triangles connected to a node u to share any nodes except for u. However, this feature of the model could be relaxed without much difficulty. A motif-based generalization of the configuration model was recently presented in [15] which provides one way of allowing triangles and other cliques to share more than one node.

Contact data increasingly provide the information necessary to parameterize network models including the one presented here. Social network studies often ascertain degree distributions and clustering coefficients [38,39] and epidemiological surveillance data often provide partnership durations and measures of concurrency [28,40]. We have demonstrated how such data can be used to parameterize the network structure parameters of our model, with a focus on the clustering introduced by household structure, and we have shown the value of explicitly considering this component of human contact patterns in epidemiological models. Without it, models may overestimate both the epidemiological risk of a population and the extent to which household contact contribute to that risk.

Supporting Information

Text S1 Supporting information. This supplement contains supporting Figures and methods. including a bond percolation solution for final epidemic size in models with generalized distributions of clique sizes, and a generalization of the ϕ_{XY} system to clique sizes >3.
(PDF)

Acknowledgments

The authors acknowledge support from NIH U01 GM087719 and thank Thomas House for valuable feedback.

Author Contributions

Conceived and designed the experiments: EMV JCM. Performed the experiments: EMV. Analyzed the data: EMV LAM. Contributed reagents/materials/analysis tools: EMV. Wrote the paper: EMV AG LAM.

References

1. Bansal S, Grenfell B, Meyers L (2007) When individual behaviour matters: homogeneous and network models in epidemiology. J R Soc Interface 4: 879.
2. Newman M (2002) Spread of epidemic disease on networks. Phys Rev E 66: 16128.
3. Newman M, Strogatz S, Watts D (2001) Random graphs with arbitrary degree distributions and their applications. Phys Rev E 64: 26118.
4. Szendroi B, Csányi G (2004) Polynomial epidemics and clustering in contact networks. Proc R Soc Lond B Biol Sci 271: S364.
5. Smieszek T, Fiebig L, Scholz R (2009) Models of epidemics: when contact repetition and clustering should be included. Theor Biol Med Model 6: 11.
6. Pastor-Satorras R, Vespignani A (2001) Epidemic spreading in scale-free networks. Phys Rev Lett 86: 3200–3203.
7. Meyers L (2007) Contact network epidemiology: Bond percolation applied to infectious disease prediction and control. Bull Am Math Soc 44: 63.
8. Newman M (2003) Properties of highly clustered networks. Phys Rev E 68: 26121.
9. Miller J (2009) Spread of infectious disease through clustered populations. J R Soc Interface 6: 1121.
10. Gleeson J, Melnik S, Hackett A (2010) How clustering affects the bond percolation threshold in complex networks. Phys Rev E 81: 066114.
11. Kiss IZ, Green DM (2008) Comment on \properties of highly clustered networks''. Phys Rev E 78: 048101.
12. Miller J (2009) Percolation and epidemics in random clustered networks. Phys Rev E 80: 020901.
13. Newman M (2009) Random graphs with clustering. Phys Rev Lett 103: 58701.
14. Molloy M, Reed B (1995) A critical point for random graphs with a given degree sequence. Random Struct Alg 6: 161–180.
15. Karrer B, Newman M (2010) Random graphs containing arbitrary distributions of subgraphs. Phys Rev E 82: 066118.
16. Ball F, Sirl D, Trapman P (2010) Analysis of a stochastic SIR epidemic on a random network incorporating household structure. Math Biosci 224: 53–73.
17. Volz E (2008) SIR dynamics in random networks with heterogeneous connectivity. J Math Biol 56: 293–310.
18. Miller J (2011) A note on a paper by Erik Volz: SIR dynamics in random networks. J Math Biol 62: 349–358.
19. House T, Keeling M (2011) Insights from unifying modern approximations to infections on networks. J R Soc Interface 8: 67.
20. Kenah E, Robins J (2007) Second look at the spread of epidemics on networks. Phys Rev E 76: 36113.
21. Durrett R (2007) Random Graph Dynamics. Cambridge: Cambridge University Press.
22. Miller JC (2008) Bounding the size and probability of epidemics on networks. J Appl Probab 45: 498–512.
23. Miller JC (2007) Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. Phys Rev E 76: 010101(R).
24. Mossong J, Hens N, Jit M, Beutels P, Auranen K, et al. (2008) Social contacts and mixing patterns relevant to the spread of infectious diseases. PLoS Med 5: e74.
25. Meyers L, Newman M, Martin M, Schrag S (2003) Applying network theory to epidemics: control measures for Mycoplasma pneumoniae outbreaks. Emerg Infect Dis 9: 204–210.
26. Guillaume J, Latapy M (2006) Bipartite graphs as models of complex networks. Physica A 371: 795–813.
27. van der Hofstad R (2010) Random Graphs and Complex Networks. Available: http://www.win.tue.nl/ rhofstad/NotesRGCN.pdf.
28. Volz E, Frost S, Rothenberg R, Meyers L (2010) Epidemiological bridging by injection drug use drives an early HIV epidemic. Epidemics 2: 155–164.
29. Meyers L, Newman M, Pourbohloul B (2006) Predicting epidemics on directed contact networks. J Theor Biol 240: 400–418.
30. Volz E, Meyers L (2007) Susceptible–infected–recovered epidemics in dynamic contact networks. Proc Biol Sci 274: 2925.
31. Trapman P (2007) On analytical approaches to epidemics on networks. Theor Popul Biol 71: 160–173.
32. Gillespie D (1977) Exact stochastic simulation of coupled chemical reactions. The J Phys Chem 81: 2340–2361.
33. Ball F, Mollison D, Scalia-Tomba G (1997) Epidemics with two levels of mixing. Ann Appl Probab 7: 46–89.
34. Longini Jr. I, Koopman J (1982) Household and community transmission parameters from final distributions of infections in households. Biometrics 38: 115–126.
35. Volz E, Meyers L (2009) Epidemic thresholds in dynamic contact networks. J R Soc Interface 6: 233.
36. Volz E (2008) Susceptible-infected-recovered epidemics in populations with heterogeneous contact rates. Eur Phys J B 63: 381–386.
37. Kamp C (2010) Untangling the Interplay between Epidemic Spread and Transmission Network Dynamics. PLoS Comput Biol 6: 1557–1561.
38. Rothenberg R, Long D, Sterk C, Pach A, Potterat J, et al. (2000) The Atlanta Urban Networks Study: a blueprint for endemic transmission. AIDS 14: 2191.
39. Abramovitz D, Volz E, Strathdee S, Patterson T, Vera A, et al. (2009) Using Respondent-Driven Sampling in a Hidden Population at Risk of HIV Infection: Who Do HIV-Positive Recruiters Recruit? Sex Transm Dis 36: 750.
40. Foxman B, Newman M, Percha B, Holmes K, Aral S (2006) Measures of sexual partnerships: Lengths, gaps, overlaps, and sexually transmitted infection. Sex Transm Dis 33: 209.