Percolation and epidemics in random clustered networks

Joel C. Miller[*]

Harvard School of Public Health, Boston, Massachusetts 02115, USA
and Fogarty International Center, National Institutes of Health, Bethesda, Maryland 20892, USA
(Received 29 April 2009; published 4 August 2009)

The social networks that infectious diseases spread along are typically clustered. Because of the close relation between percolation and epidemic spread, the behavior of percolation in such networks gives insight into infectious disease dynamics. A number of authors have studied percolation or epidemics in clustered networks, but the networks often contain preferential contacts in high degree nodes. We introduce a class of random clustered networks and a class of random unclustered networks with the same preferential mixing. Percolation in the clustered networks reduces the component sizes and increases the epidemic threshold compared to the unclustered networks.

Classical random networks contain few short cycles, and the proportion of nodes in short cycles goes to zero as the number of nodes increases. In contrast social networks typically contain many short cycles. We refer to such networks as *clustered* networks. The impact of clustering on percolation properties is usually difficult to calculate because cycles prevent the use of branching process arguments, but it is widely expected that clustering significantly alters percolation.

Typically studies of infectious disease spread assume outbreaks begin with a single infected node. The disease travels to each susceptible neighbor independently with probability T, the *transmissibility*, and the node recovers. The process repeats for each newly infected node. We focus on diseases for which recovery provides immunity. Typically the outbreak dies out stochastically or becomes an epidemic and spreads until the number of susceptible nodes is reduced.

It is well established that for fixed T, epidemic spread can be mapped to a bond percolation problem wherein each edge is kept with probability T [1–6]. If we perform percolation on the network and then choose the initial infection, the disease spreads from that initial infection along edges of the percolated network, and so an epidemic occurs if and only if the initial node is in the giant component. The size of the epidemic matches the size of the giant component. This establishes that the probability and fraction infected in epidemics are equal if T is fixed and all edges are independent [7].

Because social networks frequently exhibit clustering, a number of studies have investigated the impact of clustering on epidemics [8–15]. Some found that clustering reduces the sizes of epidemics and raises the epidemic threshold. That is, clustering reduces the size of giant components and raises the percolation threshold. However, others have shown that clustering appears to reduce the threshold. Consequently epidemics would be possible at lower transmissibility in the presence of clustering.

This discrepancy occurs because there are many ways used to generate clustered networks. It is difficult to separate the impact of clustering from other features introduced by the network generation process.

In this Rapid Communication we introduce an algorithm to generate random clustered networks [16]. The clustered networks have correlations between degrees in a well-defined manner which can lead to assortativity, the tendency for nodes to contact nodes of similar degree. We show how to generate unclustered networks with the same correlations. We make analytic comparisons between the two, clearly separating the effect of clustering from degree correlations. Although clustered networks can have a reduced threshold compared to purely random networks of the same degree distribution, this is an artifact of the assortativity. Compared to an unclustered network of the same degree correlations, clustered networks result in smaller epidemics and higher epidemic thresholds.

This Rapid Communication is organized as follows: we first introduce our clustered and unclustered networks. We then calculate and compare the epidemiological quantity $\mathcal{R}_0$ which measures how many new infections a typical infected node causes. Finally, we calculate the final size or probability of epidemics assuming constant T.

I. NETWORKS

We model our approach after standard configuration model (CM) networks [17,18]. CM networks are useful because all edges from a node are independent of one another in the sense that whether an epidemic results from following one edge is independent of the result along any other edge because short cycles are negligible.

A. Clustered networks

We begin with N nodes. To each node u we assign two degrees, an *independent edge* degree k_I and a *triangle* degree k_Δ. The joint probability of k_I and k_Δ is given by $p(k_I, k_\Delta)$. Then u will be part of k_Δ triangles and have k_I other edges. Each triangle and edge from u will be independent of other triangles, and edges in the same way that edges in CM networks are independent.

We create an independent stub list and a triangle stub list. We place each node u into the independent stub list $k_I(u)$ times and into the triangle stub list $k_\Delta(u)$ times. We shuffle

*joel.c.miller.research@gmail.com

the lists. We join the pairs of nodes in positions $2n$ and $2n+1$ of the independent list and the triples in positions $3n$, $3n+1$, and $3n+2$ of the triangle list. Some repeated edges or loops or short cycles other than the triangles we impose may appear, but their impact is negligible as $N \to \infty$ [19].

This algorithm inevitably segregates those nodes with a high proportion of triangles from those with a low proportion of triangles. If the degrees of nodes with many triangles differ from the degrees of nodes with few triangles, then this effect will cause correlation of different degrees. In order to isolate the impact of clustering, we must compare percolation in these clustered networks with percolation in networks whose nodes are segregated in the same way.

B. Unclustered segregated networks

For comparative purposes we develop a corresponding unclustered network with the same segregation as the clustered networks. Given the joint distribution $p(k_I, k_\Delta)$ of independent and triangle degrees, we create an analogous network where nodes are assigned *blue* and *red* degrees such that $k_b = k_I$ and $k_r = 2k_\Delta$. The joint distribution is given by $p_u(k_b, k_r) = p(k_b, k_r/2)$.

We proceed as before. We create a blue and a red list and pair nodes in positions $2n$ and $2n+1$ in the blue list and then repeat with the red list, joining pairs not triples. The resulting network has the same segregation as the corresponding clustered network, but short cycles are negligible.

II. $\mathcal{R}_0$

$\mathcal{R}_0$ is usually defined as the number of new infections caused by an average infected individual. Occasionally alternate definitions are used, but in some way it represents the number of new infections attributed to an average infected individual. $\mathcal{R}_0 = 1$ is the threshold below which epidemic probability is zero (i.e., the percolated network has no giant component). If $\mathcal{R}_0 > 1$ then epidemics are possible but not guaranteed.

A. Clustered networks

To simplify the analysis, first assume that u, v, and w are members of a triangle and u becomes infectious first. There are multiple ways that both v and w can become infected from edges within the triangle, but they all have the same impact on the epidemic. It is convenient to treat infections of v and w as if they came from u regardless of the actual path followed.

Thus if u becomes infected, then with probability $2T^2(1-T) + T^2 = 3T^2 - 2T^3$ it is credited with infecting both v and w, and with probability $2T(1-T)^2$ it is credited with infecting just 1. With probability $(1-T)^2$ it infects neither. The expected number of infections per triangle is $2T(1+T-T^2)$. In spirit this approach is similar to that of [20]. For bookkeeping purposes, we define the rank s of a node as follows: the index case has rank 0. Each node v is then assigned rank s to be the shortest path of infectious contacts from the index case to v under the rule above for crediting infections.

This allows us to define a 2×2 next-generation matrix [21]. We separate those nodes infected along an independent edge from those infected along a triangle edge [22]. We define c_{II} and $c_{\Delta I}$ to be the number of infections that a node infected from an independent edge is expected to cause along independent and triangle edges, respectively. We similarly define $c_{I\Delta}$ and $c_{\Delta\Delta}$. If $n_I(s)$ and $n_\Delta(s)$ are the number of nodes of rank s which were infected along independent and triangle edges, respectively, then

$$\begin{pmatrix} n_I(s+1) \\ n_\Delta(s+1) \end{pmatrix} = \begin{pmatrix} c_{II} & c_{I\Delta} \\ c_{\Delta I} & c_{\Delta\Delta} \end{pmatrix} \begin{pmatrix} n_I(s) \\ n_\Delta(s) \end{pmatrix},$$

where $c_{II} = \frac{T\langle K_I^2 - K_I \rangle}{\langle K_I \rangle}$, $c_{\Delta I} = \frac{2T(1+T-T^2)\langle K_I K_\Delta \rangle}{\langle K_I \rangle}$, $c_{I\Delta} = \frac{T\langle K_I K_\Delta \rangle}{\langle K_\Delta \rangle}$, and $c_{\Delta\Delta} = \frac{2T(1+T-T^2)\langle K_\Delta^2 - K_\Delta \rangle}{\langle K_\Delta \rangle}$. We give a sample calculation for $c_{\Delta I}$: with probability $k_I p(k_I, k_\Delta)/\langle K_I \rangle$ an infection along an independent edge reaches a node with degrees k_I and k_Δ. The expected number of infections along a triangle edge is $2T(1+T-T^2)k_\Delta$. Thus a random node infected along an independent edge creates $2T(1+T-T^2)\langle K_I K_\Delta \rangle/\langle K_I \rangle$ infections along triangle edges.

The dominant eigenvalue of this matrix is $\mathcal{R}_0$. We generally want to determine T such that $\mathcal{R}_0 < 1$. Substituting $\mathcal{R}_0 = 1$ into the characteristic equation

$$\left(T \frac{\langle K_I^2 - K_I \rangle}{\langle K_I \rangle} - \mathcal{R}_0 \right) \left(2T(1+T-T^2) \frac{\langle K_\Delta^2 - K_\Delta \rangle}{\langle K_\Delta \rangle} - \mathcal{R}_0 \right)$$
$$= 2T^2(1+T-T^2) \frac{\langle K_I K_\Delta \rangle^2}{\langle K_I \rangle \langle K_\Delta \rangle} \tag{1}$$

gives the critical transmissibility $T = T_c$ below which epidemic probability is zero. At the threshold each factor on the left-hand side is at most zero.

The original network has a giant component if $\mathcal{R}_0 > 1$ when $T = 1$. Setting $\mathcal{R}_0 = 1 + \mu$ and $T = 1$ in Eq. (1), we let $\chi(\mu)$ and ψ be the left- and right-hand side, respectively. The concave parabola χ has a negative minimum at $\hat{\mu} = \langle K_I^2 - K_I \rangle/2\langle K_I \rangle + \langle K_\Delta^2 - K_\Delta \rangle/\langle K_\Delta \rangle - 1$. Clearly $\chi(\mu) > \psi$ as $\mu \to \infty$. If $\hat{\mu} > 0$, then the intermediate value theorem guarantees a positive root (greater than $\hat{\mu}$). Similarly, even if the first condition fails, $\chi(0) < \psi$ guarantees a positive root. If both conditions fail there is no positive root. Thus a giant component exists if

$$\frac{\langle K_I^2 - K_I \rangle}{2\langle K_I \rangle} + \frac{\langle K_\Delta^2 - K_\Delta \rangle}{\langle K_\Delta \rangle} > 1$$

and/or

$$\left(\frac{\langle K_I^2 - K_I \rangle}{\langle K_I \rangle} - 1 \right) \left(2\frac{\langle K_\Delta^2 - K_\Delta \rangle}{\langle K_\Delta \rangle} - 1 \right) < 2\frac{\langle K_I K_\Delta \rangle^2}{\langle K_I \rangle \langle K_\Delta \rangle}.$$

If the first condition applies but not the second, then the network has enough independent and triangle edges that a giant component exists solely within the independent edges and another exists solely within the triangle edges. In all other cases the second condition applies.

B. Unclustered segregated network

We define $n_b(s)$ and $n_r(s)$ in the same manner except that triangles need not be considered. Then

$$\begin{pmatrix} n_b(s+1) \\ n_r(s+1) \end{pmatrix} = \begin{pmatrix} c_{bb} & c_{br} \\ c_{rb} & c_{rr} \end{pmatrix} \begin{pmatrix} n_b(s) \\ n_r(s) \end{pmatrix},$$

where $c_{bb} = \frac{T\langle K_b^2 - K_b \rangle}{\langle K_b \rangle}$, $c_{br} = \frac{T\langle K_b K_r \rangle}{\langle K_r \rangle}$, $c_{rb} = \frac{T\langle K_r K_b \rangle}{\langle K_b \rangle}$, and $c_{rr} = \frac{T\langle K_r^2 - K_r \rangle}{\langle K_r \rangle}$. To find the epidemic threshold we substitute $\mathcal{R}_0 = 1$ into the characteristic equation

$$\left(T\frac{\langle K_b^2 - K_b \rangle}{\langle K_b \rangle} - \mathcal{R}_0 \right)\left(T\frac{\langle K_r^2 - K_r \rangle}{\langle K_r \rangle} - \mathcal{R}_0 \right) = T^2 \frac{\langle K_r K_b \rangle^2}{\langle K_r \rangle \langle K_b \rangle}. \tag{2}$$

We divide Eq. (1) by $1 + T - T^2$ and substitute $\langle K_r^2 - K_r \rangle / \langle K_r \rangle = 2(\langle K_\Delta^2 - K_\Delta \rangle / \langle K_\Delta \rangle) + 1$ and $\langle K_r K_b \rangle^2 / \langle K_r \rangle \langle K_b \rangle = 2\langle K_\Delta K_I \rangle^2 / \langle K_\Delta \rangle \langle K_I \rangle$ into Eq. (2). Using the fact that the factors on the left-hand side are negative, comparing terms in the resulting equations shows that the threshold T in the unclustered network is at most the threshold in the corresponding clustered network.

III. CALCULATING GIANT COMPONENT SIZE

To calculate the fraction of nodes in the giant component, it suffices to calculate the probability that a random node is *not* part of the giant component. These calculations have been done for CM networks by [5,6,23] We assume $K_I K_\Delta \neq 0$, so the giant component is unique.

A. Clustered network

We follow the approach of [5,12]. A related approach is given by [6]. We let f be the probability a random node u is not part of the giant component. We have

$$f = \sum_{k_I, k_\Delta} p(k_I, k_\Delta) g_I^{k_I} g_\Delta^{k_\Delta},$$

where g_I and g_Δ are the probabilities that an independent edge or a triangle, respectively, does not connect u to the giant component. To find g_I, we note that there are two ways an edge can fail to connect u to the giant component: it may be deleted in the percolation process with probability $1 - T$, or it may be kept but not connect to the giant component. We have

$$g_I = 1 - T + Th_I,$$

where h_I is the probability that a node v reached along an independent edge is not part of the giant component. To calculate h_I we note that v is selected with probability proportional to k_I but only has $k_I - 1$ susceptible neighbors along independent edges. We get

$$h_I = \frac{1}{\langle K_I \rangle} \sum_{k_I, k_\Delta} k_I p(k_I, k_\Delta) g_I^{k_I - 1} g_\Delta^{k_\Delta}.$$
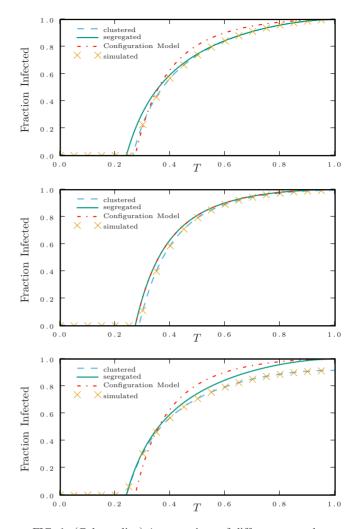
For g_Δ we get

FIG. 1. (Color online) A comparison of different network configurations. Assortative mixing reduces the epidemic threshold. Clustering reduces epidemic size. Simulations on clustered networks match the predictions.

$$g_\Delta = [1 - T + Th_\Delta]^2 - 2T^2(1 - T)h_\Delta(1 - h_\Delta),$$

where h_Δ is the probability a node reached along a triangle edge does not connect to the giant component through any edge not in the triangle. We find

$$h_\Delta = \frac{1}{\langle K_\Delta \rangle} \sum_{k_I, k_\Delta} k_\Delta p(k_I, k_\Delta) g_I^{k_I} g_\Delta^{k_\Delta - 1}.$$

This system of equations for g_I, g_Δ, h_I, and h_Δ can be solved iteratively beginning with $h_I = h_\Delta = 0$, and the result gives f.

B. Unclustered segregated network

To find f_u, the probability a random node in the unclustered network is not part of the giant component, we proceed similarly. We find

$$f_u = \sum_{k_r, k_b} p_u(k_b, k_r) g_b^{k_b} g_r^{k_r},$$

$$g_b = 1 - T + Th_b,$$

$$g_r = 1 - T + T h_r,$$

$$h_b = \frac{1}{\langle K_b \rangle} \sum_{k_r, k_b} k_b p_u(k_b, k_r) g_b^{k_b - 1} g_r^{k_r},$$

$$h_r = \frac{1}{\langle K_r \rangle} \sum_{k_r, k_b} k_r p_u(k_b, k_r) g_b^{k_b} g_r^{k_r - 1}.$$

By iterating beginning with h_b and h_r both zero we find g_b and g_r, from which f_u can be calculated. A similar approach will find f for corresponding clustered networks. At each step of the iteration, $g_r^2 \le g_\Delta$ and $g_b \le g_I$, so this holds in the limit. We conclude $f_u \le f$. Consequently the size of the giant component is smaller and the threshold higher in clustered networks than in unclustered networks of the same degree distribution and degree correlations.

IV. RESULTS

In Fig. 1 we consider outbreak spread on three networks, all of which have the same degree distribution. We compare simulated epidemic sizes with predictions from the clustered equations, the unclustered segregated equations, and the equations derived previously for configuration model networks [5,6,23]. The networks for simulations are clustered networks generated by the algorithm described above.

The nodes are equally distributed between degrees 2, 4, and 6. In each network the clustering is distributed differently. In the first, $p(0,3)=1/3$, $p(2,1)=1/3$, and $p(2,0)=1/3$. That is those nodes with degree 6 are only in triangles, nodes of degree 4 have half of their edges in triangles and independent edges, and nodes of degree 2 have just independent edges. High degree nodes tend to be clustered and contact other high degree nodes. The tendency to contact other high degree nodes reduces the epidemic threshold, but the clustering raises the threshold.

In the second network, we take $p(2,0)=1/6$, $p(0,1)=1/6$, $p(2,1)=1/3$, $p(4,1)=1/6$, and $p(0,3)=1/6$. This yields identical distributions of neighbor degrees for nodes reached by either a triangle or an independent edge. The unclustered segregated equations yield the same result as the configuration model equations. The clustered calculations have smaller epidemics.

The third network is an inversion of the first. Nodes with high degree have independent edges while nodes with low degree are clustered. We take $p(6,0)=1/3$, $p(2,1)=1/3$, and $p(0,1)=1/3$. Again the assortativity reduces the epidemic threshold while clustering reduces the epidemic size. In this particular case, it is the preference for high degree nodes (which are unclustered) to contact one another that leads to the reduction in epidemic threshold, and so it is clear that the effect is due to assortative mixing not clustering.

V. DISCUSSION

We have introduced a model of clustered networks on which we study percolation and epidemics. This model allows us to make a number of analytic predictions because the edges of the network can be partitioned into sets which are independent of one another (independent edges or triangles).

We have shown that these networks can have a lower epidemic threshold than configuration model networks with the same degree distribution. However, this is not a consequence of clustering but rather a consequence of assortative mixing. The clustering of the network can be proven to raise the epidemic threshold and reduce the epidemic size from networks with the same degree correlations but without clustering.

[1] J. L. Cardy and P. Grassberger, J. Phys. A **18**, L267 (1985).

[2] P. Grassberger, Math. Biosci. **63**, 157 (1983).

[3] E. Kenah and J. M. Robins, J. Theor. Biol. **249**, 706 (2007).

[4] D. Ludwig, Math. Biosci. **23**, 33 (1975).

[5] J. C. Miller, Phys. Rev. E **76**, 010101(R) (2007).

[6] M. E. J. Newman, Phys. Rev. E **66**, 016128 (2002).

[7] Care must be taken that no dependence between edges arises. Such a dependence can arise from, for example, heterogeneity in duration of infection [5,23].

[8] S. Bansal, Ph.D. thesis, University of Texas at Austin, 2008.

[9] T. Britton, M. Deijfen, A. N. Lageras, and M. Lindholm, J. Appl. Probab. **45**, 743 (2008).

[10] K. T. D. Eames, Theor. Popul. Biol. **73**, 104 (2008).

[11] M. J. Keeling, Proc. R. Soc. London, Ser. B **266**, 859 (1999).

[12] J. C. Miller, J. R. Soc., Interface (to be published).

[13] M. E. J. Newman, Phys. Rev. E **68**, 026121 (2003).

[14] M. Á. Serrano and M. Boguñá, Phys. Rev. E **74**, 056115 (2006).

[15] M. Á. Serrano and M. Boguñá, Phys. Rev. Lett. **97**, 088701 (2006).

[16] This algorithm was simultaneously developed by Newman [24].

[17] M. Molloy and B. Reed, Random Struct. Algorithms **6**, 161 (1995).

[18] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[19] In essence we have created a generalized edge which corresponds to a triangle. We could create other more general structures in much the same way.

[20] N. G. Becker, K. Glass, Z. Li, and G. K. Aldis, Math. Biosci. **193**, 205 (2005).

[21] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz, J. Math. Biol. **28**, 365 (1990).

[22] Without our simplification, we would need to further subdivide those infected along triangle edges into those whose other neighbor is still susceptible from those whose other neighbor is also infected.

[23] E. Kenah and J. M. Robins, Phys. Rev. E **76**, 036113 (2007).

[24] M. E. J. Newman, e-print arXiv:0903.4009, Phys. Rev. Lett. (to be published).