Percolation in clustered networks

Joel C. Miller
Fogarty Institute and Harvard School of Public Health
(Dated: May 14, 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 clustered networks, but the networks often contain preferential mixing between high degree nodes. We introduce a class of random clustered networks and another class of random unclustered networks with the same preferential mixing. We analytically show that 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 that 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. We focus on diseases for which recovery provides immunity, so recovered nodes are not susceptible. 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 $T$, the epidemic spread can be mapped to a bond percolation problem wherein each edge is kept with probability $T$ [5, 8, 10, 12, 13, 16]. 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 iff 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 [22].

Because social networks frequently exhibit clustering, a number of studies have investigated the impact of clustering on epidemic problems [11, 14, 17, 19, 20]. Some have found that clustering reduces the sizes of epidemics and raises the epidemic threshold. That is, clustering reduces the size of giant components and raise the percolation threshold. However, others have shown that clustering appears to reduce the threshold. Consequently epidemics should be possible at lower transmissibility in the presence of clustering.

This discrepancy occurs because there are many ways used to generate clustered networks, and each new network class results in different behaviors. It is difficult to separate the impact of clustering from other features introduced by the network generation process.

In this article we introduce a new algorithm to generate random clustered networks [28]. 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 can make analytic comparisons between the two, and so clearly separate the effect of clustering from degree correlations. We show that although the clustered networks can have a reduced threshold compared to purely random networks of the same degree distribution, that is entirely an artifact of the assortativity.

Compared to an unclustered network of the same degree correlations, the clustered networks result in smaller epidemics and higher epidemic threshold.

This article is organized as follows: we first introduce our clustered and unclustered networks. We then calculate and compare the epidemiological quantity $R_0$ which measures how many new infections a typical infected node causes. Finally, we calculate the final size/probability of epidemics assuming constant $T$.

I. THE NETWORKS

We model our approach after standard algorithms for Configuration Model (CM) networks [3, 15, 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 $k_\triangle$. The joint probability of $k_I$ and $k_\triangle$ is given by $p(k_I, k_\triangle)$. Then $u$ will be part of $k_\triangle$ 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 \( u \) into the independent stub list \( k_I \) times and into the triangle stub list \( k_\Delta \) times. Once all nodes are placed into the lists, we randomize them. We then take the pairs of nodes in positions \( 2n \) and \( 2n + 1 \) of the independent list and join them, and the triples in positions \( 3n, 3n + 1, \) and \( 3n + 2 \) and join them into a triangle. Some repeated edges or loops or short cycles other than the triangles we impose may appear, but their impact is negligible as \( N \to \infty \) [24].

This algorithm inevitably segregates those nodes with a high proportion of triangles from those nodes with a low proportion of triangles. If the degrees of nodes with many triangles differs from the degrees of nodes with few triangles, then this effect will cause correlation of different degrees. In order to understand the impact of clustering, we must be able to 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 a new 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 \( p_b(k_b, k_r) = p(k_b, k_r/2) \).

We proceed as before. We create a blue and 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. \( R_0 \)

\( R_0 \) 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. \( R_0 = 1 \) is the threshold below which epidemics are impossible (i.e., the percolated network has no giant component). If \( R_0 > 1 \) 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 \( T(1 - T)^2 \) it is credited with infecting just \( 1 \). With probability \( (1 - T)^2 \) it infects neither. In spirit this approach is similar to that of [2]. For book-keeping purposes, we define the rank \( s \) of a node as follows: the index case is given rank 0. Each node \( v \) is then assigned rank \( s \) to be the shortest path from the index case to \( v \), bearing in mind the rule above for crediting infections.

This allows us to define a \( 2 \times 2 \) next-generation matrix [6]. We separate those nodes infected along an independent edge from those nodes infected along a triangle edge [25]. 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 symmetrically 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 \( c_{II} = \frac{T(K_I^2 - K_I)}{K_I} \), \( c_{I\Delta} = \frac{2T(1 + T - T^2)(K_I K_\Delta)}{K_I} \), \( c_{\Delta I} = \frac{T(K_I K_\Delta)}{K_\Delta} \), and \( c_{\Delta \Delta} = \frac{2T(1 + T - T^2)(K_\Delta^2 - K_\Delta)}{K_\Delta} \).

The dominant eigenvalue of this matrix is \( R_0 \). We generally want to determine \( T \) such that \( R_0 < 1 \). Substituting \( R_0 = 1 \) into the characteristic equation gives

\[
\left( T \frac{\langle K_I^2 - K_I \rangle}{\langle K_I \rangle} - 1 \right) \left( 2T(1 + T - T^2) \frac{\langle K_\Delta^2 - K_\Delta \rangle}{\langle K_\Delta \rangle} - 1 \right) = 2T^2(1 + T - T^2) \frac{\langle K_I K_\Delta \rangle^2}{\langle K_I \rangle \langle K_\Delta \rangle},
\]

The value \( T = T_c \) that solves this equation is the threshold transmissibility below which epidemics are impossible.

The original network has a giant component if \( R_0 > 1 \) when \( T = 1 \). Thus the conditions for a giant component are

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

and/or

\[
\left( T \frac{\langle K_I^2 - K_I \rangle}{\langle K_I \rangle} - 1 \right) \left( 2 \frac{\langle K_\Delta^2 - K_\Delta \rangle}{\langle K_\Delta \rangle} - 1 \right) < 2 \frac{\langle K_I K_\Delta \rangle^2}{\langle K_I \rangle \langle K_\Delta \rangle}
\]

The only networks for which the first condition applies but not the second are networks with enough independent edges and triangle edges such that a giant component exists solely within the independent edges and a giant component exists solely within the triangle edges.
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
\[
\begin{align*}
(n_b(s+1)) & = (c_{bb} + c_{br} + c_{rr}) (n_b) \\
(n_r(s+1)) & = (c_{rb} + c_{rr}) (n_r)
\end{align*}
\]
where \( c_{bb} = \frac{T(K_b^2 - K_b)}{(K_b)} \), \( c_{br} = \frac{T(K_b K_r)}{(K_r)} \), \( c_{rb} = \frac{T(K_r K_b)}{(K_b)} \), and \( c_{rr} = \frac{T(K_r^2 - K_r)}{(K_r)} \).
Substituting \( R_0 = 1 \) into the characteristic equation finds the epidemic threshold
\[
\left( T \left( \frac{K_b^2 - K_b}{K_b} \right) - 1 \right) \left( T \left( \frac{K_r^2 - K_r}{K_r} \right) - 1 \right) = T^2 \left( \frac{K_r K_b}{K_b} \right)^2 \left( \frac{K_b}{K_b} \right).
\]
The network has a giant component when
\[
\frac{\langle K_b^2 - K_b \rangle}{2 \langle K_b \rangle} + \frac{\langle K_r^2 - K_r \rangle}{2 \langle K_r \rangle} > 1
\]
and/or
\[
\left( \frac{\langle K_b^2 - K_b \rangle}{\langle K_b \rangle} - 1 \right) \left( \frac{\langle K_r^2 - K_r \rangle}{\langle K_r \rangle} - 1 \right) < \left( \frac{K_r K_b}{K_b} \right)^2 \left( \frac{K_b}{K_b} \right).
\]
The difference between these conditions and those of the corresponding clustered network comes from the fact that
\[
2 \langle K_b^2 - K_b \rangle / \langle K_b \rangle < \langle K_r^2 - K_r \rangle / \langle K_r \rangle.
\]
From this it can be shown that the epidemic threshold occurs at smaller \( T \) for the unclustered 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 [11, 13, 16].

A. Clustered Network

We follow the approach of [13, 14]. A related approach is given by [16].

We let \( f \) be the probability a random node \( u \) is not part of the giant component. We have
\[
f = \sum_{k_I, k_\triangle} p(k_I, k_\triangle) g_I^{k_I} g_\triangle^{k_\triangle},
\]
where \( g_I \) and \( g_\triangle \) are the probabilities that an independent edge or a triangle respectively does not connect 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 \( v \), the node reached, is not part of 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 proportional to \( k_I \), but only has \( k_I - 1 \) susceptible neighbors along independent edges. We get
\[
h_I = \frac{1}{\langle K_I \rangle} \sum_{k_I, k_\triangle} k_I p(k_I, k_\triangle) g_I^{k_I} g_\triangle^{k_\triangle}.
\]
For \( g_\triangle \) we get
\[
g_\triangle = [1 - T + Th_\triangle]^2 - 2T^2(1 - T)h_\triangle(1 - h_\triangle),
\]
where \( h_\triangle \) 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
\[
h_\triangle = \frac{1}{\langle K_\triangle \rangle} \sum_{k_I, k_\triangle} k_\triangle p(k_I, k_\triangle) g_I^{k_I} g_\triangle^{k_\triangle - 1}.
\]
The resulting system of equations for \( g_I \), \( g_\triangle \), \( h_I \), and \( h_\triangle \) can be solved iteratively, 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
\[
f_u = \sum_{k_b, k_r} p_u(k_b, k_r) g_b^{k_b} g_r^{k_r},
\]
where \( g_b = 1 - T + Th_b \)
\[
g_r = 1 - T + Th_r
\]
\[
h_b = \frac{1}{\langle K_b \rangle} \sum_{k_b, k_r} k_b p_u(k_b, k_r) g_b^{k_b - 1} g_r^{k_r},
\]
\[
h_r = \frac{1}{\langle K_r \rangle} \sum_{k_b, k_r} k_r p_u(k_b, k_r) g_b^{k_b} g_r^{k_r - 1}.
\]
It can be shown that \( f_u < f \) for the equivalent degree distributions. Consequently the size of the giant component is smaller in clustered networks than in unclustered networks of the same degree distribution and degree correlations.

IV. RESULTS

In figure [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
V. DISCUSSION

We have introduced a new model of clustered networks on which we study percolation and epidemics. This model allows us to make a number of analytic prediction 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.

Acknowledgments

This work was supported by the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security and the Fogarty International Center, National Institutes of Health.

[1] Shweta Bansal. Ecology of Infectious Diseases with Contact Networks and Percolation Theory. PhD thesis, University of Texas at Austin, 2008.
[2] N.G. Becker, K. Glass, Z. Li, and G.K. Aldis. Controlling emerging infectious diseases like SARS. Mathematical biosciences, 193(2):205–221, 2005.
[3] B. Bollobás. A probabilistic proof of an asymptotic formula for the number of labelled random graphs. European Journal of Combinatorics, 1:311–316, 1980.
[4] T. Britton, M. Deijfen, A.N. Lageras, and M. Lindholm. Epidemics on random graphs with tunable clustering.
[5] John L. Cardy and Peter Grassberger. Epidemic models and percolation. Journal of Physics A: Mathematics and General, 18(6):L267–L271, 1985.
[6] O. Diekmann, J. A. P. Heesterbeek, and J. A. J. Metz. On the definition and the computation of the basic reproduction ratio R0 in models for infectious diseases in heterogeneous populations. Journal of Mathematical Biology, 28:365–382, 1990.
[7] K. T. D. Eames. Modelling disease spread through random and regular contacts in clustered populations. The-
oretical Population Biology, 73:104–111, 2008.

[8] Peter Grassberger. On the critical behavior of the general epidemic process and dynamical percolation. *Mathematical Biosciences*, 63:157–172, 1983.

[9] M. J. Keeling. The effects of local spatial structure on epidemiological invasions. *Proceedings of the Royal Society B: Biological Sciences*, 266(1421):859–867, 1999.

[10] Eben Kenah and James M. Robins. Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing. *Journal of Theoretical Biology*, 249(4):706–722, 2007.

[11] Eben Kenah and James M. Robins. Second look at the spread of epidemics on networks. *Physical Review E*, 76(3):36113, 2007.

[12] D. Ludwig. Final size distributions for epidemics. *Mathematical Biosciences*, 23:33–46, 1975.

[13] Joel C. Miller. Epidemic size and probability in populations with heterogeneous infectivity and susceptibility. *Physical Review E*, 76(1):010101, 2007.

[14] Joel C. Miller. Spread of infectious disease through clustered populations. *Journal of the Royal Society, Interface*, pages ??–??, 2009.

[15] M. Molloy and Bruce Reed. A critical point for random graphs with a given degree sequence. *Random structures & algorithms*, 6(2):161–179, 1995.

[16] Mark E. J. Newman. Spread of epidemic disease on networks. *Physical Review E*, 66(1):16128, 2002.

[17] Mark E. J. Newman. Properties of highly clustered networks. *Physical Review E*, 68(2):026121, 2003.

[18] Mark E. J. Newman. The structure and function of complex networks. *SIAM Review*, 45:167–256, 2003.

[19] MEJ Newman. Random graphs with clustering. *Arxiv preprint arXiv:0903.4009*, 2009.

[20] M. Ángeles Serrano and Marián Boguñá. Clustering in complex networks. II. Percolation properties. *Physical Review E*, 74(5):056115, 2006.

[21] M. Ángeles Serrano and Marián Boguñá. Percolation and epidemic thresholds in clustered networks. *Physical Review Letters*, 97(8):088701, 2006.

Care must be taken that no dependence between edges arises. Such a dependence can arise from, for example, heterogeneity in duration of infection [11, 13].

This algorithm was simultaneously developed by [19].

In essence we have created a generalization of an edge which corresponds to a triangle. We could create other more general structures in much the same way.

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.