Large-scale properties of clustered networks: Implications for disease dynamics

Darren M. Green[a*] & Istvan Z. Kiss[b]

September 21, 2009

[a] *Institute of Aquaculture, University of Stirling, Stirling, Stirlingshire FK9 4LA, UK.*

[b] *Department of Mathematics, University of Sussex, Falmer, Brighton BN1 9RF, UK.*

[*] Tel: +44 1786 467872; Fax: +44 1786 472133.

E-mail address: darren.green@stir.ac.uk (D.M. Green).

Abstract

We consider previously proposed procedures for generating clustered networks and investigate how these procedures lead to differences in network properties other than clustering. We interpret our findings in terms of the effect of the network structure on disease outbreak threshold and disease dynamics. To generate null-model networks for comparison, we implement an assortativity-conserving rewiring algorithm that alters the level of clustering while causing minimal impact on other properties. We show that many theoretical network models used to generate networks with a particular property often lead to significant changes in network properties other than that of interest. For high levels of clustering, different procedures lead to networks that differ in degree heterogeneity and assortativity, and in broader-scale measures such as $\mathcal{R}_0$ and the distribution of shortest path lengths. Hence, care must be taken when investigating the implications of network properties for disease transmission or other dynamic process that the network supports.

Keywords: Networks, Clustering, Epidemic Dynamics, Percolation

1 Introduction

Contact networks are a frequently used tool in epidemiological modelling: Each epidemiological unit (be it a person, animal, self-contained sub-population) is considered as a network *node*, with potentially infectious contact between nodes represented by directionless *edges* or directed *arcs*. The power of the approach is that by explicitly considering the pairwise interactions between units, one can extend the results obtained from compartmental, mean-field, spatial, and metapopulation or household-based models. Direction, strength, and (potentially) timing of contact can all be accounted for. In STI models (Anderson & Garnett, 2000), contact heterogeneity and patterns of connectivity can be accommodated in a straightforward way. They also have the benefit of being able to use epidemiological data directly, as opposed to modelling using summary parameters (e.g. variance in sexual partner count) abstracted from the data.

The principal parameter estimated in epidemiological modelling is that of the basic reproduction number $\mathcal{R}_0$. This may be defined as

> the average number of secondary infections produced when one infected individual is introduced into a [homogeneously mixed,] wholly susceptible host population at equilibrium (Anderson & May, 1991).

However, though for simple models such as the mean field, $\mathcal{R}_0$ is well defined, in general no analytic formula is available for $\mathcal{R}_0$. Moreover, one must consider whether the concept of a single $\mathcal{R}_0$ value is even appropriate in a complex population (Green et al., 2009).

Nevertheless, models of $\mathcal{R}_0$ and final epidemic size are of utility in considering the risk of infectious disease between populations with different structure . Various authors have found that epidemic spread is encouraged or hindered by different network properties. (For an overview, see Shirley & Rushton, 2005.) A node's degree – its number of contacts k – is of key importance, as is the distribution of contacts: networks with a higher variance of degree enjoy a higher $\mathcal{R}_0$ for the same between-node disease transmission rate τ (Anderson & May, 1991). Mixing patterns of contact between nodes are also important. Under proportionate mixing, nodes with contact rates u and v account for a fraction uv of contacts. Deviations from this occur where mixing is preferential, for example *assortative* mixing in homosexual and *disassortative* in heterosexual contact networks. Assortativity increases $\mathcal{R}_0$

but decreases final epidemic size (Ghani et al. 1997; Gupta et al. 1989; Anderson et al. 1990; Newman, 2003a).

In this paper, the network property of most concern is that of clustering. Clustering measures the degree to which 'any friend of yours is a friend of mine'. In clustered networks, if edges (a, b) and (a, c) exist, then connections (b, c) are more likely to exist than would be expected by chance alone in random networks. This is a form of non-random mixing associated with both assortativity and spatial structure. Clustered networks have a lower density of nodes within two steps of a focal node compared with random networks, limiting the spread of disease and reducing $\mathcal{R}_0$ (Keeling, 1999), since nodes infected by the focal node are competing for further neighbouring nodes to infect.

Ideally when comparing networks, one would like to be able to vary one parameter of interest, while keeping all other parameters constant. In this case, we are sure that any differences in network properties are due to that parameter. In practice, this proves difficult. To explore the dependency of epidemic dynamics upon network structure imposed by clustering, various authors have designed different algorithms to generate clustered networks (e.g. Watts & Strogatz, 1998; Newman, 2003b; Read & Keeling, 2003; Eames, 2007; Kiss & Green, 2008). However, these algorithms come from quite different start points, and occasionally there are notable side-effects of clustering upon other network properties (Kiss & Green, 2008).

In this paper, we consider a set of previously used algorithms for generating clustered networks (Newman, 2003b; Read & Keeling, 2003; Eames, 2007) and investigate in what ways these networks differ with otherwise similar degree and clustering coefficients. Here, we are primarily interested in parameters of epidemiological interest, in terms of transmission threshold for epidemic spread, potential epidemic size, and the time-course of disease; though one must also consider the effect of network structure on the effectiveness of control strategies (Kiss et al. 2005; Kiss et al. 2008). Some of these properties will be disease or disease model dependent, however properties such as the distribution of shortest path lengths or departures from proportionate mixing are related. We employ rewiring algorithms to change the clustering coefficients of networks while maintaining other selected network properties constant (Kiss & Green, 2008). Particularly, we wish to preserve the mean degree, degree distribution, and levels of assortative or disassortative mixing.

2 Method

2.1 Network construction

A network is described in terms of its number of nodes N and an adjacency matrix A_{ij}, elements of which are 1 where an edge (i, j) exists, and zero otherwise. The number of edges from a single node i is given by $k_i = \sum_j A_{ij}$. All networks were undirected, generated with $N = 10\,000$ nodes, with mean node degree of either $\langle k \rangle = 5$ or $\langle k \rangle = 10$ and clustering coefficients chosen from $\mathcal{C} = 0.2$, 0.4, 0.6 or no clustering 0.0. A set of 100 networks were generated for each parameter set. The clustering coefficient used is the ratio of triangles to triples, where triples are permutations of three nodes u, v, w with edges (u, v) and (u, w) and triangles are those where an additional edge (v, w) exists. Other definitions of clustering exist (Watts & Strogatz, 1998), but this measure is easy to calculate and epidemiologically useful. A selection of different network types were then generated, either using algorithms reported in the literature, or by rewiring of other networks. These algorithms are described below.

Fixed degree Each node has the same number of edges k, distributed at random by applying $50 \times N$ rewiring operations to a lattice network, where in each rewiring operation four unique nodes with edges (a, b) and (c, d) are rewired to give edges (a, d) and (b, c).

Poisson Random Poisson networks were generated by assigning edges (a, b) for each pair of nodes $a < b$ with a single constant probability.

Iterative An iterative algorithm suggested by Eames (2007) was implemented. This algorithm procedes by repeating two steps. In the first step, n_1 triples are generated by connecting unique nodes a, b, c with edges (a, b) and (a, c). In the second step, n_2 triangles are generated by selecting a node u with at least two neighbours at random, choosing two random neighbours v, w, and forming a link (v, w). Both steps are subject to the constraint that no node may have more than k connections, and duplicate edges are not allowed. The network clustering coefficient is varied by changing n_1 and n_2. Since there is potential for this algorithm to 'stall', it is considered finished if $\frac{kN}{2} \times 0.9975$ edges are successfully assigned.

Spatial This algorithm (Read & Keeling 2003) begins by assigning each node i coordinates x_i and y_i uniformly distributed across a square world of side-length $\sqrt{N}$ with toroidal boundary conditions (the top and bottom, and left and right edges are adjacent). The probability of connection p_{ij} between two nodes i and j is determined by the distance d_{ij} between them, according to $p_{ij} = p_0 \exp(-d^2/2D^2)$ where p_0 and D are parameters to be adjusted to obtain the required $\langle k \rangle$ and $\mathcal{C}$.

Group-based The clustering algorithm of Newman (2003b) has been discussed by the current authors elsewhere (Kiss & Green, 2008). The N nodes are assigned to 'groups' with connections within groups as described below. Multiple group membership by nodes leads to between-group linkages. For each of g groups, ν nodes are chosen at random (without replacement), with nodes thus enjoying a mean of $\mu = gN/\nu$ groups, binomially distributed. For every pair of nodes that are members of the same group, an edge is added with probability $p = \frac{k}{\mu(\nu-1)}$ (with higher probability where multiple groups are shared). The resulting networks have clustering coefficient $\mathcal{C} = \frac{p}{1+\mu(\nu-1)/(\nu-2)}$, adjusted by altering the number of groups per node, μ, subject to the constraint that $p \leq 1$.

Unclustered Clustered networks generated by the **iterative** algorithm had clustering removed using rewiring as for the **fixed degree** networks.

Unclustered preserving mixing Alternatively, rewiring to uncluster networks was carried out by preserving assortativity. In this case edges (u, v) and (w, x) were rewired to (u, x) and (w, v) only where edges were similar in terms of their node degrees: $k_u = k_w$ and $k_v = k_x$. This was carried out for the **spatial** algorithm.

Rewire to cluster Networks created using **fixed degree** or **Poisson** methods were clustered using an iterative rewiring algorithm. At each iteration, a chain of five random nodes u, v, w, x, y with edges (u, v), (v, w), (w, x), and (x, y) was identified without edges (u, y) or (v, x). The effect of rewiring to remove (u, v) and (x, y) and insert (u, y) and

(v, x) edges on a 'local' clustering coefficient is identified (Fig. 1). Where this is increased, the rewiring is accepted. The 'local' clustering coefficient is defined as the ratio of triangles to triples amongst triples a, b, c where node a is one of $u, \ldots, y$. This avoids calculating clustering repeatedly for the whole network.

Reclustered The **group-based** and **spatial** networks were reclustered to preserve clustering coefficients and node degree but remove other forms of structure. This was performed by first unclustering, and then using the **rewire to cluster** algorithm to return the network to its former clustering coefficient.

Small sample networks generated by some of the above procedures are shown in Fig. 2. To compare the properties of networks with the same level of clustering, generated according to different algorithms, we use a series of measures that capture the large scale properties of the network.

2.2 Network measures

Simpler network characteristics such as the distribution and average number of contacts and, for some cases, degree correlations were kept fixed to focus on the differences in large-scale network features. In particular we focus on the measures detailed below:–

Path length In addition to the adjacency matrix A_{ij}, we can calculate a matrix of shortest path lengths L_{ij}, denoting the number of edges required to be followed to travel through the network from node i to j. By definition $L_{ii} = 0$ and where no connecting path exists, $L_{ij} = \infty$. Path lengths were sampled for 10 nodes of each of the 100 networks in each set.

Correlation dimension Borrowing a term from chaos theory, we can use the correlation sum to describe the large-scale structure of a network (Grassberger & Procaccia, 1983). We can calculate this in terms of L as follows:

$$\chi(\varepsilon) = \frac{1}{N^2} \sum_{i,j=1}^{N} H_1(\varepsilon - L_{ij})$$

where $H_1(x) = 1$ for $x \geq 0$ and zero for $x < 0$ (Heaviside step function). Therefore χ

represents the proportion of nodes reached within ε steps through the network, with $\chi(0) = \frac{1}{N}$, $\chi(1) = \frac{\langle k \rangle + 1}{N}$ and $\chi(\varepsilon)$ increasing for higher ε in a manner dependent on network structure. If $\chi(\varepsilon) \propto \varepsilon^{\nu}$ (i.e. a straight line plot of χ *v.* ε on a log-log plot) then we consider the network to have dimension ν. The shape of χ determines the potential trajectory of an epidemic on the network.

Mixing measures We measure the degree to which networks depart from proportionate mixing: in assortative networks, there is preferential connection between nodes with similar degree. In contrast, in a disassortative network, edges are more likely to connect nodes of dissimilar degree than expected with random mixing. Where we write $\sum_{(i,j)} x$ in place of $\sum_{i,j=1}^{N} A_{ij} x$, iterating over all edges $(i, j) \in E$, then a measure of mixing is given by the following correlation coefficient:

$$r = \frac{M \sum_{(i,j)} k_i k_j - \left(\sum_{(i,j)} k_i\right) \left(\sum_{(i,j)} k_j\right)}{M \sum_{(i,j)} (k_i)^2 - \left(\sum_{(i,j)} k_i\right)^2}$$

where $M = \sum_{i,j=1}^{N} A_{ij}$, twice the total number of edges. The correlation r is positive for assortative networks, negative for disassortative, and zero for proportionate mixing.

Eigenvalue analysis The lead eigenvalue λ of the network adjacency matrix can be obtained through simple iteration of the following expression:

$$V^{s+1} = \frac{AV^s}{||AV^s||_1},$$

iterating until convergence, starting with $V_i^0 = 1/n \, (i = 1 \ldots N)$. The notation $|| \cdot ||_1$ indicates that for computational convenience, V is divided by its total at each step. The lead eigenvalue λ is given simply by the solution of $\lambda V^s = AV^s$ where s is large. The lead eigenvalue is related to $\mathcal{R}_0$ as discussed below (Diekmann & Heesterbeek, 2000).

Giant connected component (GCC) A network component is a set of nodes such that a path can be found between any pair of nodes within the group. The largest such component is the giant connected component (GCC). The potential resilience of a network to epidemic spread can be obtained by examining the size of the GCC when a proportion of edges are removed at random. Typically, a sharp percolation threshold is found, analogous to the epidemic threshold found with increasing transmission rate in compartmental models (Newman et al. 2001).

2.3 Simulation model

Epidemic simulation allows numerical determination of the effect of network structure on the threshold value of the transmission rate for epidemic outbreak, i.e. the point at which $\mathcal{R}_0 = 1$, as well as final epidemic size. The time-course of the spread of disease is also obtained.

Epidemic dynamics were simulated using an SIR model. At time t, nodes may be susceptible S, infectious I or removed R. Infection is transmitted at rate τ across every (S, I) edge. The epidemic is seeded with one or more infected nodes. Thereafter, the probability of a node becoming infected depends on the state of its neighbouring nodes. In a small time interval δt, a node with k^I infected neighbours becomes infected with probability $1 - \exp(-k^I \tau \delta t)$. Similarly, recovery/removal is modelled as a Poisson process with the recovery probability given by $1 - \exp(-\gamma \delta t)$, independent of neighbouring nodes. We use synchronous updating with $\gamma = 1$ throughout and a timestep of $\delta t < 0.01$, with ten randomly selected initial seeding nodes.

2.4 Estimates for $\mathcal{R}_0$

Scope of the problem Though $\mathcal{R}_0$ has a simple definition, this simple definition belies a range of problems for its calculation and applicability. In addition, for structured populations, a distinction can be made between the basic reproduction number of the simulated disease $\mathcal{R}_0$, and the transmission potential ρ_0 (May & Lloyd, 2001). The latter can be defined as the average number of secondary cases derived from an index case chosen at random from the population. Unlike $\mathcal{R}_0$, it is a function only of the properties of individual nodes (see caveat later), and independent of network mixing properties. It is therefore a useful baseline for comparison between epidemics with different transmission rates.

No general expression for the basic reproductive number $\mathcal{R}_0$ exists that can be directly calculated from basic network properties. Nevertheless, various estimates for $\mathcal{R}_0$ have been proposed which encapsulate network structure to a greater or lesser extent. Frequently, these are expressed in terms of edge transmissibility $T = \frac{\tau}{\tau+\gamma}$: the probability of transmission between an isolated (S, I) pair during the whole infectious period of the I node (Newman, 2002; Green et al. 2006). The estimates described below capture different subsets of the network properties listed in the previous sections.

Generation-based approach In this approach, an estimate of $\mathcal{R}_0$ is made using the distribution of node degrees and correlation between node degrees of adjacent nodes. Thus, data concerning the spatial or large-scale network structure and clustering are discarded. We consider the *generation* of an infected node to be the number of steps along the infection chain it lies from the index case. We let $I_{i,g}$ denote the number of nodes of degree i in generation g and $I_g = \sum_{i=0}^{\infty} I_{i,g}$ is the total number of infected nodes in generation g. The next generation, $I_{i,g+1}$ is given by

$$I_{i,g+1} = \sum_{j=0}^{\infty} Tjp(i|j)I_{j,g},$$

where $p(i|j)$ is the probability that a node with with j contacts is connected to a node with i contacts, and T is the generation-wide probability of transmission across a link (Kao, 2006). Iterating this calculation allows us to determine the number of infected nodes in consecutive generations, and based on this, calculate $\mathcal{R}_0$ (see Appendix 1). Diekmann & Heesterbeek (2000) have shown that under appropriate conditions, $\mathcal{R}_0$ is given by

$$\mathcal{R}_0 = \lim_{N,n\to\infty} \left(\prod_{g=1}^{n} I_{g+1}/I_g \right)^{1/n}.$$

In this general case, a closed expression for $\mathcal{R}_0$ is difficult to obtain, however for specific networks $p(i|j)$ can be estimated from the network adjacency matrix as follows:

$$p(i|j) = \frac{\sum_{uv} A_{uv} \left[k_u = i\right] \left[k_v = j\right]}{\sum_{uv} A_{uv} \left[k_u = i\right]},$$

where $[x = y]$ gives unity where $x = y$, and zero otherwise. The number of infected nodes in consecutive generations can then be computed under the assumption of networks of infinite size with vanishing number of loops.

Summary statistics We now briefly report on various $\mathcal{R}_0$-like measures, and how they relate to the above analytical approach. Assuming random seeding and $I_0 = 1$, from the equations above we obtain a value of $I_1 = \langle k \rangle T$ (where $\langle k \rangle$ is the mean number of edges per node), which corresponds to the transmission potential, written fully as $\rho_0 = \langle k \rangle \frac{\tau}{\tau+\gamma}$ (Keeling & Grenfell 2000; Green et al. 2006). For a specified ρ_0, the corresponding edge-based transmission rate τ can be calculated as $\tau = \gamma \frac{\rho_0}{\langle k \rangle - \rho_0}$ (Green et al. 2006).

In the case of proportionate random mixing, $p(i|j) = ip(i)/\langle k \rangle$. In this case, it can be shown that $I_2 = T^2 \langle k^2 \rangle$ (Appendix 1; Kao, 2006). In general, I_{g+1}/I_g is constant for any higher value of g and therefore $\mathcal{R}_0 = T \frac{\langle k^2 \rangle}{\langle k \rangle}$. The calculation should ideally be modified for undirected networks to account for a node losing a connection upon becoming infected from its parent case (Kiss et al., 2006). In this, case $R_0 = T \left(\frac{\langle k^2 \rangle}{\langle k \rangle} - 1 \right)$. Note that this correction is not applied to ρ_0, where infection of the focal node is assumed to happen 'by magic', not infection from a linked node.

These expressions account for degree heterogeneity (Anderson & May, 1991), but are only appropriate where there is no higher-level network structure in the form of clustering (Keeling, 1999) or assortativity (Anderson et al. 1990). A further measure is given by the lead eigenvalue of the network adjacency matrix, λ (Diekmann & Heesterbeek, 2000), whose value differs from the previous where the network is broken into dissimilar components (Green et al., 2009).

3 Results

Networks formed through the **iterative** algorithm have a slower increase in χ with path length ε and longer mean path lengths, compared to networks with similar parameter formed through the rewired **fixed-degree** networks (Fig. 3), even for $\chi(2)$, which is in a sense another measure of the degree of local clustering. Examining other network properties, no difference was found in the levels of clustering at the level of squares (proportion of quadruples a, b, c, d with edges (a, b), (b, c) and (c, d) that are also squares with edge (a, d)) with coefficients of $\mathcal{C}_{\square} = 0.44$ and 0.43 respectively. However, an interesting difference was found in the distribution of triangles at the node level (the numbers of triples being fixed at $k(k-1) = 20$), with those of the **iterative** networks having lower variance despite the same mean (17.6 versus 25.3). Local triangle counts were correlated between connected nodes, but there was no difference in the degree of correlation between network types ($r \approx 0.6$). That different measures of clustering are not consistent with each other is not unexpected: the several 'traditional' clustering measures (Soffer & Vázquez, 2005) deviate from each other in different network architectures.

The correlation sum of the **spatial** networks alone shows a trend close to a straight line on the log-log plot (Fig. 3). All other network types show exponential increase (straight line on semi-log plot, Fig. 3, inset) in proportion of network reached with distance. The **spatial** networks are therefore the only ones showing finite dimension, and power-law epidemic spread is expected to be a better model of infection than exponential in epidemics growing on these networks (Szendröi & Csányi, 2004). The ordering of the correlation sum plot slopes is reflected in the timescale of epidemic simulations shown in Fig. 4. In both plots, the slower rise of the **spatial** networks in terms of potential epidemic spread is seen, even for a particular level of clustering, and a lower rise for the clustered networks themselves.

In network-based models of disease transmission, connected components (CCs) play an important role (Newman et al., 2001; Newman, 2002; Kenah & Robins, 2007a,b). Disease seeded into any node in a CC can potentially reach any other node in that component. Thus, for undirected networks, provided that each link will transmit the infection, the size of the largest or giant CC (GCC) represents the upper limit for the potential size of an epidemic. However, for any network only a subset of all edges will be involved in the transmission process. To account for edges that will not be involved in disease transmission, the contact network can be de-constructed or diluted by removing a proportion $1 - p$ $(0 \leq p \leq 1)$ of edges at random (Cohen et al., 2002). This gives rise to a network that can be regarded as the 'epidemiological network' of truly infectious links (Kao et al., 2006).

The emergence and growth of the GCC can be investigated by increasing the value of p. In Fig. 5, the size of the GCC is plotted as a function of p for different network types. For **spatial** networks with high clustering, the GCC is only present for values of p that are considerably higher compared to the case of the re-clustered version of the same network, the **spatial** network with no clustering, and the unclustered version of the **spatial** network but with mixing preserved. This indicates that the structure of **spatial** network limits the epidemic spread and this effect is stronger than for networks with exactly the same level of clustering but obtained using the **reclustering** algorithm. Similar arguments hold for networks with **fixed degree**. However, for **group-based** networks the situation changes and the GCC emerges for smaller values of p compared to the case of **group-based** networks with no clustering. In a previous paper Kiss & Green (2008) have shown that this is a direct consequence of higher clustering leading to higher degree heterogeneity. Although, the GCC appears for smaller values of p as clustering increases, its size is limited and stays relatively small when compared to the unclustered case.

For the case of **spatial** networks, in Fig. 6 the cumulative frequency of the CCs is plotted for below and above percolation regimes. The percolation threshold is given by the value of p at which the GCC emerges (i.e. when the size of the GCC is comparable to the network size in the limit of an infinite network). Here we do not focus on the exact percolation threshold but rather on how components grow and connect together to form the GCC. Fig. 5(a) illustrates that for unclustered networks, the percolation is sharper with a clear transition from having CCs of very small sizes to a single large GCC. However, for high levels of clustering ($\mathcal{C} = 0.6$), the transition is less sharp with CCs continuing to grow almost independently and only merging in a single large GCC for high values of p (see Fig. 5(b)). This illustrates how clustering promotes the local growth of sub-clusters with few inter-cluster links that can lead to a single large component spanning most of the network.

In Table 1, numerical estimates for various $\mathcal{R}_0$-like quantities are given. Apart from $\frac{\chi(2)-\chi(1)}{\chi(1)-\chi(0)}$ all measures are based on the assumption of large networks with no loops. Moreover, $\frac{\langle k^2 \rangle}{\langle k \rangle}$ is only valid when networks are proportionally mixed. However, the value of λ and the generation-based approach captures any departure from proportionate mixing, as demonstrated by the positive correlation between these and the mixing measure r. For the

group-based model, high clustering leads to high contact heterogeneity but no assortativity. Contact heterogeneity alone gives larger $\mathcal{R}_0$ values and a fast spreading epidemic between the subset of highly connected nodes. This is reflected in high values of almost all measures. The eigenvalue approach does particularly well to capture the low level of assortativity generated by high levels of clustering in random or Poisson networks.

4 Discussion

Our results demonstrate that networks exhibiting similar levels of clustering, but generated by different algorithms, can differ significantly in their large-scale structure. This has implications for the spread of disease on such networks. Moreover, tuning a particular network property can lead to undesired but significant changes in network properties other than that of interest, and in a different manner for different network construction algorithms. This hinders accurate determination of the effect of different network properties on the dynamical processes the network supports.

To more accurately capture heterogeneity in contact at the level of individuals, models of disease transmission on contact networks – either data based or theoretical – have become more common. While accurate network data are difficult to collect, many theoretical network models have been developed simply based on partial information or general network characteristics (e.g. small-world networks (Watts & Strogatz, 1998) with short path length and high clustering). Our $\mathcal{R}_0$-like parameter estimates above fall into this category: they are an attempt to summarise the ability of the network to support an epidemic by extracting partial information from it. The information retained and utilised varies between measures, and thus so does the applicability of the measure to different network types. The ability of the measures presented above to capture particular network properties is summarised in Table 2.

The equivalence of various epidemiological network measures is epidemic model- (or rather, disease) dependent. For example, though we define ρ_0 as the number of secondary cases from a randomly chosen index case, with exponentially distributed infectious periods this is in practice an overestimate in individual-based model simulations, since the index case competes with its own secondary cases (and later) for other secondary cases to infect.

The same principle applies to $\mathcal{R}_0$. This effect is present in our network simulations as well as the mean-field model (Appendix 2) and is particularly strong in clustered networks and a large seeding population, but absent in discrete generation-based models.

Many assumptions are implicit in formulations of network epidemic models such as that presented above. For example, we assume that all edges have equal weight and that this is not affected by the number of connections an individual makes, as might be the case under the frequency dependent model paradigm. Other measures of clustering giving different weightings to nodes with dissimilar k may be more appropriate for other network types. We also assume exponentially distributed infectious period lengths, a distribution with a long tail and thus much overlap of generations of infection. With many such other – often more biologically appropriate – approaches available, there is always the danger of letting 'the tail wag the dog', that is being driven by what we usually model, rather than being driven by modelling epidemic problems that need solutions.

Simple analytical approaches can aid the analysis of complex networks. For example, Newman (2002) showed that under some appropriate conditions the transmission of diseases on networks is equivalent to a bond-percolation problem with the possibility to analytically or semi-analytically compute outbreak threshold and outbreak size distribution. Kenah & Robins (2007 a, b) have later on expanded on the precise conditions for such an agreement between the two approaches to hold. However, all these approaches are based on the assumptions of infinite networks with no loops and in some cases proportionate or random mixing. Nevertheless, such theoretical models provide a useful starting point for investigating the effect of any departure from the idealized network models.

Clustering is a local property and the triangular sub-graph structure and their frequency has been generalised to *motifs* (e.g. four nodes in a line or connected in a circle), widely studied in the context of red systems biology (Milo et al., 2002). For example, for gene regulatory networks, certain motifs are more abundant in the network compared to what would be expected at random and these frequently re-occuring small structures are regarded as the building block of networks. For our particular case, different network-generating algorithms could lead to more frequently observing motifs composed of four or more nodes. However, we found no difference in clustering at the level of squares between **iterative** and **fixed-degree** networks. Future work could examine the presence and frequency other larger motifs that could be a by-product of the generating algorithms and could have significant effect on disease transmission.

An important aspect of many disease transmission models is the exploration of the efficacy of different control measures. For example, previous studies have shown that this strongly depends on disease characteristics and contact network properties: Contact tracing performs better on clustered networks (Eames & Keeling 2003; Kiss et al., 2005) where the redundant local links offer multiple opportunities to trace and isolate individuals who have been in contact with infectious individuals. Similarly, with STIs on assortatively mixed networks, contact tracing must be performed quickly or at least at a level that is comparable to the rate of disease transmission (Kiss et al, 2008). Such studies are often based on theoretical network models and focus on investigating the effect of a particular network property. In this paper we have shown that theoretical network models must be used with care and that the analysis of the network itself merits as careful consideration as the dynamical processes that the networks support. Combining network measures that focus on local node properties with large-scale network measures can improve the transparency and accuracy of modelling predictions.

5 Appendices

5.1 Generation-based network approach

Following on from the main text, where we let $I_{i,g}$ denote the number of infected nodes of degree i in generation g, $I_{i,g+1}$ is given by

$$I_{i,g+1} = \sum_{j=0}^{\infty} Tjp(i|j)I_{j,g},$$

where $p(i|j)$ is the probability that a node with with j contacts is connected to a node with i contacts. In the case of proportionate random mixing, $p(i|j) = ip(i)/\left<k\right>$. Hence, given random seeding of initial cases in the zeroth generation such that $I_{j,0} = p(j)$, the number of individuals with degree i in the first generation is

$$I_{i,1} = \sum_j Tj \frac{ip(i)}{\langle k \rangle} I_{j,0} = \frac{Tip(i) \sum_j jp(j)}{\langle k \rangle} = Tip(i)$$

while in the second generation this is

$$I_{i,2} = \sum_j Tj \frac{ip(i)}{\langle k \rangle} I_{j,1} = \frac{T^2 ip(i) \sum_j j^2 p(j)}{\langle k \rangle} = \frac{T^2 \langle k^2 \rangle}{\langle k \rangle} ip(i).$$

Summation according to i gives $I_1 = T \langle k \rangle$ and $I_2 = T^2 \langle k^2 \rangle$. Dividing I_2 by I_1 we obtain the standard estimate for $\mathcal{R}_0$.

5.2 Generation-based mean-field model

The mean-field SIR model can be posed in a way in which the generations of infection may be identified. The infected compartment I is subdivided into compartments indexed by the generation of infection $g \in \mathbb{N}_0$. Infection by generation g produces infected individuals at generation $g+1$, with therefore no flow into the $g = 0$ index case compartment. As usual, β and γ represent the infection and removal rates.

$$\begin{array}{rcll} \frac{dI_g}{dt} & = & \beta S I_{g-1} - \gamma I_g & g > 0 \\ \frac{dI_g}{dt} & = & -\gamma I_g & g = 0 \\ \frac{dS}{dt} & = & -\beta \sum_g I_g & \\ \frac{dR_g}{dt} & = & \gamma I_g & \end{array}$$

Solving this model for $\beta = 3$ and $\gamma = 1$, and an initial infected population of $I_{0,0} = 0.0001$, we obtain a final state of $R_{1,\infty} = 0.000295$, suggesting a value of $\mathcal{R}_0 = 2.95$, less than the theoretical value of $\mathcal{R}_0 = \beta/\gamma = 3$. This theoretical value is approached as $I_{0,0}$ approaches zero.

References

Anderson, R.M., & Garnett, G. (2000). Mathematical models of the transmission and control of sexually transmitted disease. Sex. Transm. Dis. 27: 636 – 643.

Anderson, R.M., Gupta, S. & Ng,W. (1990). The significance of sexual partner contact networks for the transmission of HIV. J. AIDS 3: 417 – 429.

Anderson, R.M. & May, R.M. (1991). Infectious Diseases of Humans: Dynamics and Control. Oxford University Press, Oxford.

Britton, T., Deijfen, M., Lagerås, A.N. & Lindholm, M. (2008) Epidemics on random graphs with tunable clustering. J. Appl. Probab. 45: 743 – 756.

Cohen, R., Ben-Avraham, D. & Havlin, S. (2002) Percolation critical exponents in scale-free networks. Phys. Rev. E. 66: 036. 113.

Diekmann, O. & Heesterbeek, J.A.P. (2000). Mathematical epidemiology of infectious diseases: model building, analysis and interpretation. Wiley, Chichester, UK.

Eames, K.T.D. & Keeling, M.J. (2003). Contact tracing and disease control. Proc. R. Soc. B 270: 2565 – 2571.

Eames, K.T.D. (2007). Modelling disease spread through random and regular contacts in clustered populations. Theor. Pop. Biol. 73: 104 – 111.

Ghani, A., Swinton, J., & Garnett, G. (1997). The role of sexual partnership networks in the epidemiology of gonorrhea. Sex. Transm. Infect. 24: 45 – 56.

Grassberger, P. & Procaccia, I. (1983). Measuring the strangeness of strange attractors. Physica 9D: 189 – 208.

Green, D.M., Gregory, A., & Munro, L.A. (2009). Small- and large-scale network structure of live fish movements in Scotland. Prev. Vet. Med. in press.

Green, D.M., Kiss, I.Z., & Kao, R.R. (2006). Parameterisation of Individual-Based Models: Comparisons with Deterministic Mean-Field Models. J. Theor. Biol. 239: 289 – 297.

Gupta, S., Anderson, R.M., & May, R.M. (1989). Networks of sexual contacts: implications for the pattern of spread of HIV. AIDS 3: 807 – 818.

Kao, R.R. (2006). Evolution of pathogens towards low $\mathcal{R}_0$ in heterogeneous populations. J. Theor. Biol. 242: 634 – 642.

Kao, R.R., Danon, L., Green, D.M. & Kiss, I.Z. (2006) Demographic structure and pathogen dynamics on the network of livestock movements in Great Britain. Proc. R. Soc. Lond. B 273: 1999 – 2007.

Kenah, E. & Robins, J.M. (2007a). Second look at the spread of epidemics on networks. Phys. Rev. E 76: 036113.

Kenah, E. & Robins, J.M. (2007b). Network-based analysis of stochastic SIR epidemic models with random and proportionate mixing. J. Theor. Biol 249: 706 – 722.

Keeling, M.J. (1999). The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. Lond. B 266: 859 – 867.

Keeling, M.J. & Grenfell, B.T. (2000). Individual-based perspectives on R_0. J. Theor. Biol. 203: 51 – 61.

Kiss, I.Z., Green, D.M. & Kao, R.R. (2005). Disease contact tracing in random and clustered networks. Proc. R. Soc. Lond. B 272: 1407 – 1414.

Kiss, I.Z., Green, D.M. & Kao, R.R. (2006). The effect of network heterogeneity and multiple routes of transmission on final epidemic size. Math. Biosci. 203: 124 – 136.

Kiss, I.Z. & Green, D.M. (2008). Comment on 'Properties of highly clustered networks'. Phys. Rev. E 78: 048101.

Kiss, I.Z., Green, D.M. & Kao, R.R. (2008). The effect of network mixing patterns on epidemic dynamics and the efficacy of disease contact tracing. J. R. Soc. Interface 5: 791 – 799.

May, R.M. & Lloyd, A.L. (2001). Infection dynamics on scale-free networks. Phys. Rev. E 64: 066112.

Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D. & Alon, U. (2002). Network motifs: simple building blocks of complex networks. Science 298: 824 – 827.

Newman, M. E. J., Strogatz, S. H. & Watts, D. J. (2001) Random graphs with arbitrary degree distribution and their applications. Phys. Rev. E 64: 026 118.

Newman, M.E.J. (2002). The spread of epidemic disease on networks. Phys. Rev. E 66: 016128.

Newman, M.E.J. (2003a). Mixing patterns in networks. Phys. Rev. E 67: 026126.

Newman, M.E.J. (2003b). Properties of highly clustered networks. Phys Rev. E 68: 026121.

Read, J.M. & Keeling, M.J. (2003). Disease evolution on networks: the role of contact structure. Proc. R. Soc. Lond. B 270: 699 – 708.

Szendröi, B. & Csányi, G. (2004) Polynomial epidemics and clustering in contact networks. Proc. R. Soc. Lond. B (Suppl.) 271, S364 ?- S366.

Shirley, M.D.F. & Rushton, S.P. (2005). The impacts of network topology on disease spread. Ecological Complexity 2: 287 – 299.

Soffer & S.N., Vázquez, A. (2005). Network clustering coefficient without degree–correlation biases. Phys. Rev. E 71, 057 101.

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

Display items

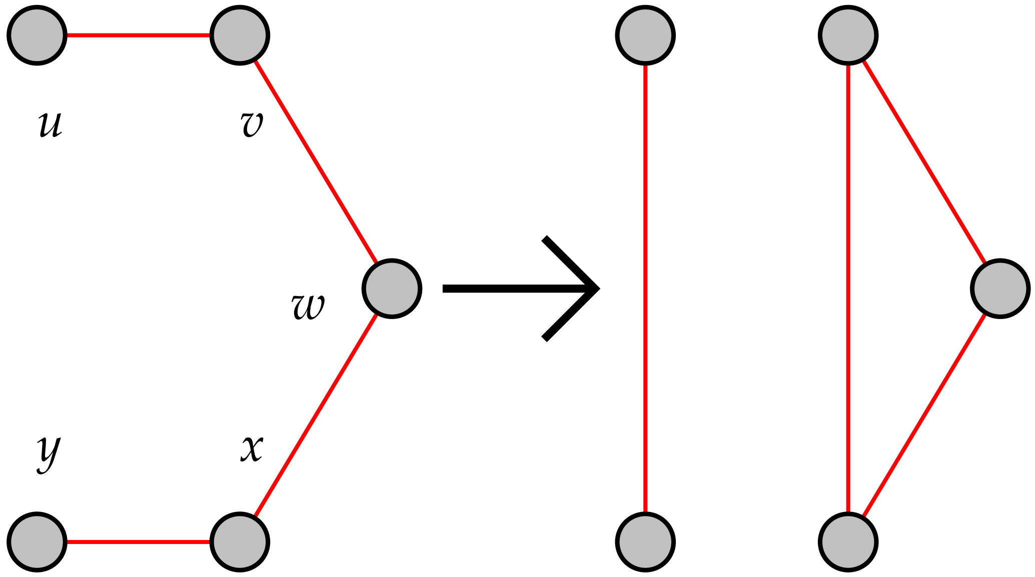

Figure 1: Rewiring algorithm step for generating clustering.

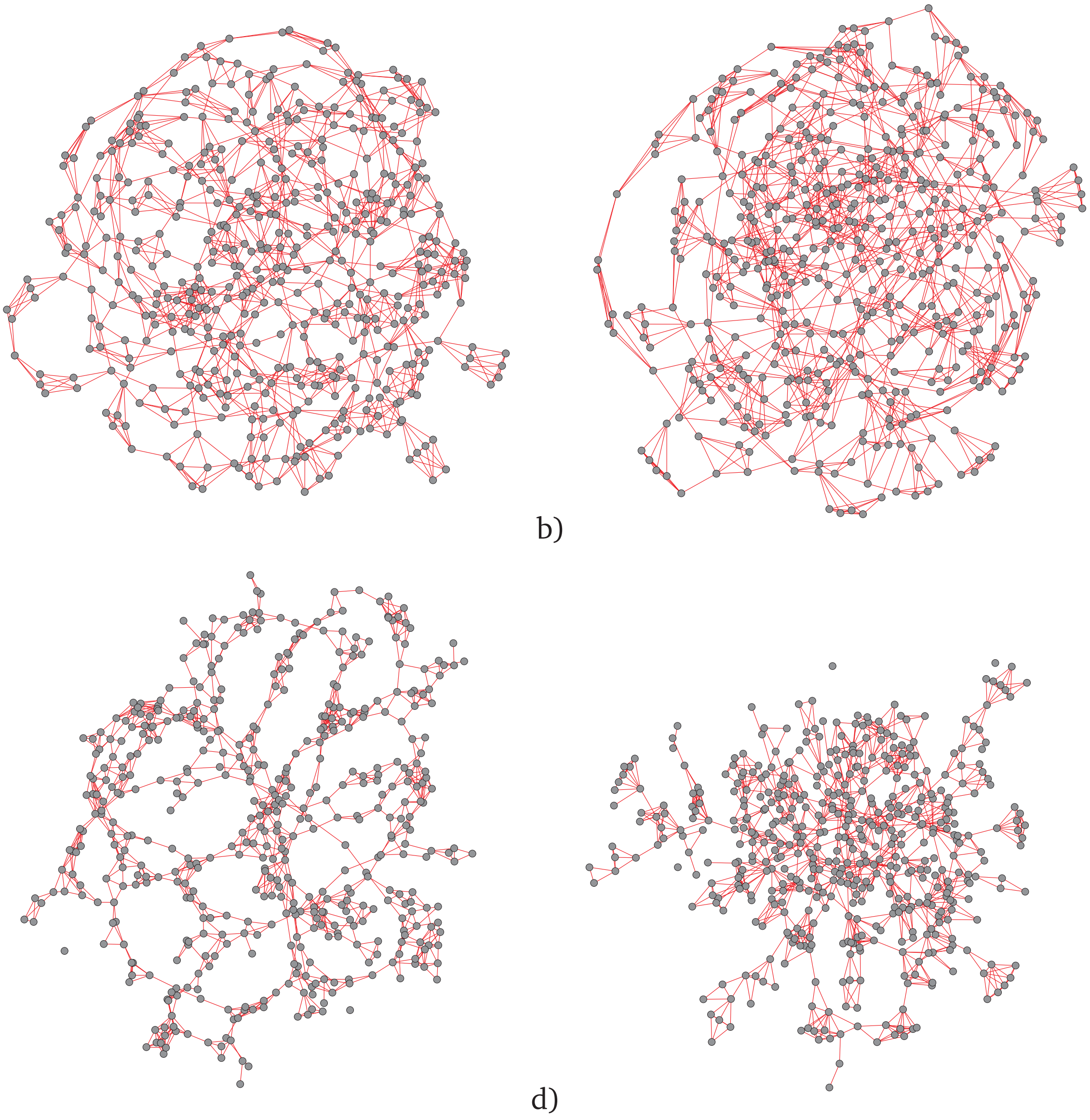

Figure 2: Sample networks with $N = 500$, $\langle k \rangle = 5$ and $\mathcal{C} = 0.6$. a) **iterative** algorithm; b) **rewire to cluster** from constant k; c) **spatial** and d) this network **reclustered**.

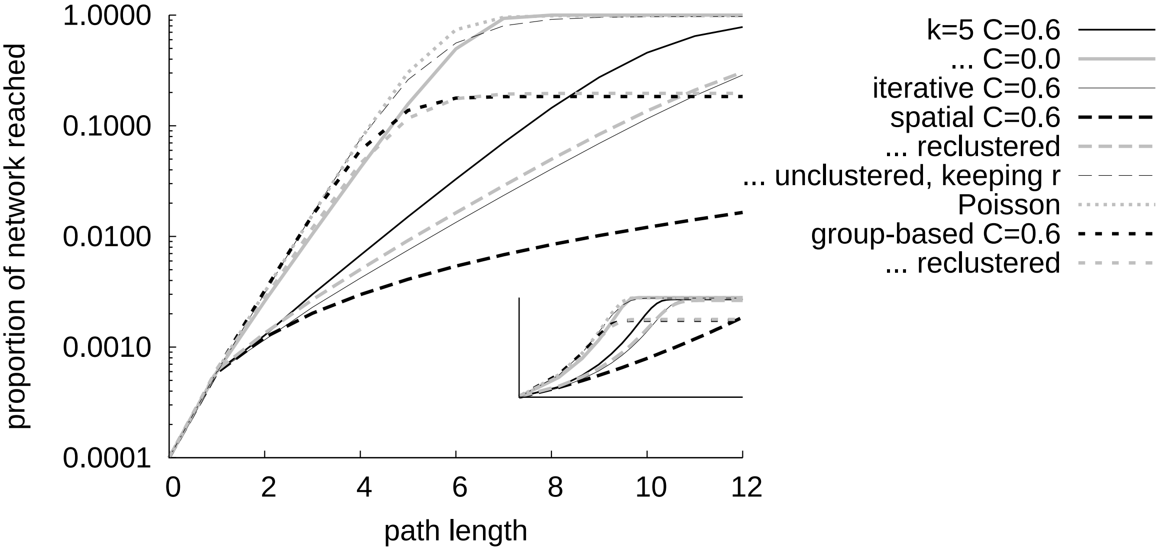

Figure 3: Correlation sum for different clustering algorithms. Inset shows same figure on a log-log scale. All lines are thick unless otherwise stated. Solid black line: fixed degree $\mathcal{C} = 0.6$; solid grey: fixed degree $\mathcal{C} = 0$; thin solid: iterative $\mathcal{C} = 0.6$; black dashed: spatial $\mathcal{C} = 0.6$; grey dashed: spatial reclustered $\mathcal{C} = 0.6$; thin dashed: spatial unclustered preserving mixing; grey dotted: Poisson; black short dash: group-based $\mathcal{C} = 0.6$; grey short dash: group-based reclustered $\mathcal{C} = 0.6$. $\langle k \rangle = 5$ throughout.

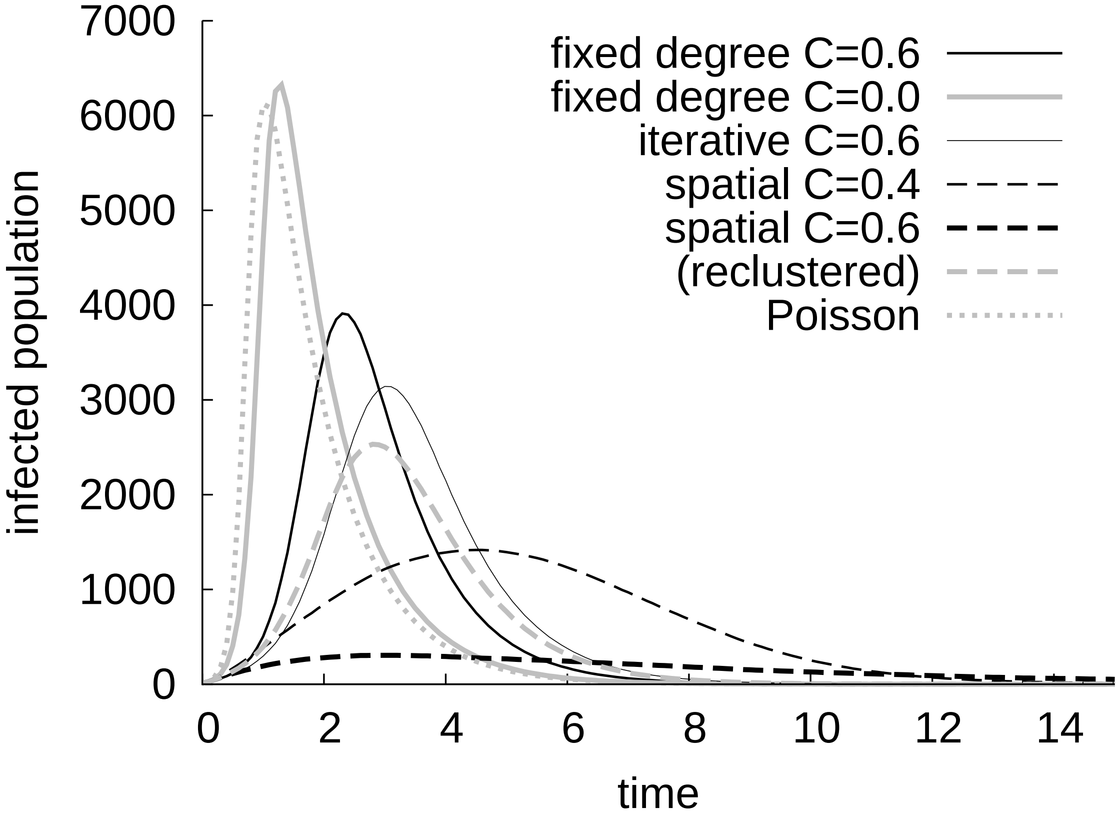

Figure 4: Time series for simulated epidemic. Results are mean prevalence for 10 simulations on each of 25 networks, with $\tau = 2.5$ and $\gamma = 1$. Line styles, mostly as in figure 3, are shown in the legend; throughout, $\langle k \rangle = 5$.

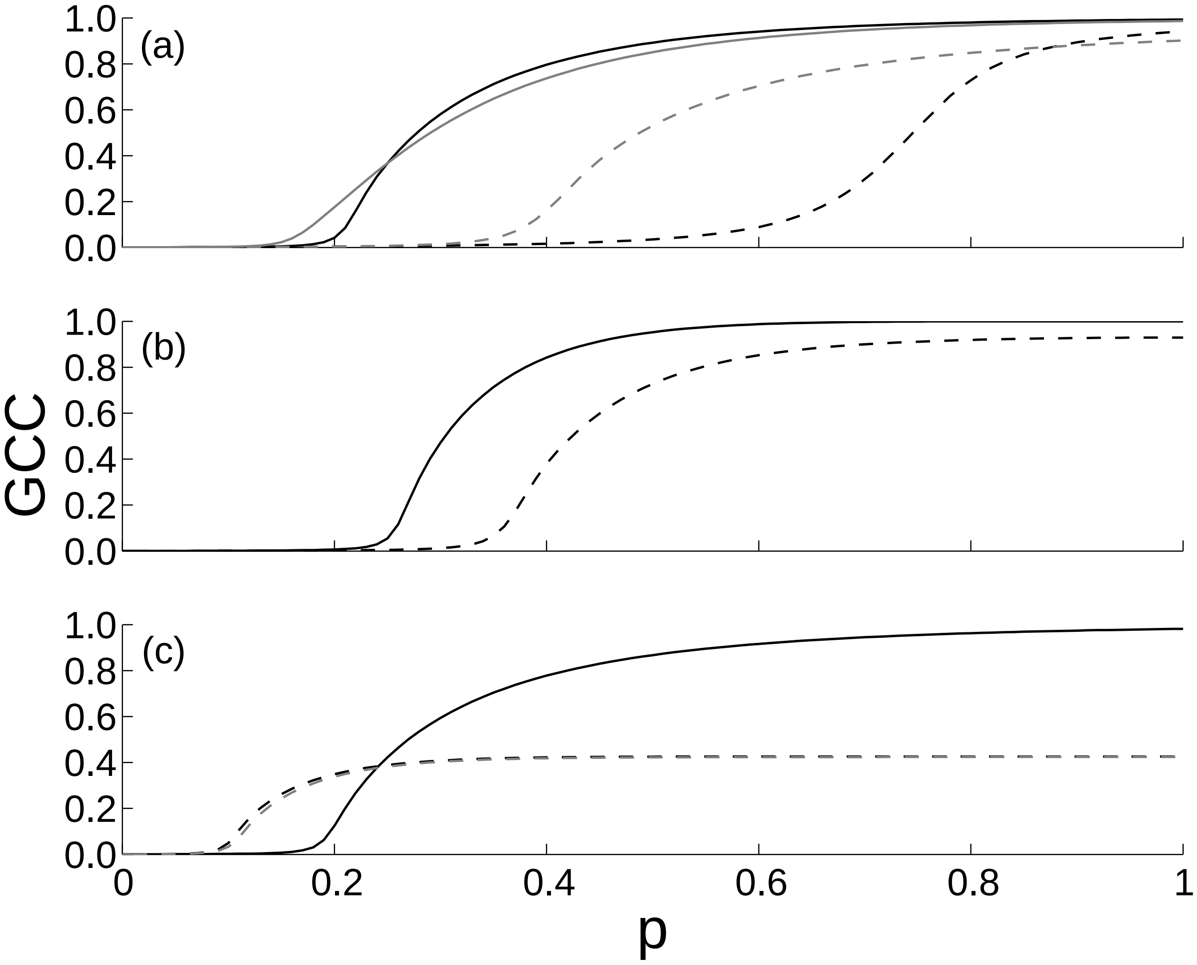

Figure 5: The size of the giant connected component (GCC) for increasing probability p of links being present. (a) spatial $\mathcal{C} = 0$ (black continuous), $\mathcal{C} = 0.6$ (black dashed), reclustered $\mathcal{C} = 0.6$ (grey dashed) and unclustered preserving mixing (grey continuous), (b) fixed degree $\mathcal{C} = 0$ (black continuous) and $\mathcal{C} = 0.6$ (black dashed), and (c) group-based $\mathcal{C} = 0$ (black continuous), $\mathcal{C} = 0.6$ (black dashed) and reclustered $\mathcal{C} = 0.6$ (grey dashed). All simulations based on networks with $\langle k \rangle = 5$.

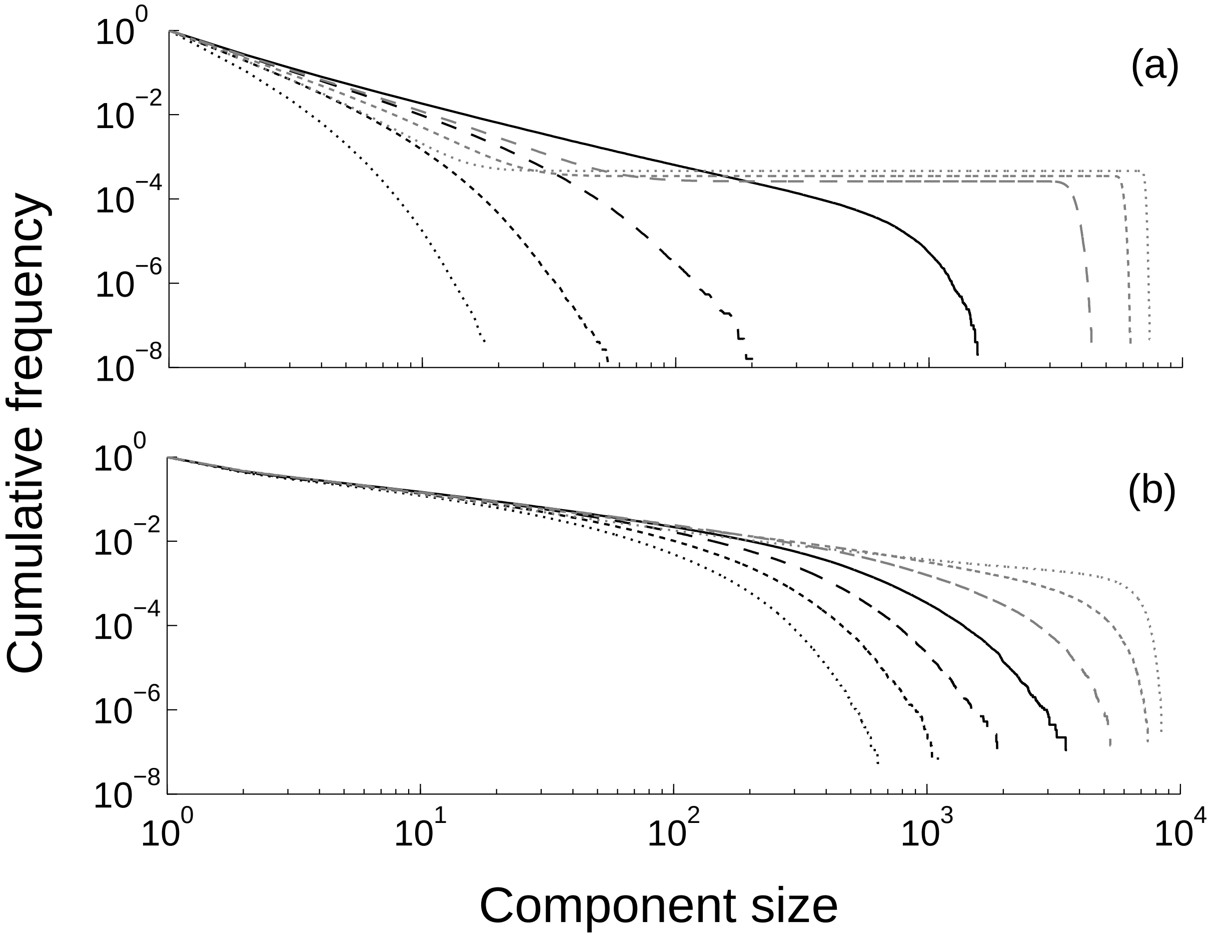

Figure 6: Cumulative distribution of the connected component size for the spatial network model with $c = 0.0$ (a) and $c = 0.6$ (b). Results are based on the outcome of 10000 simulations (100 simulations on 100 different networka). (a) Below percolation for $p = 0.05$, 0.1, 0.15, 0.2 (black: dotted, short dashed, long dashed and solid) and above percolation for $p = 0.25$, 0.3, 0.35 (grey: long dashed, short dashed, dotted). (b) Below percolation for $p = 0.45$, 0.5, 0.55, 0.6 (black: dotted, short dashed, long dashed and solid) and above percolation for $p = 0.65$, 0.7, 0.75 (grey: long dashed, short dashed, dotted). All simulations based on networks with $\langle k \rangle = 5$.

Table 1: Basic statistics of constructed networks for $\langle k \rangle = 5$. Clustering coefficient $\mathcal{C}$, mixing measure r, ratio of degree distribution first two moments $\frac{\langle k^2 \rangle}{\langle k \rangle}$, lead Eigenvalue of adjacency matrix λ, ratio of nodes within two and one step from focal node $\frac{\chi(2)-\chi(1)}{\chi(1)-\chi(0)}$, and the next-generation matrix estimate $\mathcal{R}_0 \sim \frac{I_2}{I_1}$ are shown.

Network	$\mathcal{C}$	r	$\frac{\langle k^2 \rangle}{\langle k \rangle}$	λ	$\frac{\chi(2)-\chi(1)}{\chi(1)-\chi(0)}$	$\frac{I_2}{I_1}$
Fixed degree	0.00	–	5.0	5.0	4.0	5.0
Fixed degree, clustered	0.60	–	5.0	5.0	1.4	5.0
Iterative, clustered	0.61	–	5.0	5.0	1.1	5.0
	0.21	–	5.0	5.0	3.0	5.0
Spatial, clustered	0.58	0.583	6.0	11.0	1.3	7.2
Spatial, reclustered	0.60	0.072	6.0	7.8	1.4	6.1
. . . unclustered preserving mixing	0.00	0.583	6.0	8.9	4.9	7.2
Spatial, no clustering	0.00	0.000	6.0	6.2	5.0	6.0
Group-based, clustered	0.61	0.000	14.0	14.7	5.0	14.0
Group-based, unclustered	0.01	0.000	6.5	6.7	5.4	6.5
Poisson, clustered	0.60	0.072	6.0	7.8	1.5	6.0
	0.40	0.030	6.0	6.9	2.6	6.0
	0.20	0.007	6.0	6.3	3.8	6.0
. . . no clustering	0.00	0.000	6.0	6.2	5.0	6.1

Table 2: Sensitivity of network measures to network properties (see caption to Table 1 for definitions). A tick indicates the indicated measure is sensitive to differences in the indicated network property.

Property	$\langle k \rangle$	$\frac{\langle k^2 \rangle}{\langle k \rangle}$	λ	$\frac{\chi(2)-\chi(1)}{\chi(1)-\chi(0)}$	$\frac{I_2}{I_1}$	simulation
Degree	✓	✓	✓	✓	✓	✓
Degree heterogeneity		✓	✓	✓	✓	✓
Clustering				✓		✓
Overlap of generations						✓
Non-random mixing			✓	✓	✓	✓
Community structure			✓			✓