Higher-order structure and epidemic dynamics in clustered networks

CrossMark

Martin Ritchie [a], Luc Berthouze [b,c], Thomas House [d], Istvan Z. Kiss [a,*]

[a] School of Mathematical and Physical Sciences, Department of Mathematics, University of Sussex, Falmer, Brighton BN1 9QH, UK
[b] Centre for Computational Neuroscience and Robotics, University of Sussex, Falmer, Brighton BN1 9QH, UK
[c] Institute of Child Health, University College London, London WC1E 6BT, UK
[d] Warwick Mathematics Institute, University of Warwick, Gibbet Hill Road, Coventry CV4 7AL, UK

HIGHLIGHTS

- Networks of equal clustering may show significantly different higher-order structures.
- We present an efficient motif counting algorithm.
- Novel order-four transitive-type metrics permit more accurate network description.
- We conjecture on the correct motif counting cardinality for use in transitive ratios.

ARTICLE INFO

ABSTRACT

Clustering is typically measured by the ratio of triangles to all triples regardless of whether open or closed. Generating clustered networks, and how clustering affects dynamics on networks, is reasonably well understood for certain classes of networks (Volz et al., 2011; Karrer and Newman, 2010), e.g. networks composed of lines and non-overlapping triangles. In this paper we show that it is possible to generate networks which, despite having the same degree distribution and equal clustering, exhibit different higher-order structure, specifically, overlapping triangles and other order-four (a closed network motif composed of four nodes) structures. To distinguish and quantify these additional structural features, we develop a new network metric capable of measuring order-four structure which, when used alongside traditional network metrics, allows us to more accurately describe a network's topology. Three network generation algorithms are considered: a modified configuration model and two rewiring algorithms. By generating homogeneous networks with equal clustering we study and quantify their structural differences, and using SIS (Susceptible-Infected-Susceptible) and SIR (Susceptible-Infected-Recovered) dynamics we investigate computationally how differences in higher-order structure impact on epidemic threshold, final epidemic or prevalence levels and time evolution of epidemics. Our results suggest that characterising and measuring higher-order network structure is needed to advance our understanding of the impact of network topology on dynamics unfolding on the networks.

© 2014 The Authors. Published by Elsevier Ltd. All rights reserved.

1. Introduction

Network modelling is an essential tool in characterising a wide range of phenomena: infectious diseases, brain activity, chemical reactions, social interactions, the internet, etc. Any system that involves interactions of its constituent components may be modelled as a network. The versatility of networks as a modelling paradigm may be further augmented by running dynamical processes on the network such as epidemics or neuronal activity. A network's structure can have

a dramatic effect on the processes that run on the network which is currently parameterised by low-order structure alongside the degree distribution. As we shall see, with epidemiological processes the presence of higher-order structure affects how a disease spreads through a network, and the effect of such structures on neuronal dynamics is known to be significant (Sporns et al., 2005; Honey et al., 2009; Gallos et al., 2012; Lynall et al., 2010; Kaiser and Hilgetag, 2010). In this paper we aim to go beyond open and closed triples and give a more comprehensive description of networks in terms of higher-order structure frequency (specifically order-four structures) and their distribution around nodes. In particular, we will examine existing clustered network generating algorithms with respect to their ability, or otherwise, to control higher-order network structure which

* Corresponding author.
E-mail address: i.z.kiss@sussex.ac.uk (I.Z. Kiss).

sometimes may be regarded as a by-product of generating low-order structure that can preclude a correct interpretation of the impact of clustering. The paper is structured as follows. We first introduce and describe a set of clustered network generating algorithms. We follow with a presentation of the network metrics (including a description of the motif identifying/counting algorithm) that we propose to quantify similarities and differences between the generated networks. We then analyse and discuss the impact of higher-order structural differences, at identical degree distribution and equal clustering, on SIS and SIR epidemics. Finally, we discuss how our motif-counting results and newly proposed measure for higher-order structures could be used to parameterise pairwise-like models with closure at the level of quadruples.

2. Material and methods

2.1. Network construction

A significant part of network research relies on networks with arbitrary degree distributions built using the configuration model. This algorithm generates networks where nodes mix at random and where the probability that two nodes are connected is simply proportional to the product of their degree. Such networks coupled with stochastic node dynamics such as SIS, SIR or neural dynamics are amenable to developing macroscopic low-dimensional ODE models that are in excellent agreement with values obtained from stochastic simulations. By construction, these networks are loop-less in the limit of large network size. Whilst such networks can be considered in many cases as realistic or plausible models of some real-world networks, there are many instances where networks have a high degree of structure that typically involves clusters of well connected nodes. Classic examples come from household models used in epidemiology (Ball and Lyne, 2001), and networks of social interactions in general. Motivated by this, there are a series of theoretical or synthetic network models that can be tuned to display increased levels of clustering (Volz et al., 2011; Karrer and Newman, 2010; Newman, 2009; Read and Keeling, 2003; Eames, 2008; Bansal et al., 2009), where clustering denotes the ratio of closed loops of length three with respect to all possible open triples, irrespective of whether they are closed or not.

The classic algorithms to generate networks with tunable clustering include (a) the spatial algorithm proposed by Read and Keeling (2003), (b) an iterative method proposed by Eames (2008), (c) a configuration model that includes clustering (Karrer and Newman, 2010) and (d) the Big-V rewiring algorithm (Bansal et al., 2009; House and Keeling). In a recent study, Green and Kiss (2010) showed that even under identical degree distributions and equal levels of clustering, networks built based on different algorithms can display a markedly different 'higher-order structure'. Whilst their analysis identified large scale structural differences amongst networks with identical degree distribution and clustering, it did not consider extending the concept of clustering involving three nodes to higher-order structures with four or more nodes. The concept of motifs is not

new (Sporns et al., 2005; Karrer and Newman, 2010; Volz et al., 2011; Keeling, 1999a; House and Keeling, 2011) and understanding network structure through higher-order motifs is going to provide a level of detail which cannot be articulated by open or connected triples alone. Below we provide a brief description of the clustered network construction algorithms used in this paper.

2.1.1. Big-V rewiring

The 'Big-V' is an iterative rewiring algorithm that can introduce clustering into any given network and is commonly used by network scientists (Bansal et al., 2009; House and Keeling; Green and Kiss, 2010). At each iterative step, a chain of 5 distinct nodes (u–v–w–x–y) is selected at random and a clone network is generated where the links (u–v) and (x–y) are broken and the edges (u–y) and (v–x) are created. This leads to a single chain of 5 nodes being broken into a triangle and a disconnected pair, see Fig. 1. Local clustering for each node in the chain, as well as all of its neighbours, is computed in both the original and cloned networks and the new configuration is kept only if the level of clustering has increased.

2.1.2. Motif decomposition rewiring

MD (Motif Decomposition) is an iterative rewiring algorithm that starts with a collection of complete sub-networks that are disconnected from one another and rewires edges randomly to reduce the clustering from its maximal value of 1 to the desired level. The following steps are performed:

  i. Initialise a network that is composed of m complete motifs each with n members so that $N=nm$ and $\langle k \rangle = n-1$.
 ii. Categorize every edge as 'local'.
iii. For the first step only, select at random two local edges, cut them, and swap the stubs to form new edges. Mark the pair of new edges as global.
 iv. Select a local and a global edge, cut them, and swap the stubs to form new edges. Mark the pair of new edges as global.
  v. Check the global clustering, if the desired level has not been achieved repeat step (iv).

Fig. 2 illustrates this process being performed on a complete motif with 4 members. It should be noted that this method may work with a heterogeneous degree distribution in which case the network would need to be initialised with motifs of $k+1$ nodes for each different degree k. MD has the significant advantage that it is computationally cheap and that, in the limit of large networks, network properties can be calculated analytically (see Appendix A.1).

2.1.3. CCM (Clustered Configuration Model)

It is possible to modify the configuration model (Miller, 2009; Volz, 2008) so that it constructs networks using specified motifs. Karrer and Newman (2010) and Volz et al. (2011) have shown how to build networks using a configuration model that includes triangle motifs. This idea may be easily extended to allow for larger and more exotic
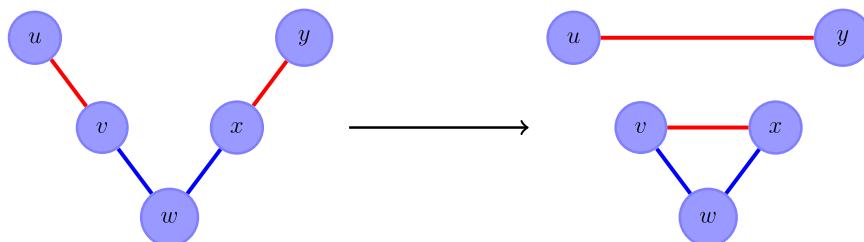
Fig. 1. A single Big-V rewiring. (a) Identify a chain of 5 nodes with 4 edges and (b) if edges (u–v) or (x–y) are already part of a triangle the cuts will not be made, otherwise rewiring is performed, and (c) independent of the outcome of (b) the algorithm will proceed to find a new chain.

motifs to be included in the networks' construction. Rather than just lines, the number of lines and corners of motifs that originate from a node can be varied. In any given motif a node can be considered as a corner and the number of stubs originating from this node that join it to the motif defines its corner type, essential in describing corners of asymmetric structures. To generate a network using this method, the following steps are performed:

1. allocate to a node a number of stubs following a given degree distribution,
2. multinomially determine the configuration of corners and single stubs,
3. create lists for each corner type where a node that is allocated κ corners of a certain type will appear κ times in the corresponding corner list,
4. draw corners at random and without replacement from the appropriate lists and connect with other corners to form motifs,
5. repeat until all lists are empty.

Fig. 3 illustrates corner allocation for an example node. Due to the nature of the configuration model self loops and double loops may be formed. The expected total number of such occurrences is constant depending only on degree; the ratio of self and double

loops to network size becomes negligibly small in the limit of large networks (Newman, 2010).

In this paper homogeneous CCM networks are used with clustering of $\phi = 0.2$ and $\phi = 0.4$. The stub configurations to generate such networks are as follows:

1. $\phi = 0.2$: with probability $p_1 = 0.5$ the quintuple of stubs is maintained as independent links, and with **probability** $p_3 = 1 - p_1$ the quintuple is arranged into one complete square corner and one triangle corner,
2. $\phi = 0.4$: every node is allocated one complete square corner and one triangle corner,
3. $\phi = 0.8$: as this algorithm does not allow overlaps between motifs, this value of clustering cannot be achieved.

Table 1 shows the expected motif allocation per node. The configuration model allows us to analytically determine some of the network structural properties, more specifically the PGF (Probability Generating Function) of the degree/motif distribution.

The CCM algorithm for this work was configured as follows. First, initialise each node with five stubs. Then let p_1 denote the probability that the five stubs form lines, p_2 two triangles and one line, p_3 one complete square and one triangle and p_4 two empty squares and one line. The probabilities are chosen such that $\sum_i p_i = 1$. Let x_i denote the dummy variables of the PGF that corresponds to corner types: x_1 (simple stubs), x_2 (triangles), x_3 (complete squares), x_4 (empty squares). The PGF of the networks degree/corner distribution may now be written as

$$\Psi(x_1, x_2, x_3, x_4) = x_1^5 p_1 + x_1 x_2^2 p_2 + x_2 x_3 p_3 + x_1 x_4^2 p_4, \qquad (1)$$

and the original stub distribution may be recovered by substituting each x_i with x_1^n where n is the corner-stub cardinality:

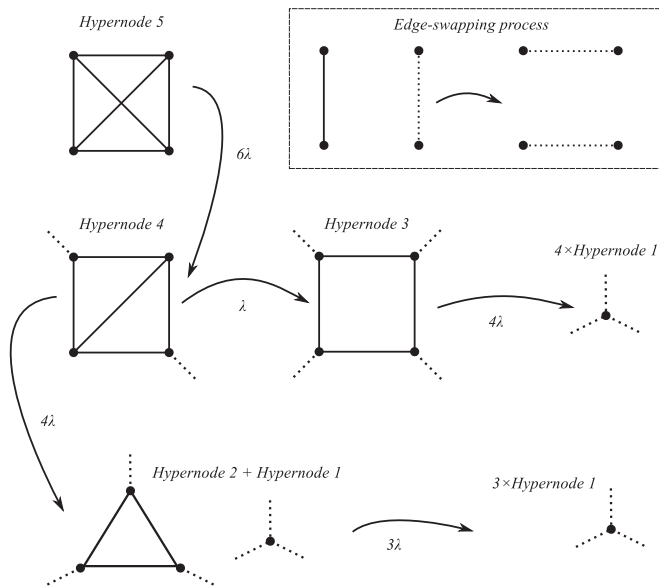$$\psi(x_1) = x_1^5 p_1 + x_1 (x_1^2)^2 p_2 + x_1^2 x_1^3 p_3 + x_1 (x_1^2)^2 p_4 \qquad (2)$$

$$= x_1^5 (p_1 + p_2 + p_3 + p_4) = x_1^5. \qquad (3)$$

$\psi'(1)$ yields the expected degree, and $N\psi''(1)$ yields the number of paths of length two in the network (counted both ways, the path $(A–B–C)$ may also be counted $(C–B–A)$) (Volz et al., 2011). The
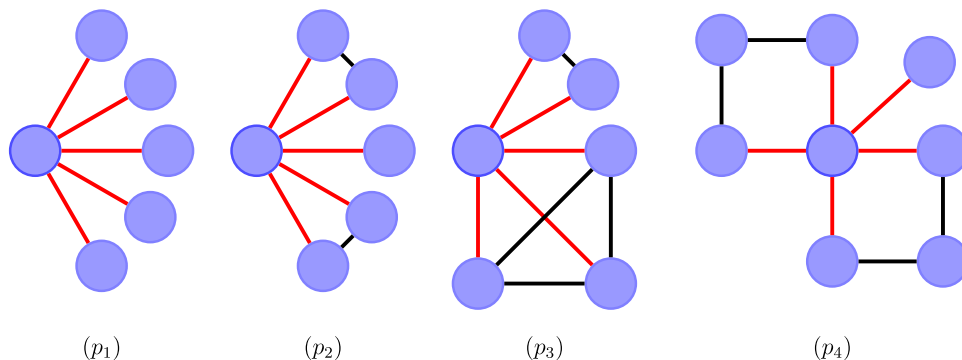

Fig. 2. MD hyper-node configurations. The different hypernode configurations of a homogeneous graph with k=3 as edges are decomposed from local to global.

Table 1
The expected number of lines, triangles and complete squares per node for each level of clustering used.

ϕ	Lines	Triangles	Complete squares
0.2	2.5	0.5	0.5
0.4	0	1	1

(p_1) (p_2) (p_3) (p_4)

Fig. 3. Corner/edge allocation. A node is initially allocated a quintuple of stubs. With probability p_1, p_2, p_3 and p_4 the node will be part of a number of different structures as shown above. In this work homogeneous networks with $\langle k \rangle = 5$ have been used. If a different degree or degree distribution is required then the configuration of motifs will need to be adjusted accordingly.

number of unique triangles in the network can be determined by Ψ

$$[\triangle] = N\left(\frac{\Psi_{x_2}(1,1,1,1)}{3} + \frac{4 \cdot \Psi_{x_3}(1,1,1,1)}{4}\right), \tag{4}$$

since each square is quadruply counted and contains four separate triangles. Clustering is measured as the ratio of three times the number of triangles to all closed and unclosed triples:

$$\phi_{global} = \frac{3N\left(\frac{\Psi_{x_2}(1,1,1,1)}{3} + \Psi_{x_3}(1,1,1,1)\right)}{N\psi''(1)} \tag{5}$$

$$= \frac{\Psi_{x_2}(1,1,1,1) + 3\Psi_{x_3}(1,1,1,1)}{\psi''(1)} \tag{6}$$

$$= \frac{p_2 + 2p_3}{5} \tag{7}$$

For the two types of CCM networks used in this study: $p_1 = 0.5$, $p_3 = 0.5$ yield $\phi = 0.2$ and $p_3 = 1$ yields $\phi = 0.4$ (see Table 2).

2.2. Network metrics: third and higher-order network structure

Here we give a succinct summary of the classic and newly proposed network metrics that will be used to compare and contrast the networks resulting from the different algorithms. Although the novelty of the paper is around order-four structure, we will first consider classic (or third-order) network measures, such as clustering in the global sense as well as distribution of clustering at the node level, nodal betweenness centrality, and connected component analysis via percolation. We then augment the classic network descriptions with an analysis of the distribution of motifs of order higher than closed and open triples both globally and on a per node basis. A network of N individuals is represented with an adjacency matrix, $A \in \{0,1\}^{N^2}$. A pair of individuals (i,j) share a connection if $A_{i,j} = 1$. The networks are undirected, $A = A^T$, and self loops are not allowed $A_{i,i} = 0$, $\forall i \in N$.

1. *Clustering*: clustering may be defined in two ways (Watts and Strogatz, 1998): local (node level) and global (network level). The local clustering of a node n, of degree n_k, is the ratio of connections between neighbours of n and potential connections of neighbours of n. Let $\mathcal{N}$ denote the sub-adjacency matrix of the neighbourhood of n then

$$\phi_{local} = \frac{\sum_{i,j} \mathcal{N}_{i,j}/2}{n_k(n_k - 1)/2}. \tag{8}$$

Global clustering is defined as the ratio of the total number of closed triples to the total number of connected structures with 3 nodes. This may be computed from the adjacency matrix as (Keeling, 1999b)

$$\phi_{global} = \frac{\text{trace}(A^3)}{\|A^2\| - \text{trace}(A^2)}, \tag{9}$$

where $\|A^2\|$ denotes the sum of all elements of A^2. Manipulating the adjacency matrix in this way yields multiplicative counts. An alternative method to obtain the equivalent counts is as follows:

$$[\vee + \triangle] = \sum_{i,j,k,i \neq j \neq k} a_{i,j}a_{j,k}, \tag{10}$$

yielding all connected structures of 3 nodes (closed and unclosed), similarly

$$[\triangle] = \sum_{i,j,k,i \neq j \neq k} a_{i,j}a_{i,k}a_{j,k}, \tag{11}$$

yielding six times the number of unique triangles. A more complete description of this approach is provided in Appendix A.3, along with a conjecture of a possible mapping between unique and multiplicative counts.

2. *Nodal betweenness centrality:* Nodal betweenness centrality measures how often a node appears in the set of shortest paths (which we shall denote s), geodesics, of the network (Freeman, 1977). Nodes with high betweenness centrality will more frequently appear in shortest paths than low ranked nodes. The betweenness centrality of a node n can be computed by

$$B_{bc}(n) = \frac{\sum_{i \neq j \neq n} s_{i,j}(n)}{|s_{i,j}|}, \tag{12}$$

where $s_{i,j}(n)$ denotes the number of shortest paths from i to j that contain node n. The removal of nodes with high betweenness centrality can significantly affect the flow of dynamical processes on the network (Newman, 2010).

3. *Connected component analysis*: CCs (connected components) are sets of nodes where any node may be reached from any other node that is a member of the set. CCs are used to describe the macroscopic structure of a network, as opposed to clustering which describes the local structure of a network. Highly clustered networks contain many components that are weakly connected to, or disconnected from, one another. It has previously been shown (Green and Kiss, 2010) that the GCC (Giant Connected Component – the component that spans almost all of the networks) of highly clustered networks are sensitive to edge removal such that removing even a low proportion of edges can be enough to isolate parts of the network. To perform the analysis, we generate a list of all edges in a network, cycle through each edge in the list and remove it with probability p_r, compute the size and frequency of all components remaining, and plot the cumulative distribution of component size.

4. *Motif frequency and distribution:* Clustering (local or global) essentially measures the occurrence of triangles in a network. It does not distinguish two separate triangles from two triangles that share an edge, neither can it describe loops of order-four or larger. From the perspective of characterising higher-order structure it is a very coarse measurement. In this paper all closed structures of order-four i.e. empty squares, diagonal squares, and complete squares motifs are considered at both network and node levels. It is possible to define new clustering type metrics using structures larger than triangles. Proceeding in a way similar to classic (third-order) clustering and limiting ourselves to 4-node structures connected in a loop, it is possible to define four new structural measurements: the ratio of unclosed quadruples $(1 - \phi_4^1)$, 'empty' squares (ϕ_4^2), squares with a single diagonal (ϕ_4^3), and complete squares (ϕ_4^4) to all connected structures of 4 nodes. We present our results in two formats: (i) global ratios of *unique* order-four structure counts to all unique paths counts, closed and unclosed, (ii) the probability distribution of finding x structures of a certain type associated with a given node. These measurements alongside clustering will provide a higher resolution analysis of network architecture. A brief synopsis of how to compute non-trivial paths of length $l+1$ is as follows (see Appendix A.2 for the full pseudo-code, all path lengths refer to the number of *edges* and $A(\cdot,\cdot)$ is the adjacency matrix):

(a) Consider a path P of length l, and identify a head $H(P)$ (1st node of the path) and a tail $T(P)$ (the last node).

(b) For each neighbour n of $T(P)$, if (i) $A(H(P), n) = 1$, (ii) it has not already been counted as a closed path, and (iii) its reverse has not been counted as a closed path then count a closed path of length $l+1$.

(c) For each neighbour n of $T(P)$, if (i) $A(H(P), n) = 0$, (ii) it has not already been counted as an open path, and (iii) its

reverse has not been counted as an open path then count an open path of length $l+1$.

(d) For all closed paths of length $l+1$ remove circular and reverse circular permutations.

(e) Categorize each closed path by its completeness, i.e., the number (if any) of diagonals in a square.

2.3. Dynamics on networks

To establish the overall impact of higher-order network structure, simulations of various dynamics are performed on the generated networks. First, we use the Markovian SIR (susceptible-infected-recovered) model with a per-contact infection rate τ and recovery rate γ. All simulations are performed using the Gillespie (1976) algorithm. To assess the impact of loops and cycles, we also simulate the SIS epidemic which is more likely to highlight differences in the cycle/motif composition. We shall see that structural differences between networks with the same degree distribution and clustering manifest in epidemiological differences with regard to dynamics on the networks. Previous work (House, 2010) used Kirkwood's super-position approximation to predict the effect of order-four structure on epidemic dynamics. Homogeneous, structured networks composed of N nodes connected in a single GCC were considered and their corresponding system of tuple-wise ODEs was derived. The equations corresponding to motifs of order one, two and three were exact; however, equations that correspond to order-four motifs were closed using order-three terms such that it was only in the limit $\lim_{N\to\infty}$ that equations became exact. For SIS dynamics it was conjectured that the presence of empty square structures reduces the endemic state for all levels of ϕ (varying only ϕ and the prevalence of empty squares whilst keeping all other parameters constant), that complete squares may increase or decrease the endemic state and that diagonal squares had very little effect on the endemic state. In the following we make comparisons between networks that use different distributions of order-four motifs. We expect that networks with markedly different order-four motif distributions produce different epidemiological behaviour.

3. Results

Using the various construction algorithms, we give an overarching analysis of structural differences between networks with the same degree distribution and same levels of classic clustering. All networks used are homogeneous with $\langle k \rangle = 5$, allowing for the formation of structures/loops whilst keeping the complexity to a manageable level. We carry out our analysis on a range of clustering values (i.e. $\phi = 0.2, 0.4, 0.8$) to measure and evaluate the extent to which clustering can emerge from, or determine, different configurations of order-four structures.

1. *Overall feature and structure of the network*: Gephi (Bastian et al., 2009) was used to visualize sample networks generated by the proposed algorithms, see Fig. 4. In these figures nodes are colour coded according to their degree of clustering, with un-clustered nodes coloured with nuances closer to the red end of the spectrum, and more highly clustered nodes coloured with shades closer to the blue end of the spectrum. The figure clearly illustrates that the CCM algorithm gives rise to networks with an extremely homogeneous structure, whilst the rewiring algorithms (i.e. Big-V and MD) construct networks with more heterogeneity in clustering at the node level. It is also evident that this difference translates into a more modular structure for the rewired networks. The CCM

networks stand out as being structurally different from the networks generated by the other algorithms; as well as being homogeneous in degree, they are also homogeneous in structure.
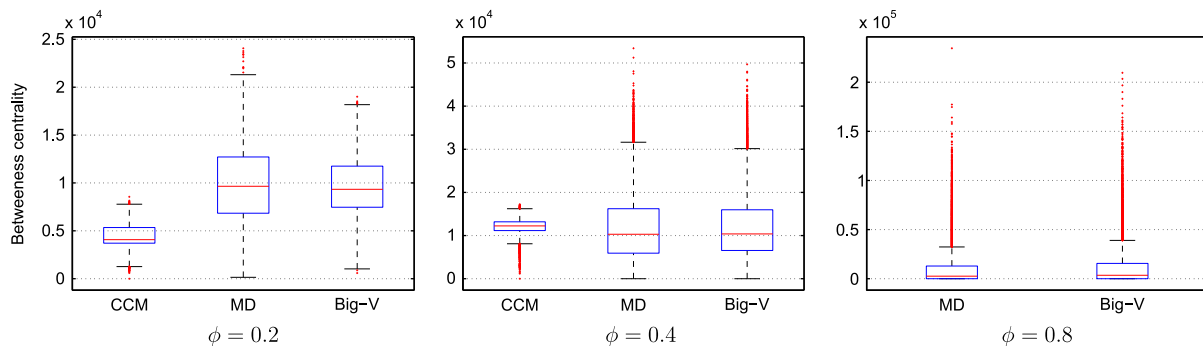
2. *Distribution of clustering and centrality*: The almost homogeneous distribution of the local clustering (see Fig. 5) and betweenness centrality (see Fig. 6) of the CCM networks is expected since by construction every node has the same local structure. For $\phi=0.2$ we know that each node has a quintuple of stubs with probability p_1 or a complete square and a triangle with probability $p_2 = 1 - p_1$. When $\phi=0.4$ every node is a member of one triangle and one complete square.

   The Big-V algorithm introduces clustering in a more heterogeneous manner, with half of the nodes having clustering in the range $0.3 \leq \phi \leq 0.5$. The MD algorithm provides the largest spread of clustering with half of the nodes having clustering in the range $0.2 \leq \phi \leq 0.6$. The box-plot (Fig. 5) of local clustering shows the tendency of the MD algorithm to leave motifs unchanged. These complete motifs must be compensated with other parts of the network being decomposed into a much more random graph-type structure. The MD algorithm relies on random edge swapping to decompose the network into a more random structure. This will tend to result in at least a few leftover fully connected motifs which are only destroyed at very low levels of clustering. Hence, when the frequency of such motifs is still large but the overall, desired, clustering is moderate, the connected parts of the network have to be left weakly clustered.

   The plot of betweenness centrality (Fig. 6) illustrates this description in a more subtle way. Nodes embedded in a motif will have a low betweenness centrality whilst those that act as bridges between structures and the rest of the network will be more highly ranked. We observe that the betweenness centrality plot shows a slightly higher spread for MD networks. The removal of nodes (with a high betweenness centrality rank) is more likely to have a bigger impact on dynamics flowing on the network constructed by the MD algorithm.

3. *Connected component analysis*: For a well connected network with low clustering, low values of p_r leave the macroscopic structure of the networks unchanged (see Fig. 7) i.e. the entire network is contained within a single GCC. Networks with low clustering are resilient to the removal of a relatively small number of edges. For example, the GCC still spans a large proportion of the network when 30% of the edges have been removed. This behaviour has been previously noted (Green and Kiss, 2010).

   The size of the GCC decreases with increasing clustering across all network-generating algorithms (reading column-wise down Fig. 7). This behaviour is expected since in any network where $\langle k \rangle \ll N$, higher clustering is only achievable through many completely isolated components (recall the initialisation of MD networks where $\phi=1$ and the network is composed of m separate motifs). This behaviour is further reflected in the step behaviour of the plots when $\phi=0.8$ (bottom row Fig. 7). Reading row-wise across Fig. 7, for $\phi=0.4$, the GCCs of CCM networks are the most robust to edges being removed. We know the networks to be highly homogeneous such that no particular edge is structurally more important than another. MD networks are extremely sensitive to edges being removed; $p_r=0.3$ has a pronounced impact revealing a strong dependency on components of size 6 or fewer. The MD algorithm tends to preserve some complete motifs, as well as to leave some motifs weakly connected to the GCC at this level of clustering. The Big-V algorithm generates networks with a relatively well connected GCC but still exhibits a mild sensitivity to the removal of edges.
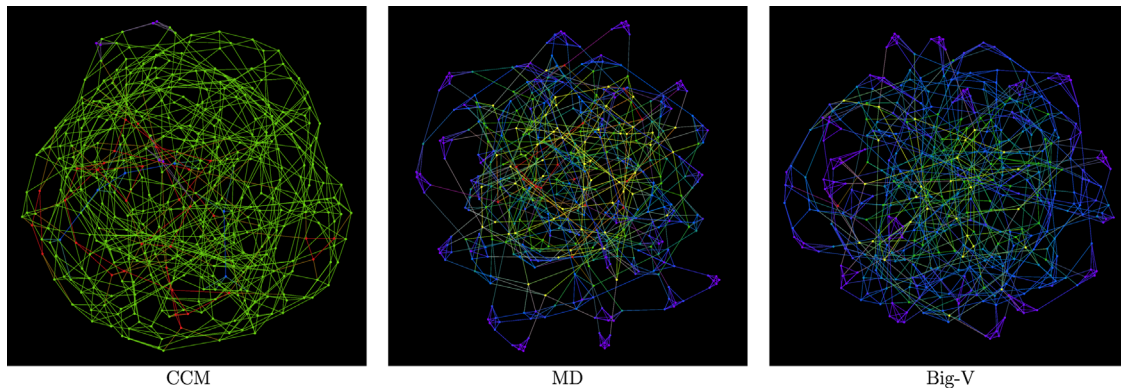
Fig. 4. Example networks. Homogeneous networks with all parameters held equal, $N=400$, $\langle k \rangle = 5$ and $\phi=0.4$. The nodes are coloured so those at the red end of the spectrum have low local clustering, and those at the blue end of the spectrum have high local clustering. (For interpretation of the references to colour in this figure caption, the reader is referred to the web version of this paper.)
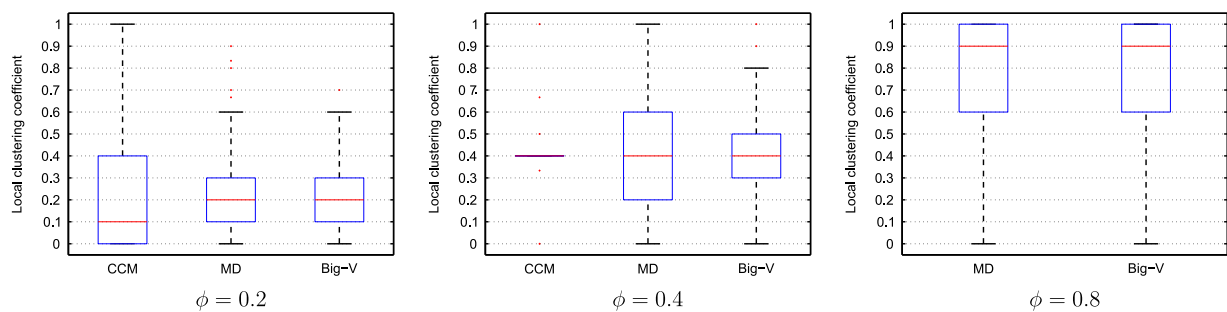
Fig. 5. Distribution of local clustering. Boxplots of local clustering measured from 20 homogeneous networks, $N=1998$, $\langle k \rangle = 5$. The size of the network was chosen to be divisible by 6 due to the MD algorithm starting with disjointed fully connected hexagons. Local clustering is a measure of interconnectivity between neighbours of a given node. CCM shows an extremely tight distribution of clustering for $\phi=0.4$, as we would expect given that each node is allocated the same number and type of structure. MD provides the largest variance in local clustering.



Fig. 6. Distribution of betweenness centrality. Box-plots of nodal betweenness centrality measured from 20 homogeneous networks, $N=1998$, $\langle k \rangle = 5$. Betweenness centrality ranks nodes on how often they appear in paths between other nodes. As clustering is tuned higher, the CCM and Big-V rewiring algorithms isolate fully connected clustered components of $\langle k \rangle + 1$ nodes away from the GCC. At this level of clustering highly connected sets of nodes are still weakly attached to the GCC, yielding the large number of outliers observed in the plot, and hence, a high spread of betweenness centrality values.

4. *Motif statistics for all network types*: Table 2 shows that third-order clustering conveys little information about order-four motifs. The proportion of closed quadruples to all connected structures of 4 nodes increases with clustering, and for high levels of clustering there is a strong dependence on complete squares. However, the algorithms' lack of control of order-four structure is apparent at moderate levels of clustering ($\phi=0.2$, 0.4) where there is no consistent presence of closed order-four structures across networks of equal clustering. The difference in ϕ_4^1 is due to triangles which do not share an edge; indeed these triangles are not measured by this metric. The distribution of triangles is important at higher levels of clustering

where they often share edges or overlap to form order-four structures.

Reading column-wise down Fig. 8, we see a more particular dependence on squares with diagonals as clustering increases. For $\phi=0.4$ there is the greatest heterogeneity in the distribution of diagonal squares. When $\phi=0.8$ we observe that diagonals can only appear in certain combinations about a node, following an almost tri-modal distribution. Again reading column-wise down, for complete squares we see a general trend of increasing complete square prevalence with increased clustering. Nodes may have a count of ten complete squares associated with them when they are members of a complete,
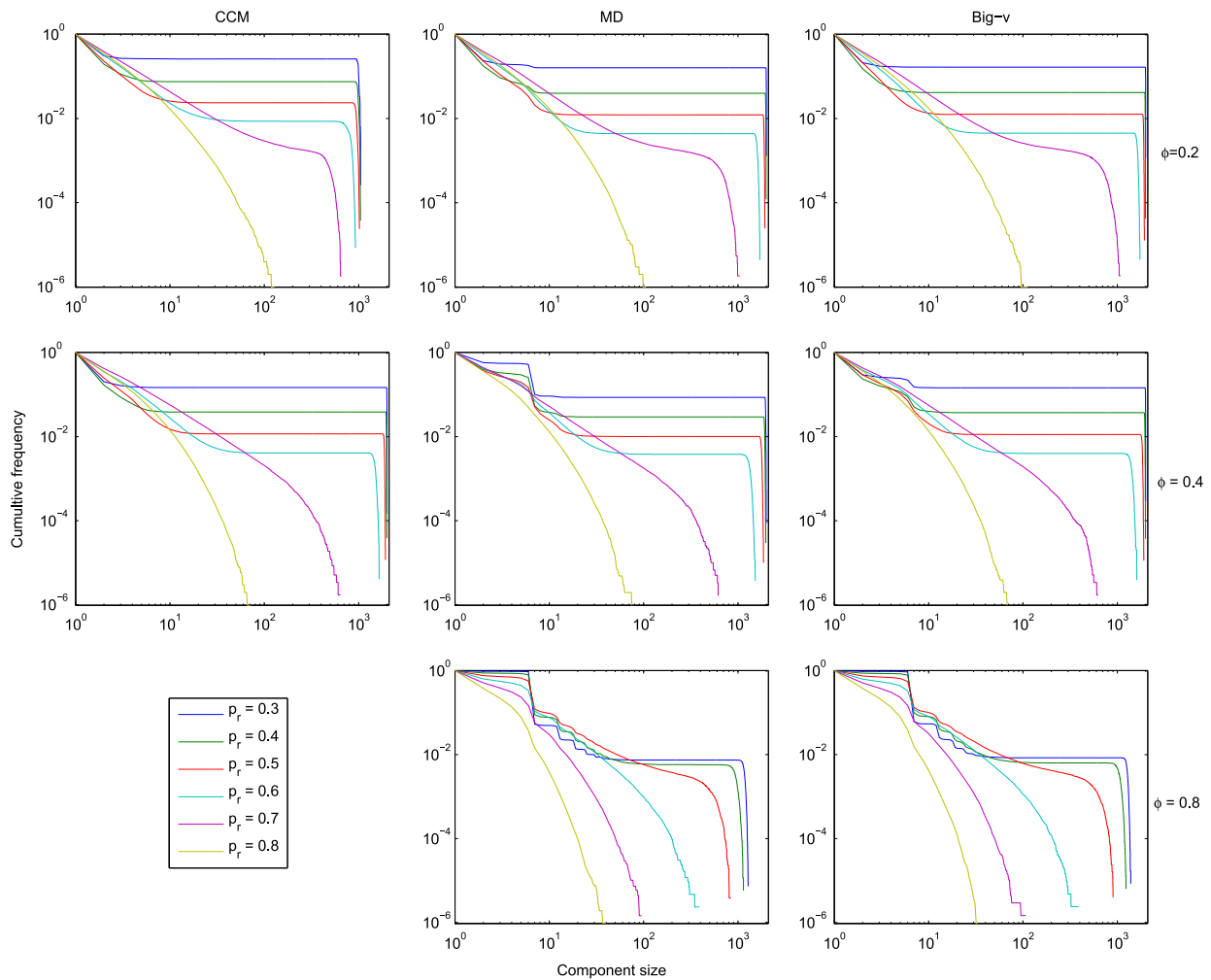
Fig. 7. Edge percolation plots. Frequency of component sizes as edges are removed from the network with probability p_r. Results are taken from homogeneous networks with $\langle k \rangle = 5$ and $N = 1998$. CCM networks are not represented when $\phi = 0.8$. The top, middle and bottom rows represent clustering of $\phi = 0.2$, $\phi = 0.4$ and $\phi = 0.8$ respectively. Each line represents a different value of p_r, varying from 0.3 to 0.8 in increments of 0.1. In each figure, p_r increases in a clockwise direction. The sudden jumps are a result of a strong dependency on a motif of certain size in the description of the network. As $p_r > 0.5$ the networks are decomposed into many disjoint components.

and isolated, six-node structure. At all levels of clustering the probability of finding an empty square associated with a node is rare. This is because, despite being a structure of order-four, squares do not contribute to clustering.

Finally, Fig. 8 also reveals that networks generated by the Big-V algorithm contain empty square motifs with very low frequency. The algorithm searches for unclosed triples contained within strings of five nodes and closes them. Only motifs that may be constructed out of triangles can be expected in Big-V networks in any significant quantity. The MD algorithm also generates few empty square motifs and the CCM algorithm will only include them by specification.

3.1. Dynamics on the networks: evaluation and comparison

To investigate the effect of high order structure we consider the two classic (e.g. SIS and SIR) epidemic dynamics on networks. Starting with the simplest structure, and as expected, for triangles it is observed that when an initially infected individual infects a second, the two infected nodes then compete for the same remaining susceptible. For empty squares and longer loops the effect is similar but less dramatic. Fig. 9 shows that the initial epidemic spread is slower for networks which exhibit loops. By opening a

closed motif whilst preserving degree, two new individuals must be added so the effect of competition is inversely proportional to the motif size. Connectivity within the motif may also negate the effect of competition.

When simulating epidemics on networks with $\phi = 0.2$, the CCM networks show a slower spread of infection (Fig. 10). At this level of clustering the CCM algorithm breaks a quintuple of stubs into all lines with $p = 1/2$, or a complete square corner and a triangle corner with $p = 1/2$. Thus, the CCM networks exhibit areas of high clustering in which the disease will spread more slowly than in areas of low clustering. At $\phi = 0.2$ the CCM networks exhibit a slower spread of infection for both SIS and SIR epidemics. Reading row-wise from left to right it is clear that higher levels of clustering slows the epidemic, see the difference between $\phi = 0.2$ and $\phi = 0.4$, with a less dramatic effect for *SIS* epidemics. Tuning clustering to an even higher level leads to the network breaking down into many disjointed components, such that connectivity within these is excellent. This means that the initial spread could be very fast, but this is quickly curtailed by limited or no connectivity between the highly connected components.

The rewiring algorithms tend to produce networks that contain clustered motifs that are poorly connected to the rest of the networks. Nodes with high betweenness centrality are important in SIR-type processes, which when recovered significantly hinder the propagation of the epidemic. Both of the rewiring algorithms
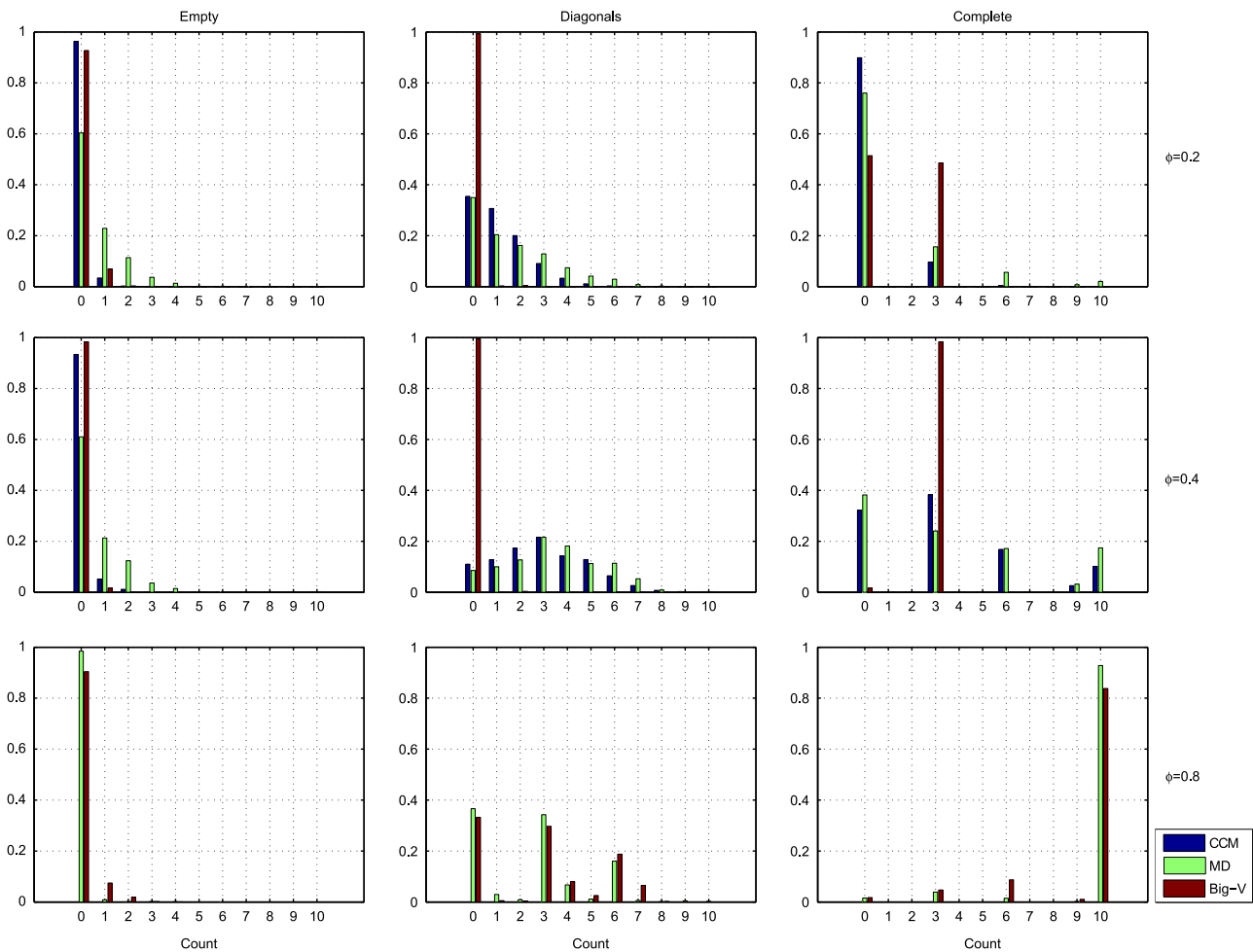
Fig. 8. Order-four motif distribution. The per-node distribution of the number of unique counts of order-four motifs, for all previously used networks. See Appendix A.2 which details how motifs are counted.

Table 2

For each level of clustering the table has been sorted in ascending ϕ then ascending ϕ_4^1, where ϕ_4^1 gives the proportion of all closed quadruples. The above table is computed using unique counts.

Netw. model/ϕ	ϕ_4^1	$\phi_4^2 \square$	$\phi_4^3 \diagup$	$\phi_4^4 \boxtimes$
CCM, 0.2	0.0053	0.0007	0.0001	0.0045
Big-V, 0.2	0.0117	0.0004	0.0104	0.0009
MD, 0.2	0.0239	0.0057	0.0149	0.0034
CCM, 0.4	0.0169	0.0003	0.0002	0.0164
Big-V, 0.4	0.0570	0.0010	0.0400	0.0160
MD, 0.4	0.0731	0.0083	0.0444	0.0204
Big-V, 0.8	0.3150	0.0013	0.0900	0.2237
MD, 0.8	0.3405	0.0044	0.1062	0.2299

produce nodes with high betweenness centrality when compared to the CCM algorithm. It has previously been noted that the MD networks are particularly dependent on isolated motifs; this is reflected in a markedly smaller final epidemic size for both SIS and SIR dynamics at moderate levels of clustering. The CCM and Big-V networks have a more consistent connectivity throughout the network yielding a slightly greater final size (see Fig. 11).

Using such simple dynamics with few states and simple transmission processes, the subtlety in the differences is somewhat expected. Nevertheless we foresee that other more complex dynamics, such as neuronal dynamics or modified voter model, where transitions may depend on membership within certain motifs and transitions do not simply scale linearly with the state of the neighbouring nodes (e.g. saturating activation for interacting neurons), will lead to more marked differences.

4. Discussion

The development of models that capture epidemic or other dynamics on networks is guided, to a great extent, by the structure of the network. Hence, models have initially sought to account for the impact of degree distribution, or heterogeneity in contact. This was closely followed by models capturing preferential mixing, where nodes of similar degrees can be either more likely (assortative mixing) or less likely (disassortative mixing) to be connected. The next stages of model development considered all the above, or at least accounting for their effect, but looked at clustering, the propensity that neighbours of a node are likely to be also neighbours of each other. In this area, progress is still being made and there are many new developments to follow.

Many network models operate on and use synthetic networks designed to be able to control and tune properties such as degree distribution, mixing, clustering and so forth. However, as shown in this paper, controlling certain lower order (node, contact, neighbourhood) properties can and will have an effect on higher-order structure and this can be significant and cannot be disregarded. For example, at high values of clustering generated based on the spatial algorithm (Read and Keeling, 2003), the networks
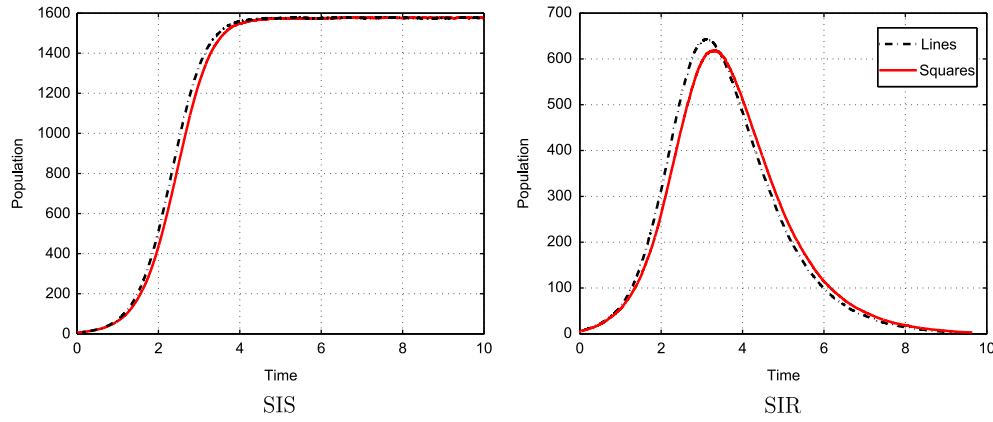
Fig. 9. Random vs square comparison. 20 homogeneous networks were generated with $N=1998$, $\langle k \rangle = 5$ and $\phi = 0.0018$. The plots correspond to averaging Gillespie simulations on each of the networks with parameters $\tau = \gamma = 1$, and 5 initially infectious nodes. The networks marked 'square' were constructed by allowing two squares to be formed out of the 5 stubs allocated to each node, compared against a random network. Plots show comparisons between the prevalence of infection for SIS and SIR dynamics.
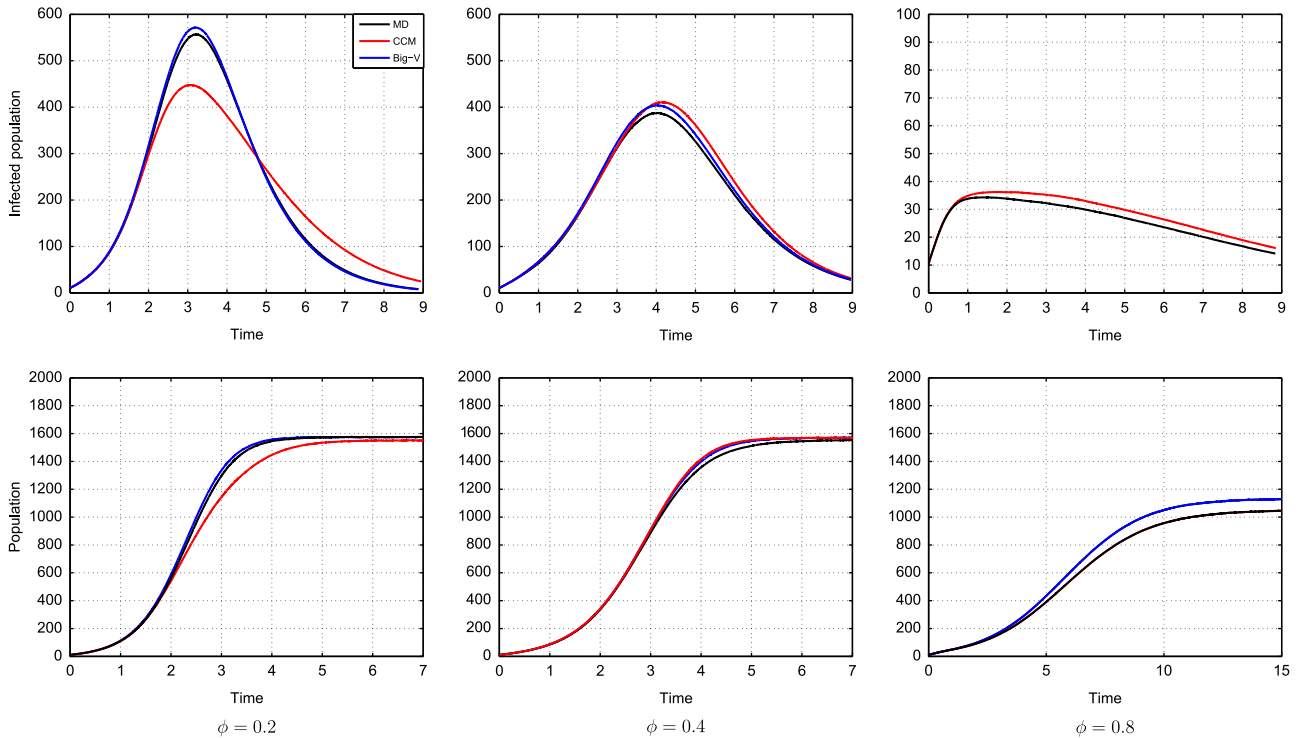

Fig. 10. SIR and SIS dynamics. 20 homogeneous networks were generated with $N=1998$, and $\langle k \rangle = 5$, and the results show the average of 100 Gillespie epidemics on each network realisation. The epidemics were run with parameters $\tau = \gamma = 1$, and were seeded with 5 infectious nodes. The top and bottom rows show the prevalence levels for a SIR and SIS epidemics, respectively..

become more assortatively mixed. Such effects are to be expected since the network is a coherent structure which reacts to each perturbation, such as rewiring or other means of tuning properties. In general, we expect that Big-V rewiring will be the most random way to introduce clustering without model-specific artefacts. However, this comes at what is occasionally prohibitive computational cost. CCM is computationally cheap, and analytically tractable, but is a long way from randomly introduced clustering. In this context, MD can be viewed as a computationally cheap and (given sufficient effort) analytically tractable alternative to Big-V that produces very similar network phenomenology.

In this study, we highlighted that synthetic algorithms that generate networks with tunable clustering do lead to different higher-order structures, such that networks with the same degree

distribution and level of clustering can yield different dynamics on the networks. In order to evaluate differences in higher-order structures we have extended the concept of clustering and proposed some measures to evaluate and quantify the frequency of structures composed of four nodes.

The measures we have proposed are ratios of the uniquely counted, closed motifs of order-four to the unique count of all connected structures of four nodes. This is conceptually convenient but these values may not be suitable for use in low-dimensional ODE approximations, such as the pairwise model. Global clustering is not defined using purely unique counts (see Appendix A.3) and yields a different value when the unique counts are used. In Appendix A.3, we hypothesise the correct counts of motifs and paths for use in clustering-type ratios. Whilst counting uniquely significantly reduces
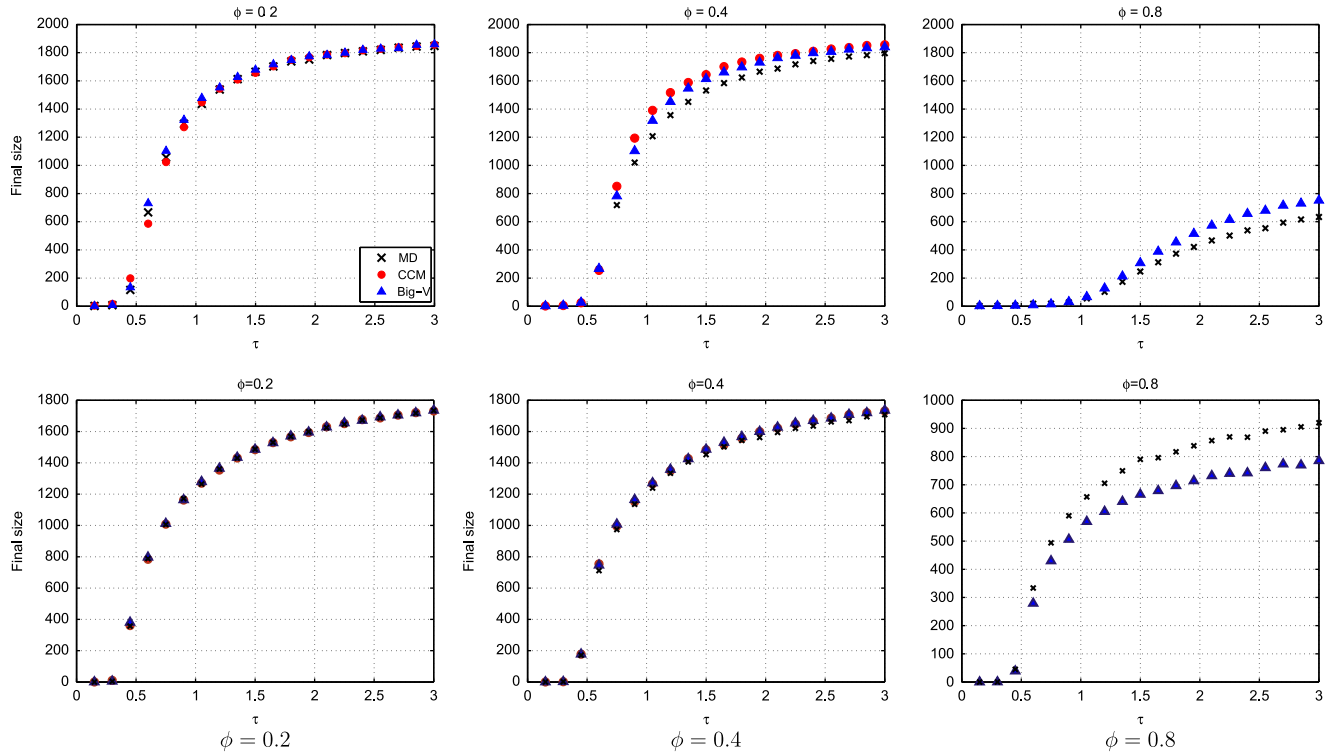
Fig. 11. Plots of final epidemic size (top row) and endemic equilibrium (bottom row) for values of τ increasing from $\tau = 0.1$ to $\tau = 3$ in increments of 0.15. Twenty networks were generated, ten Gillespie simulations were performed for each value of τ. The networks were homogeneous with $\langle k \rangle = 5$ and $N = 1998$.

computational complexity, it has the slight disadvantage that it does not provide the multiplicative type of counting used in pairwise models. In the Appendix, we conjecture that this can be easily overcome by simply multiplying unique counts with the cardinality of the automorphism group corresponding to the motif.

It has been demonstrated that care needs to be taken when trying to extend modelling to clustered networks. Whilst models for simple clustered networks composed of exclusively non-overlapping triangles and edges have been developed, it is going to be more challenging to extend to networks with more complex structures and motifs. Motifs such as a square with a diagonal or a fully connected square may fulfil some function depending on the area of application (e.g. genetic regulatory networks, cortical networks), and thus measuring and quantifying this correctly is crucial for further model development. Many natural extensions for this work exist which include considerations around higher-order structure, algorithm efficiency in measuring these and development of synthetic network models that allow robust and transparent control of not only lower, but also higher-order structures.

Acknowledgements

Appendix A

A.1. Motif decomposition, analysis

It is possible to write down the dynamics for the MD process (Section 2.1.2) in the limit of large networks by decomposing

motifs into hyper nodes (considering each motif at a higher level as a node) and considering the links between them. We now consider the process being performed in a homogeneous network with $\langle k \rangle = 3$ initialised with disjoint, complete square motifs following Fig. 2. By identifying a hyper-node with n edges as Q_n, it is possible to write equations for the normalised count of each hyper-node:

$$\frac{d}{dt} Q_5 = -6\lambda Q_5, \tag{13}$$

$$\frac{d}{dt} Q_4 = 6\lambda Q_5 - 5\lambda Q_4, \tag{14}$$

$$\frac{d}{dt} Q_3 = \lambda Q_4 - 4\lambda Q_3, \tag{15}$$

$$\frac{d}{dt} Q_2 = 4\lambda Q_3 - 3\lambda Q_2, \tag{16}$$

$$\frac{d}{dt} Q_1 = 8\lambda Q_4 + 16\lambda Q_3 + 9\lambda Q_2. \tag{17}$$

These equations can be solved for initial condition $Q(0) = (0, 0, 0, 0, 1/4)$ and evolve towards $Q(\infty) = (1, 0, 0, 0, 0)$. The re-wiring rate λ is just included for clarity and can be set to 1 with no loss of generality. The process stops at a time t^* when the desired level of clustering has been achieved

$$\frac{1}{6} \sum_i T_i Q_i(t^*) = \phi, \tag{18}$$

where T_i denotes the number of triangles associated with each hyper-node type. The above equation may be solved for t^*, and inserted into Eqs. (1)–(5) to obtain a prediction for motif structure. Such hyper-graph counting can be done for any n but quickly becomes too tedious. It is also possible to use the quantities in Table 1 to derive epidemic final sizes and other attributes. Reading

Fig. 2 along side Eqs. (13)–(17) it should be noted that for hyper-node Q_4 there is only one possible way it can decompose to Q_3, namely, through the deletion of its diagonal edge. Considering Eq. (17), any one of the Q_4's four square edges may be deleted resulting in a single Q_2 and Q_1 hyper-node, any one of the Q_3's four edges may be deleted resulting in four Q_1 hyper-nodes and any one of the Q_2's three edges may be deleted resulting in three Q_1 hyper-nodes (Table 3).

A.2. Motif counting algorithm

Below we introduce some notation in order to describe correctly and un-ambiguously the counting algorithm.

Path A path P is an ordered tuple $(i, \dots, j)$
 P_n the nth node of path P
 H^n the head operator such that $H^n(P)$ returns the n first nodes of the path P
 T^n the tail operator such that $T^n(P)$ returns the n last nodes of the path P
 R the reverse operator such that $R((i, \dots, k, \dots, j)) = (j, \dots, k, \dots, i)$
$CP(P)$ the set of circular permutations of path P
 RCP the set of all reverse circular permutations of path P (NB:
 (P) $RCP(P) \neq CP(P)$)
 A the adjacency matrix, $A = A^T$ and with $\mathrm{Tr}(A) = 0$
 $\{\cdot\}$ a set
 PLl the set of non-trivial paths of length l (l = number of edges)

Algorithm 1. Pseudo code for the motif counting algorithm.

/∗ *The following process is applied iteratively to determine non-trivial paths (open paths or closed paths) of length $l+1$ given non-trivial (non-loop) paths of length l. The description below is not specific to a single length but assumes $l \geq 2$.* **Data** *The uniquely counted set of paths of length2 is: $PL_2 = \{(i, j, k)\}$: $A[i, k] \cdot A[k, j] > 0$ and $j > i$.*
initialization;
$LL_{l+1} = \emptyset$; /∗ *Closed paths of length $l+1$*
$PL_{l+1} = \emptyset$; /∗ *Open paths of length $l+1$*
for *All paths P in PL_l* **do**
  for *All nodes n: $A[T^1(P), n] > 0$ & $\notin T^1(P)$* **do**
    $nP = (P, n)$; /∗ *new path*
    if $n = H^1(L)$ & $nP \notin LL_{l+1}$ & $R(nP) \notin LL_{l+1}$ **then**
    $LL_{l+1} \leftarrow nP$;
    end
    if $n \neq H^1(L)$ & $nP \notin LL_{l+1}$ & $R(nP) \notin LL_{l+1}$ **then**
      $PL_{l+1} \leftarrow nP$;
    end
    /∗ *These exclude symmetric paths but not circular permutations.*
  end
  if $P^* \in \{CP(LL_{l+1}), CP(PL_{l+1})\}$ **then**

Table 3
Table of hyper nodes indexed by i with the number of nodes involved n_i, the number of local links l_i, the number of global stubs σ_i, and the number of triangles T_i (counted in both directions from each node). Q_2 denotes triangles that each triangle contains 6 local links (within motif links, counted twice) and 3 global links (links that connect the hyper-node to the rest of the network).

	Q_1	Q_2	Q_3	Q_4	Q_5
n_i	1	3	4	4	4
l_i	0	6	8	10	12
σ_i	3	3	4	2	0
T_i	0	6	0	12	24

$\{LL_{l+1}\} = \{LL_{l+1}\} \backslash P^*$;
      $\{PL_{l+1}\} = \{LL_{l+1}\} \backslash P^*$;
    end
  /∗ *Removes circular permutations.*
end
for *All paths $P \in CLL_{l+1}$* **do**
  if $P^* \in CLL_{l+2}$ & $P^* \notin CRP(CLL_{l+1})$ **then**
    $LL_{l+1} \leftarrow P$;
  end
end
for *All paths $P \in CPL_{l+1}$* **do**
  if $P^* \in CPL_{l+2}$ & $P^* \notin CRP(CPL_{l+1})$ **then**
    $PL_{l+1} \leftarrow P$;
  end
end
/∗ *Removes reverse circular permutations.*

A.3. Motif counting: unique vs multiplicative

In this paper all order-four clustering-type ratios use unique counts. Ratios based on unique counts will give different values to ratios based on multiplicative counts. As an example of multiplicative counting, classic clustering is defined as

$$\phi = \frac{6 \times [\triangle]}{[\triangle + \vee]}, \tag{19}$$

where $[\triangle]$ denotes the number of triangles, and $[\triangle + \vee]$ the number of closed and unclosed length three paths (doubly counted) in the network. If unique counts are used then we have $\phi_{unique} = \phi/3$.

We have computed the unique order-four counts in order to improve the computational performance of our algorithm. However, if we wish to normalise or scale-up the unique counts to correspond to the multiplicative equivalent, correct multiplying factors need to be determined. This appears to be the number of automorphisms associated with each motif type or path length: a triangle has six and a path of length three has two automorphisms.

Let $A = (a_{i,j})$, $i, j \in \{1, \dots N\}$, be the adjacency matrix of an undirected network with no self loops i.e. $A = A^T$ and $A_{i,i} = 0$ for any $i = 1, 2, \dots, N$. It is possible to obtain the multiplicative counts from the adjacency matrix A. Summing over all nodes

$$[-] = \sum_{i,j = 1, i \neq j}^{N} a_{i,j}. \tag{20}$$

This counts twice the number of real or uniquely counted edges in the network. It is possible to count more complex paths as well

$$[\vee + \triangle] = \sum_{i,j,k = 1, i \neq j \neq k}^{N} a_{i,j} a_{j,k}, \tag{21}$$

yielding all connected structures of 3 nodes (closed and unclosed), similarly

$$[\triangle] = \sum_{i,j,k = 1, i \neq j \neq k}^{N} a_{i,j} a_{i,k} a_{j,k}, \tag{22}$$

yielding six times the number of unique triangles. It is also possible to count six different closed triples contained within a triangle: for each node we count clockwise and counter-clockwise about the triangle. This by-directional counting is important so that the method is consistent when considering directed networks. Following the same counting methodology it is possible to count order-four structures:

$$[\sqcup + \square + \boxslash + \boxtimes] = \sum_{i,j,k,l = 1 \; i \neq j \neq k \neq l}^{N} a_{ij} a_{jk} a_{kl}. \tag{23}$$

Table 4
Combinations of i, j, k and l that satisfy (27).

i	j	k	l
1	2	3	4
1	3	2	4
4	2	3	1
4	3	2	1

In this form we see that it is possible to compute the individual counts using the following identities:

$$[\boxtimes] = \sum_{i,j,k,l = 1, i \neq j \neq k \neq l}^{N} a_{ij} a_{ik} a_{il} a_{jk} a_{jl} a_{kl}, \qquad (24)$$

$$[\boxslash] = \sum_{i,j,k,l = 1, i \neq j \neq k \neq l}^{N} a_{ij} a_{ik} (1 - a_{il}) a_{jk} a_{jl} a_{kl}, \qquad (25)$$

$$[\square] = \sum_{i,j,k,l = 1, i \neq j \neq k \neq l}^{N} a_{ij} a_{ik} (1 - a_{il})(1 - a_{jk}) a_{jl} a_{kl}. \qquad (26)$$

Counting this way a single $[\boxtimes]$ is counted 24 times, $[\boxslash]$ is counted 4 times and $[\square]$ is counted 8 times, equal to the number of automorphisms associated with each motif type. By listing the different combinations of $\{i, j, k, l\}$ that satisfy (for diagonal squares)

$$a_{ij} a_{ik} (1 - a_{il}) a_{jk} a_{jl} a_{kl} = 1, \qquad (27)$$

such that $i \neq j \neq k \neq l$ it is possible to gain insight into the cardinality of this count. Consider a diagonal square as orientated in the notation $\boxslash$ labelled starting at the top left node in a clock wise direction: i, j, k and l. By listing the combinations in this way (Table 4) all automorphisms associated with a diagonal square – the identity, the orbit between j and l, the orbit between i and k and finally the permutation that includes both orbits – have been listed.

Currently, based on our intuition and numerical tests, we conjecture that this is the correct way to scale-up from unique to multiplicative motif counts. This method of counting is thorough but not practical for networks of reasonable size since it has complexity $\mathcal{O}(N^n)$ for order-n structures.

References

Ball, F., Lyne, O.D., 2001. Stochastic multi-type SIR epidemics among a population partitioned into households. Adv. Appl. Probab. 33 (1), 99–123.

Bansal, S., Khandelwal, S., Meyers, L., 2009. Exploring biological network structure with clustered random networks. BMC Bioinf. 10 (1), 405.

Bastian, M., Heymann, S., Jacomy, M., 2009. Gephi: An Open Source Software for Exploring and Manipulating Networks.

Eames, K.T., 2008. Modelling disease spread through random and regular contacts in clustered populations. Theoretical population biology 73 (1), 104–111.

Freeman, L.C., 1977. A set of measures of centrality based on betweenness. Sociometry, 35–41

Gallos, L.K., Makse, H.A., Sigman, M., 2012. A small world of weak ties provides optimal global integration of self-similar modules in functional brain networks. Proc. Natl. Acad. Sci. 109 (8), 2825–2830.

Gillespie, D.T., 1976. A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. J. Comput. Phys. 22 (4), 403–434.

Green, D.M., Kiss, I.Z., 2010. Large-scale properties of clustered networks: Implications for disease dynamics. Journal of Biological Dynamics 4 (5), 431–445.

Honey, C., Sporns, O., Cammoun, L., Gigandet, X., Thiran, J.-P., Meuli, R., Hagmann, P., 2009. Predicting human resting-state functional connectivity from structural connectivity. Proc. Natl. Acad. Sci. 106 (6), 2035–2040.

House, T.A., 2010. Generalised network clustering and its dynamical implications. Adv. Complex Syst. 13 (3), 281–291.

House, T., Keeling, M.J., The impact of contact tracing in clustered populations. PLoS Comput. Biol. 6(3), p. e1000721.

House, T., Keeling, M.J., 2011. Insights from unifying modern approximations to infections on networks. Journal of the Royal Society Interface 8 (54), 67–73.

Kaiser, M., Hilgetag, C.C., 2010. Optimal hierarchical modular topologies for producing limited sustained activation of neural networks. Front. Neuroinf. 4, 2010.

Karrer, B., Newman, M.E.J., 2010. Random graphs containing arbitrary distributions of subgraphs. Phys. Rev. E 82, 066118.

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

Keeling, M.J., 1999b. The effects of local spatial structure on epidemiological invasions. Proc. R. Soc. Lond. Ser. B: Biol. Sci. 266 (1421), 859–867.

Lynall, M.E., Bassett, D.S., Kerwin, R., McKenna, P.J., Kitzbichler, M., Muller, U., Bullmore, E., 2010. Functional connectivity and brain networks in schizophrenia. J. Neurosci. 30 (28), 9477–9487.

Miller, J.C., 2009. Spread of infectious disease through clustered populations. J. R. Soc. Interface

Newman, M.E., 2009. Random graphs with clustering. Phys. Rev. Lett. 103 (5), 058701.

Newman, M., 2010. Networks: an introduction. Oxford University Press, New York.

Read, J.M., Keeling, M.J., 2003. Disease evolution on networks: the role of contact structure. Proceedings of the Royal Society of London. Series B: Biological Sciences 270 (1516), 699–708.

Sporns, O., Tononi, G., Kötter, R., 2005. The human connectome: a structural description of the human brain. PLoS Comput. Biol. 1 (4), e42.

Volz, E., 2008. SIR dynamics in random networks with heterogeneous connectivity. J. Math. Biol. 56 (3), 293–310.

Volz, E.M., Miller, J.C., Galvani, A., Ancel Meyers, L., 2011. Effects of heterogeneous and clustered contact patterns on infectious disease dynamics. PLoS Comput. Biol. 7 (6), e1002042.

Watts, D.J., Strogatz, S.H., 1998. Collective dynamics of 'small-world' networks. nature 393 (6684), 440–442.