Degree correlations in graphs with clique clustering

Peter Mann,* V. Anne Smith, John B.O. Mitchell, and Simon Dobson
School of Computer Science, University of St Andrews, St Andrews, Fife KY16 9SX, United Kingdom
School of Chemistry, University of St Andrews, St Andrews, Fife KY16 9ST, United Kingdom and
School of Biology, University of St Andrews, St Andrews, Fife KY16 9TH, United Kingdom
(Dated: March 29, 2022)

Correlations among the degrees of vertices in random graphs often occur when clustering is present. In this paper we define a joint-degree correlation function for vertices in the giant component of clustered configuration model networks which are comprised of clique subgraphs. We use this model to investigate, in detail, the organisation among nearest-neighbour subgraphs for random graphs as a function of subgraph topology as well as clustering. We find an expression for the average joint degree of a neighbour in the giant component at the critical point for these networks. Finally, we introduce a novel edge-disjoint clique decomposition algorithm and investigate the correlations between the subgraphs of empirical networks.

I. INTRODUCTION

A network is a collection of vertices and edges [1]. The nature of the local connectivity among the vertices of a graph has a profound influence on the structural characteristics of the entire network. Common structural properties include: the clustering [2], which is the tendency for triples of vertices to be organised into triangles; subgraph composition [3], which considers the organisation of the edges into recognized motifs; nearest-neighbour degree correlation (NNDC) [4], which is the tendency for similar degree vertices to connect to one another or not; long-range degree correlations (LRDC) [5], which are nonlocal degree correlations beyond the nearest-neighbourhood; the component structure [6], the core-periphery structure, path lengths, communities, fractality and various scale phenomena. In turn, the structural characteristics determine the stability and the governing dynamics of processes occurring over the graph as well as its response to random or targeted attack. Understanding the connective microstructure of complex systems is therefore of crucial importance to a wide range of disciplines including biology, social science and physics as well as to a broad range of applications including network formation, modelling the properties of empirical networks and the observed response to processes such as epidemic spreading, synchronization, percolation or information propagation over networks. It is well known [7–9] that the structural characteristics of the giant component (GCC) of a random uncorrelated graph can be vastly different from the properties of the whole network. In particular, the GCC exhibits a negative NNDC unless the network is singly connected.

The configuration model is a method that allows the construction of uncorrelated random graphs with a prescribed distribution of degrees. Recent work has drawn attention to the generalised configuration model (GCM) which allows the construction of networks that are composed of independent subgraphs. The central object of the GCM is a joint degree distribution that describes the number of roles that a vertex plays within each subgraph on average [10–12]. The generating function formulation is an analytical technique that can be used to describe the expectation values for the properties of the ensemble of graphs that can be constructed using the GCM from a given joint degree sequence.

The GCM incorporates networks with higher-order clustering, typical of the mixing patterns in many human contact networks, as well as multilayer, modular and multiplex systems. In such empirical networks, clustering that follows a heavy tail degree distribution leads to highly clustered networks whereby the vertices can be members of several triangles among the nearest-neighbour contacts. In such cases, it is common that the triangles share one or more edges and thus, higher-order subgraphs, such as cliques, are more accurate representations of the local environment of the vertices. Organisation among cliques of different sizes plays a significant and non-trivial role in spreading processes, particularly of epidemics, over the network. Since many diseases spread through vertex-vertex interactions, effective control of an epidemic must take advantage of the understanding of the local environment of high-degree vertices in tight-knit cliques.

Clustering in complex networks has been studied previously using generating functions [10–24]. Newman found that the presence of clustering in Poisson networks led to a reduction in the critical mean degree required for the formation of a GCC as well as its size. Miller showed that this effect is due to the assortative correlations within the Poisson model and that for networks with the same degree correlations, clustering increases the critical point. Hasegawa and Mizutaka [22] considered the NNDC among the GCC of clustered networks comprised of ordinary edges and triangles. It was found that the GCC can be assortative or dissasortative depending on the details of the clustering; however, dissasortative correlations reappeared upon a characteristic renormalisation of the triangles into single *supervertices*. Thus, the GCC of random uncorrelated networks displays dissasortative NNDC by nature.

In this paper, we address how two vertices of given joint degrees are expected to connect to one another. More formally, we study NNDC in the GCC of random clus-

tered graphs that have been constructed according to the GCM prescription to include higher-order subgraphs. We examine the tendency for organisation among the subgraphs and investigate whether vertices with high subgraph degree connect preferentially to other high subgraph degree vertices or not. We then examine the properties of empirical networks by introducing a novel clique cover and compare our cover to other recent advances in the literature [25].

II. BACKGROUND

In this section, we review the generating function formulation for higher-order subgraphs [10–12, 14, 26] and the method of construction of GCM networks. We reserve bold characters for vector quantities.

The degree distribution p_k is the probability that a randomly chosen vertex in the network has degree k. A common assumption is that the edges are locally tree-like; short range cycles and connections among the nearest neighbours are prohibited. The tree-like assumption has proven very successful at describing many network properties [27]; however, the properties of random clustered networks require a generalisation to the degree of a vertex, beyond simple tree edges, to incorporate the effects of triangles and other higher-order motifs. The resulting model was developed independently by Newman [26] and Miller [14] for networks with triangles and later extended to all network motifs by Karrer and Newman [10]. The models assume that overall degree of a vertex can be partitioned into sub-degrees that correspond to the involvement of a vertex in pre-defined subgraphs. For instance, the generalised degree, $\boldsymbol{k_\tau} = (k_\perp, k_\triangle, k_\square, \dots)$, of a vertex that has six tree-like edges and is also a member of one triangle, two squares and three pentagons would be $\boldsymbol{k_\tau} = (6, 1, 2, 3)$. The probability that a randomly chosen vertex has a particular generalised degree is given by a joint degree distribution $p_{\boldsymbol{k_\tau}}$. The ordinary degree distribution is recovered from

$$p_k = \sum_{k_\perp=0}^{\infty} \cdots \sum_{k_\gamma=0}^{\infty} p_{k_\perp,\dots,k_\gamma} \delta_{k,\sum \lambda_\tau k_\tau \in \boldsymbol{\tau}} \tag{1}$$

where $\boldsymbol{\tau}$ is a vector of subgraph topologies $\{\perp, \triangle, \square, \pentagon, \cdots, \gamma\}$, up to some terminating motif topology represented by γ, k_τ is the degree of shape $\tau \in \boldsymbol{\tau}$, λ_τ is the number of edges a vertex has in shape τ, $p_{\boldsymbol{k_\tau}} = p_{k_\perp,\dots,k_\gamma}$ is the $\dim(\boldsymbol{\tau})$ joint probability distribution of degrees and $\delta_{i,j}$ is the Kronecker delta. For instance, a vertex that is part of a two tree-like edges, a triangle and a square will have the following joint degree sequence $(k_\perp, k_\triangle, k_\square) = (2, 1, 1)$, while its overall degree is $k = 6$. A network is described by its joint probability distribution of each vertex playing a certain role in a given subgraph a particular number of times [10] for all permissible combinations of joint degrees. The joint degree distribution can be generated using

$$G_0(\boldsymbol{z}) = \sum_{k_\perp=0}^{\infty} \cdots \sum_{k_\gamma=0}^{\infty} p_{k_\perp,\dots,k_\gamma} z_\perp^{k_\perp} \cdots z_\gamma^{k_\gamma} \tag{2}$$

where $\boldsymbol{z} = \{z_\perp, z_\triangle, z_\square, \dots, z_\gamma\}$. In the ordinary generating function model, the excess degree distribution q_k defines the probability that a randomly chosen edge leads to a vertex of degree $k+1$. In the generalised model we must define an excess degree distribution for each topology in $\boldsymbol{\tau}$; since, traversing an edge of a particular topology does not, in general, lead to vertices with equivalent joint degrees. The joint excess degree distribution for an edge of topology τ is

$$q_\tau(\boldsymbol{k_\tau}) = (k_\tau + 1) p_{\boldsymbol{k_\tau}\backslash\{\tau\}, k_\tau+1} / \langle k_\tau \rangle \tag{3}$$

where the notation $\boldsymbol{s}\backslash\{s\}$ excludes element s from set $\boldsymbol{s}$. Each joint excess degree distribution is generated as

$$G_{1,\tau}(\boldsymbol{z}) = \sum_{k_\perp}^{\infty} \cdots \sum_{k_\gamma}^{\infty} q_{\boldsymbol{k_\tau}} \; z_\tau^{k_\tau - 1} \prod_{\nu \neq \tau} z_\nu^{k_\nu} \tag{4}$$

and is also seen to be the partial derivative of Eq. 2 with respect to z_τ divided by the expected number of τ-motifs

$$G_{1,\tau}(\boldsymbol{z}) = \frac{1}{\langle k_\tau \rangle} \frac{\partial G_0}{\partial z_\tau} \tag{5}$$

which can also be written as

$$G_{1,\tau}(\boldsymbol{z}) = \frac{G_0'^{\tau}(\boldsymbol{z})}{G_0'^{\tau}(\boldsymbol{1})} \tag{6}$$

where $G_0'^{\tau}$ is the first derivative of $G_0(\boldsymbol{z})$ with respect to z_τ and $\langle k_\tau \rangle = G_0'^{\tau}(\boldsymbol{1})$ is the average τ-degree for a vertex in the network.

The global clustering coefficient C of a network with V vertices is defined as

$$C = \frac{3\mathcal{N}_\triangle}{\mathcal{N}_3} \tag{7}$$

where $\mathcal{N}_\triangle$ is the number of triangles in the network and $\mathcal{N}_3$ is the number of connected triples. The number of triangles involving vertices with a given joint degree $\boldsymbol{k_\tau}$ is

$$\mathcal{N}_{\triangle,\boldsymbol{k_\tau}} = V p_{k_\perp,\dots,k_\gamma} (k_\triangle + \cdots + \mu_\gamma k_\gamma) \tag{8}$$

where μ_τ is the number of triangles that a vertex belongs to as a member of a τ-motif. For instance, $\mu_\triangle = 1$ while a vertex in 4-clique has belongs to 3 triangles. The total number of triangles in the network is found by summing over the joint degree

$$\mathcal{N}_\triangle = \sum_{k_\perp=0}^{\infty} \cdots \sum_{k_\gamma=0}^{\infty} \mathcal{N}_{\triangle,\boldsymbol{k_\tau}} \tag{9}$$

The number of connected triples is given by [26]

$$\mathcal{N}_3 = V \sum_k \binom{k}{2} p_k \tag{10}$$

We can use the generating function formulation to determine the probability that a vertex selected at random belongs to the GCC. Let u_τ be the probability that a vertex reached by the traversal of an edge of topology τ does not lead to the GCC. Similarly, the probability that the entire subgraph does not connect the vertex to the GCC is $u_\tau^{m_\tau}$ where m_τ is the number of edges a vertex has in each independent subgraph of topology τ. For instance, a vertex has 3 edges in a given 4-clique. The probability that the neighbour fails to attach to the GCC is given by a self-consistent expression $u_\tau = G_{1,\tau}(\boldsymbol{u}_{\boldsymbol{\tau}}^{\boldsymbol{m_\tau}})$ where $\boldsymbol{u}_{\boldsymbol{\tau}}^{\boldsymbol{m_\tau}} = \{u_\perp, u_\Delta^2, \dots, u_\gamma^{m_\gamma}\}$. The size of the largest percolating cluster S can then be calculated as

$$S = 1 - G_0(\boldsymbol{u}_{\boldsymbol{\tau}}^{\boldsymbol{m_\tau}}) \tag{11}$$

Introducing $H(\boldsymbol{x})$ as the generating function for the GCC as

$$H(\boldsymbol{x}) = \frac{G_0(\boldsymbol{x}) - G_0(\boldsymbol{x} \cdot \boldsymbol{u}_{\boldsymbol{\tau}}^{\boldsymbol{m_\tau}})}{1 - G_0(\boldsymbol{u}_{\boldsymbol{\tau}}^{\boldsymbol{m_\tau}})} \tag{12}$$

where $\boldsymbol{v} \cdot \boldsymbol{w}$ is the scalar product $v_i w_i$. The overall degree distribution of the GCC is given by

$$p_k^{\mathrm{GCC}} = \frac{1}{k!} \frac{\partial^k}{\partial x^k} H(\boldsymbol{x}^{\boldsymbol{m_\tau}}) \Big|_{\boldsymbol{x}=\boldsymbol{1}} \tag{13}$$

where $\boldsymbol{x} = (x, x^2, \dots, x^{m_\gamma})$. The networks that we use in this paper are constructed according to the GCM which we now detail [28–31]. For each vertex in a collection of vertices, a joint degree is chosen from a distribution of joint degrees to create a joint degree sequence. Not all joint degree sequences are valid or *graphic* [23]. There is a constraint on the permissible sequence of joint degrees generated such that the sum of the number of motifs of each kind is divisible by the number of vertices in each basis motif. For instance, the number of triangles in the joint degree sequence must be divisible by 3 and so on. This ensures that when the vertices are chosen at random and connected, there are precisely the correct number of edges to construct each motif. This constraint does not impact the number of each motif in the network; however.

Once the vertices have been assigned their stub degrees, they are connected at random to form the appropriate subgraphs according to their joint degree sequence through a stub-matching process. The probability of accidental formation of short range loops or motifs that share edges (non-edge disjoint motifs) becomes vanishingly small in the limit that the networks are large. Upon renormalising each motif to its characteristic scale based on neighbouring vertex count, we recover the tree-like property of the original configuration model.

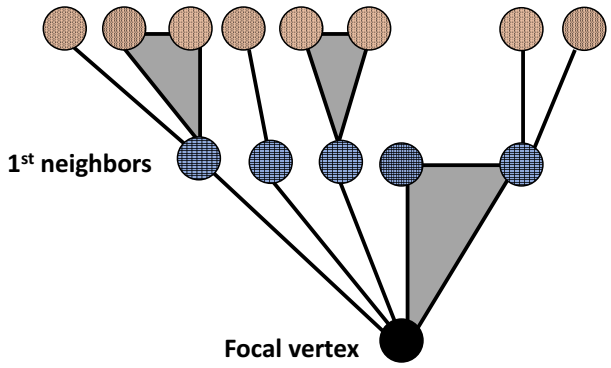

FIG. 1. A focal vertex in a 2- and 3-clique random graph with $n_{\perp,\perp,1} = n_{\perp,\Delta,0} = n_{\Delta,\perp,0} = n_{\Delta,\perp,2} = 1$ and $n_{\perp,\perp,2} = n_{\perp,\Delta,1} = n_{\Delta,\Delta,1} = 2$.

III. THEORETICAL

Consider an arbitrary set of edge topologies including ordinary edges, triangles, squares, 4-cliques, pentagons and so on, denoted by $\vec{\boldsymbol{\tau}} = \{\perp, \Delta, \square, \dots, \gamma\}$, where γ is the topology of the final element. In the following, we reserve τ and ν as indices over elements of $\boldsymbol{\tau}$. We define the number of subgraphs that a vertex plays a role in for each topology $\tau \in \boldsymbol{\tau}$ by vector $\boldsymbol{k}_{\boldsymbol{\tau},l} = \{k_\perp, k_\Delta, \dots, k_\gamma\}$ with $l = 0, 1$ representing the focal vertex and nearest-neighbour joint sequences, respectively. We reserve $k_{\nu,l} \in \boldsymbol{k}_{\boldsymbol{\tau},l}$ as an index for the number of subgraphs of topology ν around a given vertex in layer l; we drop the l label where obvious. The joint probability distribution for choosing this vertex at random is then denoted as $p_{\boldsymbol{k}_{\boldsymbol{\tau},l}}$. The number of edges that a given vertex has within each motif is defined by m_τ; for instance a vertex contributes two edges to each triangle it connects to and hence $m_\Delta = 2$.

We define n_{τ,ν,k_ν} to be the number of vertices with k_ν subgraphs of topology ν that we reach by following an edge of topology τ from the focal vertex to a nearest neighbour. There are $\dim(\boldsymbol{\tau}^2)$ of these expressions. Let a particular configuration of type ν following τ edges be $n_{\tau,\nu}$ such that

$$n_{\tau,\nu} = \{n_{\tau,\nu,1}, n_{\tau,\nu,2}, \dots\} \tag{14}$$

For instance, for a focal vertex that belongs to a GCM graph comprising of vertices with both 2- and 3-cliques such that $\boldsymbol{\tau} = \{\perp, \Delta\}$, the configuration of 3-cliques obtained by following 2-cliques to a neighbour is

$$n_{\perp,\Delta} = \{n_{\perp,\Delta,1}, n_{\perp,\Delta,2}, \dots, n_{\perp,\Delta,k_{\Delta,\max}}\} \tag{15}$$

where $k_{\Delta,\max}$ is the maximum number of triangles a single vertex belongs to, see Fig 1.

Then, we define the set of all configurations of the neighbours following τ edges to be $n_\tau = \{n_{\tau,\perp}, n_{\tau,\Delta}, \dots\}$.

For instance, returning to the mixed 2- and 3-clique example, we can also count the number of 2-cliques the neighbour has instead of enumerating the 3-cliques. Therefore, for this example we have

$$n_\perp = \{n_{\perp,\perp}, n_{\perp,\Delta}\} \tag{16}$$

Finally, the set of all configurations of neighbour motif membership is denoted by $n = \{n_\perp, n_\Delta, \dots\}$, which accounts for each edge-type we could have followed to reach the neighbour vertices.

The number of vertices reached by following all of the τ edges is

$$N_\tau = \sum_{k_\tau=1} n_{\tau,\tau,k_\tau} = \sum_{\nu=0} n_{\tau,\nu,k_\nu} \qquad \tau \neq \nu \tag{17}$$

For instance, for the focal vertex in Fig 1, we have

$$\sum_{k_\perp=1} n_{\perp,\perp,k_\perp} = \sum_{k_\Delta=0} n_{\perp,\Delta,k_\Delta} = 3 \tag{18}$$

and

$$\sum_{k_\Delta=1} n_{\Delta,\Delta,k_\Delta} = \sum_{k_\perp=0} n_{\Delta,\perp,k_\perp} = 2 \tag{19}$$

The total number of vertices 1-layer out from the focal vertex is the sum of all vertices reached by traversing each edge topology

$$N = \sum_{\tau\in\boldsymbol{\tau}} N_\tau \tag{20}$$

and hence, for the focal vertex in Fig 1, the total number of direct neighbours is given by $N = 5$.

Let $P(n \mid N)$ be the probability that the nearest-neighbour configuration is given by set n and that the total number of vertices in the first layer is N. This is given by

$$P(n \mid N) = \prod_\tau \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} q_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} q_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \tag{21}$$

where $q_{\tau,\nu,k}$ is the probability of traversing an edge of topology τ to a vertex with k_ν independent subgraphs of topology ν. We also have the understanding that each term of the product over $\nu \neq \tau$ has its own index k_ν starting from zero; we have pulled out τ from this expression since, by definition, there must be at least one τ-edge present to follow it to a nearest neighbour vertex and so the index starts at 1. The probability $P(\mathrm{GCC} \mid n)$ that the component is the GCC for a particular configuration n is given by

$$P(\mathrm{GCC} \mid n, N) = 1 - \prod_\tau \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} \left[u_\nu^{m_\nu k_\nu}\right]^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \left[u_\tau^{m_\tau(k_\tau-1)}\right]^{n_{\tau,\tau,k_\tau}} \tag{22}$$

where we have introduced u_τ as the probability that a vertex at the end of a randomly chosen edge of topology τ fails to connect to the GCC. The probability that the configuration is n, that the component is the GCC given that there are N nearest-neighbours is found from Bayes' theorem as

$$P(n, \mathrm{GCC} \mid N) = P(\mathrm{GCC} \mid n, N) P(n \mid N) \tag{23}$$

Let $P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ be the probability of there being N vertices in the 1st layer given that the joint degree of the focal vertex is $\boldsymbol{k}_{\boldsymbol{\tau},0}$ and that the component is the GCC. We can use this to find the probability $P(n, \mathrm{GCC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ that the nearest-neighbour configuration is n given the joint degree of a vertex in the GCC is $\boldsymbol{k}_{\boldsymbol{\tau},0}$ as

$$P(n, \mathrm{GCC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) P(n, \mathrm{GCC} \mid N) \tag{24}$$

where the summation is over all combinations of N_τ such that

$$\sum_N = \sum_{N_\perp} \sum_{N_\Delta} \cdots \tag{25}$$

We find

$$\begin{aligned} P(n, \mathrm{GCC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = & \sum_N P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_\tau \left(\prod_{\nu\neq\tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} q_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} q_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \\ & \times \left[1 - \prod_\eta \left(\prod_{\varphi\neq\eta} \prod_{k_\nu=0} \left[u_\varphi^{m_\varphi k_\nu}\right]^{n_{\eta,\varphi,k_\nu}} \right) \prod_{k_\tau=1} \left[u_\eta^{m_\eta(k_\tau-1)}\right]^{n_{\eta,\eta,k_\tau}} \right] \qquad \tau,\nu,\eta,\varphi \in \boldsymbol{\tau} \end{aligned} \tag{26}$$

We now generate this probability by summing over all permissible configurations of the nearest-neighbour joint degrees to obtain

$$\tilde{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \sum_{n} P(n, \mathrm{GCC} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left(\prod_{\nu \neq \tau} \prod_{k_\nu=0} X_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} X_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \tag{27}$$

where

$$\sum_{n} = \sum_{n_{\perp,\perp}} \sum_{n_{\perp,\Delta}} \cdots \sum_{n_{\Delta,\perp}} \sum_{n_{\Delta,\Delta}} \cdots \tag{28}$$

We simplify the expression by substituting Eq 26, swapping the order of the summations and collecting terms in like powers to obtain

$$\begin{aligned} \tilde{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = & \sum_{n} \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left(\prod_{\nu \neq \tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} q_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} q_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \\ & \times \left[1 - \prod_{\eta} \left(\prod_{\varphi \neq \eta} \prod_{k_\nu=0} [u_\varphi^{m_\varphi k_\nu}]^{n_{\eta,\varphi,k_\nu}} \right) \prod_{k_\tau=1} [u_\eta^{m_\eta (k_\tau - 1)}]^{n_{\eta,\eta,k_\tau}} \right] \prod_{\tau} \left(\prod_{\nu \neq \tau} \prod_{k_\nu=0} X_{\tau,\nu,k_\nu}^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} X_{\tau,\tau,k_\tau}^{n_{\tau,\tau,k_\tau}} \end{aligned} \tag{29}$$

to find

$$\begin{aligned} \tilde{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = & \sum_{n} \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left(\prod_{\nu \neq \tau} \prod_{k_\nu=0} \frac{N_\tau}{n_{\tau,\nu,k_\nu}!} (q_{\tau,\nu,k_\nu} X_{\tau,\nu,k_\nu})^{n_{\tau,\nu,k_\nu}} \right) \prod_{k_\tau=1} \frac{N_\tau}{n_{\tau,\tau,k_\tau}!} (q_{\tau,\tau,k_\tau} X_{\tau,\tau,k_\tau})^{n_{\tau,\tau,k_\tau}} \\ & \times \left[1 - \prod_{\eta} \left(\prod_{\varphi \neq \eta} \prod_{k_\nu=0} \left[u_\varphi^{m_\varphi k_\nu} \right]^{n_{\eta,\varphi,k_\nu}} \right) \prod_{k_\tau=1} [u_\eta^{m_\eta (k_\tau - 1)}]^{n_{\eta,\eta,k_\tau}} \right] \end{aligned} \tag{30}$$

The multinomial theorem can now be applied to each of the terms in the product to obtain

$$\begin{aligned} \tilde{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = & \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left[\left(\prod_{\nu \neq \tau} \sum_{k_\nu=0} q_{\tau,\nu,k_\nu} X_{\tau,\nu,k_\nu} \right) \sum_{k_\tau=1} q_{\tau,\tau,k_\tau} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \\ & - \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left[\left(\prod_{\nu \neq \tau} \sum_{k_\nu=0} q_{\tau,\nu,k_\nu} u_\nu^{m_\nu k_\nu} X_{\tau,\nu,k_\nu} \right) \sum_{k_\tau=1} q_{\tau,\tau,k_\tau} u_\tau^{m_\tau (k_\tau - 1)} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \end{aligned} \tag{31}$$

The probability that an edge of topology τ can be followed to reach a vertex with k_ν subgraphs of topology ν is given by q_{τ,ν,k_ν}. The probability that an edge of topology τ can be traversed to reach a vertex with k_ν motifs of topology ν for all $\nu \in \boldsymbol{\tau}$ is the joint excess degree distribution, $q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},l}}$. This can be constructed from the separable distributions such that

$$q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},l}} = \prod_{\nu} q_{\tau,\nu,k_{\nu,l}} \tag{32}$$

With this we can write

$$\begin{aligned} \tilde{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = & \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \\ & - \sum_{N} P(N \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} u_\nu^{m_\nu k_\nu} u_\tau^{m_\tau (k_\tau - 1)} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{N_\tau} \end{aligned} \tag{33}$$

The probability that there are N nearest-neighbour vertices given the joint degree of the focal vertex is $\boldsymbol{k}_{\boldsymbol{\tau},0}$ is simply a particular term from the $G_0(\boldsymbol{Z})$ generating function. Inserting this definition into our expression we arrive at the generating function that describes the distribution of nearest-neighbours given a particular joint degree of the focal vertex as

$$\hat{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} - p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} u_\nu^{m_\nu k_\nu} u_\tau^{m_\tau(k_\tau-1)} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} \tag{34}$$

The expectation number of nearest-neighbours with a given joint degree is found from the expectation value of $\hat{F}_{\mathrm{GCC}}(\boldsymbol{X} = Z \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$. We then find

$$\hat{F}'_{\mathrm{GCC}} = \sum_{\tau \in \boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu \in \boldsymbol{\tau} \setminus \tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})} \right) \tag{35}$$

where the derivative is evaluated at $Z_{\boldsymbol{k}_{\boldsymbol{\tau},\mathbf{1}}} = 1$ (see Appendix A for a complete derivation using the tree-triangle model). The bracket is one minus the probability that the none of the edges to the second layer lead to the GCC; whilst the prefactor describes the probability of following $k_{\tau,0}$ τ-motifs, each of which has m_τ edges to follow to reach a vertex whose joint degree is given by $q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},\mathbf{1}}}$. The exponent of u_τ is the number of neighbouring vertices that can be reached by following edges belonging to τ-subgraphs incident to two vertices at the end of an edge in a τ motif. This is the total number of τ edges minus the m_τ that belong to the focal edge's motif minus the focal edge itself.

$$m_\tau(k_{\tau,0} + k_{\tau,1} - 2) + m_\tau - 1 = m_\tau(k_{\tau,0} + k_{\tau,1} - 1) - 1 \tag{36}$$

In a similar way, we can find the generating function $F_{\mathrm{GCC}}(\boldsymbol{X})$ for the probability distribution that a randomly chosen vertex has a nearest neighbour configuration given by n and belongs to the GCC as

$$F_{\mathrm{GCC}}(\boldsymbol{X}) = \sum_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \hat{F}_{\mathrm{GCC}}(\boldsymbol{X} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \tag{37}$$

$$= \sum_{\boldsymbol{k}_{\boldsymbol{\tau},0}} p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} - \sum_{\boldsymbol{k}_{\boldsymbol{\tau},0}} p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} \prod_{\tau} \left[\prod_{\nu \neq \tau} \sum_{k_\tau=1} \sum_{k_\nu=0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} u_\nu^{m_\nu k_\nu} u_\tau^{m_\tau(k_\tau-1)} X_{\tau,\nu,k_\nu} X_{\tau,\tau,k_\tau} \right]^{m_\tau k_{\tau,0}} \tag{38}$$

which is simply $G_0(\boldsymbol{Z})$. The expectation number for the of nearest-neighbours from a random focal vertex in the GCC is given by

$$F'_{\mathrm{GCC}} = \sum_{\tau \in \boldsymbol{\tau}} m_\tau \langle k_\tau \rangle [1 - u_\tau^{m_\tau \omega_\tau}] \tag{39}$$

where ω_τ represents the number of vertices in the motif. We can use the quotient of these expectation values to define a symmetric joint-probability distribution $P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1}) = \hat{F}'_{\mathrm{GCC}}/F'_{\mathrm{GCC}}$ that two nearest-neighbours in the GCC have joint degrees $\boldsymbol{k}_{\boldsymbol{\tau},0}$ and $\boldsymbol{k}_{\boldsymbol{\tau},1}$ as

$$P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1}) = \sum_{\tau \in \boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu \in \boldsymbol{\tau} \setminus \tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})} \right) / \sum_{\tau \in \boldsymbol{\tau}} m_\tau \langle k_\tau \rangle [1 - u_\tau^{m_\tau \omega_\tau}] \tag{40}$$

where $P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1}) = P_{\mathrm{GCC}}(k_{\perp,0}, \ldots, k_{\gamma,0}, k_{\perp,1}, \ldots, k_{\gamma,1})$. This equation is a central result and can be used to compute many interesting properties of the correlation structure within configuration model networks. At any time, we can compress the information contained within $P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1})$ to find $P_{\mathrm{GCC}}(k_0, k_1)$ which is the probability that a focal vertex with overall degree k_0 attaches to a neighbour whose overall degree is k_1.

$$P^{\mathrm{overall}}_{\mathrm{GCC}}(k_0, k_1) = \sum_{\tau} \sum_{k_\tau} P_{\mathrm{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1}) \delta_{k_0, k_0^{\mathrm{overall}}} \delta_{k_1, k_0^{\mathrm{overall}}} \tag{41}$$

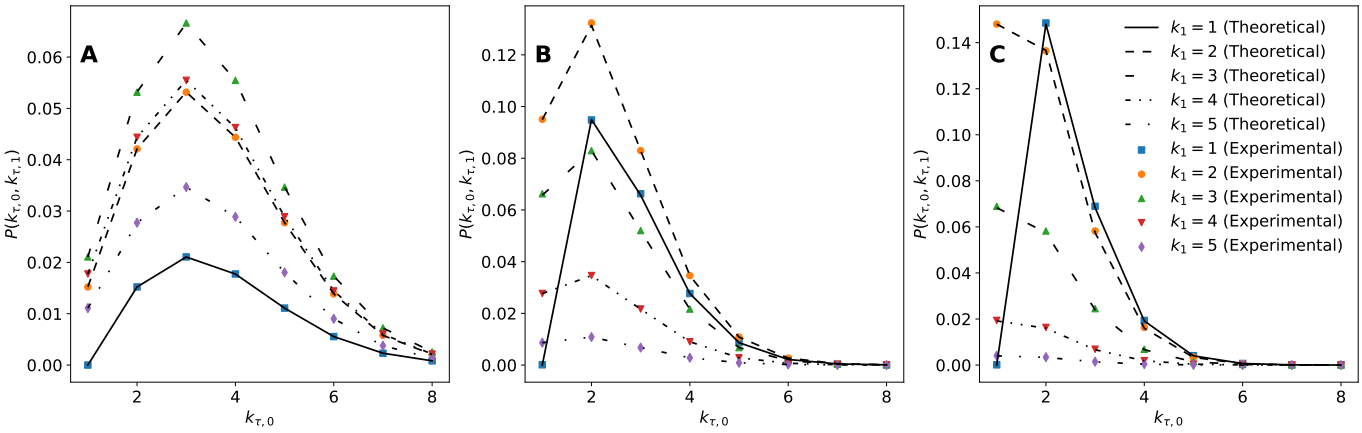

FIG. 2. The probability $P(k_{\tau,0}, k_{\tau,1})$ for Poisson random graphs comprising of a single motif topology, 2-cliques (A), 3-cliques (B) and 4-cliques (C), respectively, as a function of $k_{\tau,0}$ for several $k_{\tau,1}$. The overall mean degree is fixed at $\langle k \rangle = 2.5$ for networks with $N = 60000$ vertices. Scatter points are the average of 100 repetitions of Monte Carlo simulation while the lines are the theoretical predictions from Eq 47. The legend is the same for each plot.

where $k_0^{\text{overall}} = \sum_\tau \sum_{k_{\tau,0}} m_\tau k_{\tau,0}$ and $k_1^{\text{overall}} = \sum_\tau \sum_{k_{\tau,1}} m_\tau k_{\tau,1}$ are the overall degrees of the focal and neighbour vertices. However, this degree lumping procedure overlooks the fine structure among the correlations as many joint degrees can contribute to a given overall degree. Indeed it is precisely this structure which acts as a fingerprint of a network ensemble.

Let us introduce the conditional probability $P_{\text{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ that the nearest neighbour has joint degree $\boldsymbol{k}_{\boldsymbol{\tau},1}$ given that the focal vertex has joint degree $\boldsymbol{k}_{\boldsymbol{\tau},0}$ in the GCC. Applying Bayes' theorem to our discrete multivariate joint probability we have

$$P_{\text{GCC}}(k_{\perp,1}, \ldots, k_{\gamma,1} \mid k_{\perp,0}, \ldots, k_{\gamma,0}) = \frac{P_{\text{GCC}}(k_{\perp,0}, \ldots, k_{\gamma,0} \mid k_{\perp,1}, \ldots, k_{\gamma,1}) P_{\text{GCC}}(k_{\perp,1}, \ldots, k_{\gamma,1})}{\sum_{k_{\perp,1}, \ldots, k_{\gamma,1}} P_{\text{GCC}}(k_{\perp,0}, \ldots, k_{\gamma,0} \mid k_{\perp,1}, \ldots, k_{\gamma,1}) P_{\text{GCC}}(k_{\perp,1}, \ldots, k_{\gamma,1})} \tag{42}$$

Which simplifies to

$$P_{\text{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \frac{P_{\text{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1})}{\sum_{\boldsymbol{k}_{\boldsymbol{\tau},1}} P_{\text{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},0}, \boldsymbol{k}_{\boldsymbol{\tau},1})} \tag{43}$$

Inserting Eq 40 we find

$$P_{\text{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) = \frac{\sum_{\tau \in \boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu \in \boldsymbol{\tau} \setminus \tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})}\right)}{\sum_{\tau \in \boldsymbol{\tau}} \sum_{k_{\tau,1}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} \left(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1} \prod_{\nu \in \boldsymbol{\tau} \setminus \tau} u_\nu^{m_\nu(k_{\nu,0}+k_{\nu,1})}\right)} \tag{44}$$

We can use $P_{\text{GCC}}(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0})$ to find multivariate conditional expectation values for a given focal vertex joint degree, generalising [32] for the GCM. The expectation value for vector $\boldsymbol{X}$ given vector $\boldsymbol{Y}$ is a vector $\mathcal{E}[\boldsymbol{X} \mid \boldsymbol{Y}] = (\mathcal{E}[X_1 \mid Y], \ldots, \mathcal{E}[X_n \mid \boldsymbol{Y}])^{\text{T}}$ whose elements are the expected values of each of the variables defined as

$$\mathcal{E}[X_i \mid \boldsymbol{Y} = \boldsymbol{y}] = \sum_{x_1, \ldots, x_n} x_i P_{\text{GCC}}(x_1, \ldots, x_n \mid \boldsymbol{Y} = \boldsymbol{y}) \tag{45}$$

For instance, the average joint degree of a neighbour to a focal vertex whose joint degree is $\boldsymbol{k}_{\boldsymbol{\tau},\mathbf{0}}$ is the vector $(\mathcal{E}[k_{\perp,1} \mid \boldsymbol{k}_{\boldsymbol{\tau},\mathbf{0}}], \ldots, \mathcal{E}[k_{\gamma,1} \mid \boldsymbol{k}_{\boldsymbol{\tau},\mathbf{0}}])^T$ whose elements are

$$\mathcal{E}[k_{\tau,1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}] = \sum_{\boldsymbol{k}_{\boldsymbol{\tau},1}} k_{\tau,1} P(\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}) \tag{46}$$

We examine this expression in Appendix A for the tree-triangle model.

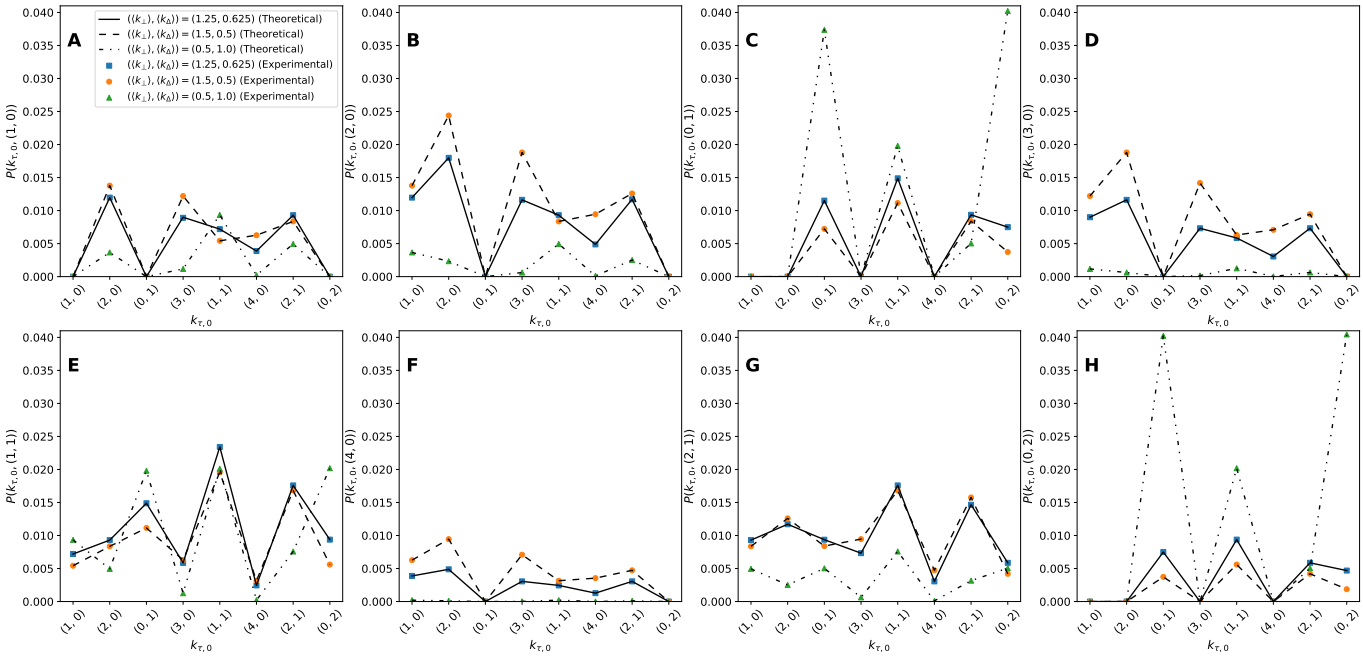

FIG. 3. The probability $P_{\rm GCC}(s_0, t_0, s_1, t_1)$ for Poisson random graphs comprising of mixed 2-clique and 3-clique topologies for three different clustering regimes. In each plot, the joint degrees of the focal vertex up to overall degree $k = 4$ are plotted on the horizontal axis for a given (s_1, t_1) neighbour. Scatter points are the average of 250 repetitions of Monte Carlo simulation on networks with 2×10^5 vertices; whilst lines are the analytical results of Eq 40. The legend is the same as tile (A) for all plots.

IV. DISCUSSION

In this paper we have introduced a theoretical model, based on generating functions, to investigate the NNDC in the GCC of random clustered graphs, constructed according to the GCM, comprising of higher-order clique clusters. We now examine a series of pertinent examples of this model.

A. Single topology

In the special case that the network consists of a single homogeneous subgraph (a homogeneous subgraph is one where all vertices are degree-equivalent), then $P_{\rm GCC}(k_{\tau,0}, k_{\tau,1})$ from Eq 40 is given by

$$P_{\rm GCC} = \frac{(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1})}{1 - u_\tau^{m_\tau \omega_\tau}} q_{\tau,k_{\tau,0}} q_{\tau,k_{\tau,1}} \qquad (47)$$

and similarly from Eq 44 we have the related conditional probability

$$P_{\rm GCC}(k_{\tau,1} \mid k_{\tau,0}) = \frac{(1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1})}{1 - u_\tau^{m_\tau k_{\tau,0}}} q_{\tau,k_{\tau,1}} \qquad (48)$$

which reproduces the results of [7, 8] for the nearest-neighbour distributions on the GCC of tree-like networks when $\tau = \perp$. We examine the NNDCs for single-topology networks with Poisson distribution participation in motifs with fixed overall mean degree $\langle k \rangle = 2.5$ in Fig 2. The networks are composed of discrete clique topologies; specifically 2, 3 and 4-cliques in Fig 2 A, B and C, respectively. The markers are the averaged results of Monte Carlo simulation while the lines are the theoretical predictions of Eq 47; both are in excellent agreement. In each case, $P_{\rm GCC}(k_{\tau,0}, k_{\tau,1})$ is plotted as a function of increasing $k_{\tau,0}$ for several $k_{\tau,1}$ values. We note that for each clique size $P_{\rm GCC}(1,1) = 0$; since, this combination cannot exist in the GCC. For networks comprised of a single topology, the average degree of a neighbour can be found from Eq 46 as

$$\mathcal{E}[k_{\tau,1} \mid k_{\tau,0}] = \frac{\sum\limits_{k_{\tau,1}} k_{\tau,1} q_{\tau,k_{\tau,1}} (1 - u_\tau^{m_\tau(k_{\tau,0}+k_{\tau,1}-1)-1})}{1 - u_\tau^{m_\tau k_{\tau,0}}} \qquad (49)$$

which is in agreement with [33] for tree-like topologies.

B. Tree-triangle model

We now examine how clustering influences the degree correlations in the GCC of the mixed topology tree-triangle model. The theoretical details of this model are derived in Appendix A. Fixing the first moment of the model to $\langle k \rangle = 2.5$ the limiting cases of $\langle k_\perp \rangle = 0$ and $\langle k_\Delta \rangle = 0$ are presented in Fig 2 and we now examine i)

an even neighbour distribution by setting $\langle k_\perp \rangle = 1.25$ and $\langle k_\Delta \rangle = 0.625$; ii) a weakly clustered regime with $\langle k_\perp \rangle = 1.5$ and $\langle k_\Delta \rangle = 0.5$ and finally iii) a strong clustering regime with $\langle k_\perp \rangle = 0.5$ and $\langle k_\Delta \rangle = 1.0$ in Fig 3. The joint degree of the horizontal axis is ordered by increasing overall degree. When a given overall degree can be formed in multiple ways, such as $k = 2$ from $(2, 0)$ or $(0, 1)$, the degenerate cases are ordered by increasing local clustering coefficient. Each tile in Fig 3 A-H plots a given neighbour joint degree (as a function of the focal vertex joint degree) for the three clustering regimes. We observe some encouraging results from these plots: firstly, as with the results of experiments with single-topology networks (Fig 2), the probabilities $P_{\mathrm{GCC}}(1, 0, 1, 0)$ and $P_{\mathrm{GCC}}(0, 1, 0, 1)$ are both zero for the vertices in the GCC (see Fig 3 A). We also notice that $P_{\mathrm{GCC}}(s_0, t_0, s_1, t_1)$ takes zero values for impossible combinations, such as neighbours whose edges are of a single, yet opposite, topology to one another. Further, the probabilities are symmetric such that $P_{\mathrm{GCC}}(k_{\boldsymbol{\tau},0}, k_{\boldsymbol{\tau},1}) = P_{\mathrm{GCC}}(k_{\boldsymbol{\tau},1}, k_{\boldsymbol{\tau},0})$ which is an expected result for undirected random graphs. Among the non-zero combinations we observe that some peaks, particularly among focal vertices with non-zero degrees in both topologies, are aligned across all series; for example $P_{\mathrm{GCC}}(1, 1, 1, 1)$ in E. Conversely, other peaks such as $P_{\mathrm{GCC}}(2, 0, 2, 1)$ in G peak in the weak and even regimes, yet trough in the strong clustered regime.

We also observe, across all tiles in Fig 3 that the correlations among the weak (blue squares) and even-neighbour (orange circles) regimes are generally of higher magnitude across all focal vertices than the strongly clustered regime (green triangles). In other words, the networks with strong clustering exhibit NNDC that have smaller magnitudes with the exception of tiles C and H, which consider neighbouring vertices that only have triangle motifs.

In tile F we notice that vertices with a high tree-like degree do not tend to connect with neighbours with triangles, especially in the strong clustering regime.

Collectively, these results give insight into how the network is held together at the microscopic level and how the presence of clustering alters this structure. This could prove useful for creating synthetic networks or for a better understanding of network resilience under targeted attack.

C. The effect of clique size on NNDC

In this section, we examine the effect of increasing the clique size on the NNDC of mixed topology GCM networks. To achieve this, we extend the calculations performed in appendix A from the 2- and 3-clique model to a binary model composed of 2- and m-cliques, whose topology we denote by σ. For this model, the NNDC for a focal vertex with s_0 ordinary edges and c_0 edge-disjoint m-cliques in the GCC of a GCM network can be obtained from

$$P_{\mathrm{GCC}}(s_0, c_0, s', c') = \frac{p_{s_0 c_0} s_0 q_{\perp,(s',c')} \left[1 - u_\perp^{s_0+s'-2} u_\sigma^{m_\sigma(c_0+c')}\right] + m_\sigma c_0 p_{s_0 c_0} q_{\sigma,(s',c')} \left[1 - u_\perp^{s_0+s'} u_\sigma^{m_\sigma(c_0+c'-1)-1}\right]}{\langle s \rangle (1 - u_\perp^2) + m_\sigma \langle c \rangle (1 - u_\sigma^{\omega_\sigma})} \tag{50}$$

The results of this expression are shown in Fig 4, where the overall neighbour degree is plotted against the overall degree of the focal vertex for several increasing clique sizes. The scatter points are the results of Monte Carlo simulation of networks with 100000 vertices, whilst the plotted lines are the theoretical results of the model; both show excellent agreement with one another. The networks are constructed according to the GCM algorithm before the GCC is selected from the possibly disconnected graph. The motifs counts at each vertex are drawn from Poisson distributions with averages chosen such that the first moment of the distribution of overall degrees is fixed at $\langle k \rangle = 6$ across all experiments whilst the average 2-clique count is held fixed at $\langle k_\perp \rangle = 1.25$ and the average clique count $\langle k_\sigma \rangle$ is the solution of $\langle k \rangle = \langle k_\perp \rangle + m_\sigma \langle k_\sigma \rangle$. From Fig 4 we observe that the average neighbour degree of networks with larger cliques increases. For cliques larger than 2-cliques, oscillations in the average neighbour degree appear at low focal vertex degree. The amplitude of the oscillations increases with clique size. In each case, the oscillations dampen to a fixed value in the limit of large focal vertex degree.

D. Emergence of correlations

At criticality, as the GCC emerges, we have that $u_\tau \to 1$; the probability of not belonging to the GCC is near unity. In this case, the multivariate limit of Eq 40 does not exist. However, in the case that the network is composed of cliques of various sizes which are each independently Poisson distributed at each vertex such that

$$p_{\boldsymbol{k}_{\boldsymbol{\tau},l}} = q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},l}} = \prod_{\tau \in \boldsymbol{\tau}} e^{-\langle k_\tau \rangle} \frac{\langle k_\tau \rangle^{k_{\tau,l}}}{k_{\tau,l}!} \qquad \forall \tau \in \boldsymbol{\tau} \tag{51}$$

we have that $u_\tau = u^{m_\tau}, \forall \tau$ [10]. In this instance Eq 40 is a univariate distribution and we can use L'Hôpital's rule

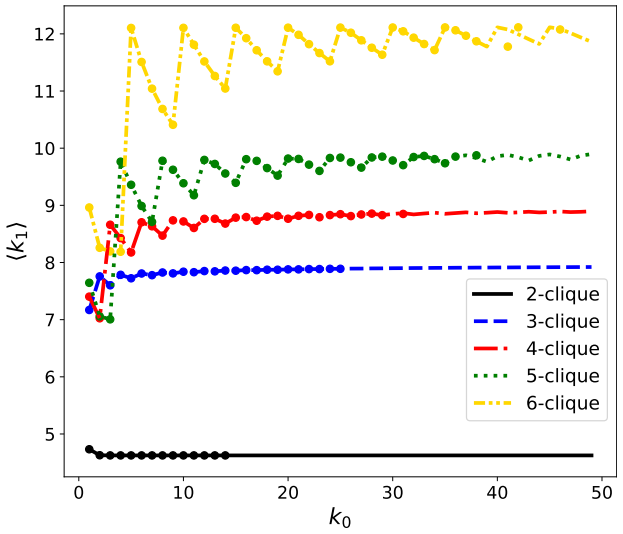

FIG. 4. The average overall degree of a neighbour for increasing focal vertex degree for binary-topology networks comprising 2-cliques and higher-order cliques. Scatter points are the average of 1000 repetitions of Monte Carlo simulation whilst the plotted lines are the result Eq 50, collected by overall degree according to Eq 41. The networks are created from the GCM algorithm with Poisson marginal distributions of each motif topology and overall average degree fixed at $\langle k \rangle = 6$ with $\langle k_\perp \rangle = 1.25$ across all experiments.

to determine the expected limit to be

$$\lim_{u\to 1} P_{\mathrm{GCC}}(k_{\tau,0}, k_{\tau,1}) = \frac{\sum_\tau m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} \Lambda_\tau q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}}}{\sum_\tau m_\tau^2 \omega_\tau \langle k_\tau \rangle} \quad (52)$$

where

$$\Lambda_\tau = m_\tau(k_{\tau,0} + k_{\tau,1} - 1) - 1 + \sum_{\nu\neq\tau} m_\nu (k_{\nu,0} + k_{\nu,1}) \quad (53)$$

The critical point can be found by linearising $u_\tau = G_{1,\tau}(\boldsymbol{u}_{\boldsymbol{\tau}}^{\boldsymbol{m}_{\boldsymbol{\tau}}})$ in a small perturbation ϵ around $u_\tau = 1 - \epsilon_\tau$ [11]. To leading order in the small parameter ϵ_τ we have $\boldsymbol{\epsilon} = \boldsymbol{A\epsilon}$ with $\boldsymbol{\epsilon} = [\epsilon_\perp, \epsilon_\Delta, \dots]^{\mathrm{T}}$. The GCC forms at the point when the determinant $\det|A - I|$ vanishes, where $A = [\partial \mathrm{G}/\partial u_\tau]$, $\mathrm{G} = [G_{1,\tau}, G_{1,\Delta}, \dots, G_{1,\gamma}]$ and identity matrix I. With mixed topology networks a GCC can form in many different ways. For instance, the GCC of a random graph model with two topologies can form by three distinct mechanisms: a GCC can emerge solely in either of the topologies or global connectivity can occur through a mixture of the binary topologies.

As we approach the critical point from below, we introduce a characteristic scale κ_τ [34] associated to the joint degrees of the focal vertex and a neighbour given by $u_\tau = e^{-1/\kappa_\tau}$. Inserting this expression into Eq 40 for finite κ_τ in each topology, the correlations fall exponentially with increasing κ_τ and hence $P_{\mathrm{GCC}}(k_{\tau,0}, k_{\tau,1})$ tends to the uncorrelated value of

$$\sum_{\tau\in\boldsymbol{\tau}} m_\tau p_{\boldsymbol{k}_{\boldsymbol{\tau},0}} k_{\tau,0} q_{\tau,\boldsymbol{k}_{\boldsymbol{\tau},1}} / \sum_{\tau\in\boldsymbol{\tau}} m_\tau \langle k_\tau \rangle \quad (54)$$

Therefore, when the joint degree exceeds the characteristic scale, the GCC is uncorrelated. It is clear that as u_τ approaches unity the scale diverges $\kappa_\tau \to \infty$ and hence, the GCC always exhibits degree correlations. In addition, approaching the critical point, the average joint degree (Eq 49) falls exponentially with increasing degree along each topology for fixed κ_τ.

$$\mathcal{E}[k_{\tau,1} \mid k_{\tau,0}] = \frac{\sum_{k_{\tau,1}} k_{\tau,1} q_{\tau,k_{\tau,1}} (1 - e^{-\phi})}{1 - e^{-m_\tau k_{\tau,0}/\kappa_\tau}} \quad (55)$$

where $\phi = m_\tau(k_{\tau,0} + k_{\tau,1} - 1) - 1/\kappa_\tau$. Thus, the correlations which are present at the critical point are negative in nature. It might happen, however, given the number of ways that the GCC of a mixed motif random graph model can emerge, that the characteristic scales of all topologies don't diverge at the critical point. For instance, consider a doubly Poisson distributed tree-triangle model with a critical average tree degree, but a sub-critical average triangle degree. A GCC will form among the tree edges, but the probability of those vertices involved only in triangles, $(0, t)$ for $t = 1, 2, 3, \dots$, connecting to this GCC is small; since, their connection requires them to connect to mixed-topology vertices, which in turn connect to the GCC. Thus, we might find that the negative degree correlation structure among the triangles has not yet formed despite there being a non-zero density of triangles in the GCC.

E. Empirical networks

We now examine the correlation properties of the GCC of the ensemble representation of empirical networks using our joint degree model. Random graphs are elements of an ensemble $\mathcal{G}$ of graphs with V vertices and E edges; each member occurring with probability $P(G)$ [7]. The average value of a property of graph G, $Z(G)$, (such as its degree distribution or average degree) can be averaged over the entire ensemble

$$\langle Z \rangle = \sum_{G\in\mathcal{G}} Z(G) P(G) \quad (56)$$

The generating function formulation describes the properties of the ensemble. Empirical networks g are particular realisations of members of $\mathcal{G}$. The properties of a particular realisation are given by

$$P(Z) = \sum_{G\in\mathcal{G}} \delta(Z - Z(G)) P(G) \quad (57)$$

If $P(Z)$ is well represented by the ensemble average then the generating function formulation can be used to describe the properties of g. To study the NNDC in the

GCC of g using generating functions, we must represent the largest component of an empirical network by a joint degree sequence of subgraphs. Whilst the choice of subgraphs is arbitrary [12], we only include cliques in the topology representation due to the vast literature on clique finding algorithms and the simplicity of calculating their properties. The clique decomposition of the GCC of g whose cliques have order less than or equal to ω can be performed in many different ways; and the resulting joint degree sequence can exhibit significantly different properties in terms of the number of subgraphs present their clustering, and other properties. Given that the method to create the joint degree distribution is not unique, and that the ensemble properties of each particular decomposition are often dissimilar, we now examine three clique decompositions and compare their properties.

The trivial decomposition is to simply cover g with 2-cliques; we refer to this as the single-edge-decomposition (SED). The degree sequence can then be used to create realisations using the ordinary configuration model. Another simple cover is the minimal cover of maximal cliques. However, it is very likely that the edges of the cliques will not be disjoint, i.e. a single edge will be a member of more than one clique. Whilst this could be an accurate representation of a vertex's local environment, the construction process for random graphs using the GCM will not work. Thus, we must impose that the cover is edge-disjoint.

One proposed method of clique decomposition is defined heuristically as follows [25]: we obtain the set C of all maximal cliques from the network; each maximal n-clique $c_i \in C$, $n \in \{1, \ldots, \omega\}$ is scored according to the fraction of edges it shares with other members of C. The largest clique within the set of lowest score cliques are included in the representation and C is recalculated. The process is repeated until the edges of the substrate network are expended. Such a covering is known as a edge-disjoint edge clique cover (EECC), see Fig 5 for details.

We propose a novel alternative clique cover as follows: the set $\mathcal{C}$ of all cliques present in the network (including those induced from subgraphs of larger cliques) is obtained from the empirical network. The set is ordered such that the largest cliques have the highest precedence. The subset of cliques within $\mathcal{C}$ that have equal size $\forall n \in \{1, \ldots, \omega\}$ are then scored in a similar fashion to the EECC algorithm and the cliques with the lowest score (and therefore the least number of overlapping edges with other motifs) are given highest precedence. The order of cliques with equivalent size and score is then randomised, thus the cover is stochastic. The largest cliques are drawn from $\mathcal{C}$ and placed on the network if their edges do not overlap other with cliques that have already been placed in the network. The list is iterated until all edges belong to an independent clique. This method draws non-maximal joint degree sequences; however, higher-order cliques are preferentially preserved, we describe it as an edge disjoint motif preserving edge clique cover (MPCC), see Fig 6. In the particular case that the set of maximal cliques are edge disjoint, the distribution obtained from both the EECC and MPCC motif decomposition algorithms are in agreement with one another. It should be mentioned that both covers are not unique when two cliques of a given size and score can be chosen. Within the MPCC, we resolve these degeneracies by retaining the cliques associated with higher degree vertices. In our implementation of the EECC, we choose cliques from the set of degenerate cliques at random. Once a suitable cover has been formed for the network, its joint-degree sequence can be extracted. This sequence is then used to create an ensemble of GCM networks. As a concrete example of this method we extract the joint degree sequences, using the SED, EECC and the MPCC, of the GCC of the network science authorship network [35] and use the GCM algorithm to construct random graph ensembles. Plotted in Fig 8 are the experimental results from the original network (red crosses), the SED (green squares), the EECC (pink triangles) and the results from the MPCC algorithm (light blue circles) as well as their average (dark blue circles). The average neighbour degree, k_1 obtained from the SED shows poor accuracy when compared to the experimental results. Instead of the detailed NNDC structure over the range of focal vertex degrees, the neighbour degrees tend to fluctuate around $k_1 = 8$. In contrast, the MPCC exhibits a rich correlation structure whose average follows the trends of the experimental data. Additionally, the average neighbour degree for the high-degree vertices is well represented; however, this is at the expense of

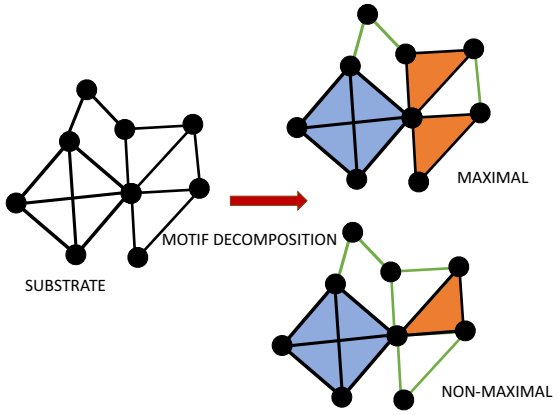

FIG. 5. The clique decomposition of a substrate network (left) can be performed in multiple ways. Two examples are shown (right). The shaded faces are higher-order cliques whilst the green edges are 2-cliques. The clustering of the resulting joint degree distributions (and their random graph ensembles) are significantly altered depending on how the decomposition is performed. The maximal representation has 6 cliques in total whilst the non-maximal representation has 8 cliques. When only maximal representations are extracted the decomposition is a EECC.

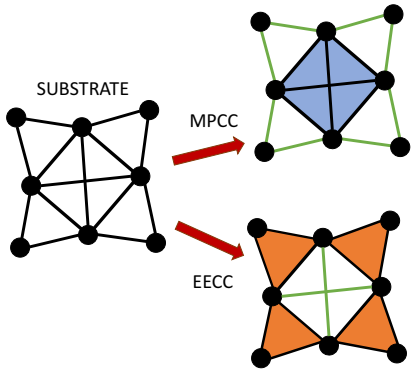

FIG. 6. The results of the to clique decomposition algorithms (MPCC) and (EECC) for a particular substrate graph. The MPCC favours the formation of large subgraphs, leading to 9 cliques (a single 4-clique and 8 2-cliques) whilst the EECC leads to 6 cliques (4 3-cliques and 2 2-cliques). The joint degree sequence obtained from the MPCC network creates a non-maximal random ensemble of GCM networks.

the lower degree information, where the representation is less accurate. The EECC shows fair agreement across the range of focal vertex degrees, outperforming the MPCC at low degrees; however, the MPCC represents the empirical network correlations for the high-degree vertices with greater accuracy than the EECC. The EECC representation of the high-degree sites is in agreement with the SED, indicating that these cliques are destroyed during the covering process. We notice from the variance of the MPCC that the NNDC of the empirical network is dense within the set of ensemble representations.

V. CONCLUSION

In this paper we have introduced a robust analytical framework to study the NNDC between vertices in the GCC of random graphs constructed according to the GCM. We have used our method to investigate the correlation properties of synthetic clustered GCM graphs in detail and found they exhibit organisation among their subgraphs. We studied the behaviour of the NNDC as the size of the substrate motif increases, along with the clustering for a fixed first moment of the overall average degree. We found that the NNDC among networks composed from larger cliques tend to be larger in magnitude for low degree vertices due to the constraint on the first moment of the overall degree.

Investigating the tree-triangle model in detail, we found that the joint degrees are negatively correlated along each topology as found for tree-like topologies in other studies [4, 7, 8].

The magnitude and the patterns of NNDC were found to vary significantly with the clustering coefficient of the network ensemble. The correlations among neighbours of mixed topology focal vertices in tree-triangle networks with larger clustering coefficients were smaller in magnitude, in general, with respect to the single-topology vertices.

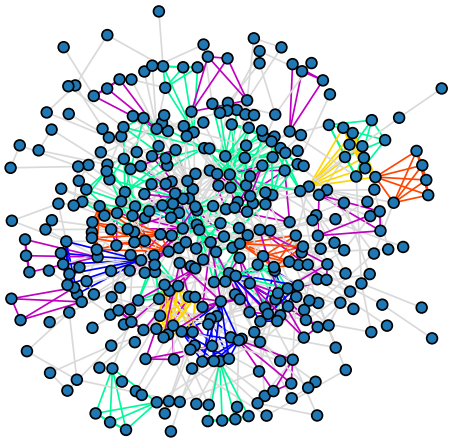

FIG. 7. A member of the MPCC random graph ensemble of the GCC of the network science authorship network with higher-order cliques (larger than 3-cliques) coloured for clarity. Specifically, the 4-cliques are magenta, 5-cliques are light green, 6-cliques are orange, 7-cliques are blue, 8-cliques are yellow and the 9-clique is cyan. Unlike random graphs constructed using the EECC method, larger cliques are preferentially retained in the ensemble.

We then investigated the role of clique size for GCM graphs and observed oscillations in the average overall neighbour degrees as a function of focal vertex degree. We found that the average neighbour degree in the GCC increases for networks composed of larger cliques.

Lastly, we studied the correlation structure of the random graph ensemble of an empirical network. To do this, we introduced a novel clique decomposition algorithm and compared it to other heuristics in the literature. We found that the manner in which the network is decomposed into motifs greatly effects the correlation substructure of the ensemble representation.

This work increases our understanding of the NNDC of clustered networks comprised of higher-order clique motifs; however, we have not addressed the long range correlation structure or defined an assortativity coefficient for these graphs, which we leave for future work.

VI. ACKNOWLEDGMENTS

This work was partially supported by the UK Engineering and Physical Sciences Research Council under grant number EP/N007565/1 (Science of Sensor Systems Software).

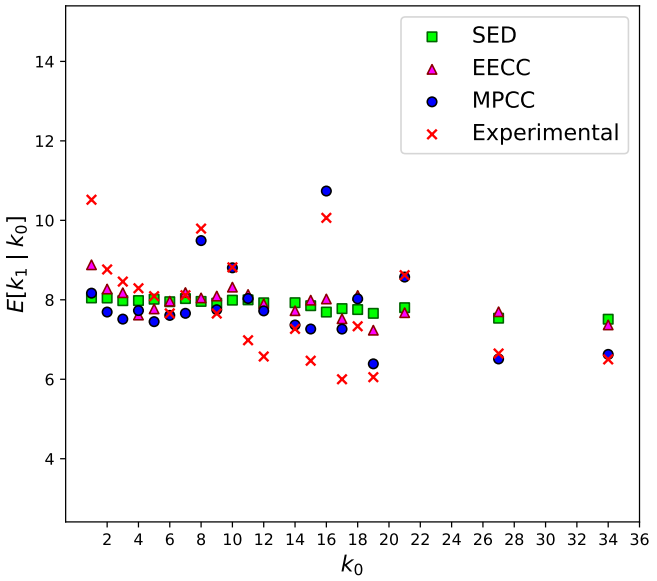

FIG. 8. The ensemble expectation value of the overall degree of a neighbour as a function of focal vertex degree for clique covers of the network science authorship network. Plotted are the experimental results (red crosses), the average EECC (pink triangles), the average MPCC (dark blue circles) and its variance (light blue circles) for each realisation. Each simulation was performed 1000 times. The SED (green squares) doesn't capture the correlation structure for this network. The MPCC accurately captures the correlation structure of the high-degree vertices due to retaining the larger motifs that a vertex belongs to; however, the low (mid) degree sites are generally under (over) predicted. Conversely, the EECC performs well for the low and mid-degree vertices, but tends to the SED for the high-degree sites.

* pm78@st-andrews.ac.uk

[1] M. E. Newman, *Networks*. Oxford University Press, 2019.

[2] M. Ángeles Serrano and M. Boguñá, "Tuning clustering in random networks with arbitrary degree distributions," *Phys. Rev. E*, vol. 72, p. 036133, Sep 2005.

[3] A. Arenas, A. Fernández, S. Fortunato, and S. Gómez, "Motif-based communities in complex networks," *Journal of Physics A: Mathematical and Theoretical*, vol. 41, p. 224001, may 2008.

[4] S. Mizutaka and T. Hasegawa, "Disassortativity of percolating clusters in random networks," *Phys. Rev. E*, vol. 98, p. 062314, Dec 2018.

[5] Y. Fujiki, T. Takaguchi, and K. Yakubo, "General formulation of long-range degree correlations in complex networks," *Phys. Rev. E*, vol. 97, p. 062308, Jun 2018.

[6] M. E. J. Newman, "Component sizes in networks with arbitrary degree distributions," *Phys. Rev. E*, vol. 76, p. 045101, Oct 2007.

[7] P. Bialas and A. K. Oleś, "Correlations in connected random graphs," *Phys. Rev. E*, vol. 77, p. 036124, Mar 2008.

[8] I. Tishby, O. Biham, E. Katzav, and R. Kühn, "Revealing the microstructure of the giant component in random graph ensembles," *Phys. Rev. E*, vol. 97, p. 042318, Apr 2018.

[9] I. Tishby, O. Biham, E. Katzav, and R. Kühn, "Generating random networks that consist of a single connected component with a given degree distribution," *Phys. Rev. E*, vol. 99, p. 042308, Apr 2019.

[10] B. Karrer and M. E. J. Newman, "Random graphs containing arbitrary distributions of subgraphs," *Physical Review E*, vol. 82, no. 6, 2010.

[11] P. Mann, V. A. Smith, J. B. O. Mitchell, and S. Dobson, "Percolation in random graphs with higher-order clustering," *Phys. Rev. E*, vol. 103, p. 012313, Jan 2021.

[12] P. Mann, V. A. Smith, J. B. O. Mitchell, and S. Dobson, "Random graphs with arbitrary clustering and their applications," *Phys. Rev. E*, vol. 103, p. 012309, Jan 2021.

[13] M. E. J. Newman, "Properties of highly clustered networks," *Phys. Rev. E*, vol. 68, p. 026121, Aug 2003.

[14] J. C. Miller, "Spread of infectious disease through clustered populations," Dec 2009.

[15] J. C. Miller, "Percolation and epidemics in random clustered networks," *Phys. Rev. E*, vol. 80, p. 020901, Aug 2009.

[16] J. P. Gleeson, "Bond percolation on a class of clustered random networks," *Physical Review E*, vol. 80, Oct 2009.

[17] A. Hackett, S. Melnik, and J. P. Gleeson, "Cascades on a class of clustered random networks," *Phys. Rev. E*, vol. 83, p. 056107, May 2011.

[18] J. P. Gleeson, S. Melnik, and A. Hackett, "How clustering affects the bond percolation threshold in complex networks," *Phys. Rev. E*, vol. 81, p. 066114, Jun 2010.

[19] A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé, "General and exact approach to percolation on random graphs," *Physical Review E*, vol. 92, Jul 2015.

[20] T. K. D. Peron, P. Ji, J. Kurths, and F. A. Rodrigues, "Spectra of random networks in the weak clustering regime," *EPL (Europhysics Letters)*, vol. 121, p. 68001, mar 2018.

[21] T. Hasegawa and Y. Iwase, "Observability transitions in clustered networks," *Physica A: Statistical Mechanics and its Applications*, vol. 573, p. 125970, 2021.

[22] T. Hasegawa and S. Mizutaka, "Structure of percolating clusters in random clustered networks," *Phys. Rev. E*, vol. 101, p. 062310, Jun 2020.

[23] P. Mann, V. A. Smith, J. B. O. Mitchell, C. Jefferson, and S. Dobson, "Exact formula for bond percolation on cliques," *Phys. Rev. E*, vol. 104, p. 024304, Aug 2021.

[24] C. Stegehuis and T. Peron, "Network processes on clique-networks with high average degree: the limited effect of higher-order structure," *Journal of Physics: Complexity*, vol. 2, p. 045011, nov 2021.

[25] G. Burgio, A. Arenas, S. Gómez, and J. T. Matamalas, "Network clique cover approximation to analyze complex contagions through group interactions," *Communications Physics*, vol. 4, no. 1, 2021.

[26] M. E. J. Newman, "Random graphs with clustering," *Phys. Rev. Lett.*, vol. 103, p. 058701, Jul 2009.

[27] S. Melnik, A. Hackett, M. A. Porter, P. J. Mucha, and J. P. Gleeson, "The unreasonable effectiveness of tree-based theory for networks with clustering," *Phys. Rev. E*, vol. 83, p. 036112, Mar 2011.

[28] B. K. Fosdick, D. B. Larremore, J. Nishimura, and J. Ugander, "Configuring random graph models with fixed degree sequences," *SIAM Review*, vol. 60, no. 2, p. 315–355, 2018.

[29] M. Ritchie, L. Berthouze, and I. Z. Kiss, "Generation and analysis of networks with a prescribed degree sequence and subgraph family: higher-order structure matters," *Journal of Complex Networks*, vol. 5, pp. 1–31, 05 2016.

[30] C. Wang, O. Lizardo, and D. Hachen, "Algorithms for generating large-scale clustered random graphs," *Network Science*, vol. 2, no. 3, p. 403–415, 2014.

[31] L. S. Heath and N. Parikh, "Generating random graphs with tunable clustering coefficients," *Physica A: Statistical Mechanics and its Applications*, vol. 390, no. 23-24, p. 4577–4587, 2011.

[32] R. Pastor-Satorras, A. Vázquez, and A. Vespignani, "Dynamical and correlation properties of the internet," *Phys. Rev. Lett.*, vol. 87, p. 258701, Nov 2001.

[33] S. Mizutaka and T. Hasegawa, "Emergence of long-range correlations in random networks," *Journal of Physics: Complexity*, vol. 1, p. 035007, sep 2020.

[34] C. Song, S. Havlin, and H. A. Makse, "Self-similarity of complex networks," *Nature*, vol. 433, no. 7024, p. 392–395, 2005.

[35] M. E. J. Newman, "Finding community structure in networks using the eigenvectors of matrices," *Phys. Rev. E*, vol. 74, p. 036104, Sep 2006.

Appendix A: Results within the tree-triangle model

In this section we derive the expectation values for the tree-triangle model. For this model the generating function for the number of nearest-neighbours given the joint degree of the focal vertex is $k_{\boldsymbol{\tau},0} = (s_0, t_0)$ is given by unpacking Eq 34 for $\boldsymbol{\tau} = \{\perp, \Delta\}$. We obtain

$$\hat{F}_{GCC}(\boldsymbol{x}, \boldsymbol{y}, s_0, t_0) = p_{s_0,t_0} f_\perp^{s_0} f_\Delta^{2t_0} - p_{s_0,t_0} g_\perp^{s_0} g_\Delta^{2t_0} \tag{A1}$$

where $f_\tau = \sum_s \sum_t q_{\tau,(s,t)} z_{st}$, $g_\perp = \sum_s \sum_t q_{\perp,(s,t)} u_\perp^{s-1} u_\Delta^{2t} x_s y_t$ and $\sum_s \sum_t q_{\Delta,(s,t)} u_\perp^{s} u_\Delta^{2(t-1)} x_s y_t$. The evaluation of the expectation values for the nearest-neighbours to a vertex of joint degree (s_0, t_0) in the tree-triangle model is given by the following derivative

$$\hat{F}'_{\mathrm{GCC}} = \left. \frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}} \right|_{z_{s't'}=1} \tag{A2}$$

We evaluate this as follows

$$\left. \frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}} \right|_{z_{s't'}=1} = \left. \frac{d}{dz_{s't'}} \right|_{z_{s't'}=1} p_{s_0 t_0} f_\perp^{s_0} f_\Delta^{2t_0} - \left. \frac{d}{dz_{s't'}} \right|_{z_{s't'}=1} p_{s_0 t_0} g_\perp^{s_0} g_\Delta^{2t_0} \tag{A3}$$

$$= p_{s_0 t_0} \left(s_0 f_\perp^{s_0-1} \frac{df_\perp}{dz_{s't'}} f_\Delta^{2t_0} + 2t_0 f_\perp^{s_0} f_\Delta^{2(t_0-1)} f_\Delta \frac{df_\Delta}{dz_{s't'}} \right)$$

$$- p_{s_0 t_0} \left(s_0 g_\perp^{s_0-1} \frac{dg_\perp}{dz_{s't'}} g_\Delta^{2t_0} + 2t_0 g_\perp^{s_0} g_\Delta^{2(t_0-1)} g_\Delta \frac{dg_\Delta}{dz_{s't'}} \right) \tag{A4}$$

At $z_{s't'}=1$ we have $f_\tau(1)=1$, $g_\tau(1)=G_{1,\tau}(u_\perp,u_\Delta^2)$ and also

$$\left.\frac{df_\tau}{dz_{s't'}}\right|_{z_{s't'}=1} = \frac{d}{dz_{s't'}}\sum_s\sum_t q_{\tau,(s,t)}z_{st} \tag{A5}$$

$$= q_{\tau,(s',t')} \tag{A6}$$

and

$$\left.\frac{dg_\perp}{dz_{s't'}}\right|_{z_{s't'}=1} = \frac{d}{dz_{s't'}}\sum_s\sum_t q_{\perp,(s,t)}u_\perp^{s-1}u_\Delta^{2t}z_{st} \tag{A7}$$

$$= q_{\perp,(s',t')}u_\perp^{s'-1}u_\Delta^{2t'} \tag{A8}$$

$$\left.\frac{dg_\Delta}{dz_{s't'}}\right|_{z_{s't'}=1} = \frac{d}{dz_{s't'}}\sum_s\sum_t q_{\Delta,(s,t)}u_\perp^{s}u_\Delta^{2(t-1)}z_{st} \tag{A9}$$

$$= q_{\Delta,(s',t')}u_\perp^{s'}u_\Delta^{2(t'-1)} \tag{A10}$$

Thus, we find

$$\begin{aligned}\left.\frac{d\hat{F}_{\rm GCC}}{dz_{s't'}}\right|_{z_{s't'}=1} = \; & p_{s_0t_0}\Big(s_0q_{\perp,(s',t')}+2t_0q_{\Delta,(s',t')}\Big) - p_{s_0t_0}\Big(s_0u_\perp^{s_0-1}q_{\perp,(s',t')}u_\perp^{s'-1}u_\Delta^{2t'}u_\Delta^{2t_0} \\ & + 2t_0u_\perp^{s_0}u_\Delta^{2(t_0-1)}u_\Delta q_{\Delta,(s',t')}u_\perp^{s'}u_\Delta^{2(t'-1)}\Big)\end{aligned} \tag{A11}$$

The evaluation of the expectation values for the nearest-neighbours to the average vertex in the tree-triangle model is given by the following derivative

$$F'_{\rm GCC} = \sum_{s'}\sum_{t'}\left.\frac{dF_{\rm GCC}}{dz_{s't'}}\right|_{z_{s't'}=1} \tag{A12}$$

where $F_{\rm GCC}$ is given by unpacking Eq 38 for $\boldsymbol{\tau}=\{\perp,\Delta\}$ to find

$$F_{GCC}(\boldsymbol{x},\boldsymbol{y}) = \sum_s\sum_t p_{s,t}f_\perp^s f_\Delta^{2t} - \sum_s\sum_t p_{s,t}g_\perp^s g_\Delta^{2t} \tag{A13}$$

To evaluate this consider the following derivative

$$\left.\frac{dF_{\rm GCC}}{dz_{s't'}}\right|_{z_{s't'}=1} = \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}G_0(f_\perp,f_\Delta) - \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}G_0(g_\perp,g_\Delta) \tag{A14}$$

$$= \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}\sum_s\sum_t p_{st}f_\perp^s f_\Delta^{2t} - \left.\frac{d}{dz_{s't'}}\right|_{z_{s't'}=1}\sum_s\sum_t p_{st}g_\perp^s g_\Delta^{2t} \tag{A15}$$

$$\begin{aligned} = \; & \sum_s\sum_t p_{st}\left\{sf_\perp^{s-1}\frac{df_\perp}{dz_{s't'}}f_\Delta^{2t} + 2tf_\perp^s f_\Delta^{2(t-1)}f_\Delta\frac{df_\Delta}{dz_{s't'}}\right\} \\ & - \sum_s\sum_t p_{st}\left\{sg_\perp^{s-1}\frac{dg_\perp}{dz_{s't'}}g_\Delta^{2t} + 2tg_\perp^s g_\Delta^{2(t-1)}g_\Delta\frac{dg_\Delta}{dz_{s't'}}\right\}\end{aligned} \tag{A16}$$

When evaluated at $z_{(s',t')}=1$ we have that $f_\tau(1)=1$ and so the first bracket simplifies significantly. The second bracket is more involved; however, using the self-consistent expressions for $u_\perp=G_{1,\perp}(u_\perp,u_\Delta^2)$ and $u_\Delta=G_{1,\Delta}(u_\perp,u_\Delta^2)$ we can write $g_\perp(1)=u_\perp$ and $g_\Delta(1)=u_\Delta$ to obtain

$$\begin{aligned}\left.\frac{dF_{\rm GCC}}{dz_{s't'}}\right|_{z_{s't'}=1} = & \sum_s\sum_t p_{st}\Big\{sq_{\perp,(s',t')}+2tq_{\Delta,(s',t')}\Big\} - \sum_s\sum_t p_{st}\Big\{su_\perp^{s-1}q_{\perp,(s',t')}u_\perp^{s'-1}u_\Delta^{2t'}u_\Delta^{2t} \\ & + 2tu_\perp^s u_\Delta^{2(t-1)}u_\Delta q_{\Delta,(s',t')}u_\perp^{s'}u_\Delta^{2(t'-1)}\Big\}\end{aligned} \tag{A17}$$

We now sum over (s',t') to obtain

$$\begin{aligned}\sum_{s'}\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} =& \sum_{s}\sum_{t}p_{st}\left\{s\sum_{s'}\sum_{t'}q_{\perp,(s',t')}+2t\sum_{s'}\sum_{t'}q_{\Delta,(s',t')}\right\}\\ &-\sum_{s}\sum_{t}p_{st}\left\{su_{\perp}^{s-1}u_{\Delta}^{2t}\sum_{s'}\sum_{t'}q_{\perp,(s',t')}u_{\perp}^{s'-1}u_{\Delta}^{2t'}\right.\\ &\left.+2tu_{\perp}^{s}u_{\Delta}^{2(t-1)}u_{\Delta}\sum_{s'}\sum_{t'}q_{\Delta,(s',t')}u_{\perp}^{s'}u_{\Delta}^{2(t'-1)}\right\}\end{aligned} \tag{A18}$$

The probability distributions are normalised and hence have the following property $\sum_s\sum_t q_{\tau,(s,t)}=1$, so the first bracket reduces trivially to the sum of the average degrees of each edge topology. The second bracket also reduces; dealing first with the double summation over dashed variables we find

$$\sum_{s'}\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1} =\sum_{s}\sum_{t}p_{st}\,(s+2t)-\sum_{s}\sum_{t}p_{st}\left\{su_{\perp}^{s-1}u_{\Delta}^{2t}u_{\perp}+2tu_{\perp}^{s}u_{\Delta}^{2(t-1)}u_{\Delta}^{2}\right\} \tag{A19}$$

before observing that

$$\sum_{s}\sum_{t}p_{st}sx^{s-1}y^{t}=\langle s\rangle G_{1,\perp}(x,y) \tag{A20}$$

$$\sum_{s}\sum_{t}p_{st}tx^{s}y^{t-1}=\langle t\rangle G_{1,\Delta}(x,y) \tag{A21}$$

to arrive at

$$\sum_{s'}\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1}=\langle s\rangle+2\langle t\rangle-\langle s\rangle G_{1,\perp}(u_{\perp},u_{\Delta}^{2})u_{\perp}-2\langle t\rangle G_{1,\Delta}(u_{\perp},u_{\Delta}^{2})u_{\Delta}^{2} \tag{A22}$$

Substituting the self-consistent relationships for $u_\perp$ and u_Δ we finalise the expression as

$$\sum_{s'}\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1}=\langle s\rangle\left(1-u_{\perp}^{2}\right)+2\langle t\rangle\left(1-u_{\Delta}^{3}\right) \tag{A23}$$

In the case that there are no triangles present in the model, then $u_\Delta=1$ and $\langle t\rangle=0$; the expression reduces to

$$\sum_{s'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s'}}\right|_{z_{s'}=1}=\langle s\rangle\left(1-u_{\perp}^{2}\right) \tag{A24}$$

which is the result of [33] in the case that $l=1$. In the opposite case, when there are no ordinary edges, we find

$$\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{t'}}\right|_{z_{t'}=1}=2\langle t\rangle\left(1-u_{\Delta}^{3}\right) \tag{A25}$$

The probability $P(k_{\tau,0},k_{\tau,1})=P((s_0,t_0),(s',t'))$ is given by the quotient of Eqs A11 and A23 where we find

$$\begin{aligned}P((s_0,t_0),(s',t')) =& \left.\frac{d\hat{F}_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1}\Bigg/\sum_{s'}\sum_{t'}\left.\frac{dF_{\mathrm{GCC}}}{dz_{s't'}}\right|_{z_{s't'}=1}\\ =& p_{s_0t_0}\left(s_0q_{\perp,(s',t')}+2t_0q_{\Delta,(s',t')}\right)-p_{s_0t_0}\left(s_0u_{\perp}^{s_0-1}q_{\perp,(s',t')}u_{\perp}^{s'-1}u_{\Delta}^{2t'}u_{\Delta}^{2t_0}\right.\\ &\left.+2t_0u_{\perp}^{s_0}u_{\Delta}^{2(t_0-1)}u_{\Delta}q_{\Delta,(s',t')}u_{\perp}^{s'}u_{\Delta}^{2(t'-1)}\right)\Bigg/\langle s\rangle\left(1-u_{\perp}^{2}\right)+2\langle t\rangle\left(1-u_{\Delta}^{3}\right)\end{aligned} \tag{A26}$$

The conditional probability that a neighbour has joint degree (s',t') given a focal vertex of joint degree (s_0,t_0) is

$$P(s',t'\mid s_0,t_0)=\frac{p_{s_0t_0}s_0q_{\perp,(s',t')}[1-u_{\perp}^{s_0+s'-2}u_{\Delta}^{2(t_0+t')}]+2p_{s_0t_0}t_0q_{\Delta,(s',t')}[1-u_{\perp}^{s_0+s'}u_{\Delta}^{2(t_0+t'-2)+1}]}{p_{s_0t_0}(s_0+2t_0)[1-u_{\perp}^{s_0}u_{\Delta}^{2t_0}]} \tag{A27}$$

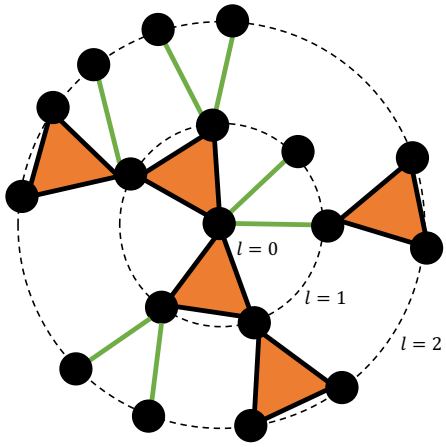

FIG. 9. An example of the degree correlation model in the tree-triangle model; 3-cliques are shaded orange whilst 2-cliques are coloured green. The joint degree of the focal vertex in layer $l = 0$ is $k_{\boldsymbol{\tau},0} = (2, 2)$. We can follow edges of topology $\perp$ or Δ to the first neighbours. The distribution of the joint degrees of vertices in layer $l = 2$ depends on the topology of the path that we choose to reach it. Note, we do not traverse edges between triangles that lead to vertices in the same layer.

Using Eq 46 we find the average joint degree of a neighbour to a (s_0, t_0) vertex as

$$\mathcal{E}[\boldsymbol{k}_{\boldsymbol{\tau},1} \mid \boldsymbol{k}_{\boldsymbol{\tau},0}] = \left(\sum_{s',t'} s' P(s', t' \mid s_0, t_0), \sum_{s',t'} t' P(s', t' \mid s_0, t_0) \right)^T \tag{A28}$$