Exact formula for bond percolation on cliques

Peter Mann[*], V. Anne Smith, John B. O. Mitchell, Christopher Jefferson, 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

We present exact solutions for the size of the giant connected component of complex networks composed of cliques following bond percolation. We use our theoretical result to find the location of the percolation threshold of the model, providing analytical solutions where possible. We expect the results derived here to be useful to a wide variety of applications including graph theory, epidemiology, percolation, and lattice gas models, as well as fragmentation theory. We also examine the Erdős-Gallai theorem as a necessary condition on the graphicality of configuration model networks comprising clique subgraphs.

I. INTRODUCTION

Bond percolation on graphs [1] is a process in which edges are randomly removed or *unoccupied* with some probability, $T - 1$, or *occupied* with probability T. As T is reduced to some critical value, T^*, the graph exhibits a second-order phase transition and fails to be globally connected. The size of the giant connected component (GCC), as well as the location of the critical point, are important quantities within the percolation process. Percolation has not only inherent theoretical interest but is also important for various applications across many disciplines [2–10]. Perhaps the most prominent utilization is the study of diseases spreading through structured populations with transmission probability T. When the infection period of a vertex is a single-valued distribution, the size of the GCC is isomorphic to the outbreak size of the disease; similarly, the critical bond occupation probability is the epidemic threshold. It is well understood how to extract the properties of graphs using the generating function formulation [2,3,11–13]. In its original form, it is assumed that there are no closed loops or cycles among the edges of the graph; it is locally treelike everywhere. When this condition is true, or approximately true, the generating function formulation yields excellent results compared to simulation. However, if the network fails to be locally treelike, then the formulation must be modified to describe correctly the emergent properties of percolation. There has been early work dedicated to finding the correct way to model percolation on graphs comprised of subgraph motifs [14]; however, these works are based on recursive formulas. Newman [15] provided analytical breakthrough in the study of configuration model networks with closed loops using the generating function formulation. The next theoretical milestone is by Miller and Newman in 2009 who independently studied three cliques along with treelike edges [16,17]. Shortly thereafter Karrer and Newman [18] developed a general framework that addressed the study of

larger subgraphs; however, it was determined that a crucial quantity, which we denote by g_τ, could only be determined by an exponentially slow exhaustive enumeration of states. This quantity is the probability that a vertex remains unattached to the GCC despite its involvement in a subgraph of topology τ. Allard *et al.* [19,20] developed a comprehensive and versatile technique based on recursive formulas to determine the percolation properties numerically through iteration. There has also been other recursive methods derived for the same purpose [21]. Within the spirit of these developments, Mann *et al.* developed an analytical approach that approximates the g_τ expression to high accuracy [22,23], affording an equation-based treatment of percolation on arbitrary subgraphs. It remains that the percolation properties can be found exactly, but slowly through Karrer and Newman's method, exactly but recursively though Allard *et al.*'s method, or approximately but analytically though Mann *et al.*'s method.

In this paper, we develop exact, closed-form analytical expressions for the probability that a clique does not lead to the GCC when edges are removed with probability $1 - T$. Cliques are cycles whose vertices are all degree equivalent to one another and connect to all other vertices in the motif. Application of our counting method to inhomogeneous cycles (cycles that contain vertices with different degrees) can readily be performed; however, the final expression depends on the details of the subgraph. We examine other types of homogeneous subgraphs in Appendix B. The method developed herein is most similar to [15–18,24] and the polynomials that our expression yields (shown in Appendix C, Table I) are equivalent to those found by [15] via recursion, when our parameter u in Eq. (27) is set to unity, thus confirming their exactness.

II. BACKGROUND

It is necessary to review both the generating function formulation and the configuration model in order to progress with contents of this paper [2,3].

*pm78@st-andrews.ac.uk

TABLE I. g_N expressions for cliques of six vertices or fewer obtained from Eq. (27).

N clique	g_N equation
3	$(1-T)^2 + 2uT(1-T)^2 + 3(uT)^2(1-T) + u^2T^3$
4	$(1-T)^3 + 3uT(1-T)^4 + 3u^2[T^3(1-T)^3 + 3T^2(1-T)^4]$ $+u^3[T^6 + 6T^5(1-T) + 15T^4(1-T)^2 + 16T^3(1-T)^3]$
5	$(1-T)^4 + 4uT(1-T)^6 + 6u^2[T^3(1-T)^6 + 3T^2(1-T)^7] + 4u^3[T^6(1-T)^4 + 6T^5(1-T)^5$ $+15T^4(1-T)^6 + 16T^3(1-T)^7] + u^4[T^{10} + 10T^9(1-T) + 45T^8(1-T)^2 + 120T^7(1-T)^3$ $+205T^6(1-T)^4 + 222T^5(1-T)^5 + 125T^4(1-T)^6]$
6	$(1-T)^5 + 5uT(1-T)^8 + 10u^2[T^3(1-T)^9 + 3T^2(1-T)^{10}] + 10u^3[T^6(1-T)^8 + 6T^5(1-T)^9$ $+15T^4(1-T)^{10} + 16T^3(1-T)^{11}] + 5u^4[T^{10}(1-T)^5 + 10T^9(1-T)^6 + 45T^8(1-T)^7$ $+120T^7(1-T)^8 + 205T^6(1-T)^9 + 222T^5(1-T)^{10} + 125T^4(1-T)^{11}] + u^5[T^{15} + 15T^{14}(1-T)$ $+105T^{13}(1-T)^2 + 455T^{12}(1-T)^3 + 1365T^{11}(1-T)^4 + 2997T^{10}(1-T)^5 + 4945T^9(1-T)^6$ $+6165T^8(1-T)^7 + 5700T^7(1-T)^8 + 3660T^6(1-T)^9 + 1296T^5(1-T)^{10}]$

A. Generating functions for treelike networks

The generating function framework is based on the degree distribution, $p(k)$, which is the probability of choosing a vertex of degree k from the graph. Two generating functions are introduced whose coefficients are (i) the probability of choosing a degree k vertex at random from the network

$$G_0(z) = \sum_{k=0}^{\infty} p(k)z^k \qquad (1)$$

and (ii) the probability of following a randomly chosen edge to a vertex with excess degree k

$$G_1(z) = \frac{G_0'(z)}{G_0'(1)}, \qquad (2)$$

where $\langle k \rangle = G_0'(1)$ is the average degree of a vertex in the graph.

Defining u as the probability that a neighbor is unattached to the GCC, the probability that a vertex fails to become attached through a single edge is $g_2 = 1 - T + uT$, which is the sum of the probability that the edge was not occupied, $1 - T$, and the probability that it *was* occupied, but the neighbor was unattached to the GCC, uT. The quantity u can be found as the solution to a self-consistent expression [12]

$$u = G_1(g_2). \qquad (3)$$

The expected size of the GCC, $\mathcal{S}$, is then given by $\mathcal{S} = 1 - G_0(g_2)$. The critical point can then be found by perturbing around $u = 1$, which corresponds to $\mathcal{S} = 0$ since $G_0(1) = 1$. From the linear stability analysis, expanding Eq. (3) with a Taylor series we have $u = 1 + uTG_1'(1) + O(u^2)$, from which we find $T^* = 1/G_1'(1)$ [2,12,13] at the fixed point $u = 1$.

B. Configuration model for treelike networks

The configuration model is a method that can be used to create a particular random graph from an ensemble of degree equivalent, uncorrelated random graphs. In the model, the vertices of the graph are assigned an integer, drawn at random from the degree distribution, which indicate its degree. The degree sequence $\{k\} = k_1, k_2, \ldots, k_{\mathcal{N}}$, where $\sum_i k_i = 2E$ for a network of $\mathcal{N} \in \mathbb{N}$ vertices and $E \in \mathbb{N}$ edges, is a sequence of the degrees of the vertices and is typically displayed in descending order such that $k_1 \geqslant k_2 \geqslant \cdots \geqslant k_{\mathcal{N}}$. However, not all degree sequences are valid, or *graphic*, such that some sequences of integers cannot be used to create a graph. The Erdős-Gallai theorem (EGT) states that, in addition to the handshaking lemma (HL), $\sum_i k_i = 2E$, a sequence is graphic if and only if the Erdős-Gallai inequality (EGI)

$$\sum_{i=1}^{n} k_i \leqslant n(n-1) + \sum_{i=n+1}^{\mathcal{N}} \min(k_i, n) \qquad (4)$$

holds for $n \in [1, \mathcal{N} - 1]$. It is trivial to construct degree sequences that satisfy the HL (EGI) but do not satisfy the EGI (HL) and are thus not graphic. For instance, with $\mathcal{N} = 3$ and $\{k\} = \{(1), (1), (1)\}$ the inequality in Eq. (4) is satisfied but the sum of degrees is not even, while $\{k\} = \{(2), (0), (0)\}$ satisfies the lemma but not Eq. (4).

To construct the networks, vertex i is inserted k_i times into a list for all $i \in \mathcal{N}$ which is then shuffled. Pairs of vertices are then drawn at random and connected together. In the limit of large and sparse networks, the probability that the construction process chooses pairs that are either already connected through another edge or belong to the same vertex is vanishingly small. The networks generated according to this process are locally treelike and contain no short-range loops; they are also absent of degree correlations.

C. GCM networks

The treelike models described in Secs. II A and II B were independently extended by Newman and Miller to incorporate triangular clustering [15–17] and subsequently motifs of arbitrary topology [18], which we refer to as the generalized configuration model (GCM).

In the GCM, the degree distribution is replaced by a joint degree distribution that describes a vertex's involvement in motifs such as triangles, squares, four cliques, etc. For instance, a vertex that is involved in $k_{\perp}$ ordinary edges and $k_{\triangle}$ triangles is specified by joint degree $(k_{\perp}, k_{\triangle})$ and the usual degree is recovered from $k = k_{\perp} + 2k_{\triangle}$. Similarly, the joint degree of a vertex that is a part of $k_{\perp}$ ordinary edges, $k_{\triangle}$ triangles, $k_{\square}$ squares, and $k_{\boxtimes}$ four cliques is given by $(k_{\perp}, k_{\triangle}, k_{\square}, k_{\boxtimes})$ and occurs with probability $p(s, k_{\triangle}, k_{\square}, k_{\boxtimes})$; its ordinary degree is recovered from $k = k_{\perp} + 2k_{\triangle} + 2k_{\square} + 3k_{\boxtimes}$, a Diophantine condition [25]. In the GCM it is important to note that the cycles are still independent of one another

(edge disjoint), in much the same way that simple edges are in the original model. This means that the accidental formation of a four clique during triangle construction through the choosing of two vertices that are already involved in a triangle vanishes with large and sparse networks. Thus, upon considering the characteristic size of each motif (i.e., its number of vertices), the GCM regenerates the locally treelike property of the subgraphs. The probability of edge sharing between independent cycles is dependent on the number of vertices and triangles in the cycles for a given number of cycles, however.

The generating function formulation for GCM networks is based on the joint degree distribution and replaces the univariate generating functions from Eqs. (1) and (2) with their multivariate analogs:

$$G_0(\mathbf{z}) = \sum_{k_\perp=0}^{\infty} \cdots \sum_{k_\gamma=0}^{\infty} p(k_\perp, \ldots, k_\gamma) z_\perp^{k_\perp} \cdots z_\gamma^{k_\gamma}, \quad (5)$$

where $\mathbf{z} = \{z_\perp, z_\triangle, z_\square, \ldots, z_\gamma\}$. 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 generalized 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 $q_\tau(\mathbf{k}_\tau) = (k_\tau + 1)p(k_{\tau\setminus\{\tau\}}, k_\tau + 1)/\langle k_\tau \rangle$, where the notation $\mathbf{s}\setminus\{s\}$ excludes element s from set $\mathbf{s}$. Each joint excess degree distribution is generated as

$$G_{1,\tau}(\mathbf{z}) = \sum_{k_\perp}^{\infty} \cdots \sum_{k_\gamma}^{\infty} q(\mathbf{k}_\tau) \, z_\tau^{k_\tau - 1} \prod_{\nu \neq \tau} z_\nu^{k_\nu}. \quad (6)$$

To obtain the fraction of the network occupied by the GCC, S, we must derive the probability $u_\tau \, \forall \tau \in \{\perp, \triangle, \square, \ldots, \gamma\}$ that neighboring vertices within that particular cycle topology fail to connect the focal vertex to the GCC. The probability that the entire τ cycle then fails to connect the focal vertex to the GCC is $g_\tau(u_\tau)$. For each edge topology that emanates from the focal vertex, this is simply given by the probability that each of the neighboring vertices within a single cycle of topology τ themselves fail to connect to the GCC, raised to the power of the number of cycles of topology τ the neighbor vertex belongs to, multiplied by the probability of reaching a neighbor with a specified joint degree and summed over all neighbor joint degree configurations. Together this is given as a self-consistent Dyson-like system of equations for each τ using $G_{1,\tau}$ as

$$u_\tau = G_{1,\tau}(g_\perp, g_\triangle, \ldots, g_\gamma). \quad (7)$$

The fraction of the network occupied by the GCC is now given by $S = 1 - G_0(g_\perp, g_\triangle, \ldots, g_\gamma)$. The critical point can be found by linearizing $u_\tau = G_{1,\tau}(\mathbf{g}_\tau^{\mathbf{m}_\tau})$ in ϵ around $u_\tau = 1 - \epsilon_\tau$. To leading order in ϵ_τ we have $\boldsymbol{\epsilon} = A\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} = [\epsilon_\perp, \epsilon_\triangle, \ldots]^{\mathrm{T}}$. The GCC forms at the point when the determinant $\det(A - I)$ vanishes, where I is the identity matrix, $A = [\partial \mathrm{G}/\partial u_\tau]$, and $\mathrm{G} = [G_{1,\tau}, G_{1,\triangle}, \ldots, G_{1,\gamma}]$.

It is the purpose of this paper to find analytically and exactly an expression for $g_\tau(u_\tau)$ where τ belongs to a restricted subset of all topologies; specifically, we restrict our attention

to clique motifs. It must be noted that, whilst the GCM allows a single vertex to play a simultaneous role in any number of motifs of different topologies, it is often easier, for our purposes, to consider reduced systems in which there is only a single topology present. In such cases, the GCM expressions reduce to univariate functions similar to Eqs. (1) and (2), albeit with topology subscripts, τ. This does not, however, diminish from the generality of our results. Further, we examine the conditions for graphicality of the joint degree sequences of clique networks and write necessary conditions by extending the Erdős-Gallai theorem to GCM networks.

III. GRAPHICALITY OF JOINT DEGREE SEQUENCES FOR GCM NETWORKS

The degree sequence of a configuration model network is a sequence of tuples

$$(k_{\perp,1}, k_{\triangle,1}, \ldots, k_{\gamma,1}), \ldots, (k_{\perp,\mathcal{N}}, k_{\triangle,\mathcal{N}}, \ldots, k_{\gamma,\mathcal{N}}), \quad (8)$$

and, as with ordinary edges, not all sequences lead to the successful creation of networks and we now consider necessary conditions on a joint degree sequence in order that it is graphic. It is natural to separate the degree tuples and order the joint sequence along each topology as $k_{\perp,1} \geqslant k_{\perp,2} \geqslant \cdots \geqslant k_{\perp,\mathcal{N}}$ for the ordinary edges, $k_{\triangle,1} \geqslant k_{\triangle,2} \geqslant \cdots \geqslant k_{\triangle,\mathcal{N}}$ for the triangles (and so on). It is clear that the EGT (the EGI and the HL) must still hold among the overall degrees of the model for the joint degree sequence to be graphic. However, the EGT is no longer sufficient to ensure the graphicality of joint degree sequences according to the extended configuration model. For example, consider an ordered joint degree sequence describing ordinary edges and triangles $\{k_\perp, k_\triangle\} = \{(0, 1), (1, 0), (1, 0)\}$, which is graphic according to the EGI, Eq. (4), and the HL applied to the overall edges, but is not according to the extended configuration model. We require the EGT to hold among the ordinary edges such that $\sum_i s_i = 2\mathcal{H}$ where $\mathcal{H} \in \mathbb{N}$ is the number of ordinary edges and that

$$\sum_{i=1}^{n} k_{\perp,i} \leqslant n(n - 1) + \sum_{i=n+1}^{N} \min(k_{\perp,i}, n) \quad (9)$$

holds for $n \in [1, \mathcal{N} - 1]$. For the triangle degree sequence to be graphical, we require that the sum of the number of triangle edges is divisible by 3,

$$2 \sum_{i=1}^{\mathcal{N}} k_{\triangle,i} = 3\mathcal{T}, \quad (10)$$

which is a modified handshaking lemma, as well as a modified inequality

$$2 \sum_{i=1}^{n} k_{\triangle,i} \leqslant n(n - 1) + \sum_{i=n+1}^{N} \min(2k_{\triangle,i}, n), \quad (11)$$

which must hold for $n \in [1, \mathcal{N} - 1]$. The factor of 2 in Eq. (11) is due to each vertex consuming two edges per triangle. Together these conditions extend the Erdős-Gallai theorem to the tree-triangle model, ensuring that the joint degree sequence is graphic. This can now readily be extended to other GCM networks. The necessary conditions for the
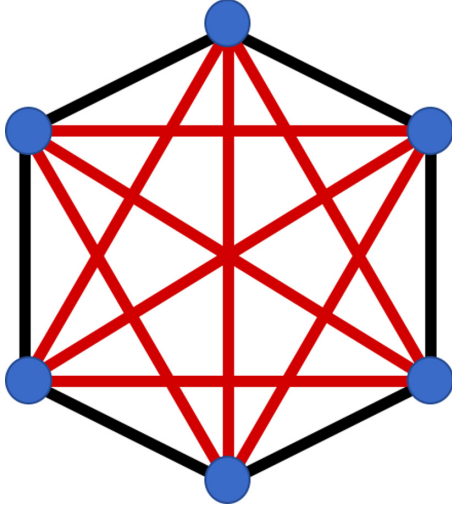
FIG. 1. Six clique has six vertices (blue), six exterior edges (black), and $6(6 - 1 - 2)/2 = 9$ interior edges (red). There are $6(6 - 1)/2 = 15$ edges in total.
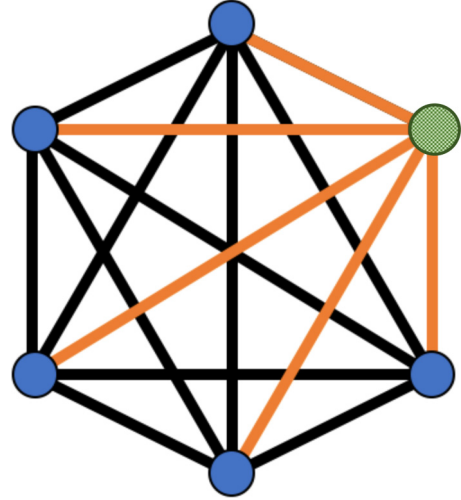

FIG. 2. (6-1)-semicomplete clique has one colored vertex (green) and $6 - 1 = 5$ ordinary vertices. The $(6 - 1) = 5$ edges that emanate from the colored vertex have also been colored (orange).

graphicality of joint degree sequences of configuration models comprising cliques can now be written by exploiting the characteristic size of each clique. While is is easy to convince ourselves that these conditions are necessary conditions for graphicality, we do not, however, know if these are sufficient conditions.

IV. CLIQUES

In this section, we derive an exact expression for the probability $g_{\tau,\text{clique}}$ (which for brevity we denote simply by g_N henceforth) that a vertex fails to become attached to the GCC when it is a constituent of a clique of size N. Cliques have been studied previously using alternative methods; however, these approaches use numerical recursion to obtain a solution [15,20]. We frame our method in layers around integer powers of u_τ (which we also simplify for notation brevity to u) in the range $[0, N - 1]$. We categorize the edges of the clique as either *exterior* or *interior* edges, depending on whether they belong to the outer skeleton of the cycle or connect vertices across the interior, through the shape, respectively; see Fig. 1. We define an *intact* graph to indicate that all edges in the cycle are intact, while a *connected* graph is one in which there exists at least one pathway between all vertices. The removal of a single edge from a clique with $N > 2$ will ruin the intact property, but it will still be fully connected.

In the following we reserve j for an index over the number of edges we have removed from the clique and we reserve r for an index over the number of vertices n we have removed from the subgraph. We define another term, a $(N - n)$-*semi-intact* graph, to be the intact clique of codimension-(n) embedded in the clique of size N with n vertices, and their edges, colored. In other words, an $(N - 1)$-semi-intact clique is a clique of size N with one vertex colored and all edges that connect to the colored vertex also colored; see Fig. 2. A $(N - 2)$-semi-intact clique is a clique of size N with two colored vertices, whose edges to all other vertices (and between the colored vertices themselves) are also marked.

With these definitions in place, let us begin the derivation of the probability that a vertex involved in an N-clique fails to connect to the GCC through its involvement in the N clique. The first and arguably the easiest case is the fully connected graph of size N. With all of its edges intact we pick a focal vertex and set the remaining $(N - 1)$ vertices to the u state. Since the graph is connected, if a neighboring vertex was *not* in the u state, by definition the focal vertex would be attached to the GCC. The fully connected, intact clique of size N occurs with probability

$$P(N, 0) = u^{N-1} T^N T^{N(N-1-2)/2}, \qquad (12)$$

where the notation $P(N, n)$ denotes the total probability of obtaining a connected graph of size N with n vertices belonging to the GCC. Examining these terms, we note that all vertices other than the focal vertex must not be in the GCC if all of their edges are occupied. There are N exterior edges and $N(N - 1 - 2)/2$ interior edges. There is only one way to pick this shape, so its *multiplicity* (the number of different ways the configuration can occur) is unity.

As remarked above, for $N > 2$, we can remove edges from this cycle and it will still be fully connected, although no longer intact. It happens that we can remove all of the interior edges, and even one of the exterior edges and still make connected graphs of size N. If we set one of the interior edges as unoccupied, we have

$$\begin{aligned} P(1 \mid N, 0) = {}& q_{N,N(N-1)/2-1} u^{N-1} T^N \\ & \times T^{N(N-1-2)/2-1}(1 - T), \qquad (13) \end{aligned}$$

where $q_{m,k}$ is the number of connected graphs that can be formed over m labeled vertices with k edges (see Appendix A). The notation $P(j \mid N, n)$ indicates the total probability that the focal vertex remains unattached to the GCC despite its involvement in a clique of size N that has n vertices attached to the GCC (whose edges are all unoccupied) and a further j unoccupied edges among the unattached vertices.

If we remove a second edge we have

$$P(2 \mid N, 0) = q_{N,N(N-1)/2-2} u^{N-1} T^N$$
$$\times T^{N(N-1-2)/2-2} (1-T)^2. \quad (14)$$

The total probability of removal of $j \in [0, \mathcal{E}(N)]$ edges, where $\mathcal{E}(N) = N(N-1-2)/2 + 1$, is given by

$$P(N, 0) = \sum_{j=0}^{\mathcal{E}(N)} P(j \mid N, 0), \quad (15)$$

which is

$$P(N, 0) = \sum_{j=0}^{\mathcal{E}(N)} q_{N,N(N-1)/2-j} u^{N-1} T^N$$
$$\times T^{N(N-1-2)/2-j} (1-T)^j. \quad (16)$$

If we were to remove another edge from the graph, we would isolate a vertex. We have therefore enumerated all combinations of graphs that exist on an N clique with $N-1$ vertices that are all unattached to the GCC.

We must now examine the case in which a single vertex within the clique is connected to the GCC, thus decreasing the largest power of u by one. There are $(N-1)$ vertices that we could remove and all edges that point to the removed vertex must now be set to $(1-T)$, of which there are $(N-1)$. Therefore, the $(N-1)$-semi-intact graph or codimension-1 subgraph in the N-clique occurs with probability

$$P(0 \mid N, 1) = (N-1) u^{N-2} T^{N-2}$$
$$\times T^{(N-1)(N-1-1-2)/2} (1-T)^{N-1}, \quad (17)$$

where the number of interior edges among the nonremoved vertices is now $(N-1)(N-1-1-2)/2$. We can imagine this as a clique of size $(N-1)$ embedded within the N clique and the remaining edges are set to $(1-T)$. We recall the $(6-1)$-semi-intact graph from Fig. 2; the removed vertex is green and the $(1-T)$ edges are orange. The leading factor of $(N-1)$ in Eq. (17) accounts for the choices of vertex we could remove other than the focal vertex.

As with the intact case, we can remove edges from this graph and still retain connectivity among the $(N-1)$ non-removed vertices. Removal of a single edge occurs with probability

$$P(1 \mid N, 1) = (N-1) q_{N-1, X_{N-1,1}} u^{N-2} T^{N-2}$$
$$\times T^{(N-1)(N-1-1-2)/2-1} (1-T)^{N-1+1}, \quad (18)$$

where $X_{N-r,j}$ is the number of edges in the $(N-r)$ clique minus j:

$$X_{N-r,j} = (N-r)(N-r-1)/2 - j. \quad (19)$$

Let us remove a second edge from this cycle to obtain

$$P(2 \mid N, 1) = (N-1) q_{N-1, X_{N-1,2}} u^{N-2} T^{N-2}$$
$$\times T^{(N-1)(N-4)/2-2} (1-T)^{N-1+2}. \quad (20)$$

The removal of $j \in [0, \mathcal{E}(N-1)]$ edges now proceeds as

$$P(N, 1) = (N-1) \sum_{j=1}^{\mathcal{E}(N-1)} q_{N-1, X_{N-1,j}} u^{N-2} T^{N-2}$$
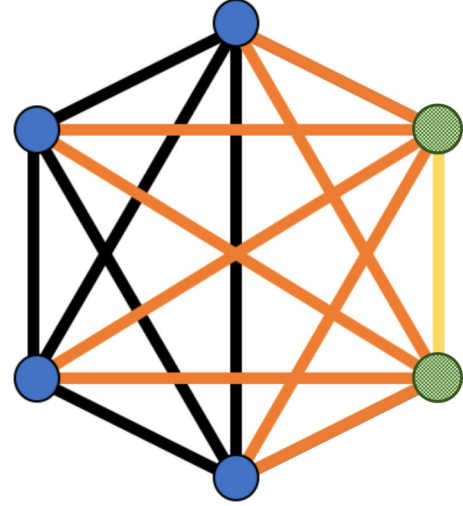

FIG. 3. (6-2)-semi-intact clique has two colored vertices (green) and $6 - 2 = 5$ ordinary vertices (blue). The $(N-1) + (N-2) = 9$ edges that emanate from colored vertices have also been colored (orange). Notice that the edge that connects the two colored vertices (yellow) has been colored differently than the other edges. Interface edges connect blue vertices to green vertices. There are 9-1 interface edges in this example.

$$\times T^{(N-1)(N-4)/2-j} (1-T)^{N-1+j}. \quad (21)$$

To be clear, this expression is the total probability that a focal vertex in an N clique with one vertex attached to the GCC and up to $j = (N-1)(N-4)/2 + 1$ edges set unoccupied fails to be attached to the GCC.

Further removal of an edge would isolate a vertex and, hence, we claim that this level is now completed as we have counted all ways that this state can occur. Although the pattern is largely the same as above, there is a complexity with the removal of a second vertex. Currently, we absorb all of the removed vertex's edges into the $(1-T)$ state, since an occupied edge would by definition connect the focal vertex to the GCC. However, when a second vertex is removed, there is a connection between the two removed vertices that need not be $(1-T)$. Therefore, we must subtract from this power those connections between removed vertices. This is simply the number of edges in a clique of size equal to the number of removed vertices, n. We introduce the term *interface edges* to be edges that connect removed vertices to nonremoved vertices; see Fig. 3.

The number of interface edges is given by the total number of edges that the removed vertices have minus the number of edges that connect removed vertices to each other. If there are $n < N$ removed vertices, then there are a total of

$$\sum_{i=1}^{n} (N-i)$$

colored edges (orange plus yellow in Fig. 3), of which a total of

$$\frac{n(n-1)}{2}$$

will point to other removed vertices (yellow in Fig. 3). Hence the total number, $\omega(r)$, of interface edges (orange) is

$$\omega(r) = \sum_{i=1}^{r}(N-i) - \frac{r(r-1)}{2}. \tag{22}$$

For subsequent vertex removals, the number of edges that are required to be $(1-T)$ is given by the number of interface edges. In the case that $n=1$ we find that the number of interface edges is $N-1$, in agreement with the previous workings.

We will now remove a second vertex from the N clique and we begin by describing the $(N-2)$-semi-intact graph, from which we will then remove edges in a sequential and now hopefully familiar fashion. There are $(N-1)$ ways to remove the first vertex followed by $\binom{N}{2}$ ways to remove the second vertex, so the binomial coefficient will lead the expression. The chain of vertices not in the GCC now occurs with probability u^{N-3} comprising N less two removed vertices and one focal vertex. The outer T skeleton of the $(N-2)$-semi-intactd graph has probability T^{N-3}. The number of interior edges among present vertices is then $(N-2)(N-2-1-2)/2$. The number of interface edges is

$$\sum_{i=1}^{2}(N-i) - \frac{2(2-1)}{2} = 2(N-2).$$

All together, the expression for a clique of size N with two removed vertices (a semi-intact graph of codimension 2) is given by

$$P(0 \mid N, 2) = \binom{N-1}{2} u^{N-3} T^{N-3} T^{(N-2)(N-2-1-2)/2}$$
$$\times (1-T)^{2(N-2)}. \tag{23}$$

We can then remove all of the interior edges among the non-removed vertices, as well as a single exterior edge, and place them into the $(1-T)$ state. Removing one edge we have

$$P(1 \mid N, 2) = \binom{N-1}{2} q_{N-2, X_{N-2,1}} u^{N-3} T^{N-3}$$
$$\times T^{(N-2)(N-2-1-2)/2-1}(1-T)^{2(N-2)+1}. \tag{24}$$

Hence the removal of $j \in [0, \mathcal{E}(N-2)]$ such edges yields the total probability that a vertex remains unattached to the GCC despite its involvement in an N clique that has two vertices attached to the GCC:

$$P(N, 2) = \binom{N-1}{2} \sum_{j=1}^{\mathcal{E}(N-2)} q_{N-2, X_{N-2,j}} u^{N-3} T^{N-3}$$
$$\times T^{(N-2)(N-2-1-2)/2-j}(1-T)^{2(N-2)+j}. \tag{25}$$

Subsequent loss of edges will isolate a further vertex and, hence, this layer is completely enumerated.

We have now encountered all the sufficient logic that we require for the correct abstraction of the formula to account for arbitrary numbers of removed vertices and edges from a clique of size N. We note that the final expression will simply be the summation of all of the total probabilities that a focal vertex fails to be attached to the GCC given its involvement in

the N clique, which is

$$g_N = \sum_{r=0}^{N-1} \sum_{j=0}^{\mathcal{E}(N-r)} P(j \mid N, r). \tag{26}$$

For a clique of size N, let there be n removed vertices. There are $\binom{N-1}{r}$ ways to remove the $r \leqslant n$ vertices sequentially. The power of u is given by $(N-r-1)$; this is the power of the exterior T also; the interior power of T is given by $(N-r)$. The final expression therefore is given by

$$g_N = \sum_{r=0}^{N-1} \binom{N-1}{r} \sum_{j=0}^{\mathcal{E}(N-r)} q_{N-r, X_{N-r,j}} (uT)^{N-r-1}$$
$$\times T^{\mathcal{E}(N-r)-1-j}(1-T)^{\omega(r)}. \tag{27}$$

This equation is the main result of this paper. We now have a closed form expression for each g_N required to compute Eq. (7) and thus can find the size of the GCC for GCM networks comprised of clique subgraphs. Upon comparison of the polynomials generated by Eq. (27) to those given by Newman in [15], we find exact agreement when $u=1$, which we show in Appendix C. Additionally, we display the results of our formula for cliques with six vertices or fewer in Appendix C, Table I.

Percolation threshold

We now turn our attention to the location of the critical point for the formation of a GCC among networks comprised entirely of N cliques during bond percolation. From the Taylor series of Eq. (3) in Sec. I, we understand that, in order to obtain the percolation properties of the network, we have to evaluate the derivative of g_N with respect to u. This derivative is found to be

$$\frac{\partial g_N}{\partial u} = \sum_{r=0}^{N-1} \binom{N-1}{r}(N-r-1) \sum_{j=0}^{\mathcal{E}(N-r)} q_{N-r, X_{N-r,j}}$$
$$\times (uT)^{N-r-2} T^{\mathcal{E}(N-r)-j}(1-T)^{\omega(r)}. \tag{28}$$

The percolation threshold is then obtained by evaluating the derivative at $u=1$, and following a similar analysis to the treelike topology we obtain

$$\left.\frac{\partial g_N}{\partial u}\right|_{u=1} \frac{\langle k^2 - k \rangle}{\langle k \rangle} = 1, \tag{29}$$

where $\langle k \rangle$ is the mean number of cliques a vertex belongs to. For example, the derivative for three cliques is found to be

$$\frac{\partial g_3}{\partial u} = 2T(1-T)^2 + 6uT^2(1-T) + 2uT^3. \tag{30}$$

Evaluated at $u=1$ and inserted into Eq. (29) we have

$$2(T^2 + T - T^3)\frac{\langle k_\Delta^2 - k_\Delta \rangle}{\langle k_\Delta \rangle} - 1 = 0. \tag{31}$$

For networks in which a vertex's membership in a given clique size is Poisson distributed, we can reduce $\langle k_\Delta^2 - k_\Delta \rangle/\langle k_\Delta \rangle$ to simply $\langle k_\Delta \rangle$. Further, factorizing the prefactor in T we have $2T(1 + T - T^2)\langle k_\Delta \rangle - 1 = 0$. Using Gauss's lemma, this cubic expression is reducible in T into the quadratic form whose

roots yield the critical transmissibilities of the model and, hence, the critical point occurs at

$$T_{\text{Poisson}}^* = -1 + 2\sqrt{1 + \frac{1}{\langle k_\Delta \rangle}}. \tag{32}$$

We repeat the calculation for the four clique to obtain the following polynomial:

$$\left.\frac{\partial g_4}{\partial u}\right|_{u=1} = 3T(-2T^5 + 7T^4 - 7T^3 + 2T + 1). \tag{33}$$

The Galois group of the quintic part is the symmetric group, S_5, and hence a root cannot be found. It is unlikely that percolation properties of larger cliques can be resolved analytically due to the Abel-Ruffini theorem.

V. CONCLUSION

In this paper we have derived an exact equation that describes the probability that a vertex fails to be connected to the percolating cluster during bond percolation despite its role in one or more clique subgraphs. In addition, we show how to calculate the critical bond occupancy probability for these graphs and unpacked the formula for cliques with six vertices or fewer in Table I. We compared our formula to other results in the literature, where they existed, and found our formula in agreement to arbitrary precision, thereby concluding its exactness. We also showed that the conditions for the existence of a GCC within these models can be found following a straightforward recipe; however, the polynomials obtained are unlikely to admit a solvable root in all but the simplest of cases. We show this condition for three and four cliques, the latter yielding a quintic polynomial with Galois group S_5. For $N > 6$, there are classes of homogeneous graphs between the weak cycle and the clique. It remains to be shown that an encompassing formula can be written to collect these classes into a single expression. We describe some of the challenges faced in pursuit of this in Appendix B.

A crucial question is how to apply this exact formulation to the study of inhomogeneous cycles whose vertices are no longer degree equivalent. In this case the expressions for g_τ are no longer univariate functions, but, instead, have as many u_v variables as there are nonisomorphic sites in the motif. Currently, the only known approach to generate these expressions exactly is an exponentially slow exhaustive enumeration [18]. This application is nontrivial and would be an extremely important addition to the model. We hope that this result will ignite new areas of research within the study of random graphs and their applications and provide meaningful contributions to a wide range of topics.

APPENDIX A: $q_{n,k}$

The number of connected graphs of n labeled vertices over k edges is given by $q_{n,k}$. This quantity has a well known recursion formula as well as a closed-form analytical solution [15,24,26,27]. Given the importance of this quantity to the contents of this paper, we will review this derivation now.

Let $\mathcal{Q}$ be the combinatorial class of connected graphs and $\mathcal{G}$ the combinatorial class of all labeled graphs [28]. The relation between these two classes is the set of relation: a graph is a set of connected components. This indicates that the mixed exponential generating function $G(z, y)$ of $\mathcal{G}$ can be generated from $Q(z, y)$ according to the following relationship:

$$G(z, y) = \exp Q(z, y). \tag{A1}$$

For m vertices, there are a total of $\binom{m}{2} = m(m-1)/2$ potential edges not allowing self-loops or multiedges between the vertices. This set has $2^{\binom{m}{2}}$ possible partitions. Therefore, counting vertices (with z) and edges (with y) we have that

$$G(z, y) = \sum_m \sum_l \frac{1}{m!}\binom{\binom{m}{2}}{l} z^m y^l. \tag{A2}$$

From the binomial theorem we find

$$G(z, y) = \sum_m \frac{z^m}{m!}(1 + y)^{\binom{m}{2}} \tag{A3}$$

or

$$G(z, y) = 1 + \sum_{m \geqslant 1}(1 + y)^{m(m-1)/2}\frac{z^m}{m!}. \tag{A4}$$

This yields an expression for the entire series of connected labeled graphs, $Q(z, y)$, since $Q(z, y) = \ln G(z, y)$ such that we obtain

$$Q(z, y) = \ln\left(1 + \sum_{m \geqslant 1}(1 + y)^{m(m-1)/2}\frac{z^m}{m!}\right). \tag{A5}$$

We can then perform a series expansion of the logarithm using

$$\ln(1 + x) = \sum_{k=0}^{\infty}\frac{(-1)^{k+1}}{k}x^k$$

to obtain

$$Q(z, y) = \sum_{l \geqslant 1}(-1)^{l+1}\frac{1}{l}\left(\sum_{m \geqslant 1}(1 + y)^{m(m-1)/2}\frac{z^m}{m!}\right)^l. \tag{A6}$$

We now examine the case of n vertices and k edges where $k \geqslant n - 1$ by extracting the coefficient $q_{n,k}$ of $[z^n][y^k]$. Note that the term in the parentheses has minimum degree l in z, allowing us to disregard the series beyond $l > n$. This yields the formula for the number of connected labeled graphs with n vertices and k edges as

$$q_{n,k} = n![z^n][y^k]\sum_{l=1}^{n}(-1)^{l+1}\frac{1}{l}$$

$$\times \left(\sum_{m=1}^{n}(1 + y)^{m(m-1)/2}\frac{z^m}{m!}\right)^l. \tag{A7}$$
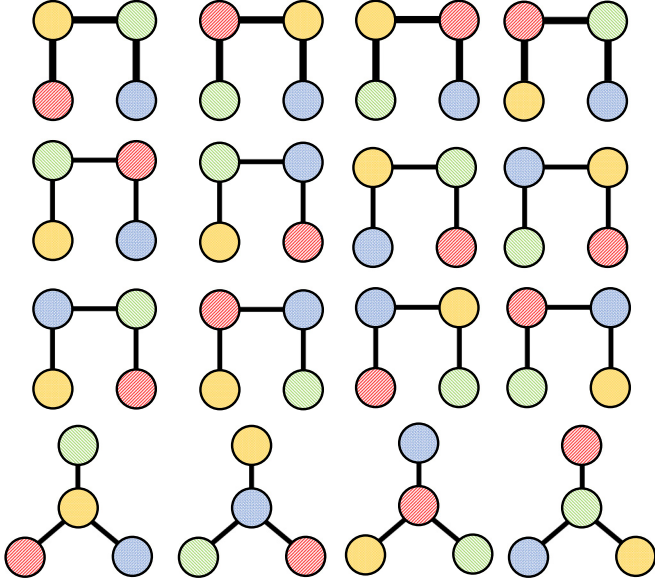
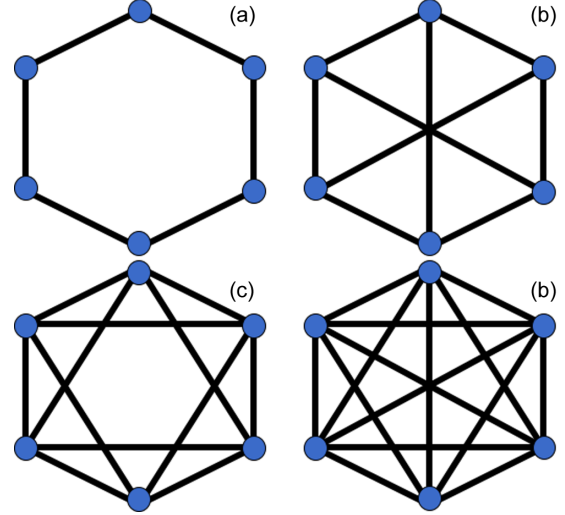FIG. 4. 16 graphs that can be made among four labeled vertices with three edges is given by $q_{4,3}$.

FIG. 5. Four homogeneous nonisomorphic graphs of size $N = 6$. The simple cycle (a) has no interior edges; (b) and (c) show two graphs with an increasing number of interior edges until the clique is obtained (d).

The coefficient of z^n is given by the integer partitions $\lambda \vdash n$ of length l, multiplied by their multiplicity (number of compositions)

$$\frac{1}{n!}\binom{n}{\lambda}\binom{l}{f}, \tag{A8}$$

where for partition λ we have $\lambda = 1^{f_1} 2^{f_2} 3^{f_3} \cdots$ and so on, such that we have

$$q_{n,k} = \sum_{\lambda \vdash n} \frac{(-1)^{l+1}}{l}\binom{n}{\lambda}\binom{l}{f}(1+y)^{\sum_{\lambda_i}\binom{\lambda_i}{2}} \tag{A9}$$

for $\lambda_i \in \lambda$. The coefficient of y^k is found from the binomial theorem to yield a final expression for $q_{n,k}$ as

$$q_{n,k} = \sum_{\lambda \vdash n} \frac{(-1)^{l+1}}{l}\binom{n}{\lambda}\binom{l}{f}\binom{\sum_{\lambda_i}\lambda_i(\lambda_i-1)/2}{k}. \tag{A10}$$

As an example of $q_{n,k}$ in Eq. (27), we examine the coefficients of the four clique when there are no removed vertices, that is, when $n = 0$. From Table I, we observe the leading coefficients of the terms in u^3 are $q_{4,k} = 1$, 6, 15, and 16, which correspond to the number of graphs that can be made with $k = 6, 5, 4, 3$ edges, respectively. The set of graphs that can be made from $q_{4,3}$ is presented in Fig. 4.

APPENDIX B: OTHER HOMOGENEOUS CYCLES

In this Appendix we explore other homogeneous cycles that are not cliques but whose vertices are all degree equivalent. We demonstrate that the percolation expression for these cycles is not readily obtained in closed form, unlike the expressions for simple cycles and cliques. Any formula that can be obtained for a class of cycles is distinct from the expressions for other, seemingly related, cycles. For this purpose, consider a cycle comprised of $N \geqslant 6 \in 2\mathbb{N}$ degree equivalent vertices in which each vertex has degree 3; see cycle B in

Fig. 5. (Note the cycle cannot be formed for odd N.) Application of the enumeration scheme developed in the text to obtain the probability that a particular focal vertex does not become attached to the GCC through its role in this cycle proceeds as follows. The cycle can lose up to $N/2 + 1$ edges before a vertex is isolated; the probability that $j \in [0, N/2 + 1]$ is removed is given by

$$P(j \mid N, 0) = q_{N,N+N/2-j}u^{N-1}T^N T^{N/2-j}(1-T)^j \tag{B1}$$

and hence the total probability $P(N, 0)$ that we can still retain a connected graph despite the removal of edges is

$$P(N, 0) = \sum_{j=0}^{N/2+1} P(j \mid N, 0). \tag{B2}$$

With any further edge removal, a vertex is pruned from the cycle. The resulting cycle has $N - 3$ deg(3) sites and 3 deg(2) sites remaining.

There are now $N/2 - 1$ interior edges and $N - 2$ exterior edges remaining from the original set of edges. It happens that we can remove all of the remaining interior edges and proceed without vertex isolation; however, we cannot remove any of the exterior edges. Thus the total probability $P(N, 1)$ that describes the cycle with one vertex removed is given by

$$P(N, 1) = (N-1) \sum_{j=0}^{N/2-1} q_{N-1,N-2+N/2-1-j}u^{N-2}$$
$$\times T^{N-2}T^{N/2-1-j}(1-T)^{j+3}. \tag{B3}$$

At this point, the cycle now contains mixed degree vertices. We must distinguish upon whether the vertex we now remove has degree 2 or degree 3 as removing either vertex will lead to different probabilities for successive counting. Further, considering $N = 10$, supposing we had removed a degree 3 vertex, it matters whether the neighbors of that vertex are degree 2 or 3 as, in both cases, the resulting probabilities for

further counting are nonequivalent. And, while it is theoretically possible to enumerate the combinations into a single expression, it seems unlikely that such a formula would be readily derived, or that it would be transferable to other related cycles. Similar complexities arise for the other homogeneous cycle in Fig. 5 and are the basis of the complexity in enumerating percolation formulas for inhomogeneous cycles.

APPENDIX C: DISPLAYED CLIQUE FORMULAS

The expressions for cliques of six vertices or fewer are shown in Table I. It is clear upon comparison that the $q_{n,k}u^x T^y$ structure of the polynomials appears to repeat across the orders of increasing clique size. For instance, the polynomial of u^4 can be compared between the $N = 5$ and $N = 6$ equations; however, in each case, the exponent of the interface edges, $(1 - T)^z$, varies.

In the Introduction we stated that Newman had also found these polynomials previously for clique networks. To compare our polynomials to those derived by Newman [15], we first set $u = 1$ in Eq. (27). Newman's method depends on the probability, $P(k \mid N)$, that a particular vertex belongs to a connected cluster of k vertices in an N clique, including itself. This is given by Eq. (7) in [15] as

$$P(k \mid N) = \binom{N - 1}{k - 1}(1 - T)^{k(N-k)}P(k \mid k), \qquad \text{(C1)}$$

where we have relabeled Newman's $p \to T$ and $q \to 1 - T$ to be in keeping with our notation. These conditional probabilities are evaluated via recursion from an initial condition of $P(1 \mid 1)$ and

$$P(k \mid k) = 1 - \sum_{l=0}^{k-1} P(l \mid k). \qquad \text{(C2)}$$

For the purpose of comparison to our closed form expression, we have the following equality:

$$g_N|_{u_N=1} = \sum_{k=1}^{N} P(k \mid N). \qquad \text{(C3)}$$

While it is trivial to confirm these formulas are equivalent to Eq. (27) for small clique sizes, we will explicitly show the agreement of our polynomials with Newman's for $N = 6$. Unpacking Newman's expressions we find

$$\begin{aligned} g_6 = {}& P(1 \mid 6) + P(2 \mid 6) + P(3 \mid 6) + P(4 \mid 6) \\ &+ P(5 \mid 6) + P(6 \mid 6), \end{aligned} \qquad \text{(C4)}$$

which evaluates to

$$\begin{aligned} g_6 = {}& \binom{6 - 1}{1 - 1}q^{1(6-1)}P(1 \mid 1) + \binom{6 - 1}{2 - 1}q^{2(6-2)}P(2 \mid 2) \\ &+ \binom{6 - 1}{3 - 1}q^{3(6-3)}P(3 \mid 3) \\ &+ \binom{6 - 1}{4 - 1}q^{4(6-4)}P(4 \mid 4) + \binom{6 - 1}{5 - 1}q^{5(6-5)}P(5 \mid 5) \\ &+ \binom{6 - 1}{6 - 1}q^{6(6-6)}P(6 \mid 6). \end{aligned} \qquad \text{(C5)}$$

Simply inserting the values of $P(k \mid k)$ from Table 1 in [15] and comparing this to the result for $N = 6$ from Eq. (27) in Table I confirms the exactness of our closed form expression.

[1] S. R. Broadbent and J. M. Hammersley, Percolation processes, Math. Proc. Cambridge Philos. Soc. **53**, 629 (1957).

[2] M. E. Newman, *Networks* (Oxford University Press, Oxford, 2019).

[3] R. Cohen and S. Havlin, *Complex Networks: Structure, Robustness and Function* (Cambridge University Press, Cambridge, UK, 2010).

[4] P. Désesquelles, Exact solution of finite size mean field percolation and application to nuclear fragmentation, Phys. Lett. B **698**, 284 (2011).

[5] A. Dávila, C. Escudero, J. López, and C. Dorso, Geometrical aspects of isoscaling, Physica A **374**, 663 (2007).

[6] K. Paech, W. Bauer, and S. Pratt, Zipf's law in nuclear multifragmentation and percolation theory, Phys. Rev. C **76**, 054603 (2007).

[7] W. Trautmann, Multifragmentation and the liquid-gas phase transition: an experimental overview, Nucl. Phys. A **752**, 407 (2005), Proceedings of the 22nd International Nuclear Physics Conference (Part 2).

[8] Y. I. Kuz'Min and I. V. Pleshakov, Percolation transitions on an electrified surface, Tech. Phys. Lett. **45**, 1167 (2019).

[9] R. Cimino, K. A. Cychosz, M. Thommes, and A. V. Neimark, Experimental and theoretical studies of scanning adsorption-desorption isotherms, Colloids Surf., A **437**, 76 (2013).

[10] D. Wilkinson, Percolation model of immiscible displacement in the presence of buoyancy forces, Phys. Rev. A **30**, 520 (1984).

[11] M. E. J. Newman, S. H. Strogatz, and D. J. Watts, Random graphs with arbitrary degree distributions and their applications, Phys. Rev. E **64**, 026118 (2001).

[12] M. Newman, Spread of epidemic disease on networks, Phys. Rev. E **66**, 016128 (2002).

[13] M. Molloy and B. Reed, A critical point for random graphs with a given degree sequence, Random Struct. Algorithms **6**, 161 (1995).

[14] E. N. Gilbert, Random graphs, Ann. Math. Stat. **30**, 1141 (1959).

[15] M. E. J. Newman, Properties of highly clustered networks, Phys. Rev. E **68**, 026121 (2003).

[16] J. C. Miller, Percolation and epidemics in random clustered networks, Phys. Rev. E **80**, 020901(R) (2009).

[17] M. E. J. Newman, Random Graphs with Clustering, Phys. Rev. Lett. **103**, 058701 (2009).

[18] B. Karrer and M. E. J. Newman, Random graphs containing arbitrary distributions of subgraphs, Phys. Rev. E **82**, 066118 (2010).

[19] A. Allard, L. Hébert-Dufresne, P.-A. Noël, V. Marceau, and L. J. Dubé, Exact solution of bond percolation on small arbitrary graphs, EPL (Europhys. Lett.) **98**, 16001 (2012).

[20] A. Allard, L. Hébert-Dufresne, J.-G. Young, and L. J. Dubé, General and exact approach to percolation on random graphs, Phys. Rev. E **92**, 062807 (2015).

[21] E. Coupechoux and M. Lelarge, How clustering affects epidemics in random networks, Adv. Appl. Probab. **46**, 985 (2014).

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

[23] P. Mann, V. A. Smith, J. B. O. Mitchell, and S. Dobson, Random graphs with arbitrary clustering and their applications, Phys. Rev. E **103**, 012309 (2021).

[24] E. N. Gilbert, Enumeration of labelled graphs, Can. J. Math. **8**, 405 (1956).

[25] 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, J. Complex Networks **5**, 1 (2017).

[26] H. S. Wilf, *Generatingfunctionology* (Academic Press, New York, 1994).

[27] R. J. Riddell and G. E. Uhlenbeck, On the theory of the virial development of the equation of state of monoatomic gases, J. Chem. Phys. **21**, 2056 (1953).

[28] P. Flajolet and R. Sedgewick, *Analytic Combinatorics* (Cambridge University Press, Cambridge, UK, 2013).