BOLYAI SOCIETY
MATHEMATICAL STUDIES, 18

Handbook of Large-Scale Random Networks
pp. 369–408.

Chapter 9

k-clique Percolation and Clustering

GERGELY PALLA, DÁNIEL ÁBEL, ILLÉS J. FARKAS,
PÉTER POLLNER, IMRE DERÉNYI and TAMÁS VICSEK*

We summarise recent results connected to the concept of k-clique percolation. This approach can be considered as a generalisation of edge percolation with a great potential as a community finding method in real-world graphs. We present a detailed study of the critical point for the appearance of a giant k-clique percolation cluster in the Erdős–Rényi-graph. The observed transition is continuous and at the transition point the scaling of the giant component with the number of vertices is highly non-trivial. The concept is extended to weighted and directed graphs as well. Finally, we demonstrate the effectiveness of k-clique percolation as a community finding method via a series of real-world applications.

1. Introduction

In recent years there has been a growing interest in the dense, highly interconnected parts of real-world graphs, often referred to as *communities, modules, clusters of cohesive groups* [74, 76, 26, 46, 32, 54, 73, 64]. These structural subunits can correspond to multi-protein functional units in molecular biology [70, 78], a set of tightly coupled stocks or industrial sectors in economy [59, 41], groups of people [74, 85, 63], cooperative players [80, 84, 79], etc. The location of such building blocks can be crucial to the

*Corresponding author.

understanding of the structural and functional properties of the systems under investigation.

Cliques (maximal complete subgraphs, in which every vertex (also referred to as node) is linked to every other vertex) correspond to the most dense parts of a network [14, 25, 12, 11], therefore, they serve as an ideal starting point to search for communities. However, limiting the community definition to cliques only would be too restrictive in most cases. *k-clique percolation* offers a similar, but more flexible alternative for network clustering [21, 64, 1]. This approach has been proven successful in many applications ranging from the study of cancer-related proteins in protein interaction networks [44, 45], through the analysis of stock correlations [41] to the examination of various social networks [34, 63]. This success has inspired the extension of the approach to *weighted* and *directed* networks as well [27, 67]. In this chapter we overview the recent results connected to this approach. At this point we note that *k-cores* are also closely related to the concepts described above and turned out to be of fundamental importance in the decomposition of large complex networks as well [13, 75, 22, 33, 23, 17]. In this approach a graph can be treated as a set of successively enclosed k-cores, similar to a Russian nested doll (getting denser and denser inside).

As k-clique percolation can be considered to be a generalisation of edge percolation, it provides a set of very interesting problems in random graph theory by itself. (The first rigorous mathematical results on this subject were given recently by Bollobás and Riordan in [15]). One of the most conspicuous early results in random graph theory was related to the (edge) percolation transition of the Erdős–Rényi (E-R) uncorrelated random graph [24, 14]. The various aspects of this classical model remain still of great interest since such a graph can serve both as a test-bed for checking all sorts of new ideas concerning complex networks in general, and as a prototype of random graphs to which all other random graphs can be compared. The mentioned percolation transition of the E-R graph takes place at $p = p_c \equiv 1/N$, where p is the probability that two nodes are connected by an edge and N is the total number of nodes in the graph. The appearance of a *giant component* in a network, which is also referred to as the *percolating component*, results in a dramatic change in the overall topological features of the graph and has been in the centre of interest for other networks as well.

A similar critical linking probability can be derived for the emergence of a giant cluster composed of adjacent k-cliques (complete subgraphs of k nodes) [21, 66]. Naturally, this critical probability grows with increasing

with k, as the conditions for the formations of a k-clique percolation cluster become more restrictive. Although k-clique percolation is a generalisation of edge percolation, it has a couple of unique features, e.g. a node can be part of several k-clique percolation clusters at the same time, and multiple order parameters can be defined to describe the percolation transitions, which show different behaviour at the critical point. These aspects are detailed in Sect. 2, which is aimed at summarising the most important results concerning k-clique percolation in the E-R graph. The concept is extended to weighted and directed networks in Sects. 3–4. Finally, in Sect. 5 we show how k-clique percolation can be applied to community finding and network clustering.

2. k-clique Percolation in the E-R-graph

We begin with a few definitions laying down the fundamentals of k-clique percolation [21, 66]:

- *k-clique:* a complete (fully connected) subgraph of k vertices [14].
- *k-clique adjacency:* two k-cliques are adjacent if they share $k-1$ vertices, i.e., if they differ only in a single vertex.
- *k-clique chain*: a subgraph, which is the union of a sequence of adjacent k-cliques.
- *k-clique connectedness:* two k-cliques are k-clique-connected, if there exists at least one k-clique chain containing the two k-cliques.
- *k-clique percolation cluster (or component):* a maximal k-clique-connected subgraph, i.e., it is the union of all k-cliques that are k-clique-connected to a particular k-clique.

The above concept of k-clique percolation can be illustrated by *"k-clique template rolling"* (see Fig. 1). A k-clique template can be thought of as an object that is isomorphic to a complete graph of k nodes. Such a template can be placed onto any k-clique of the network, and rolled to an adjacent k-clique by relocating one of its nodes and keeping its other $k-1$ nodes fixed. Thus, the k-clique-communities of a graph are all those subgraphs that can be fully explored by rolling a k-clique template in them but cannot be left by this template. We note that a k-clique percolation cluster is very much like a regular edge percolation cluster in the *k-clique adjacency graph*, where

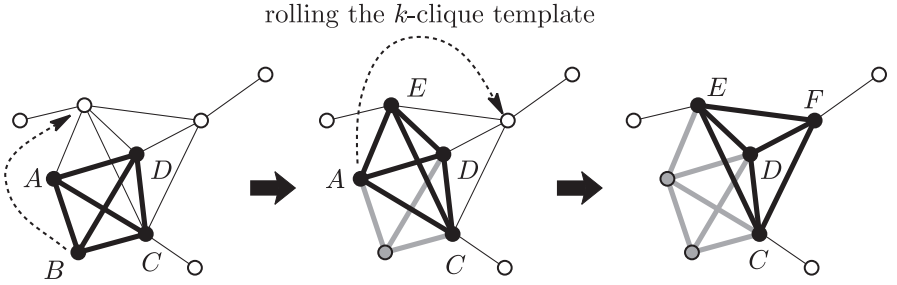

Fig. 1. Illustration of k-clique template rolling at $k = 4$. Initially the template is placed on A–B–C–D (left panel) and it is "rolled" onto the subgraph A–C–D–E (middle panel). The position of the k-clique template is marked with thick black lines and black nodes, whereas the already visited edges are represented by thick gray lines and gray nodes. Observe that in each step only one of the nodes is moved and the two 4-cliques (before and after rolling) share $k - 1 = 3$ nodes. At the final step (right panel) the template reaches the subgraph C–D–E–F, and the set of nodes visited during the process (A–B–C–D–E–F) are considered as a k-clique percolation cluster.

the vertices represent the k-cliques of the original graph, and there is an edge between two vertices, if the corresponding two k-cliques are adjacent. Moving a particle from one node of this adjacency graph to another one along an edge is equivalent to rolling a k-clique template from one k-clique of the original graph to an adjacent one. Note that a node can be part of several k-clique percolation clusters at the same time, the simplest example of this is given by two triangles (k-cliques at $k = 3$) overlapping in a single node.

A k-clique percolation cluster fulfilling the above definition is a very good candidate for a community in real networks. We shall detail this aspect in Sect. 5. Here we mention that these objects can be considered as interesting specific cases of the general graph theoretic objects defined in [25, 8] in very different contexts.

2.1. Derivation of the critical point with heuristic arguments

The threshold probability (critical point) of k-clique percolation in the E-R random graph can be obtained using the template rolling picture with the following simple heuristic arguments. At the percolation threshold we have to require that after rolling a k-clique template from a k-clique to an adjacent one (by relocating one of its vertices), the expectation value of the number of adjacent k-cliques, where the template can roll further (by

relocating another of its vertices), be equal to one. The intuitive argument behind this criterion is that a smaller expectation value would result in premature k-clique percolation clusters, because starting from any k-clique the rolling would quickly come to a halt and, as a consequence, the size of the clusters would decay exponentially. A larger expectation value, on the other hand, would allow an infinite series of bifurcations for the rolling, ensuring that a giant cluster is present in the system. The above expectation value can be estimated as $(k-1)(N-k-1)p^{k-1}$, where the first term $(k-1)$ counts the number of vertices of the template that can be selected for the next relocation, the second term $(N-k-1)$ counts the number of potential destinations for this relocation, out of which only the fraction p^{k-1} is acceptable, because each of the new $k-1$ edges (associated with the relocation) must exist in order to obtain a new k-clique. For large N, our criterion can thus be written as

$$(k-1)Np_{\mathrm{c}}^{k-1} = 1, \tag{1}$$

from which we get

$$p_{\mathrm{c}}(k) = \frac{1}{\left[(k-1)N\right]^{\frac{1}{k-1}}} \tag{2}$$

for the threshold probability. The subscript "c" throughout this Chapter indicates that the system is at the percolation threshold (or critical point). Obviously, for $k = 2$ the above result agrees with the known percolation threshold ($p_{\mathrm{c}} = 1/N$) for E-R graphs, because 2-clique connectedness is equivalent to regular (edge) connectedness.

2.2. Generating function formalism

The above results can be made stronger with the help of the generating function formalism [66] in a fashion similar to that of [58]. (Note that the following derivation is still heuristic from a rigorous mathematical point of view). We first summarise the definition and the most important properties of the generating functions. If a random variable ξ can take non-negative integer values according to some probability distribution $\mathscr{P}(\xi = n) \equiv \rho(n)$,

then the corresponding generating function is given by

$$G_\rho(x) \equiv \langle x^\xi \rangle = \sum_{n=0}^{\infty} \rho(n) x^n. \tag{3}$$

The generating-function of a properly normalised distribution is absolutely convergent for all $|x| \leq 1$ and hence has no singularities in this region. For $x = 1$ it is simply

$$G_\rho(1) = \sum_{n=0}^{\infty} \rho(n) = 1. \tag{4}$$

The original probability distribution and its moments can be obtained from the generating-function as

$$\rho(n) = \frac{1}{n!} \left. \frac{d^n G_\rho(x)}{dx^n} \right|_{x=0}, \tag{5}$$

$$\langle \xi^l \rangle = \sum_{n=0}^{\infty} n^l \rho(n) = \left[\left(x \frac{d}{dx} \right)^l G_\rho(x) \right]_{x=1}. \tag{6}$$

And finally, if $\eta = \xi_1 + \xi_2 + \cdots + \xi_l$, where $\xi_1, \xi_2, \ldots, \xi_l$ are independent random variables (with non-negative integer values), then the generating function corresponding to $\mathscr{P}(\eta = n) \equiv \sigma(n)$ is given by

$$\begin{aligned} G_\sigma(x) &= \langle x^\eta \rangle = \left\langle x^{\xi_1} x^{\xi_2} \cdots x^{\xi_l} \right\rangle = \left\langle x^{\xi_1} \right\rangle \left\langle x^{\xi_2} \right\rangle \cdots \left\langle x^{\xi_l} \right\rangle \\ &= G_{\rho_1}(x) G_{\rho_2}(x) \cdots G_{\rho_l}(x). \end{aligned} \tag{7}$$

Now, we can proceed to the derivation of the critical point in the $N \to \infty$ limit. First, let us consider the probability distribution $r(n)$ of the number of k-cliques adjacent to a randomly selected k-clique. Finding a k-clique B adjacent to a selected k-clique A is equivalent to finding a node outside A linked to at least $k-1$ nodes in A. The number of possibilities for this node is $N - k$. Edges in the E-R graph are independent of each other, therefore the probability that a given node is linked to all nodes in A is p^k, whereas the probability that it is linked to $k-1$ nodes in A is $k(1-p)p^{k-1}$. Therefore, to leading order in N the average number of k-cliques adjacent to a randomly selected one is

$$\langle r \rangle = (N-k)\left[k(1-p)p^{k-1} + p^k \right] \simeq Nkp^{k-1}. \tag{8}$$

From the independence of the edges it also follows that the probability distribution $r(n)$ becomes Poissonean, which can be written as

$$r(n) = \exp\left(-Nkp^{k-1}\right)\frac{\left(Nkp^{k-1}\right)^n}{n!}. \tag{9}$$

Let us suppose that we are below the percolation threshold, therefore, k-cliques are rare, adjacent k-cliques are even more rare, and loops in the k-clique adjacency graph are so rare that we can assume it to be tree-like[1]. In this case the size of a connected component in the k-clique adjacency graph (corresponding to a k-clique percolation cluster) can be evaluated by counting the number of k-cliques reached in a branching process as follows. We start at an arbitrary k-clique in the component, and in the first step we invade all its neighbours in the k-clique adjacency graph. From then on, whenever a k-clique is reached, we proceed by invading all its neighbours, except for the one the k-clique has been reached from. In terms of the original graph, this is equivalent to rolling a k-clique template to all adjacent k-cliques except for the one we arrived from in the previous step.

In the process described above, we can assign to each k-clique the subgraph in the k-clique percolation cluster that was invaded from it. (Note that we assumed the k-clique adjacency graph to be tree-like). Let us denote by $I(n)$ the probability, that the subgraph reached from an arbitrary starting k-clique in the branching process contains n number of k-cliques, including the starting k-clique as well. This subgraph is actually equal to a k-clique percolation cluster. Similarly, let $H(n)$ denote the probability that the subgraph reached from a k-clique appearing later in the branching process (i.e., from a k-clique that is not the starting one) contains n number of k-cliques. This is equivalent to the probability that by starting at a randomly selected k-clique and trying to roll a k-clique template via all possible subsets of size $k-1$ except for one, then by subsequently rolling the template on and on, in all possible directions without turning back, a k-clique percolation "branch" of size n is reached. And finally, let $H_m(n)$ be the probability, that if we pick m number of k-cliques randomly, then the sum of the sizes of the k-clique branches that we can reach in this way consists of n number of k-cliques. Since we are below the percolation threshold, the k-clique adjacency graph consists of many dispersed components of small size, and the probability that two (or more) k-cliques out of m belong to the same k-clique percolation cluster is negligible. Hence, according to Eq. (7), the

[1]This assumption is an approximation since the adjacency graph is weakly assortative.

generating functions corresponding to $H(n)$ and $H_m(n)$, denoted by $G_H(x)$ and $G_{H_m}(x)$ respectively are related to each other as:

$$G_{H_m}(x) = \big[G_H(x)\big]^m. \tag{10}$$

Let $q(n)$ denote the probability, that for a randomly selected k-clique, by excluding one of its possible subsets of size $k-1$, we can roll a k-clique template through the remaining subsets to n adjacent k-cliques. This distribution is very similar to $r(n)$, except that in this case we can use only $k-1$ subsets instead of k in the k-clique to roll the k-clique template further, therefore

$$q(n) = \exp\big(-N(k-1)p^{k-1}\big)\,\frac{\big(N(k-1)p^{k-1}\big)^n}{n!}. \tag{11}$$

By neglecting the loops in the k-clique adjacency graph, H_n can be expressed as

$$H(n) = q(0)H_0(n-1) + q(1)H_1(n-1) + q(2)H_2(n-1) + \dots\,. \tag{12}$$

By taking the generating function of both sides and using Eqs. (5) and (10), we obtain

$$\begin{aligned} G_H(x) &= \sum_{n=0}^{\infty}\left[\sum_{m=0}^{\infty} q(m)H_m(n-1)\right]x^n \\ &= \sum_{n=0}^{\infty}\left[\sum_{m=0}^{\infty} q(m)\frac{1}{(n-1)!}\,\frac{d^{n-1}}{dx^{n-1}}\big[G_H(x)\big]^m\Big|_{x=0}\right]x^n = \\ &= \sum_{m=0}^{\infty} q(m)\big[G_H(x)\big]^m x = xG_q\big(G_H(x)\big), \end{aligned} \tag{13}$$

where $G_q(x)$ denotes the generating function of the distribution $q(n)$.

We can write an equation similar to Eq. (12) for $I(n)$ as well, in the form of

$$I(n) = r(0)H_0(n-1) + r(1)H_1(n-1) + r(2)H_2(n-1) + \dots \tag{14}$$

Again, by taking the generating functions of both sides we arrive at

$$G_I(x) = xG_r\big(G_H(x)\big), \tag{15}$$

where $G_r(x)$ denotes the generating function of $r(n)$. From Eqs. (6) and (15) we get

$$\langle I\rangle = G_I'(1) = G_r\big(G_H(1)\big) + G_r'\big(G_H(1)\big)\,G_H'(1) = 1 + G_r'(1)G_H'(1) \tag{16}$$

for the average size of the components invaded from a randomly selected k-clique. Using Eq. (13) we can write

$$G_H'(1) = G_q\big(G_H(1)\big) + G_q'\big(G_H(1)\big)\,G_H'(1) = 1 + G_q'(1)G_H'(1), \tag{17}$$

from which $G_H'(1)$ can be expressed as

$$G_H'(1) = \frac{1}{1 - G_q'(1)}. \tag{18}$$

By substituting this back into Eq. (16) we get

$$\langle I\rangle = 1 + \frac{G_r'(1)}{1 - G_q'(1)} = 1 + \frac{\langle r\rangle}{1 - \langle q\rangle}. \tag{19}$$

The above expression for the expected size of the connected components in the k-clique adjacency graph invaded from a randomly selected k-clique diverges when

$$\langle q\rangle = N(k-1)p^{k-1} = 1. \tag{20}$$

This point marks the phase transition at which a giant component (corresponding to a giant k-clique percolation cluster) first appears. Therefore, our final result for the critical linking probability for the appearance of the giant reassures Eq. (2), found via heuristic arguments.

2.3. Partial differential equation approach to k-clique percolation

The critical point of k-clique percolation was studied in a more general framework by Ráth and Tóth [69]. (Similarly to Sects. 2.1–2.2., the results of this Section are based on heuristics and do not provide rigorous proofs in the mathematical sense). In this approach the E-R graph is constructed in a stochastic process: from an initially empty graph containing only nodes but no edges, the possible edges are introduced at a rate of $1/\sqrt{N}$. At time t, the ratio of "occupied" edges equals $1 - e^{-t/\sqrt{N}}$, therefore, for $N \to \infty$ the

resulting graph at any time t is equivalent to an E-R graph with a linking probability $p = t/\sqrt{N}$.

The above model can be naturally extended by replacing the initial state with a non-empty graph. The method used by Ráth and Tóth to derive p_c for this general case was based on partial differential equations (PDE) and can be applied to arbitrary initial component size distribution. However, the initial state must fulfil the following conditions:

- If a small subset of edges is selected, then the distribution of the sizes of the percolation clusters of these edges is asymptotically independent from the probability distribution of the sizes of the rest of the percolation clusters as $N \to \infty$.
- The k-clique percolation clusters of the initial graph correspond to trees of the k-clique adjacency graph.

The starting point of this method is the approximation of the change in the number of k-clique percolation clusters having a given number of edges m between time t and $t + dt$. Although this approach can be generalised to higher k in a straightforward way, Ráth and Tóth focused exclusively on the $k = 3$ case. By denoting the number of k-clique percolation clusters with m edges at time t by $\mathcal{C}_m(N, t)$, they introduced the following quantities:

$$c_m(t) \equiv \lim_{N\to\infty} \frac{\mathcal{C}_m(N,t)}{\frac{1}{2}N^{\frac{3}{2}}}, \tag{21}$$

$$v_m(t) \equiv m \cdot c_m(t). \tag{22}$$

The Laplace-transforms of $c_m(t)$ and $v_m(t)$ are given by

$$V(t,x) \equiv \sum_{m\in\mathbb{N}} v_m(t) \cdot e^{-m\cdot x}, \tag{23}$$

$$C(t,x) \equiv \sum_{m\in\mathbb{N}} c_m(t) \cdot e^{-m\cdot x}. \tag{24}$$

By introducing

$$E(t) \equiv \lim_{N\to\infty} \frac{\big|E(N,t)\big|}{\frac{1}{2}N^{\frac{3}{2}}}, \tag{25}$$

Ráth and Tóth showed that in the mean-field approximation and in the $N \to \infty$, $dt \to 0$ limits the $C(t,x)$ satisfies the following PDE:

$$\frac{\partial}{\partial t}C(t,x) = e^{V(t,x)^2 - E(t)^2 - x} - 2V(t,x) \cdot E(t)^2. \tag{26}$$

Equation (26) was solved with the method of characteristics [69], resulting in the following expression:

$$t_c^2 \cdot (2ab+1) + t_c \cdot \big(b + a \cdot (2ab+1)\big) = \frac{1}{2}, \tag{27}$$

where $a = V(0,0) = \sum_m v_m(0)$, $b = -\frac{\partial}{\partial x}V(0,0)$ and t_c denotes the time of the appearance of the giant k-clique percolation cluster. In the special case of an empty initial graph $a = b = 0$, yielding $t_c = 1/\sqrt{2}$ and

$$p_c = \frac{t_c}{\sqrt{N}} = \frac{1}{\sqrt{2N}}, \tag{28}$$

in agreement with the results of Sect. 2.1. From Eq. (26) Ráth and Tóth also derived an equation for the rescaled size of the giant component (number of links compared to the total number of links), denoted by $v_\infty(t)$. By introducing

$$W(t,x) \equiv e^{V(t,x)^2 - E(t)^2 - x}, \tag{29}$$

$$\widehat{V}(t,w) \equiv V\big(t, \widehat{X}(t,w)\big), \tag{30}$$

where $\widehat{X}(t,w)$ denotes the inverse function of $W(t,x)$ in the x variable, this equation was formulated as [69]

$$v_\infty(t) = \widehat{V}(0,1) + t - \widehat{V}\big(0, W(t,0)\big) - tW(t,0). \tag{31}$$

2.4. Numerical simulations

The numerical studies of k-clique percolation in the E-R graph are in full agreement with the results obtained in Sects. 2.1–2.2. The observed transition is continuous, characterised by non-universal critical exponents, which depend on both k and the way the size of the giant component is measured.

There are two plausible choices for measuring the size of the giant component: The most natural one, which we denote by N^*, is the number of vertices belonging to this cluster. We can also define an *order parameter* associated with this choice as the relative size of that cluster:

$$\Phi = N^*/N. \tag{32}$$

The other choice is the number $\mathscr{N}^*$ of k-cliques of the largest k-clique percolation cluster (or equivalently, the number of vertices of the largest component in the k-clique adjacency graph). The associated order parameter is again the relative size of this cluster:

$$\Psi = \mathscr{N}^*/\mathscr{N}, \tag{33}$$

where $\mathscr{N}$ denotes the total number of k-cliques in the graph (or the total number of vertices in the adjacency graph). $\mathscr{N}$ can be estimated as

$$\mathscr{N} \approx \binom{N}{k} p^{k(k-1)/2} \approx \frac{N^k}{k!} p^{k(k-1)/2}, \tag{34}$$

because k different vertices can be selected in $\binom{N}{k}$ different ways, and any such selection makes a k-clique only if all the $k(k-1)/2$ edges between these k vertices exist, each with probability p. Note that the classical E-R percolation is equivalent to our $k = 2$ case, and the E-R order parameter (relative number of nodes) is identical to Φ. Also note that in general the size of the largest cluster could be measured as the number of its l-cliques, $\mathscr{N}^*_{(l)}$, for $1 \le l \le k$. However, for simplicity we restrict ourselves to the two limiting cases ($N^* \equiv \mathscr{N}^*_{(1)}$ and $\mathscr{N}^* \equiv \mathscr{N}^*_{(k)}$) defined above.

Computer simulations indicate that the two order parameters behave differently near the threshold probability. To illustrate this, in Figs. 2. we plotted Φ and Ψ, respectively, as a function of $p/p_c(k)$ for $k = 4$ and for various system sizes (N), averaged over several runs. The order parameter Φ for $k \ge 3$ converges to a step function as $N \to \infty$. The fact that the step is located at $p/p_c(k) = 1$ is actually the numerical proof of the validity of the theoretical prediction (2) for $p_c(k)$. The order parameter Ψ for $k \ge 2$, on the other hand, similarly to the classical E-R transition, converges to a limit function, which is 0 for $p/p_c(k) < 1$ and grows continuously from 0 to 1 if we increase $p/p_c(k)$ from 1 to ∞. The limiting shape of this curve (with proof) is given in [15].

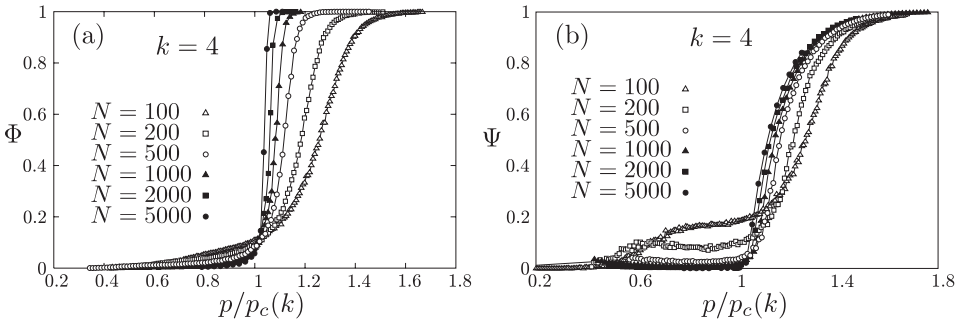

Fig. 2. (a) Simulation results for the order parameter Φ as a function of $p/p_c(k)$ at $k = 4$, averaged over several runs, such that the statistical error is smaller than the size of the symbols. Φ converges to a step function in the $N \to \infty$ limit. (b) The order parameter Ψ as a function of $p/p_c(k)$ for the same simulations as in (a). Ψ converges to a limit function (which is 0 for $p/p_c(k) < 1$ and grows continuously to 1 above $p/p_c(k) = 1$) in the $N \to \infty$ limit. Figure from [21]

The width of the steps in Fig. 2. follows a power law, $\sim N^{-\alpha}$, with some exponent α. Plotting Φ as a function of $\left[p/p_c(k) - 1\right] N^{\alpha}$, i.e., stretching out the horizontal scale by N^{α}, the data collapse onto a single curve. This is shown for $k = 3$, 4, and 5 in Fig. 3a. The exponent α is around 0.5 for every $k \geq 3$. Although for $k = 3$ a slight deviation from $\alpha = 0.5$ has been obtained, it cannot be distinguished from a possible logarithmic correction.

One of the most fundamental results in random graph theory concerns the behaviour of the largest component at the percolation threshold, where it becomes infinitely large in the $N \to \infty$ limit. Erdős and Rényi showed [24] that for the random graphs they introduced, the size of the largest component N^* (measured as the number of its nodes) at $p = p_c \equiv 1/N$ diverges with the system size as $N^{2/3}$, or equivalently, the order parameter Φ scales as $N^{-1/3}$. A similar scaling behaviour can be observed for k-clique percolation at the threshold probability $p_c(k)$ as well. If we assume, that the k-clique adjacency graph is like an E-R graph, then at the threshold the size of its giant component $\mathscr{N}_c^*$ scales as $\mathscr{N}_c^{2/3}$. Plugging $p = p_c$ from Eq. (2) into Eq. (34) and omitting the N-independent factors we get the scaling

$$\mathscr{N}_c \sim N^{k/2} \tag{35}$$

for the total number of k-cliques. Thus, the size of the giant component $\mathscr{N}_c^*$ is expected to scale as $\mathscr{N}_c^{2/3} \sim N^{k/3}$ and the order parameter Ψ_c as $\mathscr{N}_c^{2/3}/\mathscr{N}_c \sim N^{-k/6}$.

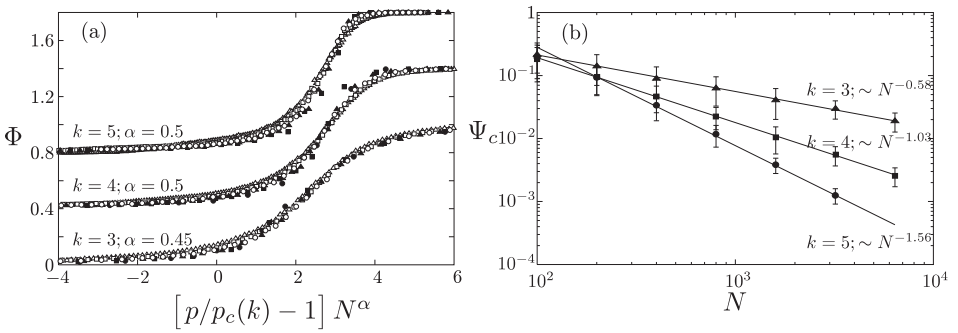

Fig. 3. (a) The width of the steps in Fig. 2a follows a power law, $\sim N^{-\alpha}$, as the steps collapse onto a single curve if we stretch them out by N^{α} horizontally. We have plotted the results obtained at $k = 3$ and $k = 5$ as well. The data for $k = 4$ and $k = 5$ are shifted upward by 0.4 and 0.8, respectively, for clarity. (b) The order parameter at the threshold, $\Psi_{\rm c}$, scales as some negative power of N, in good agreement with expression (36). Figure from [21].

This is valid, however, only if $k \leq 3$. The reason for the breakdown of the above scaling is that for $k > 3$ it predicts that the number of k-cliques of the giant k-clique percolation cluster, i.e., the number of vertices of the giant component in the k-clique adjacency graph, $\mathscr{N}_{\rm c}^{2/3} \sim N^{k/3}$, grows faster than N. On the other hand, in analogy with the structure of the giant component of the classical E-R problem, we expect that the giant component in the adjacency graph also has a tree-like structure at the threshold, with very few loops. As a consequence, almost every node of the adjacency graph corresponds to a node of the original graph. Thus, in the adjacency graph the giant component should not grow faster than N at the threshold. Therefore, for $k > 3$ we expect that $\mathscr{N}_{\rm c}^{*} \sim N$, and using Eq. (35), $\Psi_{\rm c} = \mathscr{N}_{\rm c}^{*}/\mathscr{N}_{\rm c} \sim N^{1-k/2}$. In summary:

$$\Psi_{\rm c} \sim \begin{cases} N^{-k/6} & \text{for } k \leq 3 \\ N^{1-k/2} & \text{for } k \geq 3 \end{cases}. \tag{36}$$

Numerical results showing the scaling of $\Psi_{\rm c}$ at $p_{\rm c}$ as a function of N are depicted in Fig. 3b, and the results are in good agreement with the above arguments.

3. Percolation of Weighted k-cliques

As mentioned in the introduction, k-clique percolation is important from the point of view of community finding, as k-clique percolation clusters can be considered as dense communities. For the majority of the networks occurring in nature and society, the edges connecting the nodes have an associated *weight* as well, referring to the strength/intensity of the relation between its endpoints.

Plain k-clique percolation can be used even in these cases, as described in Sect. 5.1, on a truncated graph that contains only those edges that have a weight higher than a given threshold. However, a method that incorporates edge weights is expected to produce better results. Thus, we suggest a generalisation of clique percolation to weighted networks [27]. For binary graphs, using percolating k-clique clusters instead of only full cliques is a less restrictive approach that highlights extended, less dense regions of the graph. In the same way, our aim when defining weighted clique percolation is to introduce concepts that can be used to examine those parts of the graph that are denser than a given lower limit.

3.1. Definitions

In a weighted network, to each edge, (i, j), a *weight* $w_{ij} \in \mathbb{R}$ is assigned. The *intensity* of a k-clique C is defined as the geometric mean of its edge weights [62]:

$$I(C) = \left(\prod_{\substack{i<j \\ i,j \in C}} w_{ij} \right)^{2/\left[k(k-1)\right]} .$$

Compared to unweighted k-clique percolation clusters, we define *weighted k-clique percolation clusters* by considering only k-cliques having an intensity greater than a given threshold I. In analogy with the definition of k-clique adjacency, a *weighted k-clique chain* is a k-clique chain where the intensity of all cliques is above I.

Compared to k-clique percolation on the truncated graph, the above definition is less restrictive: a k-clique containing weak edges (low weights) can be part of the percolation cluster if it contains a considerable number of strong (large weights) edges as well.

3.2. Percolation transition in the weighted E-R graph

To define a weighted version of the E-R graph, we assign to each edge of this graph a weight selected independently and randomly from a uniform distribution on the interval $(0, 1]$. At a fixed I, the critical linking probability, $p_c(I)$, of k-clique percolation is the edge probability where a giant module (containing k-cliques fulfilling the intensity condition) emerges. The special case $I = 0$ is equivalent to the unweighted case.

The results derived in Sect. 2. for unweighted clique percolation provide an upper limit for p_c in a weighted E-R graph. Consider weighted E-R graph as defined above, remove edges weaker than I and ignore the weights of the remaining edges. The resulting graph will an E-R graph with link probability $p^* = p(1 - I)$. The (unweighted) k-cliques of this truncated graph are a subset of those weighted k-cliques of the original graph which pass the intensity threshold, resulting in a higher percolation threshold. This gives the upper limit $p_c(I) < p_c(0)/(1 - I)$.

A better approximation can be given by modifying the heuristic argument considered in Sect. 2.1. by taking into account the intensities of the cliques. We keep the main idea (a k-clique template is rolled and percolation of the k-cliques is required) and modify only the condition for rolling the template further. In the process of rolling if a k-clique C_1 precedes another k-clique C_2, then we will say that C_1 is the *parent* of C_2, and C_2 is a *child* of C_1. The k-clique template can be rolled from parent to child, if the child k-clique passes the intensity threshold I.

We consider two approximations for this. First we assume that the probability distribution of edge weights in the child k-cliques is the original uniform distribution on the interval $(0, 1]$. (The actual probability distribution of an edge weight is different from this, as we shall see shortly.) Denoting by $\mathscr{P}_k < 1$ the probability that the child k-clique has an intensity larger than I, the expected number of cliques available for the template to roll to is $p^{k-1}N(k-1)\mathscr{P}_k$. Applying the criterion for percolation, i.e. that the expectation value of this number should be 1, we get $p_c(I) \simeq p_c(0)\,\mathscr{P}_k^{-1/(k-1)}$.

The probability $\mathscr{P}_k$ is simply the probability that the product of $k(k-1)/2$ independent random variables, with uniform distribution on $(0, 1]$, reaches $I^{k(k-1)/2}$, and can be expressed with straightforward but tedious integrals.

As noted above, the probability distribution of the edge weights in the child k-cliques is *not* the original uniform distribution: the parent clique has

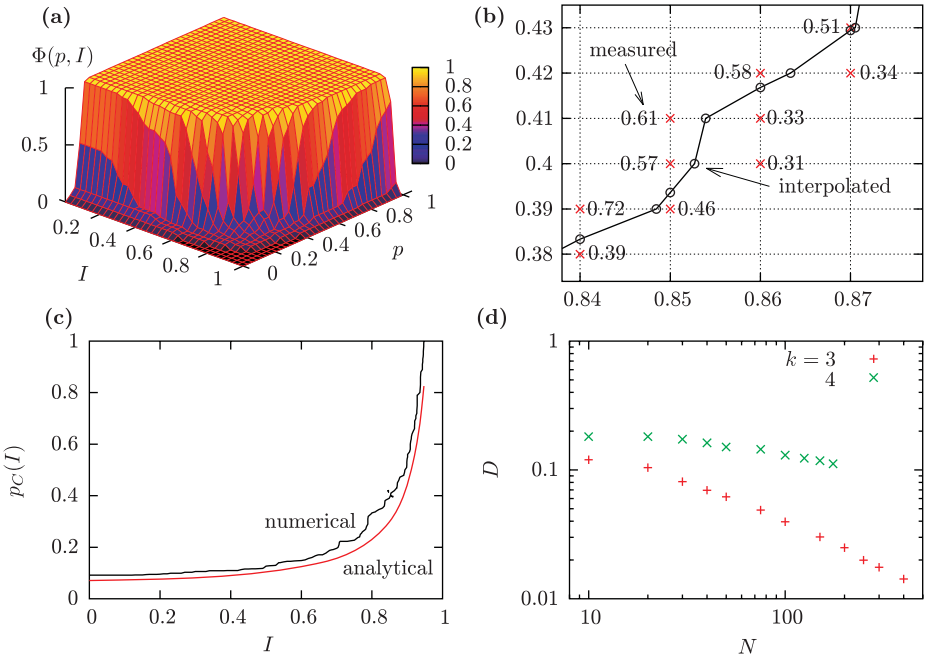

Fig. 4. Numerical analysis of the percolation of k-cliques fulfilling the intensity condition in weighted E-R graphs. The sample numerical results shown in panels (a-c) were obtained for $N = 100$ and $k = 3$ using 1 run for each (p, I) grid point. In panel (d) points were computed from 3 to 100 runs for each (k, I) parameter pair and error bars are smaller than the sizes of the symbols. **(a)** The order parameter, Φ, in the points of a grid on the (k, I) plane. **(b)** We computed the transition line, $p_c = p_c(I)$, as the curve with $\Phi = 1/2$ on the (k, I) plane. From the values of Φ at nearby grid points we increased the precision of the transition line with linear interpolation. **(c)** Numerical curve for the percolation threshold and the second order analytical approximation. The area between the two curves, D, measures the difference between the two results. **(d)** Difference between the numerical and analytical results for $p_c(I)$ at various system sizes, N, and clique size parameters. Figure from [27].

an intensity above I, and it shares $(k-1)(k-2)/2$ edges with candidate k-cliques. With an appropriate simplifying approximation for the probability distribution of edge weights of the *parent* clique, one can calculate a second-order approximation for p_c. Higher-order approximations can be generated by taking into account the probability distribution of the edge weights of the grandparent clique, grand-grandparent clique and so on.

To numerically check our results, we generated weighted E-R networks of size $N = 10 \ldots 400$, and measured $p_c(I)$ in the following way (see Fig. 4): we calculated the order parameter Φ, as defined in Eq. (32) on a square

grid, and approximated the points where $\Phi = 1/2$ by linear interpolation along the grid-lines. The algorithm we used exploited the fact that, for a given initial random seed, one can reuse the cluster structure at (p_1, I_1) for calculating the cluster structure at a different (p_2, I_2), using a Hoshen–Kopelman (also known as Union-Find) algorithm.

4. Directed k-clique Percolation

In practice many real networks contain *directed connections* among the vertices, where the direction of a single link signals either the direction of some kind of flow (e.g. the flow of information, energy), or the asymmetry of the relation between the vertices (e.g. a superior-inferior relation). This raises the question of whether a community searching algorithm that inherently takes into account the directionality of links is more suitable for directed networks than the usual undirected algorithms. Along this idea, in this section we define the notion of *directed k-cliques* (in which the configuration of the directed links has to meet certain criteria), and study the percolation of these objects in the directed equivalent of the E-R graph [67].

4.1. Definitions

In undirected graphs a pair of nodes is either connected or not, whereas in a directed graph the same pair, (A, B), can be connected in three ways: either by a "single link" as (i) $\mathrm{A} \to \mathrm{B}$ and (ii) $\mathrm{A} \leftarrow \mathrm{B}$ or by a "double link" as (iii) $\mathrm{A} \rightleftharpoons \mathrm{B}$. Multiple links (i.e., more than one link between A and B in the same direction) and self-links (such as $\mathrm{A} \to \mathrm{A}$) are not allowed.

In a complete subgraph of size k the $k(k-1)/2$ links can be directed in $3^{k(k-1)/2}$ ways. Since in the undirected k-clique percolation we treat these alternatives as identical, introducing link directions allows a large variety of possible rules for defining directed k-cliques. A natural concept, however, is to aim for objects preserving some kind of directedness as a whole, rather than just being a collection of nodes connected by directed links. Therefore, we define directed k-cliques as complete subgraphs of size k in which an ordering can be made such that between any pair of nodes there is a directed link pointing from the node with the higher order towards the lower one. Since the presence of double links usually leads to multiple

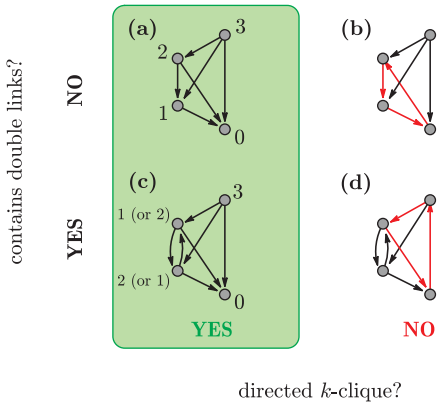

Fig. 5. Groups of nodes forming a directed k-clique **(a, c)** and groups **(b, d)** that do not. **(a)** A directed k-clique without double links. The index of each node corresponds to its order (equivalent to the number of its out-links) within the directed k-clique. **(b)** A complete subgraph without double links, but not accepted as a directed k-clique, because it contains a directed loop. **(c)** A directed k-clique with a double link. Note that the order of the nodes depends on which link is deleted from the double link. **(d)** Double link in a complete subgraph that is not a directed k clique. It is not possible to remove a link from the double link in a way that all directed loops disappear.

possibilities to order the nodes in a way fulfilling the above requirement, for simplicity we first concentrate on directed k-cliques without double links. In this case, the higher the order of a node, the more out-neighbours it has in the k-clique (see illustration in Fig. 5a). Thus, the *restricted out-degree* of a node in the k-clique (the number of its out-neighbours in the k-clique, ranging from 0 to $k-1$) can be assigned as its order. From this, it can be seen easily (for details see Appendix of [67]) that the condition for a k-clique with no double links to qualify as a directed k-clique is equivalent to the following three conditions:

- Any directed link in the k-clique points from a node with a higher order (larger restricted out-degree) to a node with a lower order.
- The k-clique contains no directed loops (where a "directed loop" is a closed directed path).
- The restricted out-degree of each node in the k-clique is different.

The overall directionality of such an object naturally follows the ordering of the nodes: the node with highest order is the one which has only out-neighbours, and can be viewed as the "source" or "top"-node of the k-clique,

whereas the node with lowest order has only incoming links from the others, and corresponds to a "drain" or "bottom" node.

None of the above three conditions holds in the presence of double links: directed loops appear in the k-clique, the restricted out-degree of at least two nodes in the k-clique become the same (see Appendix of [67]), and we can find directed links pointing in the direction of increasing order. However, based on the ordering of the nodes, it is always possible to eliminate the double links (by removing all links that point towards higher order) from a directed k-clique in such a way that the remaining single links fulfil all three conditions. See Fig. 5c as an example.

The k-clique adjacency is defined similarly to the undirected case: two directed k-cliques are adjacent if they share $k-1$ nodes. The directed k-clique percolation clusters arise as the union of directed k-cliques that can be reached from each other through a series of k-clique adjacency. The k-clique template rolling picture can be applied to illustrate these clusters in the same fashion as in the undirected case. The directed k-clique percolation clusters provide a community definition for real networks in a similar fashion to the undirected case (see Sect. 5. for details).

4.2. Percolation transition in the directed E-R graph

The directed equivalent of the E-R graph consists of N nodes providing $N(N-1)$ possible "places" for the directed links, and these are filled independently with uniform probability p, producing on average $M \simeq N(N-1)p$ links. (Note that in the original undirected E-R graph there are only $N(N-1)/2$ possibilities to introduce an edge, therefore, at linking probability p, there are only $M \simeq N(N-1)p/2$ connections). In the following we shall evaluate the critical linking probability, $p_{\mathrm{c}}^{\mathrm{dir}}$, for directed k-clique percolation using similar heuristic arguments as in Sect. 2.1.

$p_{\mathrm{c}}^{\mathrm{dir}}$ is decreasing with increasing N, and converges to zero as $N \to \infty$. We restrict ourselfs to the large N limit, and evaluate $p_{\mathrm{c}}^{\mathrm{dir}}$ to leading order only. Let us suppose that we approach the critical point from below: the directed k-cliques do not assemble yet into a giant cluster and we can find only small, isolated clusters, i.e. the system is dispersed. In our k-clique template rolling picture this means that when trying to explore the directed percolation clusters by rolling such a template on them, we must stop the rolling after a few steps as we run out of unexplored adjacent directed k-cliques.

One can estimate p_c^{dir} from the condition that at the critical point the average number of yet unexplored directed k-cliques adjacent to the k-clique we have just reached becomes equal to one. (This makes it possible to roll our template on and on for a long time). Since we are going to evaluate p_c^{dir} to leading order only, we can neglect the possibility to roll our k-clique template using double links between the same nodes: When reaching a directed k-clique, the minimal number of further links that must be present to enable the continuation of the template rolling is $k-1$. The probability of such a case is therefore proportional to p^{k-1}. Even though it is not forbidden in the first place to continue using double links as well, each double link in the new directed k-clique we are going to roll onto multiplies the probability by p. In other words, the probability to roll further to a k-clique containing one double link is smaller by a factor of p, the probability to roll further to a k-clique containing two double links is smaller by a factor of p^2, etc.

Consider the branching process exploring a directed k-clique percolation cluster at the point when we are about to roll our template further on. We can choose the next node for relocation in $k-1$ different ways, which can then be relocated to approximately N places. If there were no restrictions for directing the links inside a directed k-clique, then the $k-1$ new links connecting the new node to this $k-1$ shared nodes could be directed in 2^{k-1} ways. However, the new directed k-clique has to fulfil the three conditions detailed in Sect. 4.1. as well, therefore the actual number of allowed configurations is much smaller. The rank of the new node in the new directed k-clique can be chosen in k ways. The $k-1$ nodes shared with the previous k-clique are already ordered, and we can "insert" the new node anywhere into this hierarchy. By fixing the order of the new node we fix the direction of the new links as well, therefore, we can allow only k different configurations for the directionality of these links. By combining these factors together, the condition for reaching the critical point of the percolation transition can be written as

$$\left[p_c^{dir}\right]^{k-1} N(k-1)k = 1, \tag{37}$$

from which we gain

$$p_c^{dir} = \left[Nk(k-1)\right]^{-1/(k-1)} = p_c/k^{1/(k-1)} \tag{38}$$

as a first order approximation for the critical linking probability. Note that in the limiting case of $k = 2$ (the directed link percolation), the $p_c^{dir} = p_c/2$

relation holds, which is consistent with the 2 : 1 ratio for the number of links in the directed and undirected E-R graph, respectively.

To measure the size of the largest directed k-clique percolation cluster we can use Φ and Ψ, defined in Eqs. (32–33), in complete analogy with the undirected case. In Figs. 6a–b we display Φ and Ψ as functions of $p/p_{\rm c}^{\rm dir}$ obtained in numerical simulations, where the directed k-clique size was $k = 4$, and the system size varied between $N = 50$ and $N = 1600$. Similarly to the undirected k-clique percolation, the order parameter Φ converges to a step function for increasing system sizes, whereas Ψ converges to a limit function (which is 0 for $p/p_{\rm c}(k) < 1$ and grows continuously to 1 above $p/p_{\rm c}(k) = 1$). We have evaluated the transition point numerically as well, by computing the second moment of the distribution of $\mathcal{N}_i$ values, excluding the largest one, $\mathcal{N}_1 = \mathcal{N}^*$:

$$\chi = \sum_{i>1} (\mathcal{N}_i/\mathcal{N})^2. \tag{39}$$

Note that this quantity is analogous to the percolation susceptibility. Both below and above the transition point the $\mathcal{N}_i$ $(i > 1)$ values follow an exponential distribution, and only at $p_{\rm c}$ do they have a power-law distribution. Thus, χ is maximal at the numerical transition point, $p_{\rm c}^{\rm num}$. In Fig. 6c we show χ calculated for the curves shown in Fig. 6b, as a function of $p/p_{\rm c}^{\rm dir}$. In order to check the theoretical prediction for the critical point obtained in (38) we have carried out a finite-size scaling analysis of the numerical results. In Fig. 6d we show the ratio $p_{\rm c}^{\rm num}/p_{\rm c}^{\rm dir}$ as a function of $1/N$. Indeed, for large systems, the above ratio converges to 1 roughly as $1 + cN^{-1/2}$.

5. Applications: Community Finding and Clustering

The study of the *intermediate-scale* substructures in networks, made up of vertices more densely connected to each other than to the rest of the network, has become one of the most highlighted topics in complex network theory. A reliable method to pinpoint such objects has many potential industrial applications, e.g. it can help service providers (phone, banking, Internet, etc.) identify meaningful groups of customers (users), or support biomedical researchers in their search for individual target molecules and novel protein complex targets [47, 4]. Since communities have no widely accepted unique definition, the number of available methods to pinpoint

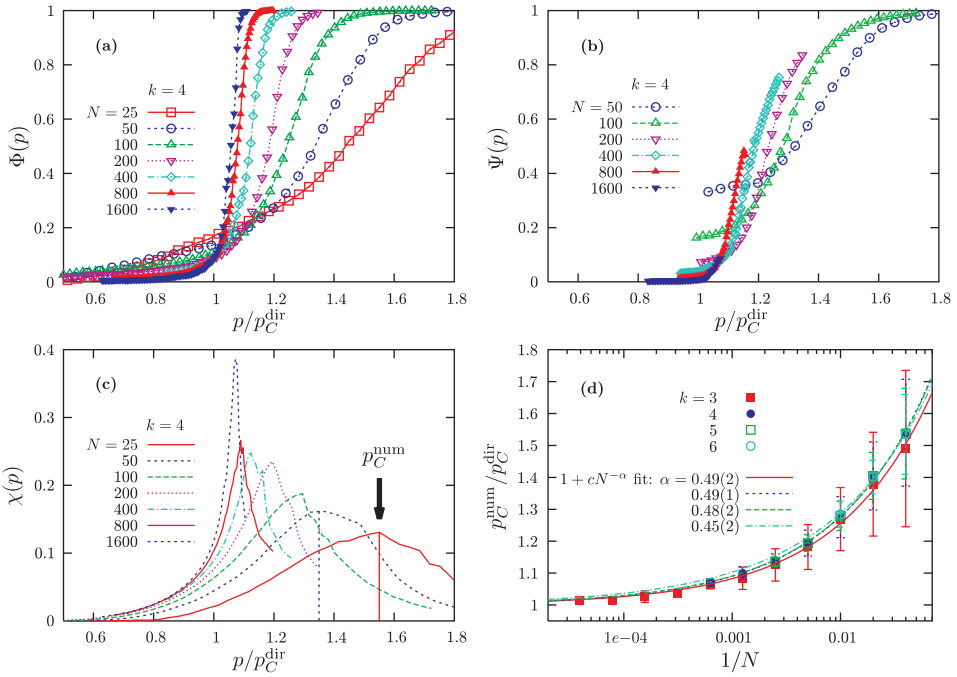

Fig. 6. Numerical results for directed k-clique percolation in ER-graphs. In each sub-figure, points show an average over 4 to 100 simulations depending on system size. **a)** The order parameter Φ (the number of nodes in the largest percolation cluster divided by N) as a function of $p/p_{\rm c}^{\rm dir}$, where $p_{\rm c}^{\rm dir}$ was obtained from Eq. (38). **b)** The order parameter Ψ (the number of directed k-cliques in the largest percolation cluster divided by the total number of directed k-cliques) as a function of $p/p_{\rm c}^{\rm dir}$. **c)** The numerically determined value for the critical linking probability, $p_{\rm c}^{\rm num}$, defined as the average location of the maximum of $\chi(p)$, playing the role of the normalised percolation susceptibility (see Eq. 39). **d)** Verification of the theoretical prediction for the critical point. The $p_{\rm c}^{\rm num}/p_{\rm c}^{\rm dir}$ ratio converges to 1 for large N. Figure from [67].

them is vast [74, 76, 26, 46, 32, 54, 73, 64, 27, 67, 71, 72, 37, 36, 38, 52]. The majority of these algorithms classify the nodes into disjoint communities, and in most cases a global quantity called *modularity* [56, 55] is used to evaluate the quality of the partitioning. However, as pointed out in [29, 49], the modularity optimisation introduces a resolution limit in the clustering, and communities containing a smaller number of edges than $\sqrt{M}$ (where M is the total number of edges) cannot be resolved. One of the big advantages of the *clique percolation method* (CPM) is that it identifies communities as k-clique percolation clusters, and therefore, the algorithm is *local*, and does not suffer from resolution problems of this type [64, 21].

Along with the rapid development of network clustering techniques, the ability of revealing overlaps between communities has become very important as well [86, 9, 39, 83, 31, 89, 57, 71, 52]. Indeed, communities in real-world graphs are often inherently overlapping: each person in a social web belongs usually to several groups (family, colleagues, friends, etc.), proteins in a protein interaction network may participate in multiple complexes [39] and a large portion of webpages can be classified under multiple categories. Prohibiting overlaps during module identification strongly increases the percentage of false negative co-classified pairs. As an example, in a social web a group of colleagues might end up in different modules, each corresponding to e.g. their families. In this case, the network module corresponding to their workgroup is bound to become lost. The other big advantage of CPM beside its local nature is that it allows overlaps between communities in a natural way: a node can be part of several k-clique percolation clusters at the same time.

5.1. The CPM in practice

In principle, the CPM detailed in Sect. 2 can be only applied to binary networks (i.e. to those with undirected and unweighted connections). However, an arbitrary network can always be transformed into a binary one by ignoring any directionality in the connections and keeping only those connections that are stronger than a threshold weight w^*. Changing the threshold is similar to changing the resolution (as in a microscope) with which the community structure is investigated: by increasing w^* the communities start to shrink and fall apart. A very similar effect can be observed by changing the value of k as well: increasing k makes the communities smaller and more disintegrated, but at the same time, also more cohesive. When we are interested in the community structure around a particular node, it is advisable to scan through a ranges of k and w^* values and monitor how the communities change. Meanwhile, when analysing the modular structure of the entire network, the criterion used to fix these parameters is based on finding a modular structure as highly structured as possible [64]. This can be achieved by tuning the parameters just below the critical point of the percolation transition. In this way we ensure that we find as many modules as possible, without the negative effect of having a giant module that would smear out the details of the modular structure by merging (and making invisible) many smaller modules.

The edge weights can be also taken into account in a somewhat refined way when using weighted k-cliques (fulfilling an edge-weight intensity criterion), as described in Sect. 3. This approach is referred to as the CPMw method [27], and the optimal k-clique intensity threshold can be adjusted similarly to the calibration of the optimal edge weight threshold described above.

The directed k-clique percolation clusters defined in Sect. 4 provide a suitable community definition for directed networks (this is the CPMd method [67]). Due to the asymmetry of the directed connections, nodes with mostly incoming links are expected to play a very different role in a given community from those having mostly outgoing links or from those having a similar amount of both kinds of links. A member (node) having only out-neighbours amongst the others can be viewed as a "source" or a "top-node", whereas a node with only incoming links from the rest of the community is a "drain" or a "bottom-node". Most nodes, however, fall usually somewhere between these two extremes. To quantify this property, we can introduce the *relative in-degree* and *relative out-degree* [67] of node i in a community α as

$$D^{\alpha}_{i,\mathrm{in}} \equiv \frac{d^{\alpha}_{i,\mathrm{in}}}{d^{\alpha}_{i,\mathrm{in}} + d^{\alpha}_{i,\mathrm{out}}}, \tag{40}$$

$$D^{\alpha}_{i,\mathrm{out}} \equiv \frac{d^{\alpha}_{i,\mathrm{out}}}{d^{\alpha}_{i,\mathrm{in}} + d^{\alpha}_{i,\mathrm{out}}}, \tag{41}$$

where $d^{\alpha}_{i,\mathrm{in}}$ and $d^{\alpha}_{i,\mathrm{out}}$ denote the numbers of in- and out-neighbours amongst the other nodes in the community, respectively. Obviously, the values of both $D^{\alpha}_{i,\mathrm{out}}$ and $D^{\alpha}_{i,\mathrm{in}}$ are in the range between 0 and 1, and the relation $D^{\alpha}_{i,\mathrm{in}}+D^{\alpha}_{i,\mathrm{out}} = 1$ holds. For weighted networks, Eqs. (40, 41) can be replaced by the *relative in-strength* and *relative out-strength* defined as

$$W^{\alpha}_{i,\mathrm{in}} \equiv \frac{w^{\alpha}_{i,\mathrm{in}}}{w^{\alpha}_{i,\mathrm{in}} + w^{\alpha}_{i,\mathrm{out}}}, \tag{42}$$

$$W^{\alpha}_{i,\mathrm{out}} \equiv \frac{w^{\alpha}_{i,\mathrm{out}}}{w^{\alpha}_{i,\mathrm{in}} + w^{\alpha}_{i,\mathrm{out}}}, \tag{43}$$

where $w^{\alpha}_{i,\mathrm{out}}$ and $w^{\alpha}_{i,\mathrm{out}}$ denote the aggregated weight of outgoing and incoming connections with other nodes in the community α. As an illustration, in Fig. 7. we show the directed communities of the word "GOLD" in a

word association network studied in [67]. The weight of a directed link in this case indicates the frequency at which people in questionnaires associated the endpoint of the link with its starting point. The communities are colour coded with the overlaps emphasised in red. According to its different meanings, the word "GOLD" participates in four, strongly internally connected communities. Beside the node labels we display the relative out-strength of the nodes in the communities using Eq. (43). Apparently, nodes with a special/particular meaning (e.g. "SAPPHIRE") tend to have high relative out-strength, whereas commonly used words with general meaning (e.g. "MONEY") have low relative out-strength. Thus, it seems that the overall directionality of the communities in this case is from special words towards more general words.

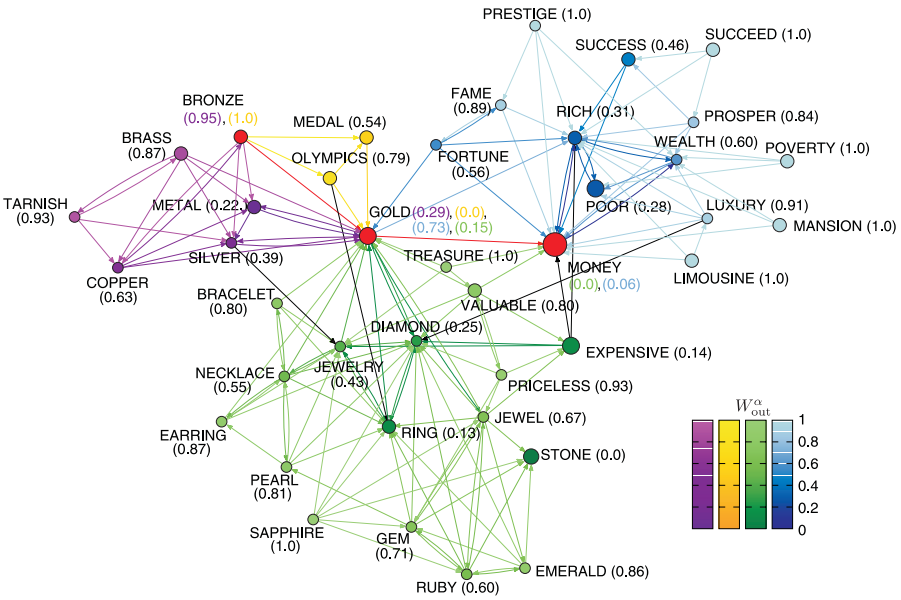

Fig. 7. The directed modules of the word "GOLD" in a word association network. The modules are colour coded and the overlaps between the modules are displayed in red. The size of each node is proportional to the number of modules it participates in (some of them are not shown in this figure). Beside the name of the nodes we display their $W^{\alpha}_{i,\text{out}} = w^{\alpha}_{i,out}/(w^{\alpha}_{i,in} + w^{\alpha}_{i,out})$ values as well. Nodes with high W (e.g. "SAPPHIRE") usually correspond to special, rarely used words, whereas nodes with low relative out-degree (e.g. "MONEY") are very common. Figure from [67].

5.2. Applying the CPM to real networks

In this section we summarise the most important results obtained so far with the help of the CPM in the analysis of real networks. These achievements are related to a wide spectrum of problems, ranging from cancer metastasis through the formation of social groups to the study of the directed communities of webpages. Here we focus solely on the results closely related to the CPM in the cited works.

5.2.1. The graph of communities. As we already pointed out, one of the big advantages of the CPM is that it allows overlaps between the communities. These overlaps naturally lead to the definition of the *community graph* [64, 68]: a network representing the connections between the communities, with the nodes referring to communities and edges corresponding to shared members between the communities of the original graph. The community graph can be treated as a "coarse-grained" view of the original network, and can be used to study the organisation of the system at a higher level. As an illustration, in Fig. 8. we show the community graph of the protein-protein interaction (PPI) network obtained from the DIP core list of protein-protein interactions of the yeast, *S. cerevisiae* [87]. The biological functions or protein complexes that can be associated with the communities shown in the left panel were looked up by using the GO-TermFinder package [16] and the online tools of the Saccharomyces Genome Database [20].

It is well known (see e.g. [5, 3, 51]) that the nodes of large real networks have a power law degree distribution. Studies of various complex systems showed that if we consider the network of communities instead of the nodes themselves, we still observe a degree distribution with a fat tail, but a characteristic scale is introduced, below which the distribution is exponential [64]. This is in agreement with our understanding of a complex system having different levels of organisation with units specific to each level. In addition, in the present case the principle of organisation (scaling) is preserved (with some specific modifications) when going to the next level in the hierarchy.

In a wide range of graph models the basic mechanism behind the emerging power law degree distribution is that new nodes appearing in the system attach to the "old" ones with a probability proportional to their degrees [5, 3, 51]. Furthermore, the occurrence of preferential attachment was directly demonstrated in several real-world networks with scale free degree

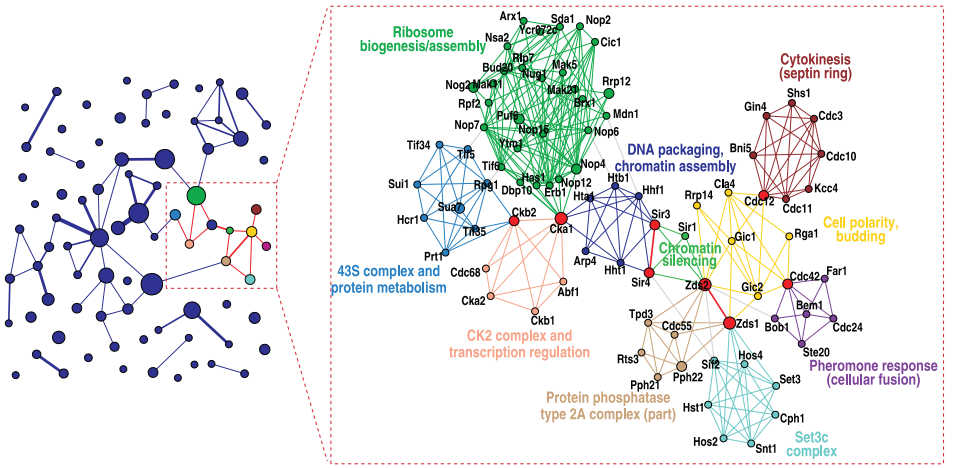

Fig. 8. The community graph at $k = 4$ for the PPI network of *S. cerevisiae* obtained from the DIP core list. The area of a node and the width of an edges are proportional to the size of the corresponding community (number of members) and to the size of the overlap (number of shared nodes), respectively. The coloured communities are cut out and magnified to reveal their internal structure in the left panel. In this magnified picture the nodes and edges of the original network have the same colour as their communities, those that are shared by more than one community are emphasised in red, and the grey edges are not part of these communities. The area of a node and the width of an edge are proportional to the total numberof communities they belong to. Figure from [64].

distribution [6, 43, 53]. In the study of the community structure in a scientific co-authorship network it has been shown that similar processes control the growth of communities and the development of the community graph as well [68].

5.2.2. Molecular biological networks. Over the past decade in biology, especially in Bioinformatics and Systems Biology, the network approach has become very popular and successful [40, 7, 10, 50, 90, 2]. The CPM (together with its free software implementation, CFinder [65, 1], capable of detecting and visualising k-clique percolation clusters) provides a flexible and handy tool for identifying modules in such graphs. In Sect. 5.2.1. we demonstrated the concept of community graphs with the help of a biological example, the PPI network of yeast. The communities detected with the CPM in such networks can be associated mostly with either protein complexes or certain functions [64, 1, 88]. For some proteins no function is available yet. Thus,

the fact that they show up as members of communities can be interpreted as a prediction for their functions.

Moreover, Jonsson et. al. used the CPM to validate the reliability of the connections in a PPI network [45]. The available direct experimental data concerning protein-protein interactions is not equally broad for the different species. Thus, in some cases the construction of the PPI network is based on other methods, too, for example, homology (DNA sequence similarity) between the proteins. The weight of an edge, A-B, in the PPI network is obtained by integrating (e.g. summing) the weights from several sources of evidence: experimentally measured physical interactions, homology and many others. Each experimental and prediction technique can identify different groups of interactions with high efficiency, therefore, to find the largest possible portion of the "real" biological list of interactions it is necessary to integrate data from a large number of sources. The scoring function used to calculate the edge weights in the integrated network can be validated in a number of ways; one is based on the assumption that interactions within densely linked communities are more acceptable than interactions between communities, i.e. a higher score is expected for intra-community connections. This assumption was confirmed by Jonsson et. al. in a study of the rat proteome, where edges inside the CPM communities were observed to be significantly stronger than edges connecting nodes in different communities [45].

In the example above one major goal was the automatic identification of protein communities involved in *cancer metastasis*. In metastasis (a cellular state) cancer cells have the ability to break away from the primary tumour and move to different organs, making the cancer more difficult to treat. Little is known about the molecular biology of metastasis, but it is now broadly accepted that these cells have an increased motility and invasiveness. These novel behaviours involve protein-protein interactions which have to be identified and characterised if an effective treatment is to be developed. The main results of [45] showed that the CPM can help to identify key protein communities involved in cancer metastasis.

A closely related study by Jonsson and Bates was aimed at the investigation of the topological properties of cancer proteins (proteins closely related to the development of cancer) as nodes of the human PPI network [44]. The community structure was examined with the help of the CPM, and the results showed that (among various other topological differences between cancer- and non-cancer proteins) cancer proteins appear in community overlaps more frequently than predicted from their overall

ratio amongst all proteins. Since communities usually represent different cellular processes, proteins in the overlaps may be participating in multiple processes, and can be considered to be at the "interface" of distinct but adjacent cellular processes. Therefore, cancer proteins seem to be mediators between different pathways. In one of the examples presented by Jonsson and Bates, four communities were tied together by cancer proteins with functions ranging from signal transduction to the regulation of cell growth and cell death. Furthermore, the ratio of cancer proteins in the communities was increasing with k, and cancer proteins seemed to take part in larger communities. A plausible explanation of this effect is that cancer proteins participate in more complex cellular processes. It is also conceivable that the larger communities correspond to larger or more complicated cellular machineries, where cancer proteins play a role [44].

A cancer-related investigation using CPM was carried out recently by Finocchiaro et. al. in [28] as well. In this case the network was constructed from gene expression data: groups of up to 10 genes with significant co-expression were fully connected. The communities in the resulting network were extracted using several methods (including CPM), and according to the results, the identified communities were enriched with genes responsible for the regulation of the cell cycle, apoptosis, phosphorylation cascades, extracellular matrix, immune and interferon response regulation. For the majority of communities, promoter searches for enriched cis-regulatory modules support the conclusion that the communities identified here reflect biologically relevant sets of co-regulated genes whose expression is altered in human cancer. As such, the identified communities may provide marker genes useful for clinical applications as well as hitherto unknown regulators of cancer signalling pathways that may constitute novel entry points for pharmacological intervention.

5.2.3. Social networks. The CPM was successfully applied to various networks related to the social contacts between people as well. The study of social networks has a long history; in its early period sociologist used questionnaires and personal interviews to reveal the graph of social ties. The spectrum of social interactions that can be probed in this approach is very wide, however, the size of the sample that can be examined in this way is rather limited. Nowadays, due to the rapid developments in computer technology, new possibilities opened up for the exploration of social ties, enabling the construction of networks on a much larger scale. A prominent example of this is given in [60, 61], where a network consisting of more than

$4 \cdot 10^6$ customers of a mobile phone company is analysed (the edges represent mutual calls between the people).

The community structure of this system was analysed with the help of CPM in [63], and according to the results, the majority of the found communities contained individuals living in the same neighbourhood, and with comparable age, a homogeneity that supports the validity of the uncovered community structure. Interestingly, the time evolution of the small communities (e.g. a smaller collaborative or friendship circles) and the large communities (e.g. institutions) showed major differences. At the heart of small cliques were a few strong relationships, and as long as these persisted, the community around them was stable. It appeared to be almost impossible to maintain this strategy for large communities, however. In contrast, the condition for stability for large communities was continuous changes in their membership, allowing for the possibility that after some time practically all members are exchanged. Such loose, rapidly changing communities are reminiscent of institutions, that can continue to exist even after all members have been replaced by new ones. For example, in a few years most members of a school or a company could change, yet the school and the company will be detectable as a distinct community at any time step throughout its existence. This effect was observed in the community evolution of a co-authorship network as well [63]. (The edges between co-authors in this case corresponded to articles published together).

Another interesting study of social networks was given in [34] by González et. al., investigating the community structure and ethnic preferences in high schools. The friendship networks between the students for 84 schools were constructed from the Add Health database [82], and the communities were extracted using the CPM. The communities at $k = 3$ covered the majority of the students in most of the schools, and the corresponding community graphs showed complex, richly interconnected structures. In contrast, at $k = 4$ the community graphs became rather sparse, and the involved communities covered less than 20% of the students. At the same time, the number of communities belonging to the different ethnic groups became balanced even for cases when the ratio of the sizes of the ethnic groups was far from unity (and, correspondingly, on the level of less cohesive groups, e.g. for $k = 3$, the students who were in majority, had much larger friendship circles). A plausible explanation of this effect is that when in minority, the students tend to form stronger ties, thus, the number of more densely interconnected communities becomes over-represented compared to what happens in the $k = 3$ case [34].

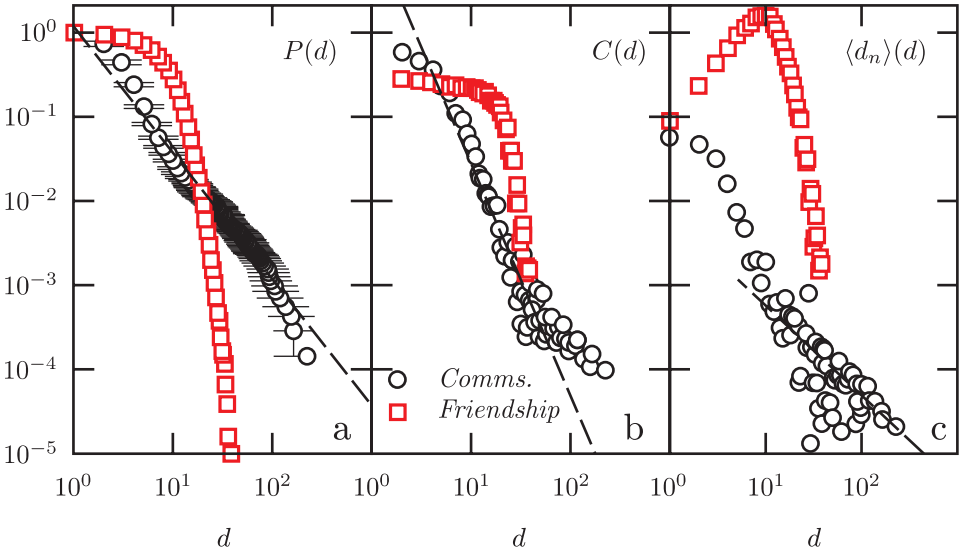

Fig. 9. Different network properties averaged over the community graphs (circles) and the underlying networks of students (squares) in the studied schools. (a) The cumulative degree distribution $P(d)$. (b) The average clustering coefficient C (the fraction of edges between the nearest neighbours of a vertex compared to the number of edges that could possibly exist between them) in function of the vertex degree d. (c) The average degree of the nearest neighbours in function of the degree. Note that the base network shows assortativity (increasing tendency at low degrees), whereas the community graph is disassortative. Figure from [34].

The other important result in this study connected to the CPM is that the graph of communities turned out to be disassortative in spite of the fact that the underlying network of friendship showed assortativity, as demonstrated in Fig. 9c. (A network is said to be assortative if the average degree of the nearest neighbours is increasing with the node degree, i.e. high degree nodes "like" to connect to high degree nodes, and disassortative in the opposite case). This is another indication of the differences in the interactions at different levels in the hierarchy of a complex system.

The results in [60, 61, 34] (described partly in this section) also inspired a couple of new models for the development of social networks [81, 48, 35]. In the works of Toivonen et. al. [81] and Kumpula et. al. [48], the emphasis is on the balance between two different type of attachment mechanisms: cyclic closure and focal closure. The first one corresponds to the formation of new ties or the enhancement of the strength of existing ties within an already densely connected neighbourhood (i.e. two people who have many

common acquaintances will get to know each other as well sooner or later). The second one refers to the formation of ties independently of the geodesic distance and is attributed to shared activities (hobbies, etc.). By changing the relative strength of the two types of attachment mechanisms, the forming network undergoes a transition from a homogeneous state (where the majority of the edges were formed by focal closure) to an inhomogeneous state with apparent, dense communities (in which the edges result mostly from cyclic closure). This transition, and the appearing communities were studied with the help of the CPM [81, 48]. According to the results, by adequately balancing the two types of attachment mechanisms, the statistical properties of the model network match the mobile-phone network in all studied aspects.

In [35] González et. al. introduced a network model based on colliding (finite sized) particles travelling in a finite cell with periodic boundary conditions. Each collision results in a new edge between the involved particles, and the updating of the velocities depend on the degree of the particles. With suitably chosen collision rules and aging scheme (particles die after a certain amount of time) the quasi-stationary states of the resulting network reproduce accurately the main statistical and topological features (e.g. the community size distribution for communities obtained by CPM) of the high school friendship networks mentioned earlier.

We close the overview of the CPM related results in social networks by mentioning the study of the collaboration network among rappers by Smith in [77]. The edges in this network correspond to co-appearance as artists in lyrics obtained from several sources. The community structure of the resulting graph was analysed with several methods including the CPM.

5.2.4. Further results. Finally, we collect a few other results related to the CPM ranging from the investigation of economical networks to the graph of certain webpages. In [41] a subset of the New York Stock Exchange was analysed by Heimo et. al. with both spectral methods and the asset graph method. In the latter case, the asset graph was constructed from the correlation matrix of the stocks: the edges represent correlations stronger than a certain threshold. The emerging graph was studied with the help of the CPM. The results show that the first few eigenvectors of the correlation matrix are localised on the communities, however their borders are fuzzy and do not define clear cluster boundaries. With increasing eigenvector index (the eigenvectors are ordered according to the corresponding eigenvalue), the eigenvectors appear to localise increasingly less regularly with respect

to the asset graph topology. Therefore it appears that identifying the strongly interacting clusters of stocks solely based on spectral properties of the correlation matrix is rather difficult; the asset graph method (coupled with the CPM) seems to provide more coherent results.

Gao and Wong applied the CPM to document clustering in [30]. The graph of the documents was constructed using document similarity (more similar documents are connected by a stronger edge). According to the results, the communities obtained via the CPM can outperform some typical algorithms on benchmark data sets, and shed light on natural document clustering.

An interesting application of the CPM is shown by Castelló et. al. in the study of the dynamics of competing opinions [19]. In the voter model, the state of the agents can be either A or B [42], whereas in the AB model a third, intermediate AB state is included as well [18]. The network of voters is constructed with the help of a variation of the social network models based on the balance between cyclic closure and focal closure [81, 48], briefly discussed in Sect. 5.2.3. At each time step, the state (opinion) of a randomly selected agent is changed with a probability depending on the states of its neighbours. Starting from a random initial opinion distribution, in both models the system converges to consensus, where all nodes are in the same state. However, in the AB model the average time needed to reach the ordered state is highly dependent on the structure of the underlying network. According to the results of Castelló et. al., when a rich, apparent community structure can be detected with the help of the CPM, the lifetime distribution of the meta-stable (disordered) states becomes a power-law, so that the mean lifetime is not representative of the dynamics. These trapped meta-stable states, which can order at all time scales, originate in the mesoscopic network structure.

Finally, we mention the comparative analysis of the directed communities in Google's own webpages (the links follow the direction of the hyperlinks), a word association network (the directions of the links indicate that people in the survey associated the end point of the link with its start point), a university e-mail network (the links point from the sender to the recipient), and a transcriptional regulatory network (the links point from the regulating protein to the protein of the regulated gene) in [67]. The identified directed modules were inherently overlapping and the investigated networks could be classified into two major groups in terms of the overlaps between the modules. In the word association network and Google's webpages overlaps were likely to contain in-hubs, whereas the modules in

the email and transcriptional regulatory networks tended to overlap via out-hubs. In other words, in these two major classes of directed graphs, directed modules "point" towards and away from their shared regions.

References

[1] B. Adamcsek, G. Palla, I. J. Farkas, I. Derényi and T. Vicsek, CFinder: Locating cliques and overlapping modules in biological networks, *Bionformatics,* **22** (2006), 1021–1023.

[2] T. Aittokallio and B. Schwikowski, Graph-based methods for analysing networks in cell biology, *Briefings in Bioinformatics,* **7** (2006), 243–255.

[3] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.,* **74** (2002), 47–97.

[4] A. V. Antonov and H. W. Mewes, Complex functionality of gene groups identified from high-throughput data, *J. Mol. Biol.,* **363(1)** (2006), 289–296.

[5] A.-L. Barabási and R. A. and, Emergence of scaling in random networks, *Science,* **286** (1999), 509–512.

[6] A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek, Evolution of the social network of scientific collaborations, *Physica A,* **311** (2002), 590–614.

[7] A.-L. Barabási and Z. N. Oltvai, Network Biology: Understanding the Cells's Functional Organization, *Nature Reviews Genetics,* **5** (2004), 101–113.

[8] V. Batagelj and M. Zaveršnik, Short cycle connectivity, *Discrete Mathematics,* **307** (2007), 310–318.

[9] J. Baumes, M. Goldberg and M. Magdon-Ismail, Efficient Identification of Overlapping Communities, *Lect. Notes in Computer Science,* **3495** (2005), 27–36.

[10] A. Beyer, S. Bandyopadhyay and T. Ideker, Integrating physical and genetic maps: from genomes to interaction networks, *Nature Reviews Genetics,* **9** (2007), 699–710.

[11] G. Bianconi and M. Marsili, Emergence of large cliques in random scale-free networks, *Europhys. Lett.,* **74** (2006), 740–746.

[12] G. Bianconi and M. Marsili, Number of cliques in random scale-free network ensembles, *Physica D-Nonlinear Phenomena,* **224** (2006), 1–6.

[13] B. Bollobás, *Graph Theory and Combinatorics: Proceedings of the Cambridge Combinatorial Conference in honour of Paul Erdős,* Academic, New York, 1984.

[14] B. Bollobás, *Random graphs,* Cambridge University Press, Cambridge, 2nd edition, 2001.

[15] B. Bollobás and O. Riordan, Clique percolation, arXiv:0804.0867, 2008.

[16] E. I. Boyle, S. A. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry and G. Sherlock, GO::TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes, *Bioinformatics,* **20** (2004), 3710–3715.

[17] S. Carmi, S. Havlin, S. Kirkpatrick, Y. Shavitt and E. Shir, A Model of Internet Topology Using k-shell Decomposition, *Proc. Natl. Acad. Sci. USA,* **104** (2007), 11150–11154.

[18] X. Castelló, V. M. Eguíluz and M. S. Miguel, Ordering dynamics with two non-excluding options: bilingualism in language competition, *New Journal of Physics,* **8** (2006), 308.

[19] X. Castelló, R. Toivonen, V. M. Eguíluz, K. Kaski and M. S. Miguel, Anomalous lifetime distributions and topological traps in ordering dynamics, *Europhys. Lett.,* **7**9 (2007), 66006.

[20] J. M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R. K. Mortimer and D. Botstein, Genetic and physical maps of Saccharomyces cerevisiae, *Nature,* **387**(6632 Suppl) (1997), 67–73.

[21] I. Derényi, G. Palla and T. Vicsek, Clique percolation in random networks, *Phys. Rev. Lett.,* **94** (2005), 160202.

[22] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes, k-core architecture and k-core percolation on complex networks, *Physica D-Nonlinear Phenomena,* **224** (2006), 7–19.

[23] S. N. Dorogovtsev, A. V. Goltsev and J. F. F. Mendes, k-core organization of complex networks, *Phys. Rev. Lett.,* **96** (2006), 040601.

[24] P. Erdős and A. Rényi, On the evolution of random graphs, *Publ. Math. Inst. Hung. Acad. Sci.,* **5** (1960), 17–61.

[25] M. G. Everett and S. P. Borgatti, Analyzing Clique Overlap, *Connections,* **21** (1998), 49–61.

[26] B. S. Everitt, *Cluster Analysis,* Edward Arnold, London, 3th edition, 1993.

[27] I. J. Farkas, D. Ábel, G. Palla and T. Vicsek, Weighted network modules, *New Journal of Physics,* **9** (2007), 180.

[28] G. Finocchiaro, F. M. Mancuso, D. Cittaro and H. Muller, Graph-based identification of cancer signaling pathways from published gene expression signatures using PubLiME, *Nucl. Ac. Res.,* **35** (2007), 2343–2355.

[29] S. Fortunato and M. Barthelemy, Resolution limit in community detection, *Proc. Natl. Acad. Sci. USA,* **104** (2007), 36–41.

[30] W. Gao and K.-F. Wong, Natural document clustering by clique percolation in random graphs, *Lect. Notes in Comp. Sci.,* **4182** (2006), 119–131.

[31] D. Gfeler, J.-C. Chappelier and P. D. L. Rios, Finding instabilities in the community structure of complex networks, *Phys. Rev. E.,* **72** (2005), 056135.

[32] M. Girvan and M. E. J. Newman, Community structure in social and biological networks, *Proc. Natl. Acad. Sci. USA,* **99** (2002), 7821–7826.

[33] A. V. Goltsev, S. N. Dorogovtsev and J. F. F. Mendes, k-core (bootstrap) percolation on complex networks: Critical phenomena and nonlocal effects, *Phys. Rev. E.,* **73** (2006), 056101.

[34] M. C. González, H. J. Herrmann, J. Kertész and T. Vicsek, Community structure and ethnic preferences in school friendship networks, *Physica A-Statistical Mechenics and its Applications,* **379** (2007), 307–316.

[35] M. C. González, P. G. Lind and H. J. Herrmann, System of mobile agents to model social networks, *Phys. Rev. Lett.,* **96** (2006), 088702.

[36] R. Guimerà and L. A. N. Amaral, Functional cartography of complex metabolic networks, *Nature,* **433** (2005), 895–900.

[37] R. Guimerà, S. Mossa, A. Turtschi and L. A. N. Amaral, The worldwide air transportation network: Anomalous centrality, community structure and cities' global roles, *Proc. Natl. Acad. Sci. USA,* **102** (2005), 7794–7799.

[38] R. Guimerà, M. Sales-Pardo and L. A. N. Amaral, Module identification in bipartite and directed networks, *Phys. Rev. E.,* **76** (2007), 036102.

[39] U. Guldener, M. Munsterkotter, G. Kastenmuller, N. Strack and J. van Helden, CYGD: the Comprehensive Yeast Genome Database, *Nucl. Ac. Res.,* **33** (2005), D364–D368.

[40] L. H. Hartwell, J. J. Hopfield, S. Leibler and A. W. Murray, From molecular to modular cell Biology, *Nature,* **402** (1999), 6761, supplement C47–C52.

[41] T. Heimo, J. Saramäki, J.-P. Onnela and K. Kaski, Spectral and network methods in the analysis of correlation matrices of stock returns, *Physica A-Statistical Mechanics and its Applications,* **383** (2007), 147–151.

[42] R. A. Holley and T. M. Liggett, Ergodic theorems for weakly interacting infinite systems and voter model, *Annals of Probability,* **3** (1975), 643–663.

[43] H. Jeong, Z. Néda and A.-L. Barabási, Measuring preferential attachment for evolving networks, *Europhysics Letters,* **61** (2003), 567–572.

[44] P. F. Jonsson and P. A. Bates, Global topological features of cancer proteins in the human interactome, *Bioinformatics,* **22** (2006), 2291–2297.

[45] P. F. Jonsson, T. Cavanna, D. Zicha and P. A. Bates, Cluster analysis of networks generated through homology: automatic identification of important protein communities involved in cancer metastasis, *BMC Bioinformatics,* **7** (2006), 2.

[46] S. Knudsen, *A Guide to Analysis of DNA Microarray Data,* Wiley-Liss, 2nd edition, 2004.

[47] N. J. Krogan, G. Cagney, H. Y. Yu, G. Q. Zhong, X. H. Guo, A. Ignatchenko, J. Li, S. Y. Pu, N. Datta, A. P. Tikuisis, T. Punna, J. M. Peregrin-Alvarez, M. Shales, X. Zhang, M. Davey, M. D. Robinson, A. Paccanaro, J. E. Bray, A. Sheung, B. Beattie, D. P. Richards, V. Canadien, A. Lalev, F. Mena, P. Wong, A. Starostine, M. M. Canete, J. Vlasblom, S. W. C. Orsi, S. R. Collins, S. Chandran, R. Haw, J. J. Rilstone, K. Gandi, N. J. Thompson, G. Musso, P. S. Onge, S. Ghanny, M. H. Y. Lam, G. Butland, A. M. Altaf-Ui, S. Kanaya, A. Shilatifard, E. O'Shea, J. S. Weissman, C. J. Ingles, T. R. Hughes, J. Parkinson, M. Gerstein, S. J. Wodak, A. Emili and J. F. Greenblatt, Global landscape of protein complexes in the yeast Saccharomyces cerevisiae, *Nature,* **440** (2006), 637–643.

[48] J. M. Kumpula, J.-P. Onnela, J. Saramäki, K. Kaski and J. Kertész, Emergence of communities in weighted networks, *Phys. Rev. Lett.,* **99** (2007), 228701.

[49] J. M. Kumpula, J. Saramäki, K. Kaski and J. Kertész, Limited resolution in complex network community detection with Potts model approach, *European Physical Journal B.,* **56** (2007), 41–45.

[50] O. Mason and M. Verwoerd, Graph theory and networks in Biology, *IET Systems Biology,* **1** (2007), 89–119.

[51] J. F. F. Mendes and S. N. D. and, *Evolution of Networks: From Biological Nets to the Internet and WWW,* Oxford University Press, Oxford, 2003.

[52] T. Nepusz, A. Petróczi, L. Négyessy and F. Bazsó, Fuzzy communities and the concept of bridgeness in complex networks, *Phys. Rev. E.,* **77** (2008), 016107.

[53] M. E. J. Newman, Clustering and preferential attachment in growing networks, *Phys. Rev. E.,* **64** (2001), 025102.

[54] M. E. J. Newman, Detecting community structure in networks, *Eur. Phys. J. B.,* **38** (2004), 321–330.

[55] M. E. J. Newman, Fast algorithm for detecting community structure in networks, *Phys. Rev. E.,* **69** (2004), 066133.

[56] M. E. J. Newman and M. Girvan, Finding and evaluating community structure in networks, *Phys. Rev. E.,* **69** (2004), 026113.

[57] M. E. J. Newman and E. A. Leicht, Mixture models and exploratory analysis in networks, *Proc. Natl. Acad. Sci. USA,* **104** (2007), 9564–9569.

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

[59] J.-P. Onnela, A. Chakraborti, K. Kaski, J. Kertész and A. Kanto, Dynamics of market correlations: Taxonomy and portfolio analysis, *Phys. Rev. E.,* **68** (2003), 056110.

[60] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, M. A. de Menezes, K. Kaski, A. L. Barabási and J. Kertész, Analysis of a large-scale weighted network of one-to-one human communication, *New Journal of Physics,* **9** (2007), 179.

[61] J.-P. Onnela, J. Saramäki, J. Hyvönen, G. Szabó, D. Lazer, K. Kaski, J. Kertész and A.-L. Barabási, Structure and tie strengths in mobile communication networks, *Proc. Natl. Acad. Sci. USA,* **104** (2007), 7332–7336.

[62] J.-P. Onnela, J. Saramäki, J. Kertész and K. Kaski, Intensity and coherence of motifs in weighted complex networks, *Phys. Rev. E.,* **71** (2005), 065103.

[63] G. Palla, A.-L. Barabási and T. Vicsek, Quantifying social group evolution, *Nature,* **446** (2007), 664–667.

[64] G. Palla, I. Derényi, I. Farkas and T. Vicsek, Uncovering the overlapping community structure of complex networks in nature and society, *Nature,* **435** (2005), 814–818.

[65] G. Palla, I. Derenyi, I. Farkas, T. Vicsek, P. Pollner and D. Ábel, Free software for finding overlapping dense groups of nodes in networks, based on the clique percolation method.

[66] G. Palla, I. Derényi and T. Vicsek, The critical point of k-clique percolation in the Erdős–Rényi graph, *J. Stat. Phys.,* **128** (2007), 219–227.

[67] G. Palla, I. J. Farkas, P. Pollner, I. Derényi and T. Vicsek, Directed network modules, *New Journal of Physics,* **9** (2007), 186.

[68] P. Pollner, G. Palla and T. Vicsek, Preferential attachment of communities: The same principle, but a higher level, *Europhys. Lett.,* **73** (2006), 478–484.

[69] B. Ráth and B. Tóth, Triangle percolation in mean field random graphs – with PDE, arXiv:0712.2646v1, 2007.

[70] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai and A.-L. Barabási, Hierarchical organization of modularity in metabolic networks, *Science,* **297** (2002), 1551–1555.

[71] J. Reichardt and S. Bornholdt, Detecting Fuzzy Community Structures in Complex Networks with a Potts Model, *Phys. Rev. Lett.,* **93** (2004), 218701.

[72] J. Reichardt and S. Bornholdt, Statistical mechanics of community detection, *Phys. Rev. E.,* **74** (2006), 016110.

[73] A. W. Rives and T. Galitski, Modular organization of cellular networks, *Proc. Natl. Acad. Sci. USA,* **100** (2003), 1128–1133.

[74] J. Scott, *Social Network Analysis: A Handbook,* Sage Publications, London, 2nd edition, 2000.

[75] S. B. Seidman, Network structure and minimum degree, *Social Networks,* **5** (1983), 269–287.

[76] R. M. Shiffrin and K. Börner, Mapping knowledge domains, *Proc. Natl. Acad. Sci. USA,* **101** (2004), 5183–5185.

[77] R. D. Smith, The network of collaboration among rappers and its community structure, *J. Stat. Mech.,* page P02006, 2006.

[78] V. Spirin and K. A. Mirny, Protein complexes and functional modules in molecular networks, *Proc. Natl. Acad. Sci. USA,* **100** (2003), 12123–12128.

[79] G. Szabó and G. Fáth, Evolutionary games on graphs, *Physics Reports-Review Section of Physics Letters,* **446** (2007), 97–216.

[80] G. Szabó, J. Vukov and A. Szolnoki, Phase diagrams for an evolutionary prisoner's dilemma game on two-dimensional lattices, *Phys. Rev. E.,* **72** (2005), 047107.

[81] R. Toivonen, J.-P. Onnela, J. Saramäki, J. Hyvönen and K. Kaski, A model for social networks, *Physica A-Statistical Mechanics and its Applications,* **370** (2006), 851–860.

[82] J. R. Udry, P. S. Bearman and K. M. Harris, Public-use data set from Add Health, funded by a grant from National Institute of Child Health and Human Development.

[83] T. Vicsek, Phase transitions and overlapping modules in complex networks, *Physica A-Statistical Mechanics and its Applications,* **378** (2007), 20–32.

[84] J. Vukov, G. Szabó and A. Szolnoki, Cooperation in the noisy case: Prisoner's dilemma game on two types of regular random graphs, *Phys. Rev. E.,* **73** (2006), 067103.

[85] D. J. Watts, P. S. Dodds and M. E. J. Newman, Identity and search in social networks, *Science,* **296** (2002), 1302–1305.

[86] D. M. Wilkinson and B. A. Huberman, A method for finding communities of related genes, *Proc. Natl. Acad. Sci. USA,* **101** (2004), 5241–5248.

[87] I. Xenarios, D. W. Rice, L. Salwinski and M. K. Baron, DIP: the Database of Interacting Proteins, *Nucl. Ac. Res.,* **28** (2000), 289–291.

[88] S. Zhang, X. Ning and X.-S. Zhang, Identification of functional modules in a PPI network by clique percolation clustering, *Comp. Biology and Chemistry,* **30** (2006), 445–451.

[89] S. Zhang, R.-S. Wang and X.-S. Zhang, Uncovering fuzzy community structure in complex networks, *Phys. Rev. E.,* **76** (2007), 046103.

[90] X. Zhu, M. Gerstein and M. Snyder, Getting connected: analysis and principles of biological networks, *Genes & Development,* **21** (2007), 1010–1024.

Gergely Palla, Illés J. Farkas, Péter Pollner and Tamás Vicsek
Statistical and Biological Physics Research Group of HAS
H-1117 Budapest, Pázmány Péter sétány 1/A
Hungary

Dániel Ábel, Illés J. Farkas, Imre Derényi and Tamás Vicsek
Department of Biological Physics Eötvös University
H-1117 Budapest, Pázmány Péter sétány 1/A
Hungary

vicsek@angel.elte.hu