New Journal of Physics
The open–access journal for physics

Weighted network modules

Illés J Farkas[1,2], **Dániel Ábel**[1], **Gergely Palla**[1,2] **and Tamás Vicsek**[1,2,3]

[1] Department of Biological Physics, Eötvös University, Budapest, H-1117, Hungary
[2] Statistical and Biological Physics Group, ELTE-HAS, Pázmány P. Stny. 1A, Budapest, H-1117, Hungary
E-mail: vicsek@angel.elte.hu

New Journal of Physics **9** (2007) 180
Received 2 March 2007
Published 28 June 2007
Online at http://www.njp.org/
doi:10.1088/1367-2630/9/6/180

Abstract. The inclusion of link weights into the analysis of network properties allows a deeper insight into the (often overlapping) modular structure of real-world webs. We introduce a clustering algorithm clique percolation method with weights (CPMw) for weighted networks based on the concept of percolating k-cliques with high enough intensity. The algorithm allows overlaps between the modules. First, we give detailed analytical and numerical results about the critical point of weighted k-clique percolation on (weighted) Erdős–Rényi graphs. Then, for a scientist collaboration web and a stock correlation graph we compute three-link weight correlations and with the CPMw the weighted modules. After reshuffling link weights in both networks and computing the same quantities for the randomized control graphs as well, we show that groups of three or more strong links prefer to cluster together in both original graphs.

[3] Author to whom any correspondence should be addressed.

PII: S1367-2630(07)45454-4

Contents

1. Introduction

Networks provide a ubiquitous mathematical framework for the analysis of natural and man-made systems [1]–[5]. They allow one to picture, model and understand in a simple and rather intuitive way the high diversity of phenomena ranging from technological webs [6] to living cells [7], ecological interactions [8] and our societies [9]. The key to the applicability of the network approach is one's ability to dissect the phenomenon under analysis into a list of meaningful interacting units connected by pairwise connections.

Over the past decade several fields of science have been reshaped by a flood of strongly structured experimental information. Due to this transition, algorithms extracting compact, informative statements from measured data receive steadily increasing attention: among such techniques the clustering of data points has become a widely used one [10]. In networks clustering methods locate network modules [11] (also called clusters or communities), i.e. internally densely linked groups of nodes, and lead the observer intuitively to a transformation replacing the original network by its modules. The resulting web of modules contains 'supernodes' (the modules) and a link between two supernodes, if the corresponding modules of the original network are linked [11] or overlap [12]. Interestingly, this mapping resembles a renormalization step from statistical physics [13]. Recent practical applications of network clustering techniques include the grouping of titles in a web of co-purchased books (each cluster represents a topic) [14], the description of cancer-related protein modules in a web of protein–protein interactions [15] and in stock correlation graphs the identification of business sectors or the analysis of links between different sectors [16, 17].

A major success of the network approach to the analysis of large complex systems has been its ability to pinpoint key local and global characteristics based on not more than the bare list of

interactions. This list is a 'plain' graph, i.e. it describes nodes and links without any additional properties, and has been often referred to as the topology of interactions or the static backbone of the underlying complex system. The most pronounced and widely observed static features are the small-world property [1], the scale-free degree distribution [18] and overrepresented small subgraphs (motifs) [19]. In addition, correlations between neighbouring degrees were found to define distinct types of real-world webs [20, 21]. However, several important aspects of the investigated systems can be described only by incorporating additional measurables, e.g. link weights [22]–[25], link directions [26, 27] or node fitness [28, 29] into the models. Examples for the use of these characteristics are large-scale tomographic measurements of the Internet identifying heavily congested sections together with possible alternative routes [30] and the decomposition of multi-million social webs into groups of individuals with common activity patterns [31].

The additional graph property often providing the deepest insight into the dynamical behaviour is the weight of links. In the Internet and transportation webs link weights describe traffic [6, 22], in social systems they represent the frequency and intensity of interactions [9, 32, 33] and in metabolic networks they encode fluxes [34]. Generalizations of several graph properties to the weighted case have revealed that, e.g. in air transportation webs strong links tend to connect pairs of hubs, while in scientific collaboration graphs the degree of a node (number of co-workers) has almost no influence on the average weight of the node's connections (co-operation intensities) [22]. In [35] motifs were generalized to the weighted case using the geometric mean of a subgraph's link weights. With this definition the total intensity of triangles, i.e. a generalized clustering coefficient, was successfully applied for a weighted net of New York Stock Exchange (NYSE) stock correlations to find the structural characteristics and precise time of a major crash. Global modelling approaches to weighted graphs include a weight-driven preferential attachment growth rule [23] and the embedding of nodes into Euclidean space [36]. As for weighted correlation functions, in empirical networks they often depend both on the unweighted link structure (the backbone) and the distribution of weights on these links. Maximally random weighted networks [25] provide a null model to separate these two effects.

As a step towards the characterization of the modules of complex networks, we introduce in this paper a clustering algorithm locating overlapping modules in weighted graphs (nodes connected with weighted links). This technique, that we call the clique percolation method with weights (CPMw), extends the (unweighted) CPM [37] by applying the concept of subgraph intensity [35] to k-cliques (fully connected subgraphs on k nodes). Similarly to the CPM, by definition the CPMw permits overlaps between the modules, a property increasingly recognized in several types of complex networks [38]–[40]. To illustrate the use of the CPMw, we compute the weighted modules of two empirical networks and investigate the correlation properties of their link weights. Also, we provide detailed analytical and numerical results for the percolation of k-cliques with intensities above a fixed threshold, I, in the weighted Erdős–Rényi (ER) graph.

2. Definitions

2.1. Local properties and correlations

Probably, the most basic properties of a node (i) in a weighted network are its degree, d_i (number of neighbours), and its strength, s_i (sum of link weights). In several real systems node degrees

(or strengths) are correlated: the network is assortative if adjacent nodes have similar degrees, and it is disassortative if adjacent nodes have dissimilar degrees. The correlation between link weights can be studied in a very similar way. Two links are adjacent if they have one end node in common, and link weights are assortative (disassortative) in a network if the weights of neighbouring links are correlated (anti-correlated). Moving from pairs of links to triangles, one can quantify the assortativity of link weights in triangles (with nodes i, j and k) by measuring the weight of a link, $w_{i,j}$, as a function of the geometric mean of the other two links' weights, $w_{i,k}$ and $w_{j,k}$:

$$w_{i,j} = F\left([w_{i,k}w_{j,k}]^{1/2}\right). \tag{1}$$

If the link weights in a triangle are similar (or very different), then F is an increasing (or decreasing) function. This definition is closely related to the intensity, $I(g)$, of a subgraph, g, defined as the geometric mean of its link weights [35].

2.2. CPM

In many complex networks internally densely connected groups of nodes (also called modules, clusters or communities) overlap. The importance of module overlaps is illustrated in figure 1. A recently introduced, link density-based module finding technique allowing module overlaps is the CPM [41].

The strongest possible coupling of k nodes with unweighted links is a k-clique: the $k(k-1)/2$ possible pairs are all connected. However, natural and social systems are inherently noisy, thus, when detecting network modules, one should not require that all pairs be linked. In any k-clique a few missing links should be allowed. Removing one link from a $(k+1)$-clique leads to two k-cliques sharing $(k-1)$ nodes, called two *adjacent* k-cliques. Motivated by this observation, one can define a k-clique *percolation cluster* as a maximal set of k-cliques fully explorable by a walk stepping from one k-clique to an adjacent one. In the CPM modules are equivalent to k-clique percolation clusters and overlaps between the modules are allowed by definition (one node can participate in several k-clique percolation clusters).

With the help of the CPM, one can define in a natural way the web of modules as well. In this web, the nodes represent modules and two nodes are linked, if the corresponding modules overlap. In addition, the CPM has been successfully applied to, e.g. tracing the evolution of a social net with over four million users [31] and for highlighting which proteins—beyond the already characterized ones—are possibly involved in the development of certain types of cancer [15].

2.3. The CPMw networks

The search method described in the previous section is applicable to binary graphs only (a link either exists or not). Therefore, in weighted networks the CPM has been used to search for modules by removing links weaker than a fixed weight threshold, W, and considering the remaining connections as unweighted. Here we introduce an extension of CPM that takes into account the link weights in a more delicate way by incorporating the subgraph intensity defined in [35] into the search algorithm. As mentioned in section 2.1, the intensity of a subgraph is equal to the geometric mean of its link weights. In the CPMw approach we include a k-clique into a module only if it has an intensity larger than a fixed threshold value, I.

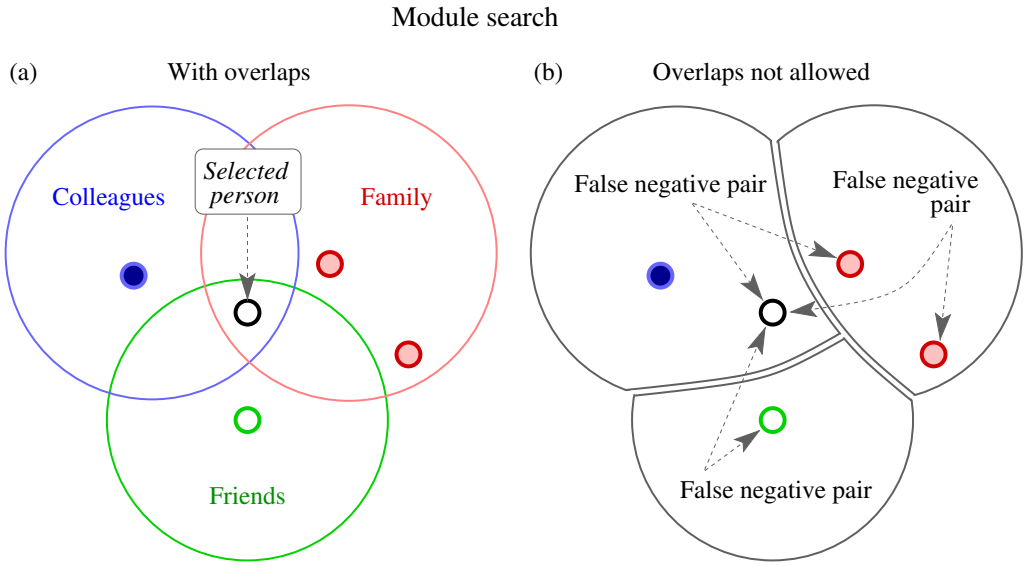

Figure 1. Schematic illustration of the difference between module search methods. Divisive module search techniques do not allow a node to belong to more than one group, which can produce a classification with high numbers of false negative pairs. Algorithms allowing overlaps between the modules can significantly reduce this problem. (a) Example for the overlapping social groups of a selected person. (b) Network modules around the same person as identified by several divisive clustering techniques. Observe the occurrence of false negative pairs.

A k-clique, $\mathcal{C}$, has $k(k-1)/2$ links among its nodes (i, j) and its intensity can be written as

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

Note that this definition is conceptually different from using a simple link weight threshold and then the original CPM. Most importantly, here we allow k-cliques to contain links weaker than I as well.

The k-clique adjacency in the CPMw is defined exactly the same as in the CPM: two k-cliques are adjacent if they share $k-1$ nodes. Finally, a weighted network module is equivalent to a maximal set of k-cliques, with intensities higher than I, that can be reached from each other via series of k-clique adjacency connections.

2.4. Comparing the CPM and the CPMw

The most important difference between the CPM and CPMw is that all links included in a CPM module must have weights higher than the link weight threshold W. However, the modules obtained by the CPMw often contain links weaker than the intensity threshold, I, too. In a weighted network where strong links prefer to be neighbours, the above two algorithms provide

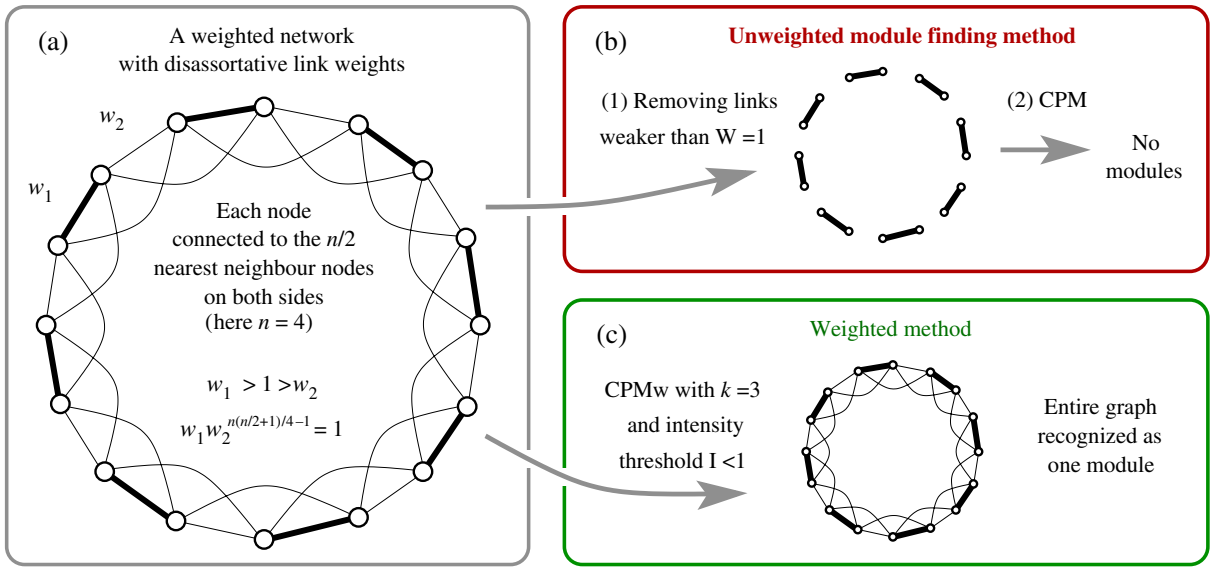

Figure 2. In weighted networks with disassortative link weights, i.e. where strong links tend to have weak links as neighbours, the results of unweighted and weighted module finding can differ strongly. (a) Sample network with equal node degrees, $d = n$, and node strengths, $s = w_1 + (n - 1)w_2$. Each strong connection (w_1) has only weak (w_2) links as neighbours. (b) The unweighted module finding method consists of two steps and finds no modules in the example network. (1) Links weaker than the selected threshold, $W = 1$ in this case, are deleted. (2) Applying the (unweighted) CPM to the remaining links. (c) The CPMw keeps all links and finds one module containing all nodes of the sample graph.

similar results. Note, however, that the edges discarded by the first method (weight cut + CPM) are often registered (measured) to be weaker than W only because of the inherently high noise level of the investigated complex system. In comparison, the CPMw with an intensity threshold $I = W$ is more permitting and produces modules with 'smoother' contours. It expands slightly the modules located by the CPM and may attach to each module additional k-cliques containing weaker links.

Results from the CPM and the CPMw differ strongly for graphs where strong links prefer to have weak links as neighbours, i.e. links are disassortative with respect to their weights. The assortativity of neighbouring node degrees (or strengths) and that of adjacent link weights are conceptually different measures in a network. For example, consider a circular path with an even number of nodes and alternating w_1, w_2 link weights ($w_1 > 1 > w_2$; $w_1 w_2^{n(n/2+1)/4-1} = 1$; $n = 4, 6, \ldots$) and add weaker (w_2) connections between 2nd, 3rd, ... , ($n/2$)th neighbour nodes (see figure 2). In this graph node degrees and node strengths are neither assortative nor disassortative. Each node has a degree $d = n$ and a strength $s = w_1 + (n - 1)w_2$. However, the strong edges (w_1) have exclusively weak (w_2) neighbours, therefore, link weights are clearly disassortative. With clique size and intensity threshold parameters $k = n$ and $I < 1$ the CPMw recognizes the entire graph as one weighted module (figure 2(c)). The corresponding unweighted search finds no modules: If all links with weights below the link weight threshold $W = 1$ are removed, then the remaining links will be all isolated and the CPM finds no modules (figure 2(b)).

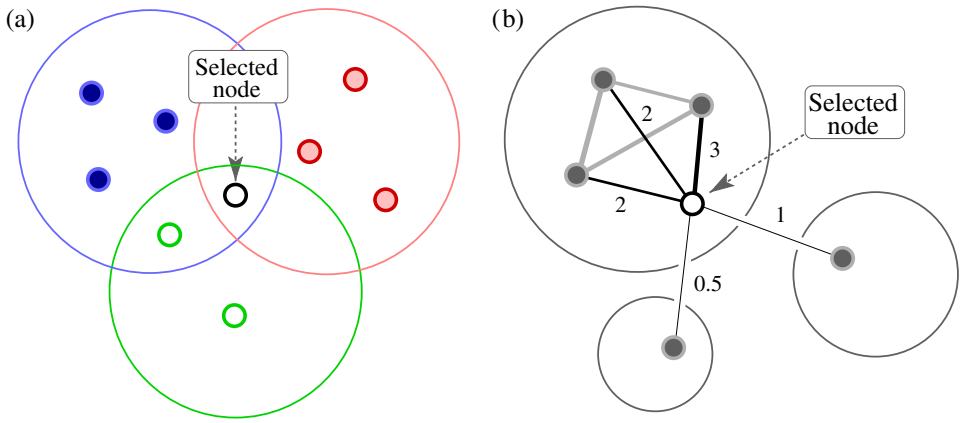

Figure 3. Schematic illustrations of further module-related quantities. (a) The selected node participates in 3 modules, i.e. its module membership number is $m_i = 3$. The total number of its module neighbour nodes is $t_i = 8$. (b) The sum of link weights (strength) connecting the selected node to its module neighbours is $s_{i,\mathrm{in}} = 7$ and the total weight of links connecting it to other modules' nodes is $s_{i,\mathrm{out}} = 1.5$.

2.5. *Further module-related definitions*

The number of modules that the ith node is contained by is called the node's module membership number (m_i) [12]. We define here the *module neighbours* of the ith node as the set of nodes contained by at least one of the modules of that node and we will denote the number of module neighbours by t_i. The total weight of links (strength) connecting the ith node to module neighbours is $s_{i,\mathrm{in}}$ and the total weight of links connecting the same vertex to nodes in other modules is $s_{i,\mathrm{out}}$. See figure 3 for illustrations.

2.6. *Selecting the parameters of the CPMw in real-world graphs*

The CPMw has two parameters: k (clique size) and I (intensity threshold). The optimal choice of k and I is the one with which the CPMw detects the richest structure of weighted modules. Here we discuss this condition from the statistical physics point of view.

Consider a fixed k-clique size parameter, k, and a weighted graph with link weights $w_1 \geqslant w_2 \geqslant \ldots \geqslant w_L$. If $I > w_1$, then the intensity of each k-clique is below the threshold, therefore no weighted modules are found. If, however, $I < w_L$, then any k-clique fulfils the condition for the intensity in the CPMw. In this case often one can observe a very large weighted module (a giant cluster) spreading over the major part of the network. The emergence of this giant module (when lowering I below a certain critical value) is analogous to a percolation transition. The optimal value of I is just above the critical point: on the one hand, the threshold is low enough to permit a huge number of k-cliques to participate in the modules, resulting in a rich module structure. On the other hand, we prohibit the emergence of a giant module that would smear out the details of smaller modules. At the critical point the size distribution of the modules, $p(n_\alpha)$ is broad, usually taking the form of a power-law, analogously to the distribution of cluster sizes at the transition point in the classical edge percolation problem on a lattice.

When I is below the critical point, the size of the largest module, n_1, 'brakes away' from the rest of the size-distribution and becomes a dominant peak far from the rest of distribution $p(n_\alpha)$. This effect allows one to determine the optimal I parameter in a rather simple way. One should start with the highest meaningful value of $I = w_1$ and then lower I until the ratio of the two largest module sizes, n_1/n_2 reaches 2. However, for small networks this ratio can have strong fluctuations, therefore, in such cases it is preferable to determine the transition point by using $\chi = \sum_{n_\alpha \neq n_{\max}} n_\alpha^2/(\sum_\beta n_\beta)^2$, which is similar to percolation susceptibility. To find the weighted modules of real-world graphs (section 4), we first identified separately for each fixed k the optimal I value and then we selected the k parameter with the broadest $p(n_\alpha)$ distribution at its optimal I.

3. Percolation threshold of weighted k-cliques in ER graphs

An (unweighted) ER graph with N nodes has $N(N-1)/2$ possible links, each filled independently with probability p. To obtain a weighted ER graph, we assign to each link (i, j) a weight, w_{ij}, picked independently and randomly from a uniform distribution on the interval $(0, 1]$. Similarly to the previous section, we denote by I the intensity threshold. At a fixed I, the critical link probability, $p_C(I)$, of k-clique percolation is the link probability where a giant module (containing k-cliques fulfilling the intensity condition) emerges. A special case is $I = 0$, i.e. k-clique percolation on ER graphs without weights, for which the critical link probability can be written as [41]

$$p_C(I = 0) = [(k-1)N]^{-1/(k-1)}. \tag{3}$$

3.1. Analytical results

Below we show three analytical approximations for the critical point of clique percolation at $I > 0$. The first is an upper bound obtained by link removal, while the second and third are (cluster) mean-field methods.

3.1.1. Upper bound by link removal. Consider a weighted ER graph, $\mathcal{G}$, with link weights as above and remove all of its links weaker than I. The edges of the truncated weighted graph, $\mathcal{G}^*$, form an unweighted ER network with link probability $p^* = p(1-I)$. As already noted, the intensity of a k-clique can exceed I even when it contains links that are weaker than I. This link removal step discards a finite portion of the k-cliques $\mathcal{C}$ having $I_\mathcal{C} > I$ from the giant (percolating) cluster of $\mathcal{G}$, and changes the percolation threshold to $p_C^*(I) > p_C(I)$. In $\mathcal{G}^*$ there are no link weights below I, therefore, the list of k-cliques with an intensity above I is identical to the list of all unweighted k-cliques. In other words, the critical point of k-clique percolation in $\mathcal{G}^*$ is the same for any value of the intensity threshold between 0 and I. Specifically, $p_C^*(I) = p_C^*(0)$. Moreover, the link deletion step keeps a random $1-I$ portion of all links from $\mathcal{G}$ and modifies the unweighted percolation threshold from $p_C(0)$ to $p_C^*(0) = p_C(0)/(1-I)$. Combining the above gives the following upper bound for $p_C(I)$:

$$p_C(I) < p_C^*(I) = p_C^*(0) = \frac{p_C(0)}{1-I}. \tag{4}$$

3.1.2. Branching process, intensity condition for child k-cliques In the second approximation, we treat the percolation of k-cliques fulfilling the intensity condition as a branching process visiting k-cliques via k-clique adjacency connections. We investigate one branching event: having arrived at a k-clique (parent), we try to move on to further ones fulfilling $I_C > I$ as well (children). Consider one of these child k-cliques and assume that the probability distribution of each link weight in the parent k-clique is the original uniform distribution on the interval (0, 1]. (The actual probability distribution of a link weight in the parent k-clique is different from this.)

The expected number of all neighbouring k-cliques, including those with intensities below I, is $p^{k-1}N(k-1)$ in the large N limit. Now apply the intensity condition (section 2.3) to each child k-clique separately: we denote by $\mathcal{P}_k(<1)$ the probability that the child k-clique has an intensity larger than I. With this notation the expected number of accepted child k-cliques available at the current branching step is $p^{k-1}N(k-1)\mathcal{P}_k$. On the other hand, being at the critical point means that the expectation value of this number should be one. In summary, compared to the $I = 0$ (unweighted) case, we get the following approximation:

$$p_C(I) \simeq p_C(0)\mathcal{P}_k^{-1/(k-1)}, \tag{5}$$

where $\mathcal{P}_k$ is the probability that the product of $k(k-1)/2$ independent link weights, with uniform distribution on (0, 1], reaches $A = I^{k(k-1)/2}$. For $k = 3$ and 4, the $\mathcal{P}_k$ probabilities are

$$\begin{aligned}\mathcal{P}_3 &= \int_A^1 \mathrm{d}w_3 \int_{A/w_2}^1 \mathrm{d}w_2 \int_{A/w_3w_2}^1 \mathrm{d}w_1 = 1 - A\left(1 - \ln A + \frac{\ln^2 A}{2}\right),\\ \mathcal{P}_4 &= \int_A^1 \mathrm{d}w_6 \int_{A/w_6}^1 \mathrm{d}w_5 \ldots \int_{A/w_6\ldots w_2}^1 \mathrm{d}w_1 = 1 - A\sum_{i=0}^{5}\frac{(-\ln A)^i}{i!}.\end{aligned} \tag{6}$$

In summary, the transition point, $p_C(I)$, can be approximated in the $k = 3$ and 4 cases (with $n = k(k-1)/2$) as

$$\left.\frac{p_C(I)}{p_C(0)}\right|_{k=3,4} \simeq \left[1 - I^n \sum_{i=0}^{n-1}\frac{(-n\ln I)^i}{i!}\right]^{-1/(k-1)}. \tag{7}$$

3.1.3. Branching process, child and first parent k-cliques We improve the previous approximation and modify $\mathcal{P}_k$ by taking into account that the parent k-clique has an intensity above I. Due to this condition the distributions of the $(k-1)(k-2)/2$ link weights in the overlap (connecting the $(k-1)$ shared nodes of the parent k-clique and its child) are not independent from each other. The distribution density of the product, t, of these link weights is

$$\tilde{p}_k(t) = \frac{f_k(t)}{C} = \frac{1}{C}\int_{A/t}^1 \mathrm{d}w_1 \int_{A/(tw_1)}^1 \mathrm{d}w_2 \ldots \int_{A/(tw_1\ldots w_{k-2})}^1 \mathrm{d}w_{k-1} \tag{8}$$

Each of the integrations is an averaging for one of the $k-1$ links of the parent k-clique not contained by the overlap. The normalization constant is $C = \int_A^1 \mathrm{d}t\, f_k(t)$. To compute the

probability that the child k-clique's intensity is above I, the same integrations should be performed for the $k-1$ links of the child k-clique outside the overlap. Therefore, we get

$$\frac{p_C(I)}{p_C(0)} \simeq \mathcal{P}_k^{-1/(k-1)} = \left(\frac{\int_A^1 \mathrm{d}t f_k(t)}{\int_A^1 \mathrm{d}t f_k^2(t)} \right)^{1/(k-1)} . \tag{9}$$

Again, as an example, we have performed the integrals and computed $p_C(I)$ for $k = 3$ and 4:

$$f_3(t) = 1 - \frac{A}{t}\left(1 - \ln\frac{A}{t}\right),$$

$$f_4(t) = 1 - \frac{A}{t}\left(1 - \ln\frac{A}{t} + 12\ln^2\frac{A}{t}\right) = 1 - \frac{A}{t}\sum_{i=0}^{2}\frac{(-\ln A/t)^i}{i!},$$

which gives

$$\frac{p_C(I)}{p_C(0)}\Big|_{k=3} \simeq \left[\frac{1 - I^3(1 - 3\ln I + 9/2\ln^2 I)}{1 + I^3[4 - 5I^3 + 6(1 + 2I^3)\ln I - 9(1 + I^3)\ln^2 I]}\right]^{1/2} \tag{10}$$

and

$$\frac{p_C(I)}{p_C(0)}\Big|_{k=4} \simeq \left[\frac{F_4(I)}{G_4(I)}\right]^{1/3}, \tag{11}$$

where

$$F_4(I) = 1 - I^6\left[1 - 6\ln I + 18\ln^2 I - 36\ln^3 I\right],$$

$$G_4(I) = 1 + I^6[18 - 19I^6 + 12(1 + 9I^6)\ln I - 36(1 + 8I^6)\ln^2 I + 72(1 + 6I^6)\ln^3 I - 324I^6\ln^4 I]. \tag{12}$$

3.2. Numerical results

We generated weighted ER graphs as described above, and extracted the k-clique percolation clusters emerging from the k-cliques fulfilling the intensity condition for several threshold (I) and clique size parameter (k) values. Denoting again by n_α the number of nodes in a module (percolation cluster) $n_1 \geqslant n_2 \geqslant \ldots$, we used as an order parameter the relative number of nodes in the largest module:

$$\Phi = \frac{n_1}{\sum_\alpha n_\alpha}. \tag{13}$$

It is known that in the classical ER link percolation problem below the critical link probability, p_C, all clusters contain significantly fewer nodes than the total (N), while above p_C there is one module with size $\mathcal{O}(N)$ and all others are much smaller [42]. One can measure the transition point between these two regimes in several ways that are equivalent in the large system size limit.

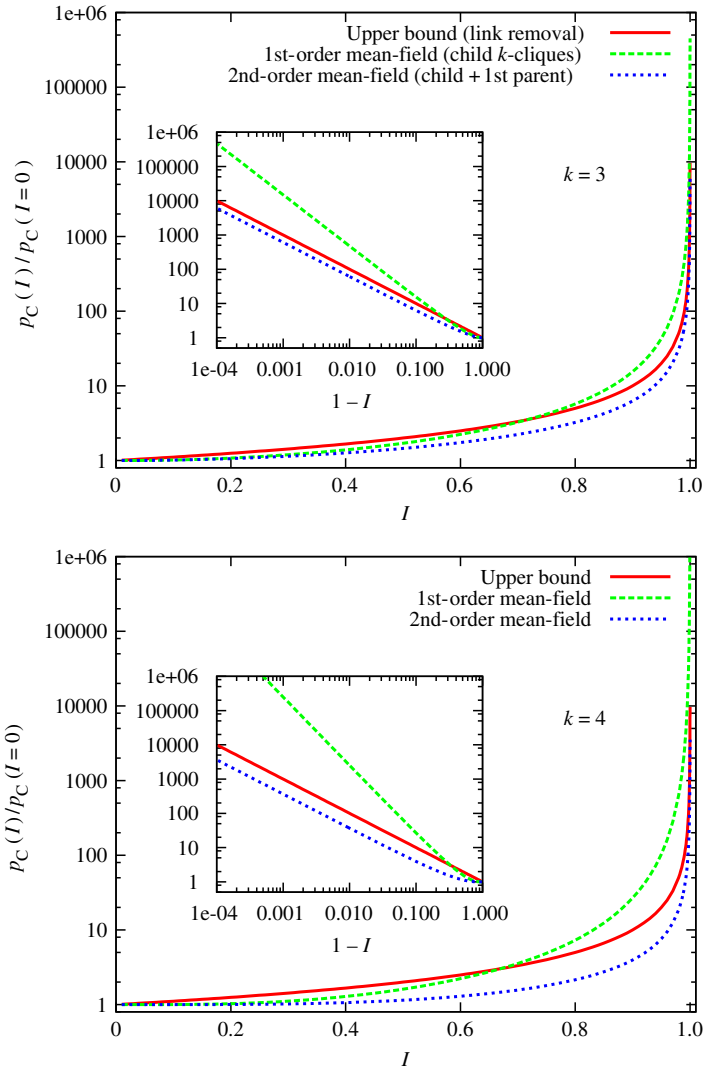

Figure 4. Main panels. Analytical approximations for the critical link probability, $p_C(I)$, of k-clique intensity percolation in weighted ER graphs as a function of the intensity threshold, I (see text for details). Clique size parameters are $k = 3$ (top) and $k = 4$ (bottom). We plotted the ratio between $p_C(I)$ and the critical link probability, $p_C(0)$, of clique percolation without weights [41]. In the ER graph each link is filled with probability p and link weights are randomly and uniformly selected from the interval (0, 1]. Insets: the same curves transformed. At low I the first-order (dashed green) and second-order mean-field (dotted blue) approximations are below the upper bound (solid red), while for $I \to 1$, the first order approximation diverges faster than the strict upper bound. We suggest that for each k increasing the precision of the approximations in section 3.1 (to 3rd, 4th, etc order) will make the solution converge to the exact one. We predict that for the exact solution $p_C(I)/p_C(0)$ diverges as $(1 - I)^{-1}$ when $I \to 1$.

Here we decided to identify the critical point as the link probability where the order parameter, Φ, becomes $1/2$. Figure 4 shows our numerical results for the critical point of intensity k-clique percolation in ER graphs and a comparison with the analytical result from section 3.1.3. To quantify the distance between the numerical and analytical results, we computed the difference integral, D, between the two curves. With growing system size D decreases indicating that the second-order approximation converges to the actual transition curve, $p_C(I)$.

Compared to our generic CPMw search method, the numerical work presented in this section was accelerated by a factor of $\approx$100 with the help of two algorithmic improvements constructed for this purpose. We computed the order parameter, Φ, in all > 1000 points of a grid on the (p, I) plane (figure 5). Depending on the total number of nodes, N, we used in each grid point 3–100 samples (weighted ER networks).

The first algorithmic improvement was based on the observation that for a fixed graph and a fixed clique size parameter, k, the weighted modules at two intensity thresholds ($I_1 > I_2$) differ only in the k-cliques with intensities between I_1 and I_2. Recall that the weighted modules at I_1 (or I_2) contain the k-cliques with intensities above I_1 (I_2). Knowing all k-cliques with intensities above I_1, one can compute the weighted modules for the threshold I_2 by adding k-cliques between I_1 and I_2 and then assembling the percolation clusters of k-cliques. Thus, to find the weighted modules in a given ER graph at each of the intensity threshold values $I_1 > I_2 > \ldots > I_n$, one does not need to perform the entire CPMw and consider all k-cliques again at each I_i. We first listed all k-cliques with intensities above I_n, and then sequentially inserted them (into an empty graph) in the descending order of their intensities. Whenever we reached an I_i threshold, we assembled the weighted modules based on those already computed for the previous threshold, I_{i-1} in an analogous way to the Hoshen–Kopelman algorithm [43]. During the process of inserting k-cliques if the size of the largest module reached N, i.e. the order parameter, Φ, became 1, then we set $\Phi = 1$ for all lower I_i thresholds and proceeded to the next parameter set.

The second algorithmic improvement allowed us to find k-clique adjacencies in shorter time and thereby to assemble the percolation clusters of k-cliques faster. If a k-clique overlaps with another k-clique, then they share one of the $(k-1)$-cliques contained by the first. Thus, we listed the $(k-1)$-cliques occurring in all considered k-cliques, and for each we listed its containing k-clique(s). More than one containing k-clique for a $(k-1)$-clique means that the containing k-cliques are all pairwise adjacent. Note also that all k-clique adjacency connections can be located this way.

4. Results for real-world graphs

As opposed to the ER model, in real-world graphs local properties (e.g. node degree, strength and link weight) are often correlated giving rise to small-, intermediate- and large-scale network structures. Below, we analyse link weight correlations and the structure of weighted modules in two types of real webs. The first is a social (scientific co-authorship) net and the second is a set of two stock correlation graphs.

4.1. Scientific co-authorship network (SCN)

Social networks were among the first few where the small-world [44] and scale-free [18] properties were observed. Since then several models have been constructed to describe these

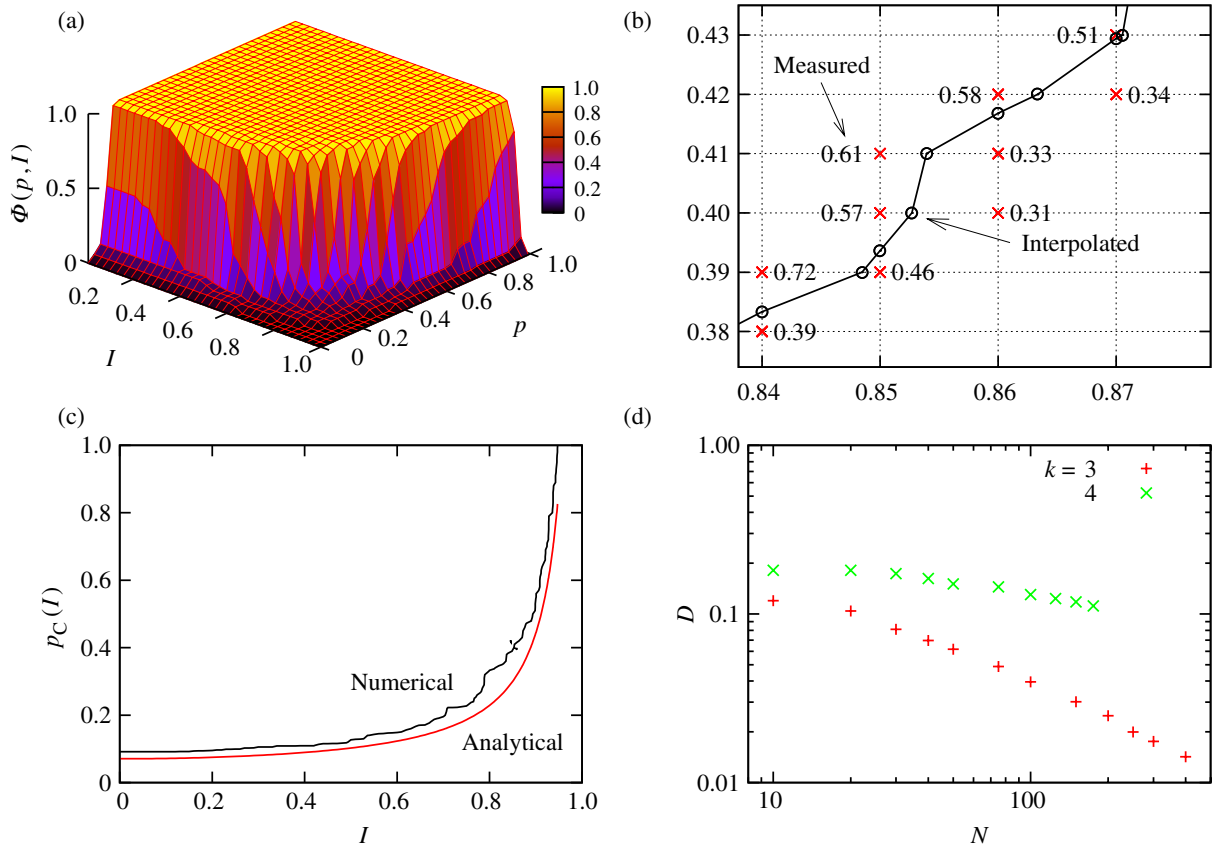

Figure 5. Numerical analysis of the percolation of k-cliques fulfilling the intensity condition in weighted ER graphs. The sample numerical results shown in panels (a)–(c) were obtained for $N = 100$ and $k = 3$ using 1 run for each (p, I) grid point. In panel (d) points were computed from 3 to 100 runs for each (k, I) parameter pair and error bars are smaller than the sizes of the symbols. (a) The order parameter, $\Phi = n_1/\sum_\alpha n_\alpha$, 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 from section 3.1.3. 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.

and further characteristics [1, 18] and some of the microscopic rules of the models have been verified by direct measurements on real graphs [45]. Scientific collaboration networks, as webs of professional contacts, are usually 'measured' through lists of joint publications. Here we consider the weighted co-authorship network of researchers appearing on the 50 634 e-prints of the Los Alamos cond-mat archive [46] between April 1992 and February 2004. In this graph a paper with r authors contributes by $1/(r-1)$ to the weight of the link connecting any two of its authors (nodes) and thus, the strength of a node is equal to the number of papers of the author. In the resulting weighted co-publication graph there are 31 319 non-isolated nodes with

136 065 links between them; these nodes have an average degree (collaborator number) of 8.69 and an average strength (paper number) of 4.47.

Several correlation properties of the SCN (both unweighted and weighted) are well-known from previous studies. As for the unweighted case, node degrees are assortative and the clustering coefficient is high [32]. Moreover, nodes with the highest degrees tend to form so-called rich-clubs [21, 47], i.e. they are more likely to be linked to each other than in the corresponding fully uncorrelated (ER) model. The weighted correlation measures of the SCN analysed so far have been 2- and 3-point correlation functions, which were found to be influenced mainly by the positions of the graph's links, but not the weights of the links [22, 25]. The expected weight of a link is almost independent from its end point degrees. Weighted nearest-neighbour degree correlations and weighted clustering coefficients have highly similar distributions to the analogous unweighted quantities both as a function of node degree and strength. The difference between the investigated weighted and unweighted measures was found to be much smaller in the SCN than in other types of real webs, e.g. air transportation and trade networks.

Here we show that there are correlation properties of the SCN significantly influenced by the links' weights, not only by the positions of the links. The information contained by the link weights can be decomposed into two parts. The first is the (heavy-tailed) distribution of the weights and the second is how these numbers are arranged on the links of the underlying unweighted graph. We constructed a randomized null model, a control graph, of the SCN. We kept the positions of links (a list of node pairs) and the list of link weights (non-negative numbers) unchanged and shuffled the weights on the links of the graph. Comparing the SCN to its control graph, we found a strong assortativity of link weights in triangles (figure 6(a)): two links with high weights have a third neighbouring link with a high weight, too.

The tendency of high link weights to stay close to each other can be measured for groups containing more than three links as well. A standard tool for analysing such correlations is provided by enumerations methods listing each possible subgraph of a fixed size. Along this approach, we used the CPMw to compute the weighted overlapping modules for the SCN and its randomized counterpart, and inferred link weight correlation properties by comparing the sizes of the obtained modules in the two systems. The optimal intensity threshold and k-clique size parameters for the SCN were found to be $I = 0.439$ and $k = 4$. The largest weighted module contained $n_1^{(\mathrm{SCN})} = 714$ authors, whereas in case of the randomized graph (at the same I, k parameters) we observed $n_1^{(\mathrm{rnd})} = 1946$ (figure 6(b)). The $n_1^{(\mathrm{SCN})} < n_1^{(\mathrm{rnd})}$ relation indicates that large link weights cluster together more strongly in the largest component of the SCN than expected by chance: the more closely large ($w_{ij} > I$) link weights cluster together, the smaller the number of k-cliques will fulfil the intensity condition and the smaller the largest weighted module becomes. For comparison, we computed the modules of the original CPM in the SCN as well, at the same k-clique size ($k = 4$) and a link-weight threshold $W = I$. About 32% of the CPM communities were exactly the same in the CPMw approach, and a further 27% were contained in a larger CPMw module.

The CPMw allows overlaps between the modules which enables the investigation of further weighted correlation properties. In figures 6(c) and (d) we quantify the influence of strong hubs (researchers with many publications) on the densely internally coupled modules of their co-authors. We find that except for authors with very large paper numbers both the number of communities, m_i, and the number of module neighbours, t_i, of a scientist grow roughly linearly with the number of his/her publications. (Note that t_i is the number of co-authors in dense communities, which is usually smaller than the total number of co-authors, d_i, i.e. the degree of

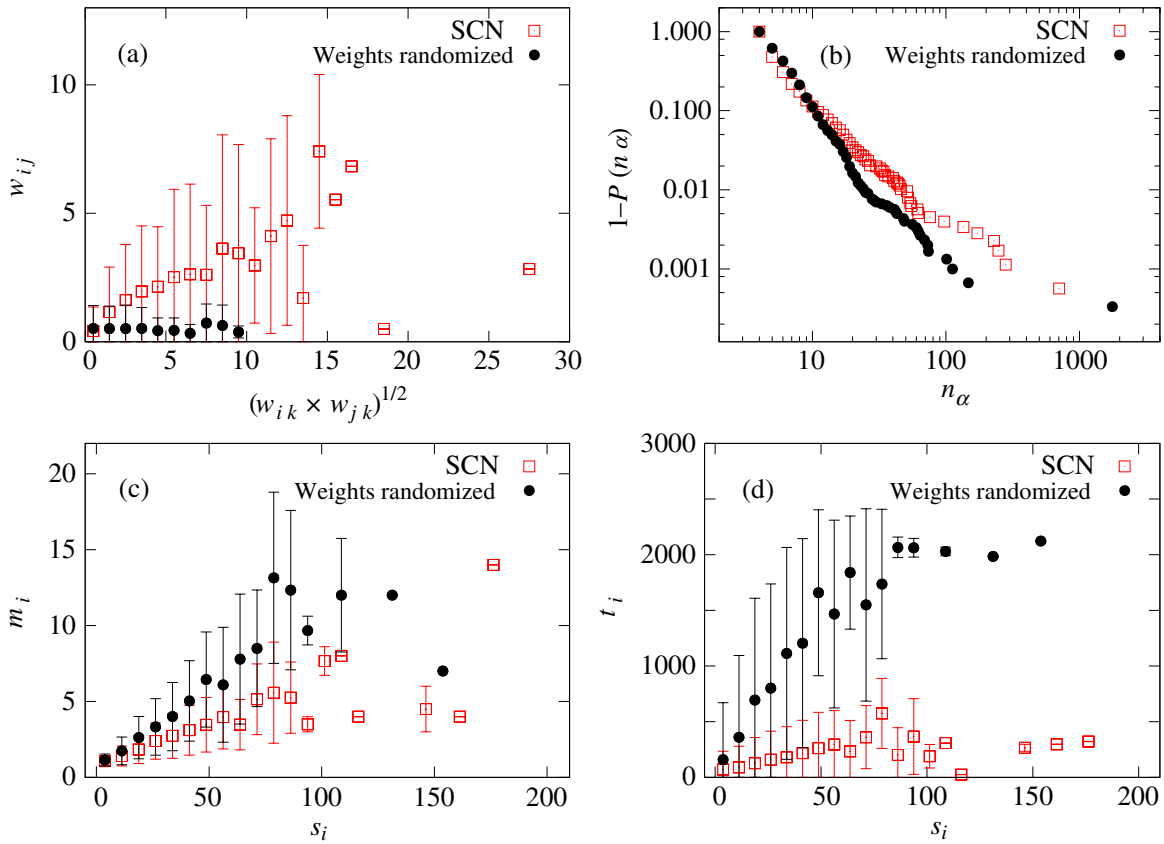

Figure 6. Link weight correlations (in triangles) and weighted modules in the weighted co-publication network of cond-mat authors. The randomized control graph was constructed by shuffling the weights of the links. Different instances of the randomized control graph (with other random seeds) produced similar results. (a) In triangles (nodes i, j, k) the weight of a link, w_{ij}, grows roughly linearly with the geometric mean of the other two link weights, w_{ik} and w_{jk}. (b) Cumulated size distribution of weighted modules. Observe that the largest weighted module of the randomized graph is significantly larger than that of the SCN. (c) Except for scientists with $s_i > 80$ publications, the number of communities (modules) of a node (author) grows linearly with its strength (paper number), similarly to (d) the number of co-authors, t_i, contained by these communities.

the node.) However, both m_i and t_i remain well below the values obtained for the randomized case. These findings indicate that authors remain focussed over time and maintain tight collaborations only with a relatively small number of colleague groups. This weighted correlation behaviour can be quantified more accurately with intermediate-scale methods, e.g. weighted module finding algorithms, than previous 2- or 3-node weighted correlation measurements. Figure 6(c) shows that among authors with $s_i > 80$ publications the average number of modules of one author is above 4.

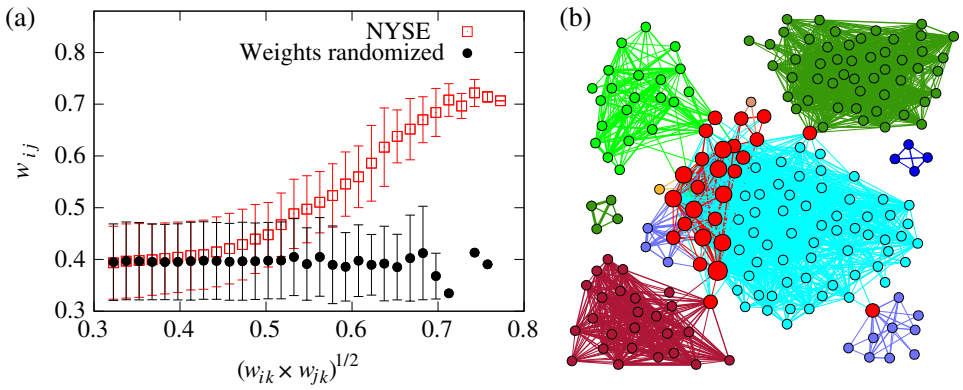

Figure 7. (a) In the NYSE stock graph two strong links of a triangle have a strong third neighbour. (b) Weighted modules of the stock graph. Each node is coloured according to its module. A node contained by more than one module is coloured red and its size is proportional to the number of modules it is contained by.

4.2. Correlation graphs of NYSE stocks

Financial markets, similarly to the participants of a social web, integrate information from a multitude of sources and are truly complex systems. The most widely investigated subunits of a market are its individual stocks (i) and their performances are measured by their prices, $P_i(t)$, over time. Common economic factors influencing the prices of two selected stocks (nodes) are usually detected from the (absolute) value of their correlation (weighted link), which allow one to assemble a network of stocks. In the statistical physics literature minimum spanning trees and asset graphs defined on this web have been have been applied to uncover the hierarchical structure of markets [48] and their clustering properties [16]. Notably, the correlations in their original, matrix, form also provide useful insights when compared to random matrix ensembles as controls [49, 50].

We have analysed a pre-computed stock correlation matrix [35] containing averaged correlations between the daily logarithmic returns, $r_i(t) = \ln P_i(t) - \ln P_i(t-1)$, of $N = 477$ NYSE stocks. Considering a time window of length T, one can compute the equal time correlation coefficients between assets i and j as

$$c_{ij}(t) = \frac{\langle r_i(t) r_j(t)\rangle - \langle r_i(t)\rangle\langle r_j(t)\rangle}{\left[\langle r_i^2(t)\rangle - \langle r_i(t)\rangle^2\right]^{1/2}\left[\langle r_j^2(t)\rangle - \langle r_j(t)\rangle^2\right]^{1/2}}. \tag{14}$$

The pre-computed matrix contained the time averages, c_{ij}, of the correlation coefficients over a four-year period, 1996–2000 ($T = 1000$ days). We used each correlation coefficient, c_{ij}, as a link weight between nodes i and j. As observed and analysed in detail previously in, e.g. [16], only the strongest links (correlations) convey significant information, thus, in both cases we kept only the strongest 3% of all link weights. The resulting network had 301 nodes and 3405 weighted links, the highest and lowest link weights were 0.786 and 0.321.

Similarly to the previous section, we constructed a randomized control graph by reshuffling link weights to analyse weight correlations in groups of three and more weights (figure 7). We found that in triangles the presence of two strong links implies that the third link is also

strong, i.e. groups of three strong links prefer to cluster together. We computed the weighted modules of the stock graph and its randomized control with the CPMw using the same (k, I) parameters and found that the largest modules contained $s_1^{(\mathrm{NYSE})} = 84$ and $s_1^{(\mathrm{rnd})} = 190$ nodes, i.e, the largest module is bigger in the randomized control graph than in the original one. Following the reasoning in section 4.1, this indicates that groups of 2, 3 and more strong links prefer to cluster together in the stock correlation network.

5. Conclusions

We have introduced a module identification technique for weighted networks based on k-cliques having a subgraph intensity higher than a certain threshold, and allowing shared nodes (overlaps) between modules. With this algorithm, the CPMw, we first considered the percolation of k-cliques fulfilling the intensity condition on (weighted) ER graphs. For the critical link probability we showed analytical approximations together with detailed numerical results and found a quickly decaying difference between the two with growing system size.

For two weighted real-world graphs we analysed link weight correlations within groups of 3 and more links. The first was a scientific co-authorship network and the second was a stock correlation graph (NYSE). In the SCN the weighted 2 and 3-point correlation functions studied earlier showed only minor differences from the analogous unweighted correlation functions. Here we investigated the correlations of weights in triangles and computed the weighted modules of the empirical graphs (SCN and NYSE) with the CPMw. We found that in both graphs groups of 3 and more strong links cluster together, i.e. the weighted correlation functions of 3 or more links significantly differ from their randomized counterparts.

Acknowledgments

We thank I Derényi for helpful suggestions and critical reading of the analytical results. We thank S Warner for the ArXiv preprint listings and J-P Onnela and J Kertész for the stock market data. We acknowledge financial support from the Hungarian Scientific Research Fund (grants no. K068669, PD048422 and T049674).

References

[1] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[2] Albert R and Barabási A-L 2002 *Rev. Mod. Phys.* **74** 47
[3] Dorogovtsev S N and Mendes J F F 2002 *Adv. Phys.* **51** 1079
[4] Newman M E J 2003 *SIAM Rev.* **45** 167
[5] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 *Phys. Rep.* **424** 175
[6] PastorSatorras R and Vespignani A 2004 *Evolution and Structure of the Internet: A Statistical Physics Approach* (Cambridge: Cambridge University Press)
[7] Barabási A-L and Oltvai Z N 2004 *Nat. Rev. Gen.* **5** 101
[8] Solé R V and Montoya J M 2001 *Proc. R. Soc.* **268** 2039
[9] Wasserman S and Faust K 1994 *Social Network Analysis: Methods and Applications* (Cambridge: Cambridge University Press)
[10] Everitt B S, Landau S and Leese M 2001 *Cluster Analysis* (London: Arnold)
[11] Girvan M and Newman M E J 2002 *Proc. Natl Acad. Sci. USA* **99** 7821

[12] Palla G, Derényi I, Farkas I and Vicsek T 2005 *Nature* **435** 814
[13] Song C, Havlin S and Makse H A 2005 *Nature* **433** 392
[14] Clauset A and Newman M E J 2004 *Phys. Rev.* E **70** 066111
[15] Jonsson P F, Cavanna T, Zicha D and Bates P A 2006 *BMC Bioinformatics* **7** 2
[16] Onnela J-P, Kaski K and Kertész J 2004 *Eur. Phys. J.* B **38** 353
[17] Kim D-H and Jeong H 2005 *Phys. Rev.* E **72** 046133
[18] Barabási A-L and Albert R 1999 *Science* **286** 509
[19] Milo R *et al* 2002 *Science* **298** 824
[20] Maslov S and Sneppen K 2002 *Science* **296** 910
[21] Colizza V, Flammini A, Serrano M A and Vespignani A 2006 *Nat. Phys.* **2** 110
[22] Barrat A, Barthélemy M, Pastor-Satorras R and Vespignani A 2004 *Proc. Natl Acad. Sci. USA* **101** 3747
[23] Barrat A, Barthélemy M and Vespignani A 2004 *Phys. Rev. Lett.* **92** 228701
[24] Newman M E J 2004 *Phys. Rev.* E **70** 056131
[25] Serrano M A, Boguña M and Pastor-Satorras R 2006 *Phys. Rev.* E **74** 055101(R)
[26] Yu H and Gerstein M 2006 *Proc. Natl Acad. Sci. USA* **103** 14724
[27] Bernhardsson S and Minnhagen P 2006 *Phys. Rev.* E **74** 026104
[28] Bianconi G and Barabási A-L 2001 *Phys. Rev. Lett.* **86** 5632
[29] Fortunato S, Flammini A and Menczer F 2006 *Phys. Rev. Lett.* **96** 218701
[30] Claffy K C, Monk T and McRobb D 1999 Internet tomography http://www.caida.org/tools/measurement/skitter
[31] Palla G, Barabási A-L and Vicsek T 2007 *Nature* **446** 664
[32] Newman M E J 2001 *Phys. Rev.* E **64** 016131
[33] Newman M E J 2001 *Phys. Rev.* E **64** 016132
[34] Almaas E, Kovács B, Vicsek T, Oltvai Z N and Barabási A-L 2004 *Nature* **427** 839
[35] Onnela J-P, Saramäki J, Kertész J and Kaski K 2005 *Phys. Rev.* E **71** 065103
[36] Mukherjee G and Manna S S 2006 *Phys. Rev.* E **74** 036111
[37] Adamcsek B, Palla G, Farkas I J, Derényi I and Vicsek T 2006 *Bioinformatics* **22** 1021
[38] Luscombe N M *et al* 2004 *Nature* **431** 308
[39] Wuchty S and Almaas E 2005 *Proteomics* **5** 444
[40] Pollner P, Palla G and Vicsek T 2006 *Europhys. Lett.* **73** 478
[41] Derényi I, Palla G and Vicsek T 2005 *Phys. Rev. Lett.* **94** 160202
[42] Bollobás B 1985 *Random Graphs* 2nd edn (Cambridge: Cambridge University Press)
[43] Hoshen J and Kopelman R 1976 *Phys. Rev.* B **14** 3438
[44] Milgram S 1967 *Psychol. Today* **2** 60
[45] Jeong H, Neda Z and Barabasi A-L 2003 *Europhys. Lett.* **61** 567
[46] Warner S 2003 *Library Hi Tech* **21** 151
[47] Zhou S and Mondragon R J 2004 *IEEE Commun. Lett.* **8** 180
[48] Mantegna R N 1999 *Eur. Phys. J.* B **11** 193
[49] Laloux L, Cizeau P, Bouchaud J-P and Potters M 1999 *Phys. Rev. Lett.* **83** 1467
[50] Plerou V, Gopikrishnan P, Rosenow B, Amaral L A N and Stanley H E 1999 *Phys. Rev. Lett.* **83** 1471