Available online at www.sciencedirect.com
SCIENCE DIRECT®
Physica A 364 (2006) 581–594
PHYSICA A
www.elsevier.com/locate/physa

Subgraph centrality and clustering in complex hyper-networks

Ernesto Estrada[a,*], Juan A. Rodríguez-Velázquez[b]

[a]*Complex Systems Research Group, X-rays Unit, RIAIDT, Edificio CACTUS, University of Santiago de Compostela, 15706 Santiago de Compostela, Spain*
[b]*Department of Mathematics, University Carlos III de Madrid, 28911 Leganés, Madrid, Spain*

Received 28 September 2004; received in revised form 1 December 2005
Available online 3 January 2006

Abstract

The representation of complex systems as networks is inappropriate for the study of certain problems. We show several examples of social, biological, ecological and technological systems where the use of complex networks gives very limited information about the structure of the system. Consequently, we extend the concepts of subgraph centrality and clustering for complex networks represented by hypergraphs: *complex hyper-networks*. The first parameter characterizes the node participation in different sub-hypergraphs and the second one characterizes the transitivity in the hyper-network through the proportion of hyper-triangles to paths of length two. Another measure characterizing the formation of triples of mutually adjacent groups in the hyper-network is also introduced. All of these characteristics are studied in three different hyper-networks: a scientific collaboration hyper-network, an ecological competition hyper-network and the hyper-network formed by the American corporate elite in 1999.

Keywords: Centrality; Clustering; Hypergraphs; Graph spectra

1. Introduction

The study of complex networks represents an important area of multidisciplinary research involving physics, mathematics, chemistry, biology, social sciences, and information sciences, among others [1–5]. These systems are commonly represented by mean of simple or directed graphs that consist of sets of nodes representing the objects under investigation, e.g., people or groups of people, molecular entities, computers, etc., joined together in pairs by links if the corresponding nodes are related by some kind of relationship. These networks include the Internet [6], the World Wide Web [7], social networks [8–11], information networks [12,13], neural networks [14], food webs [15], reaction and metabolic networks [16], and protein–protein interaction networks [17].

In some cases the use of simple or directed graphs to represent complex networks does not provide a complete description of the real-world systems under investigation. For instance, in a collaboration network represented as a simple graph we only know whether scientists have collaborated or not, but we can not know

*Corresponding author. Tel.: +34 981563100; fax: +34 981547077.
E-mail address: estrada66@yahoo.com (E. Estrada).

0378-4371/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.physa.2005.12.002

whether three or more authors linked together in the network were coauthors of the same paper or not. A possible solution to this problem is to represent the collaboration network as a bipartite graph in which a disjoint set of nodes represents papers and another disjoint set represents authors. However, in this case the "homogeneity" in the definition of nodes is lost, because we have certain nodes that represent papers and others that represent authors. In the study of connectivity, clustering and other topological properties, this distinction between two classes of nodes with completely different interpretations may lead to artifacts in the data [18].

A natural way of representing these systems is to use a generalization of graphs known as *hypergraphs* [19,20]. In a graph a link relates only a pair of nodes, but the edges of the hypergraph—known as hyper-edges—can relate groups of more than two nodes. Thus, it is useful to represent the collaboration network as a hypergraph in which nodes represent authors and hyper-edges represent the groups of authors that have published papers together. For the sake of abbreviation we will call *complex hyper-networks* to these hypergraphs representing complex systems.

The topological characterization of complex systems has played an important role in the understanding of the construction principles, evolution and robustness of real-world complex networks. For instance, the clustering coefficient and the average path length appear intimately related to the concept of "small-world" networks [14]. The clustering coefficient has become one of the global standard parameters used to characterize the topology of complex networks. It has been extended to higher order analogues [21] as well as modified to filter out the degree correlation biases [22] as well as by considering correlated networks [23] and mean-field theory in Barabási-Albert networks [24]. Another important topological parameter for complex networks is related to the number of subgraphs or "network motifs", which are patterns of interconnections occurring in complex networks at numbers that are significantly higher than those in random networks [25,26]. This topological characteristic has been generalized by defining "roles" in a subgraph according to structural equivalence [27]. Subgraphs and network motifs has also been analyzed in random and in geometric networks [28,29].

The principal objective of the current work is to extend the concept of clustering coefficient and subgraph centrality to complex hyper-networks. The *subgraph centrality* [30] is a topological parameter for complex networks which is related to the number of subgraphs giving higher importance to the smallest ones, which have been revealed as important motifs in real-world complex networks [31]. The extension of these concepts to complex hyper-networks will open new possibilities for the topological analysis of complex systems represented by hypergraphs. In this work we first give several examples of complex systems for which hypergraph representation is necessary, then we introduce the concepts of subgraph centrality and clustering coefficients for complex hyper-networks and finally apply them to three real-world complex hyper-networks.

2. Examples of complex hyper-networks

2.1. Social networks

In social networks nodes represent people or groups of people, normally called actors, that are connected by pairs according to some pattern of contact or interactions between them [8]. Such patterns can be of friendship, collaboration, sexual contact, business relationships, etc. There are some cases in which hypergraph representations of the social network are indispensable. These are, for instance, the supra-dyadic transactions in social networks in which it is necessary to consider the coordinated actions of more than two actors, such as a buyer, a seller and a broker. Other examples include the scenarios in which not only the actors taking part in the actions are important, but other factors such as places or times in which the actions taking place are essential to describe such acts. Bonacich et al. [32] have taken such additional characteristics into account in extending the eigenvector centrality for hypergraphs representing supra-dyadic transactions. In such hyper-networks the nodes represent actors related by a common process, which is represented by a hyper-edge, such as a commercial transaction. The first application of hypergraphs for representing social networks appears to be dated on 1981 as reported by Seidman [33].

2.2. Reaction and metabolic networks

A chemical reaction is a process in which a set of chemical compounds known as educts, E_i, react in certain *stoichiometric* proportions, e_i, to be transformed into a set of other chemical compounds named products, P_i, which are produced in certain *stoichiometric* quantities p_i:

$$e_1E_1 + e_2E_2 + \cdots \rightarrow p_1P_1 + p_2P_2 + \cdots.$$

A chemical reaction can be described as a weighted directed hyper-edge in a directed hypergraph where nodes are the chemicals and hyper-edges are the reactions [34]. The absence of a well-developed theory for the structural analysis of (directed) hypergraphs means that two alternative representations of a chemical reaction are commonly used. The first is the bipartite graph, in which a set of nodes represents educts and products and the other set represents the reaction itself. The other representation consists of the *substrate graph*, which considers educts and products as nodes—two nodes are connected if the corresponding chemical compounds take part in the same reaction.

Metabolic networks can be considered as particular cases of reaction networks that are structurally well-characterized as they can be reconstructed for many organisms up to genome-scale. Metabolic pathways are represented in the form of graphs in which nodes represent molecular entities and edges represent reactions or processes relating the molecules involved in the reaction. However, since a reaction may have more than one substrate, and more than one product, the pathway is better represented by a hyper-network in which hyper-edges represent reactions and nodes represent molecular entities [35]. As reaction graphs, metabolic networks are normally represented as substrate or bipartite graphs. Some problems arise when these representations are used for the analysis of potential failure modes in the metabolic network [36].

2.3. Protein complex networks

The systematic characterization of multi-protein complexes in the whole proteome of an organism requires the data to be organized in the form of protein membership lists of the protein complexes. The most common forms of this organization are the protein–protein interaction networks and the complex intersection graphs. In the first representation the nodes of the network represent proteins and an edge links two proteins that interact with each other. This representation, however, does not take into account the multi-protein complexes. In the complex intersection graph the nodes represent complexes, and a link exists between two complexes if they have one or more proteins in common. Clearly, this second representation does not provide information about proteins. A natural way of accounting for the information about both proteins and common protein membership in the complexes, such as common regulation, localization, turnover, or architecture, is to use a hypergraph representation. In the protein complex hyper-networks each protein is represented by a node and each complex by a hyper-edge [37]. These kinds of hyper-networks can be visualized as bipartite graphs.

2.4. Food webs

Trophic relations in ecological systems are normally represented through the use of food webs, which are oriented graphs (digraphs) whose nodes represent species and links represent trophic relations between species [38]. Another way of representing food webs is by means of competition graphs $C(G)$, which have the same set of nodes as the food web but in which two nodes are connected if, and only if, the corresponding species compete for the same prey in the food web [39]. In the competition graph we can only know if two linked species have some common prey, but we can not know the composition of the whole group of species that compete for common prey. In order to solve this problem a competition hypergraph has been proposed in which nodes represent species in the food web and hyper-edges represent groups of species that compete for common prey [40]. It has been shown that in many cases competition hyper-networks yield a more detailed description of the predation relations among the species in the food web than competition graphs. A food web and its competition network and hyper-network are illustrated in Fig. 1.

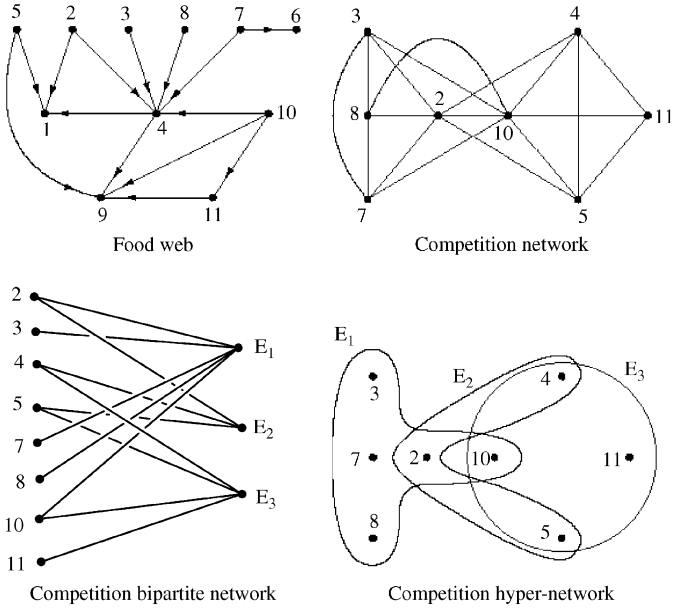

Fig. 1. Food web for a Malaysian rain forest, its competition network, bipartite network and hyper-network.

2.5. Other complex systems

There are other fields in which hypergraphs have been used to study complex systems and there are many others in which the use of hyper-networks may provide an interesting alternative. Harn et al. [41] used hyper-networks to study the software evolution process in which an *evolutionary hypergraph* was defined as a labeled, directed and acyclic hypergraph. A different application of hypergraphs was developed by Krackhardt and Kilduff, who studied the so-called Simmelian tied dyads, which are dyads embedded in three-person cliques, in three entrepreneurial firms [42]. Hypergraphs have also appeared as a natural consequence of an L-percolation process in complex networks, as studied by da Fontoura Costa [43], in the multi-index matching problem [44] as well as in the detection of hidden groups in communication networks [45]. All of these applications clearly indicate the importance of hypergraphs for representing and studying complex systems.

3. Sub-hypergraph centrality of hyper-networks

Throughout this article $H = (V, E)$ denotes a simple and finite hypergraph with vertex set $V = \{v_1, v_2, \ldots, v_N\}$ and hyper-edge set $E = \{E_1, E_2, \ldots, E_m\}$. We will refer to *complex hyper-networks* as hypergraphs that represent a complex system, such as those previously described in this work. The adjacency matrix, $\mathbf{A}(H)$, of the hypergraph $H = (V, E)$ is a square symmetric matrix whose entries a_{ij} are the number of hyper-edges that contain both nodes v_i and v_j: the diagonal entries of $\mathbf{A}(H)$ are zero. For a set $J \subset \{1, 2, \ldots, m\}$ and a set $A \subset V$ the family $H_{(A,J)} = (E_j \cap A : j \in J, E_j \cap A \neq \phi)$ is called *sub-hypergraph* of $H = (V, E)$ induced by the sets A and J. The reader is referred to the literature for more details on hypergraphs [19,20].

Let H be a simple hypergraph of order N. Since the adjacency matrix, A, of H is a symmetric matrix with real entries, there exists an orthogonal matrix $U = (u_{ij})$ such that $A = UDU^{\mathrm{T}}$, where $D = diag(\lambda_1, \lambda_2, \ldots, \lambda_N)$ whose diagonal entries are the eigenvalues of A, and the columns of U are the corresponding eigenvectors that form an orthogonal basis of the Euclidean space $\Re^N$. It must be emphasized that, if the hypergraph H is connected, then the symmetric and non-negative matrix A is irreducible. As a consequence, the main eigenvalue A has a positive eigenvector of multiplicity one. This fact facilitates the extension, to the case of hypergraphs, of the use of the main eigenvector as a measure of centrality. The following result that was obtained in Ref. [46] will be useful in extending the definition of subgraph centrality [30] to hypergraphs. Let v_i

and v_j be vertices of a hypergraph H. Let A be the adjacency matrix of H. Then, the number of walks of length k in H, from v_i to v_j, is the entry in position (i,j) of the matrix A^k.

From the above result we can see that walks of length k in H, from v_i to v_j, are $\mu_k(ij) = (A^k)_{ij} = \sum_{s=1}^{N} u_{is}u_{js}\lambda_s^k$. Hence, the number W_k of walks of length k in H is given by

$$W_k = \sum_{i,j} \mu_k(ij) = \sum_{s=1}^{N} \left(\sum_{i=1}^{N} u_{is} \right)^2 \lambda_s^k.$$

Moreover, the number of closed walks of length k starting and ending on vertex v_i in H is given by the local spectral moments $\mu_k(i)$, which are simply defined as the ith diagonal entry of the kth power of the adjacency matrix, $\mathbf{A}$:

$$\mu_k(i) = (\mathbf{A}^k)_{ii} = \sum_{s=1}^{N} (u_{is})^2 \lambda_s^k \tag{1}$$

and the number CW_k of closed-walks of length k in H is given by

$$CW_k = \sum_{i} \mu_k(i) = \sum_{s=1}^{N} \lambda_s^k, \quad \text{i.e., the trace of } A^k.$$

We define the *sub-hypergraph centrality* of the vertex v as the "sum" of closed walks of different lengths in the network starting and ending at vertex v. As this sum includes both trivial and non-trivial closed walks, we must consider all sub-hypergraphs, i.e., acyclic and cyclic. The contribution of these closed walks decreases as the length of the walks increases. In other words, shorter closed walks have more influence on the centrality of the vertex than longer closed walks. This rule is based on the observation that motifs in real-world networks are small sub-hypergraphs. On the other hand, the use of the sum of closed walks to define sub-hypergraph centrality presupposes a mathematical problem as the series $\sum_{k=0}^{\infty} \mu_k(i) = \infty$ diverges. Consequently, we avoid this problem by scaling the contribution of closed walks to the centrality of the vertex by dividing them by the factorial of the order of the spectral moment. The *sub-hypergraph centrality* of vertex v_i in the network is then given by

$$C_{SH}(i) = \sum_{k=0}^{\infty} \frac{\mu_k(i)}{k!}. \tag{2}$$

Let λ be the main eigenvalue of $\mathbf{A}$. For any non-negative integer k and any $i \in \{1, \ldots, n\}, \mu_k(i) \leqslant \lambda^k$, series (2)—whose terms are non-negative—converges.

$$\sum_{k=0}^{\infty} \frac{\mu_k(i)}{k!} \leqslant \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{\lambda}. \tag{3}$$

Thus, the sub-hypergraph centrality of any vertex v_i is bounded above by $C_{SH}(i) \leqslant e^{\lambda}$. The following result shows that the sub-hypergraph centrality can be obtained mathematically from the spectrum of the adjacency matrix of the network.

Let $H = (V, E)$ be a simple hypergraph of order N. If $v_i \in V$, then the sub-hypergraph centrality $C_{SH}(i)$ may be expressed as follows:

$$C_{SH}(i) = \sum_{j=1}^{N} (u_{ij})^2 e^{\lambda_j}. \tag{4}$$

The expression (4) is obtained by using expressions (1) and (2), from which we obtain the following:

$$C_{SH}(i) = \sum_{k=0}^{\infty} \left(\sum_{j=1}^{N} \frac{\lambda_j^k (u_{ij})^2}{k!} \right). \tag{5}$$

By reordering the terms of series (5), we obtain the absolutely convergent series:

$$\sum_{j=1}^{N}\left((u_{ij})^2\sum_{k=0}^{\infty}\frac{\lambda_j^k}{k!}\right)=\sum_{j=1}^{N}((u_{ij})^2 e^{\lambda_j}), \tag{6}$$

which, clearly, also converges to $C_{SH}(i)$, obtaining the main result.

A global characterization of the network H can be carried out by obtaining the mean of the average sub-hypergraph centrality: $\langle C_{SH}\rangle=\frac{1}{N}\sum_{i=1}^{N}C_{SH}(i)$. It has been recommended that the use of the term centralization instead of centrality is more appropriate for these sorts of global measures [8]. An analytical expression for $\langle C_{SH}\rangle$ can be obtained using a procedure analogous to that described to prove the previous result, showing that $\langle C_{SH}\rangle$ depends only on the eigenvalues and size of the adjacency matrix of the network:

$$\langle C_{SH}\rangle=\frac{1}{N}\sum_{i=1}^{N}C_{SH}(i)=\frac{1}{N}\sum_{i=1}^{N}e^{\lambda_i}. \tag{7}$$

4. Clustering coefficient of hyper-networks

The clustering coefficient was first introduced by Watts and Strogatz to describe "what proportion of acquaintances of a vertex know each other" [14]. In this respect, the global clustering of a network is obtained as the average of the local clustering coefficients for all nodes in the network. It has been stated that this kind of "average of an average" [47] is often not very informative and that a better alternative is to use the following definition of clustering coefficient for a network, which is also known in the sociology literature as the transitivity coefficient [8]:

$$C_2(G)=\frac{6\times\text{number of triangles}}{\text{number of paths of length two}}. \tag{8}$$

The factor of six in the numerator compensates for the fact that each triangle contributes six paths of length two and ensures that $C_2(G)=1$ for the complete graph K_N. In the case of multigraphs, i.e., graphs with multiple edges, this proportion is not maintained and the multigraph is represented as a simple graph to calculate the clustering coefficient. A similar situation is presented for hypergraphs. Thus, in those cases in which the hypergraph has multi-hyper-edges we will consider them as simple hyper-edges. This is equivalent to removing the multiple links between vertices in H. In order to account for the cliquishness of a hypergraph we have to modify (8) to the following expression:

$$C_2(H)=\frac{6\times\text{number of hyper-triangles}}{\text{number of 2-paths}}, \tag{9}$$

where a hyper-triangle is defined as a sequence of three different vertices and three different hyper-edges of the form: v_i, E_p, v_j, E_q, v_k, E_r, v_i, in which the three nodes are mutually adjacent and a 2-path is path of length 2, i.e., a sequence of the type v_i, E_p, v_j, E_q, v_k (we recall that in a path all vertices and hyper-edges are distinct).

We can use the number of closed walks of length three in H to count the number of hyper-triangles in the hyper-network. However, we have to exclude those CWs of length three that are not hyper-triangles. For instance, in Fig. 1 it is shown that v_2, E_2, v_4, E_3, v_{10}, E_1, v_2 is an example of hyper-triangle but neither v_2, E_2, v_4, E_2, v_5, E_2, v_2 nor v_2, E_2, v_4, E_3, v_5, E_2, v_2 are hyper-triangles despite the fact that they are CWs of length three. The CWs of length three that are not hyper-triangles come from the nodes that are on the same hyper-edge. For the sake of simplicity in the terminology we will call them "false" hyper-triangles.

The number of CW of length three containing only one hyper-edge E_i is given by $t_i=\binom{|E_i|}{3}$. These are the number of CWs of length three formed inside a single hyper-edge and, consequently, are not hyper-triangles, e.g., $v_2, E_2, v_4, E_2, v_5, E_2, v_2$ in Fig. 1. In general, to calculate the number of false hyper-triangles we need to use the inclusion–exclusion principle. The cardinal of the intersection of hyper-edges $E_{i_1}, E_{i_2}, \ldots, E_{i_k}$, $\alpha_{i_1 i_2 \ldots i_k}=|\cap_{r=1}^{k}E_{i_r}|$ is denoted by $\alpha_{i_1 i_2 \ldots i_k}$. The number of false hyper-triangles is then $t=\sum_{j=1}^{m}(-1)^{j+1}a_j$, where $a_k=\sum_{i_1,i_2\ldots ik}\binom{\alpha_{i_1 i_2 \ldots i_k}}{3}$ (we only consider the terms in which the combinatorial expression makes sense).

On the other hand, we have to count the number of 2-paths in the hyper-network, which is the denominator of the expression of the clustering coefficient (8). In this respect, we need to identify the number of walks of length two between nodes in the same hyper-edge E_i because they do not constitute a path of length two, i.e., they are "false" 2-paths. The number of false 2-paths is $p = 3t$, where t is the number of false hyper-triangles. The clustering coefficient of the hyper-network is now given by:

$$C_2(H) = \frac{CW_3(H) - 6t}{W_2(H) - CW_2(H) - 6t} = \frac{\sum_i \mu_3(i) - 6t}{W_2(H) - \sum_i \mu_2(i) - 6t}, \tag{10}$$

which, by substitution, gives the expression of the clustering coefficient for a hyper-network. More formally, let t be the number of false hyper-triangles of a hyper-network H. Let $\lambda_1, \lambda_2, \ldots, \lambda_N$ be the eigenvalues of H and let $U = (u_{ij})$ denote an orthogonal matrix whose columns are the corresponding eigenvectors, which form an orthogonal basis of the Euclidean space $\Re^N$. The clustering coefficient of H is given by

$$C_2(H) = \frac{\sum_{s=1}^{N} \lambda_s^3 - 6t}{\sum_{s=1}^{N}((\sum_{j=1}^{N} u_{is})^2 - 1)\lambda_s^2 - 6t}. \tag{11}$$

The clustering coefficient for the competition network in Fig. 1 is $C_2(H) = 0.25$, which indicates that 1/4 of the 18 triples of nodes participating in at least two different competition groups do participate in three different groups. They are v_2, v_4, v_{10} and v_2, v_5, v_{10}, which form the hyper-triangles $v_2, E_2, v_4, E_3, v_{10}, E_1, v_2$ and $v_2, E_2, v_5, E_3, v_{10}, E_1, v_2$, respectively.

It is feasible that we could be interested in knowing the proportion of triples of groups that are mutually adjacent in the hyper-network forming triangles with respect to the number of triples of groups that only form two pairs of adjacent groups. For instance, if we consider three different groups E_1, E_2, E_3 in a hyper-network it is possible to form three pairs of adjacent groups: E_1, E_2, E_1, E_3 and E_2, E_3 and only one triangle of mutually adjacent groups. If we now define the clustering coefficient of the dual, H^*, of the hyper-network H, we will obtain the proportion of triples of groups forming triangles with respect to the number of triples forming adjacent groups:

$$C_2(H^*) = \frac{3 \times \text{number of triangles}}{\text{number of pairs of adjacent edges}}, \tag{12}$$

which in terms of the graph spectrum is given by the following expression:

$$C_2(H^*) = \frac{\sum_{s=1}^{m} \mu_s^3}{\sum_{s=1}^{m}((\sum_{j=1}^{m} b_{is})^2 - 1)\mu_s^2}, \tag{13}$$

where $\mu_1, \mu_2, \ldots, \mu_m$ are the eigenvalues of H^* and $B = (b_{ij})$ denotes an orthogonal matrix whose columns are the corresponding eigenvectors that form an orthogonal basis of the Euclidean space $\Re^{N+m}$.

In the competition hyper-network represented in Fig. 1 the values of $C_2(H^*) = 1$, which indicates that the three competition groups in the network are mutually adjacent. In this case the trophic species 2 predates together with competitors in the competition groups E_1, E_2; species 4 and 5 participate in the competition groups E_2, E_3, and species 10 competes with predators in groups E_1, E_3.

5. Hypernetwork vs. bipartite graph representation

Complex systems of the type described in the Section 2 of this work can also be represented in the form of bipartite networks. In this kind of representation one set of nodes represents the nodes of the hypernetwork and the other set of nodes represents the hyperedges of the hypernetwork. For instance, in the case of a scientific collaboration network the authors form one of the disjoint sets of nodes and the papers form the other, where connections between authors and papers are only allow but neither author–author nor paper–paper connections appear. In the case of the competition network illustrated in Fig. 1 the species form one of the disjoint set of nodes and the three competition groups form the other as illustrated at the end of this figure. As the bipartite graph representation appears to be more intuitive than the hypernetwork the following question is of great importance for the further use of these representations of complex systems: Is the

topological information contained in both representations of a complex system the same? In other words, can we obtain the same topological information using the node degrees, clustering and subgraph centrality if we indistinctly use the bipartite graph or the hypergraph representation?

As a matter of example we will consider the hypernetwork and bipartite graph given in Fig. 1 for the competition of species in the Malaysian rain forest ecosystem. In the case of the node degree it is easy to see that in the bipartite network there are only two kind of species, i.e., those having degree 1 (species 3, 7, 8 and 11) and those having degree 2 (species 2, 4, 5 and 10). However, in the hypernetwork representation there are five groups of species according to the degree centrality, which imposes the following ranking of nodes: $10(7) > 2(6) > 4(5) = 5 > 3(4) = 7 = 8 > 11(3)$, we give the values of the degree in parenthesis. In closing, while the node degree in the bipartite network represents the number of competition groups in which the species i takes place, in the hypernetwork it represents the number of species that competes with i in one or more of the existing competition groups.

Concerning the clustering coefficient it is easy to see that an identical value to that obtained by using expression (11) can be obtained by using the following expression for the bipartite network:

$$C_2(BG) = \frac{3 \times \text{number of cycles of length 6}}{\text{number of paths of length 4}}. \tag{14}$$

However, while expression (11) can be solved analytically from the spectrum of the hypergraph, the calculation based on the bipartite network requires a computational counting of the number of cycles of length 6 and 4, which cannot be computed directly from the spectra of the adjacency matrix of this graph.

The subgraph centrality of both network representations can be obtained from the spectra of the, respective, adjacency matrices of the bipartite graph (BG) and hypergraph (H), respectively. By definition, a bipartite graph does not contain cycles of odd length, which makes that the expression for $SC(i)_{BG}$ can be expressed as follows:

$$C_{BG}(i) = \sum_{j=1}^{N+m} (b_{ij})^2 \cosh(\eta_j), \tag{15}$$

where $\eta_1, \ldots, \eta_{N+m}$ are the eigenvalues of the bipartite graph and (b_{ij}) is an orthogonal matrix whose columns are the corresponding eigenvectors that form an orthogonal basis of the Euclidean space $\Re^{N+m}$.

The values of $SC(i)$ for both representations are given below in the form of vectors corresponding to the node in the following order: 2, 3, 4, 5, 7, 8, 10, 11:

$$SC(i)_{BG} = (2.512\ \ 1.750\ \ 2.559\ \ 2.559\ \ 1.750\ \ 1.750\ \ 2.568\ \ 1.699),$$

$$SC(i)_{H} = (26.698\ \ 14.878\ \ 23.293\ \ 23.293\ \ 14.878\ \ 14.878\ \ 31.716\ \ 10.628).$$

The most important thing here is not the differences in the values of $SC(i)$ according to both representations but whether or not they induce the same ordering of nodes according their centrality. While $SC(i)_{BG}$ ranks nodes 4 and 5 as more central than node 2, the reverse is observed by using $SC(i)_H$, where 2 is ranked as the second most central node after node 10. The specie represented by node 10 takes place in the most numerous competition groups, E_1 and E_3, having 5 and 4 competitors, respectively. This makes this node the most central according to its participation in substructures which involves species in its competition groups and external groups simultaneously. According to this criterion the following most central node is 2, which participates in groups E_1 and E_2 and then nodes 4 and 5, which take part in the competitions groups E_2 and E_3.

On the other hand, in the bipartite network, simple links indicates the number of competition groups in which the node participates. Paths of length two correspond to either the participation of two nodes in one competition group or the participation of one node in two competition groups. As there are not odd cycles, the following substructures present are the squares, which represent the simultaneous participation of two nodes in two competition groups. The only one square present in the bipartite graph is the one formed by: $4 - E_2 - 5 - E_3 - 4$, which indicates the simultaneous participation of nodes 4 and 5 in the competition groups E_2 and E_3. This simple fact makes the difference in favour of ranking these two nodes as more central than node 2 in the bipartite network.

In closing, it is evident that both kind of representation of complex systems are not equivalent and that they contain important but different topological information that can be useful in one or another situation.

6. Analysis of real-world hyper-networks

We will consider here three complex hyper-networks representing a collaboration network, a competition graph in an ecological system and an economical system composed by the corporate elite of US principal companies in 1999. The collaboration network was extracted from the bibliography of the book "Product Graphs" by Imrich and Klavžar [48]. The original network is a bipartite author-by-paper network where link (i,j) represents a situation where author i is the (co)author of the paper j [49]. There are 244 authors and 361 papers, which form a main connected component of 86 authors and 166 papers. We transformed this information into a complex hyper-network in which nodes are authors and hyper-edges are papers in such a way that all coauthors of a paper are linked by the same hyper-edge. The second network consists of the trophic relation between species in the marine ecosystem of Benguela, which is off the southwest coast of South Africa [50]. In this system there are 29 species (3 of which are isolated) in 21 competition groups. We represent this network as a competition graph, as explained before, and in this way we obtain the Benguela competition hyper-network in which nodes are species and hyperlinks join together all species competing for the same prey. The third network is composed by the listed companies and their directors forming a bipartite network in which a director is tied to a company if he/she participates in the board of director of such company. This data set composed by 1591 directors (5 of which are isolated nodes) and 824 companies was previously analyzed as a complex network by Davis et al. [51] and then by Caldarelli and Catanzaro [52]. Here we will account for the hyper-network structure of this system in which nodes represent directors and hyper-edges represent companies in such a way that two hyper-edges are adjacent if they share at least one director. For the sake of comparison we also consider complex networks representing these systems in the form of the author-by-author complex network, representing those pairs of authors that have published a joint paper, the competition graph of the Benguela food web and the director-by-director network composed by pairs of directors that share position at the same company.

We begin by analyzing the characteristics of the degree centrality in the hyper-networks studied and compare them with the corresponding network representations of the same complex systems. One of the distinctive characteristics of centrality measures is that they allow the nodes of the network to be ranked in order to determine the most central ones in such a system. The relative degree centrality for the most central authors in the "Graph Product" collaboration network are plotted in Fig. 2A according to two different representations of the system: network and hyper-network. As can be seen, the ranking obtained for the authors is completely different for both representation methods. It was found that Hell is the most central author in the simple network, whereas Imrich is the most prominent in the hyper-network. There are some other authors that appear highly ranked in one of the representations but do not appear among the top ten authors in the others. For instance, Zhu is ranked as the fifth author according to the hyper-network representation but does not appear among the top ten authors in the network. Imrich and Klavžar are the most central authors in the complex hyper-network. This means that they appear in the largest number of hyper-edges, which represent the different collaboration groups in the hyper-network. We give the term *collaboration group* to the set of authors that collaborate together in a paper. Imrich and Klavžar participate in 26 and 22 collaboration groups, respectively, while Hell participates in only 8 collaboration groups—despite the fact that he has 12 collaborators. It is clear that you can have N collaborators but participate in only one collaboration group if all your collaborators participate in the same paper. The importance of participating in different collaboration groups is evident from the following perspective. If your N collaborators are in only one group all of them can share the same series of ideas, i.e., they form a "school". However, if you are taking part in two or more collaboration groups you will be in touch with more than one school of thought and share ideas from different perspectives. This makes nodes participating in larger numbers of collaboration groups (hyper-edges) the most central nodes in the hyper-network.

A similar situation is represented in Fig. 2B, where the relative degrees based on simple network and hyper-network are plotted for all species in the competition network and hyper-network of the Benguela ecosystem. In the competition network there are 10 species ranked as the most central ones, all of which have a degree

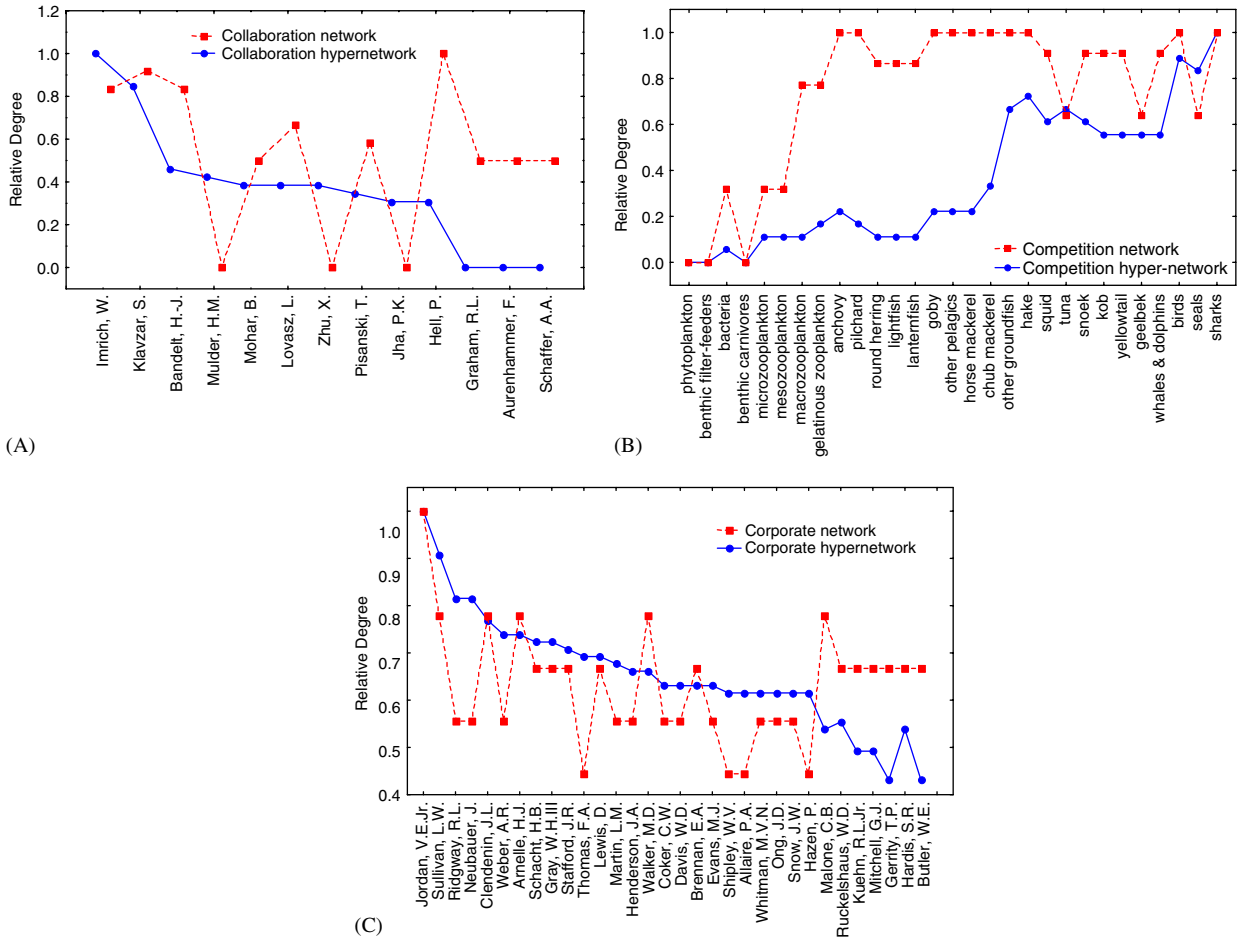

Fig. 2. Ranking of nodes according to the relative degree centrality. Authors in the collaboration network and hyper-network of "Graph Product" (A), species in the competition network and hyper-network for Benguela food web (B) and directors in the corporate elite network and hyper-network (C).

equal to 22. This means that each of these species competes for 22 types of prey. For instance, anchovy, horse and chub mackerel all compete for the same number of prey as sharks and birds because they all have a centrality degree equal to 22 in the competition network. The possibility exists that most of the prey for which a group of species compete are the same. We assign the term *competition group* to this group of species that compete for the same prey. The competition network does not allow us to know in how many competition groups a particular species is participating. However, this information can be obtained in a straightforward way from the competition hyper-network. In this case the node degree corresponds to the number of competition groups in which a species is participating. In the hyper-network representation of the Benguela ecosystem, sharks are the most central species followed by birds and seals. Sharks participate in 18 different competition groups while birds participate in 16 and seals in 15. In contrast, anchovy, horse and chub mackerel participate in only 4 competition groups, which makes sharks, birds and seals the most central species in the competition hyper-network.

The analysis of the degrees of corporate elite network and hyper-network also reveals interesting differences. In both representations V. E. Jordan, Jr. and L. W. Sullivan are ranked as the first two individuals with the most *Fortune* 1000 board membership in 1999 as well as sharing more positions at different companies. However, the directors occupying the third place in the ranking according to the network representation, R. L. Ridgway and J. Neubauer, having 53 links to other directors, are only ranked among the

18–64 positions according to the hyper-network representation. These two directors share most of their 53 links with individuals which are occupying positions in only 5 different companies. On the other hand, there are directors with lower degree in the network representation which are highly ranked in the corporate hyper-network. The most visible case is that of C. B. Malone who has 35 links to other directors, which ranks him among the 41–49 top directors in the corporate network. However, he is sharing his connections with directors in 7 different companies, which makes him the second most important individual according to his degree in the hyper-network representation. In a similar way there are 6 other directors which are ranked as the third most connected individuals in the hyper-network and are only ranked at positions between 37 and 137 in the network representation (see Fig. 2C). The case exists of individuals which are linked to several other directors but participate in very few different companies. This is the case of directors which participate in the boards of one or two companies with large number of other directors. While they have contacts with many influential individuals this influence is only reflected in the mark of a limited space of companies. However, those directors which take part in many different boards are participating in the decisions about long-term strategy of several corporations. For instance, Shirley Young have direct contacts with 33 other directors but they share positions in only two corporations with large number of directors: Bank of America Corp. and Bell Atlantic Corp. On the other side Gerald Tsai, Jr. has contacts with only 8 directors but they are spread in four different corporations: Rite Aid Corp., Saks Inc., Sequa Corp. and United rentals Inc. N.Y.

We also studied the sub-hypergraph centrality of complex hyper-networks by calculating $C_{SH}(i)$ and $\langle C_{SH}\rangle$ for the three hyper-networks studied here as well as their equivalent values for the corresponding simple networks. In the "Product Graphs" collaboration hyper-network there are significant differences in the ranking of nodes compared to that observed in the collaboration network. $C_{SH}(i)$ in the hyper-network ranks Klavžar as the most central author, followed by Imrich, Mohar and Gutman (see Fig. 3A). Hell, who is the most central author in the collaboration network, is not among the top ten authors in the hyper-network. In general, 50% of the authors ranked in the top ten most central nodes in the hyper-network do not appear in the network and vice-versa. On the other hand, there are significant differences between the ranking introduced by $C_{SH}(i)$ in the hyper-network and that obtained by node degrees—as can be seen by comparing Figs. 2A and 3A. In the Benguela ecosystem the competition hyper-network clearly identifies sharks, birds and seals as the most central species according to the sub-hypergraph centrality. In the competition network there are 7 species that are ranked as the most central ones. They include sharks and birds—but not seals—along with horse and chub mackerel, hake, other pelagics and other groundfish. There are also interesting results produced by the analysis of the $C_{SH}(i)$ of the corporate hyper-network. The most central individual according to this measure is John R. Stafford for the hyper-network (see Fig. 3C), who participates in the boards of directors of 5 different companies: Bell Atlantic Corp., Chase Manhattan Bank N.Y., Deere & Co., Allied Signal Inc. and American Home Productions Corp. According to the sub-hypergraph centrality these corporations appear to be forming highly interconnected clusters among them and with other companies. For instance, Bell Atlantic Corp. and Allied Signal inc., which are bonded together by the participation of Stafford in their boards of directors form a hypertriangle with Champion International Corp., which share a director (W. V. Shippley) with Bell and another (L. A. Bossidy) with Allied Signal. These clusters are important for understanding the relationships of influence that two or more companies not directly bonded by the same directors can have. For instance, Champion International Corp. share a director neither with Deere & Co. nor with American Home Productions Corp. However, these two corporations are bonded to Bell and Allied Signal, which share directors with Champion.

The clustering coefficient of the Benguela competition hyper-network is $C_2(H) = 0.067$ because there are 14 hyper-triangles and 625 paths of length two. Only two hyper-triangles exist in the "Graph Product" collaboration network, which has a clustering of $C_2(H) = 0.016$ (there are 377 paths of length 2). These values indicate a low transitivity in the hyper-network as the number of hyper-triangles formed is low compared to the number of paths of length 2. Each hyper-triangle in a hyper-network represents a triple of nodes that join together three different groups (hyper-edges). For instance, the hyper-triangle $v_2, E_2, v_4, E_3, v_{10}, E_1, v_2$ is formed by three trophic species (v_2, v_4, v_{10}) that join together three different competition groups (E_1, E_2, E_3). Thus, the clustering coefficient of the hyper-network $C_2(H)$ measures the proportion of triples of nodes that join three different groups with respect to the number of triples that only join two different groups. With the aim of extracting more conclusive results concerning the role played by transitivity in complex

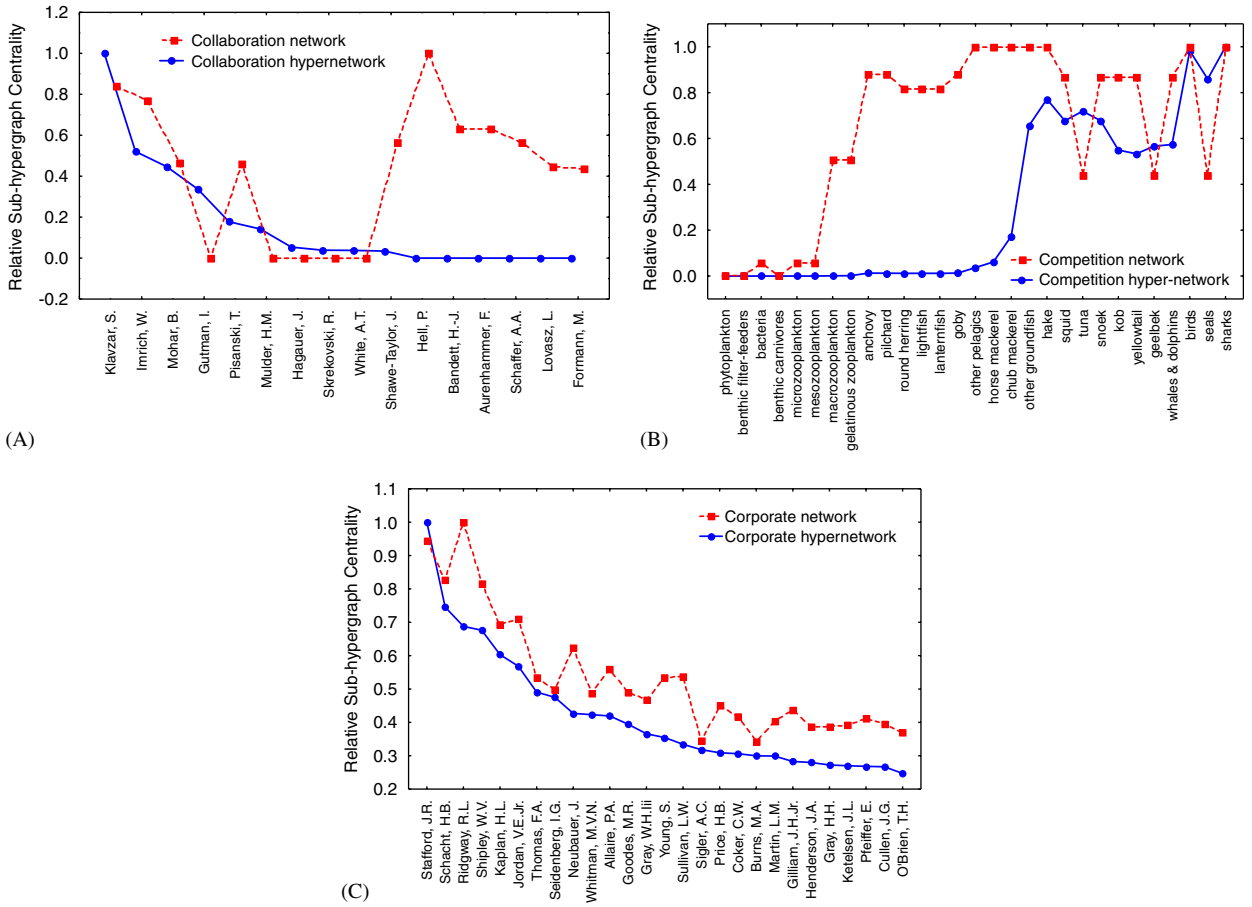

Fig. 3. Ranking of nodes according to the relative subgraph centrality. Authors in the collaboration network and hyper-network of Benguela rain forest food web (B) and directors in the corporate elite network and hyper-network (C).

hyper-networks, we propose the further study of random hyper-networks in order to show whether real-world hyper-networks show, for instance, "small-world" characteristics as observed for complex networks. A value of $C_2(H^*) = 0.963$ is obtained for the competition graph of the Benguela ecosystem, which is close to 1. This result indicates a high level of interrelationship between the different competition groups in these ecosystems. An analysis of this factor for a greater dataset of food webs is necessary to obtain definitive conclusions about the role of this interrelation in ecological systems. The collaboration network on "Product Graphs" also shows a high value of the clustering coefficient between collaborating groups, $C_2(H^*) = 0.758$. This result is not unexpected given the limited scope of the collaboration topic, which makes the different groups working in the field collaborate to a large extent.

Another question of great importance for the study of complex hyper-networks which is related to that of clustering and subgraphs is the identification of communities in such systems [53,54]. The study of communities in complex networks has received great attention in the literature and several approaches have been proposed [55,56]. In hypergraphs this problem has also been studied in particular with relation to the clustering of very large scale integration (VLSI) systems, which are typically represented by hypergraphs [57]. The first step in identifying communities in a complex hyper-network is to represent it as a graph. Several transformation have been proposed to represent hypergraphs as graphs (see [57] for a survey). They include the use of directed graphs, weighted undirected graphs, intersection graphs, clique transformations, etc. The choice of a particular representation depends on the objective or algorithmic approach to be used. In the case of complex hyper-networks their representation as weighted undirected graphs as well as clique graphs

appears to be convenient. In both cases the graph and the hypergraph use the same series of nodes. In a weighted graph the edge weights represent the connectivities of the nodes in the hyper-network. The clique graph is build by constructing an edge (v, w) for every pair of nodes $v, w \in E_j$ for every hyperedge E_j in the hyper-network. The "standard" clique model assigns a weight of $1/(|E| - 1)$ to each clique edge motivated by the linear placement into fixed locations at unit separation. Thus, both representations are weighted graphs but while in the first the weights are integers in the second one they are fractions. In general, several algorithms for clustering in weighted graphs can be used for finding communities in these representations of complex hyper-networks. Another approach that can be useful in this direction is the one introduced by Everett and Borgatti for regular colouring in hypergraphs [58]. This, however, deserves a separate publication for analysing the feasibility of such algorithms for particular complex hyper-networks like the ones studied here.

7. Conclusions

We have introduced here some basic principles for the use of more general representations of complex systems based on hypergraphs. We have coined the term *complex hyper-networks* to designate such systems in which nodes are grouped together in multi-dyadic relationships represented by hyper-edges. We have extended several valuable measures for studying complex hyper-networks, such as node and sub-hypergraph centralities as well as clustering coefficients for both hyper-networks and their duals. The application of these measures to the study of three real-world complex systems—representing a collaboration hyper-network, a competition hyper-network in an ecological system and the American corporate elite in 1999—has shown some of the main differences between these measures for complex networks and hyper-networks. These measures should be necessary for the further study of topological and organizational properties of complex hyper-networks, such as their "small-world" and "scale-free" characteristics, robustness to random fails and attacks, identification of communities, etc.

Acknowledgements

EE thanks "Ramón y Cajal" program, MEC-Spain, for partial financial support. We thank J. Dunne and G. F. Davis for making available datasets used in this work.

References

[1] S.H. Strogatz, Nature (London) 410 (2000) 268.
[2] R. Albert, A.-L. Barabási, Rev. Mod. Phys. 74 (2002) 47.
[3] S.N. Dorogovtsev, J.F.F. Mendes, Adv. Phys. 51 (2002) 1079.
[4] M.E.J. Newman, SIAM Rev. 45 (2003) 167.
[5] A.-L. Barabási, Z.N. Oltvai, Nature Rev. Genet. 5 (2004) 101.
[6] M. Faloutsos, P. Faloutsos, C. Faloutsos, Comp. Commun. Rev. 29 (1999) 251.
[7] R. Albert, H. Jeong, A.-L. Barabási, Nature (London) 401 (1999) 130.
[8] S. Wasserman, K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, UK, 1994.
[9] F. Liljeros, C.R. Edling, L.A.N. Amaral, H.E. Stanley, Y. Åberg, Nature (London) 411 (2001) 907.
[10] A. Schneeberger, C.H. Mercer, S.A.J. Gregson, N.M. Ferguson, C.A. Nyamukapa, R.M. Anderson, A. Johnson, G.P. Garnett, Sex. Transm. Dis. 31 (2004) 380.
[11] M.E.J. Newman, Proc. Natl. Acad. Sci. USA 98 (2001) 404.
[12] R. Ferrer i Cancho, R.V. Solé, Proc. R. Soc. London B 268 (2001) 2261.
[13] M. Sigman, G.A. Cecchi, Proc. Natl. Acad. Sci. USA 99 (2002) 1742.
[14] D.J. Watts, S.H. Strogatz, Nature (London) 393 (1998) 440.
[15] R.J. Williams, N. Martinez, Nature (London) 404 (2000) 180.
[16] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, A.L. Barabási, Nature (London) 407 (2000) 651.
[17] S. Wuchty, Mol. Biol. Evol. 18 (2001) 1694.
[18] G. Benko, C. Flamm, P.F. Stadler, Lect. Notes Comput. Sci. 2801 (2003) 10.
[19] C. Berge, Graphs and Hypergraphs, Elsevier, New York, 1973.
[20] C. Berge, Hypergraphs: The Theory of Finite Sets, North-Holland, Amsterdam, 1989.
[21] A. Fronczak, J.A. Holyst, M. Jedynak, J. Sienkiewicz, Physica A 316 (2002) 688.
[22] S. Nadiv Soffer, A. Vázquez, Phys. Rev. E 71 (2005) 057101.

[23] S.N. Dorogovtsev, Phys. Rev. E. 69 (2004) 027104.
[24] A. Fronczak, P. Fronczak, J.A. Holyst, Phys. Rev. E 68 (2003) 046126.
[25] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, U. Alon, Science 298 (2002) 824.
[26] R. Milo, S. Itzkovitz, N. Kashtan, R. Levitt, S. Shen-Orr, I. Ayzenshtat, M. Sheffer, U. Alon, Science 303 (2004) 1538.
[27] N. Kashtan, S. Itzkovitz, R. Milo, U. Alon, Phys. Rev. E. 70 (2004) 031909.
[28] S. Itzkovitz, R. Milo, N. Kashtan, G. Ziv, U. Alon, Phys. Rev. E. 68 (2003) 026127.
[29] S. Itzkovitz, U. Alon, Phys. Rev. E. 71 (2005) 026117.
[30] E. Estrada, J.A. Rodríguez-Velázquez, Phys. Rev. E 71 (2005) 056103.
[31] E. Estrada, Proteomics, 2006, in press.
[32] P. Bonacich, A.C. Holdren, M. Johnston, Social Networks 26 (2004) 189.
[33] S.B. Seiden, Math. Social Sci. 1 (1981) 381.
[34] O.N. Temkin, A.V. Zeigarnik, D. Bonchev, Chemical Reaction Networks: A Graph-Theoretical Approach, CRC Press, Boca Raton, FL, 1996.
[35] L. Krishnamurthy, J. Nadeau, G. Ozsoyoglu, M. Ozsoyoglu, G. Schaeffer, M. Tasan, W. Xu, Bioinformatics 19 (2003) 930.
[36] S. Klamt, E.D. Gilles, Bioinformatics 20 (2004) 226.
[37] E. Ramadan, A. Tarafdar, A. Pothen, in: Proceedings of the. HICOMB 2004, IPDPS Workshop on High-Performance Computational Biology, Santa Fe NM, April 2004.
[38] S.L. Pimm, Food Webs, Chapman & Hall, London, 1982.
[39] J.E. Cohen, Rand Corp. Document 17696-PR, Santa Monica, CA, 1968.
[40] M. Sonntag, H.-M. Teichert, Discrete Appl. Math. 143 (2004) 324.
[41] M. Harn, V. Berzins, Luqi, A. Mori, in: Proceedings of the IEEE/IEEJ/JSAI International Conference on Intelligence Transportation Systems, Tokyo, Japan, October 5–8, 1999.
[42] D. Krackhardt, M. Kilduff, Social Networks 24 (2002) 279.
[43] L. da Fontoura Costa, Phys. Rev. E 70 (2004) 056106.
[44] O.C. Martin, M. Mézard, O. Rivoire, Phys. Rev. Lett. 93 (2004) 217305.
[45] J. Baumes, M. Goldberg, M. Magdon-Ismail, W.A. Wallace, Lect. Notes Comput. Sci. 3073 (2004) 378.
[46] J.A. Rodríguez, Linear Multilinear Algebra 51 (2003) 285.
[47] B. Bollobás, O.M. Riordan, in: S. Bornholdt, H.G. Schuster (Eds.), Handbook on Graphs and Networks, Wiley-VCH, Berlin, 2004, p. 1.
[48] W. Imrich, S. Klavžar, Product Graphs: Structure and Recognition, Wiley, New York, 2000.
[49] 〈http://vlado.fmf.uni-lj.si/pub/networks/data/2mode/Sandi/Sandi.htm〉.
[50] P. Yodzis, Ecology 81 (2000) 261.
[51] G.F. Davis, M. Yoo, W.E. Baker, Strateg. Organ. 1 (2003) 301.
[52] G. Caldarelli, M. Catanzaro, Physica A 338 (2004) 98.
[53] R. Radicchi, C. Castellano, F. Cecconi, V. Loreto, D. Parisi, Proc. Natl. Acad. Sci. USA 101 (2004) 2658.
[54] A. Barrat, M. Barthelemy, R. Pastor-Satorras, A. Vespignani, Proc. Natl. Acad. Sci. USA 101 (2004) 3747.
[55] M. Girvan, M.E.J. Newman, Proc. Natl. Acad. Sci. USA 99 (2002) 7821.
[56] R. Guimera, L. Danon, A. Díaz-Guilera, F. Giralt, A. Arenas, Phys. Rev. E 68 (2003) 065103.
[57] C.J. Alpert, A.B. Kahng, Integration VLSI J. 19 (1995) 1.
[58] M.G. Everett, S.P. Borgatti, Social Networks 15 (1993) 237.