ELSEVIER

Physica A 314 (2002) 25–34

PHYSICA A

www.elsevier.com/locate/physa

Networks in life: scaling properties and eigenvalue spectra

I. Farkas[a], I. Derényi[a,b], H. Jeong[c], Z. Néda[c,d], Z.N. Oltvai[e], E. Ravasz[c], A. Schubert[f], A.-L. Barabási[c], T. Vicsek[a,*]

[a] *Department of Biological Physics, Eötvös Loránd University, H-1117 Budapest, Hungary*
[b] *Institut Curie, UMR 168, 26 rue d'Ulm, F-75248 Paris 05, France*
[c] *Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA*
[d] *Department of Theoretical Physics, Babes-Bolyai University, RO-3400, Cluj, Romania*
[e] *Department of Pathology, Northwestern University, Medical School, Chicago, IL 60611, USA*
[f] *Bibliometric Service, Library of the Hungarian Academy of Sciences, H-1245 Budapest, Hungary*

Abstract

We analyze growing networks ranging from collaboration graphs of scientists to the network of similarities defined among the various transcriptional profiles of living cells. For the explicit demonstration of the scale-free nature and hierarchical organization of these graphs, a deterministic construction is also used. We demonstrate the use of determining the eigenvalue spectra of sparse random graph models for the categorization of small measured networks.

PACS: 89.65.−s; 89.75.−k; 05.10.−a; 02.70.Hm

Keywords: Random networks; Collaboration graphs; Graph spectra; Spectral analysis of real-world graphs

1. Introduction

Numerous natural, social and technological systems develop large complex structures made up from many similar, but still specific and individual units connected in a stochastic way. The simplest approach still rich in details to such phenomena uses *random network models* built upon ideas from random graph theory and statistical physics.

* Corresponding author.
E-mail address: vicsek@angel.elte.hu (T. Vicsek).

0378-4371/02/$ - see front matter
PII: S0378-4371(02)01181-0

In this paper, we give an overview of the major random graph models and also show an example for a deterministic scale-free network. We analyze and model the *social networks* derived from scientific co-authorship data in mathematics and in neuroscience. For a more detailed description of random networks, we suggest the usage of *spectral properties*. Using a *molecular biological network*, we demonstrate that the spectral characterization is appropriate for the categorization of measured networks even in the case of small systems made up of a few hundred nodes.

2. Network models

2.1. Stochastic graphs

The *uncorrelated random graph model* of Erdős and Rényi [1] treats the network as an assembly of *identical units*, where the number of edges grows quadratically with the size of the system, N. However, in realistic cases, the number of edges grows less rapidly, e.g., linearly with system size. The *small-world graph* [2,3] can be created by connecting each 1st, 2nd, $\ldots, k$th neighbor pair of a one-dimensional periodic lattice and then randomly rewiring a given portion, p_r, of the edges of the original regular graph. The resulting graph shows the *small-world property*: neighbors of an arbitrary vertex are often connected to each other, as well, but the number of steps (taken on the edges of the graph) connecting two arbitrary vertices tends to be low. In the *stochastic scale-free model*, one starts from m isolated vertices, and at each time step one new vertex is added by connecting it to m previous vertices. For a connection, any previous vertex is chosen with a probability proportional to its degree, k_i: $\Pi_i = k_i / \sum_{j=1,N} k_j$. In the infinite size limit, the distribution of degrees converges to a power-law, i.e., the system has *no characteristic length scale*.

2.2. Deterministic scale-free model

In the deterministic scale-free model [4] the network is built iteratively, each iteration repeating and reusing the already existing network. We start from a single node called the *root* of the network. Next, we add two nodes and connect each of them to the *root* of the network. In the nth time step, we add two units identical to the already existing network (containing 3^{n-1} nodes each), and connect each of the 2^n bottom points (vertices) of these two units to the root point.

Due to the deterministic nature of the model, the degree distribution of hubs can be obtained exactly. In the limit, the asymptotic behavior of the degree distribution will depend only on the degrees of *hubs*,[1] i.e., vertices having at least one further vertex below themselves. If we exclude the root point, the number of hubs with $k = 2^{n-i+1} - 2$ links after the nth iteration will be $N_k = (\frac{2}{3})\, 3^i$ $(i = 1, \ldots, n-1)$. With the N_k values known, a standard tool for obtaining the probability density, $p(k)$, of the graph's

[1] The degrees of non-hub points do never exceed the index of the iteration, n, therefore, in the $n \to \infty$ limit, their degrees will not modify the asymptotic power-law behavior of the degree distribution.

degrees would be the cumulative density function, $\Phi(k) = N^{-1} \sum_{k}^{k_{\max}} p(k')$ (N is the total number of hubs). However, for a clear illustration of the underlying mathematical idea, here we will give a slightly different derivation of the probability density of the graph's degrees.

The above formulas of k and N_k will give $N_k \propto k^{-\log 3/\log 2}$ (which can be simplified to $N_{k/2} = 3N_k$). When using this relation as a histogram of the probability density, the size of histogram bins (i.e., the separation between adjacent k values) is proportional to k itself. Thus, in the $N \to \infty$ limit we have

$$p(k) \propto k^{-1} N_k = k^{-(1+\log 3/\log 2)} \tag{1}$$

for the degree distribution and $\gamma = 1 + \log 3/\log 2$ for the scaling exponent. For further recent proposals of deterministic scale-free models, see Refs. [5,6].

3. Mathematics and neuroscience co-authorship networks

Social networks have been largely studied in social sciences [7,8]. A general feature of these studies is that they are restricted to rather small systems, and view networks as static graphs, whose nodes are individuals and links represent various social interactions. Recent statistical physics approaches focus instead on large networks, searching for universalities both in the topology of the web and in the dynamics governing its evolution. These combined theoretical and empirical results have opened unsuspected directions in many fields ranging from computer science to biology [3,7,9–15].

To illustrate the power of these advances, here we summarize our results for the collaboration network of scientists. For each research field one can define a co-authorship network which reflects the professional links between the scientists. In this network nodes represent individual scientists, and two scientists are connected if they have ever published together. In order to gather information on the topology of a scientific co-authorship web, ideally, one would need a complete dataset of the published papers starting from the beginnings of the considered discipline until today. However, computer databases cover at most the past several decades. Thus, any study of this kind needs to be limited to only a recent segment of the database, imposing unexpected challenges.

The databases considered by us contain article titles and authors of all relevant journals in the field of mathematics (M) and neuro-science (NS), published in the period 1991–1998. In mathematics, our database contains 70,975 different authors and 70,901 papers for an interval spanning eight years. In NS, the number of different authors is 209,293 and the number of published papers is 210,750.

It is also important to mention that recently, Newman also applied modern network ideas to collaboration networks [16,17]. He studied large databases focusing on several fields of research over a five-year period, finding that collaboration networks possess all general ingredients of small-world networks [16].

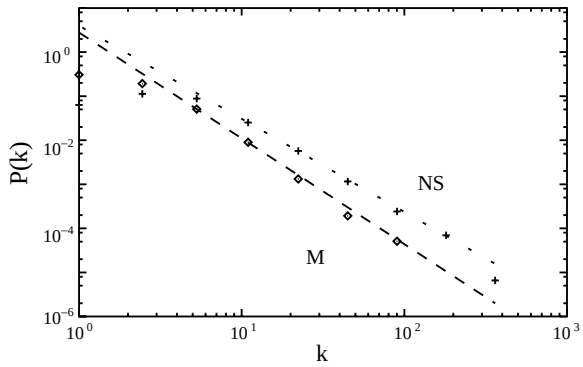

Fig. 1. Degree distribution for the M and NS databases shown with logarithmic binning computed from the full dataset, cumulative up to 1998. The lines correspond to the best fits and have the slopes 2.1 (NS, dotted) and 2.4 (M, dashed).

3.1. Data analysis

In this section, we analyze the topology and dynamics of the M and NS databases.

A quantity that has been much studied lately for various networks is the *degree distribution*, $P(k)$, giving the probability that a randomly selected node has k links. The degree distributions of both the M and NS data indicate that collaboration networks are scale-free. The power-law tail is evident from the raw, uniformly binned data, but the scaling regime is better seen on the plot when logarithmic binning is applied to reduce the noise in the tail (Fig. 1).

Preferential attachment is part of all network models aiming to explain the emergence of the inhomogeneous network structure and power law connectivity distributions [18–20]. For the networks considered by us preferential attachment appears at two levels:

(i) *New nodes*: For a new author preferential attachment means that it is more likely that his/her first paper will be co-authored with somebody who already has a large number of co-authors (links) than with another researcher having fewer collaborators. As a result "old" authors with more links will increase their number of co-authors at a higher rate than those with fewer links. To investigate this process in quantitative terms, we determined the probability that an old author with connectivity k is selected by a new author for co-authorship. This probability defines the $\Pi(k)$ distribution function. Calling "old authors" those present up to the last year, and "new authors" those who were added during the last year, we computed the change in the number of links, Δk, for an old author with k links at the beginning of the previous year, Plotting Δk as a function of k gives the function $\Pi(k)$, describing the nature of preferential attachment involved. Since measurements are limited to only a finite ($\Delta T = 1$ year) interval, we improve the statistics by plotting the integral of $\Pi(k)$:

$$\kappa(k) = \int_1^k \Pi(k')\,\mathrm{d}k' \,. \tag{2}$$

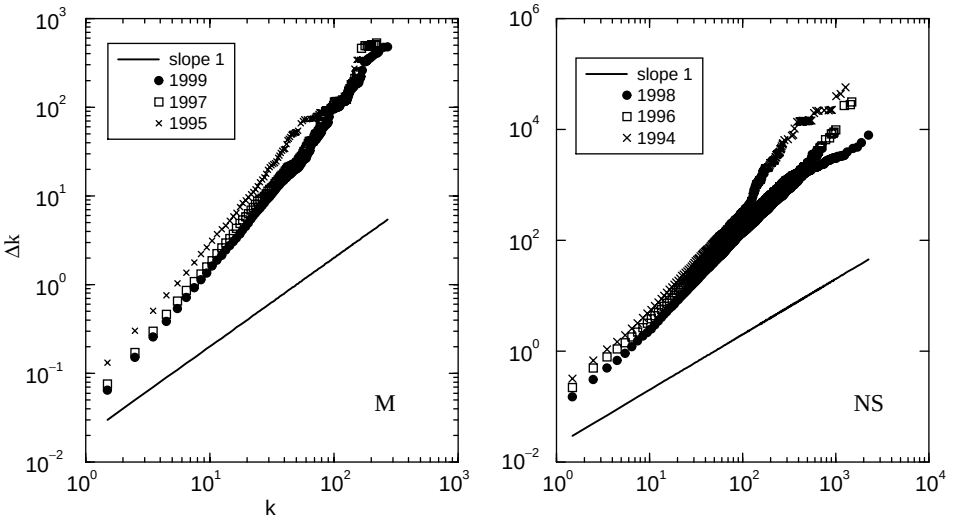

Fig. 2. Cumulated preferential attachment ($\kappa(k)$) of incoming new nodes for the M and NS databases. In the absence of preferential attachment $\kappa(k) \sim k$, which shown as a continuous line on the figures.

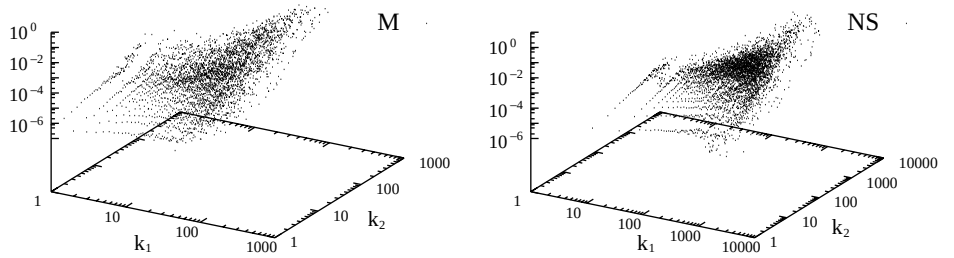

Fig. 3. Internal preferential attachment for the M and NS databases, 3D plots: Δk as a function of k_1 and k_2. Results computed on the cumulative data in the last considered year.

If preferential attachment is absent, $\Pi(k)$ should be independent of k, as each node grows independently of its degree, and $\kappa(k)$ is expected to be linear. As Fig. 2 shows, we find that $\kappa(k)$ is nonlinear, increasing as $\kappa(k) \sim k^{\nu+1}$, with the best fits giving $\nu \simeq 0.8$ for M and $\nu \simeq 0.75$ for NS.

(ii) *Internal links*: As the network evolves, a large number of new links appear between old nodes representing papers written by authors already part of the network, but having not collaborated before. These internal links are also subject to preferential attachment. We studied the probability $\Pi(k_1, k_2)$ that an old author with k_1 links forms a new link with another old author with k_2 links. The three-dimensional plot of $\Pi(k_1, k_2)$ is shown in Fig. 3, the overall behavior indicating that $\Pi(k_1, k_2)$ increases as either k_1 or k_2 increases. Analyzing in detail our data we also found that $\Pi(k_1, k_2)$ for internal links is approximately linear in $k_1 k_2$.

4. Spectral characterization of random graphs

Suprisingly, not only human social networks, one of the highest levels of organization among living systems, but also food webs [21], molecular biological networks [14,22] and the network of the similarities among various genetic programs of a single cell [23] display small-world and scale-free behavior.

However, until now, most analyses of the more realistic models and the analyses of data sets have been confined to the computation of quantities which are only loosely connected to structural properties: e.g., degree sequences, shortest connecting paths and clustering coefficients. Here, we will carry out a more detailed analysis using *algebraic tools* intrinsic to random graphs. Also, we will show that using algebraic tools suprisingly *small measured networks*, consisting of not more than a few hundred nodes, can be successfully *classified* into one of the realistic network model categories shown above.

4.1. Definitions and algorithms

Any graph G can be represented by its *adjacency matrix*, $A(G)$, which is a real symmetric matrix: $A_{ij} = A_{ji} = 1$, if vertices i and j are connected, or 0, if these two vertices are not connected.

The *spectrum of a graph* is the set of eigenvalues of the graph's adjacency matrix. The largest eigenvalue, λ_1, is also called the *principal eigenvalue* of the graph. To illustrate the meaning of the graph's eigenvalues, consider the following example. Write each component of a vector $\vec{v}$ on the corresponding vertex of the graph: v_i on vertex i. Next, on every vertex write the sum of the numbers found on the neighbors of vertex i. If the resulting vector is a multiple of $\vec{v}$, then $\vec{v}$ is an eigenvector, and the multiplier is the corresponding eigenvalue of the graph.

The *spectral density*, $\rho(\lambda)$, *of a graph* is the density of the eigenvalues of its adjacency matrix. For a finite system, this can be written as a sum of delta functions

$$\rho(\lambda) := \frac{1}{N} \sum_{j=1}^{N} \delta(\lambda - \lambda_j) \,, \tag{3}$$

which converges to a continuous function with $N \to \infty$. The spectral density of a graph can be *directly related to the graph's topological features*: the kth moment, M_k, of $\rho(\lambda)$ can be written as

$$M_k = \frac{1}{N} \sum_{j=1}^{N} (\lambda_j)^k = \frac{1}{N} \operatorname{Tr}(A^k) = = \frac{1}{N} \sum_{i_1, i_2, \ldots, i_k} A_{i_1,i_2} A_{i_2,i_3} \cdots A_{i_k,i_1} \,. \tag{4}$$

From the topological point of view, $D_k = NM_k$ is the *number of directed paths* (loops) of the underlying—undirected—graph, that return to their starting vertex after k steps (see Ref. [24] for a detailed explanation).

4.2. Spectral densities of random graph models

In the infinite system size limit, the spectral density of an uncorrelated random graph—if rescaled as $\lambda' = \lambda[Np(1-p)]^{-1/2} \propto \lambda N^{-1/2}$—converges to a semi-circle:

$$\rho(\lambda') = \begin{cases} (2\pi)^{-1}\sqrt{4-\lambda'^2}, & \text{if } |\lambda'| < 2\sigma\,, \\ 0, & \text{otherwise}\,. \end{cases} \tag{5}$$

Suprisingly, the ideal semi-circular spectral density is not valid for any of the above-mentioned realistic graph models [24–26]. In the *sparse uncorrelated random graph* the largest eigenvalue remains constant: $\lambda_1/pN \to c$, and $\rho(\lambda)$ will be symmetric in the infinite system size limit. Since the number of isolated clusters of any size will grow linearly with system size, the spectral density will contain an infinite number of singularities[2] in the limit. Therefore, in the limit all odd moments (M_{2k+1}) converge to 0. In other words: the number of all loops with odd length (D_{2k+1}) disappear. Since on a tree a path returning to its starting point must contain any edge an even number of times, the absence of loops with odd length indicates that the sparse uncorrelated random graph becomes more and more tree-like with $N \to \infty$. In other words, *except for a few shortcuts a sparse uncorrelated random graph looks like a tree*.

We conclude, that—from the spectrum's point of view—the *high number of triangles is one of the most basic properties of the small-world model*, and it is preserved much longer, than regularity or periodicity, if the level of randomness, p_r, is increased. Note that the high number of triangles is equivalent to a high average clustering, C, of the graph.

For $m = m_0 = 1$, the scale-free graph is a tree by definition and its spectrum is symmetric [27]. In the $m > 1$ case, $\rho(\lambda)$ consists of several well distinguishable parts (see Fig. 4). The "bulk" part of the spectral density—the set of the eigenvalues $\{\lambda_2, \ldots, \lambda_N\}$—converges to a symmetric continuous function, which has a triangle-like shape for $|\lambda'| < 1.5$ and has power-law tails.

The *central part* of the spectral density *converges to a triangle-like shape* with its top lying well above the semi-circle. Since the scale-free graph is fully connected by definition, the increased number of eigenvalues with small magnitudes cannot be accounted to small isolated clusters. (All N eigenvalues of a finite-sized isolated cluster with N vertices fall between $-\sqrt{N}$ and $\sqrt{N}$.) As an explanation, we suggest, that the *eigenvectors* of these eigenvalues are *localized on a small subset of the graph's vertices*.

The inset of Fig. 4 shows the *tail of the bulk part* of the spectral density for a graph with $N = 20,000$ vertices and $100,000$ edges (i.e., $pN = 10$).

[2] In a graph with N vertices, the contribution of one isolated vertex to the spectral density is $N^{-1}\delta(\lambda)$, and the contribution of an isolated cluster with two vertices is $N^{-1}[\delta(\lambda+1)+\delta(\lambda-1)]$. An isolated cluster with three vertices and two edges will give $N^{-1}[\delta(\lambda+\sqrt{2})+\delta(\lambda)+\delta(\lambda-\sqrt{2})]$, and one with three vertices and three edges adds $N^{-1}[2\delta(\lambda+1)+\delta(\lambda-2)]$ to the spectral density.

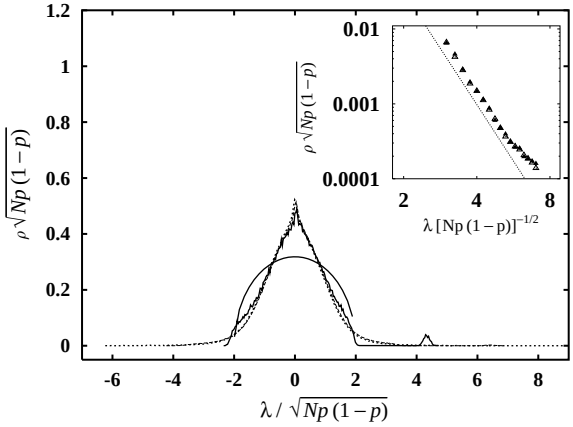

Fig. 4. Main panel: The average spectral densities of scale-free graphs with $m = m_0 = 5$, and $N = 100$ (—), $N = 1000$ (– –) and $N = 7000$ (- - -) vertices. (For $N = 100$ and $N = 1000$, the complete spectrum of 1000 graphs, and for $N = 7000$, the complete spectrum of 25 graphs was used.) Another continuous line shows the semi-circular distribution for comparison. Observe that (i) the central part of the scale-free graph's spectral density is triangle-like, not semi-circular and (ii) the edges show a power-law decay, whereas the semi-circular distribution's edges decay exponentially [28]. Inset: The upper and lower tails of $\rho(\lambda)$ (open and full triangles) for scale-free graphs with $N = 40,000$ vertices the average degree of a vertex being $\langle k_i \rangle = 2m = 10$, as before. Note that both axes are logarithmic and $\rho(\lambda)$ has a power-law tail with the same decay rate at both ends of the spectrum. The line with the slope -5 in this figure is a guide to the eye.

4.3. Application: classification of a molecular biological network by spectral methods

Fig. 5 shows a biochemical network derived from recent gene expression data (see Refs. [23,29] for an explanation). Nodes of the graph represent individual genetic programs (also called transcriptomes) of a cell in response to various internal or external perturbations. A connection between two nodes indicates that according to the applied similarity search algorithm [23], which compares each pair of the genetic programs individually, the two indicated transcriptomes contain a high number of regions with strong correlation ($|C| > 0.8$) of gene expression.

Besides the transcriptome similarity graph, Fig. 5 also shows its spectral analysis by comparing the largest component to three idealized test graphs with the same number of edges and vertices. Note that on the inverse participation ratio vs. eigenvalue plots the *best fit is given by the scale-free graph*, which almost completely overlaps with the measured graph's data. The principal eigenvalue and the inverse participation ratio of the first eigenvector are both high in the measured and the scale-free graphs, whereas they are both significantly lower in the two other models. This indicates that the largest component of the transcriptome similarity graph is scale-free and a handful of its vertices are structurally dominant.

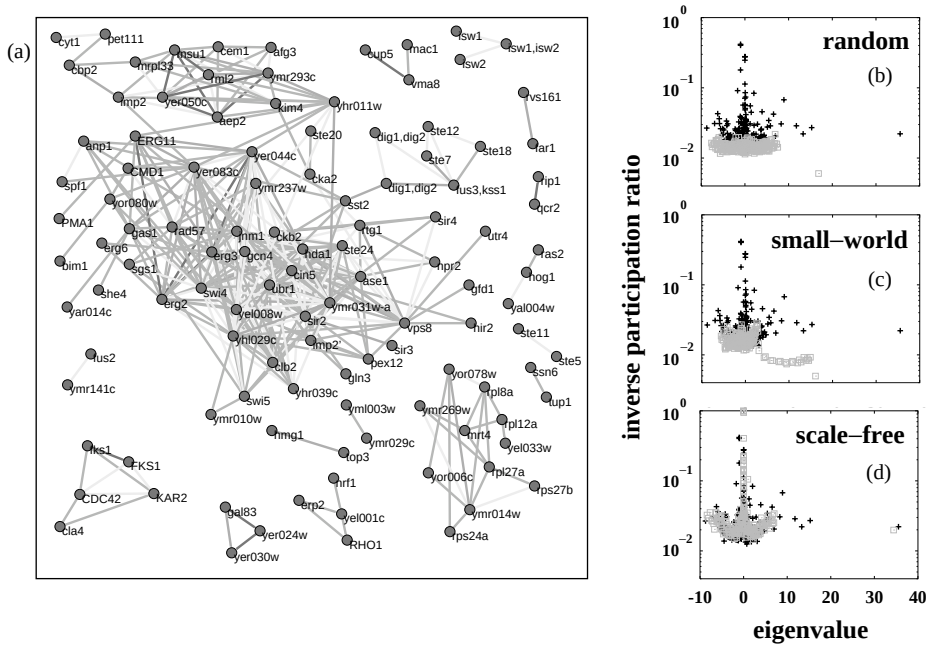

Fig. 5. (a) The transcriptome similarity network of Ref. [23]. Each link is colored according to its predicted strength: dark links indicate strong coupling. (b–d) Analysis of the graph's largest component (the data graph). On each analysis plot black + signs show data points and open gray boxes show data points of the three random graph models. The data graph and all three test graphs have the same number of edges and vertices. The almost complete overlapping of data and test points on subfigure (d) indicates that the transcriptome similarity graph is scale-free.

References

[1] P. Erdős, A. Rényi, On the evolution of random graphs, Publ. Math. Inst. Hung. Acad. Sci. 5 (1960) 17–61.
[2] D.J. Watts, Small World, Princeton University Press, Princeton, NJ, 1999.
[3] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (1998) 440–442.
[4] A.-L. Barabási, E. Ravasz, T. Vicsek, Deterministic scale-free networks, Physica A 299 (2001) 559–564.
[5] S.N. Dorogovtsev, A.V. Goltsev, J.F.F. Mendes, Pseudofractal scale-free web, Phys. Rev. E 65 (2002) 066122.
[6] S. Jung, S. Kim, B. Kahng, Phys. Rev. E 65 (2002) 056101.
[7] S. Wasserman, K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, 1994.
[8] M. Kochen (Ed.), The Small World, Ablex, Norwood, NJ, 1989.
[9] L.A.N. Amaral, A. Scala, M. Barthélémy, H.E. Stanley, Proc. Nat. Acad. Sci. USA 97 (2000) 11149.
[10] R. Albert, A.L. Barabási, Phys. Rev. Lett. 85 (2000) 5234.
[11] R. Albert, H. Jeong, A.L. Barabási, Nature 400 (1999) 130.
[12] S. Lawrence, C.L. Giles, Nature 400 (1999) 107.
[13] B.A. Huberman, L.A. Adamic, Nature 401 (1999) 131.
[14] H. Jeong, et al., Nature 407 (2000) 651.
[15] R.V. Solé, J.M. Montoya, 2000, cond-mat/0011196.
[16] M.E.J. Newman, Proc. Nat. Acad. Sci. USA 98 (2001) 404.

[17] M.E.J. Newman, Phys. Rev. E 64 (2001) 016131.
[18] S.N. Dorogovtsev, J.F.F. Mendes, Europhys. Lett. 52 (2000) 33.
[19] S.N. Dorogovtsev, J.F.F. Mendes, Phys. Rev. E 62 (2000) 1842.
[20] P.L. Krapivsky, S. Redner, F. Leyvraz, Phys. Rev. Lett. 85 (2000) 4629.
[21] J.M. Montoya, R.V. Solé, Small World Patterns in Food Webs, cond-mat/0011195.
[22] S. Wuchty, Scale-free behavior in protein domain networks, Mol. Biol. Evol. 18 (2001) 1694–1702.
[23] I.J. Farkas, H. Jeong, T. Vicsek, A.-L. Barabási, Z.N. Oltvai, The topology of the transcription regulatory network in the yeast, *S. cerevisiae*, to be published.
[24] I.J. Farkas, I. Derényi, A.-L. Barabási, T. Vicsek, Phys. Rev. E 64 (2001) 026704:1-12.
[25] M. Bauer, O. Golinelli, Random incidence matrices: moments of the spectral density J. Statist. Phys. 103 (2001) 301–337.
[26] K.-I. Goh, B. Kahng, D. Kim, Spectra and eigenvectors of scale-free networks, Phys. Rev. E 64 (2001) 051903.
[27] D.M. Cvetković, M. Doob, H. Sachs, Spectra of Graphs—Theory and Applications, 3rd revised and enlarged edition, Johann Ambrosius Barth Verlag, Heidelberg-Leipzig, 1995.
[28] M.L. Mehta, Random Matrices, 2nd Edition, Academic Press, New York, 1991.
[29] T.R. Hughes, et al., Functional discovery via a compendium of expression profiles, Cell 102 (2000) 109–126.