

Extracting the hierarchical organization of complex systems

Marta Sales-Pardo, Roger Guimerà, André A. Moreira, and Luís A. Nunes Amaral*

Department of Chemical and Biological Engineering and Northwestern Institute on Complex Systems, Northwestern University, Evanston, IL 60208

Edited by H. Eugene Stanley, Boston University, Boston, MA, and approved July 22, 2007 (received for review April 23, 2007)

Extracting understanding from the growing “sea” of biological and socioeconomic data is one of the most pressing scientific challenges facing us. Here, we introduce and validate an unsupervised method for extracting the hierarchical organization of complex biological, social, and technological networks. We define an ensemble of hierarchically nested random graphs, which we use to validate the method. We then apply our method to real-world networks, including the air-transportation network, an electronic circuit, an e-mail exchange network, and metabolic networks. Our analysis of model and real networks demonstrates that our method extracts an accurate multiscale representation of a complex system.

cellular metabolism | complex networks | multiscale representation

The high-throughput methods available for probing biological samples have drastically increased our ability to gather comprehensive molecular-level information on an ever-growing number of organisms. These data show that these systems are connected through a dense network of nonlinear interactions among its components, and that this interconnectedness is responsible for their efficiency and adaptability. This interconnectedness, however, poses significant challenges to researchers trying to interpret empirical data and to extract the “systems biology” principles that will enable us to build new theories and to make new predictions (1).

A central idea in biology is that life processes are hierarchically organized (2–4). Additionally, it seems plausible that this hierarchical structure plays an important role in the system’s dynamics (5). However, given a set of genes, proteins, or metabolites, and their interactions, we still do not have an objective manner to assess whether such hierarchical organization does indeed exist or to identify the different levels in the hierarchy.

Here, we report a method that identifies the levels in the organization of complex systems and extracts the relevant information at each level. To illustrate the potential of our method, it is useful to think of electronic maps such as those provided by Google Maps [see [supporting information \(SI\) Fig. 5](#)]. Electronic maps are powerful tools because they present information in a scalable manner—despite the increase in the amount of information as we “zoom out,” the representation displays the information that is relevant at the new scale. In the same spirit, our method will enable researchers to characterize each scale with the relevant information at that scale. This achievement is key for the development of systems biology, but it will encounter application in many other areas.

Background

Complex networks are convenient representations of the interactions within complex systems (6). Here, we focus on the identification of inclusion hierarchies in complex networks, that is, to the unraveling of the nested organization of the nodes in a network into modules, which in turn are composed of submodules and so on.[†]

A method for the identification of the hierarchical organization of nodes in a network must fulfill two requirements: (i) It must be accurate for many types of networks, and (ii) it must identify the different levels in the hierarchy as well as the number of modules and their composition at each level. The first condition may appear trivial, but we make it explicit to exclude algorithms that only work

for a particular network or family of networks, but that will otherwise fail. The second condition is more restrictive, as it excludes methods whose output is subject to interpretation. Specifically, a method does not fulfill the second condition if it organizes nodes into a tree structure, but it is up to the researcher to find a “sensible” criterion to establish which are the different levels in that tree. An implication of the previous two requirements is that any method for the identification of node organization must have a null output for networks, such as Erdős-Rényi random graphs (10), which do not have an internal structure.

To our knowledge, there is no procedure that enables one to simultaneously assess whether a network is organized in a hierarchical fashion and to identify the different levels in the hierarchy in an unsupervised way. Ravasz *et al.* (11) studied the hierarchical structure of metabolic networks, but in their analysis the authors put emphasis on detecting “global signatures” of a hierarchical network architecture. Specifically, they reported that for the metabolic networks studied and for certain hierarchical network models the clustering coefficient of nodes appears to scale with the connectivity k as k^{-1} . This scaling, however, is neither a necessary nor a sufficient condition for a network to be hierarchical (12).

More direct methods to investigate the hierarchical organization of the nodes in a network have also been recently proposed (13–15). Although useful in some contexts, these methods do not clearly identify hierarchical levels and thus fail to satisfy condition *ii* above. Furthermore, all of these methods yield a tree even for networks with no internal structure.

In the following, we define inclusion hierarchies in complex networks and describe an ensemble of hierarchically nested random graphs. We then introduce a method that is able to accurately extract the hierarchical organization of graphs in such an ensemble. Last, we apply our method to several real-world networks.

Inclusion Hierarchies

We start by explicitly defining “networks with a nested hierarchical organization” (see *SI Text* for a mathematical formulation). We focus on networks that have groups of nodes (modules) that are more densely connected between themselves than they are to other groups of nodes. Each module can in turn have its own internal organization if there are subgroups of nodes within the module (submodules) that are more interconnected than to other nodes in

Author contributions: M.S.-P. and L.A.N.A. designed research; M.S.-P. performed research; M.S.-P., R.G., and A.A.M. analyzed data; and M.S.-P., R.G., and L.A.N.A. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

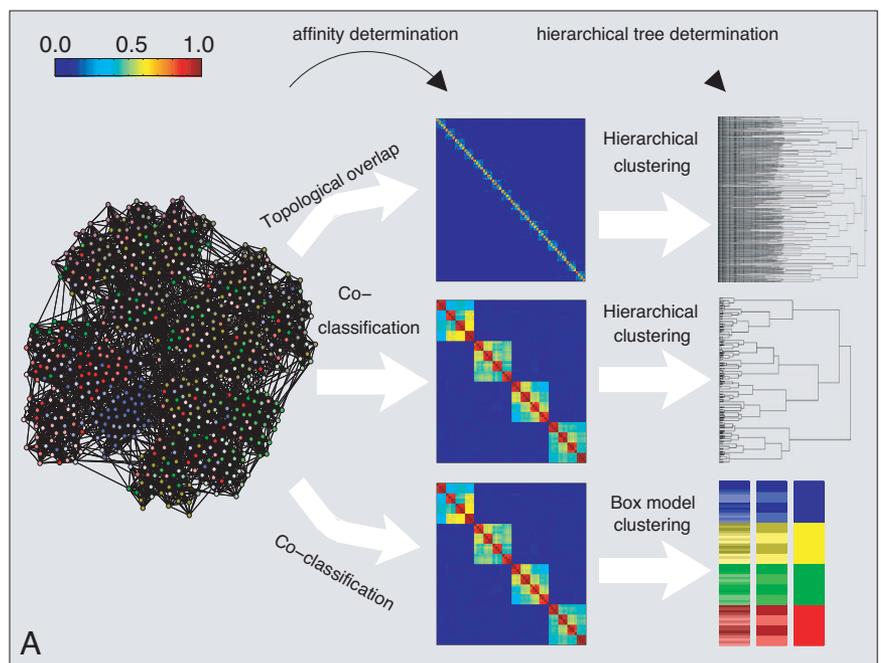
*To whom correspondence should be addressed. E-mail: amaral@northwestern.edu.

[†]We do not consider other hierarchical schemes that classify nodes according to, for instance, their importance (7). Another issue that we do not address here is that of “overlapping” modules. Note also that some authors refer to the existence of “soft” boundaries between communities (8, 9). However, there has been so far no rigorous connection between the soft boundaries and the overlap between communities. Moreover, at present, there is no theoretical model that includes overlapping modules, that is, modules that share nodes, as opposed to communities that share edges.

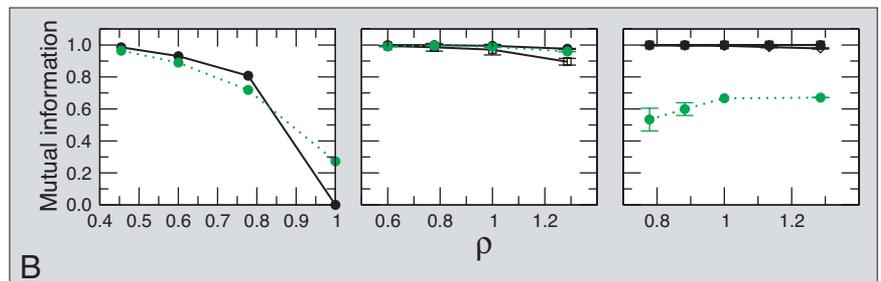
This article contains supporting information online at www.pnas.org/cgi/content/full/0703740104/DC1.

© 2007 by The National Academy of Sciences of the USA

Fig. 2. Affinity measures and clustering methods. (A) We generate a model network comprised of 640 nodes with average degree 16 and with a three-level hierarchical structure (see *SI Fig. 8* for results for a network with a “flat” organization of the nodes). We show the affinity matrices A_{ij} obtained for two different measures: (i) topological overlap (11) and (ii) coclassification (see text and Supplementary Information). The color scale goes from red for an affinity of one to dark blue for an affinity of zero. At the far right, we show the hierarchical tree obtained by using two different methods: hierarchical clustering and the “box clustering” method we propose. In the hierarchical clustering tree, the vertical axis shows the average distance, $\bar{d}_{ij} = 1 - A_{ij}$, of the matrix elements that have already merged. In the box-model clustering tree, each row corresponds to one hierarchical level. Different colors indicate different modules at that level. To better identify which are the submodules at a lower level, we color the nodes in the submodules with shades of the color used for the modules in the level above. Note that topological overlap fails to find any modular structure beyond a locally dense connectivity pattern. In contrast, the coclassification measure clearly reveals the hierarchical organization of the network by the “nested-box” pattern along the diagonal. Significantly, the hierarchical tree obtained via hierarchical clustering fails to reproduce the clear three-level hierarchical structure that the affinity matrix displays, whereas the box-model clustering tree accurately reproduces the three-level hierarchical organization of the network.



(B) Accuracy of the method. We generate networks with 640 nodes and with built-in hierarchical structure comprising one (Left), two (Center), and three (Right) levels. The top level always comprises four modules of 160 nodes each. For networks with a second level, each of the top-level modules is organized into four submodules of 40 nodes. For the networks with three levels, each level-two module is further split into four submodules of 10 nodes. We build networks with different degrees of level cohesiveness by tuning a single parameter ρ (see *SI Text*). For low values of ρ , the levels are very cohesive, for high values of ρ the levels are weakly cohesive. Because we know *a priori* which are the nodes that should be coclassified at each level, we measure the accuracy as the mutual information between the empirical partition of the nodes and the theoretical one (23). We plot the mutual information versus ρ and, for comparison, we also plot the accuracy of a standard community detection algorithm (24) in finding the top level of the networks (dashed green line). Each point is the average over 10 different realizations of the network. Filled circles, empty squares, and filled diamonds represent the accuracy at the top, middle, and lowest levels, respectively. Note that our method is as good at detecting communities as a standard community detection algorithm for networks with a flat organization of the nodes. Additionally, our method is able to detect the top level for all cases analyzed, whereas standard modularity optimization algorithms are not.



Let then \mathcal{P}_{\max} be the set of partitions for which the modularity M is a local maxima, that is, partitions for which neither the change of a single node from one module to another nor the merging of two modules will yield a higher modularity (Fig. 1B). The most straightforward way to calculate A_{ij} would be to consider all partitions $\tilde{P} \in \mathcal{P}_{\max}$, and find the fraction for which (i, j) are placed in the same module. However, such a procedure would not take into consideration the size of the basins of attraction of the different maxima. To understand the importance of this fact, consider the “landscape” in Fig. 1 in which each node represents a partition of the network, and for simplicity, we connect two partitions if the change of a single node transforms one partition into the other. This landscape has two local maxima, partitions 1 and 15. Therefore, if we were only to consider those partitions, we would conclude that those partitions are equally important. However, there is no reason to assume that all partitions have the same importance. Actually, for networks with a very clear modular structure, one expects that a few local maxima will yield the most relevant information about the organization of the network. This idea is formalized through the concept of basin of attraction.

Consider again the landscape in Fig. 1B. Suppose we wanted to find a partition for which the modularity is a maximum with no *a priori* information on the landscape. We would start by grouping the nodes into a randomly chosen partition; let us say, partition 13. In

partition 13, nodes a and c are placed in one group, whereas nodes b and d are placed into their own groups. There are two single node changes that increase the modularity. Node b can be placed in the same group as node d ; this is partition 15, which is a local maxima. Instead, node b can be placed in the same group as nodes a and c ; this is partition 14. Partition 14 is not a modularity maximum; thus one would continue our random ascent of the modularity landscape. From partition 14, one could move to partition 1 or to partition 15, both local maxima. This example illustrates that from partition 13, one has a 25% chance of ending in partition 1 and a 75% chance of ending in partition 15. If one repeats this calculation for every possible starting partition, one obtains the size of the basin of attraction of the two local modularity maxima.

Formally, the size of the basin of attraction of \tilde{P} is

$$b(\tilde{P}) = \sum_{P \in \mathcal{P}} \frac{b(P, \tilde{P})}{\|\mathcal{P}\|} \quad [2]$$

where $b(P, \tilde{P})$ is the probability that starting from partition P one ends at partition $\tilde{P} \in \mathcal{P}_{\max}$ and $\|\mathcal{P}\|$ is the number of possible partitions (Fig. 1B).

We propose that the affinity A_{ij} of a pair of nodes (i, j) is then the probability that when local maxima partition $\tilde{P} \in \mathcal{P}_{\max}$ are sampled with probabilities $b(\tilde{P})$, nodes (i, j) are classified in the same module.

Note that, in contrast to other affinity measures proposed in refs. 9, 15, and 18, the measure we propose does not necessarily coincide with the “optimal” division of nodes into modules, that is, the partition that maximizes M (20). In fact, the modules at the top level of the hierarchy do not necessarily correspond to the best partition found for the global network, even for relatively simple networks (Fig. 2C).

Statistical Significance of the Hierarchical Organization. Given a set of elements and a matrix of affinities between them, a commonly used tool to cluster the elements and, presumably, uncover their hierarchical organization is hierarchical clustering (25, 26). Hierarchical clustering methods have three major drawbacks: (i) They are only accurate at a local level—at every step a pair of units merge and some details of the affinity matrix are averaged with an inevitable loss of information. (ii) The output is always a hierarchical tree, regardless of whether the system is indeed hierarchically organized or not. (iii) There is no statistically sound general criterion to determine the relevant levels on the hierarchy.

To overcome the first caveat of agglomerative methods such as hierarchical clustering, one necessarily has to follow a top-to-bottom approach that keeps all of the information contained in the affinity matrix. That is the spirit of divisive methods such as k -means or principal component analysis (25), which group nodes into “clusters” given an affinity matrix. However, these methods have a significant limitation: The number of clusters is an external parameter, and, again, there is no sound and general criterion to objectively determine the correct number of clusters.

Because of the caveats of current agglomerative and divisive methods, we propose a “box-clustering” method that iteratively identifies in an unsupervised manner the modules at each level in the hierarchy. Starting from the top level, each iteration corresponds to a different hierarchical level (Fig. 2).

First, to assess whether the network under analysis has an internal organization, we need to compare it with the appropriate null model, which in this case is an ensemble of “equivalent” networks with no internal organization. These equivalent networks must have the same number of nodes and an identical degree sequence. A standard method for generating such networks is the Markov-chain switching algorithm (27, 28). Despite their having no internal structure, these randomized networks have numerous local modularity maxima (19). Thus, to quantify the level of organization of a network, one needs to compare the modularities of the sampled maxima from the original network and its corresponding random ensemble; if the network has a nonrandom internal structure, then local maxima in the original landscape should have significantly larger modularities than local maxima in the landscapes of the randomized networks.

Specifically, for a given network, we compute the average modularity M_{av} from $\{M(\hat{P}) : \hat{P} \in \mathcal{P}_{\text{max}}\}$. Then, we compute the same quantity M_{av}^i for each network in the equivalent random ensemble. In virtue of the central limit theorem, the set of average modularities for the whole ensemble $\{M_{\text{av}}^i\}$ is normally distributed with mean M_{rand} and variance $\sigma_{M_{\text{rand}}}^2$ (see SI Fig. 6). To quantify the level of organization of a network, we thus compute the z -score of the average modularity $z = (M_{\text{av}} - M_{\text{rand}})/\sigma_{M_{\text{rand}}}$.

If z is larger than a threshold value z_t , then we conclude that the network has internal structure, and we proceed to identify the different modules; otherwise, we conclude that the network has no structure (Fig. 1D). In what follows, we show results for $z_t = 2.3267$, which corresponds to a 1% significance level[‡] (SI Text and SI Fig. 9).

Building the Hierarchical Tree. In networks organized in a hierarchical fashion, nodes that belong to the same module at the bottom level of the hierarchy have greater affinity than nodes that are together at a higher level in the hierarchy. Thus, if a network has a hierarchical organization, one will be able to order the nodes in

Table 1. Top-level structure of real-world networks

Network	Nodes	Edges	Modules	Main modules
Air transportation	3,618	28,284	57	8
E-mail	1,133	10,902	41	8
Electronic circuit	516	686	18	11
<i>Escherichia coli</i> KEGG	739	1,369	39	13
<i>E. coli</i> UCSD	507	947	28	17

We show both the total number of modules and the number of main modules at the top level. Main modules are those composed of more than 1% of the nodes. Note that there is no correlation between the size or number of edges of the network and the number of main modules. KEGG, Kyoto Encyclopedia of Genes and Genomes; UCSD, University of California at San Diego.

such a way that groups of nodes with large affinity are close to each other. With such an ordering, the affinity matrix will have a “nested” block-diagonal structure. This is indeed what we find for networks belonging to the ensemble of hierarchically nested random graphs (Fig. 2).

For real-world networks, we do not know *a priori* which nodes are going to be coclassified together; that is, we do not know which is the ordering of the nodes for which the affinity matrix has a nested block-diagonal structure. To find such an ordering, we use simulated annealing (29) to minimize a cost function that weighs each matrix element with its distance to the diagonal (30)

$$C = \frac{1}{N} \sum_{i,j=1}^N A_{ij}|i - j|, \quad [3]$$

where N is the order of the affinity matrix (see SI Text and SI Fig. 7). This problem belongs to the general class of quadratic assignment problems (31). Other particular cases of quadratic assignment problems have been suggested to uncover different features of similarity matrices (32). Our algorithm is able to find the proper ordering for the affinity matrix and to accurately reveal the structure of hierarchically nested random graphs (Fig. 2).

The computational cost of this step, the slowest one in our algorithm, limits network sizes to $\approx 10,000$ nodes. However, the cost can be reduced by using faster, but less accurate, methods for ordering the matrix, such as principal component analysis.

Unsupervised Extraction of the Structure. Given an ordered affinity matrix, the last step is to partition the nodes into modules at each relevant hierarchical level. An ansatz that follows naturally from the considerations in the previous section and the results in Fig. 2 is that, if a module at level ℓ (or the whole network at level 0) has internal modular structure, the corresponding affinity matrix is block-diagonal: At level ℓ , the matrix displays boxes along the diagonal, such that elements inside each box s have an affinity A_{ℓ}^s , whereas matrix elements outside the boxes have an affinity $B_{\ell} < A_{\ell}^s$. Note that the number of boxes for each affinity matrix is not fixed; we determine the “best” set of boxes by least-squares fitting of the block-diagonal model to the affinity matrix.

Importantly, we want to balance the ability of the model to accurately describe the data with its parsimony; that is, we do not want to over-fit the data. Thus, we use the Bayesian information criterion to determine the best set of boxes (33).[§]

To find the modular organization of the nodes at the top level (level 1), we fit the block diagonal model to the global affinity

[‡]Results for real networks at a 5% significance level are identical; however, the more stringent threshold is more efficient at detecting the last level in the hierarchy for model networks. Only for a 1–3% of the cases—depending on the cohesiveness of the levels—does the algorithm find one more level than expected.

[§]We have also applied Akaike’s information criterion (34), obtaining the same results for nearly all cases.

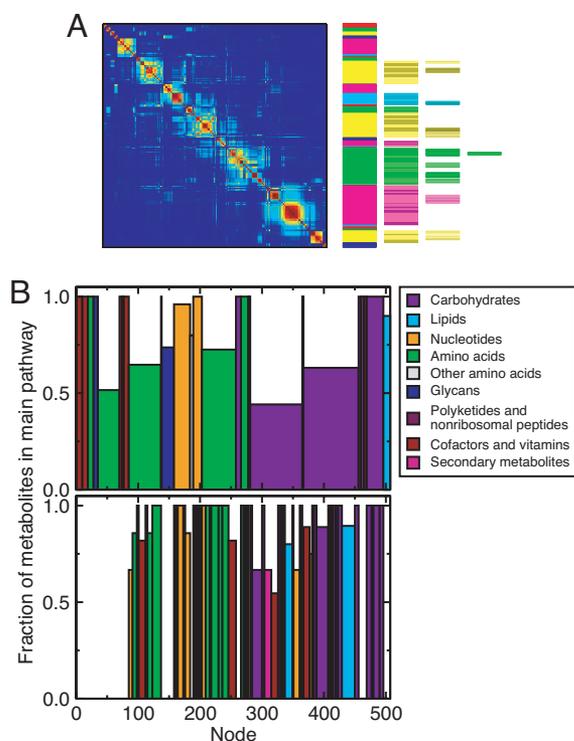


Fig. 4. Hierarchical structure of metabolic networks. (A) Global-level affinity matrices and hierarchical trees for the UCSD reconstruction of the metabolic network of *E. coli* (43). The overall organization of the network is similar and independent of the reconstruction used to build the network (see SI Fig. 11). (B) We analyze the within-module consistency of metabolite pathway classification for the first (Upper plot) and the second (Lower plot) levels. For each module, we first identify the pathway classifications of the corresponding metabolites; then, we compute the fraction of metabolites that are classified in the most abundant pathway. In the plots, each bar represents one module, its width being proportional to the number of nodes it contains. We color each bar according to its most abundant pathway following the color code on the right-hand side. At the second level (Lower plot), we show each submodule directly below its corresponding top level module. Again, the width of each submodule is proportional to its size. Note that, at the first level (Upper), for all modules except one, the most abundant pathway is composed of more than 50% of the metabolites in the module. Remarkably, at the second level (Lower), for most of the modules all of the metabolites are classified in the same pathway. Moreover, at the second level, we detect smaller pathways that are not visible at the top level.

module), we find that the most abundant pathway comprises >50% of the metabolites in the module.

For the second level, we find that for most of the modules, all of the metabolites are classified in the same pathway. We also detect smaller pathways that are not visible at the top level (such as those for polyketides and nonribosomal peptides, and for secondary metabolites).

Conclusion

Our analysis of model and real-world networks demonstrates that our algorithm is able to provide an objective multiscale description of complex systems that, although not affected by human subjectivity, captures our current understanding of these systems. For example, for the air-transportation network, our method extracts features all the way from continents to country boundaries. Further, our algorithm can be easily generalized to other classes of graphs such as bipartite graphs (e.g., collaboration networks or protein interaction networks obtained from bait-prey data) and directed graphs (e.g., citation networks, food webs, or gene-regulatory networks). The steps of the method would remain unchanged; one would only have to replace Eq. 1 by a suitable measure of modularity for the graph under analysis (45).

Interestingly, for metabolic networks, the algorithm reveals that “known” pathways do not correspond to a single module at the top level, implying that large pathways are in fact composed of smaller units. Intriguingly, these units are not necessarily uniform in “pathway composition” but are a mixture of submodules associated to different pathways. Our results thus prompt the question of how the modules we identify relate to metabolism evolution (46).

More generally, our results have significant implications for systems level approach to the study of cellular processes. A systems approach will only be successful if we are able to develop methods that enable us to extract the small set of information that is significant at the chosen scale of observation, whether this scale is molecular or organismal. A scalable multiscale representation of a biological process, such as the one we demonstrate here, will guide the purposeful design and re-engineering of biological systems for therapeutic purposes.

We thank U. Alon, A. Arenas, and S. Itzkovitz for providing network data and W. Jiang for advice with the statistical analysis. M.S.-P. and R.G. thank the Fulbright Program and the Spanish Ministry of Education, Culture & Sports. L.A.N.A. gratefully acknowledges the support of the Keck Foundation, the J. S. McDonnell Foundation, and a National Institutes of Health/National Institute of General Medical Sciences K-25 Award.

- Pennisi E (2005) *Science* 309:94.
- Oltvai ZN, Barabási A-L (2002) *Science* 298:763–764.
- Alon U (2003) *Science* 301:1866–1867.
- Itzkovitz S, Levitt R, Kashtan N, Milo R, Itzkovitz M, Alon U (2005) *Phys Rev E* 71:016127.
- Arenas A, Diaz-Guilera A, Pérez-Vicente CJ (2006) *Phys Rev Lett* 96:114102.
- Amaral LAN, Ottino J (2004) *Eur Phys J B* 38:147–162.
- Trusina A, Maslov S, Minnhagen P, Sneppen K (2004) *Phys Rev Lett* 92:178702.
- Reichardt J, Bornholdt S (2004) *Phys Rev Lett* 93:218701.
- Ziv E, Middendorf M, Wiggins CH (2005) *Phys Rev E* 71:046117.
- Bollobás B (2001) *Random Graphs* (Cambridge Univ Press, Cambridge, UK), 2nd Ed.
- Ravasz E, Somera AL, Mongru DA, Oltvai, ZN, Barabási AL (2002) *Science* 297:1551–1555.
- Soffer SN, Vázquez A (2005) *Phys Rev E* 71:057101.
- Guimerà R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2003) *Phys Rev E* 68:065103.
- Clauset A, Moore C, Newman MEJ (2006) *Lecture Notes Comput Sci* 4503:1–13.
- Pons P (2006) *Post-Processing Hierarchical Community Structures: Quality Improvements and Multi-Scale View*, cs.DS/0608050.
- Ye Y, Godzik A (2004) *Genome Res* 14:343–353.
- Ihmels J, Bergmann S, Berman J, Barkai N (2005) *PLoS Genet* 1:e39.
- Newman MEJ, Girvan M (2004) *Phys Rev E* 69:026113.
- Guimerà R, Sales-Pardo M, Amaral LAN (2004) *Phys Rev E* 70:025101.
- Fortunato S, Barthélemy M (2007) *Proc Natl Acad Sci USA* 104:36–41.
- Kumpula JM, Saramaki J, Kaski K, Kertesz J (2007) *Eur Phys J B* 56:41–45.
- Stilling FH, Weber TA (1982) *Phys Rev A* 25:978–989.
- Danon L, Diaz-Guilera A, Duch J, Arenas A (2005) *J Stat Mech Theory Exp*, P09008.
- Newman MEJ (2006) *Proc Natl Acad Sci USA* 103:8577–8582.
- Everitt BS, Landau S, Leese M (2001) *Cluster Analysis* (Arnold, London).
- Schena M (2002) *Microarray Analysis* (Wiley, New York).
- Maslov S, Sneppen K (2002) *Science* 296:910–913.
- Guimerà R, Sales-Pardo M, Amaral LAN (2007) *Nat Phys* 3:63–69.
- Kirkpatrick S, Gelatt CD, Vecchi MP (1983) *Science* 220:671–680.
- Wasserman S, Faust K (1994) *Social Network Analysis* (Cambridge Univ Press, Cambridge, UK).
- Koopmans T, Beckmann M (1975) *Econometrica* 25:53–76.
- Tsafirir D, Tsafirir I, Ein-Dor L, Zuk O, Notterman DA, Domany E (2005) *Bioinformatics* 21:2301–2308.
- Schwarz G (1978) *Ann Stat* 6:461–464.
- Akaike H (1974) *IEEE Trans Automatic Control* 19:716–723.
- Barrat A, Barthélemy M, Pastor-Satorras R, Vespignani A (2004) *Proc Natl Acad Sci USA* 101:3747–3752.
- Guimerà R, Amaral LAN (2004) *Eur Phys J B* 38:381–385.
- Guimerà R, Mossa S, Turtchi A, Amaral LAN (2005) *Proc Natl Acad Sci USA* 102:7794–7799.
- Guimerà R, Amaral LAN (2005) *Nature* 433:895–900.
- Thiele I, Vo TD, Price ND, Pálsson BØ (2005) *J Bacteriol* 187:5818–5830.
- Goto S, Nishioka T, Kanehisa M (1998) *Bioinformatics* 14:591–599.
- Kanehisa M, Goto S (2000) *Nucleic Acids Res* 28:27–30.
- Ma H, Zeng A-P (2003) *Bioinformatics* 19:270–277.
- Reed JL, Vo TD, Schilling CH, Pálsson BØ (2003) *Genome Biol* 4:R54.
- Guimerà R, Sales-Pardo M, Amaral LAN (2007) *Bioinformatics* 23:1616–1622.
- Guimerà R, Sales-Pardo M, Amaral LAN (2007) *Phys Rev E* 76:036102.
- Raymond J, Segrè D (2006) *Science* 311:1764–1767.