

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

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.

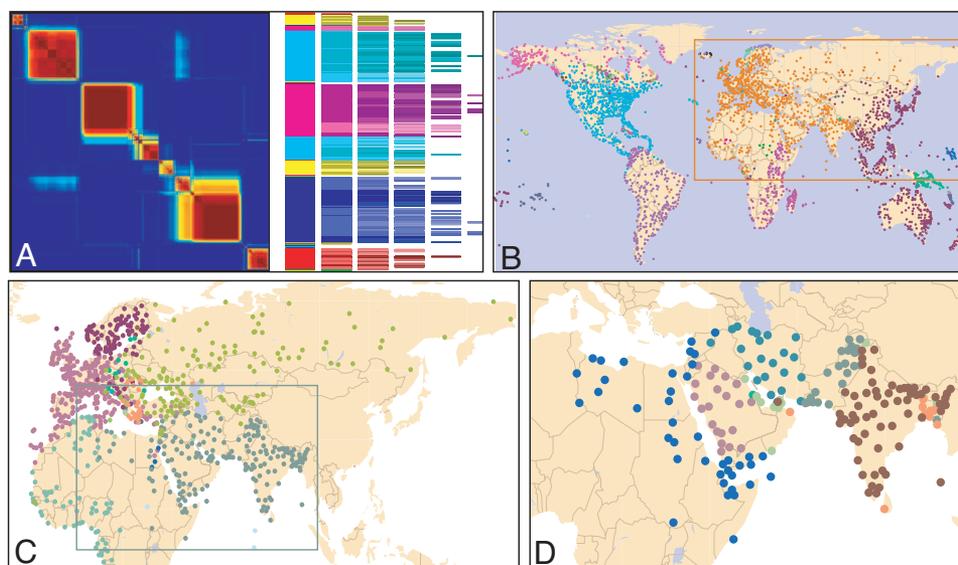


Fig. 3. Hierarchical organization of the air-transportation network. (A) Global-level affinity matrix and hierarchical tree (the representation is the same used in Fig. 2). (B) Top-level modules. Each dot represents a city and different colors represent different modules. Note that the top level in the hierarchy corresponds to major geopolitical units. (C) The “Eurasian” module (which is composed of the majority of European countries, ex-Soviet Union countries, Middle-Eastern countries, India, and countries in Northern half of Africa) splits for levels $\ell = 2$ into five submodules. (D) The “Near and Middle East” submodule further splits into seven submodules for $\ell = 3$ (D). Note that the air-transportation network is large and very dense (Table 1), and thus the organization of the network is not at all apparent (SI Fig. 11). Remarkably, the modules our method detects show a clear agreement with geopolitical units.

matrix. As we said previously, we assume that the information at different levels in the hierarchy is decoupled, thus to detect submodules beyond the first level, one needs to break the network into the subnetworks defined by each module and apply the same procedure from the start. The algorithm iterates these steps for each identified box until no subnetworks are found to have internal structure.

Method Validation

We validate our method on hierarchically nested random graphs with one, two, and three hierarchical levels. We define the accuracy of the method as the mutual information between the empirical partition and the theoretical one (23). Fig. 2C shows that the algorithm uncovers the correct number of levels in the hierarchy.

Moreover, our method always detects the top level, even for the networks with three hierarchical levels. In contrast, because the partition that globally maximizes M corresponds to the submodules in the second level, even the more accurate module identification algorithms based on modularity maximization would fail to capture the top level organization (20).

The hierarchically nested random graphs considered above have a homogeneous hierarchical structure; however, real-world networks are not likely to be so regular. In particular, for real-world networks, one expects that some modules will have deeper hierarchical structures than others. We thus have verified that our method is also able to correctly uncover the organization of model networks with heterogeneous hierarchical structures (see SI Fig. 10).

Analysis of Real-World Networks

Having validated our method, we next analyze different types of real-world networks for which we have some insight into the network structure: the worldwide air-transportation network (35–37), an e-mail exchange network of a Catalan university (13), and an electronic circuit (4).

In the air-transportation network, nodes correspond to cities (that is, all airports around major cities would be merged into a single node), and two nodes are connected if there is a nonstop flight connecting them. In the e-mail network, nodes are people and two people are connected if they send e-mails to each other. In the electronic network, nodes are transistors and two transistors are connected if the output of one transistor is the input of the other (Table 1).

We find that the air-transportation network is strongly modular and has a deep hierarchical organization (Fig. 3). This finding does

not come as a surprise because historical, economic, political, and geographical constraints shape the topology of the network (35–37). We find eight main modules that closely match major continents and subcontinents and major political divisions, and thus they truly represent the highest level of the hierarchy.[†]

The electronic circuit network is comprised of eight D-flip-flops and 58 logic gates (4). Our method identifies two levels in the network (SI Fig. 12A). At the top level, modules comprise either a D-flip-flop plus some additional gates, or a group of logic gates. At the second level, the majority of modules comprise single gates.

For the e-mail network, five of the seven major modules at the top level (SI Fig. 12B) correspond to schools in the university, with >70% of the nodes in each of those modules affiliated with the corresponding school. The remaining two major modules at the top level are a mixture of schools and administration offices (often collocated on campus), which are distinctly separated at the second level. The second level also identifies major departments and groups within a school, as well as research centers closely related to individual schools.

Application to Metabolic Networks

Finally, we analyze the metabolic networks of *E. coli* obtained from three different sources[‡] (Fig. 4 and SI Fig. 13): the KEGG database (40, 41), the Ma-Zeng database (42), and the reconstruction compiled by Palsson’s Systems Biology Laboratory at the UCSD (43). In these networks, nodes are metabolites and two metabolites are connected if there is a reaction that transforms one into the other (44).

To quantify the plausibility of our classification scheme, we analyze the within-module consistency of metabolite pathway classification for the top and the second levels of the metabolic network of *E. coli* (43). For each module, we first identify the pathways represented; then, we compute the fraction of metabolites that are classified in the most abundant pathway. We find that there is a clear correlation between modules and known pathways: At the top level, for all of the modules except one (the central metabolism

[†]The ability of the present method to detect the top level is significant. A previous study coauthored by two of us identified 19 modules in the worldwide air-transportation network (37) by using the most accurate modularity maximization algorithm in the literature (38).

[‡]In the SI Text, we also show the organization obtained for the UCSD reconstruction of the metabolic network for *Helicobacter pylori* (39).

