Cartography of complex networks: modules and universal roles

Roger Guimerà and **Luís A Nunes Amaral**

NICO and Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL 60208, USA E-mail: rguimera@northwestern.edu and amaral@northwestern.edu

Abstract

Integrative approaches to the study of complex systems demand that one knows the manner in which the parts comprising the system are connected. The structure of the complex network defining the interactions provides insight into the function and evolution of the components of the system. Unfortunately, the large size and intricacy of these networks implies that such insight is usually difficult to extract. Here, we propose a method that allows one to systematically extract and display information contained in complex networks. Specifically, we demonstrate that one can (i) find modules in complex networks and (ii) classify nodes into universal roles according to their pattern of within- and between-module connections. The method thus yields a 'cartographic representation' of complex networks.

Keywords

network dynamics

1. Introduction

Integrative approaches to the study of complex systems demand that one knows the manner in which the parts comprising the system are connected. Recent studies have revealed that these interaction networks often display complex features themselves, thus deviating from purely random or purely ordered topologies [1]-[8]. The Internet [9], food webs [10]-[12], and metabolic networks [13] are examples of complex networks.

To extract the relevant information from the topology of large complex networks, knowledge of the role of each node is of crucial importance. A cartographic analogy helps to illustrate this point. Consider the network formed by all cities and towns in a country—the nodes—and all the roads that connect them—the links. It is clear that a map in which each city and town is represented by a circle of fixed size and each road is represented by a line of fixed width is of little use. Rather, real maps emphasize capitals and important communication lines so that one can obtain scale-specific information at a glance. Similarly, it is difficult, if not impossible, to obtain information from a network with hundreds or thousands of nodes and links, unless the information about nodes and links is presented in a scale-specific context.

Here, we propose a methodology, which is based on the connectivity of the nodes, that yields a 'cartographic representation' of a complex network. The first step in our method is to identify the modules [14] in the network. In the cartographic picture, modules are analogous to countries or regions, and enable a coarse-grained, and thus simplified, description of the network. Then, we classify the nodes in the network into a small number of *system-independent* 'universal roles'.

2. Modules in complex networks

It is a matter of common experience that social networks have communities of highly interconnected nodes that are less connected to nodes in other communities. Such modular structures have been reported not only in social networks [14]-[17], but also in the Internet [18], in food webs [19,20] and in biochemical networks [21]-[24]. It is widely believed that the modular structure of complex networks plays a critical role in their functionality [21,22, 25]. There is therefore a clear need to develop algorithms for identifying modules accurately [14,16,18,26,27].

2.1. Modularity

Here we propose a method that is based on the maximization of the *modularity* [16,26,28]. For a given partition of the nodes of a network into modules, the modularity M of this partition is

$$M \equiv \sum_{s=1}^{N_\mathrm{M}} \left[\frac{l_s}{L} - \left(\frac{d_s}{2L} \right)^2 \right],$$

(1)

where N_M is the number of modules, L is the number of links in the network, l_s is the number of links between nodes in module s, and d_s is the sum of the degrees of the nodes in module s. The rationale for this definition of modularity is the following. A good partition of a network into modules must comprise many within-module links and as few as possible between-module links. However, if one just tries to minimize the number of between-module links (or, equivalently, maximize the number of within-module links) the optimal partition consists of a single module and no between-module links. Equation (1) addresses this difficulty by imposing that $M = 0$ if nodes are placed at random into modules *or* if all nodes are in the same cluster [16,26,28].

2.2. Simulated annealing for module identification

The objective of a module identification algorithm is to find the partition with largest modularity, and several methods have been proposed for attaining such a goal. Most of them rely on heuristic procedures and use M—or a similar measure—only to assess their performance. In contrast, we use simulated annealing (SA) [29,28] to obtain the best determination of the modules of a network by direct maximization of M.

Simulated annealing [29] is a stochastic optimization technique that enables one to find 'low cost' configurations without getting trapped in 'high cost' local minima. This is achieved by introducing a *computational temperature T*. When T is high, the system can explore configurations of high cost while at low T the system only explores low cost regions. Starting at high T and slowly decreasing it, the system descends gradually toward deep minima, overcoming small cost barriers.

When identifying modules, the objective is to maximize the modularity and, thus, the cost is $C = -M$, where M is the modularity as defined in equation (1). At each temperature, we perform a number of random updates and accept them with probability

$$p = \begin{cases} 1 & \text{if } C_f \leq C_i \\ \exp\left(-\frac{C_f - C_i}{T} \right) & \text{if } C_f > C_i \end{cases}$$

(2)

where C_f is the cost after the update and C_i is the cost before the update.

Specifically, at each T we propose $n_i = fS^2$ individual node movements from one module to another, where S is the number of nodes in the network. We also propose $n_c = fS$ collective movements, which involve either merging two modules or splitting a module. For f we typically choose $f = 1$. After the movements are evaluated at a certain T, the system is cooled down to $T' = cT$, with $c = 0.995$.

Both the individual movement of a node from a module to another, and the merging of two modules are straightforward movements. The split movement is, however, more involved. In principle, any heuristic procedure that generates non-deterministic plausible splits of a module —that is, splits that have a finite probability of being accepted—would suffice.

We have considered a number of different schemes. The one that performs best consists in isolating the module from the rest of the network, and performing a 'nested' SA, entirely independent of the 'global' one. We start by isolating the module that we intend to split. We then partition its nodes into two random groups, and set the initial temperature of the nested SA to a high value. Then, we make individual node movements according to the modularity of the isolated module and decrease the temperature of the nested SA until it reaches the value of the global SA. The result of the nested SA is a partition of the module into two new modules, which we accept or reject according to the global modularity and the global temperature.

In using M as a 'fitness function', our method is more 'transparent' than those relying on heuristic procedures. Furthermore, SA enables us to carry out an exhaustive search and to minimize the problem of finding sub-optimal partitions. It is noteworthy that, in our method, one does not need to specify *a priori* the number of modules; rather, the number of modules is an outcome of the algorithm.

To test the performance of the method, we build 'random networks' with known module structure. Each test network comprises 128 nodes divided into 4 modules of 32 nodes. Each node is connected to the other nodes in its module with probability p_{in}, and to nodes in other modules with probability $p_{out} < p_{in}$. On average, thus, each node is connected to $k_{out} = 96p_{out}$ nodes in other modules and to $k_{in} = 31p_{in}$ in the same module. Additionally, p_{in} and p_{out} are selected so that the average degree of the nodes is $k = 16$. Our algorithm, which significantly outperforms the best algorithm in the literature [14], is able to reliably identify modules in a network whose nodes have as many as 50% of their connections outside their own module (figure 1).

3. Roles in complex modular networks

Already in 1957, Nadel argued that 'roles' are the central elements in the analysis of social systems [30,31], and in the 1970s White and co-workers introduced the concepts of *structural equivalence* and *block model* to address this issue from a network perspective [32]-[34], [31]. Two nodes are structurally equivalent if they are connected to the same nodes [32,34]. Therefore, any network can be divided into blocks of structurally equivalent nodes in such a way that the structure of the network is *summarized* in a block model by stating the relations between the blocks (figure 2).

Usually, structural equivalence is too strong a requirement for a large complex network; it is very unlikely that two nodes are connected to exactly the same set of other nodes. Regular structural equivalence [35,34] relaxes this requirement by requiring that regularly equivalent nodes have identical links to other *equivalent* nodes. Formally, if nodes i and j are regularly equivalent and i has a link to/from some node k, then node j must have a link to/from some node l, and nodes k and l must be, also, regularly equivalent [34].

Real networks are likely to have both a modular structure and a block structure. This fact raises serious concerns about the conceptual relationship between blocks and roles. Although blocks certainly give interesting information about the overall structure of the network, simple examples, such as the one shown in figure 3, demonstrate that, in general, *blocks cannot be interpreted as roles*.

Motivated by this handicap of the block scheme, we propose a new method for determining the role of a node in a complex network. Our approach is not based on the idea of blocks but on the general idea that nodes with the same role should have similar topological properties (figure 3(b)).

3.1. Within-module degree and participation coefficient

Each module can be organized in very different ways, ranging from totally centralized—with one or a few nodes connected to all the others—to totally decentralized—with all nodes having similar connectivities. Nodes with similar roles are expected to have similar relative within-module connectivity. If κ_i is the number of links of node i to other nodes in its module s_i, $\bar{\kappa}_{s_i}$ is the average of κ over all the nodes in s_i, and $\sigma_{\kappa_{s_i}}$ is the standard deviation of κ in s_i, then

$$z_i = \frac{\kappa_i - \bar{\kappa}_{s_i}}{\sigma_{\kappa_{s_i}}}$$

(3)

is the so-called *z*-score. The within-module degree *z*-score measures how 'well connected' node *i* is to other nodes in the module.

Different roles can also arise because of the connections of a node to modules other than its own. For example, two nodes with the same *z*-score will probably play different roles if one of them is connected to several nodes in other modules while the other is not. We define the participation coefficient P_i of node *i* as

$$P_i = 1 - \sum_{s=1}^{N_M} \left(\frac{\kappa_{is}}{\kappa_i} \right)^2$$

(4)

where k_{is} is the number of links of node *i* to nodes in module *s*, and k_i is the total degree of node *i*. The participation coefficient of a node is therefore close to one if its links are uniformly distributed among all the modules and zero if all its links are within its own module.

We hypothesize that the role of a node can be determined, to a great extent, by its *within-module degree* and its *participation coefficient*, which define how the node is positioned in its own module and with respect to other modules [36,37]. Note that these two properties are easily computed once the modules of a network are known.

3.2. Arguments for the definition of a universal set of discrete roles

We surmise that the role of a node is defined mainly by its within-community degree and its participation coefficient. Our definition of the roles is firstly determined by the within-module degree. We classify nodes with $z \geq 2.5$ as module hubs and nodes $z < 2.5$ as non-hubs. Both hub and non-hub nodes are then more finely characterized by using the values of the participation coefficient. Simple calculations suggest that non-hub nodes can be naturally assigned into four roles:

- *Ultra-peripheral nodes* (role R1).

If a node has all its links within its module ($P \approx 0$).

- *Peripheral nodes* (role R2).

   If a node has at least 60% its links within the module, then for $k < 4$ it follows that $P < 0.625$ (figure 4(a)).

- *Non-hub connectors* (role R3).

   If a node with $k < 4$ has half of its links (or at least two links, whichever is larger) within the module, then it follows that $P < 0.8$ (figure 4(a)). Thus, a plausible region for non-hub connectors is $0.62 < P < 0.8$.

- *Non-hub kinless nodes* (role R4).

   If a node has fewer than 35% of its links within the module, it implies that $P > 0.8$. We surmise that such nodes cannot be clearly assigned to a single module. We thus classify them as kinless nodes. We will demonstrate later that non-hub kinless nodes are found in most network growth models, but not in real-world networks.

Similarly, hubs can be naturally assigned into three different roles:

- *Provincial hubs* (role R5).

   If a node with a large degree, $k \gg 1$, has at least $5/6$ of its links within the module, then it follows that $P = 1 - (5/6)^2 - (k/6)(1/k^2) = 0.31 - 1/(6k) \approx 0.30$.

- *Connector hubs* (role R6).

   If a node with a large degree has at least half of its links within the module, then it follows that $P = 1 - 1/4 - (k/2)(1/k^2) = 0.75 - 1/(2k)$. Since $k \gg 1$, $P < 0.75$ for such nodes.

- *Kinless hubs* (role R7).

   If a hub has fewer than half its links within the module, i.e., $P > 0.75$, then we surmise that it may not be clearly associated with a single module. We then classify it as a kinless hub. We will demonstrate later that hubs in most network growth models are actually kinless hubs.

In total, we are left with seven roles that correspond to seven regions of the zP parameter space (figure 5).

4. Roles in real networks: validation of the role definitions

Our definition of a set of distinct roles has been, so far, based on mathematical arguments. A question that we need to address is, therefore, how this definition relates to the roles of nodes in real networks.

In order to obtain as complete as possible a picture of how the nodes in a given network might populate the zP parameter space, we calculate z and P values for all the nodes in a large number of networks (figure 6). Specifically, we obtain these values for (i) the metabolic networks of three organisms, (ii) the proteome of *C. elegans*, (iii) the North-American airport network, (iv) the collaboration networks of chemical engineers as defined by publications in two different journals, (v) the Internet at the autonomous system level. Additionally, we obtain these values for nodes in model networks generated by the Barabási–Albert network growth model [3] and the Erdös–Rényi model [40]. In all, we consider in our analysis 26 771 nodes.

4.1. Uncertainty in the position of nodes in parameter space and the density landscape

In our analysis, we estimate the value of the within-module degree of each node and its participation coefficient. Since we have access to these networks at a single moment in time, it is plausible to assume that the values that we measure for z_m and P_m for a given node are not error free. To take this uncertainty into consideration, we assume that each node could be in a *region* of the zP space, which is centred in the measured (z_m, P_m) value. Specifically, we assign to each node a Gaussian distribution centred at (z_m, P_m) and with widths σ_z and σ_P, which gives the probability of finding that particular node at any point of the zP parameter space.

By adding the distributions of all nodes, one obtains a 'density landscape' that represents the probability of finding a node at a certain point of the zP space. In figure 7, we plot the density landscape obtained for the 26 771 nodes with $\sigma_P = 0.035$. In the density landscape, high probability regions are valleys and low probability regions are peaks. Then, at (almost) every point of the landscape, one can 'follow' the gradient to reach a local minimum. The region of the space that 'flows' toward a certain minimum is what we call a 'basin of attraction'.

4.2. Non-hub nodes

As discussed above, we define non-hub nodes as those with $z < 2.5$. We then calculate the node density plot for different choices of the values of σ_z and σ_P and identify the basins of attraction for the different node density plots (figure 8). These plots confirm that our definition of non-hub roles with boundaries at $P = 0.62$ and $P = 0.80$ is a sensible one and that, indeed, these regions of the zP space correspond to distinct universal roles in real networks.

4.3. Hub nodes

We define hub nodes as those with $z \leq 2.5$. We then calculate the node density plot for different choices of the values of σ_z and σ_P and identify the basins of attraction for the different node density plots (figure 9).

In this case, there are many more basins of attraction than for the non-hub region because of the scarcity of data points. However, the density plots are compatible with a selection of three regions corresponding to distinct roles, with boundaries at $P = 0.30$ and 0.75, as estimated before.

5. Conclusions

Computational and high throughput techniques are leading to an explosive and unprecedented growth in the amount of information available for some physical, biological, and socio-economic systems. These advances are creating the opportunity to revolutionize our understanding of nature, life and disease, and social organization. Interpretation of these data remains, however, a major scientific challenge.

Here, we presented a methodology for extracting relevant scale-specific information from complex networks. Our method is based on the analysis of the connectivity patterns of the nodes, and yields a 'cartographic representation' of a complex network. The first step in our method is to identify the modules in the network. In the cartographic picture, modules are analogous to countries or regions, and enable a coarse-grained, and thus simplified, description of the network. Then, we classify the nodes in the network into a small number of *system-independent* 'universal roles'. A node's role is determined from its pattern of inter- and intra-module connections.

Our 'cartographic method' provides a way to process the information contained in the structure of complex networks, and to extract knowledge about the function carried out by the network

and its constituents. This should allow us, in turn, to identify key players in the network. Some of these key nodes are likely to be already known. For example, hubs are highly visible due to their large number of connections. More interestingly, our method also enables one to identify more 'subtle' roles, such as non-hub connectors, which play important structural roles in spite of their small number of connections. In metabolic networks, for example, it seems that these nodes are highly conserved compared to provincial hubs [24].

Acknowledgments

References

1. Watts DJ, Strogatz SH. Nature 1998;393:440. [PubMed: 9623998]

2. Barthélémy M, Amaral LAN. Phys. Rev. Lett 1999;82:3180.

3. Barabási AL, Albert R. Science 1999;286:509. [PubMed: 10521342]

4. Amaral LAN, Scala A, Barthélémy M, Stanley HE. Proc. Nat. Acad. Sci 2000;97:11149. [PubMed: 11005838]

5. Albert R, Barabási AL. Rev. Mod. Phys 2002;74:47.

6. Dorogovtsev SN, Mendes JFF. Adv. Phys 2002;51:1079.

7. Newman MEJ. SIAM Rev 2003;45:167.

8. Amaral LAN, Ottino J. Eur. Phys. J. B 2004;38:147.

9. Vázquez A, Pastor-Satorras R, Vespignani A. Phys. Rev. E 2002;65:066130.

10. Camacho J, Guimerà R, Amaral LAN. Phys. Rev. Lett 2002;88:228102. [PubMed: 12059454]

11. Camacho J, Guimerà R, Amaral LAN. Phys. Rev. E 2002;65:030901.

12. Stouffer DB, Camacho J, Guimerà R, Ng CA, Amaral LAN. Ecology. 2005at press

13. Jeong H, Tombor B, Albert R, Oltvai ZN, Barabási AL. Nature 2000;407:651. [PubMed: 11034217]

14. Girvan M, Newman MEJ. Proc. Nat. Acad. Sci 2002;99:7821. [PubMed: 12060727]

15. Guimerà R, Danon L, Díaz-Guilera A, Giralt F, Arenas A. Phys. Rev. E 2003;68:065103.

16. Newman MEJ, Girvan M. Phys. Rev. E 2004;69:026113.

17. Arenas A, Danon L, Díaz-Guilera A, Gleiser PM, Guimerà R. Eur. Phys. J. B 2004;38:373.

18. Eriksen KA, Simonsen I, Maslov S, Sneppen K. Phys. Rev. Lett 2003;90:148701. [PubMed: 12731952]

19. Pimm SL. Theor. Popul. Biol 1979;16:144. [PubMed: 538731]

20. Krause AE, Frank KA, Mason DM, Ulanowicz RE, Taylor WW. Nature 2003;426:282. [PubMed: 14628050]

21. Hartwell LH, Hopfield JJ, Leibler S, Murray AW. Nature 1999;402:C47. [PubMed: 10591225]

22. Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. Science 2002;297:1551. [PubMed: 12202830]

23. Holme P, Huss M. Bioinformatics 2003;19:532. [PubMed: 12611809]

24. Guimerà R, Amaral LAN. Nature. 2005at press

25. Alon U. Science 2003;301:1866. [PubMed: 14512615]

26. Newman MEJ. Phys. Rev. E 2004;69:066133.

27. Radicchi F, Castellano C, Cecconi F, Loreto V, Parisi D. Proc. Nat. Acad. Sci 2004;101:2658. [PubMed: 14981240]

28. Guimerà R, Sales-Pardo M, Amaral LAN. Phys. Rev. E 2004;70:025101.

29. Kirkpatrick S, Gelatt CD, Vecchi MP. Science 1983;220:671. [PubMed: 17813860]

30. Nadel, SF. The Theory of Social Structure. Cohen and West; London: 1957.

31. Scott, J. Social Network Analysis: A Handbook. Vol. 2nd edn. SAGE Publications; London: 2000.

32. Lorrain F, White HC. J. Math. Sociol 1971;1:49.

33. White HC, Boorman SA, Breiger RL. Am. J. Sociol 1976;81:730.

34. Wasserman, S.; Faust, K. Social Network Analysis. Cambridge University Press; Cambridge: 1994.

35. Sailer LD. Soc. Networks 1978;1:73.

36. Rives AW, Galitski T. Proc. Nat. Acad. Sci 2003;100:1128. [PubMed: 12538875]

37. Han JDJ, Bertin N, Hao T, Goldberg DS, Berriz GF, Zhang LV, Dupuy D, Wallhout AJM, Cusick ME, Roth FP, Vidal M. Nature 2004;430:88. [PubMed: 15190252]

38. Guimerà R, Mossa S, Turtschi A, Amaral LAN. 2003submitted

39. Guimerà R, Amaral LAN. Eur. Phys. J. B 2004;38:381.

40. Bollobás, B. Random Graphs. Vol. 2nd edn. Cambridge University Press; Cambridge: 2001.
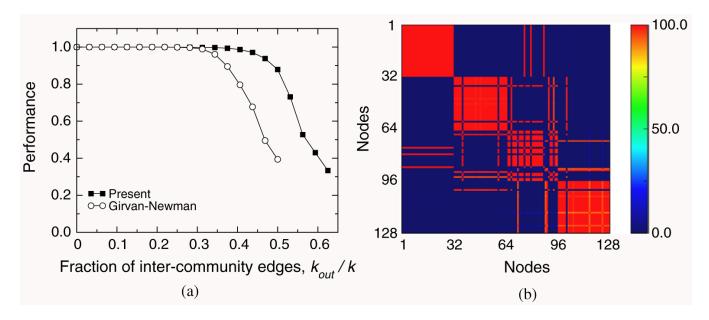
Figure 1.
(a) The performance of a module identification algorithm is typically defined as the fraction of correctly classified nodes. We compare our algorithm to the Girvan–Newman algorithm [14,26], which is the reference algorithm for module identification [16,26,27]. Note that our method is 90% accurate even when half of a node's links are to nodes in outside modules. (b) Our module identification algorithm is stochastic, so different runs yield, in principle, different partitions. To test the robustness of the algorithm, we obtain 100 partitions of a network with $k_{in} = k_{out} = 8$ and plot, for each pair of nodes in the network, the fraction of times that they are classified in the same module. As shown in the figure, most pairs of nodes are either always classified in the same module (red) or never classified in the same module (dark blue), which indicates that the solution is robust.
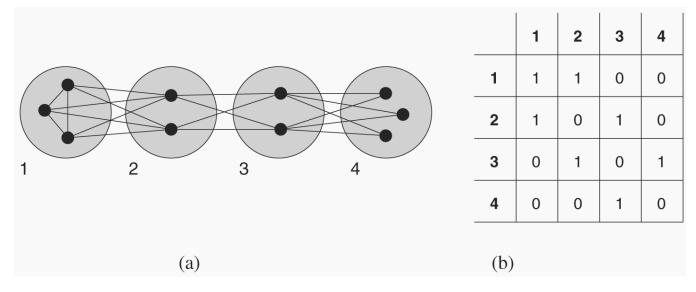
	1	2	3	4
1	1	1	0	0
2	1	0	1	0
3	0	1	0	1
4	0	0	1	0

(a)   (b)

Figure 2.
(a) Structural equivalence and blocks. The network depicted can be divided into four blocks of structurally equivalent nodes. All nodes in block 1 are connected to each other and to all nodes in block 2. All nodes in block 2 are connected to all nodes in blocks 1 and 3, and so forth. (b) The structure of the network can be conveniently summarized using a block model matrix.
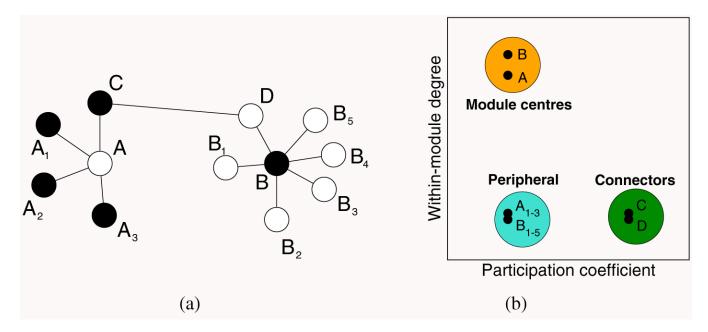
Figure 3.
Weaknesses of the block approach to the identification of roles in modular networks. (a) To illustrate the weaknesses of the block model approach to the identification of roles in modular networks, consider the network shown. Black nodes are connected to white nodes only, and white nodes are connected to black nodes only. Therefore, black nodes are regularly structurally equivalent to each other and white nodes are regularly structurally equivalent to each other. An ideal block detection algorithm may thus partition the nodes into two blocks, black and white. Significantly, this partition fails to capture the truly significant roles of the nodes in the network. Namely, nodes A and B are the 'centres' of their modules, nodes C and D are module connectors, and all the other nodes are peripheral. (b) Identification of roles based on the 'within-module degree' and the 'participation coefficient' (see the text for definitions).
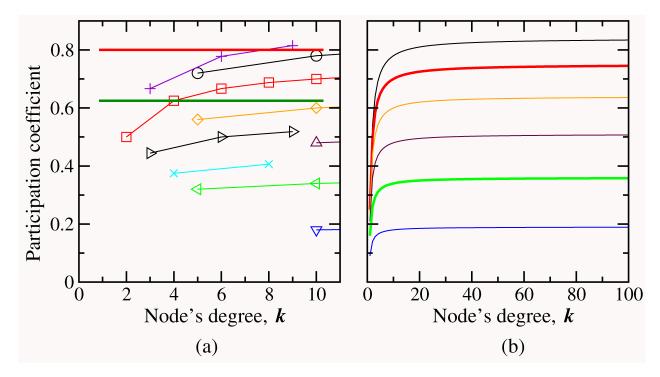
(a)   (b)

Figure 4.
The dependence of the value of the participation coefficient on the total degree and fraction of within-module links. (a) P for, from top to bottom, 1/3, 0.4, 1/2, 0.6, 0.66, 0.7, 0.75, 0.8, and 0.9 of within-module links. The red horizontal line corresponds to $P = 0.8$ and the dark green one to $P = 0.625$. These results suggest that $P > 0.8$ occurs only for cases in which the assignment of a node to a role is mostly a matter of chance. (b) P for, from top to bottom, 0.4, 1/2, 0.6, 0.7, 0.8, and 0.9 of within-module links. The red curve, which corresponds to half of the links being within the module, converges to $P = 0.75$. The green curve, which corresponds to 80% of the links being within the module, converges to $P = 0.35$.
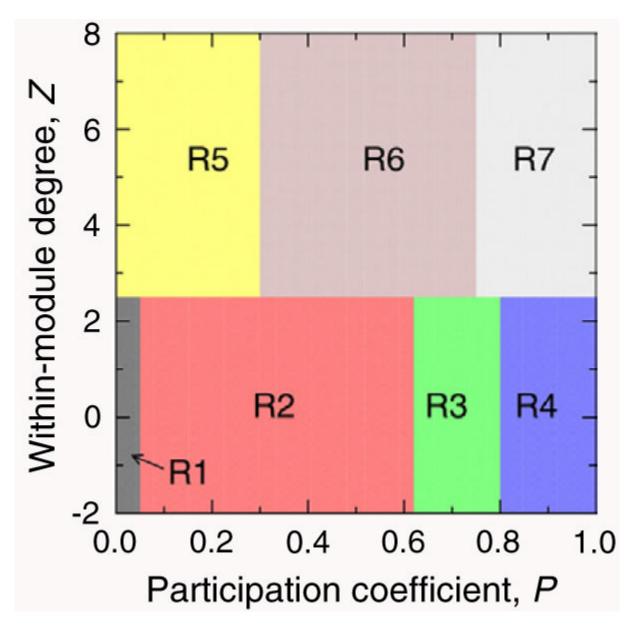
Figure 5.
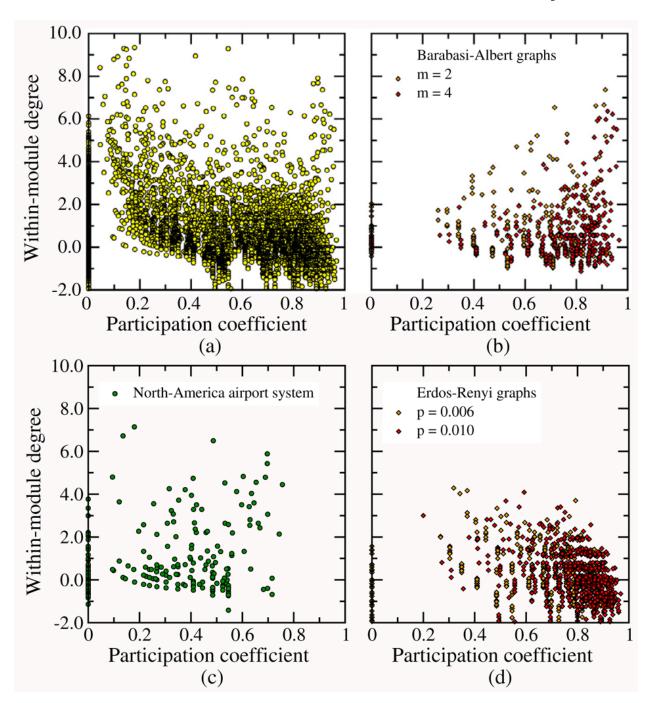Role-specific regions in the *zP* parameter space.

Figure 6.
(a) Values of z and P for 26 771 nodes from 16 networks, including the metabolic networks of three organisms, the proteome of *C. elegans*, the North-American airport network [38,39], the collaboration networks of chemical engineers obtained from two journals (*Chemical Engineering Science* and the *AIChE Journal*), the Internet at the autonomous system level [9], four Erdös–Rényi graphs with $p = 0.004$, 0.006, 0.008, and 0.010, and four Barabási–Albert graphs with $m = 1$, 2, 3, and 4. (b) Values of z and P for two Barabási–Albert graphs with 1000 nodes each. (c) Values of z and P for 940 nodes in the largest fully connected component of the North-American airport network. (d) Values of z and P for two Erdös–Rényi graphs with 1000 nodes each.
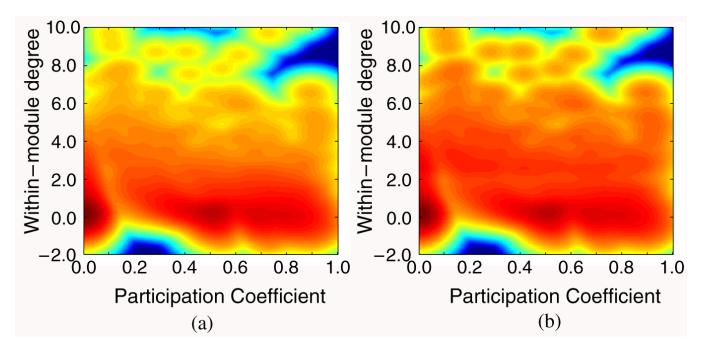
Figure 7.
The density landscape for the nodes belonging to eight real-world networks and eight model networks. Due to the fact that more than 98% of the nodes have $z < 2.5$, one finds that the density landscape for $z > 2.5$ is quite 'washed' down by the background of the non-hub region. For this reason, we obtain the density landscape under two distinct conditions: (a) in the first, we weigh each hub with a weight of one; (b) in the second, we weigh each hub with a weight of five.
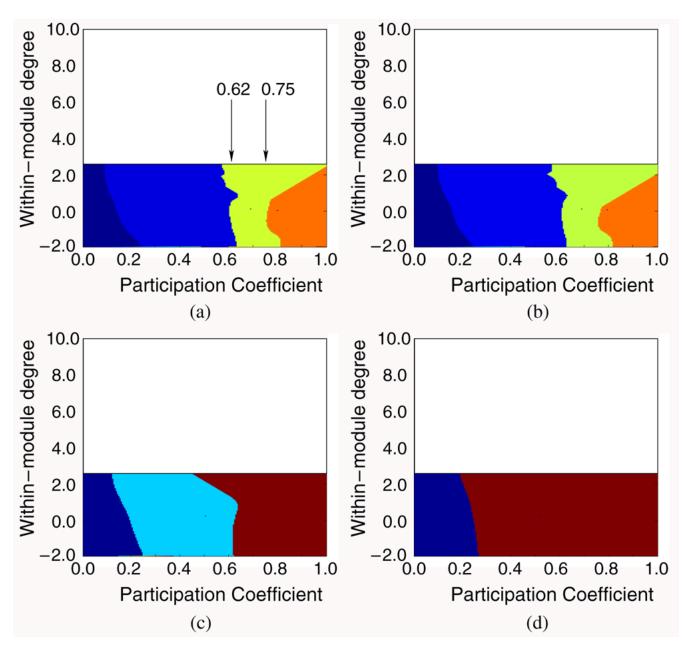
Figure 8.
Basins of attraction for density landscapes obtained for non-hub nodes obtained with (a) $\sigma_P = 0.03$, (b) $\sigma_P = 0.035$, (c) $\sigma_P = 0.05$, and (d) $\sigma_P = 0.08$. Note how the values of P identified in our simple analysis provide a good match to the boundaries of the basins of attraction in the node density landscapes.
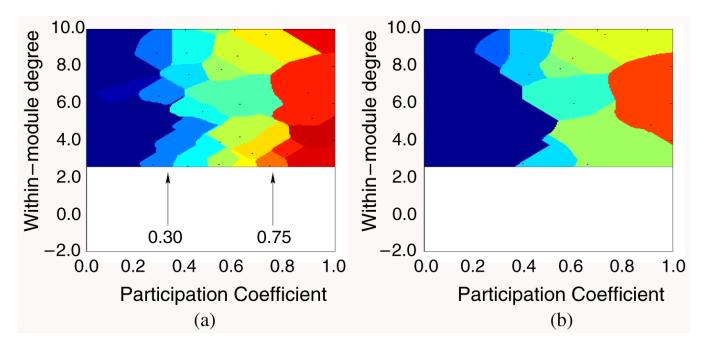
Figure 9.
Basins of attraction for density landscapes obtained for hubs with (a) $\sigma_P = 0.03$ and (b) $\sigma_P = 0.05$.