SUPPLEMENTARY DISCUSSION

Functional cartography of complex metabolic networks

Roger Guimerà and Luís A. Nunes Amaral

1 Roles and blocks

Already in 1957, Nadel argued that "roles" are the central elements in the analysis of social systems[1,2], and in the 1970s White and coworkers introduced the concepts of *structural equivalence* and *blockmodel* to address this issue from a network perspective[3,4,5,2]. Two nodes are structurally equivalent if they are connected to the same nodes[3,5]. 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 blockmodel by stating the relations between the blocks. Usually, structural equivalence is too strong a requirement for a large complex network; it is very unlikely that two nodes are connected to the exact same set of other nodes. Regular structural equivalence[6,5] relaxes this requirement by requiring that regularly equivalent nodes have identical links with other *equivalent* nodes (Fig. 1). 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[5].

Real networks are likely to have both modular structure and 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 Fig. 1c, demonstrate that, in general, *blocks cannot be interpreted as roles*.

Motivated by this handicap of the block-scheme, we propose a new method to determine

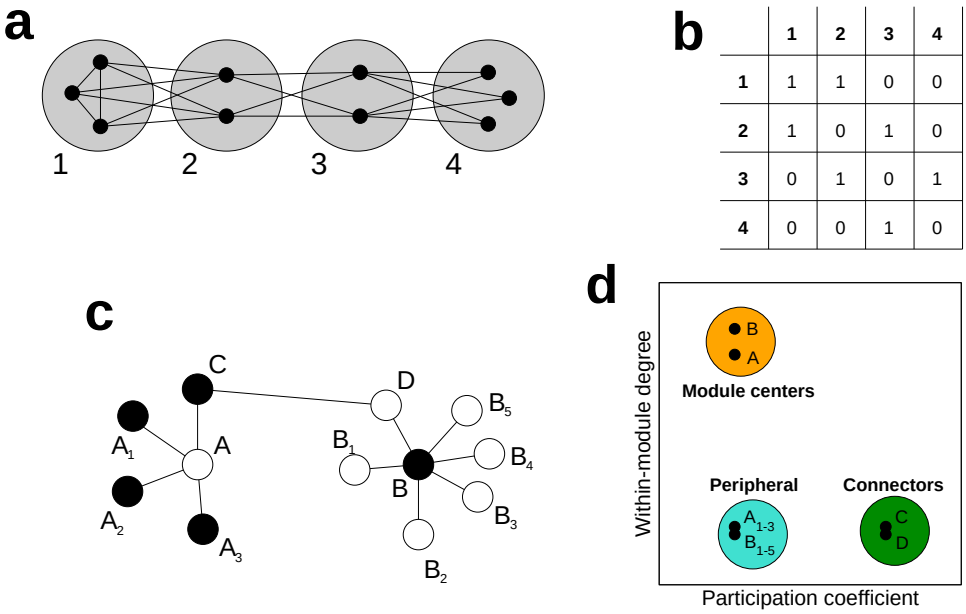

Figure 1: Weaknesses of the block approach to the identification of roles in modular networks. **a,** Structural equivalence and blocks. The network depicted can be divided into four blocks of structurally equivalent nodes. All nodes in block 1 are connected to each other and to all nodes in block 2. 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 blockmodel matrix. **c,** To illustrate the weaknesses of the blockmodel 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 "centers" of their modules, nodes C and D are module connectors, and all the other nodes are peripheral. **d,** Identification of roles based on the "within-module degree" and the "participation coefficient" (see Methods for definitions.)

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 (Fig. 1c,d).

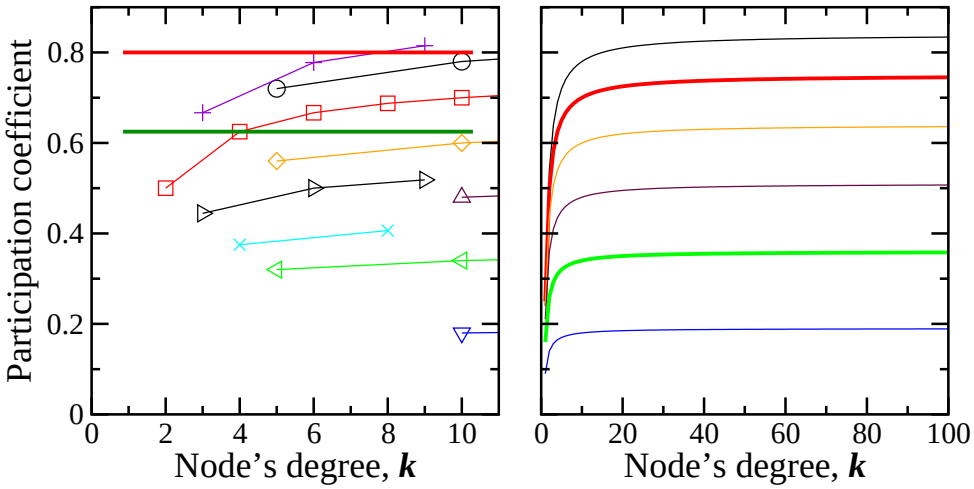

Figure 2: Dependence of value of participation coefficient on total degree and fraction of within-module links. **a,** P for, from top to bottom, 1/3, 0.4, 1/2, 0.6, 0.66, 0.7, 0.75, 0.8, and 0.9 of within-module links. The red horizontal line corresponds to $P = 0.8$ and the dark green to $P = 0.625$. 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 correspond to half of the links within-module, converges to $P = 0.75$. The green curve, which correspond to 80% of the links within-module, converges to $P = 0.35$.

2 Heuristic determination of a 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-module, then for $k < 4$ it follows that $P < 0.625$

(Fig. S2a).

- *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-module, then it follows that $P < 0.8$ (Fig. S2a). 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-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 80% of its links within-module, then it follows that $P = 1 - (0.8)^2 - (k/5)(1/k^2) = 0.36 - 1/(5k)$.

- *Connector hubs* (Role R6).

  IF a node with a large degree has at least half of its links within 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-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.

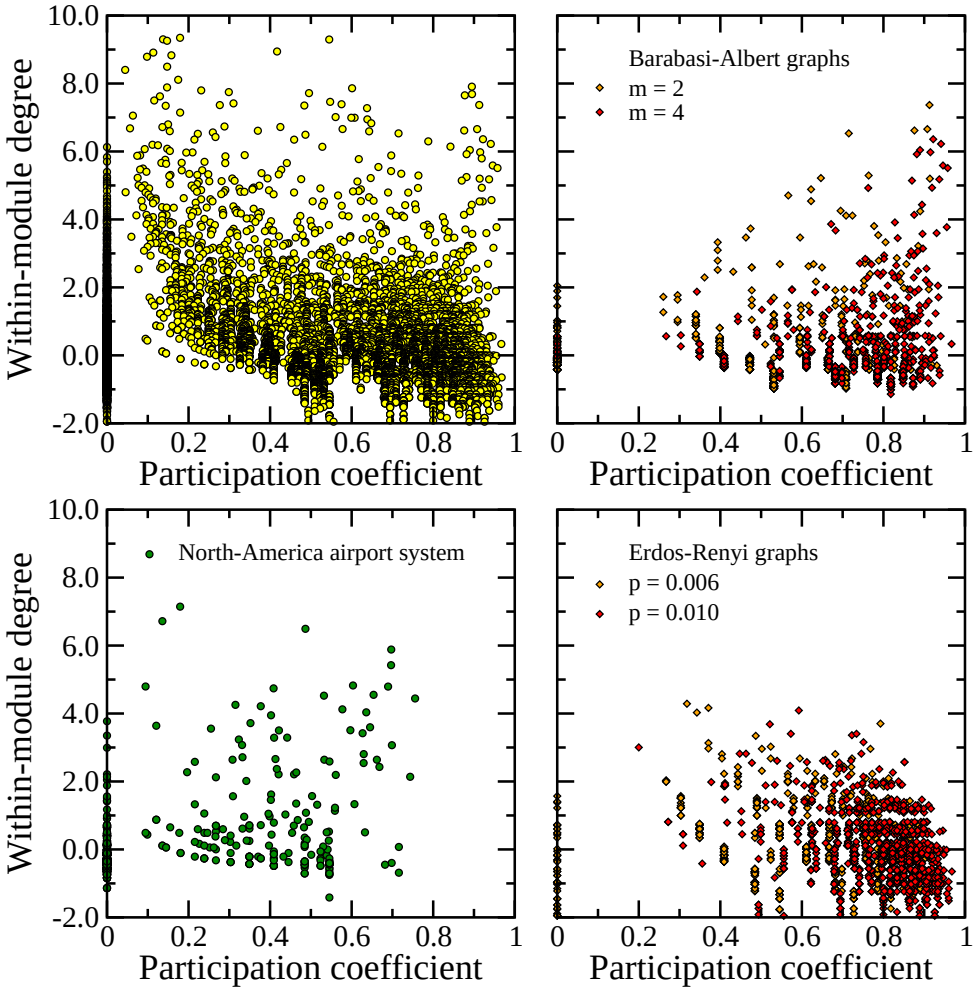

Figure 3: **a,** Values of z and P for 26771 nodes from 16 networks, including the metabolic networks of three organisms, the proteome of *C. elegans*, the North-American airport network, the collaboration networks of chemical engineers obtained from two journals (Chemical Engineering Science and AIChE Journal), the Internet at the autonomous system level, 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.

We will demonstrate later that hubs in most network growth models are actually kinless hubs.

2.1 Uncertainty in node position in parameters-space

In our analysis, we estimate the z value of the intra-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 we measure for z and P for a given node are not error-free. To take this uncertainty into consideration, we assign to each node a Gaussian peak with specific widths σ_z and σ_P. Figure S4 show the spread of the Gaussian peaks for two values of these parameters.

In order to obtain as complete as possible a picture of how the nodes in a given network might populate the zP parameters-space, we calculate z and P values for all the nodes in a large number of networks. 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 and the Erdös-Rényi model. In all, we consider in our analysis 26,771 nodes. We plot the density

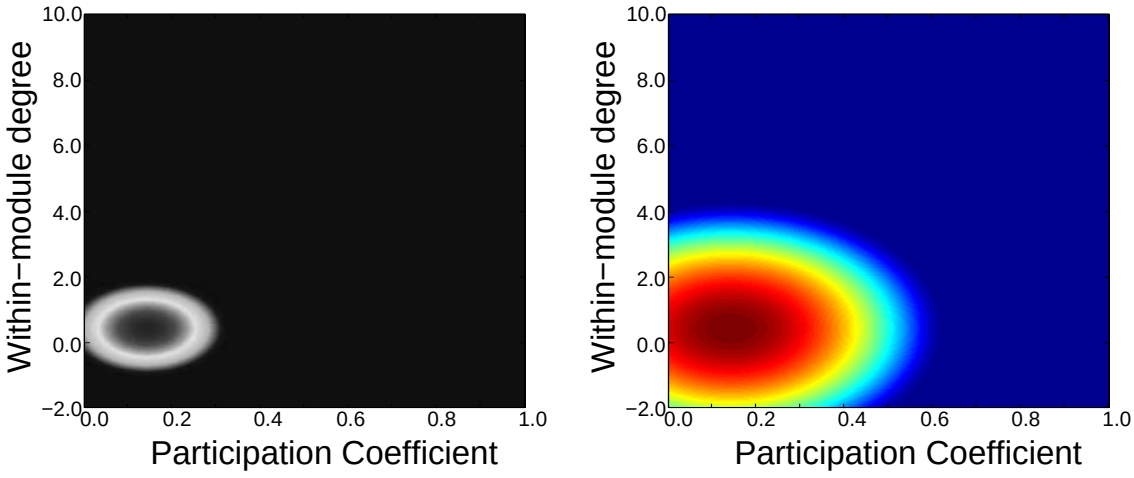

Figure 4: Gaussian peak decay for a single node as a function of the peak width. Note that the scale is logarithmic Node density for **a,** $\sigma_P = 0.03$, and **b,** $\sigma_P = 0.08$. In both cases, $\sigma_z = 10\sigma_P$.

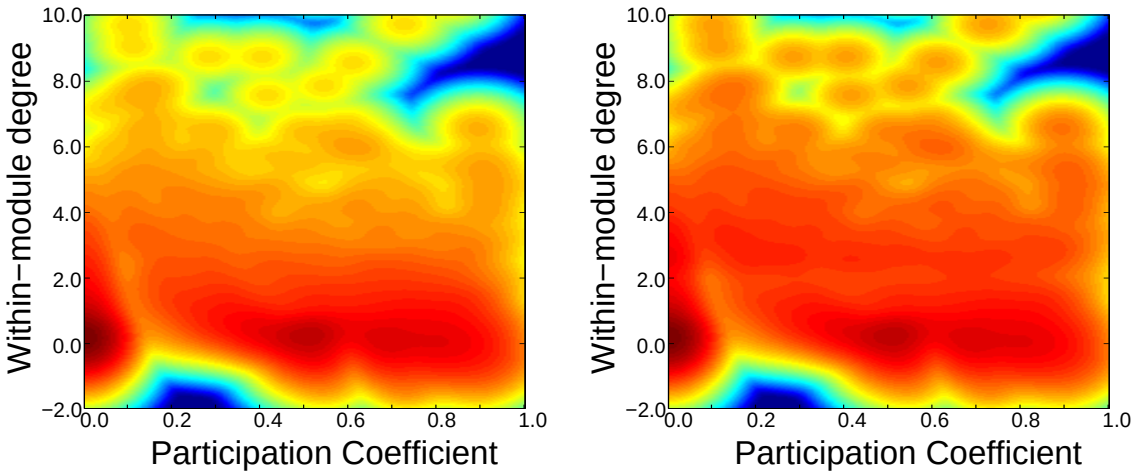

Figure 5: Density landscape for the nodes belonging to 8 real-world networks and 8 model networks. Due to the fact that more than 98% of the nodes have $z < 2.5$, one finds that the density landscape for $z > 2.5$ is quite "washed" down by the background of the non-hub region. For this reason, we obtain 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.

landscape obtained for these nodes with $\sigma_P = 0.035$ in Fig. S5.

2.2 "Basins of attraction" for non-hub nodes

One can see the probability of finding a node with given values of z and P as a density landscape, with high probability regions as valleys and low probability regions as 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."

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 (Fig. S6).

Based on the results of Figs. S2–S6, we partition the zP parameters space for $z < 2.5$ into four regions with boundaries at $P = 0.05, 0.62$ and 0.8.

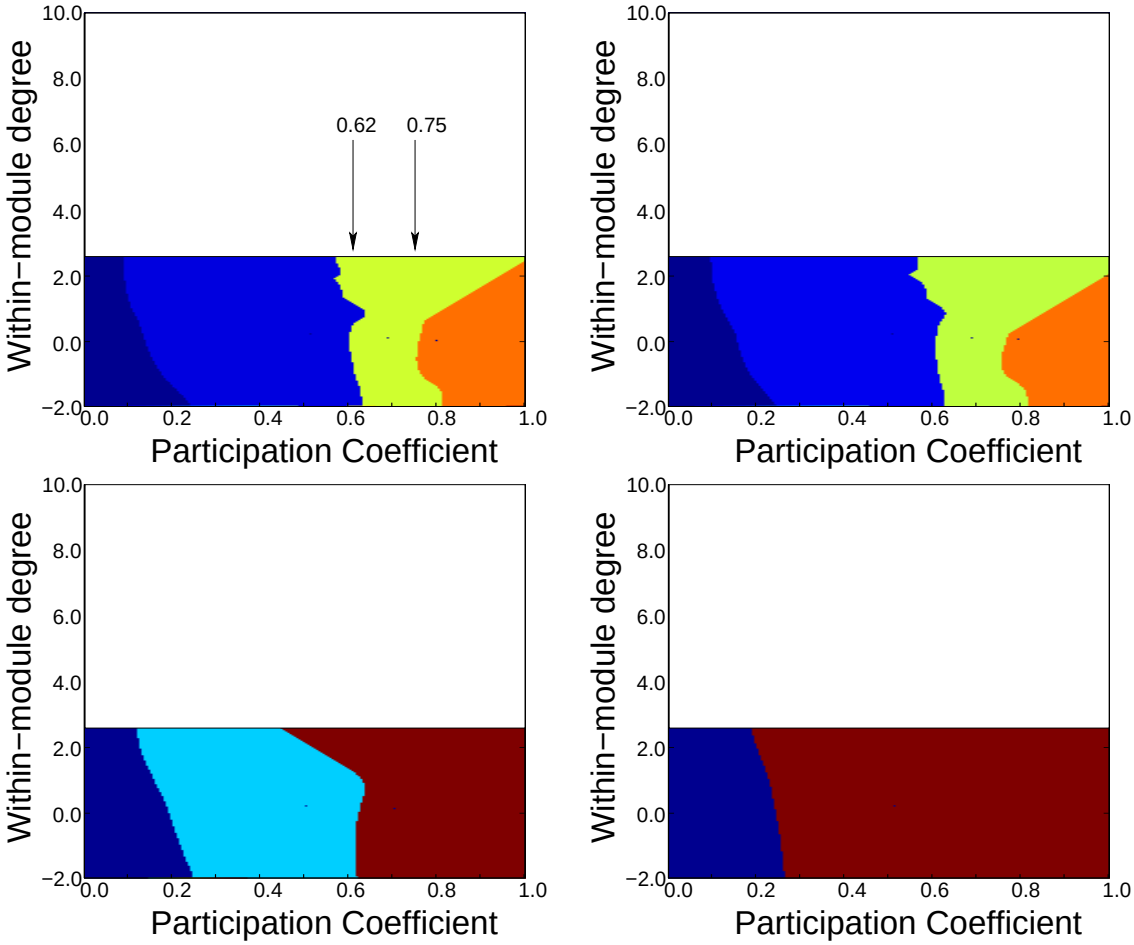

Figure 6: Basin of attraction identification for density landscapes obtained with **a,** $\sigma_P = 0.03$, **b,** $\sigma_P = 0.035$, **c,** $\sigma_P = 0.05$, and **d,** $\sigma_P = 0.08$. Note how the values of P identified in our simple analysis provide a good match to the boundaries of the basins of attraction in the node density landscapes.

2.3 "Basins of attraction" for hubs

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 (Fig. S7) .

In this case the results are not as clear as for the non-hub region because of the scarcity of data points. However, the density plot are compatible with a selection of three regions corresponding to distinct roles. The boundary of these regions are at $P = 0.30$ and 0.75.

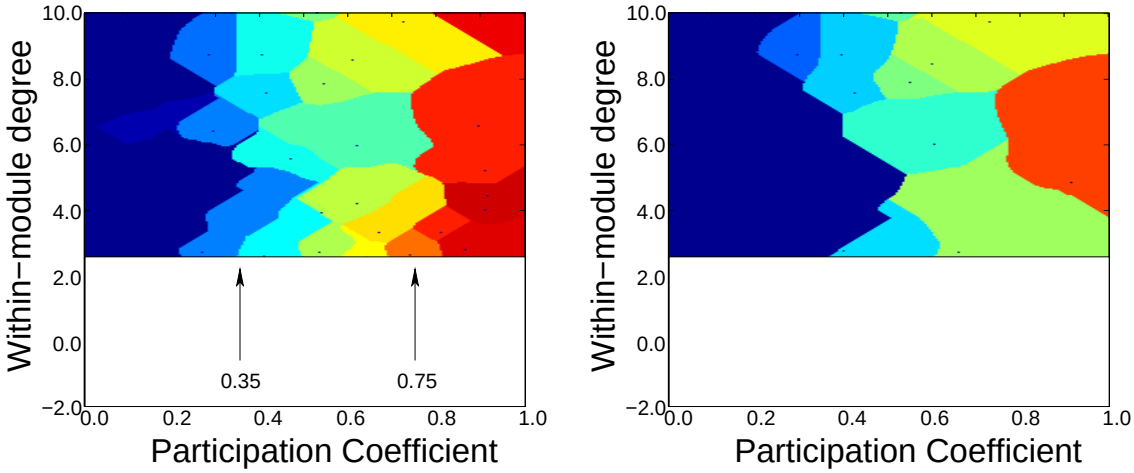

Figure 7: Basin of attraction identification for density landscapes obtained with **a,** $\sigma_P = 0.03$, and **b,** $\sigma_P = 0.05$.

The seven roles we identify and the corresponding regions in the zP parameters-space are displayed in Fig. S8.

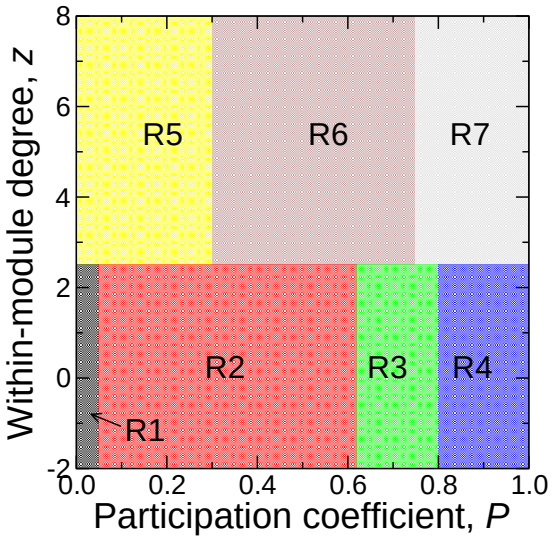

Figure 8: Role-specific regions in the zP parameters-space.

3 Metabolic networks

3.1 Modules in metabolic networks

Two issues related to the accuracy of the identification of modules in metabolic networks are worth analyzing in detail. First, nodes are expected to be more densely connected to nodes in the same module than to nodes in other modules. To quantify to which extent the algorithm accomplishes this task, we depict the within- and between-module connectivity density, that is, the ratio between the actual number of links and the maximum number of links within a module and between a pair of modules (Fig. S9a). As expected, the within-module density is much larger than the between-module density, typically 100-1000 times larger.

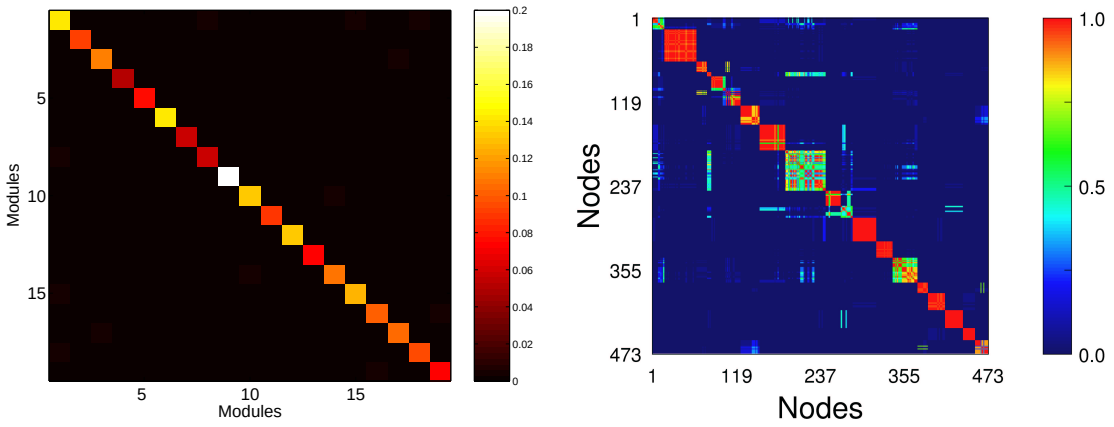

Figure 9: Accuracy of the module identification algorithm. **a,** Connection density between and within modules. The connection density is defined as the ratio between the actual number of links and the maximum possible number of links. **b,** To test the robustness of the algorithm when applied to metabolic networks, we obtain 25 partitions of the metabolic network of *E. coli* and plot, for each pair of nodes in the network, the fraction of times that they are classified in the same module.

Second, since our module-identification algorithm is stochastic, different runs yield, in general, different partitions of the nodes into modules. To test the robustness of the algorithm when applied to metabolic networks, we obtain 25 partitions of the metabolic network of *E. coli* 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 Fig. S9b, modules are robustly and consistently identified.

3.2 Role assignment for the metabolites in 12 organisms

We analyze the metabolic networks of 12 organisms: four archaea—*P. furiosus*, *A. pernix*, *A. fulgidus*, and *S. solfataricus*—, four prokaryotes—*E. coli*, *B. subtilis*, *L. lactis*, and *T. elongatus*—, and four eukaryotes—*S. cerevisiae*, *C. elegans*, *P. falciparum*, and *H. sapiens*.

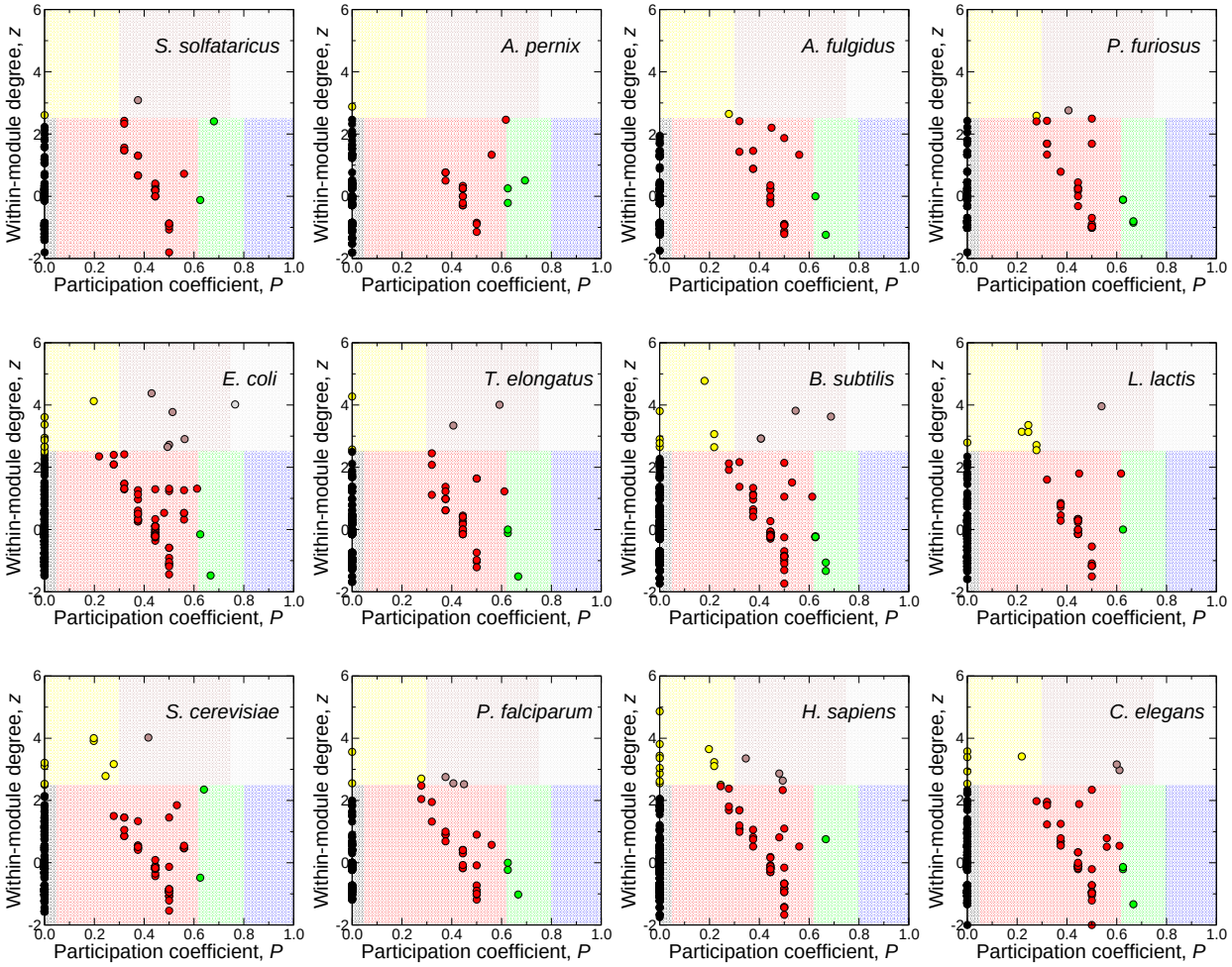

Figure 10: Metabolite role determination for the 12 organisms. Each metabolite is represented as a point in the zP phase-space, and is colored according to its role.

An issue of potential concern is the accuracy of the metabolic network database. The metabolic networks used in the paper were originally compiled by Ma and Zeng[7] (MZ). These authors carefully considered the reactions contained in the KEGG database and manually cor-

rected inconsistencies and errors, removed current metabolites from reactions, and excluded polymerization reactions as well as reactions with macromolecule participation. Although this database is probably the most reliable to date[8], it may be argued that it is very restrictive and that our results may be contingent on its use.

To assess this possibility, we analyze the complete, unfiltered, KEGG database. To build the metabolic networks, we use the data compiled in the LIGAND section of the KEGG database, publicly available at ftp://ftp.genome.ad.jp/pub/kegg/ligand/. In particular, we consider all bio-

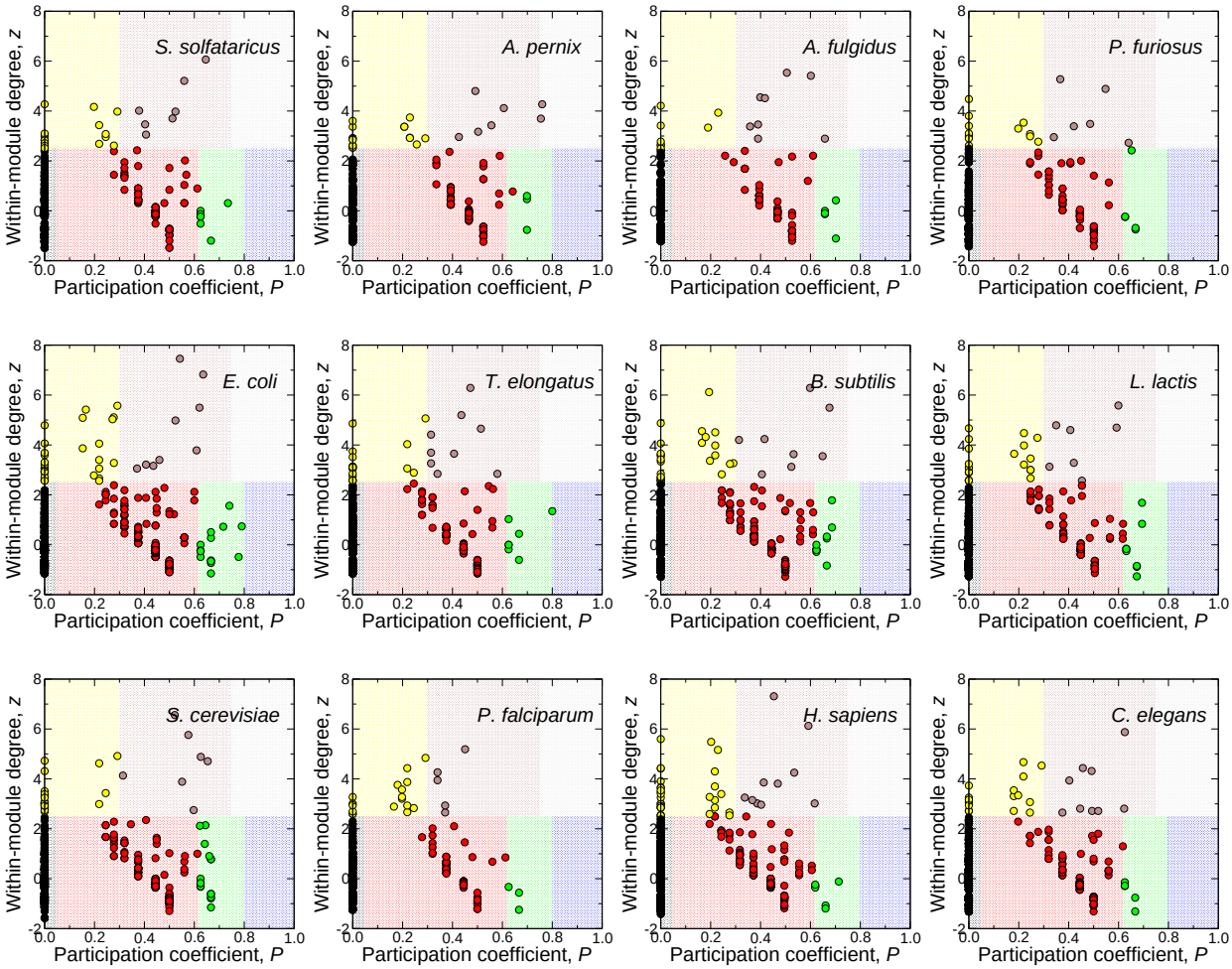

Figure 11: Metabolite role determination for the 12 organisms using unfiltered data from KEGG. Each metabolite is represented as a point in the zP phase-space, and is colored according to its role.

chemical reactions included in the *reaction_main.lst* file, which includes the main metabolites for each reaction. Then, for each organism, we only take into account reactions that are catalyzed by an enzyme that the organism is able to synthesize. We obtain the enzymes necessary for each reaction from the *reaction* file, and the enzymes synthesized by each organism from the organism database. For example, for *E. coli* the database is available at ftp://ftp.genome.ad.jp/pub/kegg/genomes/genes/E.coli.ent.

Since the KEGG metabolic networks are not subject to manual filtering, they contain many more metabolites and reactions than the networks in the MZ database. For *E. coli*, for example, the MZ network contains 473 nodes while the unfiltered network contains 1200 metabolites. This difference is due to two factors. First, some metabolites, such as macromolecules, are removed from the MZ database. Second, some reactions are also discarded, so the connectivity of the MZ networks is smaller and the giant component of the network contains a smaller fraction of nodes.

Remarkably, the distribution of nodes in the different regions of the zP space is similar in networks obtained from the MZ and the unfiltered databases (Figs. S10 and S11). In particular, roles R4 and R7 are unpopulated, and there are usually only a few connectors—both hub and non-hub. These results are in stark contrast to those observed in some other networks (Fig. S3). The North American airport network, for example, contains many more connector hubs than metabolic networks (Fig. S3c), but fewer non-hub connectors. On the other hand, module-less model networks contain many nodes in role R4—Erdös-Rényi and Barabási-Albert networks—and in role R7—Barabasi-Albert networks— but few ultra-peripheral nodes (Fig. S3b and d).

3.3 Node degree and metabolite conservation

We find that non-hub connector metabolites are more conserved than provincial hubs, which have larger within-module degree z-score. This is a remarkable result and its soundness and

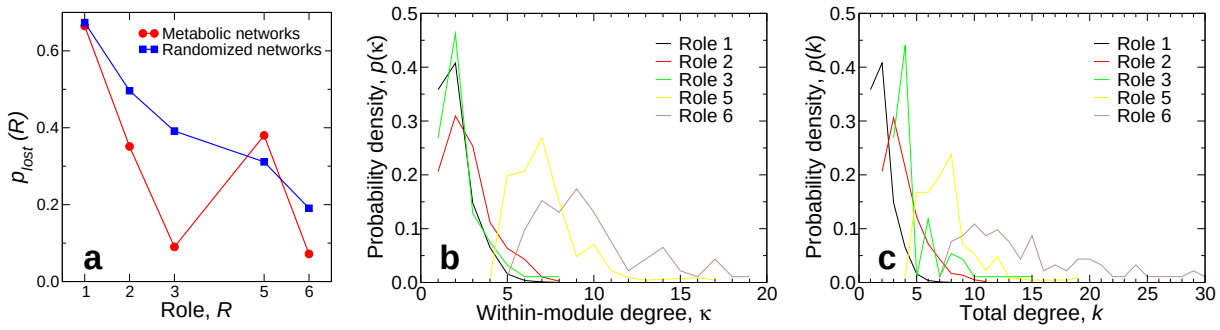

Figure 12: Node degree and role conservation results. **a,** Comparison of role conservation for real metabolic networks and for the corresponding randomized networks. **b,** Distribution $p(\kappa)$ of within-module degrees for nodes with different roles. **c,** Distribution $p(k)$ of total degrees for nodes with different roles.

robustness deserve to be considered carefully.

First, it is worth pointing out that studies of other complex biological networks have stressed the relevance of the degree of nodes. Jeong and coworkers showed that, in protein interaction networks, nodes with high degree are more essential than those with low degree[9]. In our analysis, then, one may expect that the total degree of a node is a key determinant of its inter-species conservation. Our claim is, however, that not only degree but also the position—role—of the node in the network is a crucial consideration regarding essentiallity.

To determine to which extent degree can account for the reported results, we randomize the metabolic networks by keeping constant the degree of each node. By doing this, all connectivity correlations are "removed" from the network, and the total degree of a node is the only relevant property kept. In Fig. S12a, we show that the conservation pattern is significantly altered for the randomized networks. In this case, $p_{\rm lost}(R3) > p_{\rm lost}(R5)$ because nodes in R3 have lower degree than those in R5. It is also illustrative to compare the conservation of each role in the real and randomized networks. Roles with some connector function—R2, R3, and R6—are more conserved in the real case than in the random case, while provincial hubs—R5—are less conserved. Ultra-peripheral nodes are similarly conserved in the random and real networks.

The second point that needs to be checked is the following. It may be that, due to the use

of the z-score instead of the raw degree, some low-degree nodes in very sparse modules were classified as provincial hubs. Then, one may expect the loss rate to be large for provincial hubs just as an artifact, because of the influence of these nodes.

We have tested this possibility and found that this is not the case. In Fig. S12b, we show that the distribution $p(\kappa)$ of within-module degrees is essentially identical for roles 1-3, with approximately 95% of the nodes having $\kappa \in [1, 4]$. In contrast, less than 1% of the nodes with role 5 have $\kappa = 4$ and none have $\kappa < 4$. Similarly, if one considers the total degree k, 71% of the nodes in role 3 have $k < 5$, while less than 1% of the nodes in role 5 have such small degrees (Fig. S12c). The conclusion is therefore sound: connectors are more conserved than provincial hubs in spite of having smaller within-module degree and smaller total degree. We have added this discussion to the Supplementary Material.

3.4 Metabolite conservation results and role definition

In the first section of this Supplementary Information, we have analyzed the considerations that lead to the definition of seven system-independent roles. One could, however, divide the zP phase-space in other manners. Next, we show that the metabolite conservation results are independent of the particular definition of roles one uses, as long as the definition of roles, specially for non-hubs, takes into account certain fundamental considerations. This also serves as further evidence that our definition of roles is a parsimonious definition.

Let us start by considering an alternative partition of the zP space, in which nodes are divided, first, into three classes according to their within-module degree: anti-hubs ($z < 0$), non-hubs ($0 \leq z < 2.5$), and hubs ($z \leq 2.5$). Then, each one of these three classes is further subdivided into five groups according to the participation coefficient: $P < 0.2$, $0.2 \leq P < 0.4$, $0.4 \leq P < 0.6$, $0.6 \leq P < 0.8$, and $P \geq 0.8$. This is a finer definition of roles than the one we propose because it divides the zP phase-space into 15 regions instead of 7. In Fig. S13 we

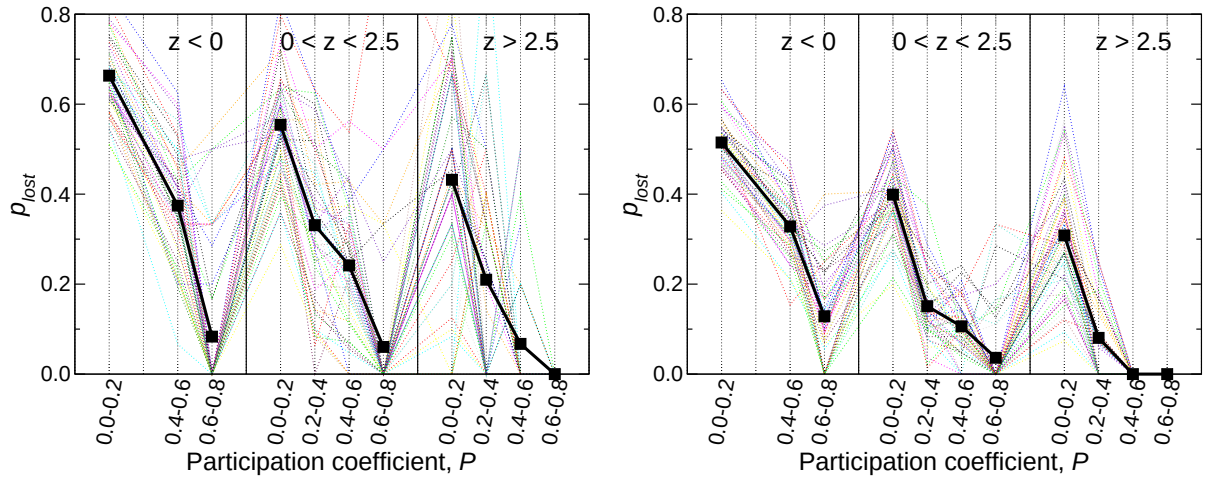

Figure 13: Robustness of the role conservation results. To test the robustness of our results for role-dependent conservation of metabolites, we investigate alternative definitions of the roles. Specifically, we partition the nodes into three grouping according to their degree: "anti-hubs" ($z < 0$), non-hubs ($0 < z < 2.5$) and hubs ($z > 2.5$). We then further subdivide the nodes according to the value of P into five equally-wide regions. Inter-species metabolite conservation as a function of the values of z and P for the **a,** MZ database and the **b,** unfiltered KEGG database. We find that: (i) the results are pretty insensitive to the database used, (ii) there is a very strong dependence of p_{lost} on the participation coefficient, and a somewhat weaker dependence on z, i.e., the degree, (iii) the seven roles described in the manuscript fully capture the results obtained with a finer definition of the roles.

show that such a definition of the roles does not alter the main conclusions drawn in the paper. Namely, connector nodes with low degree and high participation coefficient are systematically and consistently more conserved than hubs with low participation coefficient. Even anti-hubs with high P are considerably more conserved than hubs with low P. Thus, the results obtained from this finer partition of the zP phase-space emphasize, even more clearly, to which extent the participation coefficient is more relevant than the degree.

Now, a relevant methodological question is: would *any* partition of the zP phase-space yield *correct* results? The answer is that, specially for non-hubs, one must be careful. Non-hubs are nodes with low degree, and therefore integer constraints on the possible values of P become important. For example, as discussed in the first section of this Supporting Information and in Fig. S2, a node with degree $k = 2$ can only have $P = 1$ or $P = 0.5$.

To see how this fact potentially affects the definition of roles, consider a situation in which

one divides the the zP space into hubs ($z \leq 2.5$) and non-hubs ($z < 2.5$), and each of these two classes into five groups according to P as before: $P < 0.2$, $0.2 \leq P < 0.4$, $0.4 \leq P < 0.6$, $0.6 \leq P < 0.8$, and $P \geq 0.8$. Then, the region that includes $P = 0.5$ will contain many nodes with $k = 2$, while the region that includes $P = 0.3$ will include none. Since degree also plays a role in metabolite conservation, it may happen that nodes in the region containing $P = 0.5$ are less conserved than those in the region that contains $P = 0.3$, but this would be entirely due to a poor sampling of the degrees within each of the regions, and not to P itself.

In our definition of roles described in the first section, this problem is addressed by assigning a larger range of P to peripheral nodes (R2), in such a way that degrees are more uniformly sampled. Other partitions of the zP space into 7 or a similar number of roles must take this into consideration and, consequently, should be similar to our "universal roles."

References and Notes

1. Nadel, S. F. *The Theory of Social Structure* (Cohen and West, London, UK, 1957).

2. Scott, J. *Social Network Analysis: A Handbook* (SAGE Publications Ltd., London, UK, 2000), 2 edn.

3. Lorrain, F. & White, H. C. Structural equivalence of individuals in social networks. *J. Math. Sociol.* **1**, 49–80 (1971).

4. White, H. C., Boorman, S. A. & Breiger, R. L. Social structure from multiple networks. I. Blockmodels of roles and positions. *Am. J. Sociol.* **81**, 730–780 (1976).

5. Wasserman, S. & Faust, K. *Social Network Analysis* (Cambridge University Press, Cambridge, U.K., 1994).

6. Sailer, L. D. Structural equivalence: Meaning and definition, computation and applications. *Soc. Networks* **1**, 73–90 (1978).

7. Ma, H. & Zeng, A.-P. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* **19**, 270–277 (2003).

8. Hatzimanikatis, V., Li, C., Ionita, J. A. & Broadbelt, L. Metabolic networks: enzyme function and metabolite structure. *Curr. Opin. Struc. Biol.* **14**, 300–306 (2004).

9. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* 41–42 (2001).