

## SUPPLEMENTARY DISCUSSION

### **Classes of complex networks defined by role-to-role connectivity profiles**

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

#### **Description of the datasets**

We consider four types of networks: (i) metabolic networks<sup>1,2,3,4</sup>, (ii) protein interactomes<sup>5,6,7,8</sup>, (iii) air transportation networks<sup>9,10,11</sup>, and (iv) the Internet at the autonomous system (AS) level<sup>12,13</sup>.

To build the metabolic networks, we use the data compiled in the LIGAND section of the KEGG database<sup>14,15</sup>. In particular, we consider all biochemical 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, and those reactions that are explicitly labeled as spontaneous. We obtain the enzymes necessary for each reaction from the *reaction* file, and the enzymes synthesized by each organism from the organism databases in KEGG. We consider 18 organisms: six archaea (*A. fulgidus*, *A. pernix*, *M. jannaschii*, *P. aerophilum*, *P. furiosus*, and *S. solfataricus*), six bacteria (*B. subtilis*, *E. coli*, *F. nucleatum*, *H. pylori*, *M. leprae*, and *T. elongatus*), and six eukaryotes (*A. thaliana*, *C. elegans*, *H. sapiens*, *P. falciparum*, *S. cerevisiae*, and *S. pombe*). To obtain the role-to-role connectivity profiles, we average the results for the six networks in each group.

We analyze the protein interactome of two species: *S. cerevisiae*<sup>6</sup> and *C. elegans*<sup>7</sup>. The interaction data is obtained directly from the cited references.

We analyze the world-wide air transportation network<sup>9</sup>, as well as two regions<sup>9</sup>: North

America and Asia and Middle East. The data is obtained as described in the cited reference.

Finally, we analyze the Internet at the autonomous system level<sup>12,13</sup>. We consider three datasets, corresponding to January 2, 1998, January 15, 1999, and January 2, 2000. We obtained the data from the NLANR project web page (<http://moat.nlanr.net/AS/Data/>).

## Modularity and link type distribution in model networks

We find that typical model networks, such as Erdős-Rényi random graphs<sup>16</sup> or preferential attachment (PA) scale-free networks<sup>17</sup>, do not possess a significant modular structure. This fact is not surprising, considering that these models include no mechanism that could give rise to a modular structure. We also find that the profile of link types for model PA scale-free networks is orthogonal to networks in both the chain periphery class and the multi-star class (Fig. S1).

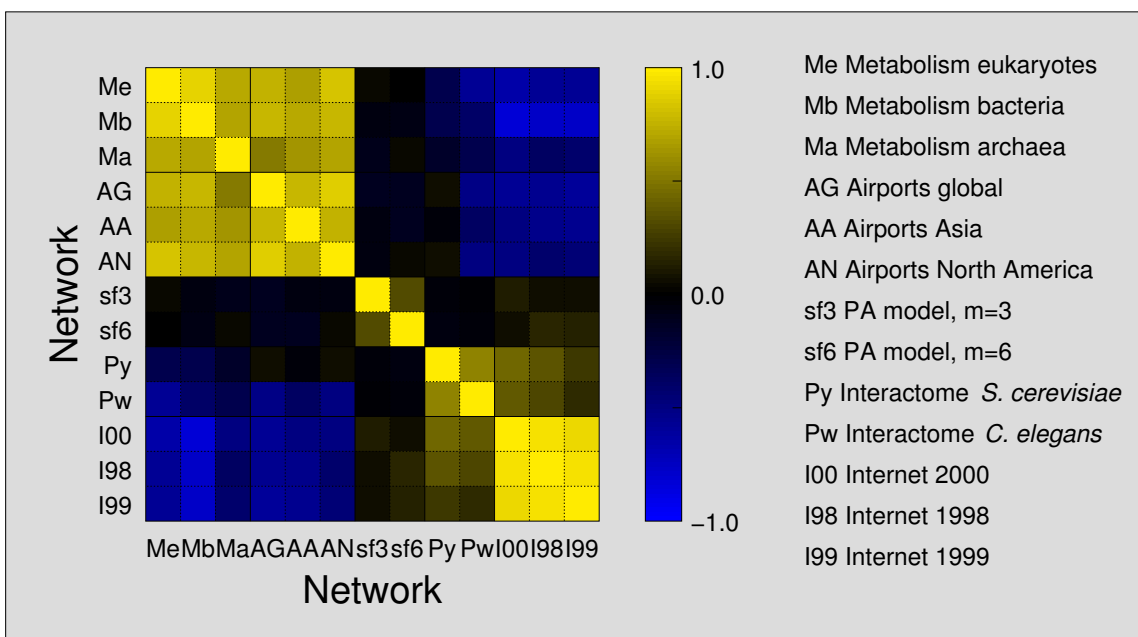


Figure S1: We quantify the similarity between two  $z$ -score profiles by means of the correlation coefficient (Methods), with yellow corresponding to high correlation, blue to high anti-correlation, and black to no correlation. The profile of link types for model scale-free networks is orthogonal to networks in both the chain periphery class and the multi-star class.

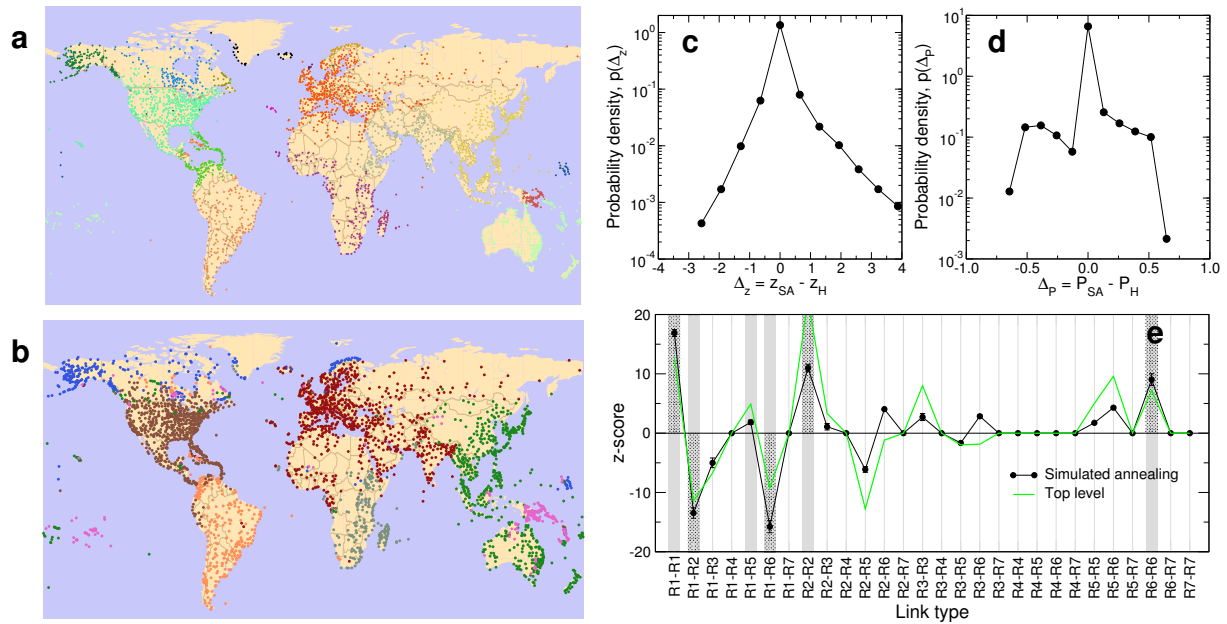


Figure S2: Robustness of the results to alternative partitions of the network for the global air transportation network. **a**, Partition of the network obtained using simulated annealing<sup>9</sup>. **b**, Partition of the network corresponding to the top level of the hierarchical modular structure of the network<sup>18</sup>. **c**, Distribution of the difference between the within module relative degree  $z$  obtained from the partitions in **a** and **b**. **d**, Same as **c** for the participation coefficient  $P$ . **e**, Role-to-role connectivity profiles obtained from the partitions in **a** and **b**. The fact that the distributions in **c** and **d** are heavily peaked around  $\Delta = 0$  (note the logarithmic scale) and the fact that the two profiles in **e** are very similar to each other indicate that our results are robust even when using partitions of the network at different levels of coarse-graining.

## Robustness of the results

### Alternative partitions of the networks

An issue of potential concern is how different divisions of the network into modules affect our findings. Specifically, some networks have more than one sensible partition corresponding to different levels of modular organization<sup>18</sup>. The question we address here is whether these different sensible partitions yield the same qualitative results or not. We focus on the global air transportation network, which is a network that clearly has different levels of modular organization<sup>18</sup>.

The best partition of the global air transportation network obtained using simulated annealing consists of 25 modules, while the top level of modular organization consists of only 7 modules<sup>18</sup>. Despite this considerable difference, the participation coefficient  $P$  and the within module relative degree  $z$  change very little for most nodes, and so does the role-to-role connectivity profile (Fig. S2).

## **Alternative definition of the roles**

We have previously given ample justification for the definition of roles we use in the paper<sup>19</sup>. Still, it is important to verify that alternative definitions would leave unchanged our main findings in this work, namely, that networks with the same function have more similar role-to-role connectivity profiles, and that networks can be classified into two main classes.

To investigate this issue, we use an alternative definition of the roles, in which the  $zP$ -space of within-module relative degrees and participation coefficients is split in 9 regions. The  $z$  axis is divided at  $z = 2.5$  and  $z = 0.0$ , and the  $P$  axis at  $P = 0.33$  and  $P = 0.67$ . Note that, although one could in principle define more roles (bins), such a division would likely result in lots of quasi-empty roles and, therefore, in poor statistics.

As we show in Fig. S3, this alternative definition of the roles does not alter, in any significant way, the results reported in the main article.

## **Errors in the databases**

Finally, we check whether errors in the databases from which the networks are obtained may significantly affect our results. We investigate this issue using the 1998 instance of the Internet, which is one of the biggest networks in our study.

We find that, when 5% of the links of the network are randomly rewired to simulate errors in the database, the role-to-role connectivity profile is still strongly correlated to the original

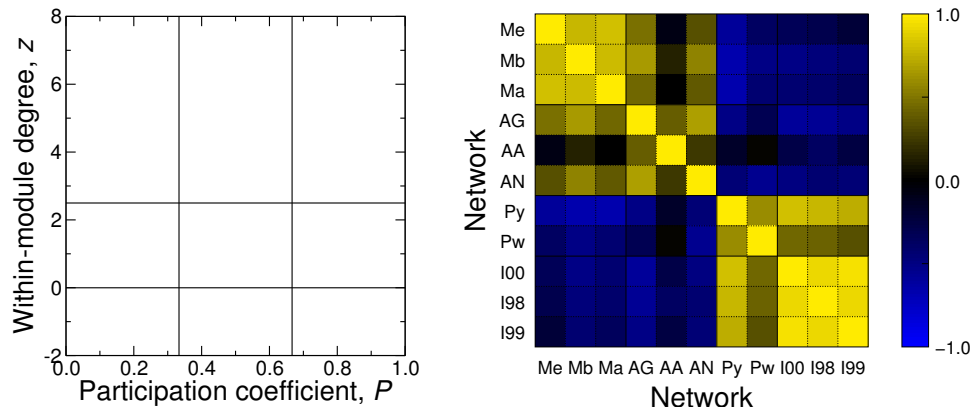


Figure S3: Robustness of the results to alternative definitions of the roles. The left panel illustrates the boundaries for the alternative definition of roles (no nodes are represented). The right panel is equivalent to Fig. 3c of the main text, but obtained with the alternative definition of the roles.

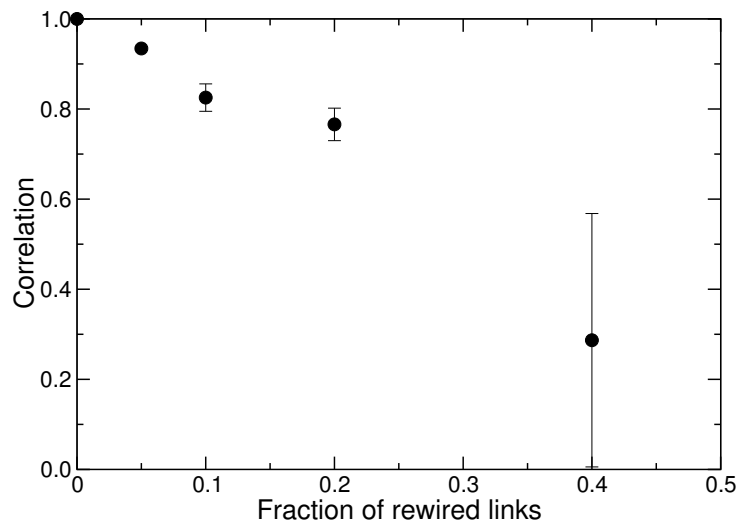


Figure S4: Robustness of the results to errors in the 1998 Internet network. We simulate errors in the databases by randomly rewiring a fraction of the links in the network. Even when as many 20% of the links of the network are rewired, the correlation turns out to be quite strong. Only when we rewire 40% of the links we observe a severe drop in the correlation with the original profile.

one (correlation  $c = 0.93$ , Fig. S4). Even when as many 20% of the links of the network are rewired, the correlation turns out to be quite strong. Only when we rewire 40% of the links we observe a severe drop in the correlation with the original profile.

## Constraints in the randomization process

In the two random network ensembles we consider, there are topological constraints that need to be satisfied. In the  $\mathcal{D}$  ensemble, the degree of each node has to be the same as in the original network. In the  $\mathcal{M}$  ensemble, in addition to this constraint, the number of connections each node has to nodes in each module has to be preserved, too.

Because of these constraints (and due to other effects such as increasing link density), a fraction of links in the randomized networks coincide with links that actually exist in the real network under consideration. While this is to be expected, a potential source of concern is whether the original network is being “sufficiently randomized” or, in other words, if the fraction of links that is different in the randomized networks is “large enough.” This consideration becomes particularly important in light of the fact that most degree-degree correlations of real networks (as measured by the degree of the nodes’ neighbors) are explained by the  $\mathcal{M}$  ensemble (Fig. 1 of the main text).

As we show in Table S1, for the networks we study an average of 96.1% of the links in the  $\mathcal{D}$  ensemble and of 74.9% in the  $\mathcal{M}$  ensemble do not exist in the original network.

Combined with the results in Fig. 1 of the main text, this means that, at most, 25.1% of the links in the networks we study are responsible for all the degree-degree correlations measured by the average neighbors’ degree. Assuming that this figure is representative of other real-world complex networks, our finding highlights the need for module-based measures. Indeed, by considering degree-degree correlations, research on complex networks has so far focused on around one quarter of the links in networks and implicitly (and unknowingly) ignored the remaining three quarters. Far from being irrelevant, these remaining links are responsible for the differences between the two network classes reported in the main text.

Network type	Network	$f_{\mathcal{D}}[\%]$	$f_{\mathcal{M}}[\%]$
Metabolism Archaea	<i>A. fulgidus</i>	98.5	74.3
	<i>A. pernix</i>	97.5	74.0
	<i>M. jannaschii</i>	97.2	67.6
	<i>P. aerophilum</i>	97.9	76.3
	<i>P. furiosus</i>	97.7	72.4
	<i>S. solfataricus</i>	97.8	73.6
Metabolism Bacteria	<i>B. subtilis</i>	98.7	80.6
	<i>E. coli</i>	98.9	83.9
	<i>F. nucleatum</i>	97.9	76.8
	<i>H. pylori</i>	98.3	78.4
	<i>M. leprae</i>	98.2	78.9
	<i>T. elongatus</i>	98.6	78.8
Metabolism Eukaryotes	<i>A. thaliana</i>	98.8	80.8
	<i>C. elegans</i>	98.2	77.7
	<i>H. sapiens</i>	99.2	82.3
	<i>P. falciparum</i>	98.7	76.5
	<i>S. cerevisiae</i>	98.8	80.8
	<i>S. pombe</i>	98.4	79.2
Air transportation	Global	92.1	71.5
	Asia & Middle East	86.4	58.2
	North America	78.0	54.2
Interactome	<i>S. cerevisiae</i>	99.3	78.0
	<i>C. elegans</i>	98.2	80.6
Internet	2000	92.1	72.8
	1998	90.8	69.1
	1999	91.3	69.0

Table S1: For each network, we show the percentage  $f$  of links that are randomized in ensembles  $\mathcal{D}$  and  $\mathcal{M}$ .

### Nodes in Figure 3

In Figs. S5 and S6 we display the name of the hubs and of all the nodes, respectively, for the network modules plotted in Fig. 3 of the main text.

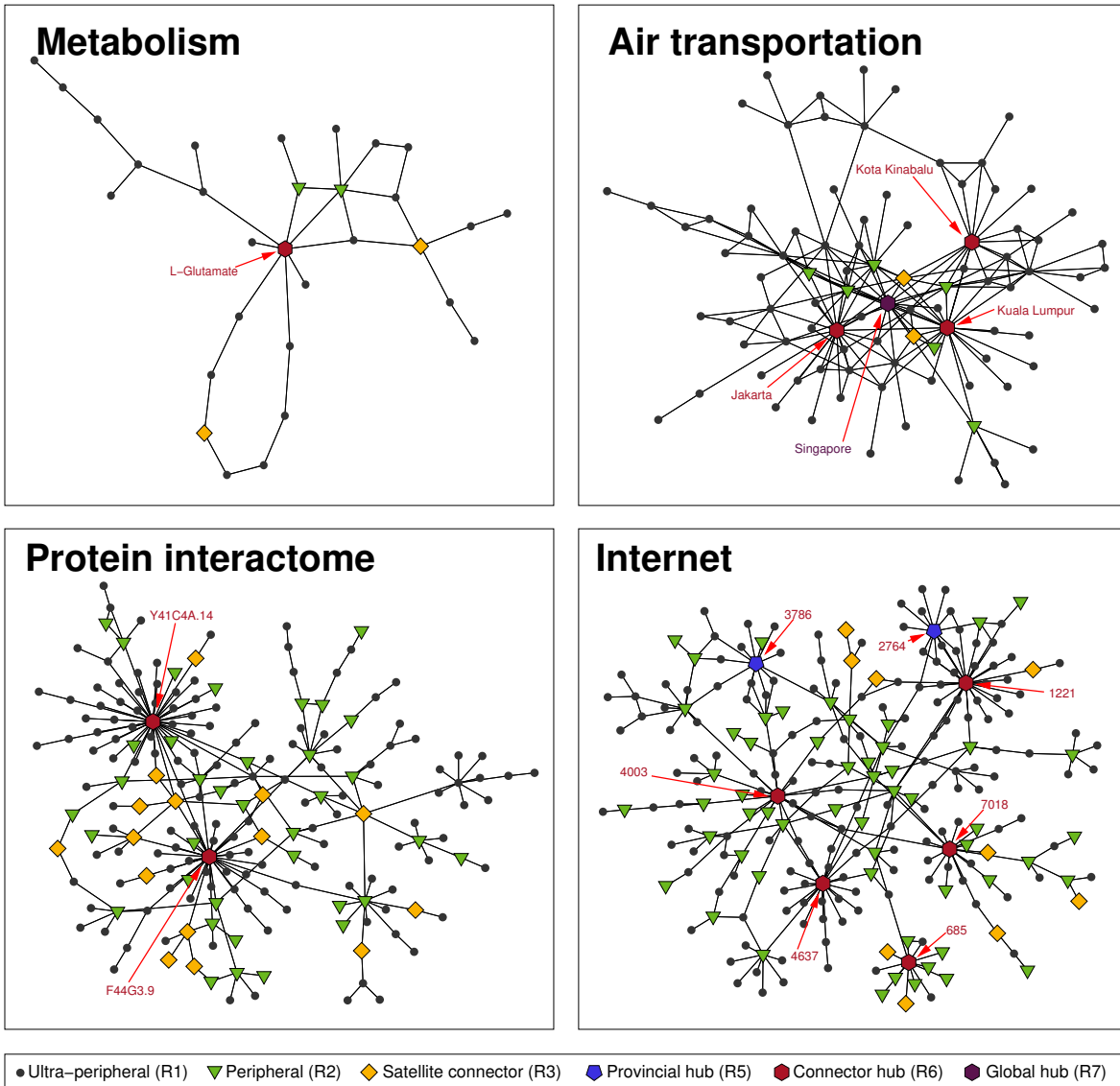


Figure S5: Same as in Fig. 3 of the main text, but with hubs labeled.



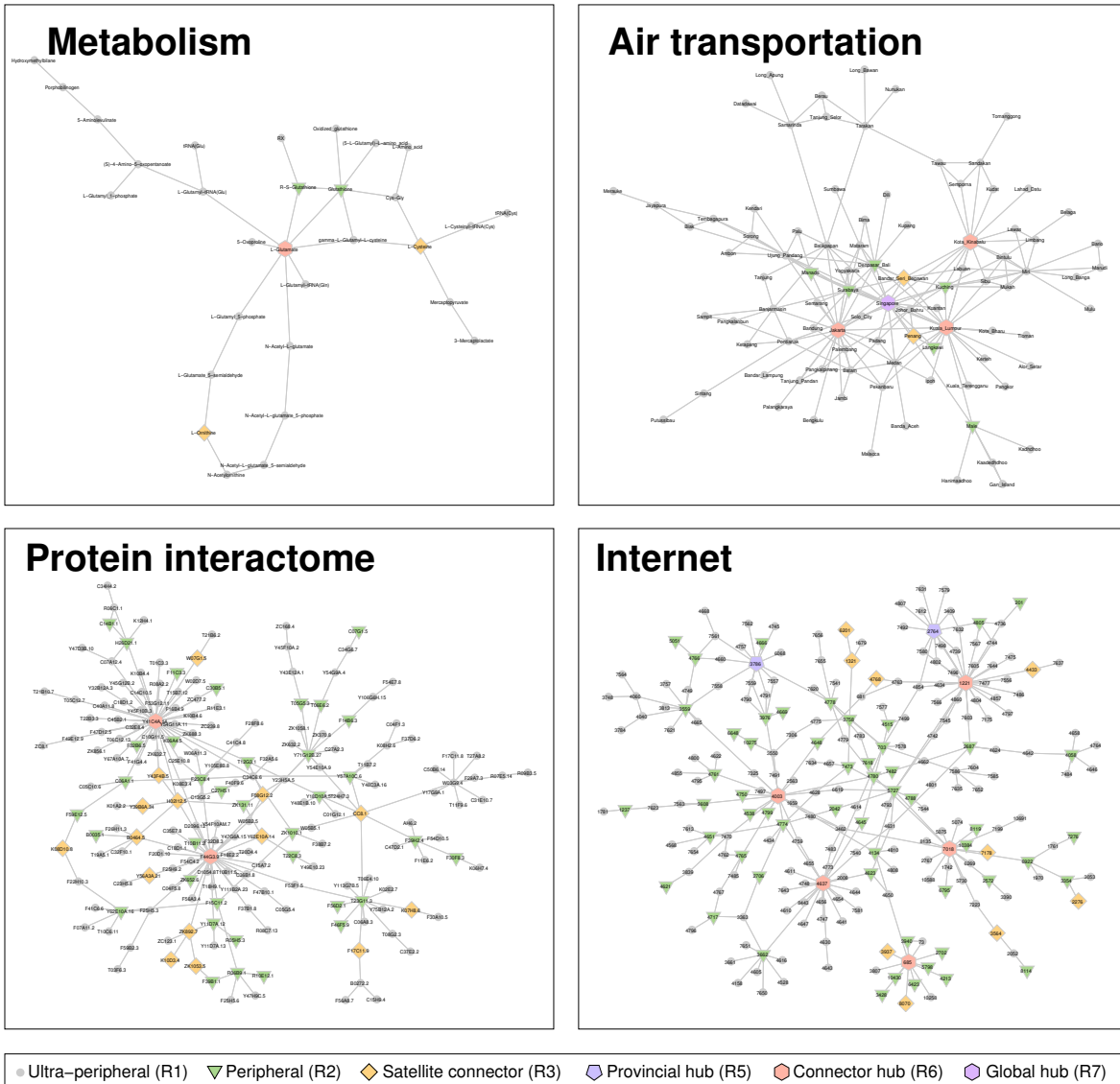


Figure S6: Same as in Fig. 3 of the main text, but with all nodes labeled (PDF file enables zooming in to read the labels).

## References

1. Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N. & Barabási, A.-L. The large-scale organization of metabolic networks. *Nature* **407**, 651–654 (2000).
2. Wagner, A. & Fell, D. A. The small world inside large metabolic networks. *Proc. Roy. Soc. B* **268**, 1803–1810 (2001).
3. 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).
4. Guimerà, R. & Amaral, L. A. N. Functional cartography of complex metabolic networks. *Nature* **433**, 895–900 (2005).
5. Uetz, P. *et al.* A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403**, 623–627 (2000).
6. Jeong, H., Mason, S. P., Barabási, A.-L. & Oltvai, Z. N. Lethality and centrality in protein networks. *Nature* **411**, 41–42 (2001).
7. Li, S. *et al.* A map of the interactome network of the metazoan *C. elegans*. *Science* **303**, 540–543 (2004).
8. Han, J.-D. J. *et al.* Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature* **430**, 88–93 (2004).
9. Guimerà, R., Mossa, S., Turtschi, A. & Amaral, L. A. N. The worldwide air transportation network: Anomalous centrality, community structure, and cities' global roles. *Proc. Natl. Acad. Sci. USA* **102**, 7794–7799 (2005).

10. Barrat, A., Barthélemy, M., Pastor-Satorras, R. & Vespignani, A. The architecture of complex weighted networks. *Proc. Natl. Acad. Sci. USA* **101**, 3747–3752 (2004).
11. Li, W. & Cai, X. Statistical analysis of airport network of China. *Phys. Rev. E* **69**, art. no. 046106 (2004).
12. Pastor-Satorras, R., Vázquez, A. & Vespignani, A. Dynamical and correlation properties of the Internet. *Phys. Rev. Lett.* **87**, art. no. 258701 (2001).
13. Vázquez, A., Pastor-Satorras, R. & Vespignani, A. Large-scale topological and dynamical properties of the Internet. *Phys. Rev. E* **65**, art. no. 066130 (2002).
14. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**, 27–30 (2000).
15. Goto, S., Nishioka, T. & Kanehisa, M. LIGAND: chemical database for enzyme reactions. *Bioinformatics* **14**, 591–599 (1998).
16. Bollobás, B. *Random graphs* (Cambridge University Press, Cambridge, UK, 2001), 2nd edn.
17. Barabási, A.-L. & Albert, R. Emergence of scaling in random networks. *Science* **286**, 509–512 (1999).
18. Sales-Pardo, M., Guimerà, R., Moreira, A. A. & Amaral, L. A. N. Extracting and representing the hierarchical organization of complex systems. *submitted* (2006).
19. Guimerà, R. & Amaral, L. A. N. Cartography of complex networks: modules and universal roles. *J. Stat. Mech.: Theor. Exp.* P02001 (2005).