# Global constraints on local interaction patterns in cellular networks

A. Vázquez[1], R. Dobrin[2], D. Sergi[3], J.-P. Eckmann[3,4], Z. N. Oltvai[2] and A.-L. Barabási[1]

[1]Department of Physics, University of Notre Dame, Notre Dame, IN 46556, USA
[2]Department of Pathology, Northwestern University, Chicago, IL 60611, USA
[3]Département de Physique Théorique and [4]Section de Mathématiques, Université de Geneve, Geneve, CH-1211, Switzerland

**Recent evidence indicates that the abundance of certain local interaction patterns, or subgraphs, in complex biological networks may carry significant information about their function and overall organization. Yet, the underlying reason for the variable number of different types of subgraphs remains poorly understood. Here we show that the large-scale architecture of five well-characterized cellular networks impose significant constraints on their local structure, and the knowledge of only two global statistical parameters is sufficient to correctly predict the density of most of their subgraphs. We also find that highly abundant subgraphs do not exist in isolation, but undergo a percolation transition and aggregate into subgraph clusters. These results are likely to have important implications on the evolutionary origin and function of subgraphs in all cellular networks.**

Overrepresented subgraphs (motifs) are characteristic features of all networks (*1, 2*), representing recurring elementary interaction patterns that potentially carry out distinct tasks, from information processing to auto-regulation (*1-7*). The high degree of evolutionary conservation of motif constituents within the protein-protein interaction network (*8,9*), and the convergent evolution towards the same motif types in the transcriptional regulatory network of diverse species (*10, 11*) suggest that motifs are of direct functional relevance, and that a systematic characterization of complex biological networks can be achieved by studying their local, subgraph-based structure (*4*). Yet, most

motifs in biological networks are inherently non-separable, aggregating into larger topological superstructures that melt into the network's global architecture (*12*), characterized by a highly inhomogeneous scale-free connectivity and a hierarchically modular topology (*12-18*)

A coherent understanding of a network's topological and functional organization requires the development of a single framework that can explain the appearance of subgraphs and motifs, the mechanisms responsible for their aggregation into larger superstructures, and their relationship with the universal large-scale features of complex networks. Here we present such a unifying framework by focusing on five well-characterized cellular networks of a prokaryotic and an eukaryotic model organism, the metabolic and transcriptional-regulatory networks of *Saccharomyces cerevisiae* and *Escherichia coli*, respectively, and the protein-protein interaction (PPI) network of *S. cerevisiae*. We show that the subgraph density in these five cellular networks can be fully predicted based on knowledge of the two parameters characterizing their scale-free and hierarchical topology, indicating that the network's large scale characteristics uniquely determine its local, subgraph-based topology.

Table 1a lists the density of several *n*-node subgraphs of the five studied cellular networks, indicating that the density of specific subgraphs in the corresponding *E. coli* and *S. cerevisiae* networks are comparable, underscoring the absence of significant subgraph-based differences between the two organisms. There are notable differences, however, among the different types of molecular interaction webs even within the same organism: the metabolic and the PPI networks display a much higher subgraph density than transcriptional regulatory networks. The observed paucity of certain subgraphs types and the abundance of others suggest two possible scenarios for their origin: their number may be largely determined by local functional constraints, such as the desirable signal filtering properties of feedforward motifs (*3*), or, alternatively, may simply represent a natural consequence of the networks' large-scale topology. To distinguish between these alternatives we start by focusing on the two most relevant topological parameters of a cellular network's large-scale structure: the degree exponent, $\gamma$ (*13*), and the hierarchical exponent, $\beta$ (*13*). The degree exponent ($\gamma$) characterizes the overall inhomogeneity in the connectivity of cellular networks: while most molecules are engaged in only a few

interactions, a few hubs are linked to a significantly higher number of other molecules (nodes). These wide degree variations are captured by the degree distribution, which for the studied cellular networks follows a power law, $P(k) \sim k^{-\gamma}$ (*1, 16, 19-21*). In contrast, the hierarchical exponent ($\beta$) characterizes the networks' innate modularity, indicating that many small, highly interconnected groups of nodes form larger, but less cohesive topological modules (*12, 13, 22*). This hierarchical modularity is captured by the scaling law $C(k) \sim C_0 k^{-\beta}$ (*23-26*), where $C(k)=2T(k)/k(k-1)$ is the clustering coefficient of a node with $k$ links, denoting the probability that a node's neighbors are linked to each other (*27*), and $T(k)$ is the number of direct links between the node's $k$ neighbors. Empirical studies indicate that each cellular network is characterized by a unique pair of ($\gamma$,$\beta$) parameters, listed in Table Ib, which were determined from the scaling of $P(k)$ and $C(k)$ functions describing the undirected version of these networks (*12, 13, 22*).

To examine the role of these two parameters in determining the observed subgraph density, we calculated analytically the number $N_{nm}$ of subgraphs with $n$ nodes and $m$ interactions expected for a network of $N$ nodes, in which the nodes -apart from fixed ($\gamma$,$\beta$) parameters-, are randomly connected to each other. Our calculations predict the existence of two subgraph classes. Type I subgraphs are those that satisfy $\gamma(m-n+1)-(n-\beta)<0$, their number being given by $N^{I}_{nm} \sim N k_{\max}^{-[\gamma(m-n+1)-(n-\beta)]}$, where $k_{\max}$ denotes the degree of the most connected node in the network. Type II subgraphs are those that satisfy $\gamma(m-n+1)-(n-\beta)>0$, and their number is given by $N^{II}_{nm} \sim N$. As even for finite networks $k_{\max} >> 1$, the typical number of Type I subgraphs is significantly larger than the number of Type II subgraphs ($N^{I}_{nm}/N^{II}_{nm} >> 1$). Moreover, for infinite systems ($N \to \infty$) the relative number of Type II subgraphs is vanishingly small compared to Type I subgraphs, as $N^{I}_{nm}/N^{II}_{nm} \to \infty$. Table I fully supports these predictions, indicating that the density of the subgraphs with a minimal number of connections (extreme Type I) (4,3), (5,4), (6,5), (7,6) is in the range $10\text{-}10^5$ ($N^{I}_{nm} >> 1$). In contrast, the density of the subgraphs with a maximal number of connections (extreme Type II) (4,6), (5,10), (6,15), (7,21) is either zero or close to zero, and always negligible compared to their Type I counterparts.

The ($n$,$m$) phase diagrams of Fig. 1, in which each discrete point corresponds to a different subgraph, offer a summary of our main predictions. The $\gamma(m-n+1)-(n-\beta)=0$ condition, separating the Type I and II subgraphs, appears as the stepped yellow phase

boundary in the phase diagram. For example, for the *E. coli* transcriptional regulatory network the phase boundary corresponds to a stepped-line with approximate overall slope $1+1/\gamma_{out}=2.0$ and intercept $-1-\gamma_{out}\log\langle k_{out}\rangle=-3.1$. The Type II subgraphs are those above this boundary, and should be either absent, or present only in very low numbers in the transcriptional regulatory network. In contrast, the Type I subgraphs below the boundary are predicted to be abundant.

To visually highlight the validity of these predictions we color-coded Fig. 1 according to the normalized count of each subgraph in each cellular network. We find a good agreement between the predictions and the actual subgraph count: the normalized count of the Type I subgraphs below the phase boundary is in the $10^{-2}$-1 range, in contrast with the Type II subgraphs above the predicted boundary, whose normalized count is either zero, or in the $10^{-9}$-$10^{-3}$ range. Comparing Figs. 1a-e indicates that while the stepped phase boundaries for the different cellular networks are predicted to differ due to the differences in the ($\gamma_{out}$,$\langle k_{out}\rangle$) exponents (Table 1b), the observed densities in the real networks follow relatively closely the predicted phase boundaries. Occasional local deviations from the predictions can be attributed to the high error bars for the ($\gamma_{out}$,$\langle k_{out}\rangle$) exponents (Table 1b), which allow for some local uncertainties for the phase boundary. Figures 1a-e also indicate that, in agreement with the empirical findings (*1, 2, 4*), each cellular network is characterized by a distinct set of over-represented Type I subgraphs, raising the possibility of classifying networks based on their local structure (*4*). Yet, the phase diagrams provide direct proof that the global (scale-free, hierarchical) topology determines the local (subgraph) structure of cellular networks, indicating that a subgraph or motif based classification must be rooted in the different ($\gamma_{out}$,$\langle k_{out}\rangle$) exponents characterizing these networks.

The concept of *motifs* was recently introduced to denote the subgraphs whose number exceeds with a preset threshold their expected count in a randomized network (*1-4*). However, our results indicate that the overrepresented Type I subgraphs is a fundamental topological feature of cellular networks, and thus we do not need to invoke a comparison to a randomized graph, nor introduce a threshold parameter to identify them. Indeed, the signature of Type I subgraphs is that their density increases with the number of nodes in the network ($N^{I}_{nm}/N\to\infty$ as $N\to\infty$), compared with the Type II subgraphs, whose density

is independent of the network size ($N^{II}_{nm}/N \to$ const). The existence of the Type II subgraphs is a direct consequence of the hierarchical topology: the decreasing $C(k)$ reduces the likelihood that the neighbors of a highly connected node are linked to each other, therefore limiting the chance that these nodes participate in highly connected subgraphs. If $C(k)$ were independent of $k$ (i.e., $\beta$=0), only Type I subgraphs would exist, as in the $\beta \to 0$ limit the $1+1/\beta$ slope of the yellow phase boundary diverges, eliminating all Type II subgraphs. As the absolute count of the subgraphs is the most fundamental quantity for evaluating a local interaction pattern's topological and functional role in a network, we will continue focusing on the direct subgraph count, limiting the discussion on motifs and the role of the randomized reference frame to the Supplementary Material.

The very large densities we observe for some Type I subgraphs (Table 1) requires us to explain how can we distribute as many as $10^{11}$ subgraphs in a network with only $10^3$ nodes. We start addressing this question by calculating the number of distinct subgraphs a given node (gene, metabolite, or protein) participates in. We first focus on the triangle subgraph (3,3), the elementary building block of many higher order subgraphs. The average number of triangles a node with $k$ links participates in is given by $T(k)=C(k)k(k-1)/2$, which for large $k$ scales as $T(k)\sim k^{2-\beta}$. Therefore, we predict that the probability that exactly $T$ triangles pass through a node follows $P(T)\sim T^{-\gamma}$, where $\gamma=1+(\beta+1)/(2-\beta)$, a power-law dependence that indicates that while the majority of nodes participate in at most one or two triangles, a few nodes take part in a very large number of triangle subgraphs. The monotonic nature of $T(k)$ indicates that the triangles are not distributed uniformly within the network, but tend to aggregate around the hubs. As a node with $k$ links can carry up to approximately $k^2$ triangles, the aggregation around the high $k$ hubs, visible e.g. in Fig. 2a and b, allows the network with a modest number of nodes to absorb a very large number of subgraphs. The calculations can be extended to arbitrary $(n,m)$ subgraphs, in each case predicting a power law form for both $T(k)$ and $P(T)$, with exponents that depend on the $(n,m)$ parameters (see the Supplementary Material). To test the validity of these analytical predictions we determined numerically $P(T)$ and $T(k)$ for several subgraphs in each of the studied cellular networks. As shown in Figure 2c and d, the results not only support the predicted power law nature of $P(T)$, but the numerically

determined exponent $\gamma$ is also in good agreement with the analytically predicted values (Table 1).

The fact that the $P(T)$ distribution of the individual subgraphs is uniquely determined by the ($\alpha$,$\beta$) exponents has a quite unexpected consequence: it indicates that the relationship between the network's global architecture and its subgraph densities is reciprocal, such that the network's large-scale topology can be uncovered from the inspection of the local subgraph structure. Indeed, by measuring the $P(T)$ distribution and the corresponding $\gamma$ exponent for any *two* subgraphs (e.g., those shown in Figure 2), and using the derived relationship between $\alpha$, $\beta$ and $\gamma$, we can determine the $\alpha$ and $\beta$ exponents of the overall network. As the scaling region of $P(T)$ is more extended than that of $P(k)$ or $C(k)$, displaying, for example, over five order of magnitudes of scaling in Fig. 2d, such subgraph-based determination of $\alpha$ and $\beta$ can be at times more precise than the direct fitting of $P(k)$ and $C(k)$.

The developed analytical tools allow us to uncover how the various subgraphs relate to each other, an issue that is likely to exert a significant influence on their functional properties. The topological relationship between various subgraphs is illustrated in Figure 3, where we show all nodes participating in several six-node subgraphs ($n$=6) for each of the studied *S. cerevisiae* cellular networks. The figure indicates that the underrepresented Type II subgraphs, shown on the right, are either absent or form small fragmented islands with only a few nodes. As we move towards the Type I subgraphs shown on the left, we not only observe a rapid increase in the subgraph density, but also a spectacular aggregation process, forcing all the high density Type I subgraphs into a single giant cluster, consisting of thousands to millions of highly interconnected subgraphs.

Our analytical methods permit us to uncover the mechanisms of the observed subgraph aggregation, predicting the existence of a percolation condition given by the equation $\beta(m-n+1)-(n-2)<0$, such that the subgraphs satisfying this condition should form a giant cluster. The subgraphs that do not satisfy this condition, however, are allowed to break into isolated islands and/or vanish in size. Direct quantitative evidence for the percolation-like transition is provided by the measurement of the relative size of the largest cluster (shown as squares in Fig. 3), indicating that as we move away from the abundant Type I subgraphs, from left to right, the size of the largest cluster shrinks,

falling particularly rapidly in the vicinity of the predicted percolation transition. The analytical prediction, shown as a continuous line, indicates a good agreement between the predicted and the measured cluster sizes for the two larger networks (metabolic and protein network) (*28*).

The demonstrated equivalence between the local and global topological organization has a number of important consequences for our understanding of cellular networks. First, a series of studies have indicated that the evolution of the large-scale structure of cellular networks may be the consequence of two genome-level mechanisms: gene duplication and the divergence of duplicated molecular interactions due to subsequent mutations (*29-32*). In contrast, the network's local wiring diagram may be shaped by selection towards subgraphs with desirable functional properties (*1-5*). Therefore, while the global structure reflects the network's growth and buildup, it is implied that the local properties reflect evolutionary selection towards desirable functional traits. Our results indicate, however, that a sharp distinction between the local and global structure is not justified: determining the large-scale exponents ($\gamma$ and $\gamma$) is equivalent with specifying the number of subgraphs, while providing the distribution of any two subgraphs uniquely identifies the system's large-scale organization. Thus, if gene duplication and subsequent interaction divergence indeed determines $\gamma$ and $\gamma$(*29-32*), these two mechanisms must define the statistical relevance of the subgraphs as well. Consequently, while it is likely that selection for function plays an important role in shaping the directionality and/or strength of the links (e.g., of the molecular interactions for information processing in transcriptional-regulatory networks (*1, 2*)), the density and topology of motifs and subgraphs cannot be dissociated from the evolution of the overall network. Similarly, the inevitable aggregation of Type I subgraphs into clusters is equally important, implying that statistically abundant subgraphs cannot be viewed as independent network building blocks. Consequently, their potential functional properties need to be evaluated not only at the level of a single subgraph (*1-4*), but also at the level of integrated subgraph clusters.

In conclusion, the demonstrated equivalence of a network's large-scale and local, subgraph-based structure underlies the need to understand the properties and evolution of cellular networks as fully integrated systems, where the achievable local changes are

constrained by the network's global organization. Also, the interdependence between the local and global architecture is by no means limited to cellular networks, but is expected to apply to all networked systems, from the World Wide Web to social networks. Therefore, while there appears to be significant freedom in the construction and function of various complex networks, the kind and abundance of local interaction patterns are uniquely determined by the global characteristics of the emerging network, raising intriguing questions about the role of the local, individual events to shape a network's overall behavior.

## References


1. S. Shen-Orr, R. Milo, S. Mangan, U. Alon, *Nat Genet* **31**, 64-8 (2002).
2. R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, U. Alon, *Science* **298**, 824-27 (2002).
3. S. Mangan, A. Zaslaver, U. Alon, *J Mol Biol* **334**, 197-204 (2003).
4. R. Milo *et al.*, *Science* **303**, 1538-42 (2004).
5. D. M. Wolf, A. P. Arkin, *Curr Opin Microbiol* **6**, 125-134 (2003).
6. J. Hasty, D. McMillen, J.J. Collins, *Nature* **420**, 224-230 (2002).
7. T. I. Lee *et al*, *Science* **298**, 799-804 (2002).
8. S. Wuchty, Z. N. Oltvai, A.-L. Barabási, *Nat Genet* **35**, 176-9 (2003).
9. A. Vespignani, *Nat Genet* **35**, 118-9 (2003).
10. G. C. Conant, A. Wagner, *Nat Genet* **34**, 264-6 (2003).
11. V. F. Hinman, A. T. Nguyen, R. A. Cameron, E. H. Davidson, *Proc Natl Acad Sci U S A* **100**, 13356-61 (2003).
12. R. Dobrin, Q. K. Beg, A.-L. Barabási, Z. N. Oltvai, *BMC Bioinformatics* **5**, 10 (2004).
13. E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, A.-L. Barabási, *Science* **297**, 1551-5 (2002).
14. L. H. Hartwell, J. J. Hopfield, S. Leibler, A. W. Murray, *Nature* **402**, C47-52 (1999).
15. D. Bray, *Science* **301**, 1864-65 (2003).
16. A. Wagner, D. A. Fell, *Proc R Soc Lond B Biol Sci* **268**, 1803-10. (2001).
17. L. Giot *et al.*, *Science* **302**, 1727-36 (2003).
18. S. Li *et al.*, *Science* **303**, 540-3 (2004).
19. H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, A.-L. Barabási, *Nature* **407**, 651-654 (2000).
20. H. Jeong, S. P. Mason, A.-L. Barabási, Z. N. Oltvai, *Nature* **411**, 41-42 (2001).
21. A. Wagner, *Mol Biol Evol* **18**, 1283-92. (2001).
22. S.-H. Yook, Z. N. Oltvai, A.-L. Barabási, *Proteomics* **4**, 928-42 (2004).
23. S. N. Dorogovtsev, A. V. Goltsev, J. F. F. Mendes, *Phys Rev E* **65**, 066122 (2002).
24. J.-P. Eckman, E. Moses, *Proc Natl Acad Sci U S A* **99**, 5825-9 (2002).
25. G. Caldarelli, R. Pastor-Satorras, A. Vespignani, *arXiv.org:cond-mat/*0212026.
26. A. Vázquez, *Phys Rev E* **67**, 56104-15 (2003).
27. D. J. Watts, S. H. Strogatz, *Nature* **393**, 440-2 (1998).

28. In the calculations leading to Fig. 1-3 we have made some simplifications that can be easily relaxed (see Supplementary Material). First, while we focused on the undirected version of cellular networks, the Type I and II subgraphs can be generalized to directed networks as well, which represents a better approximation of the function of the regulatory and metabolic networks. Second, while Fig. 1 is limited to the subset of *n*-node subgraphs that contain triangles, the results can be generalized to other elementary subgraphs as well, such as those containing cycles of four or more nodes.

29. A. Rzhetsky, S. M. Gomez, *Bioinformatics* **17**, 988-96 (2001).

30. J. Qian, N. M. Luscombe, M. Gerstein, *J Mol Biol* **313**, 673-81 (2001).

31. A. Bhan, D. J. Galas, T. G. Dewey, *Bioinformatics* **18**, 1486-93 (2002).

32. R. V. Solé, R. Pastor-Satorras, D. E. Smith, T. Kepler, *Adv Comp Syst* **5**, 43 (2002).

**Acknowledgement**

We thank G. Balázsi for comments on the manuscript. Research at the University of Notre Dame and Northwestern University were supported by grants from the U.S. Department of Energy, the NIH , and the NSF. Correspondence and request for materials should be sent to A.-L.B. (E-mail: alb@nd.edu).

**(a)**

| (*n*,*m*) | | Transcription | | Metabolic | | Protein Interaction |
|---|---|---|---|---|---|---|
| | | *E. coli* | *S. serevisiae* | *E. coli* | *S. cerevisiae* | *S. serevisiae* |
| **(3,2)** |  | 12 | 19 | 101 | 72 | 70 |
| **(3,3)** |  | 0.30 | 0.31 | 5.0 | 5.8 | 4.1 |
| **(4,3)** |  | 169 | 220 | 4412 | 2041 | 2395 |
| **(4,6)** |  | 0.00 | 0.00 | 0.44 | 0.77 | 0.97 |
| **(5,4)** |  | 2492 | 2587 | $2.1\times10^5$ | $5.9\times10^4$ | $1.2\times10^5$ |
| **(5,10)** |  | 0.00 | 0.00 | 0.055 | 0.20 | 0.66 |
| **(6,5)** |  | $3.2\times10^4$ | $2.8\times10^4$ | $8.8\times10^6$ | $1.5\times10^6$ | $5.7\times10^6$ |
| **(6,15)** |  | 0.00 | 0.00 | 0.00 | 0.03 | 0.36 |
| **(7,6)** |  | $3.4\times10^5$ | $2.7\times10^5$ | $3.5\times10^8$ | $3.7\times10^7$ | $2.4\times10^8$ |
| **(7,21)** |  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

**(b)**

| | | | | | | |
|---|---|---|---|---|---|---|
| $\gamma$ | | 2.1±0.3 | 2.0±0.2 | 2.0±0.4 | 2.0±0.1 | 2.4±0.4 |
| $\gamma_{\Delta}$ | | 1.0±0.2 | 1.0±0.2 | 0.8±0.3 | 0.7±0.3 | 1.3±0.5 |
| $\beta$ | **meas.** | 1.0±0.2 | 0.8±0.2 | 1.1±0.2 | 1.4±0.2 | 0.7±0.2 |
| | **pred.** | 0.97 | 0.95 | 1.2 | 1.3 | 0.7 |
| $\beta_{\Delta}$ | **meas.** | 2.1±0.2 | 2.2±0.2 | 1.8±0.2 | 1.7±0.2 | 2.3±0.2 |
| | **pred.** | 2.0 | 1.9 | 1.8 | 1.8 | 3.0 |

**Table 1. Local and global properties of cellular networks:** Panel **a** shows the relative count $N_{nm}/N$ of the least and most connected subgraphs in each of the five studied cellular networks, where $N_{nm}$ represents the number of the given (*n*,*m*) subgraph found in the network, and *N* is the total number of nodes in the network. The first and second columns list the subgraph codes and show a representative topology. Panel **b** lists the $\gamma$ and $\gamma_{\Delta}$ exponents for each of the studied cellular networks, determined from a direct fit to the $P(k)$ and $C(k)$ functions of the undirected network representation (see Supplementary Material). We also provide the measured and predicted values of the $\beta$ and $\beta_{\Delta}$ exponents, characterizing the average number of triangle (3,3) motifs in which a node with *k* links participates ($T(k)\sim k^{\beta}$) and the distribution of the number of triangle motifs in which a node participates ($P(T)\sim T^{\beta_{\Delta}}$).

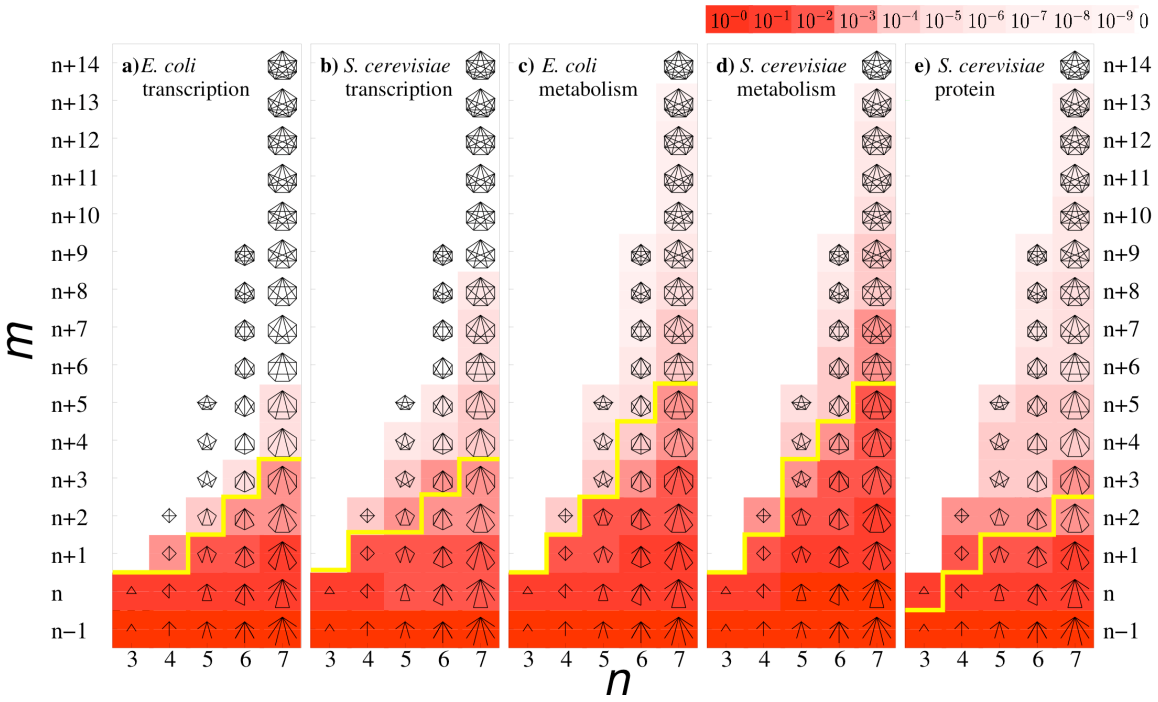


**Figure 1** **Subgraph phase diagrams:** The phase diagrams organize the subgraphs based on the number of nodes ($n$, horizontal axis) and the number of links ($m$, vertical axis), each discrete point explicitly depicting the corresponding subgraph. The stepped yellow line corresponds to the predicted phase boundary separating the abundant Type I subgraphs (below the line) from the constant density Type II subgraphs (above the line). The background color is proportional to the relative subgraph count $C_{nm}=N_{nm}/\sum_s N_{ns}$ of each $n$-node subgraph, the color code being shown in the upper right corner. Note that some $(n,m)$ points in the phase diagram may correspond to several topologically distinguishable subgraphs. For simplicity, we depict only one representative topology in such cases. As the yellow phase boundary depends on the $\square$ and $\square$ exponents of the corresponding network, each phase diagram is slightly different. Yet, there is a visible similarity between the networks of the same kind: the phase diagrams of the two transcription or the two metabolic networks are almost indistinguishable.

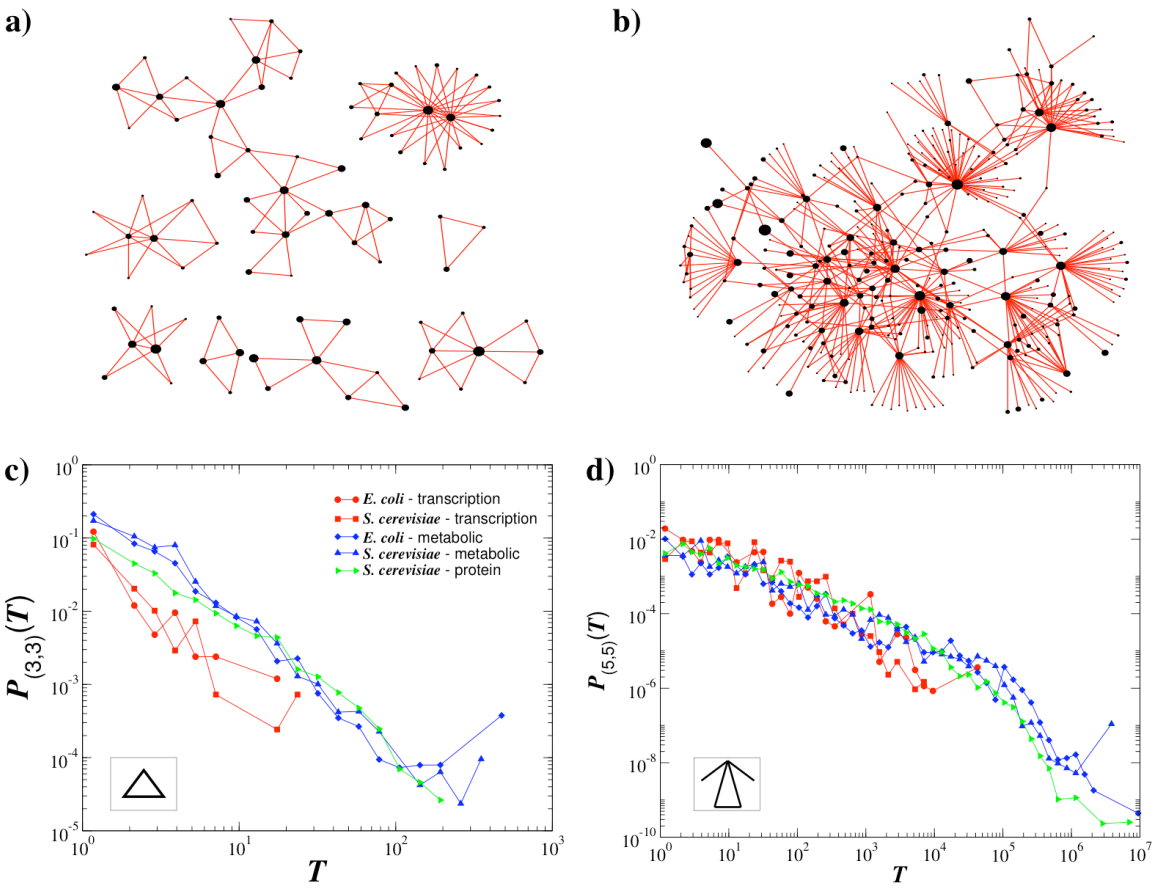


**Figure 2 Subgraph distributions in cellular networks:**

Panels **a** and **b** show all nodes in the *S. cerevisiae* transcription regulatory network that participate in triangle (3,3) and (5,5) subgraphs (depicted in the insets of **c** and **d**). The size (area) of each node is drawn proportional to its degree $k$ in the full network, indicating that subgraphs tend to aggregate around the hubs. Indeed, while there are hubs that have only a few subgraphs around them, in most cases subgraph aggregation is seen only around highly connected nodes. Note that the (3,3) subgraph of *S. cerevisiae* is above the percolation boundary (Fig. 3a), and therefore they are broken into small islands. In contrast, the (5,5) subgraph is well below the boundary, forming a fully connected giant component, with no isolated subgraphs, as predicted. The bottom panels show the $P(T)$ distribution of the number of **(c)** (3,3) and **(d)** (5,5) subgraphs passing by a node, the different colors corresponding to the different cellular networks. The plot indicates that for both subgraphs $P(T)$ approximates a power law $P(T) \sim T^{-\alpha}$. Note the quite extended scaling regimes for some networks: for example for the (5,5) subgraph the scaling extends over four-five orders of magnitude. The $\alpha$ exponents measured and predicted for each network are summarized in Table 1b and the Supplementary Material.

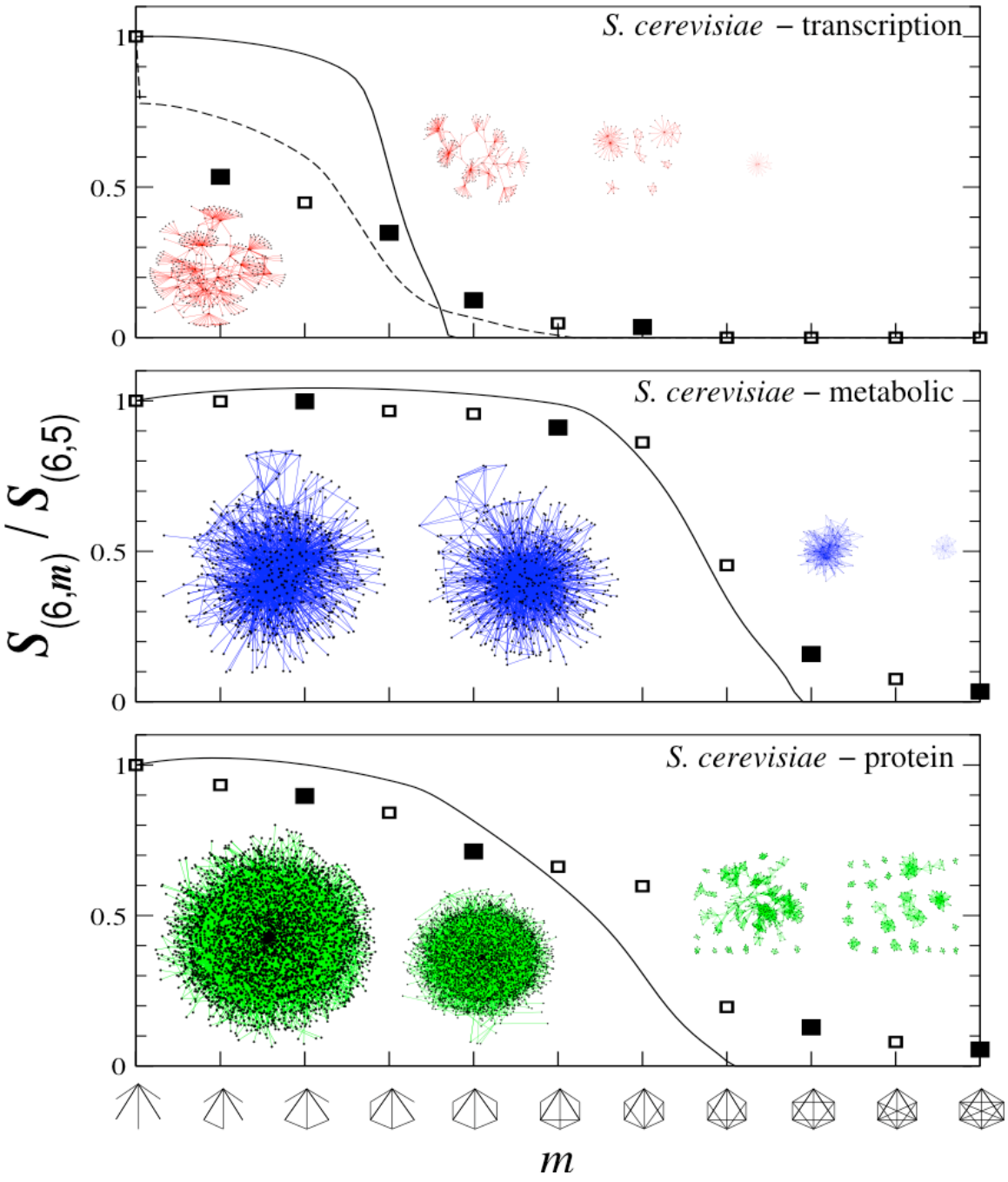


**Figure 3.** **Subgraph aggregation and percolation:** The horizontal axis shows the sequence of $n$=6 subgraphs, the number of links ($m$) increasing from left to right. The vertical axis corresponds to the relative size of the largest cluster for the subgraphs shown on the horizontal axis, being determined by the $S_{(6,m)}/ S_{(6,5)}$ ratio, where $S_{(6,m)}$ represents the number of nodes participating in ($n$=6,$m$) subgraphs and $S_{(6,5)}$ represents the total number of nodes participating in the first and most abundant subgraph of the $n$=6 subgraph family. The square symbols represent the measured value of the $S_{(6,m)}/ S_{(6,5)}$ ratio for the *S. cerevisiae* networks listed in the upper right corner, indicating that the relative size of the subgraph cluster shrinks from close to one to zero as we move from the highly abundant

Type I subgraphs to the low density Type II subgraphs. The topological consequences of the predicted transition can be seen on the inserted network maps, each corresponding, in order, to the four filled symbols. The sequence of maps demonstrates that while the Type I subgraphs all aggregate into a giant subgraph cluster, as we move towards the Type II subgraphs, the cluster shrinks rapidly in the vicinity of the predicted percolation transition, and disappears by either shrinking to close to zero size (see e.g. the metabolic network) or by breaking into many small islands, which also disappear by further shrinking (see e.g. the transitional-regulatory and protein interaction networks). The continuous line, corresponding to our analytical prediction for the relative cluster size is in quite good agreement with the measured curve for the relatively large protein interaction and metabolic networks. The particular shape of the curve depends, however, on the functional form we use for $C(k)$. For example, the continuous curves were obtained using the analytic approximation $C(k)=C_0/[1+(k/k_0)^{\square}]$. In contrast, the agreement for the transcriptional-regulatory network can be significantly improved by replacing this fit with the directly measured $C(k)$ (dashed line), reproducing even the sharp drop for the relative density of the least connected cluster (first symbol in the top panel).