Global constraints on local interaction patterns in cellular networks Supplementary material

Alexei Vázquez,[1] Radu Dobrin,[2] Danilo Sergi,[3] Jean-Pierre Eckmann,[3,4] Zoltán N. Oltvai,[2] and Albert-László Barabási[1]

[1]*Department of Physics, University of Notre-Dame, Notre Dame, IN 46556, USA*
[2]*Department of Pathology, Northwestern University, Chicago, IL 60661, USA*
[3]*Département de Physique Théorique, Université de Genève, 1211 Genève 4, Switzerland*
[4]*Section de Mathématiques, Université de Genève, 1211 Genève 4, Switzerland*

Contents

I. BIOLOGICAL NETWORK REPRESENTATIONS AND DATABASES

A. Transcriptional regulatory networks

In transcription regulatory networks the nodes represent transcription factors and genes while the links describe regulation-based interactions. Each link is directed, pointing from the transcription factor to the gene that it regulates. To compare the topological properties of transcription-regulatory networks to those of metabolic and protein-protein interaction networks, we have studied their undirected version where each directed link is replaced by an undirected one.

Data source: The transcription network data of *E. coli* has been reported in [2] and it is available from http://www.weizmann.ac.il/mcb/UriAlon/. The transcription network data of *S. cerevisiae* has been reported in [8], and it is available from http://www.weizmann.ac.il/mcb/UriAlon/Papers/networkMotifs/sma.html.

B. Metabolic networks

Metabolic networks represent the sum of all biochemical reactions taking place in an organism. Different network representations are possible, depending on the purpose [4]. In our work we associate nodes with metabolites. Undirected links connect each substrate to the products of the reaction in which the substrate participates.

Data source: The metabolic networks of *E. coli* and *S. cerevisiae* where obtained from the WIT database (http://igweb.integratedgenomics.com/IGwit/), and they are also available at http://www.nd.edu/∼networks/.

C. Protein-protein interaction networks

In the protein-protein interaction network nodes represent proteins and undirected links describe the pairwise interactions between them.

Data source: We focus on the protein-protein interaction network of *S. cerevisiae*. The data source is the Database of Interacting Proteins (DIP) available at http://dip.doe-mbi.ucla.edu/. For a comparison between the different protein interaction databases see [1].

In summary, we constructed an undirected network representation of the transcription, the metabolic and the protein-protein interaction networks of *E. coli* and *S. cerevisiae*. In order to have a coherent basis of comparison we have reduced the original networks by removing self-loops, double links (i.e., if we have simultaneus links from node A to node B and from B to A, we draw only one undirected link joining the two A and B) and isolated nodes. The basic quantities characterizing the resulting networks are summarized in Tab. I.

II. BASIC TOPOLOGICAL PROPERTIES

We denote by N the number of nodes in the network, usually referred to as the network size, and by E the number of links (or edges). Each node i is characterized by the degree k_i and the clustering coefficient C_i.

The *degree* k_i counts the number of interactions that node i has with the other nodes in the network. By averaging k_i over all nodes in the network we obtain the average degree

Network	N	E	$\langle k\rangle$	$\langle c\rangle$	$\langle k\rangle/N$
EC Transcriptional	418	519	2.5	0.086	0.0059
SC Transcriptional	688	1078	3.13	0.047	0.0046
EC Metabolic	884	2739	6.2	0.169	0.0070
SC Metabolic	551	1698	6.2	0.227	0.0112
SC Protein	5068	15117	5.97	0.085	0.0012

TABLE I: Network statistics after the removal of double links, self-loops and isolated nodes. N denotes the number of nodes, E the number of edges, $\langle k\rangle$ the average degree, and $\langle c\rangle$ the average clustering coefficient of the studied networks. The ratio $\langle k\rangle/N$ gives the average clustering coefficient for a random network with the same pair of parameters (N, E).

$$\langle k\rangle \equiv \frac{1}{N}\sum_{i=1}^{N} k_i = \frac{2E}{N} \ . \tag{1}$$

The last equality follows from the fact that each link connects two nodes, and hence contributing to the degree of two nodes. The values of N, E and $\langle k\rangle$ for the investigated biological networks are given in Table I. Note that for each network $\langle k\rangle \ll N$, meaning that biological networks are sparse, i.e. on average each node is connected to only a few other nodes.

The *clustering coefficient* C_i is a measure of the fraction of connected neighbors of node i. A node with k_i links can have at most $\binom{k_i}{2} = k_i(k_i-1)/2$ pairs of its neighbors connected to each other. If we denote by t_i the number of links among the neighbors of node i, then the clustering coefficient is defined as

$$C_i \equiv \frac{2t_i}{k_i(k_i-1)} \ . \tag{2}$$

The average of C_i over all nodes in the network is

$$\langle C\rangle \equiv \frac{1}{N}\sum_{i=1}^{N} C_i \ . \tag{3}$$

The values of $\langle C\rangle$ for the analyzed biological networks are given in Table I. For comparison we also list the clustering coefficient of a random network with the same parameters N and E. Note that $\langle C\rangle \gg \langle k\rangle/N$; therefore, biological networks are characterized by a high degree of local clustering [5].

The average values $\langle k\rangle$ and $\langle C\rangle$ do not characterize exhaustively the topology since biological networks exhibit marked degree fluctuations [3, 4]. Indeed, the probability $P(k)$ for a node to have degree k is given by the power law

$$P(k) = Ak^{-\gamma} \ , \tag{4}$$

where A is a constant and γ is the *degree exponent* (see Fig. 1).

The highly inhomogeneous nature of $P(k)$ prompted the study of the clustering coefficient as a function of the degree. Instead of taking the average over all nodes in the network, as in Eq. (3), the average is taken over nodes with the same degree, providing

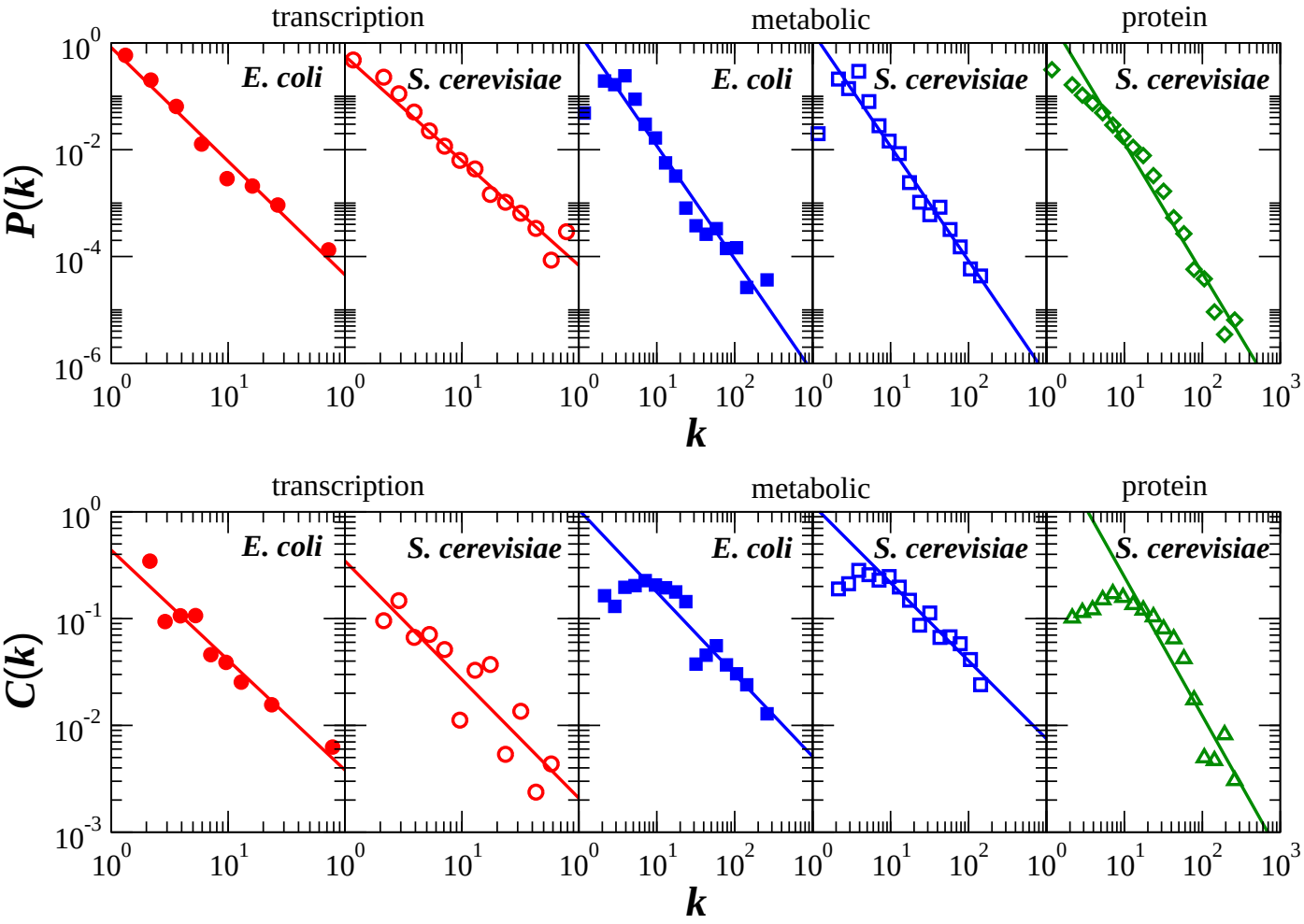

FIG. 1: Connectivity distribution and clustering coefficient for the networks discussed in the paper. The solid lines correspond to the best fit to a power law.

$$C(k) = \frac{\sum_{i=1}^{N} C_i \delta_{k_i k}}{\sum_{i=1}^{N} \delta_{k_i k}} \ , \tag{5}$$

where $\delta_{k_i k} = 1$ if $k_i = k$, and zero otherwise. For biological networks $C(k)$ follows a power law

$$C(k) = C_0 k^{-\alpha} \ , \tag{6}$$

where C_0 is a constant and α is the *hierarchical exponent* (see Fig. 1). The hierarchical exponent characterizes the network's overall modularity, indicating the presence of many small, highly interconnected modules forming larger, less cohesive topological modules.

The degree and hierarchical exponents, γ and α respectively, rather than the average degree $\langle k \rangle$ and the average clustering coefficient $\langle C \rangle$, are the two important topological quantities characterizing biological networks. Their values for the studied biological networks are summarized in Table I of the paper.

III. STATISTICAL PROPERTIES OF SUBGRAPHS

A. Number of triangles passing by a node

To develop an intuition for the statistical properties of general subgraphs we first study the occurrence of triangles. Since each connected pair of neighbors forms a triangle with the central node i, then t_i defined in (2) also represents the number of triangles passing by node i. The average of t_i over nodes of same degree k, denoted by $T(k)$, is

$$T(k) = \binom{k}{2} C(k) \ . \tag{7}$$

Using $\binom{k}{2} \sim k^2/2$, we find

$$T(k) \sim k^{\beta} \ , \tag{8}$$

with

$$\beta = 2 - \alpha \ . \tag{9}$$

Combining Eqs. (4) and (8), we predict that the probability that a randomly selected node participates in T triangles follows a power law

$$P(T) \sim T^{-\delta} \ , \tag{10}$$

where

$$\delta = 1 + \frac{\gamma - 1}{\beta} \ . \tag{11}$$

$P(T)$ and $T(k)$ for the studied biological networks are plotted in Fig. 2 together with the power law fits. The β and δ exponents provided by the fit are listed in Table I of the paper, indicating that the predicted values are in good agreement with the measured ones.

B. Number of small subgraphs passing by a node

The above analysis based on triangles can be extended to other small subgraphs as well. We consider subgraphs with n nodes and m links that can be decomposed into a central node and $n-1$ neighbors. More precisely, we have $n-1$ interactions from the central node to its neighbors and

$$t = m - (n-1)$$

interactions among the $n-1$ neighbors. In the paper we have used m together with n to classify each motif; here, to simplify the calculations we use t instead of m. The total number of n-node subgraphs that can pass by a node with degree k is $\binom{k}{n-1}$, which tells us in how many different ways we can choose a group of $n-1$ nodes out of the total k. Each of these n-node subgraphs can have at most $n_{\rm p} = (n-1)(n-2)/2$ links between the $n-1$ neighbors of the central node. The probability that there is a link between two neighbors of a node with degree k is by definition $C(k)$, and the

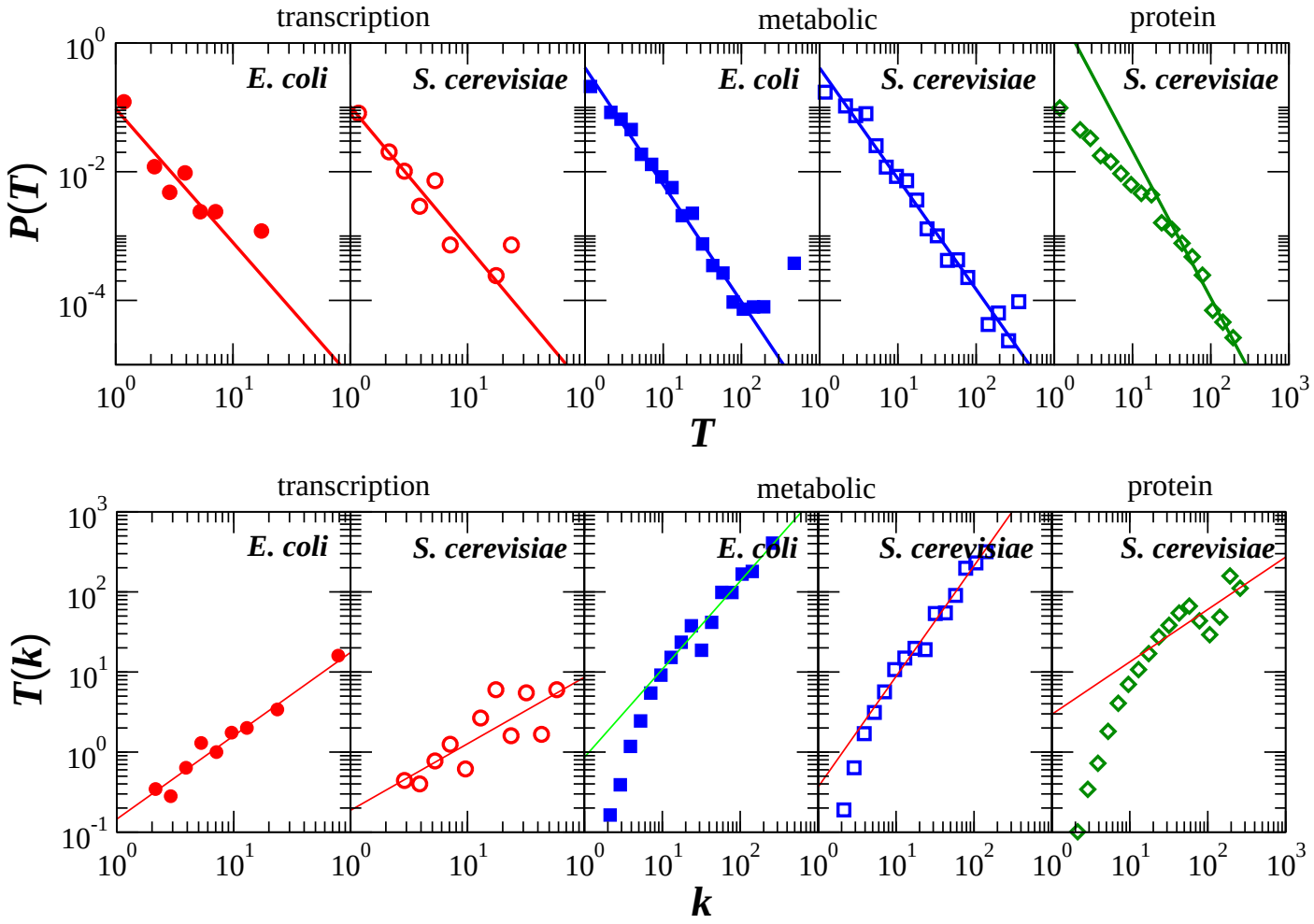

FIG. 2: The distribution $P(T)$ of the number of triangles passing by a node and the average number of triangles $T(k)$ passing by a node with degree k. The solid lines correspond to the best fit to a power law. Note that the scaling of $T(k)$ for the protein interaction network of *S. cerevisiae* is not partially good, a result of the cutoffs in the $P(k)$ and $C(k)$ functions in this network. The scaling of the subgraphs in (Fig. 2 in the paper), however, is not affected by these cutoffs, offering a better way to determine the scaling exponents.

probability that two neighbors are not connected is $1 - C(k)$. Therefore, the probability to obtain t connected pairs and $n_{\rm p} - t$ disconnected pairs is given by the binomial distribution

$$b_{nt}(k) = \binom{n_{\rm p}}{t} C(k)^t [1 - C(k)]^{n_{\rm p}-t} \ . \tag{12}$$

The average number of subgraphs formed by $n-1$ neighbors and t interactions among them and centered at a node with degree k is given by

$$N_{nt}(k) = \binom{k}{n-1} b_{nt}(k) \ . \tag{13}$$

We readily obtain that if k is large, then $N_{nt}(k)$ scales as

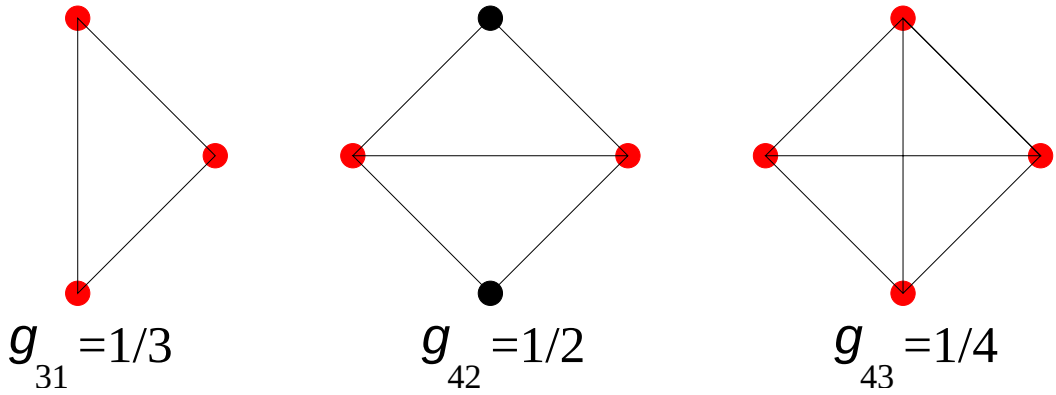

FIG. 3: Some examples illustrating the over-counting of subgraphs. In our approach we count subgraphs with a central node connected to all other nodes in the subgraphs. In some cases a subgraph may have more than one node satisfying this condition. The number of potential central nodes of an (n,t) subgraph is counted by the combinatorial factor $1/g_{nt}$.

$$N_{nt}(k) \sim k^{\beta_{nt}} \ , \tag{14}$$

where

$$\beta_{nt} = n - 1 - \alpha t = n - 1 - \alpha(m - n + 1) \ . \tag{15}$$

Using Eqs. (4) and (14) we find that the probability for a randomly selected node to participate in T_{nt} subgraphs (n,t) scales as

$$P(T_{nt}) \sim T_{nt}^{-\delta_{nt}} \ , \tag{16}$$

where

$$\delta_{nt} = 1 + \frac{\gamma - 1}{\beta_{nt}} \ . \tag{17}$$

In particular, triangles correspond to $n = 3$ and $t = 1$ (i.e., $m = 3$), leading to $\beta_{31} = 2 - \alpha$ and $\delta_{31} = 1 + (\gamma - 1)/\delta$, and we thus recover the expressions (9) and (11) derived earlier for triangles.

Note that this approach can be generalized to higher order cycles, such as squares, pentagons and others.

C. Over- (Type I) and under-represented (Type II) subgraphs

Here we use the scaling relations derived above to address the abundance of (n,t) subgraphs in scale-free and hierarchical networks.

The expected number of (n,t) subgraphs in a network is given by Eq. (13)

$$N_{nt} = g_{nt} N \sum_{k=1}^{k_{\max}} P(k) N_{nt}(k) \ , \tag{18}$$

where $k_{\max}$ is the maximum degree and the factor g_{nt} takes into account that the same subgraph can have more than one node as a center (see Fig. 3). Eqs. (4), (6) and (12-18) lead to

$$N_{nt} = NAg_{nt}\binom{n_{\rm p}}{t}C_0^t\sum_{k=1}^{k_{\max}}k^{-\gamma-\alpha t}\binom{k}{n-1}[1-C_0k^{-\alpha}]^{n_{\rm p}-t}\ . \tag{19}$$

We recognize the existence of two regimes. There is a regime where the sum in the above expression is dominated by the large k region; this means that we can approximate $\binom{k}{n-1}$ by $k^{n-1}/(n-1)!$ and $1-C_0k^{-\alpha}$ by 1, being left with

$$N_{nt} \approx N\frac{Ag_{nt}\binom{n_{\rm p}}{t}C_0^t}{(n-1)!}\sum_{k=1}^{k_{\max}}k^{n-1-\gamma-\alpha t}\ . \tag{20}$$

Type I subgraphs: if $n-\gamma-\alpha t>0$ the sum in the above expression diverges with $k_{\max}$, allowing us to approximate it by an integral yielding

$$N_{nt}^I \approx \eta_{nt}^I N k_{\max}^{n-\gamma-\alpha t}\ , \tag{21}$$

where

$$\eta_{nt}^I = \frac{Ag_{nt}\binom{n_{\rm p}}{t}C_0^t}{(n-\gamma-\alpha t)(n-1)!}\ . \tag{22}$$

Type II subgraphs: If $n-\gamma-\alpha t<0$, we are in the second regime, as the sum in (19) does not diverge with $k_{\max}$, leading to

$$N_{nt}^{II} \approx \eta_{nt}^{II}N\ , \tag{23}$$

where

$$\eta_{nt}^{II} = Ag_{nt}\binom{n_{\rm p}}{t}C_0^t\sum_{k=1}^{\infty}k^{-\gamma-\alpha t}\binom{k}{n-1}[1-C_0k^{-\alpha}]^{n_{\rm p}-t}\ . \tag{24}$$

Note that η_{nt}^I and η_{nt}^{II} do not depend on the network size N. Since $k_{\max}$ is large in biological networks, the number of Type I subgraphs (21) will be significantly larger than the number of Type II subgraphs (23). In particular, we find that the density (number of subgraphs/N) of Type II subgraphs is independent of N, and the density of Type I subgraphs increases with the total number of nodes in the system.

To emphasize the difference between Type I and Type II subgraphs we introduce the normalized subgraph count

$$C_{nt} = \frac{N_{nt}}{\sum_{s=0}^{n_{\rm p}}N_{ns}}\ , \tag{25}$$

representing the fraction of interaction patterns of n nodes and t interactions relative to the total number of n-node subgraphs (within the subgraph class defined above). For large networks, the sum

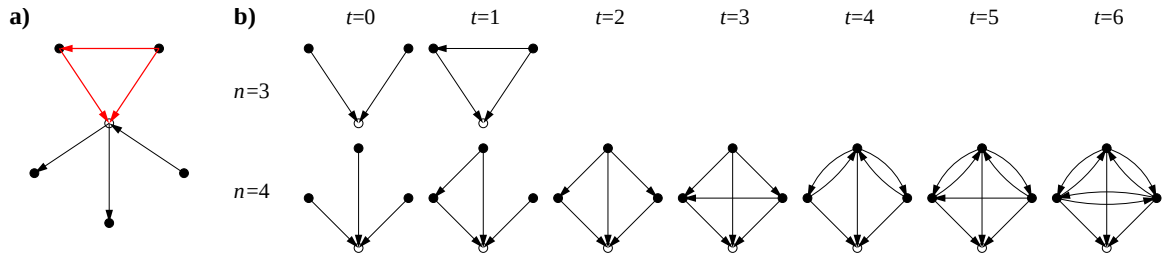

FIG. 4: **a)** An example of a node with in-degree $k_{\rm in} = 2$, out-degree $k_{\rm out} = 3$, and one feed-forward-loop (FFL) passing by it (red arrows). The nodes with a link pointing to the central node (open circle) are called *in-neighbors*, while nodes that are pointed at by the central node are called *out-neighbors*. **b)** Examples of directed n node subgraphs, with a central node (empty circle) and $n-1$ in-neighbors, forming t FFLs.

in the denominator is dominated by the $s = 0$ term, and thus substituting Eqs. (21) and (23) into (25) we obtain

$$C^I_{nt} \approx \frac{\eta^I_{nt}}{\eta^I_{n0}} k_{\rm max}^{-\alpha t} \ , \tag{26}$$

$$C^{II}_{nt} \approx \frac{\eta^{II}_{nt}}{\eta^I_{n0}} k_{\rm max}^{-(n-\gamma)} \tag{27}$$

The comparison of these predictions with the direct subgraph count for each biological network is shown in Fig. 1 of the paper.

D. Directed subgraphs

Many biological networks, from the transcription regulatory to the metabolic network, are directed. Next we generalize our calculations to describe directed subgraphs as well, showing that distinction between Type I and Type II subgraphs is relevant for directed networks as well. As an example, we study the aggregation of the three node subgraph shown in Fig. 4a, often called feed-forward loop (FFL). For simplicity, we will consider the set of subgraphs made of n nodes, with one central node and $n-1$ in-neighbors (see Fig. 4). In this case the central node (open circle in Fig. 4) is characterized by in-degree $k_{\rm in}$. Each subgraph will be characterized by the number of nodes n, as well as t copies of FFLs in the subgraph (see Fig. 4b). In analogy with the clustering coefficient, we define an FFL clustering coefficient, denoted by $C_{\rm FFL}$, as the number t of FFLs passing by a node divided by the maximum number of FFLs that can pass by it, given $k_{\rm in}$. Since there may be directed links between the in-neighbors (see Fig. 4), the total number of FFLs that can pass by a node with in-degree $k_{\rm in}$ is $k_{\rm in}(k_{\rm in}-1)$, obtaining

$$C_{\rm FFL}(k_{\rm in}) = \frac{t}{k_{\rm in}(k_{\rm in}-1)} \ . \tag{28}$$

Note that $C_{\rm FFL}(k_{\rm in})$ represents the probability that two in-neighbors form an FFL, allowing us to compute the number of n node subgraphs with t FFLs (see Fig. 4b) passing by a node with $k_{\rm in}$.

The total number of n node subgraphs that can pass by a node with in-degree $k_{\rm in}$ is given by $\binom{k_{\rm in}}{n-1}$, representing the number of groups of $n-1$ in-neighbors chosen from $k_{\rm in}$ nodes. Among the $n-1$ in-neighbors we can form $n_{\rm p}=(n-1)(n-2)$ pairs (note that $i\to j$ is different from $j\to i$). The probability that t pairs form an FFL and $n_{\rm p}-t$ pairs do not is given by the binomial distribution

$$b_{nt}(k_{\rm in})=\binom{n_{\rm p}}{t}C_{\rm FFL}(k_{\rm in})^t[1-C_{\rm FFL}(k_{\rm in})]^{n_{\rm p}-t}\ . \tag{29}$$

Therefore, the average number of (n,t) directed subgraphs passing by a node with in-degree $k_{\rm in}$ is given by

$$N_{nt}(k_{\rm in})=\binom{k_{\rm in}}{n-1}b_{nt}(k_{\rm in})\ . \tag{30}$$

Finally, summing over the in-degree distribution we obtain the total number of (n,t) directed subgraphs in the network

$$N_{nt}=N\sum_{k_{\rm in}}P(k_{\rm in})N_{nt}(k_{\rm in})\ , \tag{31}$$

Using $C_{\rm FFL}(k_{\rm in})\sim k_{\rm in}^{-\alpha_{\rm FFL}}$ and $P(k_{\rm in})\sim k^{-\gamma_{\rm in}}$ (which we have verified numerically for selected examples) and Eqs. (29)-(31), we obtain two different subgraph phases for undirected networks. The Type I subgraphs are those satisfying $n-\gamma_{\rm in}-\alpha_{\rm FFL}>0$, their number growing faster than N. The Type II subgraphs are in turn those satisfying $n-\gamma_{\rm in}-\alpha_{\rm FFL}<0$, their number growing as N. These results indicate that the distintion between Type I and Type II subgraphs obtained for undirected networks is present in directed networks as well. While the set of directed subgraphs considered above is incomplete (i.e. one needs to derive the scaling for other directed subgraphs as well), it is sufficient to show the apperance of Type I and II subgraphs for directed networks.

E. Overrepresentation relative to randomized networks: Motifs

Recently, the "motif" concept has been introduced to refer to subgraphs that are overrepresented with respect to a randomized network [2, 8]. The aim of this subsection is to study the abundance of Type I subgraphs relative to such randomized networks and make contact with the motif concept.

Some care is necessary in relation with the meaning of "randomized" networks, which is not unique. In general we can consider the ensemble of random networks that preserve any of the following topological properties: (1) number of nodes and links, (2) degree distribution, (3) total number of triangles. Depending on the choice of randomization, one obtains different results on the abundance of specific subgraphs relative to the randomized networks. Our approach, however, can be developed for any particular randomization.

The abundance of Type I subgraphs relative to randomized networks can be measured by [8]

$$Z_{nt}=\frac{N_{nt}^{\rm (rd)}}{\sigma_{nt}^{\rm (rd)}}\left(\frac{N_{nt}^{I}}{N_{nt}^{\rm (rd)}}-1\right)\ , \tag{32}$$

where $N_{nt}^{\rm (rd)}$ and $\sigma_{nt}^{\rm (rd)}$ are the average number and the standard deviation of the (n,t) subgraphs in the randomized graph. $N_{nt}^{\rm (rd)}$ can be computed as shown in III C. We assume that the degree

Network	$C_{\rm rd}$	$C_{k_{\rm max}}$	$C_{k_{\rm max}}/C_{\rm rd}$
E. coli-transcription	0.10	0.0063	0.063
S. cerevisiae-transcription	0.071	0.0044	0.062
E. coli-metabolic	0.21	0.013	0.062
S. cerevisiae-metabolic	0.19	0.024	0.12
S. cerevisiae-protein	0.021	0.0030	0.14

TABLE II: Clustering coefficient of the node of maximum degree, $C_{k_{\rm max}}$, and average clustering coefficient for the randomized graphs, $C_{\rm rd}$, for the different biological networks studied in the paper.

distribution is preserved, and that the clustering coefficient is independent of the node degree, as it is the case in randomized graphs usually used [2, 8]. This means that $C(k) \approx C_{\rm rd}$, or equivalently $\alpha = 0$, which, from Eq. (19), leads to

$$N_{nt}^{(\rm rd)} \approx k_{\rm max}^{n-\gamma}\ \eta_{nt}^{(\rm rd)}\ , \tag{33}$$

where

$$\eta_{nt}^{(\rm rd)} = \frac{A g_{nt} \binom{n_{\rm p}}{t} C_{\rm rd}^t}{(n-\gamma)(n-1)!}\ . \tag{34}$$

Note that, in contrast to hierarchical networks with a finite α, the randomized networks are characterized by a single phase, Type I. Substituting (21), (22), (33) and (34) into (32) we obtain

$$Z_{nt}^I \approx \frac{N_{nt}^{(\rm rd)}}{\sigma_{nt}^{(\rm rd)}} \left(r_{nt}^I - 1\right)\ , \tag{35}$$

where

$$r_{nt}^I = \frac{n-\gamma}{n-\gamma-\alpha t} \left(\frac{C_{k_{\rm max}}}{C_{\rm rd}}\right)^t\ , \tag{36}$$

$C_{k_{\rm max}} = C_0 k_{\rm max}^{-\alpha}$ is the clustering coefficient of the node with maximum degree and $\alpha t - (n-\gamma) < 0$.

The strength of the over- or under-representation is modulated by the prefactor in Eq. (35) and its sign is determined by Eq. (36). If $r_{nt}^I > 1$ then the (n,t) subgraph is overrepresented with respect to the randomized graph; on the other hand, if $r_{nt}^I < 1$ then the (n,t) subgraphs is underrepresented. To compute r_{nt}^I we need only the exponents γ and α, and the clustering coefficients $C_{k_{\rm max}}$ and $C_{\rm rd}$ for the original and randomized networks. The clustering coefficient of a random graph with an arbitrary degree distribution is [6]

$$C_{\rm rd} = \frac{\langle k\rangle}{N} \left(\frac{\langle k^2\rangle - \langle k\rangle}{\langle k\rangle^2}\right)^2\ . \tag{37}$$

Since the randomized network has the same degree distribution as the original one, the averages $\langle k\rangle$ and $\langle k^2\rangle$ are exactly the same as those for the original graph. The values of $C_{k_{\rm max}}$ (estimated from Fig. 1) and $C_{\rm rd}$, Eq. (37), for the biological networks studied in the paper are shown in Table

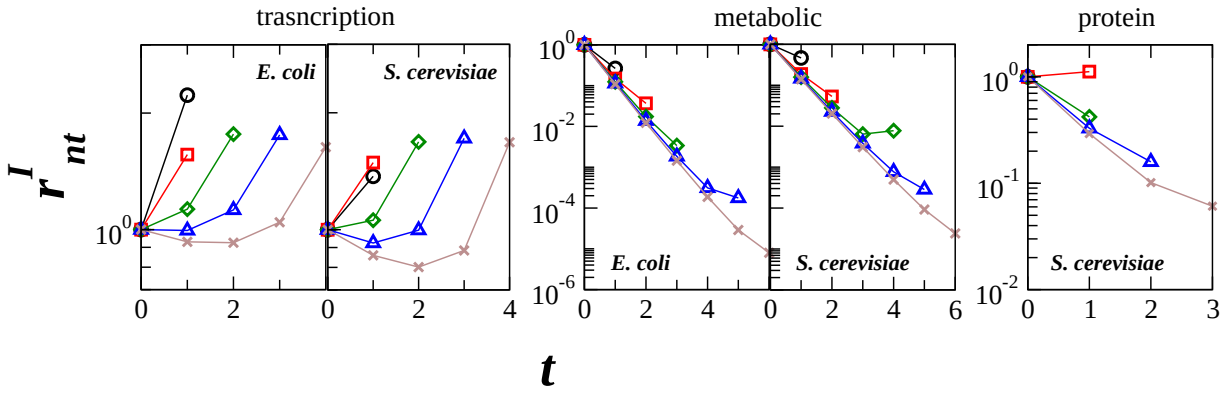

FIG. 5: Subgraph over- under-representation relative to the randomized network for Type I subgraphs ($t \leq (n-\gamma)/\alpha$). Positive and negative values of z^I_{nt} correspond to overrepresented and underrepresented subgraphs, respectively. The symbols correspond to $n = 3$ (circles), $n = 4$ (squares), $n = 5$ (diamonds), $n = 6$ (triangles) and $n = 7$ (crosses). Note that the value of t is given by the x-axis.

II. Using these values we have computed r^I_{nt}, Eq. (36), as shown in Fig. 5. The figure indicates that, when compared to a randomized network, Type I subgraphs can be over or underrepresented, depending on the detailed network parameters. The absence of monotonic trends in the relative subgraph count indicates that in order to fully understand the absence or abundance of certain subgraphs, it is simpler to inspect their absolute number. A comparison to a randomized graph masks the clear trends seen in the absolute subgraph counts, enhancing subgraphs that may not be particularly abundant in the network, and supressing some abundant subgraphs. As this is important in some applications, should one wish to determine the subgraph abundance relative to a randomized network, the methods discussed above allow us to do so in a systematic fashion.

IV. SUBGRAPH PERCOLATION AND SUBGRAPH CLUSTERS

The observed agglomeration of subgraphs around the hubs does not exclude the possibility that many isolated subgraphs are still scattered around the less connected nodes. Indeed, if certain subgraphs are selected because they have desirable signal processing properties, the underlying selection principles work only if subgraphs are indeed isolated. In the following we show, however, that subgraphs do not occur independently in biological networks, but the large-scale constraints forces them together into large subgraph clusters. We first focus on triangles, removing from the network all nodes and links that do not participate in at least one triangle. If the triangle subgraphs were independent of each other, the network would break into many independent, or lightly connected triangles. Fig. 3 in the paper shows that this is not the case: subgraphs typically form a large subgraph cluster, with at most a few subgraphs existing in smaller clusters, and almost surely none individually.

Subgraph clusters are an inevitable consequence of the high subgraph density. Indeed, the probability that a link connected to a node of degree k participates in a triangle is given by $q(k) = 1 - [1 - C(k)]^{k-1}$, which is the probability that the neighbor at the other end of the link is connected to at least one of the $k - 1$ remaining neighbors. If $C(k) = C_0 k^{-1}$, for large k we obtain $q \approx 1 - \exp(-C_0)$, independent of the node degree. As a consequence, calculating the size of the largest triangle cluster is equivalent with determining the largest cluster of connected nodes, after each link is removed with probability $1-q$. As scale-free networks do not break apart under random node or link removal [3], the removal of a $1 - q$ fraction of links should leave behind a giant cluster

of connected nodes.

This analysis can be extended to networks whose clustering coefficient scales as $C(k) = C_0 k^{-\alpha}$ and other subgraphs. Let us compute the probability that a link going out from a node with degree k participates in an (n,t) subgraph, by proceeding as in the calculation of N_{kt}. Consider a node i with degree k and one of its neighbors, say j. We consider all groups of n nodes containing i, j and any other neighbor of i. The number of such groups is given by the number of groups of $n-2$ nodes (the missing two nodes are i and j) that can be made out of $k-1$ nodes (the missing neighbor is j), i.e. $\binom{k-1}{n-2}$. On each group, we can form $n_{\rm p} = (n-1)(n-2)/2$ pairs of neighbors, each pair connected with probability $C(k)$ and disconnected with probability $1-C(k)$. The probability to obtain t connected pairs and $n_{\rm p}-t$ disconnected pairs is given by the binomial distribution $b_{nt}(k)$, Eq. (12), which is in general small. The probability that the link connecting i and j participates in at least one (n,t) subgraph having node i at its center is given by

$$\begin{aligned} q_{nt}^{(0)}(k) &= 1 - [1 - b_{nt}(k)]^{\binom{k-1}{n-2}} \\ &\approx 1 - \exp\left[\binom{k-1}{n-2} b_{nt}(k)\right] \ . \end{aligned} \tag{38}$$

Using (13) this expression can be approximated by

$$q_{nt}^{(0)}(k) \approx 1 - \exp\left(-\frac{n-1}{k} N_{nt}(k)\right) \ . \tag{39}$$

Furthermore, the link (i,j) can participate in an (n,t) subgraph with j being the central node. In the absence of degree correlations between the conncted nodes, the probability that j belongs to at least one (n,t) subgraph is the average of (39) over the degrees of its immediate neighbors, distributed as $p(k)k/\langle k\rangle$, yielding

$$q_{nt}^{(1)} = \frac{\sum_k p(k) k q_{nt}^{(0)}(k)}{\langle k\rangle} \ . \tag{40}$$

Finally, the probability that a link going out from a node with degree k participates in at least one (n,t) subgraph, centered either at i or j, is given by

$$q_{nt}(k) = 1 - \left(1 - q_{nt}^{(0)}(k)\right)\left(1 - q_{nt}^{(1)}\right) \ . \tag{41}$$

Eqs. (39)-(41) tell us that the participation in an (n,t) subgraph of a link going out from a node with degree k is related to the average number $N_{nt}(k)$ of subgraphs (n,t) passing by a node with degree k. Furthermore, since the cohesiveness of scale-free networks is essentially determined by the hubs, the existence or not of a giant cluster of connected (n,t) subgraphs is determined from the asymptotic behavior of $q_{nt}(k)$ for large k. From Eq. (13) follows that

$$\frac{1}{k} N_{nt}(k) \sim k^{n-2-\alpha t} \tag{42}$$

for $k \gg n$. By using this expression we can distinguish two regimes. For $n-2-\alpha t > 0$ we have

$$q_{nt} \approx 1 \ , \tag{43}$$

and thus the links incident to high degree nodes participate with probability very close to 1 in at least one (n,t) subgraph. Therefore, the existence of a giant cluster of (n,t) connected subgraphs is predicted. In the other regime, $n-2-\alpha t<0$, we have

$$q_{nt}\approx 0\ , \tag{44}$$

and the links incident to high degree nodes do not participate with probability almost one in any (n,t) subgraph, suggesting the possible fragmentation of the network. The condition $n-2-\alpha t=0$ gives a lower bound for the value of t at which the relative size of the largest cluster becomes zero, but there may still be a giant component for $n-2-\alpha t<0$. It is worth noting that in most cases the boundary $n-2-\alpha t=0$ coincides with the phase boundary associated with the subgraph's representation $n-\gamma-\alpha t=0$. This overlap is determined by the fact that γ is quite close to 2 in biological networks and that t takes only discrete values.

To determine whether there is a giant component in the regime $n-2-\alpha t<0$ we compute the size of the largest cluster made by nodes participating in (n,t) subgraphs. To this end we renormalize the degree distribution taking into account that some links do not participate in any (n,t) subgraph. The effective degree distribution, incorporating only those links that participate in at least one (n,t) subgraph, has the form

$$P_{nt}(k)=\sum_{s=k}^{k_{\max}}P(s)\binom{s}{k}q_{nt}(s)^k\left[1-q_{nt}(s)\right]^{s-k}\ . \tag{45}$$

Assume that the graphs are random with the degree distribution given by $P_{nt}(k)$, we can compute the relative size of the giant component S_{nt}, made of (n,t) subgraphs by applying the generating function formalism [7]. We obtain that

$$S_{nt}=1-G_{nt}^{(0)}(u_{nt})\ , \tag{46}$$

where u_{nt} is the solution to the equation

$$u_{nt}=G_{nt}^{(1)}(u_{nt})\ , \tag{47}$$

and $G_{nt}^{(0)}(x)$ and $G_{nt}^{(1)}(x)$ are the generating functions of the node and first neighbor degree, respectively. More explicitly

$$G_{nt}^{(0)}(x)=\sum_{k=1}^{k_{\max}}P_{nt}(k)x^k\ , \tag{48}$$

$$G_{nt}^{(1)}(x)=\frac{\sum_{k=1}^{k_{\max}}kP_{nt}(k)x^{k-1}}{\sum_{k=1}^{k_{\max}}kP_{nt}(k)}\ . \tag{49}$$

Using (45) we obtain

$$G_{nt}^{(0)}(x)=\sum_{k=1}^{k_{\max}}P(k)[1-(1-x)q_{nt}(k)]^k\ , \tag{50}$$

$$G_{nt}^{(1)}(x) = \frac{\sum_{k=1}^{k_{\max}} kP(k)q_{nt}(k)[1-(1-x)q_{nt}(k)]^{k-1}}{\sum_{k=1}^{k_{\max}} kP(k)q_{nt}(k)} \ , \tag{51}$$

which provide the generating functions in terms of the original degree distribution $P(k)$. For each (n,t) we solve (47) by successive approximations and substitute the result into (46); the corresponding plots are shown in Fig. 3 in the paper.

---

[1] Yook, S. Y. Oltvai, Z. N. Barabási, A.-L. Functional and topological characterization of protein interaction networks. Proteomics (accepted).

[2] Shen-Orr, S. S. Milo, R. Mangan, S. Alon, U. Network Motifs in the transcription regulatory network of Escherichia coli. Nat. Genet. **31**, 64-68 (2002).

[3] Jeong, H. Mason, S. Barabási, A.-L. Oltvai, Z. N. Lethality and centrality in protein networks. Nature **411**, 41-42 (2001).

[4] Jeong, H. Tombor, B. Albert, R. Oltvai, Z. N. Barabási, A.-L. The large scale organization of metabolic networks. Nature **407**, 651-654 (2000).

[5] Watts, D. J. and Strogatz, S. H. Collective dynamics of small world networks Nature **393**, 440 (1998).

[6] Newman, M. E. J. The structure and function of complex networks. SIAM Review **45**, 167-256 (2003)

[7] M. E. J. Newman, S. H. Strogatz, D. J. Watts, Random graphs with arbitrary degree distributions and their applications Phys. Rev. E **64**, 026118 (2001).

[8] R. Milo, S. S. Shen-Orr, S. Itzkovitz, N. Kashtan, U. Alon, Network motifs: simple building blocks of complex networks, Science **298**, 824-827 (2002).