Information content: beyond modularity in complex networks

M. Zanin,1,2 P. A. Sousa,1 and E. Menasalvas3

1Facultade de Ciencias e Tecnologia, Departamento de Engenharia
Electrotécnica, Universidade Nova de Lisboa, Lisboa, Portugal
2Innaxis Foundation & Research Institute, José Ortega y Gasset 20, 28006, Madrid, Spain
3Center for Biomedical Technology, Universidad Politécnica de Madrid, 28223 Pozuelo de Alarcón, Madrid, Spain

We propose a novel measure to assess the presence of meso-scale structures in complex networks. This measure is based on the identification of regular patterns in the adjacency matrix of the network, and on the calculation of the quantity of information lost when pairs of nodes are iteratively merged. We show how this measure is equivalent to standard modularity, but also how it is able to detect the presence of other structures, such as bipartite and core-periphery configurations. Results corresponding to a large set of real networks are used to validate its ability to detect non-trivial topological patterns.

PACS: 89.75.-k, 05.45.Tp, 02.10.Ox, 87.18.Vf

In the last decade, complex network theory has unveiled several topological characteristics that are ubiquitous among many real-world systems. Initially the attention was directed towards two global, macro-scale network structures, i.e., small-world and scale-free topologies. But soon it was found that complex networks typically possess non-trivial patterns of connectivity at a meso-scale level, i.e., in between micro and macroscopical scales, which have been shown to have an important impact on, for instance, spreading and synchronization processes.

In spite of their importance, only one type of meso-scale structure has extensively been studied: communities, that is, the organization of nodes in clusters, with many links connecting nodes belonging to the same cluster and comparatively few joining nodes of different clusters. The pervasiveness of a community structure can, in principle, be characterized by quantifying the network modularity. Yet, this metric suffers from two main drawbacks: first of all, it is a posteriori metric, in that it can only be calculated after a community structure has been defined. Furthermore, modularity is not robust to the presence of different topological scales, e.g., when one community is much smaller than the others. While the concept of modularity can be generalized to include other meso-scale structures, as for instance bipartite networks, it still inherits the previously discussed drawbacks.

In this Letter, we address the following question: is it possible to define a single metric able to detect the presence of different kinds of meso-scale structures? We propose a novel metric, called Information Content, which is simultaneously (i) capable of detecting generic regularities in the adjacency matrix of a network, (ii) a priori metric, i.e., not requiring any previous computation like community detection, and (iii) robust to different topological scales.

The guiding hypothesis here is that important meso-scale structures are associated with regularities in the corresponding adjacency matrix. For instance, in the simplest case of a network with a perfect modular structure, nodes connect to all peers belonging to the same community: the resulting adjacency matrix is composed of four blocks, two containing only ones, two only zeros (see Eq. 5 below). In this case, erasing nodes within one community causes no loss of information, as their connections are equivalent; thus, measuring the information lost when pairs of nodes are merged can be used as a way of detecting such kind of regularities - and hence meso-scale structures.

Given an initial network, the proposed algorithm identifies the pair of nodes whose merging would suppose the smallest information loss, a quantity which is a function of the number of common links to / from other nodes shared by the pair. Once the best pair has been detected, both nodes are merged (thus yielding a network one node smaller), and the quantity of information I lost in the process is calculated. When this process is iteratively repeated, the Information Content IC of the network is defined as the sum of all Is, i.e., of all information contained in the network. The lower IC, the more regular is the link arrangement, indicating the presence of meso-scale structure.

As such, the calculation of the Information Content can be seen as a type of network renormalization procedure, characterized by two specific features. First of all, the objective is the estimation of the quantity of information lost in the process, while classical renormalization focuses on how some properties of the system are conserved at different scales. Furthermore, the renormalization transformation is guided by information theory criteria, instead of geometrical (topological) rules.

The calculation of the Information Content starts with a network of n nodes, which is fully defined by its adjacency matrix A, whose elements \( a_{ij} \) are equal to one when a link exists between nodes \( i \) and \( j \), and zero otherwise. The amount of information that would be lost if two nodes were merged together is first estimated for each pair of nodes \( k, l \) (with \( k \neq l \)). This is performed by comparing the connections departing from and arriving at both nodes, i.e. the vectors \( a_k, a_{-k}, a_l, \) and \( a_{-l} \), and...
by creating a new vector \( \mathbf{m} \) of size \( 2n \), representing the links that should be modified to recover the connections of node \( l \) given the connections of node \( k \), and thus the information lost when both nodes are merged together. In the first half of \( \mathbf{m} \), the \( i \)-th element (with \( i \in [1, n] \)) is defined as one if \( a_{kl} \neq a_{li} \), and zero otherwise, thus accounting for different outgoing links; the second half of \( \mathbf{m} \) accounts for different incoming links: thus \( m_{i+n} \) (again with \( i \in [1, n] \)) is set to one when \( a_{ik} \neq a_{il} \), and zero otherwise. In the two extreme situations, when two nodes either share all links or none, \( \mathbf{m} \) will either take all values 0 or 1 respectively.

Once the vector \( \mathbf{m} \) is constructed, the probability of finding an element equal to one (zero) is given by

\[
p_1 = \frac{1}{2n} \sum_{i=1}^{2n} m_i,
\]

\[
p_0 = 1 - p_1.
\]

Finally, the information contained in \( \mathbf{m} \) is assessed through the Shannon’s entropy \( I \):

\[
I_{kl} = 2n \left( -p_0 \log_2 p_0 - p_1 \log_2 p_1 \right).
\]

\( I_{kl} \) is defined in \([0, 2n]\), being \( I_{kl} = 0 \) when \( p_0 = 1 \) or \( p_1 = 1 \), meaning that all links are respectively equal or different, and \( I_{kl} = 2n \) when there is no correlation between the links of nodes \( k \) and \( l \).

Once \( I \) has been assessed for all possible pairs of nodes, the algorithm identifies the pair whose merging will suppose minimum information loss. Such pair is then merged by deleting one of its nodes, and the original network is transformed into a new one composed of \( n - 1 \) nodes (see Fig. 11 for an example). The whole process is then repeated iteratively, until one single node remains.

Each merging step supposes some loss of information (previously denoted by \( I_{kl} \)): the Information Content \( IC \) is given by the total amount of information lost as a result of the merging steps leading from the initial network to a single node. Conversely, it can be seen as the amount of information needed to reconstruct the full topology of the network, once it is reduced to a single node, by the merging process.

Two aspects of this metric should be clarified. Firstly, the information included in \( IC \) is not complete, as for instance at each step it would be necessary to track which pair of nodes has been merged: yet, the quantity of information required for this is constant, as does not depend on the topology of the network, and is thus discarded. Secondly, the Shannon entropy only provides a lower bound to the quantity of information required to encode vector \( \mathbf{m} \), which may be lower than what required in real applications.

For a network with a completely random structure, no correlation is expected on average between incoming and outgoing links of any pair of nodes: thus, merging pairs of nodes will result in a nearly maximal \( I \), and a maximal \( IC \) is expected. This can be used to normalize the Information Content of any network, such that

\[
IC_{norm} = IC / \langle IC_{random} \rangle,
\]

\( \langle IC_{random} \rangle \) being the average \( IC \) obtained for an ensemble of random networks with the same number of nodes and links of the original graph.

If \( \langle IC_{random} \rangle \) provides the upper bound of \( IC \), it is easy to find regular structures that will result in a very low Information Content. Clearly \( IC = 0 \) both for empty \((a_{ij} = 0, \forall i,j)\) and fully connected networks \((a_{ij} = 1, \forall i,j)\), as merging two nodes would suppose no information loss. More interestingly, the same will occurs with a fully modular network, such that

\[
A = \begin{bmatrix}
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1
\end{bmatrix}.
\]

The fact that all pairs of nodes have either the same or the opposite connections, thus either \( p_1 = 0 \) or \( p_1 = 1 \) and \( I_{kl} = 0 \) for any \( k \) and \( l \), and \( IC = IC_{norm} = 0 \), can be used to assess the modularity of a network: moving from a perfectly modular to a random structure, the \( IC_{norm} \) smoothly raises from zero to one. Contrary
to traditional community detection algorithms, $IC_{norm}$ is unaffected by the presence of multiple, widely separated, scales. Both ideas are demonstrated in Fig. 2, in which different rewiring probabilities are applied to an initial network of 400 nodes, comprising two communities of different sizes.

More generally, $IC$ can be used to assess the presence of any regular mesoscale structure. Consider for instance a bipartite network, i.e. networks where nodes belong to two groups, with nodes belonging to one of them being connected only to nodes of the other. The resulting adjacency matrix would thus have the following structure:

$$A = \begin{bmatrix}
0 & 0 & 1 & 1 \\
0 & 0 & 1 & 1 \\
1 & 1 & 0 & 0 \\
1 & 1 & 0 & 0 \\
\end{bmatrix}. \tag{6}$$

Similar results can also be obtained for networks showing a core-periphery structure, with a densely connected inner core, and a set of peripheral nodes sparsely connected with the core \cite{12}. In this case, merging nodes in the network core will result in low information loss, with a $IC_{norm}$ lower than expected for random graphs.

In summary, a low value of $IC_{norm}$ indicates the presence of some kind of meso-scale regularity, although it gives no information about the specific type of structure detected; in other words, one knows that a structure is present, but not if it is a modular structure, a bipartite one, etc. Thus it is natural to complement the information yielded by $IC$ with other common topological metrics. In order to stress this point, Fig. 3 Top presents a phenospace of 55 real networks, covering social, biological and technological systems \cite{13}. Each network is represented as a point in the plane, whose coordinates are given by their modularity (as calculated with the Blondel’s community detection algorithm \cite{11}) and $IC_{norm}$. If both metrics were similar, all points should be expected to lay on a line crossing the plane from the top-left to the bottom-right corner. On the contrary, it can be appreciated how they cover the whole plane, indicating that the information they provide is not equivalent.

Of special interest are biological networks, represented by black squares. They form a cluster in the lower part of the plane, indicating a low modularity; yet, their $IC_{norm}$ spans from 0.25 to 0.75, suggesting the presence of other types of meso-scale structures. As an example, Fig. 3 Bottom represents the network of the Lake Michigan food web. Such network has a low modularity (aprox 0.13), but also a low $IC_{norm}$ of 0.23: has can easily be seen, although no modular, it presents a strong core-periphery

![Figure 2: (Color online) Modularity vs. $IC_{norm}$. (Top) Modularity (as calculated with the Blondel’s community detection algorithm \cite{11}) for a network of 400 nodes organized in two communities. The different lines represent different sizes of the two communities: 1 : 1 (black line) two communities of 200 nodes, 1 : 2 (red line) 134 and 266 nodes respectively, and so forth. (Bottom) Normalized Information Content for the same networks.](image1)

![Figure 3: (Color online) Phenospace of 55 real networks. (Top) Each network is represented by a point, whose coordinates are given by their modularity and $IC_{norm}$. Colors encode the type of system represented by each network: black squares for biological systems, red circles for social, and blue triangles for other types of systems (mainly technological). (Bottom) Representation of a biological network - black square at (0.23, 0.13) in the top graph.](image2)
structure.

Information Content can also be used to assess the presence of different structures in weighted networks, by applying different thresholds and track how the $IC_{norm}$ evolves. As a test case, here we consider three brain functional networks [14], obtained through magneto-encephalographic (MEG) recordings of three healthy subjects performing a Sternberg’s letter-probe task. For each subject, a weighted clique of size $148 \times 148$ was computed using the MEG time series, where the weights are given by the correlation between each pair of sensors as calculated by means of a Synchronization Likelihood (SL) algorithm [15].

Fig. 4 reports the evolution of the modularity and of the normalized Information Content for the three subjects as a function of the applied threshold. While the former has a monotonous behavior (except for high thresholds, where the reduced amount of links results in strong fluctuations), the $IC_{norm}$ presents a clear maximum corresponding to a threshold of $0.2 - 0.25$. This region of reduced topological regularity points to a change in the structure of the networks, which is consistent with the varying fractal topology of the human brain at different synchronization thresholds [16].

In conclusion, this Letter reports on the definition of a new metric designed to assess the presence of regular meso-scale structures in complex networks. While other metrics, e.g. modularity, are defined a posteriori, that is the community structure should be detected before the calculation of the modularity of a network, Information Content can be obtained directly from the adjacency matrix. Furthermore, it is an exact metric, not requiring any optimization process whose result depends on the specific algorithm used. Finally, it enables the simultaneous assessment of different meso-scale structures, providing information complementary to standard measures. For all this, Information Content is expected to provide important benefits in tasks requiring the systematic and automatized analysis of large sets of networks, as in the case of classification tasks, for instance when a network representation is used to assess the health status of different patients [17].

---

[1] R. Albert and A.L. Barabási, Rev. Mod. Phys. 74, 47 (2002); S. Boccaletti, V. Latora, Y. Moreno, M. Chavez and D.-U. Hwang, Phys. Rep. 424, 175 (2006).

[2] J. A. Almendral, R. Criado, I. Leyva, J. M. Buldú and I. Sendiña-Nadal, Chaos 21 (1), 016101 (2011).

[3] S. Gil and D. H. Zanette, Phys. Lett. A 356 (2), 89 (2006); X. Wu and Z. Liu, Physica A 387 (2), 623 (2008).

[4] A. Arenas, A. Díaz-Guilera and C. J. Pérez-Vicente, Phys. Rev. Lett., 96 (11), 114102 (2006); A. Arenas, A. Díaz-Guilera, J. Kurths, Y. Moreno and C. Zhou, Phys. Rep. 469 (3), 93 (2008).

[5] M. E. Newman and M. Girvan, Phys. Rev. E 69 (2), 026113 (2004). G. Palla, I. Derényi, I. Farkas and T. Vicsek, Nature 435 (7043), 814 (2005). S. Fortunato, Phys. Rep. 486 (3), 75 (2010).

[6] M. E. Newman, Proc. Nat. Acad. Sci. USA 103 (23), 8577-8582 (2006).

[7] L. Danon and A. Díaz-Guilera and A. Arenas, Journal of Statistical Mechanics 2006 (11), P11010 (2006); S. Fortunato and M. Barthelemy, Proc. Nat. Acad. Sci. USA 104 (1), 36 (2007).

[8] R. Guimerà, M. Sales-Pardo and L. A. N. Amaral, Phys. Rev. E 76 (3), 036102 (2007).

[9] F. Radicchi, J. J. Ramasco, A. Barrat and S. Fortunato, Phys. Rev. Lett. 101 (14), 148701 (2008); H. D. Rozenfeld, C. Song and H. A. Makse, Phys. Rev. Lett. 104 (2), 025701 (2010).

[10] C. E. Shannon, Bell system technical journal 28 (4), 656-715 (1949).

[11] V. D. Blondel, J. L. Guillaume, R. Lambiotte and E. Lefebvre, Journal of Statistical Mechanics: Theory and Experiment 2008 (10), P10008 (2008).

[12] P. Holme, Phys. Rev. E 72 (4), 046111 (2005).

[13] B. Killworth and H. Bernard, Human Organization 35, 269-286 (1976); N. P. Hummon and P. Doreian, Social Networks 11, 39-63 (1989); S. Wasserman and K. Faust, Social Network Analysis: Methods and Applications, CUP (1994); V. Batagelj and A. Mrvar, Pajek data sets, http://pajek.imfm.si/doku.php?id=data:index (2003); S. Sun, L. Ling, N. Zhang, G. Li and R. Chen, Nucleic Acids Research 31 (9), 2443-2450 (2003); D. Lusseau, Proc. R. Soc. London B 270, S186-S188 (2003); C. J. Melia’n and J. Bascompte, Ecology 85 (2), 352-358 (2004); T. Opsahl, F. Agneessens and J. Skvoretz, Social Networks 33 (32), 245-251 (2010).

[14] E. Bullmore and O. Sporns, Nat. Rev. Neurosc. 10 (3), 186-198 (2009).

[15] C. J. Stam and B. W. van Dijk B W, Physica D 163, 236-251 (2002).

[16] D. S. Bassett, A. Meyer-Lindenberg, S. Achar, T. Duke and E. Bullmore, Proc. Nat. Acad. Sci. USA 103 (51), 19518-19523 (2006). L. K. Gallosa, H. A. Makse and M. Sigman, Proc. Nat. Acad. Sci. USA 109 (8), 2825-2830 (2012).

[17] M. Zanin and S. Boccaletti, Chaos 21 (3), 033103-033103 (2011). J. M. Buldú, R. Bajo, F. Maestú, N. Castellanos, I. Leyva, P. Gil, I. Sendiña-Nadal, J. A. Almendral, A. Nevado, F. del-Pozo and S. Boccaletti, PLoS One 6 (5), e19584 (2011). M. Zanin, P. Sousa, D. Papo, R. Bajo, J. García-Prieto, F. del Pozo, E. Menasalvas and S. Boccaletti, Sci. Rep. 2 (2012).
FIG. 4: (Color online) Modularity and $IC_{norm}$ in weighted functional brain networks. Evolution of the modularity (Left) and of the normalized Information Content (Right) for three human brain functional networks, as a function of the applied threshold. Dotted gray lines represent the corresponding link density (right axes).