Characterization of subgraph relationships and distribution in complex networks

Lucas Antiqueira and Luciano da Fontoura Costa

Instituto de Física de São Carlos, Universidade de São Paulo, Avenida Trabalhador São Carlense 400, Caixa Postal 369, CEP 13560-970, São Carlos, São Paulo, Brazil
E-mail: lantiq@ursa.if.sc.usp.br and luciano@if.sc.usp.br

New Journal of Physics 11 (2009) 013058 (34pp)
Received 8 September 2008
Published 30 January 2009
Online at http://www.njp.org/
doi:10.1088/1367-2630/11/1/013058

Abstract. A network can be analyzed at different topological scales, ranging from single nodes to motifs, communities, up to the complete structure. We propose a novel approach which extends from single nodes to the whole network level by considering non-overlapping subgraphs (i.e. connected components) and their interrelationships and distribution through the network. Though such subgraphs can be completely general, our methodology focuses on the cases in which the nodes of these subgraphs share some special feature, such as being critical for the proper operation of the network. The methodology of subgraph characterization involves two main aspects: (i) the generation of histograms of subgraph sizes and distances between subgraphs and (ii) a merging algorithm, developed to assess the relevance of nodes outside subgraphs by progressively merging subgraphs until the whole network is covered. The latter procedure complements the histograms by taking into account the nodes lying between subgraphs, as well as the relevance of these nodes to the overall subgraph interconnectivity. Experiments were carried out using four types of network models and five instances of real-world networks, in order to illustrate how subgraph characterization can help complementing complex network-based studies.

1 Author to whom any correspondence should be addressed.
1. Introduction

Because of their flexibility to represent, model and simulate virtually any discrete structure, complex networks [1]–[5] have been extensively studied and applied to the most diverse problems [6], ranging from transportation (e.g. flights [7]) to communications (e.g. Internet [8]). Complex networks are ‘complex’ because they exhibit particularly intricate and heterogeneous connectivity, e.g. by involving hubs [8, 9] or communities [10, 11]. As shown recently [12, 13], most real-world complex networks also include in their structure regular patches of connectivity, i.e. subgraphs whose nodes present similar topological measurements. All in all, the heterogeneity of complex networks tends to range along several topological scales, extending from the individual node level through mesoscopic structures such as modules and regular subgraphs, up to the whole network level. As a matter of fact, it is precisely the heterogeneous distribution of structural features along the several scales which defines the intricate organization and most interesting structural and dynamical properties of complex networks.

Although several works have investigated mesoscopic features of the connectivity of complex networks, e.g. by considering their respective communities [10, 11] and/or paths between different portions of the networks [14, 15], few works [13, 16] have focused on the study and characterization of the distribution of nodes and subgraphs within a given network. Such nodes and subgraphs of special interest arise in several situations, not only as communities or regular patches, but also with respect to extreme values of specific topological measurements. For instance, the nodes with degree higher than a given threshold can give rise to several subgraphs inside a network. Furthermore, motifs (e.g. cliques and loops) are intrinsically subgraphs and thus can also be analyzed regarding their distribution within a network. Another possibility for subgraph analysis derives from the fact that many real-world networks often have information assigned to their nodes or edges, such as topics of World Wide Web (WWW) pages and morphosyntactic categories of words in a linguistic network. Therefore, subgraphs can be obtained by selecting portions of a network that share some specific feature, e.g. pages.
about music in the WWW or verbs in a linguistic network. It should be observed that such
subgraphs do not necessarily yield a partition of the original network, as typically they do not
ever encompass all the original nodes. At the same time, it is assumed in this work that subgraphs
are connected components and do not overlap one another. Since the definition of connected
components allows the existence of trivial subgraphs (i.e. composed of single nodes [17]), the
proposed methodology is flexible enough to encompass subgraphs of any size.

Given a set of disjoint subgraphs of a network expressing specific properties of interest,
it becomes critically important to characterize how these subgraphs are distributed through the
network, as such information can be particularly important regarding the overall organization
of the network and its dynamics. Going back to the above example with node degree, if highly
connected subgraphs are found to be close to one another (in terms of shortest path length
between them), the portion of the original network containing such subgraphs can be understood
as corresponding to a critical bottleneck for the whole system under analysis. On the other hand,
a more uniform distribution of subgraphs with highest degree suggests a system less critically
structured for communications. Several similar situations can be characterized with respect to
other types of subgraphs, including other measurements as well as communities and regular
patches. Yet, few works have addressed the specific issue of how critical nodes or subgraphs are
distributed through the network topology.

The objective of the present work is to develop and apply a comprehensive framework
for characterization of the distribution of subgraphs of specific interest within a given complex
network. In order to do so, we resource to the distances, quantified in terms of the shortest path
lengths, between each pair of given subgraphs. Such distances are organized into a histogram,
which can provide valuable information about the topological distribution of the subgraphs.
For instance, a sharp peak in such a histogram at a small value of distance will indicate that
the subgraphs are all close to one another. Moreover, subgraph sizes are also organized into a
histogram, thus allowing the analysis of another basic subgraph information. Though subgraph
distances and sizes provide valuable information about the overall subgraph distribution and
topology, it is also interesting to have the means for progressively merging subgraphs in
order to obtain connected components surrounding the critical regions. Such progressively
increasing components can then be analyzed regarding their structure or function. Moreover,
this progressive merging allows the analysis of nodes surrounding subgraphs according to
their relevance to subgraph interconnectivity. Therefore, we report an algorithm based on the
morphological operation called *dilation*, which allows the progressive merging of the subgraphs
in terms of successive distance values, up to the point of containing all the given subgraphs.
Figure 1 depicts a network with eight subgraphs and some bidirectional arrows that denote the
distance and merging relationships we want to characterize. The potential of the aforementioned
histograms and subgraph merging algorithm is illustrated with respect to both theoretical as well
as real-world complex networks, including the Barabási–Albert model [9] as well as the power
grid of the western states of the USA [18]. Each of these networks was analyzed according
to subgraphs generated from cliques (complete subgraphs) and from extreme values of the
measurements degree and clustering coefficient.

This paper starts by presenting the basic concepts and methods (section 2) and proceeds
by describing the histograms and the merging algorithm (section 3), which are then illustrated
with respect to theoretical and real-world networks (section 4).
2. Basic concepts and methods

A network can be represented by a graph $G(V, E)$, where $V = \{v_1, v_2, \ldots, v_N\}$ is its set of $N$ vertices (or nodes) and $E = \{e_1, e_2, \ldots, e_L\}$ is its set of $L$ edges (or links). An edge $e_k$ is a pair $(v_i, v_j)$ that represents a connection between nodes $v_i \in V$ and $v_j \in V$. The set of edges $E$ can be encoded into an adjacency matrix $A$, of dimension $N \times N$, with elements $A(i, j) = 1$ whenever there is an edge from node $i$ to node $j$, with $A(i, j) = 0$ being imposed otherwise. Notice that $A$ is symmetric for undirected graphs.

In what follows, we present the definitions of the adopted measurements for undirected graphs, as well as the details of the considered artificial and real-world networks.

2.1. Network measurements

The network measurements reviewed in this section have been frequently employed in the field of complex networks. For more details, please refer to the review paper [5].

2.1.1. Degree. The node degree $k(i)$ corresponds to the number of edges attached to a node $i$. Using the adjacency matrix $A$, the degree can be obtained by

$$k(i) = \sum_{j=1}^{N} A(i, j). \quad (2.1)$$
2.1.2. Clustering coefficient. This measurement reflects the density of connections between the neighbors of a node \( i \). Let 
\[
\eta(i) = \{ j \mid A(i, j) = 1, i \neq j \}
\]
be the set of neighbors of \( i \), and 
\[
\epsilon(i) = \sum_{u, v \in \eta(i), u \neq v} A(u, v)
\]
be the number of edges between the neighbors of \( i \). The clustering coefficient of a node \( i \) is defined as 
\[
c(i) = \frac{2\epsilon(i)}{|\eta(i)|(|\eta(i)| - 1)},
\]
(2.2)
where \(|\eta(i)|\) is the cardinality of \( \eta(i) \), i.e. it is the number of neighbors of \( i \).

2.1.3. Length of shortest paths. The proximity between nodes is usually quantified in terms of shortest paths. A path \( p(i, j) \) extending from node \( i \) to node \( j \) is denoted by a sequence of neighboring nodes:
\[
p(i, j) = (v_1, v_2, \ldots, v_l, v_{l+1}),
\]
where \( A(v_l, v_{l+1}) = 1, v_1 = i, v_{l+1} = j \) and the length of the path is \( \omega(p(i, j)) = l \). Note that the length of a path is the number of edges along it. The length of the shortest path between two nodes \( i \) and \( j \) is thus given by
\[
s(i, j) = \min\{\omega(p(i, j))\},
\]
(2.3)
which is the minimum number of steps (edges) needed to reach node \( j \) after starting at node \( i \) (or vice versa in the case of undirected networks).

2.1.4. Clique. A subgraph \( g \) of a graph \( G \) is called a clique if every pair of distinct nodes of \( g \) is linked by an edge, i.e. if \( g \) is a complete subgraph [17]. The size of a clique is equal to its number of nodes, and the clique number \( \omega(G) \) of a graph \( G \) is the size of the largest clique it contains.

2.2. Network models

Four theoretical network models were chosen in this work in order to construct undirected networks. For each model, 100 realizations (networks) were performed with \( N = 1000 \) nodes and mean degree \( \langle k \rangle = 6 \). The general characteristics of these models are given below.

2.2.1. Erdős–Rényi (ER). In this model, every possible pair of nodes \( (i, j) \) is connected with uniform probability \( p \) [19]. For an ER network, the mean node degree is given by \( \langle k \rangle = p(N - 1) \) in the large network limit \( N \to \infty \). Moreover, this model yields random networks with a Poisson degree distribution, which implies a characteristic mean degree, i.e. the node degrees do not greatly deviate from \( \langle k \rangle \).

2.2.2. Watts–Strogatz (WS). The WS model generates networks exhibiting the small-world property, i.e. high average clustering coefficient and low average shortest paths [18]. In order
Table 1. Basic information about the real networks used in our experiments. For each network, we show its number of nodes $N$, number of edges $L$ and respective mean degree $\langle k \rangle$.

| Network       | $N$  | $L$  | $\langle k \rangle$ |
|---------------|------|------|----------------------|
| NetScience    | 379  | 914  | 4.82                 |
| *E. coli*     | 763  | 2412 | 6.32                 |
| Email         | 1133 | 5451 | 9.62                 |
| Power grid    | 4941 | 6594 | 2.67                 |
| Internet-AS   | 22963| 48436| 4.22                 |

to obtain a WS network, we start with a regular ring-shaped network with $N$ nodes, where every node is connected to its $\kappa$-nearest neighbors in both directions. Then, each edge is moved (rewired) to another position with probability $p$. Depending on $p$, an ER network can approach the features of the initial regular network (for $p \to 0$) or of a random network (for $p \to 1$). In our experiments, we employed $p = 0.2$. Also notice that the mean node degree of a WS network is $\langle k \rangle = 2\kappa$.

2.2.3. Barabási–Albert (BA). Networks with a power-law degree distribution can be obtained by considering the BA model [9]. This type of network contains a few nodes, called hubs, concentrating many connections, while the majority of nodes have only a few links. A BA network is generated by adding new nodes to an initial network of $m_0$ nodes. Each newly added node is connected to $m$ previous nodes, with the probability of connections being proportional to the respective degrees. The average node degree of a BA network is $\langle k \rangle = 2m$.

2.2.4. Geographical (GG). In contrast to the ER, WS and BA models, a geographical model considers the spatial position of nodes to create edges [5], so that the spatial adjacency between nodes often strongly influences the respective connectivity. In the geographical model adopted here (called GG), randomly placed nodes are distributed through a bi-dimensional grid of size $L \times L$, and edges are established among nodes geographically close to each other, i.e. separated by a distance not greater than $R$. Thus, long-range connections are not created, implying longer paths than in the previous models presented in this section. The mean degree of a GG network can be estimated as $\langle k \rangle \approx \pi R^2 N / L^2$.

2.3. Real-world networks

A set of real networks, as described below, has also been used in our experiments. These networks have been chosen so as to provide a representative sample of several types of real-world networks of general interest. Table 1 provides a quick reference with basic information about each network.

2.3.1. Coauthorship in network science. This network, called NetScience, expresses the coauthorship relationships between scientists that published papers in the field of complex...
networks [20]. It was compiled by M E J Newman in May 2006\(^2\) using the references cited in two surveys of the field [3, 4], plus some manually added references. Each scientist is a node in this network, while an undirected edge is created between two scientists whenever they have published at least one joint paper. NetScience has 1589 nodes, of which 379 are inside the largest connected component, which is the part we used in our experiments. Henceforth, whenever we mention NetScience, we refer to its largest connected component. Moreover, we do not take into consideration the original weights of this network.

2.3.2. Escherichia coli metabolism. The set of biochemical reactions that constitute the metabolism of the bacterium *E. coli* is considered in this case. The substrates are represented as nodes, and undirected links are placed between all incoming substrates (i.e. the educts of a reaction) and respective outgoing substrates (i.e. the products of the same reaction) [21]. The network we have used is based on a similar network that includes the direction of reactions (from educts to products) as well as temporary educt–educt complexes and enzymes as nodes [22]. Since the former network is not publicly available, we have downloaded the latter\(^3\) and removed temporary complexes and enzymes, beyond creating the corresponding undirected links between incoming and outgoing substrates. This procedure resulted in a network with \(N = 763\) nodes, very similar to the one used in [21].

2.3.3. Email communications. We also considered a graph reflecting the flow of email messages exchanged among the members of the University at Rovira i Virgili (Spain) [23]. This network, compiled in the research group of A Arenas\(^4\), has a single connected component, where each email address is identified by a node (there are \(N = 1133\) addresses), and a message sent from node \(i\) to node \(j\) is represented by an undirected unweighted edge \((i, j)\). The authors removed bulk emails, i.e. messages sent to more than 50 addresses, before defining the edges in this network.

2.3.4. Power grid. This network represents the topology of the power grid of the western states of the USA [18], and was compiled by D Watts and S Strogatz\(^5\). A power grid is the structure that underlies the transmission of electricity from power plants to consumers. The power grid of the USA western states is a single connected component with 4941 nodes interconnected by undirected and unweighted links.

2.3.5. Internet-AS. The connections that associate ASs (Autonomous Systems) in the Internet were considered by M E J Newman in the compilation of this network\(^6\). An AS is a group of computer networks that share the same routing policy and have a centralized administration. Using BGP (Border Gateway Protocol) data of 22 July 2006, Newman reconstructed the links

---

\(^2\) The coauthorship network is available at the website of M E J Newman: http://www-personal.umich.edu/~mejn/netdata.

\(^3\) The *E. coli* network is available at the website of the Center for Complex Network Research (CCNR): http://www.nd.edu/~networks/resources.htm.

\(^4\) The Email network is available at the website of A Arenas: http://deim.urv.cat/~aarenas/data/welcome.htm.

\(^5\) The power grid network is available at the website of D Watts: http://cdg.columbia.edu/cdg/datasets.

\(^6\) The Internet snapshot is available at the website of M E J Newman: http://www-personal.umich.edu/~mejn/netdata.
between 22,963 ASs, which yielded a connected graph with unweighted and undirected edges. Due to the nature of BGP, which is a gateway protocol used to route data packets between ASs, it was possible to retrieve information about the physical links of the Internet at the AS level.

3. Characterization of subgraphs

The method for analyzing subgraph interconnectivity introduced in this section assumes that the graph/network $G(V, E)$ under study is undirected, unweighted and connected. It also requires that $C$ subgraphs $G_i = (V_i, E_i), 1 \leq i \leq C$, be defined such that:

(i) $V_i \subset V$,
(ii) $V_i \neq \emptyset$,
(iii) $V_i \cap V_j = \emptyset$ for every $i \neq j$,
(iv) $(v_i, v_j) \in E_i$ if and only if $v_i \in V_i, v_j \in V_i$ and $(v_i, v_j) \in E$,
(v) $G_i$ is a connected component,
(vi) two different subgraphs $G_i, G_j$ are not direct neighbors.

In order to create a subgraph $G_i$, it is enough to define a valid $V_i$, since $E_i$ must contain all the edges of $E$ that connect pairs of nodes included in $V_i$ (rule (iv)). We refer to a ‘valid’ $V_i$ as a non-empty subset of $V$ (rules (i) and (ii)) that results in a connected component (rule (v)) not sharing any nodes or edges with other subgraphs, i.e. each connected component must be maximal (rules (iii) and (vi)). It is worth pointing out that these rules allow the existence of trivial subgraphs $G_i$ (i.e. with only one node), thus being consistent with the definition of subgraphs in graph theory [17]. Figure 2 shows an example of an arbitrary graph $G$ with $N = 64$ and four subgraphs $G_1, \ldots, G_4$.

The selection of nodes for subgraphs $G_i$ depends on the specific study being carried out, given that the above conditions are followed. For instance, nodes with extreme values of a given topological measurement may be selected to form subgraphs, such as nodes with high degree, high clustering coefficient or any other meaningful measurement. In this case, subgraphs are defined as connected components encompassing all the selected nodes. Another option would be to directly use specific subgraph patterns, such as motifs. In general, any property of nodes or edges can be used to define subgraphs, as long as each selected subgraph be a maximal connected component. In section 4, we give detailed examples of some approaches to define subgraphs, although alternative approaches are also possible.

3.1. Histograms of subgraph sizes and distances

Since subgraphs are the main elements of the methodology described in this section, it is important to understand their structure before analyzing their interrelationships. In our case, we want to know the size of each subgraph $G_i$, i.e. the number of nodes $N_i = |V_i|$ of all $C$ subgraphs. These sizes can be concisely organized into a histogram, thus giving a complete picture of the $N_i$ distribution. Depending on the case, especially if the histogram is normalized or the number of subgraphs $C$ is large, it would be useful to provide $C$ alongside the histogram. Examples of this histogram of subgraph sizes are given in section 4.

One way to analyze subgraph distribution is by computing the distance between every pair of subgraphs. We define this distance to correspond to the length of the shortest path between
Figure 2. Graph $G = (V, E)$ with $N = 64$ nodes, where each node is identified by a number $v$. This graph has four subgraphs $G_1$, $G_2$, $G_3$ and $G_4$, whose set of nodes are $V_1 = \{6, 7, 8, 15, 16, 23\}$, $V_2 = \{18, 19, 20, 27, 28, 29, 37\}$, $V_3 = \{26, 33, 34\}$ and $V_4 = \{48, 55, 56, 64\}$. These subgraphs correspond to the four connected components with black nodes.

subgraphs and, therefore, (2.3) can be used for this purpose. Note that there are at least $|V_i||V_j|$ different paths between two different subgraphs $G_i$ and $G_j$, because each path may start at any node of the source subgraph and end at any node of the destination subgraph. The length of the shortest path between two different subgraphs can now be defined as

$$s(G_i, G_j) = \min_{w_i \in V_i, w_j \in V_j} \{s(w_i, w_j)\} - 1,$$

imposing that $s(G_i, G_j) = 0$ whenever $i = j$. Note that the length of the shortest path is decremented by one, which changes the length of the distance from an edge orientation to a node orientation. This modification has been done because there must be at least one node, or two edges, between a pair of subgraphs (i.e. the distance would start at two). Thus, a node-oriented distance is preferred because it is more intuitive. We therefore define the matrix $D_s$ of order $C \times C$ with elements $D_s(i, j) = s(G_i, G_j)$, i.e. it encodes every distance between all $C$ subgraphs of $G$. As an example, the matrix $D_s$ for the four subgraphs in figure 2 is given as

$$D_s = \begin{bmatrix} 0 & 1 & 5 & 4 \\ 1 & 0 & 2 & 2 \\ 5 & 2 & 0 & 5 \\ 4 & 2 & 5 & 0 \end{bmatrix}.$$
where $D_s$ is symmetric because $G$ is an undirected graph. If these distances (excluding the diagonal of $D_s$) are placed in a histogram, the overall proximity between subgraphs can be examined more easily than just observing $D_s$, as will become clearer in the experiments reported in section 4.

3.2. Subgraph merging

The procedure detailed in this subsection aims at gradually merging subgraphs $G_i$ inside the original graph $G$, while giving special attention to the relationships between them. We implement this progressive merging, or expansion, in terms of the gradual growth of the subgraphs $G_i$ toward graph $G$, which is accomplished by adding to the subgraphs $G_i$ nodes of $G$ that do not belong to any $G_i$ yet (and also adding the necessary edges, as specified in the definition of the subgraphs $G_i$ presented earlier in this paper). The progressive merging is based on the relationships between all subgraphs and the remaining nodes of the network. Therefore, besides providing means to sequentially join subgraphs, which can be useful for multilevel graph analysis, the merging algorithm also allows the characterization of subgraphs by assessing the relevance of each network node regarding subgraph interconnectivity.

In order to achieve a gradual subgraph merging, some vertices of $G$ need to be included in the expansion earlier than others. In our methodology, higher relevance is given to the nodes inside a short path connecting some pair of different subgraphs $(G_i, G_j)$. In this manner, the merging is controlled by the length of paths between subgraphs. More specifically, for a node $v$ outside every subgraph $G_i$, we compute the length of the minimum path between all pairs of different subgraphs $(G_i, G_j)$ that necessarily pass through node $v$. This is understood as the relevance of a node $v$ in the merging of subgraphs. In other words, a node that is close to only one subgraph is considered a member of a weak tie, because it does not take part in short paths connecting that subgraph with others.

The aforementioned merging can be carried out by applying consecutive dilations [12, 24] in the subgraphs of $G$. The dilation is a morphological operation $\delta(g)$, defined over a subgraph $g$ of $G$, that yields another subgraph that is equal to the union of the original subgraph and its neighbors in $G$ (plus the respective edges). Figure 3 illustrates the dilation of the subgraph $G_1$ of figure 2, which is formed by vertices $V_1 = \{6, 7, 8, 15, 16, 23\}$. The dilation $\delta(G_1)$ results in a subgraph whose nodes are $V_1 \cup \{14, 22, 24\}$ (i.e. nodes $\{14, 22, 24\}$ are neighbors of $G_1$), along with the respective edges that connect these nodes in $G$.

Dilations are employed as an intermediate step in our method. When $\delta(G_i)$ is applied sequentially and recursively inside $G$, until no more dilations are possible, a distance map is traced between the subgraph $G_i$ and the other nodes of $G$. This recursive dilation is denoted by

$$\delta_d(G_i) = \delta \left( \ldots \left( \delta(G_i) \right) \right), \quad (3.2)$$

and the nodes that are included in $\delta_d(G_i)$, but not in $\delta_{d-1}(G_i)$, are said to be at distance $d$, in number of edges, from $G_i$. For $d = 0$, the recursive dilation is defined as $\delta_0(G_i) = G_i$, and because the dilation $\delta_{-1}(G_i)$ is not possible, the nodes inside $G_i$ are naturally defined to be at a distance 0 from $G_i$. It is worth pointing out that these distances are different from those given in the previous section (from matrix $D_s$), which are only calculated between subgraphs, not between a subgraph and every node of the network.
Figure 3. Dilation $\delta(G_1)$ of the subgraph $G_1$ of figure 2. $G_1$ has nodes $V_1 = \{6, 7, 8, 15, 16, 23\}$, and the dilation is formed by nodes $V_1 \cup \{14, 22, 24\}$, represented in black in the figure.

Since we are going to dilate all $C$ subgraphs, it is necessary to apply the dilation $\delta_d(G_i)$ without considering the nodes of other subgraphs $G_j$, $i \neq j$. This particular behavior is required because the merging is made outwards the set of subgraphs, and thus it is not necessary to consider nodes that already belong to a subgraph. Therefore, when dilating a subgraph $i$, some other subgraph $j$ may block the accessibility of $i$ to some nodes in the graph. For example, in figure 2, subgraph $G_2$ cannot communicate with nodes 25 and 41 because subgraph $G_3$ is blocking its access to these nodes. In fact, in this example, only subgraph $G_3$ is able to communicate with nodes 25 and 41, or, conversely, these nodes can only access subgraph $G_3$.

In what follows, we define the set of subgraphs accessible from a node $v$ as

$$Q(v) = \{G_i \mid v \in \delta_d(G_i), \quad 0 \leq d < \infty\},$$

(3.3)

and the total number of subgraphs accessible from a node $v$ is

$$q(v) = |Q(v)|,$$

(3.4)

where $|Q(v)|$ is the cardinality of $Q(v)$. Note that $q(v) \geq 1$ for any $v$, due to the fact that $G$ is a connected graph.

The complete set of dilations $\delta_d(G_i)$, which takes into account every subgraph $G_i$ and every possible dilation starting from $d = 1$ (until there are no more nodes to be added by the dilation), allows the definition of a distance matrix $D_d$, of order $N \times C$. An element $D_d(v, i)$ of $D_d$ indicates the distance between node $v$ and subgraph $G_i$. As observed before, this distance
is given by the dilation in number of edges. There is an exception requiring special treatment: for a node $v$ not accessible from a subgraph $G_i$, i.e. $G_i \not\ni Q(v)$, we define $D_{\delta}(v, i) = d_{\text{max}} + 1$, where $d_{\text{max}} = N - 1$ is the maximum possible distance between two nodes in a graph with $N$ nodes.

The definition of the matrix $D_{\delta}$ is the last step before specifying a relevance value for each vertex. The shortest path between any two subgraphs $G_i$ and $G_j$, $i \neq j$, that necessarily pass through node $v$, is then defined as the relevance $r(v)$ of node $v$. Consequently, the lower $r(v)$, the higher is the relevance of node $v$ in the merging of the subgraphs. More formally, $r(v)$ is given by

$$r(v) = \begin{cases} \min_{1 \leq i, j \leq C, i \neq j} \{D_{\delta}(v, i) + D_{\delta}(v, j)\} - 1 & \text{if } q(v) > 1, \\ \min_{1 \leq i \leq C} \{D_{\delta}(v, i)\} & \text{if } q(v) = 1, \end{cases} \quad (3.5)$$

where the second case is an exception that occurs when node $v$ can access only one subgraph (or it is inside a subgraph), thus $r(v)$ is equal to $D_{\delta}(v, i)$, where $i$ is the index of the only subgraph to which $v$ is connected. In this special case, $r(v)$ is solely controlled by the consecutive dilations of a single subgraph. Observe also that, in the first case, the minimum expression is decreased by one. This decrement was included because otherwise $r(v)$ would always be greater than one for nodes with $q(v) > 1$, odd behavior for a quantity that starts at zero. In other words, the relevance was changed from an edge-oriented to a vertex-oriented one, similarly to what was done with matrix $D_{\delta}$ in the computation of the distance histogram (section 3.1). Therefore, (3.5) allows relevance values greater or equal than one for every node outside subgraphs $G_i$, whereas zero relevance is reserved for the nodes inside some subgraph.

Figure 4 shows $r(v)$ for every node of the graph illustrated in figure 2. In this example, $r(v)$ is shown both numerically and graphically. The latter approach uses a gray scale proportional to $r(v)$, ranging from black (when $r(v) = 0$, which refers to the more relevant nodes) to white (when $r(v) = 8$, representing the less relevant nodes in this example). Note that the darkest nodes are placed in the shorter paths that connect subgraphs $G_1, \ldots, G_4$. The nodes of these subgraphs correspond to those with $r(v) = 0$.

Finally, the merging of the subgraphs of $G$ is performed by thresholding relevance values as follows:

$$V^+ = \{v \mid r(v) \leq T\},$$

where $T \geq 0$ is an integer threshold. In addition, $C^+$ subgraphs $G_i^+ = (V_i^+, E_i^+), 1 \leq i \leq C^+$, are created such that:

(i) $V_i^+ \cup V_j^+ \cup \ldots \cup V_{C^+}^+ = V^+$,
(ii) $V_i^+ \cap V_j^+ = \emptyset$ for every $i \neq j$,
(iii) $(v_i, v_j) \in E_i^+$ if and only if $v_i \in V_i^+, v_j \in V_i^+$ and $(v_i, v_j) \in E$,
(iv) $G_i^+$ is a connected component,
(v) two different subgraphs $G_i^+, G_j^+$ are not direct neighbors.

These rules are similar to those given in the definition of the original subgraphs $G_i$, with the difference that the merged subgraphs are restricted to the vertices belonging to the thresholded set $V^+$. To summarize the process of merging, it suffices to take into account that the new subgraphs are the connected components that remain when nodes $v \not\in V^+$ (and their edges) are excluded from $G$. 

New Journal of Physics 11 (2009) 013058 (http://www.njp.org/)
Thresholding with $T = 0$ gives the original subgraphs, as $r(v)$ is equal to zero if and only if $v$ belongs to some subgraph $G_i$. Since greater thresholds include other nodes, two or more subgraphs can then be joined into one single connected component. An example of a merging for a threshold $T = 2$ is given in figure 5, where the input for the merging is the graph $G$ and its subgraphs depicted in figure 2. Figure 5 shows in black the nodes that belong to the merging, which form a single connected component joining all four subgraphs of $G$.

It is worth pointing out that recursive dilations could be used as a method of subgraph merging (and not only as an intermediate step), where each subgraph would be simultaneously dilated until the entire network was covered. Nevertheless, figure 3 shows that the first dilation of the subgraph $G_1$ includes nodes 14 and 24, whereas our merging does not include any of them until $T = 3$ (see figure 5). Although nodes 14 and 24 are neighbors of $G_1$, they are farther from other subgraphs, and thus they do not participate in short paths linking a pair of subgraphs. This comparison shows that dilations do not discriminate the relative position of nodes between subgraphs, and this is the reason why we have defined the matrix of distances $D_δ$ and the relevance values $r(v)$.

If the merging is computed for sequentially increasing thresholds starting at $T = 0$, $V^+$ grows until $V^+ = V$, i.e. the subgraphs $G_i$ expand until they form a single connected subgraph that is equal to $G$. We refer to this procedure as gradual merging. The number of subgraphs (or connected components) $C^+$ in the merging can be monitored until the end of the sequential thresholding, when $C^+$ must be equal to one. In this case, $C^+$ is a monotonically non-increasing function of $T$. 

Figure 4. Values of $r(v)$ for each node in the graph $G$ of figure 2. The number next to each node denotes $r(v)$ (please refer to figure 2 for the number $v$ of each node). The color of each node is derived from a gray scale ranging from black ($r(v) = 0$) to white ($r(v) = 8$).
In a gradual merging, it is possible to quantify the overall proximity between the original subgraphs by observing when $C^+$ drops to one. This approach complements the distance histograms explained in section 3.1 by looking at a fraction of the possible minimum paths existing between all subgraphs, i.e. only the paths necessary to merge subgraphs are taken into account. Furthermore, the distance histogram does not give any information about nodes lying between subgraphs. Monitoring the growth of the set $V^+$ for sequentially increasing thresholds is important to identify which nodes are required to join subgraphs into a single connected component, which can then be thought as a core group of nodes regarding subgraph interconnectivity. In a more general sense, the overall relevance $r(v)$ of nodes can be analyzed by monitoring all variations in the set $V^+$ for increasing $T$. In order to explain in a more detailed fashion the use of the gradual merging and also to illustrate its potential for complementing the distance histogram, we give in the next section examples of subgraph characterization for both real-world and artificial networks.

4. Experimental results and discussion

We now illustrate the application of subgraph characterization on a set of artificial and real-world networks. As already mentioned in section 2.2, 100 realizations of models ER, WS, BA and GG were obtained, each one with $N = 1000$ nodes and mean degree $\langle k \rangle = 6$. The real
networks were introduced in section 2.3, namely NetScience, *E. coli*, Email, Power Grid and Internet-AS, with *N* ranging from 379 to 22,963 and \(\langle k \rangle\) between 2.67 and 9.62 (please refer to table 1 for more details about real networks). The chosen networks cover a considerable range of types of structures usually studied in the field of complex networks [1]–[5], therefore providing a representative basis for the illustration of our method.

An important step in the experimental setup is the definition of the subgraphs to be analyzed. To perform this task, we have independently used three network concepts: the measurements degree \(k(i)\), clustering coefficient \(c(i)\) and the complete subgraph called clique, all explained in section 2.1. For the first two cases, the nodes with the highest \(k(i)\) or \(c(i)\) were chosen to form subgraphs \(G_i\) according to the definition presented in section 3. Thus, subgraphs \(G_i\) correspond to the connected components (possibly containing only one node) existing between the nodes with highest degree (or highest clustering coefficient), limited to 2.5% of the total number of nodes. The use of degree, already mentioned in the introductory section of this paper, is particularly important in which concerns proximity between groups of critical nodes. In other words, if these subgraphs are close to each other, the network may become particularly sensitive to attacks directed to highly connected nodes. Subgraphs with high clustering coefficient are also interesting to analyze because they tend to be more cohesive than others, i.e. showing a variety of different paths between the nodes. The characterization of the connectivity between dense subgraphs may lead to the improvement of search and transport strategies and also of network designs. It is important to observe that the analysis of the overall distribution of the critical subgraphs through the network can provide complementary information to the already important insights provided by those measurements at the individual node level, as done traditionally.

The third class of subgraphs to be analyzed in this section is formed by nodes taking part in cliques of at least size 6. Thus, every node contained in these cliques is selected, and connected components among the selected nodes are defined as the subgraphs to be studied, since two or more cliques may form one connected component when combined. Size 6 was chosen in order to select cliques as large as possible for the majority of networks considered here. Nevertheless, some networks have maximal cliques of size lower than 6: for ER and BA, the clique number is \(\omega(G) = 3\), and for WS, \(\omega(G) = 4\). Therefore, for these exceptions, we have selected only the cliques of maximal size to be analyzed. Cliques are portions of a network that implement interconnectivity at the maximum possible level, i.e. all nodes connect to each other. Therefore, being such an extreme type of motif, cliques are useful to indicate the level of robustness of a network, similarly to what may be done when analyzing subgraphs of high degree or high clustering coefficient (as suggested in the last paragraph). For instance, cliques may be thought as centers of activity in transport, search, surfing or any other dynamical process, as they concentrate connections at the maximum level. Congestion can be avoided if traffic is shared inside these centers of activity, thus reducing the load on each node/edge of their neighborhood. Nevertheless, if cliques are attacked and entirely broken, great portions of the network may stop working. Therefore, they can be an important bottleneck if tightly coupled to the entire network, as explained in the following experiments.

In the following subsections, we report the results obtained for the aforementioned artificial and real-world networks regarding size and distance histograms and subgraph merging. For the merging procedure, we monitor the number of subgraphs \(C^+\) and the number of nodes \(|V^+|\) as threshold \(T\) increases. Since 100 realizations of each network model have been performed, the following results are presented in terms of average measurements and respective standard
deviations. Note that for the real networks this procedure was not necessary because only one
network was available for each case.

4.1. Network models

4.1.1. Subgraphs with high degree. Figure 6 shows the average size and distance histograms
for the subgraphs consisting of nodes with the highest degree in ER, WS, BA and GG networks.
The considered ER networks have 17.48 subgraphs on average, most of them of size 1. Larger
subgraphs, with at most five nodes, occur less frequently. Moreover, the distance distribution
for highly connected subgraphs in ER shows a well-defined peak at distance 2, whereas other
possible distances do not highly deviate from this value. The same behavior, i.e. small subgraphs
close to each other, was found in WS networks, with a slightly higher probability to find trivial
subgraphs (with only one node), and a distance peak at 3–4, instead of 2. At the opposite
side is the BA model, which tends to group selected hubs into a single subgraph of size
25. Therefore, distance values are more frequently equal to 0 in this model. An intermediary
situation was found in the GG model, which shows 8.11 subgraphs on average comprising
at most 16 nodes each one, with smaller subgraphs being more frequent. Another distinctive
feature of the investigated geographical networks is the high heterogeneity of their subgraph
distances, with an average distance histogram ranging from 1 to 45.

Before discussing these results, we turn our attention to the gradual merging of subgraphs,
in order to complement the observations made in the last paragraph. The plots in the left column
of figure 6 show, for each network model, the number of subgraphs $C^+$ as a function of the
merging threshold $T$, while the right column shows the number of nodes $|V^+|$ as a function
of $T$. By inspecting variations in $C^+$, one could find that the ER and WS models have their
subgraphs joined into one connected component when $T = 1$ and 2, respectively. Thus, only the
lower distances in the histograms of figure 6 are taken into account in the gradual merging
of ER and WS at this point, since relevance values of the merging algorithm are directly
related to shortest paths. This algorithm allows us to identify to what extent the distances in
the histogram contribute to efficient communication between subgraphs (in terms of creating a
single connected component). Furthermore, the merging procedure allows us to identify which
nodes are necessary to bring subgraphs together. In the case of ER model, about 50 nodes
are included in the merging when $C^+ = 1$. Since 25 nodes are inside subgraphs, the remaining
nodes (also 25) are those implementing subgraph interconnectivity at the minimum possible
cost in terms of path lengths. We call these nodes a core group of nodes regarding subgraph
interrelationship. In the case of WS networks, nearly 75 nodes form the core group, since
approximately 100 nodes are covered by the merging procedure when $C^+ = 1$. Therefore, ER
and WS networks have subgraphs of high degree quickly merged (i.e. with low $T$) using a small
fraction of network nodes.

An analysis of nodes beyond the core group is also worthwhile. For instance, ER and
WS networks are entirely covered by the gradual merging when $T = 5$ and 7, respectively.
These threshold values are the maximum relevance values found in these networks, which
are themselves minimum paths between subgraphs with the restriction of visiting a node $i$
in the middle. Thus, the information obtained from the distance histograms can again be
complemented by observing that nodes not lying inside the core group do not impose large costs
for subgraph interconnectivity in ER and WS models. An empirical observation that supports
this conclusion is that the maximum values of the distance histograms for models ER and WS
Figure 6. Distributions of average subgraph sizes (left column) and average distances between subgraphs (right column) along with respective standard deviations considering 100 realizations of each network model (ER, WS, BA and GG). The left column also contains averages and standard deviations for the number of subgraphs $C$ in each model. The subgraphs were created using the nodes (2.5% of $N$) with the highest degree.
Figure 7. Average number of subgraphs $C^+$ (left column) and average number of nodes $|V^+|$ (right column) in the gradual merging performed for 100 realizations of models ER, WS, BA and GG (standard deviations are denoted by vertical bars). The subgraphs were created with respect to the nodes (2.5% of $N$) with the highest degree.
are, respectively, 4 and 6 (figure 6), whereas the maximum relevance values for these networks are 5 and 7, only one unit of distance greater than the maximum distances depicted in the histograms.

Highly connected nodes in BA, widely known as hubs, tend to form a single subgraph. Thus, the merging algorithm acts as a sequential dilation, expanding that unique subgraph until the entire network is covered at $T = 5$ (figure 7). This means that every node in BA is connected to the hub subgraph by a path made of not more than five nodes. Therefore, in ER, WS and BA networks, nodes or subgraphs with high degree are easily accessible from each other, as well as from the remaining nodes of the network, despite the existence of some important differences in the size and number of subgraphs between BA and the other two models. Since nodes with high degree concentrate connections and tend to shorten the paths, the ease of access to them is of primary importance in complex networks. For instance, diseases (or news) may spread fast in a social network (typically well-modeled by WS) if the first infected (or informed) person is a highly connected node, or if it is just a few steps away from such a node. Moreover, revisiting well-connected nodes is also crucial for faster spreading processes. Therefore, ER, WS and BA networks facilitate these processes, since all nodes are relatively close to highly connected subgraphs (see maximum $T$ discussed before), which in turn are close to each other (low $T$ for $C^+ = 1$), especially for BA networks, which have all hubs in the same subgraph. Procedures to stop epidemics may also become more successful if applied mostly at the highly connected subgraphs. If we consider transport processes, these subgraphs need to deal with considerable higher traffic than the rest of the network, not to mention the need for stronger security policies against attacks, otherwise a critical bottleneck may arise (especially in BA networks [25]).

Rather different features were found in GG networks. A considerably higher threshold is necessary to merge all subgraphs (at $T = 35$, figure 7), even with a low number of subgraphs (8.11 on average, figure 6). Furthermore, nearly 950 nodes are included in the merging when $C^+ = 1$, i.e. almost the entire network is required to merge subgraphs into an unique connected component, thus forming the aforementioned core group of nodes. The remaining nodes of the network are only included in the merging when $T$ is around 60. Indeed, GG networks generally impose longer distances between pair of nodes than the ones usually found in the other network models. Nevertheless, the exact configuration and distribution of subgraphs is only grasped using some specific methodology, such as the one presented in this paper. We also observe that nearly 50% of nodes in GG networks have relevance higher than 15, another indication that highly connected subgraphs are sparsely connected in these geographical structures. These results suggest that spreading processes can be much slower in GG networks than in the other models considered in this paper, besides making more difficult the execution of tasks based on transport or search processes.

4.1.2. Subgraphs with high clustering coefficient. The histograms shown in figure 8 were obtained by taking nodes with high clustering coefficient as references. All models, especially the BA, have most of their subgraphs with size 1. Thus, the number of subgraphs in this model is almost the highest possible on average ($C \approx 25$). Model ER has the lowest number of subgraphs (about 18 on average) since it is not rare to find subgraphs of size 2 or 3 in this model. In general, clustered subgraphs are farther from each other than highly connected subgraphs in all four models considered (see figures 6 and 8). For instance, model ER had its distance peak increased from 2 to 3, and WS from 4 to 5, when comparing histograms of figures 6 with those of figure 8. Nevertheless, these changes are relatively small when taking into account the total number of
Figure 8. Distributions of average subgraph sizes (left column) and average distances between subgraphs (right column) along with respective standard deviations considering 100 realizations of each network model (ER, WS, BA and GG). Averages and standard deviations for the number of subgraphs $C$ are also shown in the left column. The subgraphs were created using the nodes (2.5% of $N$) with the highest clustering coefficient.
nodes in these networks ($N = 1000$). Subgraphs in BA networks now tend to be 2 or 3 units of distance apart, rather than being all connected as in the BA histogram for hub subgraphs. Nevertheless, clustered subgraphs can be considered close to each other in all these models, given the much larger size of the networks when compared with the values in the distance histograms. Figure 9, which shows the number of subgraphs $C^+$ and the number of nodes $|V^+|$ in the gradual merging, reinforces this tendency. ER and BA networks have $C^+ = 1$ when $T = 2$ ($C^+$ drops faster in BA than in ER), using less than 100 nodes (the core group of nodes). Thus, both these network models have their clustered subgraphs quickly joined. Nevertheless, $T = 1$ joins the biggest portion of subgraphs in model BA, indicating that subgraphs are even more cohesive in BA than in ER. Subgraphs in WS networks are all merged when $T = 4$, with approximately 230 nodes included in the merging, thus demanding more nodes than ER and BA to join all subgraphs, although still requiring a low threshold to reach $C^+ = 1$. In summary, these three models (ER, WS and BA) tend to have trivial subgraphs close to each other, with less than 23% of the network inside the core group of nodes. Moreover, as the threshold necessary to cover the entire network is not higher than 9 in these models, it is possible to say that all $N$ nodes are relatively close to clustered subgraphs.

Higher subgraph distances were observed in the geographical model, where subgraphs are as far as 60 nodes apart (see figure 8). A peak around 20–30 was found in the respective distance histogram, with a slow decay for higher distance values. GG networks also present the most distinct results when concerning the merging algorithm (figure 9): a threshold $T \approx 25$ is necessary to bring together all subgraphs, when almost the entire network (nearly 950 nodes) is included in the merging. Furthermore, only when $T \approx 40$ is the whole network covered by the merging procedure, which means that the GG model has highly clustered (trivial) subgraphs sparsely distributed.

The proportion of triangles (cycles of length three) is, by definition, high around nodes of high clustering coefficient. Triangles are related to the property of transitivity, which means that if node $i$ is connected to node $j$ and node $j$ is connected to node $k$, then node $i$ must be connected to node $k$. Triangles contribute to increase the redundancy in a network, since short alternative paths exist between nearby nodes. Therefore, ER, WS and BA networks show link redundancy at close locations, all connected by a core of at most 23% of network nodes. Furthermore, as all nodes are quickly covered by the merging algorithm (at most at $T = 9$), nodes outside this core are also close to clustered regions. These results suggest that dynamical processes taking place in these networks may take advantage of the easy accessibility to highly redundant regions. For instance, congestion can be avoided by using alternative paths whenever needed. Moreover, link redundancy may prevent cascading failures when some node or edge becomes inoperable. GG networks, on the other hand, have their clustered subgraphs more dispersed, implying that link redundancy is neither concentrated nor easy accessible. Therefore, the network is more restricted regarding alternative paths, suggesting a worse topology for some dynamical processes (i.e. it may become more difficult to avoid congestion and cascading failures).

4.1.3. Clique subgraphs. Figure 10 shows size and distance histograms for subgraphs based on cliques. It is possible to notice that all cliques, when combined, form one single subgraph in BA networks, thus distances are always equal to 0 in these networks. The size of this single subgraph varies between 178 and 253, with peaks around 215–225 (i.e. it has nearly 20% of network nodes). In the gradual merging, BA networks are quickly covered at $T = 2$, with a steep
Figure 9. Average number of subgraphs $C^+$ (left column) and average number of nodes $\left| V^+ \right|$ (right column) in the gradual merging performed for 100 realizations of models ER, WS, BA and GG. The subgraphs were created with respect to the nodes (2.5% of $N$) with the highest clustering coefficient.
Figure 10. Distributions of average subgraph sizes (left column) and average distances between subgraphs (right column) and standard deviations (not for some insets in the left column, for better visualization) considering 100 realizations of each network model (ER, WS, BA and GG). Averages and standard deviations for the number of subgraphs $C$ are written in the left column. The subgraphs were created using complete subgraphs (cliques).
increase of nodes covered by the merging at $T = 1$ (please see figure 11). Therefore, all nodes in BA networks are intimately associated with cliques, which may be a good feature for several dynamical processes, since densely connected centers are easily accessible from any node. On the other hand, as cliques in these BA networks are very small (size 3), they can be easily broken by removing a single node or edge. It is possible that cliques may be critical bottlenecks for these networks, as a sequential removal of cliques directly affects the whole structure.

Cliqu es do not form a single subgraph in ER and WS networks (the average number of subgraphs is 7.12 and 2.99, respectively), although subgraphs are very close to each other in these networks, with distances peaked at 1 (figure 10). Subgraph sizes are frequently equal to 3 in ER and 4 in WS networks, with smaller probabilities of finding large subgraphs, including much larger ones (please observe the insets in the left column of figure 10, which show that some cliques are combined forming larger subgraphs). Despite the fact that only 40% of network nodes are outside subgraphs in WS networks, the merging procedure covers the entire network only at $T = 10$, whereas ER networks, with nearly 90% of nodes outside subgraphs, are entirely covered by the merging procedure at $T = 6$ (figure 11). Therefore, ER and WS networks have clique subgraphs very close to one another, although nodes lying outside subgraphs in WS networks are more distant from subgraphs. ER networks are similar to BA networks in the sense that cliques are very close to the whole network, although at different intensities (also notice that both networks have maximal cliques of size 3). WS networks, on the other hand, may have a slightly worse topology for dynamical processes than ER and BA, since a large portion of these networks is not so close to cliques. Nevertheless, when considering attacks directed to cliques, WS networks may not be entirely affected due to the fact that many of its nodes are not so closely connected to cliques.

GG networks have on average 24.23 subgraphs, more than observed for the other models (figure 10). The most frequent size of subgraphs is 6 in these networks, while sometimes sizes as high as 45 appear. Nearly 28% of network nodes are inside subgraphs, which are separated by at most 50 units of distance (with a peak at about 15), the highest distances observed in this experiment. Moreover, the merging procedure is able to join all subgraphs in one connected component at $T = 10$ (figure 11), comprising nearly 850 nodes (i.e. 85% of all nodes). These results show that most nodes in GG networks have a weak relationship with clique subgraphs, thus suggesting a less efficient topology for dynamical processes (such as transportation and communications) when compared with the other network models. On the other hand, if cliques are removed, other portions of GG networks can still function properly due to their loose connection with cliques.

4.2. Real-world networks

4.2.1. Subgraphs with high degree. The size and distance histograms for real-world networks considering subgraphs with high degree are shown in figure 12. The Email network concentrates all highly connected nodes in one single subgraph of size 28, so that the respective histogram presents only null distances. Similarly, the metabolic network of *E. coli* also organizes highly connected nodes in only one subgraph (of size 19). The Internet, a much larger network, gathers virtually all hubs into one subgraph of size 569, coexisting with five other trivial subgraphs. The distances between the six subgraphs of the Internet vary between 1 and 4, relatively short distances. Opposite behavior can be observed in the Power Grid, where trivial subgraphs prevail. Nearly 75% of Power Grid subgraphs have only one node. Some bigger subgraphs
Figure 11. Average number of subgraphs $C^+$ (left column) and average number of nodes $|V^+|$ (right column) in the gradual merging performed for 100 realizations of models ER, WS, BA and GG. The subgraphs were created using complete subgraphs (cliques).
Figure 12. Distributions of subgraph sizes (top plot) and distances between subgraphs (bottom plot) for the real-world networks. The top plot also contains the number of subgraphs $C$ in each network. In the bottom plot histograms for *E. coli* and Email are superimposed on one another. The subgraphs were created using the nodes (2.5% of $N$) with the highest degree.

...can also be found, mainly of size 2 and 3, and some of sizes between 10 and 20. A distinctive behavior of this network is the large variation of distance values, ranging from 1 to 35 with a peak at distance 11. NetScience has only four subgraphs of sizes 1, 2 and 4 and distances ranging from 1 to 5.

Before interpreting these results, we point out some features of the respective gradual mergings (figure 13). For the *E. coli* and Email networks, where highly connected nodes form one subgraph, the merging procedure acts as a sequential dilation until the whole network is covered at $T = 3$ and 4, respectively. Therefore, every node in these networks is close to the highly connected subgraph, behavior similar to that observed in BA networks. The Internet also resembles results obtained for BA, since almost every highly connected node of this network is in the same subgraph, and $C^+ = 1$ is quickly achieved at $T = 1$, when 18 468 nodes (80% of $N$) are selected by the merging procedure. Thus, a great part of the Internet is intrinsically associated with hubs, despite the fact that the other 20% of nodes is completely covered by the merging only at $T = 27$. Hence, the *E. coli*, Email and Internet networks have highly connected nodes easily accessible to one another, and from other nodes as well. Thus, as already discussed for models ER, WS and BA, dynamical processes on these real-world networks can take great advantage of these properties. For instance, highly connected substrates in the *E. coli* network play an important role by connecting modules that perform different metabolic functions [21, 22]. Thus, the observed ease of access to hubs is essential to the robustness of a reaction pathway. Information transferred by email can quickly cover the entire network by sequentially accessing well connected people. On the other hand, we know that computer viruses
attached to emails usually take advantage of this topology to quickly infect as many systems as possible. In the case of the Internet, data packets can be transmitted faster by reaching a nearby hub as needed. On the other hand, the Internet is extremely fragile to attacks on hubs, as well as BA networks [25] (note that both have a power-law distribution of degrees [8]).

The Power Grid takes a longer time to reach \( C^+ = 1 \), since a threshold \( T = 7 \) is responsible for joining all subgraphs, where 2386 nodes are inside the merging (48% of \( N \)). Thus, highly connected subgraphs are not cohesively connected to each other in the Power Grid, as the distance histogram already indicated (figure 12). Nevertheless, this threshold \( T = 7 \) is much lower than the overall distances depicted in its histogram, although not as small as in the other real-world networks. Moreover, the Power Grid slowly covers the entire network (the merging finishes at \( T = 27 \)). This network roughly resembles GG networks, where highly connected subgraphs are sparsely distributed over the network. Therefore, this behavior probably results from the geographical nature of the Power Grid. The last real network of this subsection, NetScience, reaches \( C^+ = 1 \) quickly (at \( T = 2 \)) using 27% of its nodes (102 nodes), thus forming a tightly connected core group of highly connected scientists. The remaining portion of NetScience takes a much longer time to be entirely included in the merging (only at \( T = 16 \), a high value for this small network), therefore being loosely connected to subgraphs. The core group of scientists is able to associate, through a very short path, two or more highly connected scientists, as the results for \( T = 2 \) illustrate. If we consider that knowledge is better transmitted from one scientist to another through collaborations (possibly including personal meetings) rather than by just reading written publications, this group of nearly 100 scientists have prime access to leading research on complex networks conducted by highly connected scientists.

**Figure 13.** Number of subgraphs \( C^+ \) (top plot) and number of nodes \( |V^+| \) (bottom plot) for the gradual merging performed in the real-world networks. *E. coli* and Email have \( C^+ = 1 \) for every \( T \). The subgraphs were created with respect to the nodes (2.5% of \( N \)) with the highest degree.
Figure 14. Distributions of subgraph sizes (top plot) and distances between subgraphs (bottom plot) for the real-world networks. Observe that the top plot also contains the number of subgraphs \( C \) in each network. Moreover, the curves for \( E. \ coli \), Email and Internet-AS are superimposed on one another in the top plot. The subgraphs were created using the nodes (2.5% of \( N \)) with the highest clustering coefficient.

4.2.2. Subgraphs with high clustering coefficient. Figure 14 shows the distance and size histograms for subgraphs based on high clustering coefficient. As already observed for network models, clustered subgraphs in real-world networks tend to comprise single nodes, particularly \( E. \ coli \), Email and Internet. The NetScience and Power Grid networks have some slightly bigger subgraphs, although most subgraphs have only one node. Clustered subgraphs are remarkably close to one another in the Internet, at most distant by two nodes. The Email and \( E. \ coli \) networks also present subgraphs near one another, with all distances lower than or equal to 4. On the other hand, Power Grid and NetScience have clustered subgraphs that are more distributed, mainly the former one, with subgraphs separated by at most 35 units of distance, with peaks at distances 9 and 18. NetScience shows distances varying from 1 to 9, high values when considering the small size of this network.

When observing figure 15, it is possible to note that the Internet subgraphs are merged when the relevance threshold is \( T = 2 \). Remarkably, only 653 nodes (2.84% of \( N \)) are included in the merging at this time (the core group), i.e. only 0.34% of \( N \), or 81 nodes, are capable of merging all 572 subgraphs of the Internet, since 2.5% of network nodes are initially selected to form subgraphs. Other networks also do not demand high relevance thresholds to reach \( C^+ = 1 \), although including relatively more nodes in the core group: \( E. \ coli \) demands \( T = 3 \) to merge all its subgraphs, requiring 345 nodes (45% of \( N \)); Email subgraphs are joined at \( T = 3 \), using 368 nodes (32% of \( N \)) and NetScience maintains its subgraphs separated until \( T = 5 \), when 90 nodes are included in the merging (24% of \( N \)). Observe that even though the NetScience
subgraphs are separated by distances as high as 9 (figure 14), a much lower relevance threshold (almost half of this value) is sufficient to reach \( C^+ = 1 \). This fact shows that clustered subgraphs in this network are in fact closer than the respective distance histogram shows, since we only consider the core group of nodes necessary to join subgraphs. Nevertheless, outside this group, it is possible to find nodes with relevance values as high as 13 (see figure 15), a large value for such a small network. Thus, outside the core group of nodes, highly clustered subgraphs are not so easy to access in NetScience.

Generally speaking, Internet, \( E. \ coli \) and Email present similar behavior as observed in the models ER, WS and BA (for subgraphs with high clustering coefficient), since all of them reach \( C^+ = 1 \) with relatively low thresholds, and neither of them demand high thresholds to cover the entire network in the merging procedure (mainly the model BA and the Internet). For the Internet, the interpretation of these results is similar to that of network models: data packet transportation can take advantage of link redundancy in order to avoid congestion and prevent cascading failures, since alternative paths (triangles) are frequently available when traversing the network. Assuming that the Email network encodes relationships at least at the acquaintance level, since bulk emails were not considered in the network formation, we can conclude that acquaintance cycles of length 3 (i.e. person \( a \) knows person \( b \) who knows \( c \) who in turn knows \( a \)) are fairly distributed all over the network. In the case of \( E. \ coli \), the nodes (substrates) with high clustering coefficient were found to be intrinsically related to known metabolic functions [21]. Since connectivity between modules is an important feature of metabolic structures, the proximity between highly clustered subgraphs shows that the

**Figure 15.** Number of subgraphs \( C^+ \) (top plot) and number of nodes \( |V^+| \) (bottom plot) for the gradual merging performed in the real-world networks. The subgraphs were created with respect to the nodes (2.5% of \( N \)) with the highest clustering coefficient.
$E. \ coli$ network is especially arranged to maximize intermodule communication. This is not the case of NetScience, where cycles are less frequently found outside the core group of nodes, thus separating the network into a region of well-developed social ties and another where it is more difficult to find a common collaborator of a pair of scientists (i.e. a cycle of length 3).

Exhibiting completely different behavior, the Power Grid only has its subgraphs merged at $T = 9$ (figure 15). In addition, the merging procedure takes much longer to cover the entire Power Grid (only when $T = 38$) than the other real networks. These results suggest that the Power Grid may be more sensitive to overload and also more likely to suffer cascading failures. It is worth pointing out that these conclusions do not take into account the transmission capacity of each element of the Power Grid.

4.2.3. Clique subgraphs. Results for clique subgraphs in real-world networks can be seen in figures 16 and 17. $E. \ coli$, the Power Grid and the Internet group cliques in a single subgraph that includes a small fraction of network nodes (respectively, 3.67, 0.24 and 2.44%). Nevertheless, only in $E. \ coli$ and the Internet are nodes not lying inside subgraphs close to subgraphs, i.e. the threshold necessary to cover all the network is $T = 3$ and 5, respectively (for the Power Grid, the threshold is $T = 39$). Therefore, in this respect $E. \ coli$ and the Internet bear some resemblance to the BA model, since there is only one subgraph easily accessible from other nodes. However, cliques are smaller in BA (its maximum clique is of size 3), and also more frequent (i.e. nearly 20% of BA nodes are inside cliques). The Email network has two subgraphs.

Figure 16. Distributions of subgraph sizes (top plot) and distances between subgraphs (bottom plot) for the real-world networks. The bottom plot also shows the number of subgraphs $C$ in each network. The curves for $E. \ coli$, Power Grid and Internet-AS are superimposed on one another in the bottom plot. The subgraphs were created using complete subgraphs (cliques).
very close to each other which, when combined, comprise 14% of network nodes. Furthermore, other nodes are not much distant from the subgraphs \((T = 7)\) covers all the network in the merging), thus presenting behavior similar to the observed for ER networks. Therefore, three real-world networks (E. coli, the Internet and Email) have all nodes strongly associated with cliques, a feature that, as explained for model networks, may facilitate dynamical processes such as transport, spreading, search and so on, since groups of maximum interconnectivity are easily accessible from anywhere in the network. However, especially if cliques cover a small portion of the network, as in E. coli and the Internet, an attack directed to cliques may rapidly undermine the whole network.

NetScience has nearly 18% of its nodes inside six clique subgraphs, which are all quickly merged at \(T = 3\). Nevertheless, some nodes are far from cliques, as a threshold \(T = 14\) is necessary to cover all the network (which is a high value for this small network). Similarly to the Power Grid, NetScience contains cliques very close to each other, while a considerable portion of the network is far away from cliques (thus somewhat resembling WS networks). As explained earlier, the ease of access to cliques is an important feature in dynamical processes unfolding in the network, since cliques can easily distribute traffic to accelerate transmission. As the Power Grid and NetScience networks have a considerable number of nodes far from cliques, we argue here that the distribution of electricity and the spreading of knowledge may be facilitated if shortcuts are created between cliques and far away nodes (naturally, we should also consider geographical distances and power capacities in the Power Grid).

At this point, it is worthwhile to make some comments integrating all the experiments reported in this section. The Power Grid and the GG networks consistently show similar results.

Figure 17. Number of subgraphs \(C^+\) (top plot) and number of nodes \(|V^+|\) (bottom plot) for the gradual merging performed in the real-world networks. The subgraphs were based on complete subgraphs (cliques).
for both highly connected and highly clustered subgraphs. Though geographically constrained, the Power Grid has been associated with the WS model because of its small-world property [18]. Indeed, some similarities appear between the Power Grid and WS networks when considering subgraphs based on cliques. The Internet at the autonomous system level, although also being geographically constrained, shows very different results than the GG model in all experiments performed. In fact, the Internet is closely associated with the BA model, since it shows a power-law degree distribution [8, 9]. Our results reinforce this belief, as the Internet presents results similar to those obtained for the BA model in all experiments reported in this paper. The Email and E. coli networks are also similar to BA networks in almost all experiments reported in this paper (with the exception that the Email network is more similar to ER networks in the experiment with cliques). Nevertheless, both networks significantly deviate from the BA model: the Email network has an exponential distribution of degrees [23] and the E. coli network is better reproduced by a hierarchical model [21] (more details on this model are given below). It is possible, on the other hand, to associate NetScience with the BA model, since power-law degree distributions were observed in scientific collaboration networks [26]. Despite this fact, our results show little resemblance between these two networks.

The E. coli network also has a power-law degree distribution [22], but it is better associated with a deterministic hierarchical model that reproduces well many of its features, including the scale-free distribution of degrees [21]. The hierarchical model has highly clustered nodes distributed at its bottom hierarchical levels, whereas highly connected nodes recursively join lower levels up to the point where all nodes are inside the whole hierarchy (hubs are thus at the top levels). Some of our results agree with this model by showing that hubs are all inside the same connected component, thus forming the highest hierarchy from which all other branches derive. Highly clustered nodes are all separated from each other in our experiment, deviating from the hierarchical model where some of these nodes are interconnected. Nevertheless, since there is strong evidence of a hierarchical organization in metabolic networks [21], our results show that highly clustered nodes are still near one another. Besides, this result suggests that even when clustered nodes are placed inside distinct hierarchical branches, it is common to find a short path connecting them. Finally, we also found that different cliques in the E. coli network are all combined in one subgraph.

5. Concluding remarks

In this paper, we presented a framework for characterization of the distribution of critical subgraphs through complex networks. We adopted distance histograms to assess the overall distribution of subgraphs and also developed an algorithm to progressively merge subgraphs according to a metric of node relevance. The merging approach complements the distance histogram by providing means to analyze every node in the network regarding their relevance to subgraph interconnectivity, as well as for allowing the identification of the connectivity surrounding each critical subgraph.

Rather than characterizing single nodes exclusively, the proposed framework operates at a higher topological level by allowing the analysis of groups of nodes and their interconnectivity. Closely related subgraphs have been the focus of many network-based studies, such as in the analysis of communities and motifs. However, differently from communities, the method proposed in this paper does not require a partition of the network, and also does not identify small subgraph patterns (i.e. motifs). Our main motivation is to analyze the interconnectivity and
dispersion of similar groups of nodes (according to any desired criteria, particularly nodes with critical topological features) independently of their size. The proposed method was analyzed with respect to both theoretical and real-world networks. Subgraphs comprising nodes with high degree and high clustering coefficient, as well as nodes that form cliques, were taken into account independently. The experimental findings reported in this paper contribute to a better understanding of the structure of the studied networks, allowing us to draw some conclusions about the importance of these findings with respect to each considered case.

Further developments could focus on similar analysis using other networks, as well as different types of subgraphs. Another interesting application of the merging algorithm would be in multilevel graph analysis, where each subgraph would be considered a single entity to which a number of measurements could be applied. Thus, at each relevance threshold of the merging procedure, it would be possible to analyze the new subgraphs and trace the evolution of their properties. Dendrograms depicting the subgraph merging dynamics could be obtained, providing a complete characterization of the subgraph hierarchy. The application of concentric measurements [12] also provides interesting possibilities for the analysis of subgraphs along different neighborhoods.

Acknowledgments

L da F Costa is grateful to FAPESP (05/00587-5) and CNPq (301303/06-1) for financial support. L Antiqueira thanks the sponsorship of FAPESP (06/61743-7). The authors also thank M E J Newman, A Arenas, D Watts and CCNR for making their datasets available on the WWW.

References

[1] Albert R and Barabási A-L 2002 Statistical mechanics of complex networks Rev. Mod. Phys. 74 47–97
[2] Dorogovtsev S N and Mendes J F F 2002 Evolution of networks Adv. Phys. 51 1079–187
[3] Newman M E J 2003 The structure and function of complex networks SIAM Rev. 45 167–256
[4] Boccaletti S, Latora V, Moreno Y, Chavez M and Hwang D-U 2006 Complex networks: structure and dynamics Phys. Rep. 424 175–308
[5] Costa L da F, Rodrigues F A, Travieso G and Villas Boas P R 2007 Characterization of complex networks: a survey of measurements Adv. Phys. 56 167–242
[6] Costa L da F, Oliveira O N Jr, Travieso G, Rodrigues F A, Villas Boas P R, Antiqueira L, Viana M P and da Rocha L E C 2008 Analyzing and modeling real-world phenomena with complex networks: a survey of applications arXiv:0711.3199v3
[7] Barabási A-L 2007 The architecture of complexity IEEE Control Syst. Mag. 27 33–42
[8] Faloutsos M, Faloutsos P and Faloutsos C 1999 On power-law relationships of the Internet topology Proc. Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication pp 251–62
[9] Barabási A-L and Albert R 1999 Emergence of scaling in random networks Science 286 509–12
[10] Clauset A, Newman M E J and Moore C 2004 Finding community structure in very large networks Phys. Rev. E 70 066111
[11] Newman M E J 2006 Modularity and community structure in networks Proc. Natl Acad. Sci. USA 103 8577–82
[12] Costa L da F and da Rocha L E C 2006 A generalized approach to complex networks Eur. Phys. J. B 50 237–42
[13] Costa L da F and Rodrigues F A 2008 Seeking for simplicity in complex networks arXiv:physics/0702102v3
[14] Lopez E, Parshani R, Cohen R, Carmi S and Havlin S 2007 Limited path percolation in complex networks Phys. Rev. Lett. 99 188701
[15] Costa L da F and Rodrigues F A 2008 What is there between any two nodes in a complex network? arXiv:0801.4068v3
[16] Makarov L I 2005 Estimations of subgraph positions in molecular graphs and their common subgraph peculiarities J. Struc. Chem. 46 738–43
[17] Bondy J A and Murty U S R 1976 Graph Theory with Applications (New York: Elsevier)
[18] Watts D J and Strogatz S H 1998 Collective dynamics of ‘small-world’ networks Nature 393 440–2
[19] Erdős P and Rényi A 1959 On random graphs I Publicationes Math. (Debrecen) 6 290–7
[20] Newman M E J 2006 Finding community structure in networks using the eigenvectors of matrices Phys. Rev. E 74 036104
[21] Ravasz E, Somera A L, Mongru D A, Oltvai Z N and Barabási A-L 2002 Hierarchical organization of modularity in metabolic networks Science 297 1551–5
[22] Jeong H, Tombor B, Albert R, Oltvai Z N and Barabási A-L 2000 The large-scale organization of metabolic networks Nature 407 651–4
[23] Guimerà R, Danon L, Díaz-Guilera A, Giralt F and Arenas A 2003 Self-similar community structure in a network of human interactions Phys. Rev. E 68 065103
[24] Heijmans H J A M, Nacken P, Toet A and Vincent L 1992 Graph morphology J. Vis. Commun. Image Represent. 3 24–38
[25] Albert R, Jeong H and Barabási A-L 2000 Error and attack tolerance of complex networks Nature 406 378–82
[26] Newman M E J 2001 Scientific collaboration networks. I. Network construction and fundamental results Phys. Rev. E 64 016131

New Journal of Physics 11 (2009) 013058 (http://www.njp.org/)