Chain motifs: The tails and handles of complex networks

Paulino R. Villas Boas, Francisco A. Rodrigues, Gonzalo Travieso, and Luciano da Fontoura Costa
Institute of Physics at São Carlos, University of São Paulo, P.O. Box 369, São Carlos, São Paulo, 13560-970 Brazil

(Received 21 June 2007; revised manuscript received 26 October 2007; published 13 February 2008)

A great part of the interest in complex networks has been motivated by the presence of structured, frequently nonuniform, connectivity. Because diverse connectivity patterns tend to result in distinct network dynamics, and also because they provide the means to identify and classify several types of complex network, it becomes important to obtain meaningful measurements of the local network topology. In addition to traditional features such as the node degree, clustering coefficient, and shortest path, motifs have been introduced in the literature in order to provide complementary descriptions of the network connectivity. The current work proposes a different type of motif, namely, chains of nodes, that is, sequences of connected nodes with degree 2. These chains have been subdivided into cords, tails, rings, and handles, depending on the type of their extremities (e.g., open or connected). A theoretical analysis of the density of such motifs in random and scale-free networks is described, and an algorithm for identifying these motifs in general networks is presented. The potential of considering chains for network characterization has been illustrated with respect to five categories of real-world networks including 16 cases. Several interesting findings were obtained, including the fact that several chains were observed in real-world networks, especially the world wide web, books, and the power grid. The possibility of chains resulting from incompletely sampled networks is also investigated.

DOI: 10.1103/PhysRevE.77.026106 PACS number(s): 89.75.Fb, 02.10.Ox, 89.75.Da, 87.85.Ng

I. INTRODUCTION

A large number of interesting dynamic systems can be studied and modeled by first representing them as networks and then considering specific dynamic models. Because the latter depend greatly on the connectivity of the network, it becomes critical to obtain good characterizations of the corresponding connectivity structure. This characterization is even more important in cases when the dynamics is not considered, e.g., while analyzing a frozen instant of systems such as the internet and protein-protein interaction networks. Therefore, it is hardly surprising that a great deal of effort (e.g., [1]) has been invested in developing new measurements capable of providing meaningful and comprehensive characterization of the connectivity structure of complex networks.

Traditional measurements of the topology of complex networks include the classical vertex degree and the clustering coefficient (e.g., [2]). Both these features are defined for each vertex in the network and express the connectivity only at the immediate neighborhood of that reference vertex. Other measurements such as the minimum shortest path and betweenness centrality reflect the connectivity of broader portions of the network. Hierarchical measurements (e.g., [3–6]) such as the hierarchical vertex degree and hierarchical clustering coefficient, also applicable to individual reference vertices, have been proposed in order to reflect the connectivity properties along successive hierarchical neighborhoods around the reference vertex. Another interesting family of measurements of the topological properties of complex networks involves the quantification of the frequency of basic motifs in the network (e.g., [7–10]). Motifs are subgraphs corresponding to the simplest structural elements found in networks, in the sense of involving small numbers of vertices and edges. Examples of motifs include feedforward loops, cycles of order 3, and bifans.

Preliminary studies of chains of nodes in networks have been made. Costa [11] studied the effect of chains in affecting the fractal dimension as revealed by dilations along networks. Kaiser and Hilgetag [12] studied the vulnerability of networks involving linear chains with an open extremity. In another work [13], they addressed the presence of this same type of motif in a sparse model of a spatial network. More recently, Levnajić and Tadić [14] investigated the dynamics in simple networks including linear chains of nodes.

Although several measurements are now available in the literature, their application will always be strongly related to each specific problem. In other words, there is no definitive or complete set of measurements for the characterization of the topology of complex networks. For instance, in case one is interested in the community structures, measurements such as the modularity are more likely to provide valuable and meaningful information [15]. In this sense, specific new problems will likely continue to motivate novel, especially suited, measurements. The reader is referred to the survey [1] for a more extensive discussion of measurement choices and applications.

The current work proposes a complementary way to characterize the connectivity of complex networks in terms of a special class of motifs defined by *chains* of vertices, which are motifs composed by vertices connected in a sequential way, where the internal vertices have degree 2. These motifs include *cords*, *tails*, *rings*, and *handles*. While tails and handles have at least one extremity connected to the remainder of the network, cords and rings are disconnected, being composed of groups of vertices connected in a sequential way. Additional motifs such as two or more handles connected to the remainder of the network, namely, n-handles with $n \geqslant 2$, can also be defined, but they are not considered in this work.

Figure 1 illustrates six types of chain, namely, (a) a cord, (b) a tail, (c) a two-tail, (d) a ring, (e) a handle, and (f) an n-handle. The main difference between the traditional motifs

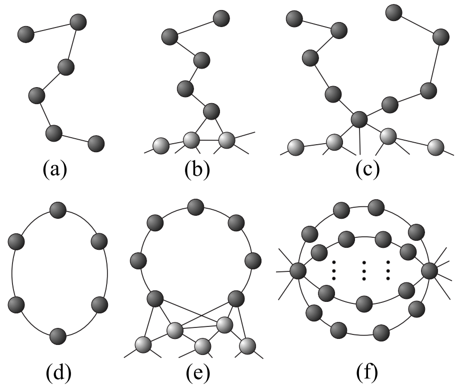


FIG. 1. Chains can be classified into different types, depending on the connections among their external vertices. Here are shown six types of chains (dark gray vertices): (a) a cord, (b) a tail, (c) a two-tail, (d) a ring, (e) a handle, and (f) an n—handle.

and those defined and characterized in this paper is that the latter may involve a large number of vertices and edges.

The main motivation behind the introduction of the concept of chains in complex networks provided in this paper is that such a structure is odd in the sense that it can be conceptualized as an edge containing a series of intermediate vertices which make no branches. In several aspects, such as flow, the incorporation of such intermediate vertices along an edge will imply virtually no change in the overall dynamics of that substructure of the network. In other words, the same flow capacity will be offered by either the isolated edge or its version incorporating a series of intermediate vertices. Interestingly, vertices with only two neighbors—henceforth called *articulations*—seem to have a rather distinct nature and role in complex networks, which suggests that they may have distinct origins. For instance, as explored further in this work, articulations seem to appear in networks generated by sequential processes (e.g., word adjacency in books), but can also be a consequence of incompleteness of the building process of networks. The latter possibility is experimentally investigated in this work by considering incompletely sampled versions of network models.

In addition to introducing the concept and a theory of chains and articulations in complex networks and presenting means for their identification, the present work also illustrates the potential of considering the statistics of cords, tails, and handles for characterizing real-world networks (social, information, technological, word adjacency in books, and biological networks). This paper starts by presenting the definition of chains and their categories (i.e., cords, tails, and handles), and proceeds by developing an analytical investigation of the density of chains in random and scale-free models. Next, an algorithm for the identification of such motifs is described, following by a discussion of the chain statistics obtained. The application of such a methodology considers the characterization of real-world complex networks in terms of chain motifs.

II. CHAINS, CORDS, TAILS, HANDLES, AND RINGS

Given a network with N vertices, consider a sequence $(n_1, n_2, \ldots, n_{m+1})$ of $m+1$ vertices n_i. If the sequence has the following properties: (1) There is an edge between vertices n_i and n_{i+1}, $1 \leqslant i \leqslant m$; (2) vertices n_1 and n_{m+1} have degree not equal to 2; and (3) intermediate vertices n_i, $2 \leqslant i \leqslant m$, if any, have degree 2; we call the sequence a chain of length m. Vertices n_1 and n_{m+1} are called the *extremities* of the chain.

Chains can be classified in four categories (k_{n_i} is the degree of vertex n_i): *Cords* are chains with $k_{n_1}=1$ and $k_{n_{m+1}}=1$; *handles* are chains with $k_{n_1}>2$ and $k_{n_{m+1}}>2$; *tails* are chains with $k_{n_1}=1$ and $k_{n_{m+1}}>2$ (or equivalently $k_{n_1}>2$ and $k_{n_{m+1}}=1$); *rings* (of length m) are sequences $(n_1, n_2, \ldots, n_m)$ of m vertices where the degree of each vertex is $k_{n_i}=2$, $1 \leqslant n \leqslant m$, n_i is adjacent to n_{i+1} (for $1 \leqslant i \leqslant m-1$), and n_m is adjacent to n_1. Rings are a special case of chains in which there are no extremities, and the category was included in the chain classification only for completeness.

Including the trivial cases with $m=1$, it is easy to see that each vertex of degree 1 is at an extremity of a cord or a tail and each vertex of degree greater than 2 is at an extremity of a tail or a handle. Note that the definition of handles includes the degenerate case where the extremities are the same vertex: $n_1 = n_{m+1}$.

With these definitions and writing N_C, N_H, N_T, and N_R for the total number of cords, handles, tails, and rings, respectively, and $N(k)$ for the number of vertices of degree k, we have

$$N(1) = 2N_C + N_T, \tag{1}$$

$$\sum_{k>2} kN(k) = 2N_H + N_T. \tag{2}$$

To evaluate the number of vertices of degree 2, we introduce the notation $N_C(m)$ for the number of cords of length m, and similarly $N_H(m)$ for handles, $N_T(m)$ for tails, and $N_R(m)$ for rings. Each chain of length m has $m-1$ and each ring of length m has m vertices of degree 2, giving

$$N(2) = \sum_{m=1}^{\infty} \{mN_R(m) + (m-1)[N_C(m) + N_H(m) + N_T(m)]\}. \tag{3}$$

Isolated vertices (vertices with degree 0) have no effect on such structures, and it is considered hereafter that the network has no isolated nodes.

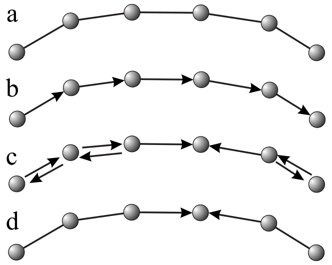


FIG. 2. The chain can be (a) undirected, (b) directed, or (c) mixed. Mixed chains have arcs in any direction. Note that (c) and (d) are equivalent.

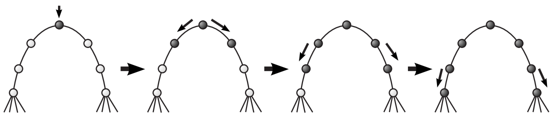


FIG. 3. The main steps to identify handles of size greater than 2 in networks are the following. (i) Choose a vertex of degree 2 and add it to a list (dark gray vertex); (ii) go to its neighbors and also add them if they have degree 2; (iii) go to the next neighbors, excluding the vertices already added to the list, and also add them if they have degree 2; (iv) stop adding vertices to the list after finding two vertices of degree greater than 2. In this case, the size of the obtained handle is 6. The same procedure can also be applied to find cords and tails, but at least one extremity should have degree equal to 1.

The chains can also be classified according to the nature of their connections as in Fig. 2. In undirected networks, the chains are *undirected* (Fig. 2). In directed networks, on the other hand, the chains can be classified into three types.

(1) *Directed chains* are those whose arcs of inner vertices follow just one direction, i.e., there is a directed path from one extremity to the other [Fig. 2(b)].

(2) *Undirected chains* are defined as for undirected networks, which have undirected arcs between inner vertices [Fig. 2(a)]. An undirected arc between vertices i and j exists if there is an arc from i to j and another from j to i.

(3) *Mixed chains* are those with any other combination of arc directions as in Fig. 2(c).

In our analysis we consider just undirect networks, but the extension for direct networks is straightforward.

III. ALGORITHM FOR CHAIN IDENTIFICATION

The algorithm to identify chains of vertices includes two steps, one for finding chains of size greater than 1 and the other for finding chains of unit size. The first step is illustrated in Fig. 3 and described as follows.

```
(1) Input: graph $G$.
(2) Output: list containing all chains of size greater than 2.
(3) Calculate the degree of vertices in $G$ and store them in a list $K$.
(4) Find vertices $i$ such that $k_i=2$, $k_i \in K$, and store them in a list $Q2$.
(5) While $Q2$ is not empty,
    (a) remove a vertex ($A$) from $Q2$ and then insert its first neighboring vertex ($B$), $A$, and its second neighboring vertex ($C$) in a queue $P$ (in this order).
    (b) While the first and last elements of $P$ have degree equal to 2 or are not the same do the following.
        (i) Let $D$ be the neighboring node of the first element in $P$. In case $D$ is not already in $P$, include it into that queue in the first position.
        (ii) If $D$ is in $Q2$, remove it.
        (iii) Let $E$ be the neighboring node of the last element in $P$. In case $E$ is not already in $P$, include it into that queue in the last position.
        (iv) If $E$ is in $Q2$, remove it.
    (c) Insert $P$ in a list $L$ and clear $P$.
```

The list L contains all chains of size greater than 2. They can now be classified into cords, tails, and handles according to the degree of the first and last elements of the corresponding queue.

The second step, required for identifying the chains of unit length, is as follows.

```
(1) Input: graph $G$, list $K$, and list $L$.
(2) Output: list of cords, tails, and handles of unit size.
(3) Find all vertices of degree equal to 1 which were not in $L$ and store them in a list $Q1$.
(4) While $Q1$ is not empty,
    (a) remove a vertex from $Q1$ and insert it in a queue $P$;
    (b) if the neighboring node of $A$ has degree also equal to 1, remove it from $Q1$, insert it in $P$, and insert $P$ in a list $C1$;
    (c) else insert its neighbor in $P$ and insert $P$ in a list $T1$.
(5) include all pairs of connected vertices that are not in $L$, $C1$, or $T1$ in a list $H1$.
```

The lists $C1$, $T1$, and $H1$ contain, respectively, all cords, tails, and handles of unit size in the network.

IV. STATISTICS

Consider an ensemble of networks completely determined by the degree-degree correlations $P(k,k')$.[1] Given $P(k,k')$ and the number of vertices in the network, we want to evaluate the number of each chain type and the rings. The degree distribution $P(k)$ and the conditional neighbor degree distribution $P(k'|k)$, i.e., the probability that a neighbor of a vertex with degree k has degree k', are easily computed:

$$P(k) = \frac{\sum_{k'} P(k,k')/k}{\sum_{k',k''} P(k',k'')/k'}, \tag{4}$$

$$P(k'|k) = \frac{\langle k\rangle P(k,k')}{kP(k)}, \tag{5}$$

where $\langle k\rangle = \Sigma_k kP(k)$ is the average degree of the network.

A. Rings

For a ring of length m, we start at a vertex of degree 2, go through $m-1$ vertices of degree 2, and come back to the original vertex. Each transition from a vertex of degree 2 to another, with the exception of the last one that closes the ring, has probability $P(2|2)$; the closing of the ring requires reaching one of the vertices of degree 2 [probability $P(2|2)$] and, among them, exactly the first one [probability

[1] For such an ensemble to be possible, connections from a vertex to itself (self-connections) and multiple connections between two vertices must be allowed, in contrast to many network models. Such self- and multiple connections will be rare provided the network is sufficiently large.

$1/(NP(2))$]. If we start from all vertices of degree 2, each ring will be counted m times, resulting in

$$N_R(m) = \frac{1}{m} P(2|2)^m. \tag{6}$$

This expression is valid only for the case of small m and large N, such that the vertices already included in the ring do not significantly affect the conditional probabilities. Such an approximation is used throughout this work. Note that, under these circumstances, when computing Eq. (3), $N_R(m)$ is of the order of the approximation error in the expressions of $N_C(m)$, $N_T(m)$, and $N_H(m)$.

B. Cords

Starting from a vertex of degree 1, a cord is traversed by following through a set of vertices of degree 2 until a vertex of degree 1 that ends the cord is reached. A cord of length 1 has no intermediate vertices; starting in a vertex of degree 1, the probability of finding a cord of length 1 is therefore given by $P(1|1)$. For a cord of length 2, the edge from the initial vertex should go through a vertex of degree 2 before arriving at a new vertex of degree 1, giving $P(2|1)P(1|2)$. For lengths greater than 2, each new intermediate vertex is reached with probability $P(2|2)$, and therefore we have $P(2|1)P(2|2)^{m-2}P(1|2)$ for a cord of length m. Considering that there are $NP(1)$ vertices of degree 1 in the network, but only half of them must be taken as the starting vertex to find a cord, we arrive at

$$N_C(m) = \begin{cases} \frac{1}{2} NP(1)P(1|1) & \text{if } m = 1, \\ \frac{1}{2} NP(1)P(2|1)P(2|2)^{m-2}P(1|2) & \text{if } m > 1. \end{cases} \tag{7}$$

C. Tails

The number of tails can be computed similarly. We need to either start at a vertex with degree 1 and reach a vertex of degree greater than 2 or vice versa; only one of these possibilities must be considered. We arrive at

$$N_T(m) = \begin{cases} NP(1)P(>2|1) & \text{if } m = 1, \\ NP(1)P(2|1)P(2|2)^{m-2}P(>2|2) & \text{if } m > 1, \end{cases} \tag{8}$$

where the notation $P(>2|k) = \Sigma_{k'>2} P(k'|k)$ is used.

D. Handles

A handle starts in a vertex of degree $k>2$ and ends in a vertex of degree $k'>2$. Starting from one of the $NP(k)$ vertices of degree $k>2$ of the network, there are k possibilities to follow a chain, each characterized by a sequence of vertices of degree 2 until reaching a vertex of degree $k'>2$. This gives a total of $NkP(k)P(>2|k)$ handles of length 1 and $NkP(k)P(2|k)P(2|2)^{m-2}P(>2|2)$ handles of length $m>1$. Summing up for all values of $k>2$, using $\Sigma_k kP(k)P(k'|k) = k'P(k')$, which can be deduced from relations (4) and (5), and considering that each handle is counted twice when starting from all nodes of degree greater than 2, we have

$$N_H(m) = \begin{cases} \frac{1}{2} N\{\langle k \rangle - P(1)[2 - P(1|1) - P(2|1)] - P(2)[4 - P(1|2) - P(2|2)]\} & \text{if } m = 1, \\ \frac{1}{2} N[2P(2) - P(1)P(2|1) - 2P(2)P(2|2)]P(2|2)^{m-2}P(>2|2) & \text{if } m > 1. \end{cases} \tag{9}$$

Using Eqs. (7)–(9) we have

$$\sum_{m=1}^{\infty} \{(m-1)[N_C(m) + N_H(m) + N_T(m)]\} = N(2).$$

Comparing this result[2] with Eq. (3) we see that the rings are already counted in the number of chains, as hinted at the end of Sec. IV A. This happens because, while computing the probability of chains, we ignore the fact that the presence of rings decreases the number of possible chains. For a large enough network, the number of rings should be small compared with the number of the other structures, validating the approximation.

Note that all expressions are proportional to $P(2|2)^m$, and therefore large chains should be exponentially rare, if they are not favored by the network growth.

V. THEORETICAL ANALYSIS FOR UNCORRELATED NETWORKS

For uncorrelated networks, where the degree at one side of an edge is independent of the degree at the other side of the edge, $P(k,k')$ can be factored as

[2] In these expressions and the following, we assume that the network is sufficiently large, such that the inclusion of some vertices in the chain does not affect the probabilities of reaching new vertices in the next step.

$$P(k,k') = \frac{kP(k)k'P(k')}{\langle k\rangle^2}. \tag{10}$$

The conditional probability is simplified to

$$P(k'|k) = \frac{k'P(k')}{\langle k\rangle}. \tag{11}$$

Using this last expression, we have for uncorrelated networks

$$N_R(m) = \frac{1}{m}\left(\frac{2P(2)}{\langle k\rangle}\right)^m, \tag{12}$$

$$N_C(m) = \frac{2^{m-2}NP(1)^2P(2)^{m-1}}{\langle k\rangle^m}, \tag{13}$$

$$N_T(m) = NP(1)\left(\frac{2P(2)}{\langle k\rangle}\right)^{m-1}\alpha, \tag{14}$$

$$N_H(m) = \frac{N\langle k\rangle}{2}\left(\frac{2P(2)}{\langle k\rangle}\right)^{m-1}\alpha^2. \tag{15}$$

where $\alpha=[1-P(1)/\langle k\rangle-2P(2)/\langle k\rangle]$.

A. Erdős-Rényi networks

Erdős-Rényi networks have no degree correlations and a Poissonian degree distribution:

$$P(k) = \frac{e^{-\langle k\rangle}\langle k\rangle^k}{k!}. \tag{16}$$

This gives the following expressions for the number of rings, cords, tails, and handles:

$$N_R(m) = \frac{\langle k\rangle^m e^{-m\langle k\rangle}}{m}, \tag{17}$$

$$N_C(m) = \frac{N}{2}\langle k\rangle^m e^{-(m+1)\langle k\rangle}, \tag{18}$$

$$N_T(m) = N\langle k\rangle^m e^{-(m+1)\langle k\rangle}\varepsilon, \tag{19}$$

$$N_H(m) = \frac{N}{2}\langle k\rangle^m e^{-(m+1)\langle k\rangle}\varepsilon^2, \tag{20}$$

where $\varepsilon=(e^{\langle k\rangle}-\langle k\rangle-1)$. Figure 4 shows a comparison of the results for networks with $N=10^6$ vertices and $L=972\,941$ edges (this number of edges was chosen to give the same average degree as for the scale-free network discussed below). A total of 1000 realizations of the model were used to compute the averages and standard deviations.

B. Scale-free networks

We now proceed to uncorrelated scale-free networks with degree distribution given as

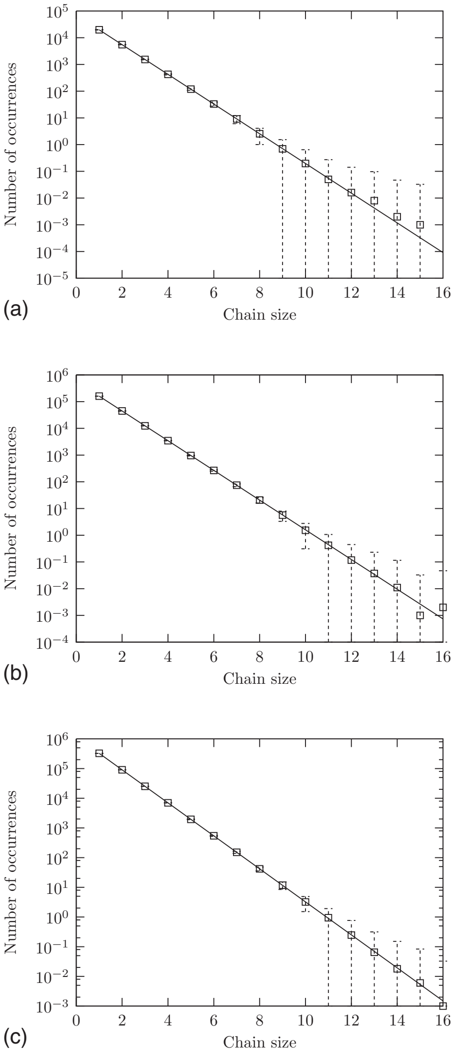


FIG. 4. Number of cords (a), tails (b), and handles (c) of different sizes in the model with Poisson degree distribution. The points are the averaged measured values (each of the error bars corresponds to one standard deviation); the lines are the values computed analytically. Note that the abrupt increase of the width of the error bars is a consequence of the logarithmic scale.

$$P(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}, \tag{21}$$

where γ is the power law coefficient and $\zeta(x)$ is the Riemann zeta function. This distribution describes a strictly scale-free network, with the power law valid for all values of k and a minimum $k_{\min}=1$. The results are therefore not directly applicable to scale-free real networks or models. The average degree is $\langle k\rangle=\zeta(\gamma-1)/\zeta(\gamma)$. The resulting expressions are

$$N_R(m) = \frac{2^{-m(\gamma-1)}}{m\zeta(\gamma-1)^m}, \tag{22}$$

$$N_C(m) = \frac{N}{2}\frac{2^{-(m-1)(\gamma-1)}}{\zeta(\gamma)\zeta(\gamma-1)^m}, \tag{23}$$

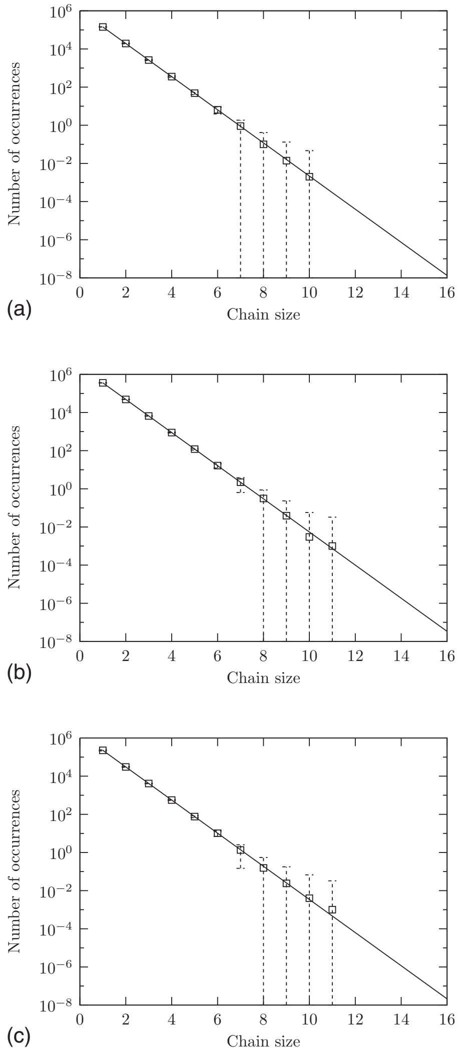


FIG. 5. Number of cords (a), tails (b), and handles (c) of different sizes in the model with scale-free degree distribution. The points are the averaged measured values (each of the error bars corresponds to one standard deviation); the lines are the values computed analytically.

$$N_T(m) = N\frac{2^{-(m-1)(\gamma-1)}}{\zeta(\gamma)\zeta(\gamma-1)^m}\beta, \qquad (24)$$

$$N_H(m) = \frac{N}{2}\frac{2^{-(m-1)(\gamma-1)}}{\zeta(\gamma)\zeta(\gamma-1)^m}\beta^2, \qquad (25)$$

where $\beta=[\zeta(\gamma-1)-1-2^{-(\gamma-1)}]^2$.

Figure 5 shows the comparison of the results for networks with $N=10^6$ vertices and $\gamma=2.5$. A total of 1000 realizations of the model were used to compute the averages and standard deviations. A comparison with Fig. 4 shows that the Poisson degree distribution with the same average degree presents larger chains. This is due to the relation between the constants in the exponential dependency on m: $\langle k\rangle/e^{\langle k\rangle}\approx 0.278$ for the Poisson model and $2^{1-\gamma}/\zeta(\gamma-1)\approx 0.135$ for the scale-free model.

The results presented in this section addressed the issue of validating the theory for analytical models. In Sec. VI, the theory is used to evaluate chains in real-world networks.

VI. REAL-WORLD NETWORKS

It is known that networks belonging to the same class may share similar structural properties [8,16]. So, to study the presence of handles in networks, we considered five types of complex networks, namely, social networks, information networks, word adjacency networks in books, technological networks, and biological networks.

A. Social networks

Social networks are formed by people or groups of people (firms, teams, economic classes) connected by some type of interaction, as friendship, business relationship between companies, collaboration in science, and participation in movies or sport teams [2], to cite just a few examples. Below we describe the social networks considered in our analysis.

Scientific collaboration networks are formed by scientists who are connected if they have authored a paper together. In our investigations, we considered the astrophysics collaboration network, the condensed matter collaboration network, the high-energy theory collaboration network, all collected by Newman from [38], and the scientific collaboration of complex networks researchers, also compiled by Newman from the bibliographies of two review articles on networks (by Newman [2] and Boccaletti *et al.* [17]). The astrophysics collaboration network is formed by scientists who post preprints on the astrophysics archive, between the years 1995 and 1999 [18]. The condensed matter collaboration network, on the other hand, is composed by scientist posting preprints on the condensed matter archive from 1995 until 2005 [18]. Finally, the high-energy theory collaboration network is composed by scientists who posted preprints on the high-energy theory archive from 1995 until 1999 [19,20].

B. Information networks

Roget's Thesaurus network is constructed by associating each vertex of the network to one of the 1022 categories in the 1879 edition of Peter Mark Roget's *Thesaurus of English Words and Phrases*, edited by John Lewis Roget [21]. Two categories i and j are linked if Roget gave a reference to j among the words and phrases of i, or if such two categories are directly related to each other by their positions in Roget's book [21]. This network is available at Pajek data sets [22].

Wordnet is a semantic network which is often used as a form of knowledge representation. It is a directed graph consisting of concepts connected by semantic relations. We collected the network from the Pajek data sets [22].

The world wide web (www) is a network of web pages belonging to the nd.edu domain linked together by hyperlinks from one page to another [23]. The data considered in our paper are available at the Center for Complex Network Research [24].

C. Word adjacency in books

Word adjacency in books can be represented as a network of words connected by proximity [25]. A directed edge is established between two words that are adjacent and its weight is the number of times the adjacent words appear in the text. Before constructing a network, the text must be preprocessed. All stop words (e.g., articles, prepositions, conjunctions, etc.) are removed, and the remaining words are lemmatized [25]. In our analysis, we considered the books *David Copperfield* by Charles Dickens, *Night and Day* by Virginia Woolf, and *On the Origin of Species* by Charles Darwin, compiled by Antiqueira *et al.* [26].

D. Technological networks

The internet or autonomous systems (AS) network is a collection of internet protocol (IP) networks and routers under the control of one entity that presents a common routing policy to the internet. Each AS is a large domain of IP addresses that usually belongs to one organization such as a university, a business enterpriser, or an internet service provider. In this type of network, two vertices are connected according to Border Gateway Protocol (BGP) tables. The considered network in our analysis was collected by Newman in July 2006 [27].

The U.S. airlines transportation network is formed by U.S. airports connected by flights in 1997. This network is available at Pajek data sets [22].

The western states power grid represents the topology of the electrical distribution grid [28]. Vertices represent generators, transformers, and substations, and edges the high-voltage transmission lines that connect them.

E. Biological networks

Some biological systems can be modeled in terms of networks, such as the brain, genetic interaction, and the interaction between proteins.

The neural network of Caenorhabditis elegans is composed of neurons connected according to synapses [28,29].

The transcriptional regulation network of Escherichia coli is formed by operons (an operon is a group of contiguous genes that are transcribed into a single mRNA molecule). Each edge is directed from an operon that encodes a transcription factor to another operon, which is regulated by that transcription factor. This kind of network plays an important role in controlling gene expression [7].

The protein-protein interaction network of Saccharomyces cerevisiae is formed by proteins connected according to identified directed physical interactions [30].

VII. RESULTS AND DISCUSSION

We analyzed the real-world networks by comparing their numbers of cords, tails, and handles with random networks generated by the rewiring procedure as described in [31] and with the theory proposed in Sec. IV.

A. Comparison between real-world networks and their randomized counterparts

For each considered real-world network, we generated 1000 randomized versions (100 for WWW) by the rewiring process described in [31]. The networks generated have the same degree distribution as the original, but without any degree-degree correlation. In order to compare the chain statistics obtained for the real-world and the corresponding randomized versions, we evaluated the Z-score values for each size of the cords, tails, and handles. The Z score is given by

$$Z = \frac{X_{\mathrm{real}} - \langle X \rangle}{\sigma}, \tag{26}$$

where X_{real} is the number of cords, tails, or handles with a specific size of the original (real-world) analyzed network, and $\langle X \rangle$ and σ are, respectively, the average and the standard deviation of the corresponding values of its randomized counterparts. A null value of the Z score indicates that there is no statistical difference between the number of occurrences of cords, tails, or handles in the considered network and in its randomized versions.

The results of the Z scores for all considered networks can be seen in Fig. 6. The cases in which the Z-score values are not defined ($\sigma = 0$) were not considered.

The majority of the results presented in Fig. 6 can be explained by the fact that the rewiring process tends to make uniform the distribution of cord, tail, and handle sizes. In this way, the excess of these structures on real networks will be reduced in the random counterparts. For instance, if a network has many large handles, its random version will present few large handles but many small ones. The next discussion will not take into account the shape of the distribution of chains, but just the most important results.

In the case of collaboration networks, there is a large quantity of cords. This fact suggests that researchers published papers with just one, two, or three other scientists. Cords may appear because many researchers can publish in other areas and, therefore, such papers are not included in the network. If other research areas had been considered, this effect would not occur and the number of small cords would be less significant. Thus, the presence of cords in collaboration networks can be the result of database incompleteness. Another possible cause of cords in such networks concerns the situations of authors who publish only among themselves.

The information networks do not present such well-defined patterns as observed in the collaboration network. The Roget thesaurus network is different from the others, but the results obtained for this network are not informative enough to be discussed. It is important to note that in the Wordnet and WWW, there is a large occurrence of tails of size 1. In the case of Wordnet, this happen because specific words have connections with more common words which have connections with the remainder of the network. In the case of the WWW, this structure is a consequence of characteristic URL documents which have just one link. In addition to small tails, the WWW has long tails and handles. This fact can be associated with the way in which the network were

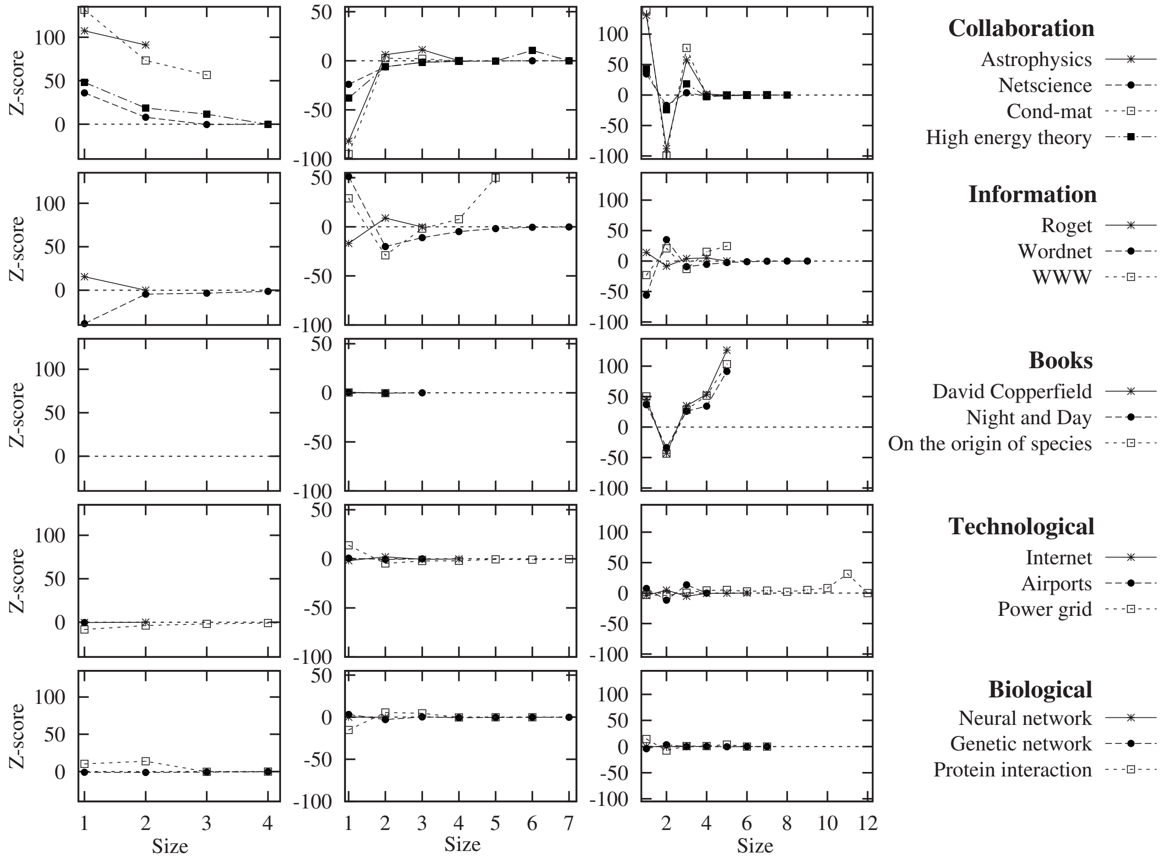


FIG. 6. Z scores of the number of cords (first column), tails (second column), and handles (third column) for each size. The number of generated random networks was 1000 for all considered networks, except for the WWW, for which it was 100 (because of the substantially larger size of this network).

constructed, by considering a *web crawler* [23]—a program designed to visit URL documents inside a given domain and get links between them in a recursive fashion. When pages are visited by the crawler, the wandering path can originate chains. If the program is not executed for a long time interval, long chains can appear. Thus, this effect can be the result of incomplete sampling (see Sec. VII C). In addition, as the process of network construction is recursive, isolated components do not occur in the database and therefore there are no cords and rings.

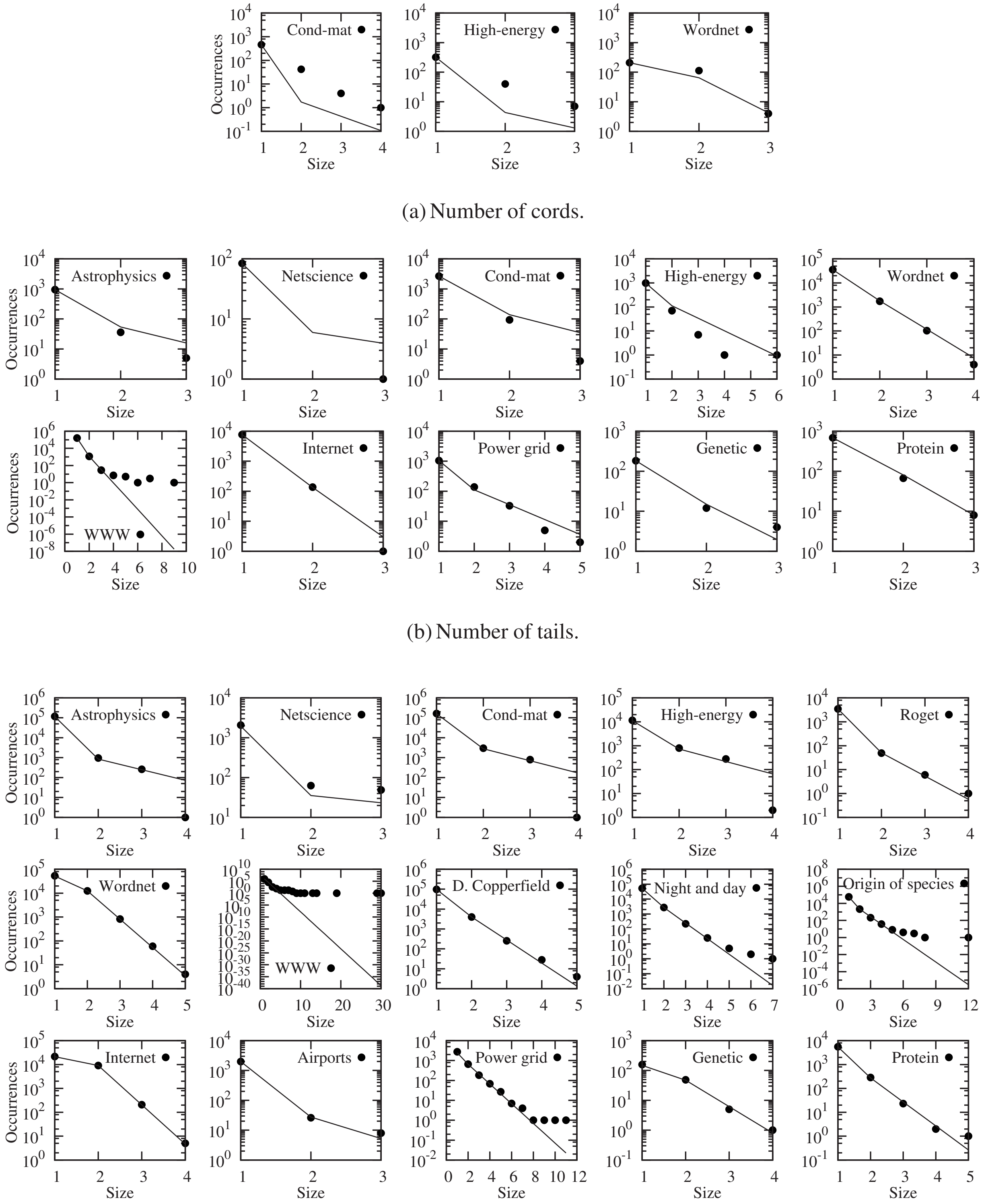
The book adjacency network presents a characteristic pattern of chains: no cords, the same quantity of tails of sizes 1, 2, and 3 as observed in the random counterparts, and many handles of sizes 1, 3, 4, and 5. The increase in the quantity of handles of size 2 in random versions is a consequence of the fact that, when the rewiring process is performed, many handles of size 1 can be put together. This fact explain why book networks present more handles of size 1 than random counterparts. On the other hand, the long handles are a consequence of the sequential process considered to obtain the network.

In technological networks, the chain patterns are more significant in the power grid. This network presents a high quantity of tails of size 1 and handles of size 11. While the first occurrence appears to be related to the geographical effect, where new vertices needed to cover a new region tend to connect with near vertices, the second can be the result of geographical constraints (e.g., the transmitters may be allocated in a strategic way in order to contour a mountain, lake, or other geographical accident).

The results obtained for biological networks are not so informative. However, the protein interaction network of the yeast *S. cerevisiae* has many cords of sizes 1 and 2. The presence of small cords in this network is a consequence of isolated chains of proteins which interact only with a small number of other proteins. This fact can be due to incompleteness [32], where many real connections may not be considered, or highly specialized proteins, which have lost many connections because of the mutation process—protein interaction networks evolve from two basic processes, duplication and mutation [33].

B. Theoretical analysis of real-world networks

Going back to the analysis presented in Sec. IV, we applied those theoretical developments to the real-world networks considered. We obtained their degree-degree correlations and computed the expected number of cords, tails, and handles as functions of their sizes by using Eqs. (7)–(9), respectively. The number of rings was not taken into account because of their very low probability of appearing in real-world networks. The results concerning the theoretical analysis are shown in Fig. 7. The cases not shown are those that have all chains smaller than 2. Due to the low probability of



(a) Number of cords.

(b) Number of tails.

(c) Number of handles.

FIG. 7. Distributions shown in (a), (b), and (c) correspond to the most significant data (each distribution has at least three points). Points correspond to the real data, and the solid lines correspond to the theoretical predictions.

finding cords in networks, only three networks are shown [Fig. 7(a)], namely, cond-mat, high-energy collaborations, and the Wordnet. The theoretical prediction does not work well for these networks, except for the Wordnet, predicting fewer cords than those found in the real networks. An opposite situation was found for the numbers of tails and handles, shown in Figs. 7(b) and 7(c), respectively. However, there are more larger tails and handles in the real-world networks than predicted by theory, except for the astrophysics, cond-mat, and high-energy collaboration networks.

Despite the fact that, for some cases, the number of small cords, tails, and handles of the real-world networks were far from the values obtained from their respective randomized counterparts (see Fig. 6), the theoretical results were accurate for several cases, except for the astrophysics (handles), net-science (tails), cond-mat (cords and handles), high-energy (cords, tails, and handles), WWW (tails and handles), the book *On the Origin of Species* (handles), and power grid (handles) networks (see Fig. 7). The good agreement between theoretical and experimental results indicates that in some networks chains are a consequence of degree correlations.

C. Analysis of incomplete networks

In order to investigate the possibility that incomplete networks presents many tails and handles, we sampled two theoretical network models, namely, the Erdős-Rényi (ER) [34] and the Barabási-Albert (BA) scale-free model [35] by performing random walks [36,37], and analyzing the corresponding distributions of tails and handles. The ER and BA models included 100 000 vertices with average degree 6. The results of the random walks in these theoretical networks are shown in Fig. 8. Each point of the mesh grid is the average value considering 1000 realizations.

For the ER and BA models the results are very similar, with the difference that the tails tend to vanish with larger random walks (almost 10^7 steps) in the BA model. This is not the case for the ER network because its original structure already had vertices with unit degree. Therefore, this network already had small tails (sizes 1 and 2). Conversely, BA networks of average vertex degree 6 do not have tails, and with large random walks these structures tend to vanish.

The results from Fig. 8 clearly indicate that there are many large tails and handles for both models when the random walks are relatively short. As the size of the random walks is increased, the number of large tails and handles tends to decrease, but the number of small tails and handles increases, because with large random walks the probability of breaking large tails and handles into smaller parts is increased. As the length of the random walks increases further, the large tails and handles tend to vanish, and the original networks are recovered.

VIII. CONCLUSIONS

One of the most important aspects characterizing different types of complex networks concerns the distribution of specific connecting patterns, such as the traditionally investigated motifs. In the present work we considered specific connecting patterns including chains of articulations, i.e., linear sequences of interconnected vertices with only two neighbors. This type of motif has been subdivided into cords (i.e., chains with free extremities), rings (i.e., chains with no free extremities but disconnected from the remainder of the network), tails (i.e., chains with only one free extremity), and handles (i.e., chains with no free extremity). By considering a large number of representative theoretical and real-world networks, we identified that many specific types of such net-

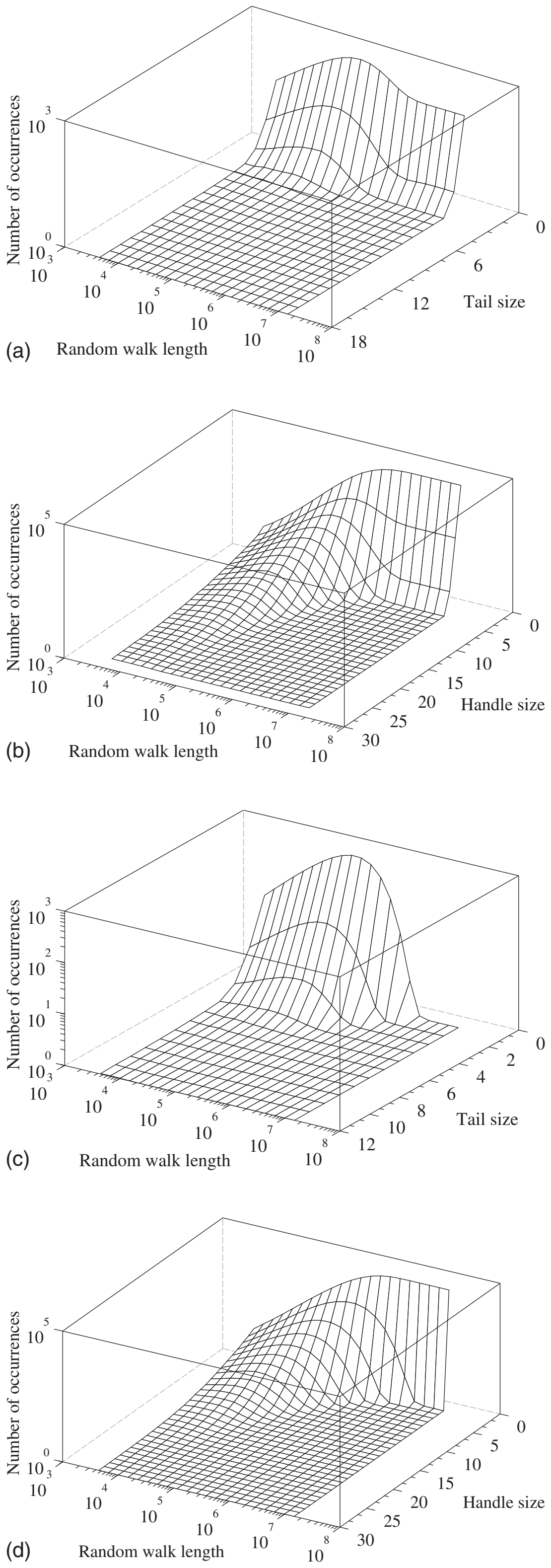


FIG. 8. (a) and (b) present the number of tails and handles of different sizes in the Erdős-Rényi model, respectively. (c) and (d), on the other hand, present the number of tails and handles for the Barabási and Albert scale-free model, respectively. Each point in the mesh grid is the average obtained by considering 1000 realizations of each random walk.

works tend to exhibit specific distributions of cords, tails, and handles. We provide an algorithm to identify such motifs in generic networks. Also, we developed an analytical framework to predict the number of chains in random network models, scale-free network models, and real-world networks, which provided accurate approximations for several of the considered networks. Finally, we investigated the presence of chains by considering *Z*-score values (i.e., comparing the presence of chains in real networks and the corresponding random counterparts). The specific origins of handles and tails are likely related to the evolution of each type of network, or incompleteness arising from sampling. In the first case, the handles and tails in geographical networks may be a consequence mainly of the chaining effect obtained by connecting vertices with are spatially near or adjacent to one another. In the second, we showed that incomplete sampling of networks by random walks can produce specific types of chains.

All in all, the results obtained in our analysis indicate that handles and tails are present in several important real-world networks, while being largely absent in the randomized versions. The study of such motifs is particularly important because they can provide clues about the way in which each type of network was grown. Several future investigations are possible, including the proposal of models for the generation of networks with specific distributions of handles and tails, as well as additional experiments aimed at studying the evolution of handles and tails in growing networks such as the WWW and the internet.

ACKNOWLEDGMENTS

The authors thank Lucas Antiqueira for providing the book networks. L.F.C. thanks CNPq (Grant No. 301303/06-1) and FAPESP (Grant No. 05/00587-5): F.A.R. is grateful to FAPESP (Grant No. 07/50633-9); P.R.V.B. is grateful to CNPq (Grant No. 141390/2004-2); and G.T. is grateful to FAPESP (Grant No. 03/08269-7).

[1] L. da F. Costa, F. A. Rodrigues, G. Travieso, and P. R. V. Boas, Adv. Phys. **56**, 167 (2007).

[2] M. E. J. Newman, SIAM Rev. **45**, 167 (2003).

[3] L. da F. Costa, Phys. Rev. Lett. **93**, 098702 (2004).

[4] L. da F. Costa and L. E. C. da Rocha, Eur. Phys. J. B **50**, 237 (2006).

[5] L. da F. Costa and F. Silva, J. Stat. Phys. **125**, 841 (2006).

[6] J. S. Andrade, Jr., H. J. Herrmann, R. F. S. Andrade, and L. R. da Silva, Phys. Rev. Lett. **94**, 018702 (2005).

[7] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, Nat. Genet. **31**, 64 (2002).

[8] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, Science **298**, 824 (2002).

[9] U. Alon, *An Introduction to Systems Biology: Design Principles of Biological Circuits* (Chapman & Hall/CRC, Boca Raton, FL 2007).

[10] I. Lodato, S. Boccaletti, and V. Latora, Europhys. Lett. **78**, 28001 (2007).

[11] L. da F. Costa, Int. J. Mod. Phys. C **15**, 175 (2004).

[12] M. Kaiser and C. Hilgetag, Biol. Cybern. **90**, 311 (2004).

[13] M. Kaiser and C. C. Hilgetag, Phys. Rev. E **69**, 036103 (2004).

[14] Z. Levnajić and B. Tadić, in *Proceedings of the International Conference on Computer Science*, edited by Yong Shi,G. Dick van Albada,Jack Dongarra andPeter M. A. Sloot, (Springer-Verlag, Berlin, 2007), pp. 633–640.

[15] M. E. J. Newman and M. Girvan, Phys. Rev. E **69**, 026113 (2004).

[16] M. E. J. Newman and J. Park, Phys. Rev. E **68**, 036122 (2003).

[17] S. Boccaletti, V. Latora, Y. Moreno, M. Chaves, and D.-U. Hwang, Phys. Rep. **424**, 175 (2006).

[18] M. E. J. Newman, Proc. Natl. Acad. Sci. U.S.A. **98**, 404 (2001).

[19] M. E. J. Newman, Phys. Rev. E **64**, 016131 (2001).

[20] M. E. J. Newman, Phys. Rev. E **64**, 016132 (2001).

[21] P. Roget and A. Robert, *Roget's Thesaurus of English Words and Phrases* (Longman, Harlow, U.K., 1982).

[22] V. Batagelj and A. Mrvar, http://vlado.fmf.uni-lj.si/pub/networks/data

[23] R. Albert, H. Jeong, and A.-L. Barabási, Nature (London) **401**, 130 (1999).

[24] A.-L. Barabási, http://www.nd.edu/~networks/resources.htm

[25] L. Antiqueira, M. Nunes, O. Oliveira, Jr., and L. da F. Costa, Physica A **373**, 811 (2007).

[26] L. Antiqueira, T. A. S. Pardo, M. das G. V. Nunes, and O. N. de Oliveira, Jr., Inteligencia Artificial, Revista Iberoamericana de IA, 11, 51 (2007).

[27] M. E. J. Newman, http://www-personal.umich.edu/~mejn/netdata

[28] D. J. Watts and S. H. Strogatz, Nature (London) **393**, 440 (1998).

[29] J. White, E. Southgate, J. Thomson, and S. Brenner, Philos. Trans. R. Soc. London, Ser. B **314**, 1 (1986).

[30] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai, Nature (London) **411**, 41 (2001).

[31] R. Milo, N. Kashtan, S. Itzkovitz, M. Newman, and U. Alon, e-print arXiv: cond-mat/0312028.

[32] J. Han, D. Dupuy, N. Bertin, M. Cusick, and M. Vidal, Nat. Biotechnol. **23**, 839 (2005).

[33] A. Vazquez, A. Flammini, A. Maritan, and A. Vespignani, ComPlexUs **1**, 38 (2003).

[34] P. Erdös and A. Rényi, Publ. Math. (Debrecen) **6**, 290 (1959).

[35] A.-L. Barabási and R. Albert, Science **286**, 509 (1999).

[36] J. D. Noh and H. Rieger, Phys. Rev. Lett. **92**, 118701 (2004).

[37] L. da F. Costa and G. Travieso, Phys. Rev. E **75**, 016102 (2007).

[38] http://www.arxiv.org