Border trees of complex networks

IOP Publishing Journal of Physics A: Mathematical and Theoretical

J. Phys. A: Math. Theor. **41** (2008) 224005 (12pp) doi:10.1088/1751-8113/41/22/224005

Border trees 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, PO Box 369, São Carlos, São Paulo 13560-970, Brazil

Received 8 October 2007, in final form 19 December 2007
Published 21 May 2008
Online at stacks.iop.org/JPhysA/41/224005

Abstract

The comprehensive characterization of the structure of complex networks is essential to understand the dynamical processes which guide their evolution. The discovery of the scale-free distribution and the small-world properties of real networks were fundamental to stimulate more realistic models and to understand important dynamical processes related to network growth. However, the properties of the network borders (nodes with degree equal to 1), one of its most fragile parts, remained little investigated and understood. The border nodes may be involved in the evolution of structures such as geographical networks. Here we analyze the border trees of complex networks, which are defined as the subgraphs without cycles connected to the remainder of the network (containing cycles) and terminating into border nodes. In addition to describing an algorithm for identification of such tree subgraphs, we also consider how their topological properties can be quantified in terms of their depth and number of leaves. We investigate the properties of border trees for several theoretical models as well as real-world networks. Among the obtained results, we found that more than half of the nodes of some real-world networks belong to the border trees. A power-law with cut-off was observed for the distribution of the depth and number of leaves of the border trees. An analysis of the local role of the nodes in the border trees was also performed.

PACS numbers: 89.75.Fb, 02.10.Ox, 89.75.Da, 87.80.Tq

1. Introduction

Complex networks are characterized by uneven distribution of connectivity which suggests that their growth is not governed by random events (e.g. [1]). Some special patterns in these networks, also known as network motifs, have been found to strongly affect dynamical aspects related to resilience, transport, network maintenance and even more specific functions of the networks. While the smaller motifs are believed to be the building blocks of complex networks [2], other motifs may appear as a consequence of specific network requirements and growth dynamics. For instance, chains of nodes [3, 4] can appear between two nodes or at the border of networks.

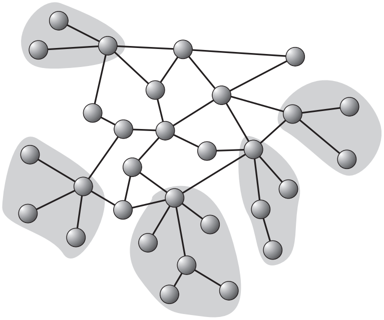

Figure 1. Some examples of border trees (gray regions) in a small network.

Although many network motifs have been largely characterized in the last few years (e.g. [2–5]), some remain uncharacterized, implying that their role in the network function is not known yet. One such example is the border tree, defined as subgraphs without cycles connected to the remainder of the network (see figure 1 for some examples). Such motifs (as well as other peripheral motifs) can be the result of the peripheral growth of the network, i.e. the network can evolve as a tree, where each 'branch' of nodes emerges from the main connected component toward the outside of the network.

Border trees are related to k-core decomposition (e.g. [6]) of the outmost layers. The k-core approach is an interesting way to describe the topology of real networks in terms of subgraphs. The k-core is obtained by removing from the network all vertices with degree smaller than k. This process is called *k-core decomposition*. After such a removal, the vertices in the resulting network that have degree lesser than k are removed and the network is analyzed again. When no further removal is possible, the non-empty resulting subgraph corresponds to the k-core of the original network [6]. Fundamental statistical properties of k-core are discussed by Dorogvtesev *et al* [6] and investigations about topology of the Internet using k-core decomposition are presented by Carmi *et al* [7]. Besides, protein interaction networks have been analyzed in terms of k-cores by Wuchty and Almaas [8], who considered the relation between k-cores and lethality. The k-core approach has also been applied in order to predict the function of proteins [9]. In the case of tree identification, if a network has no 1-shell, it means that no border trees can be found in its structure. Therefore, k-core decomposition can be considered as a preliminary and complementary investigation about the presence of border trees in networks. The number of nodes in the border trees can be computed as the size of the 1-shell plus the number of roots.

Besides the importance of understanding network growth, the study of border trees can also help to better characterize complex networks. In this work we provide an algorithm to find border tree motifs as well as a statistical method to characterize such structures through two measurements: their depth and number of leaves. The first property corresponds to the largest distance between the root (node which also belongs to a loop) and the leaves (nodes with degree 1) of a border tree. The other measurement is the number of leaves of the border tree, i.e. the number of possible paths from the root to the outer nodes. We applied the methodology in order to investigate the occurrence of such motifs in real-world networks as well as networks generated by traditional theoretical models.

The paper is organized as follows. We start by defining border trees and follow by presenting an algorithm for their detection, which is then applied for the characterization of several theoretical and real-world networks.

2. Border trees: basic concepts

In graph theory, a tree is defined as a graph in which every two vertices are connected by exactly one path. We define the *kernel* of any network as the main connected component whose vertices belong to at least one loop. The border tree is therefore a tree located at the periphery of the kernel, but with just one connection to it, provided by its root. Another way of defining the root of a border tree is as a vertex which also belongs to at least one loop. The leaves of a border tree are its extremities, i.e. vertices with unit degree. The largest distance between the root and the leaves of a border tree gives its depth, and the number of paths from its root to its leaves gives its number of branches, which is the same as its number of leaves. Vertices at the maximum distance from the root are considered to be at level 0.

In order to find the border trees in a given network, the following algorithm can be applied. We start by their leaves, going up to vertices of higher levels until their roots are reached. Thus, initially we find all vertices of degree 1 and start a tree from each of them. At this stage each tree has only one vertex and its neighbor, which is at one level higher, is added to the respective tree and becomes its root. The next stage is to recursively verify whether the vertex at the top of each tree has more than one neighbor, ignoring those at lower levels. If there are two or more neighbors, keep this tree in a waiting list. If there is just one, add it to the tree and join any other trees in the auxiliary list which has this vertex at its top. The algorithm ends when all trees are in the auxiliary list and the trees can no longer be joined. The isolated trees found by this algorithm are ignored and the resulting roots of all the trees are those vertices which also belong to at least one loop.

3. Databases

The models considered in this work are the Erdős and Rényi (ER) random graph [10], the Watts and Strogatz (WS) small-world model [11], Barabási and Albert (BA) scale-free model [1] and a geographical network (GN). The ER network model defines N vertices and a probability p of connecting each pair of such vertices. The degree distribution of networks generated by this model is a Poisson distribution. To construct a SW network, one starts with a regular lattice of N vertices in which each vertex is connected to κ nearest neighbors in each direction, resulting in 2κ connections. Next, each edge is randomly rewired with probability p. In this way, when $p = 0$ the graph is an ordered lattice with a high number of loops but large shortest distances and when $p \rightarrow 1$, the network becomes random with short distances but few loops. Note that when $p = 1$, WS networks do not become an ER network. Watts and Strogatz have shown that, in an intermediate case ($0 < p < 1$), both short distances and a large number of loops are present. The BA network model is based on two rules: (i) *growth*, the network is generated starting with a set of m_0 vertices; afterwards, at each step of the construction the network grows with the addition of new vertices with m edges; and (ii) *preferential attachment*, the vertices which receive the new edges are chosen following a *linear preferential attachment* rule, i.e. the probability of the new vertex i to connect with an existing vertex j is proportional to the degree of $j, \mathcal{P}(i \rightarrow j) = k_j \big/ \sum_u k_u$. Therefore, the most connected vertices have greater probability of receiving new vertices, which leaves to emergence of hubs. The GN model generates networks as described in [4] where N vertices are randomly distributed inside a $L = \sqrt{N}$ length square and two vertices are connected with probability $p \sim \mathrm{e}^{-\lambda d}$, where d is the geographical distance between them and λ is a model parameter chosen to generate the desired average vertex degree. Such model was initially proposed by Waxman [12] in 1998 to explain the Internet topology. Though this model tries to reproduce the connection structure between routers, the degree distribution is not a power law, but similar to that obtained for the ER network model.

All analyzed models had $N = 1000$ vertices and were designed to have average degrees $\langle k \rangle = 2$, 4 and 6. The probability of the rewiring process in the WS model was 0.2 and $\kappa = 1$, 2 and 3. For the GN model, the λ used was 1.7, 1.22 and 0.97. A total of 100 realizations of each model were considered.

We also considered 16 real-world networks divided into the following five classes.

(1) *Scientific collaboration networks*. Collaboration networks are formed by scientists that publish papers together. We considered four different networks: (i) astrophysics collaboration network, formed by scientists who posted preprints on the astrophysics archive, between the years 1995 and 1999 [13]; (ii) condensed matter collaboration network, composed by scientist posting preprints on the condensed matter archive from 1995 until 2005 [13]; (iii) high-energy theory collaboration network, composed by scientists who posted preprints on the high-energy theory archive from 1995 until 1999 [14, 15] and (iv) scientific collaboration of complex network researches compiled by Newman from the surveys [16, 17].

(2) *Information networks*. We considered three basic information networks: (i) Roget's thesaurus network [18, 19], formed by 1022 categories in the 1879 edition of Peter Mark Roget's Thesaurus of English Words and Phrases [18, 19], where two categories i and j are linked if Roget gives a reference to j among the words and phrases of i; (ii) Wordnet [19], composed of vertices, which represent concepts, and edges, which represent semantic relations between the concepts; and (iii) World Wide Web, which is a network of Web pages belonging to `nd.edu` domain connected by hyperlinks from one page to another [20][1].

(3) *Word adjacency networks in books*. In this kind of network, a directed edge is established between two words that are adjacent in the text. Stop words (e.g. articles, prepositions, conjunctions, etc) were removed and the remaining words were lemmatized [21]. We considered the books David Copperfield by Charles Dickens, Night and Day by Virginia Woolf and On the Origin of Species by Charles Darwin [21, 22].

(4) *Technological networks*. These networks are designed typically for distribution of some commodity or resource such as power or information. We considered three technological networks: (i) Internet at the level of autonomous systems (AS), where two AS are connected according to BGP (Border Gateway Protocols) tables posted by the University of Oregon Route Views Project[2]; (ii) the US Airlines Transportation Network, composed of US airports connected by flights [19] and (iii) the Western States Power Grid, composed of generators, transformers and substations connected by the high-voltage transmission lines [11].

(5) *Biological networks*. Biological interactions are naturally modeled by complex networks. We take into account the following networks: (i) the neural network of *Caenorhabditis elegans*, where neurons are represented by nodes and the synapses by edges [11, 23]; (ii) a transcriptional regulation network of the *Escherichia coli*, which is formed by operons (an operon is a group of contiguous genes that are transcribed into a single mRNA molecule), where each edge is directed from an operon that encodes a transcription factor to another operon which is regulated by that transcription factor; and (iii) a protein–protein interaction network of *Saccharomyces cerevisiae*, which is formed by proteins connected

[1] A L Barabási, Center for Complex Network Research, available at http://www.nd.edu/~networks/resources.htm.

[2] M E J Newman, Mark Newman's Network data, available at http://www-personal.umich.edu/~mejn/netd ata.

according to physical interactions. This network was constructed by Sprinzak *et al* [24] using the non-redundant databases of interacting proteins.

The originally directed networks were transformed into undirected versions by using the symmetrization method, which corresponds to reciprocating all directed edges.

4. Results and discussion

Each considered network had all border trees identified and separated, being subsequently measured with respect to their depth and number of leaves. Figures 2 and 3 present the distribution of depth and number of leaves found in the real networks and in the respective random counterparts. It is interesting to note that most distributions follow a power law with an exponential cut-off ($\mathrm{P(x)} \approx (\mathrm{x} + \mathrm{x}_0)^{\gamma} \mathrm{e}^{-(\mathrm{x}+\mathrm{x}_0)/\mathrm{x}_c}$) [25]. The WWW, Wordnet, protein interaction and AS presented power-law distributions ($\mathrm{P(x)} \approx (\mathrm{x} + \mathrm{x}_0)^{\gamma}$). This effect suggests that trees with large depth and number of leaves are very rare.

The importance of border trees is shown in table 1. As we can see, in collaboration networks, the border trees represent 11.2 to 25.2 % of the nodes in the network. In this case, the border trees tend to have small depth and small number of leaves. For information networks, specifically for the Wordnet and WWW, more than half of the nodes of these networks belong to border trees. Therefore, border trees are fundamental to define the structure of these networks. The same happens with biomolecular networks, namely genetic networks and protein interaction networks. For technological networks, the border trees are fundamental to define the structure of the Internet and the power grid. On the other hand, for book networks, the border trees represent a small fraction of the network. This effect is a result of the sequential process in which such networks are generated.

In order to know if a range of depths is important (the same applies to the number of leaves), a set of 1000 randomized networks were obtained by the rewiring process [26, 27] for each considered network. The rewiring starts with a network that already has the desired degree distribution, and then iteratively chooses two edges and interchanges the corresponding attached vertices. This process is frequently used in sociology [28, 29]. The border trees of each real network were found, measured and characterized in terms of the respective Z-scores. This statistical measurement is given as

$$Z = \frac{X_{\mathrm{Real}} - \langle X \rangle}{\sigma}, \tag{1}$$

where Z is the Z-score, X_{Real} is the number of border trees with a certain range of depths (the same for the number of leaves), and $\langle X \rangle$ and σ are, respectively, the average and the standard deviation of the randomized networks for the same range of depths. In the case of the theoretical networks, for each of their 100 realizations, 1000 randomized versions were created. The results concerning the Z-scores of the depth and the number of leaves of the considered networks are shown in table 2. Generally speaking, positive values of the Z-score index suggest that the network of interest contains more border trees than the respective randomized counterpart; the opposite is observed for negative values. Note that 95% of the randomized networks are comprised between Z-score values from -2 to 2. Therefore, only Z-scores with absolute values larger than 2 will be considered henceforth.

As can be seen in table 2, the rewiring procedure tends to eliminate the larger structures in the network, such as large border trees. The networks obtained after such a randomization tend to present a large quantity of small border trees, with few branches as well as a small quantity of large border trees with many branches. Z-score values smaller than -2 indicate that the

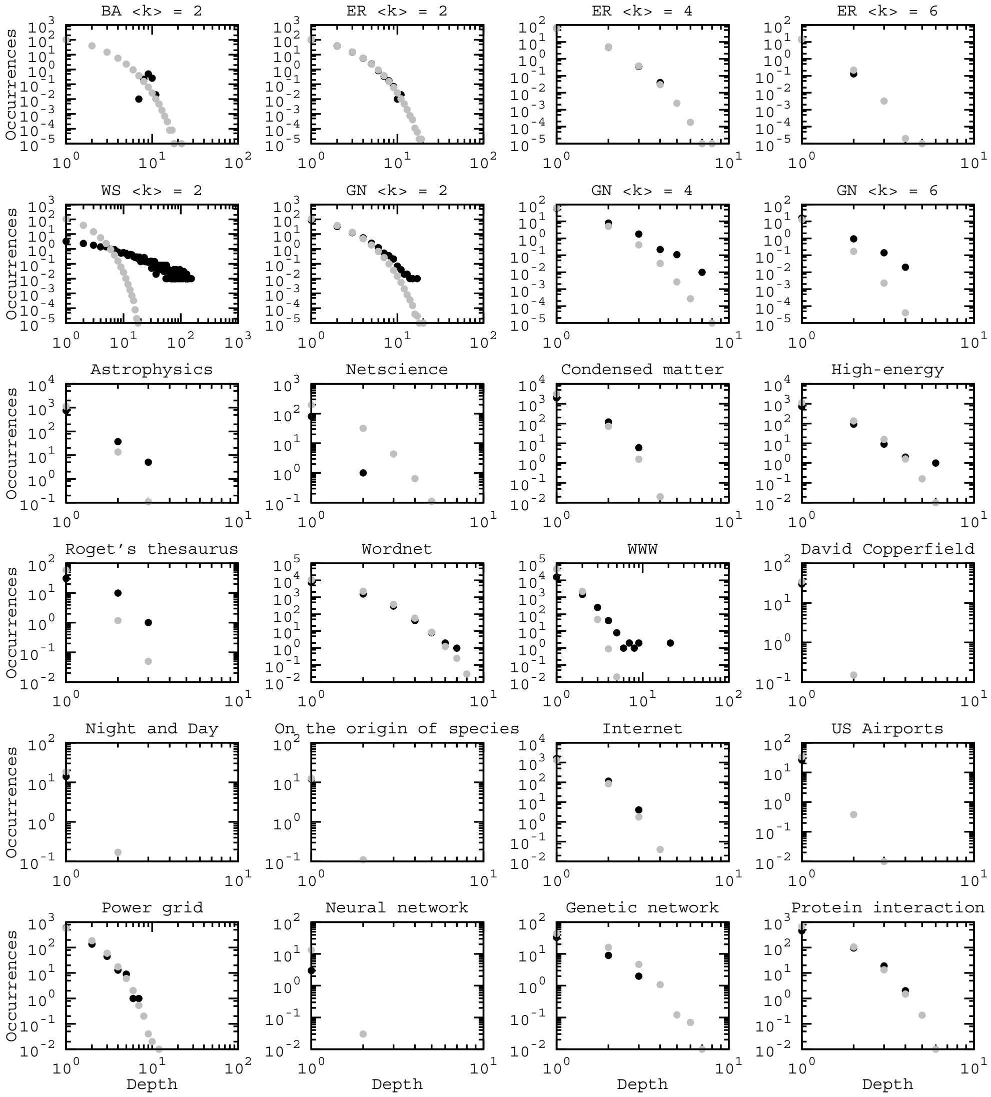

Figure 2. Distribution of depth of the border trees obtained for each real network (black points) and of the average of depth for the respective random counterparts (gray points).

respective network presents a total number of border trees which is substantially smaller than the respective randomized counterparts.

Among the analyzed models, significant *Z*-score values were obtained only for the BA networks with average degree 2, WS networks with average degree 2 and GN with average degrees of 2.08, 3.97 and 6.18. The BA network with average degree 2 is, by itself, organized as a large tree. In the case of the WS networks with average degree 2, the growth mechanism implies the creation of long chains, resulting in deep trees. At the same time, portions of the network are rewired giving rise to long trees with many leaves. The GN networks also present larger trees with many leaves as a consequence of border effects during the growth dynamics. GN networks start as isolated nodes inside a box, so that the nodes near the border tend to establish few connections, mainly with the nearest node. As a consequence, trees appear along

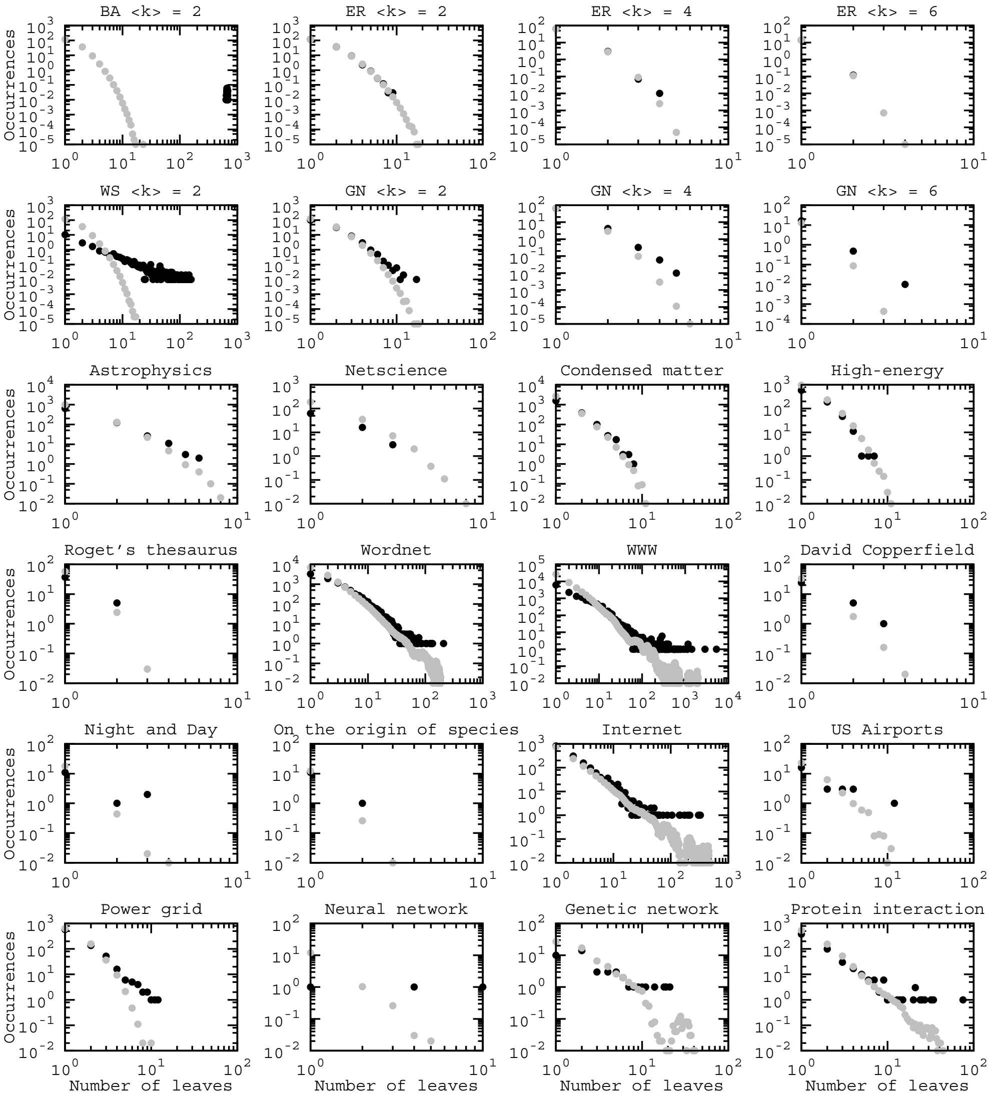

Figure 3. Distribution of number of leaves of the border trees obtained for each real network (black points) and of the average number of leaves for the random counterparts (gray points).

the border of the bounding box. The smaller the average degree of the network, the larger the probability of obtaining chains of nodes, implying the appearance of deep border trees with many leaves. The larger the average degree, the smaller the chance of getting deep border trees.

The majority of the real-world networks exhibit deeper and highly branched border trees than the respective random counterparts. In the case of the collaboration networks, this effect is observed particularly for the networks defined by collaborations between researchers from the areas of astrophysics, condensed matter and high-energy theory. However, the border trees of these networks are not particularly deep or branched (depths between 2 and 3 for astrophysics and condensed matter, and 6 for high theory energy; 3 to 6 leaves for astrophysics and from 2 to 8 leaves for condensed matter). A possible cause of these structures is the fact that the majority of researchers from these areas maintain collaborations with researchers from areas

Table 1. The statistics of border trees in complex networks. The column P_{nodes} gives the percentage of nodes in border trees, N_{trees} the number of border trees, N_{nodes} the average number of nodes in the border trees, depth the average depth of trees and N_{leaves} the average number of leaves of the border trees present in each network.

Class	Networks	Network size	P_{nodes}	N_{trees}	N_{nodes}	Depth	N_{leaves}
Models	BA $\langle k\rangle = 2$	1000	100.0%	1	1000 ± 0.0	9.06 ± 0.01	663.9 ± 0.1
	ER $\langle k\rangle = 2$	1000	51.0%	166	3.1 ± 0.0	1.62 ± 0.01	1.4 ± 0.0
	$\langle k\rangle = 4$	1000	14.4%	68	2.1 ± 0.0	1.08 ± 0.00	1.0 ± 0.0
	$\langle k\rangle = 6$	1000	2.9%	14	2.0 ± 0.0	1.01 ± 0.00	1.0 ± 0.0
	WS $\langle k\rangle = 2$	1000	95.2%	21	44.6 ± 1.1	13.89 ± 0.20	7.6 ± 0.2
	GN $\langle k\rangle = 2$	1000	47.0%	144	3.3 ± 0.0	1.69 ± 0.01	1.5 ± 0.0
	$\langle k\rangle = 4$	1000	16.0%	71	2.3 ± 0.0	1.18 ± 0.00	1.1 ± 0.0
	$\langle k\rangle = 6$	1000	3.6%	17	2.1 ± 0.0	1.07 ± 0.00	1.0 ± 0.0
Collaboration	Astrophysics	16 706	11.2%	797	2.3 ± 0.8	1.06 ± 0.26	1.3 ± 0.7
	Netscience	1461	12.7%	81	2.3 ± 0.6	1.01 ± 0.11	1.3 ± 0.5
	Cond-mat	40 421	12.6%	2095	2.4 ± 0.8	1.06 ± 0.25	1.4 ± 0.8
	High-energy	8361	25.2%	828	2.5 ± 0.9	1.15 ± 0.43	1.4 ± 0.7
Information	Roget	1022	9.9%	42	2.4 ± 0.7	1.29 ± 0.51	1.1 ± 0.3
	Wordnet	82 670	60.6%	9248	5.4 ± 7.5	1.25 ± 0.55	4.1 ± 6.8
	WWW	325 729	56.2%	17 070	10.7 ± 60.5	1.13 ± 0.47	9.5 ± 60
Books	David Copperfield	11 378	0.6%	30	2.2 ± 0.5	1.00 ± 0.00	1.2 ± 0.5
	Night and day	7959	0.4%	14	2.4 ± 0.7	1.00 ± 0.00	1.4 ± 0.7
	On the origin of species	6973	0.4%	12	2.1 ± 0.3	1.00 ± 0.00	1.1 ± 0.3
Technological	Internet (AS)	22 963	42.3%	1712	5.7 ± 15.9	1.07 ± 0.27	4.6 ± 15.8
	US airports	332	24.4%	26	3.1 ± 2.3	1.00 ± 0.00	2.1 ± 2.3
	Power grid	4941	48.4%	805	3.0 ± 2.0	1.39 ± 0.80	1.5 ± 1.2
Biological	Neural network	297	6.1%	3	6.0 ± 4.6	1.00 ± 0.00	5.0 ± 4.6
	Genetic network	423	63.8%	44	6.1 ± 5.4	1.30 ± 0.55	4.6 ± 5.0
	Protein interaction	4135	49.5%	570	3.6 ± 5.3	1.25 ± 0.52	2.3 ± 4.6

not considered in the construction of the collaboration networks used in this work. In other words, the border trees of these networks are partly a consequence of the incompleteness of the data. Also observe that border trees arise from papers with two authors, but not all papers with two authors became a border tree.

The presence of a considerable number of deep and branched border trees was also verified for the information networks, though in a lesser scale than for Roget's thesaurus. The tree heights and extensive branching are more expressive in the Wordnet and WWW. In the latter case, there is a large number of deep border trees and branches (between 9 and 5324). Such a connectivity structure is likely a consequence of the sampling of the WWW (see [4]). More specifically, the leaves are likely to have connections with sites outside the analyzed network.

The book networks present some intrinsic properties in the sense that there are a very small number of border trees, presenting few leaves (2 or 3). Such properties are a direct consequence of the way in which these networks were constructed. More specifically, all pairs of subsequent words in a paragraph generate a respective edge; however, no edges are generated when proceeding from a paragraph to the next. Therefore, the sequential linking of the nodes promotes the generation of several loops, while the trees are a consequence of the paragraphs' structure.

The technological networks are also substantially distinct from the others as far as the border trees are concerned. The Internet and power grid network present a particularly large

Table 2. Statistical evaluation of border trees in the considered networks. N_{real} indicates the number of trees encountered in each network with a given depth or number of leaves, N_{rand} is the corresponding values for average of the number of trees in the random counterparts and SD is the standard deviation. The ranges correspond to depths (or the number of leaves) which depart significantly from the respective randomized versions (positive and negative values indicate that the border trees are above or below the random counterparts).

	Depth				Number of leaves			
Networks	Range	N_{real}	$N_{\text{rand}} \pm SD$	Z-score	Range	N_{real}	$N_{\text{rand}} \pm SD$	Z-score
				Models				
BA $\langle k \rangle = 2$	1–7	0	166.0 ± 0.8	−201.1	1–634	0	166.3 ± 0.8	−203.1
	8–11	1	0.2 ± 0.1	14.6	635–684	1	0.0 ± 0.0	ND
WS $\langle k \rangle = 2$	1–5	10	164.7 ± 0.8	−186.3	1–5	16	165.8 ± 0.8	−184.6
	6–152	11	1.6 ± 0.1	77.9	6–157	5	0.5 ± 0.1	65.0
GN $\langle k \rangle = 2$	1–3	134	159.1 ± 0.9	−28.3	1–2	131	156.4 ± 0.9	−27.5
	4–20	11	7.8 ± 0.3	11.0	3–18	13	10.5 ± 0.3	8.7
GN $\langle k \rangle = 4$	1–1	61	63.8 ± 0.7	−4.6	1–1	66	66.3 ± 0.7	−0.4
	2–8	10	5.5 ± 0.2	20.6				
GN $\langle k \rangle = 6$	1–4	17	12.9 ± 0.3	11.8	1–4	17	12.9 ± 0.3	11.8
				Collaborations				
Astrophysics	1	755	1142 ± 13	−30.8	1–2	755	1128 ± 16	−23.8
	2–3	42	13.7 ± 3.4	8.3	3–6	42	28.2 ± 4.8	2.9
Netscience	1–5	81	233.1 ± 7.4	−20.5	1–8	81	233.1 ± 7.4	−20.5
Cond-mat	1	1971	3062 ± 23	−47.6	1	1561	2666 ± 31	−36.2
	2–3	124	73.2 ± 8.6	5.9	2–8	534	468.5 ± 13.2	5.0
High-energy	1–5	827	1277 ± 16	−29.1	1–6	827	1276 ± 16	−28.8
	6–6	1	0.0 ± 0.1	9.9	7–7	1	0.5 ± 0.7	0.7
				Information				
Roget	1	31	58.7 ± 2.1	−13.5	1	37	57.5 ± 2.9	−7.1
	2–3	11	1.2 ± 1.1	9.0	2	5	2.4 ± 1.3	2.0
Wordnet	1–5	9245	13 304 ± 51	−79.2	1–3	6323	11 041 ± 62	−76.2
	6–7	3	1.5 ± 1.3	1.2	4–208	2925	2264 ± 23	28.7
WWW	1–2	16 762	48 108 ± 108	−290.9	1–8	13 402	45 449 ± 115	−278
	3–21	308	49.6 ± 6.2	41.9	9–5324	3668	2709 ± 24	40.3
				Books				
David Copperfield	1–2	30	34.9 ± 1.5	−3.3	1	24	33.0 ± 2.8	−3.2
					2–3	6	1.9 ± 1.3	3.0
Night and Day	1–2	14	18.5 ± 0.7	−6.3	1	11	18.0 ± 1.3	−5.2
					2–3	3	0.5 ± 0.6	4.0
On the origin of species	1–2	12	12.7 ± 0.6	−1.3	1	11	12.4 ± 1.1	−1.3
					2	1	0.3 ± 0.5	1.5
				Technological				
Internet (AS)	1–3	1712	1427 ± 22	12.9	1–214	1710	1423 ± 22	13.0
					215–492	2	4.2 ± 0.8	−2.9
US airports	1–3	26	34.0 ± 2.8	−2.8	1–2	19	29.4 ± 3.7	−2.8
					3–12	7	4.6 ± 1.3	1.8
Power grid	1–4	794	836 ± 14	−3.1	1–2	715	797 ± 16	−5.3
	5–5	9	6.1 ± 2.5	1.2	3–12	90	48.5 ± 5.6	7.4
				Biological				
Neural network	1–2	3	13.3 ± 1.4	−7.5	1–3	1	13.2 ± 1.5	−8.4
					4–10	2	0.1 ± 0.2	8.9
Genetic network	1–7	44	65.3 ± 4.3	−5.0	1–8	37	62.0 ± 4.6	−5.5
					9–21	7	2.4 ± 1.2	3.9
					22–41	0	1.0 ± 0.2	−5.7
Protein interaction	1–2	549	771 ± 15	−14.7	1–19	560	784 ± 14	−15.7
	3–4	21	14.7 ± 3.4	1.9	20–75	10	1.3 ± 0.5	16.1

number of small trees and just a few deep trees. The Internet also exhibits many border trees with few leaves. This phenomenon can be explained by the jelly-fish structure of this network, which includes many nodes connected to the central kernel [30]. The power grid network presents many shallow border trees with many leaves (star-like trees). This seems to be related to the fact that the new nodes needed to cover a new region tend to be connected to the nearest existing node. A similar organization was verified for the airport network.

The biological networks do not seem to exhibit a well-defined pattern of border trees. Several of these networks have many highly branched border trees, though the protein–protein interaction network incorporates an expressive number of border trees with depth 2 or 3. The considered neuronal network presents features similar to those of the geographical theoretical model. In the case of the genetic transcription network, the presence of many small trees with many ramifications is a consequence of the fact that some genes participate in the regulation of a large quantity of other genes. A similar situation is observed for the protein–protein interaction network.

Observe that several networks such as the WWW, Wordnet, BA and WS (average degree 2) yielded particularly small Z-score values (for both depth and number of leaves). This effect can be explained by the fact that the original networks contained large and deep trees (see figures 2 and 3), which were split into many smaller trees by the rewiring procedure.

4.1. Local analysis of border trees

In some of the considered networks, the vertices are identified by labels and therefore it is possible to make a functional analysis of the border trees. These networks are the protein–protein interaction, the US air transportation network, Roget's thesaurus network and the Wordnet.

In the case of protein–protein interactions, the border trees are composed of proteins that have a similar function along the trees, where proteins with similar functions tend to be connected, as suggested by the majority rule [31]. On the other hand, proteins emerging from different branches tend to have distinct functions. Also, the root proteins tend to be less specific than the proteins at the leaves. For instance, the root protein P33418 (involved in pre-tRNA splicing) is connected to protein P46672, which binds specifically G4 quadruplex nucleic acid structures. This protein is linked to P00958 and P46655 proteins, which form a complex with glutamyl-tRNA synthetase and increase the catalytic efficiency of tRNA synthetases [32]. The protein P46655 is connected to proteins P21957 and P34246. The latter protein is connected to proteins P21957 and P34246, which are both uncharacterized proteins (see figure 4(*a*)). Let us now consider the tree whose root is the protein P43609, which is a component of the chromatin structure-remodeling complex—involved in transcription regulation and nucleosome positioning. This protein is connected to proteins P32832, P47102 and Q03124. The protein P47102 is connected to P39993. These two proteins have similar functions: they activate the ARF proteins by exchanging bound GDP for free GTP [32]. The protein Q03124, which is also a component of the chromatin structure-remodeling complex, is connected to P53101, which converts cystathionine into homocysteine (see figure 4(*b*)). Therefore, we can see from these examples that the root proteins are more general than the proteins at the leaves. Since proteins at the same level in the tree do not share connections, they tend to be different with respect to their functions.

The US air transportation networks is basically composed of three types of airports: international, regional and small airports. The border trees tend to have many leaves and low depth (every border three has depth one), which suggests that airports with a small number of links tend to be linked with a more connected one, instead of to have connections between

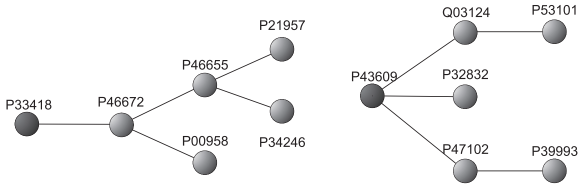

Figure 4. Example of border trees present in the considered protein–protein interaction network. The root proteins are indicated by dark gray nodes.

them, which contributes to make the network small world ($\ell = 2.74$). We observed that most root airports tend to be international and the airports at the border, regional or small. In fact, 20 of the 26 roots are international airports. Among the exceptions, the Bethel Airport, localized in Alaska, has properties of international airports, as it has connections to airports in other states of US. Likewise, among the 55 airports at the leaves, only 5 are international. Therefore, there is a high relation between the importance of the airports and their position on the border trees.

In Roget's thesaurus network, two words i and j are connected whenever they are directly related. Therefore, the words in the border trees tend to be very specific, irrespective of the remainder of the network. Also, just the words in the same branches are associated. For instance, the border tree whose root is the word 'demon' is connected to words 'Jupiter' and 'Satan'. The latter two words are not related. The word 'Satan' is connected to the word 'angel', which has no association with 'Jupiter'. Most of the border trees in Roget's thesaurus network are trees without branches. These trees are also known as tails [4].

The Wordnet is another type of information network. Again, the same effect discussed before tends to be observed: words at the same level of the tree tend not to be related. On the other hand, words at the same branch tend to have similar functions. A good example is the border tree associated with the word 'sport'. At the first level, this word is connected to 'archery', 'team sport', 'cycling', 'nonresident', 'sledding', 'skating' and 'racing'. As we can see, these words have no relation between them, but are associated with sports. Among the ramifications, the word 'cycling' is connected to 'bicycling', 'motorcycling' and 'dune cycling'. The word 'skating' is connected to 'roller skating', 'skateboarding' and 'ice skating'. Therefore, words in different branches tend to have no semantic association.

Thus, border trees can be seen as structures whose leaves have functions which are more specific than those of the roots. Therefore, while hubs are very general structures on the networks, having associations with many other nodes, the leaves represent an opposite situation.

5. Conclusions

This work has introduced the concept of border trees and presented a simple and effective algorithm for their identification. Statistics of the presence of such motifs in several real-world and theoretical networks were obtained and shown to provide valuable information regarding the general structure of the analyzed networks. Unlike recent results obtained for chain motifs [4], border trees were found in both theoretical and real-world networks. Among the former, we obtained the largest tree for the BA with average degree equal to 2, while the WS model exhibited the largest depth values. In the case of the real-world networks, the WWW presented the largest overall measurements, suggesting that this network involves a larger number of significant border trees, possibly corresponding to the more recently included nodes. The Internet and power-grid network (a geographical structure) presented similar properties, though exhibiting the shortest depths. Though no well-defined, systematic pattern was identified in the case of the biological networks, some specific structures emerged with important implications. These concern a large number of shallow border trees with many branches observed for the protein–protein interaction networks. A more in-depth study of these structures has potential for unveiling important properties of the protein trees. We also observed that the distribution of the depth, number of leaves and number of nodes of the border trees follow a power-law with an exponential cut-off. The local analysis of border trees show that the nodes in the periphery of the trees tend to be the most specific in the network. In this way, the study of border trees can help in the investigations about the structure and function in complex networks.

Acknowledgments

The authors thank Lucas Antiqueira for providing the book networks. Luciano da F Costa thanks CNPq (301303/06-1) and FAPESP (05/00587-5). Francisco A Rodrigues is grateful to FAPESP (07/50633-9) and Paulino R Villas Boas is grateful to CNPq (141390/2004-2).

References

[1] Barabási A-L and Albert R 1999 *Science* **286** 509
[2] Shen-Orr S S, Milo R, Mangan S and Alon U 2002 *Nature Genet.* **31** 64
[3] Guimerà R, Sales-Pardo M and Amaral L 2007 *Nature Phys.* **3** 63
[4] Villas-Boas P R, Rodrigues F A, Travieso G and Costa L da F 2007 *Preprint* 0706.2365
[5] Huang C, Sun C, Cheng C and Hsieh J 2007 *Physics* A **377** 340
[6] Dorogovtsev S N, Goltsev A V and Mendes J F F 2006 *Phys. Rev. Lett.* **96** 40601
[7] Carmi S, Havlin S, Kirkpatrick S, Shavitt Y and Shir E 2006 *Proc. Natl. Acad. Sci. USA* **104** 11150
[8] Wuchty S and Almaas E 2005 *Proteomics* **5** 444
[9] Altaf-Ul-Amin M, Nishikata K, Koma T, Miyasato T, Shinbo Y, Arifuzzaman M, Wada C, Maeda M, Oshima T and Mori H 2003 *Genome Inf.* **14** 498
[10] Erdős P and Rényi A 1959 *Publ. Math.* **6** 290
[11] Watts D J and Strogatz S H 1998 *Nature* **393** 440
[12] Waxman B 1988 *IEEE J. Sel. Areas Commun.* **6** 1617
[13] Newman M E J 2001 *Proc. Natl Acad. Sci.* **98** 404
[14] Newman M E J 2001 *Phys. Rev.* E **64** 16131
[15] Newman M E J 2001 *Phys. Rev.* E **64** 16132
[16] Newman M E J 2003 *SIAM Rev.* **45** 167
[17] Boccaletti S, Latora V, Moreno Y, Chaves M and Hwang D-U 2006 *Phys. Rep.* **424** 175
[18] Roget P and Robert A 1982 *Roget's Thesaurus of English Words and Phrases* (Essex: Longman Harlow)
[19] Batagelj V and Mrvar A 2006 *Pajek datasets* http://vlado.fmf.uni-lj.si/pub/networks/data
[20] Albert R, Jeong H and Barabási A-L 1999 *Nature* **401** 130
[21] Antiqueira L, Nunes M, Oliveira Jr O and Costa L F 2007 *Physics* A **373** 811
[22] Antiqueira L, Pardo T A S, Nunes M G V, de Oliveira Jr O N and Costa L da F 2006 *4th Workshop in Information and Human Language Technology*
[23] White J, Southgate E, Thomson J and Brenner S 1986 *Phil. Trans. R. Soc.* B **314** 1
[24] Sprinzak E, Sattath S and Margalit H 2003 *J. Mol. Biol.* **327** 919
[25] Costa L da F, Rodrigues F A and Travieso G 2006 *Appl. Phys. Lett.* **89** 174101
[26] Kannan R, Tetali P and Vempala S 1999 *Random Struct. Algorithms* **14** 293
[27] Milo R, Kashtan N, Itzkovitz S, Newman M and Alon U 2003 *Preprint* cond-mat/0312028
[28] Holland P and Leinhardt S 1976 *Sociol. Methodol.* **7** 1
[29] Roberts J 2000 *Soc. Netw.* **22** 273
[30] Faloutsos M, Faloutsos P and Faloutsos C 1999 *Proc. Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communication* pp 251–62
[31] Schwikowski B, Uetz P and Fields S 2000 *Nature Biotechnol.* **18** 1257
[32] Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel R and Bairoch A 2003 *Nucl. Acids Res.* **31** 3784