Applied Mathematics and Computation 218 (2012) 10393–10405

Contents lists available at SciVerse ScienceDirect

Applied Mathematics and Computation

journal homepage: www.elsevier.com/locate/amc

Distance-sum heterogeneity in graphs and complex networks

Ernesto Estrada *, Eusebio Vargas-Estrada

Department of Mathematics and Statistics, University of Strathclyde, 26 Richmond Street, Glasgow G1 1XQ, UK

ARTICLE INFO

Keywords:
Distance-sum
Complex networks
Balaban index
Graph distances
Distance distributions

ABSTRACT

The heterogeneity of the sum of all distances from one node to the rest of nodes in a graph (distance-sum or status of the node) is analyzed. We start here by analyzing the cumulative statistical distributions of the distance-sum of nodes in random and real-world networks. From this analysis we conclude that statistical distributions do not reveal the distance-sum heterogeneity in networks. Thus, we motivate an index of distance-sum heterogeneity based on a hypothetical consensus model in which the nodes of the network try to reach an agreement on their distance-sum values. This index is expressed as a quadratic form of the combinatorial Laplacian matrix of the network. The distance-sum heterogeneity index $\varphi(G)$ gives a natural interpretation of the Balaban index for any kind of graph/network. We conjecture here that among graphs with a given number of nodes $\varphi(G)$ is maximized for a graph with a structure resembling the agave plant. We also found the graphs that maximize $\varphi(G)$ for a given number of nodes and links. Using this index and a normalized version of it we studied random graphs as well as 57 real-world networks. Our findings indicate that the distance-sum heterogeneity index reveals important structural characteristics of networks which can be important for understanding the functional and dynamical processes in complex systems.

1. Introduction

The study of complex networks has become one of the fastest growing areas of interdisciplinary research in the XXI century [1,2]. In a complex network nodes represent entities and links represent interactions among these entities in a complex system. Examples of these networks are ubiquitous in natural (molecular, cellular, ecological) and man-made (social, technological, infrastructural) systems [3]. One important challenge for the study of complex networks is that many techniques developed for the analysis of small graphs are computationally intractable for gigantic complex networks found in the real-world. On the other hand, some statistical approaches developed so far for the analysis of these huge networks are not applicable to small graphs. An example of the last situation is the analysis of degree heterogeneity in networks [4], which is frequently carried out by studying the distributions of node degrees [5–7]. In large networks it is possible to analyze the distribution of the probabilities $p(\delta)$ of finding a node with degree δ as a function of the node degree. However, in a small graph there is not enough data points as for having a good fit for these distributions. Other difficulties found in studying degree distributions include the selection of the best fit, and the way to compare the heterogeneity of networks with different types of distributions [4,7].

Statistical distributions of other graph-theoretic parameters have also been studied for complex networks, such as the eigenvalue distributions [8] and the node–node distance distribution [9–11]. The distance-based analogous of the node degree distribution in a network is the distance-sum distribution. This kind of distribution has not previously been studied for

* Corresponding author.
E-mail address: ernesto.estrada@strath.ac.uk (E. Estrada).

0096-3003/$ - see front matter © 2012 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.amc.2012.03.091

networks. It consists of the distribution of the probabilities $p(s)$ of finding a node with distance-sum equal to s, where s is the sum of all distances from a given node (see further section for formal definitions). The distance-sum is an important characterization of a node that can be found in many graph-theoretic invariants. For instance, the Wiener [12] and Balaban [13] indices which are frequently used for the analysis of molecular graphs [14], and the average path length [15] and the closeness centrality [16] of a node which are commonly applied to the analysis of complex networks [4], are all based on the distance-sum.

In a similar way to the analysis of any kind of statistical distributions for the nodes of a graph, distance-sum distributions are difficult to find for small graphs where the number of data points is very scarce as well as the other difficulties mentioned before. In those cases where the distributions can be found the previously mentioned difficulties for analyzing the heterogeneity of networks also apply to the analysis of distance-sums. Consequently, we propose here the derivation of an index quantifying the distance-sum heterogeneity of a graph/network in such a way that it can be applicable for a graph of any size. The paper is organized as follows. First we give the preliminary definitions needed for the rest of the paper. Then, we introduce the analysis of the distance-sum distribution and show some examples of them for random and real-world networks of different sizes. In the next section we motivate and introduce an index of distance-sum heterogeneity and relate it to the well-known Balaban index [14] of a graph. We continue by developing a spectral representation of this index on the basis of the Laplacian spectra of the corresponding graphs. In the next two sections we illustrate the results of the distance-sum heterogeneity index for random and real-world networks with different topologies. The work is finished with some conclusions about the applications of this index for the analysis of complex networks.

2. Preliminary definitions

Let $G = (V,E)$ be a simple, undirected and unweighted graph having $n = |V|$ vertices or nodes and $m = |E|$ links or edges. The adjacency matrix **A** of the graph G is a square, symmetric matrix whose entries are $A_{i,j} = 1$ if $\{i,j\} \in E$ and $A_{i,j} = 0$ otherwise. The degree of the node i is given by $\delta_i = \sum_{j=1}^{n} A_{ij}$. The density of a graph is defined as $d = 2m/n(n-1)$ and the Laplacian matrix of the graph is defined as $\mathbf{L} = \mathbf{\Delta} - \mathbf{A}$, where $\mathbf{\Delta}$ is the diagonal matrix of node degrees. This matrix is positive semidefinite with eigenvalues $0 = \mu_1 < \mu_2 \leqslant \cdots \leqslant \mu_n$ for a connected graph.

A *path* of length l between v_1 and v_{l+1} is any sequence of nodes $v_1, v_2, \ldots, v_l, v_{l+1}$ such that for each $i = 1,2,\ldots,l$ there is a link from v_i to v_{i+1} and all the nodes (and all the edges) are distinct. Among all the paths between v_1 and v_{l+1} the ones having the minimum length are called *shortest-paths*. The length of a shortest path between v_i and v_j is called the shortest-path distance (or simply the distance) between nodes v_i and v_j, and denoted by $d_{i,j}$. The distance matrix **D** of the graph $G = (V,E)$ is a square, symmetric $n \times n$ matrix whose i,j entry is given by $d_{ij} = d(i,j)$. The *status* or *distance-sum* $s(i)$ of a node i in G is the sum of all distances from i to every other node in G [17]. That is,

$$s(i) = \sum_{j \in V(G)} d(i,j). \tag{1}$$

A vector of distance-sums can be obtained as

$$\mathbf{s} = \mathbf{1}^T \mathbf{D}, \tag{2}$$

where **1** is a column vector of ones.

As mentioned in the Introduction, the distance-sum is the basis for several graph-theoretic invariants, such as the Wiener [12] and Balaban [13] indices, average shortest path [15] and closeness centrality [16]. The Wiener index is defined as follow [12,18]

$$W(G) = \sum_{i} \sum_{j>i} d(i,j) = \frac{1}{2} \sum_{i=1}^{n} s(i) \tag{3}$$

The Balaban index is defined as [13]

$$J(G) = \frac{m}{C+1} \sum_{(i,j) \in E} (s_i s_j)^{-1/2}, \tag{4}$$

where $C = m - n + 1$ is the cyclomatic number.

The average path length is defined as [3]

$$\bar{l} = \frac{\sum_{i=1}^{n} s(i)}{n(n-1)} = \frac{2W(G)}{n(n-1)}. \tag{5}$$

The so-called 'small-world' effect is present in a given network when $\bar{l}$ is small compared to the size of the network, i.e., $\bar{l} \sim \ln n$ [15]. The small-world effect impacts directly on the properties of networked systems and dynamical processes in networks, particularly those related with communications and synchronization [19].

Another graph-theoretic measure related to the distance-sum of a given node is the closeness centrality, which is defined as:

$$CC(i) = \frac{n-1}{\sum_{j \in V(G)} d(i,j)} = \frac{n-1}{s(i)}, \tag{6}$$

which characterizes how close a node is from the rest of nodes in a network [16,20].

3. On distance-sum distributions

We start by defining the probability $p(s)$ of selecting uniformly at random a node with distance-sum s in a network:

$$p(s) = n(s)/n, \tag{7}$$

where $n(s)$ is the number of nodes having distance-sum equal to s, and n is the size of the network. Then, the plot of $p(s)$ versus s represents the probability distribution function (PDF) of the distance-sum in a network. The cumulative distribution function (CDF) can be obtained by plotting the probability $P(s)$ of choosing at random a node with distance-sum larger or equal than s versus the distance-sum, where

$$P(s) = \sum_{s'=s}^{\infty} p(s'). \tag{8}$$

We study here the CDF instead of the PDF for the distance-sum of the nodes. The main reason is that the PDF is very noisy for both random and real-world networks, which makes difficult to find good fits for the distribution. We start by studying several random networks with different degree distributions. The first class corresponds to the 'classical' random networks built by using the Erdös–Rényi (ER) model [21]. The second group corresponds to networks with power-law degree distributions, known as 'scale-free' (SF) networks [5], which were constructed by using the algorithm developed by Hagberg et al. [22]. In the ER graphs a group of nodes are connected randomly forming a graph with Poisson degree distribution. In the case of SF model, the resulting graph displays a power-law degree distribution of the type $p(\delta) \sim \delta^{-\gamma}$, where $p(\delta)$ is the probability of finding a node of degree δ in the graph [22]. We have generated SF random networks with exponents $\gamma = 1.8, 2.5, 3.0$. The last ones are known as the Barabási–Albert (BA) networks [5,23]. In Fig. 1 we illustrate the cumulative distance-sum distributions (CDSD) for the networks with the previously mentioned topologies and having 1000 nodes and average degree $\bar{\delta} = 8$.

As can be seen in Fig. 1 the shapes of the CDSDs for all the random networks studied here are very similar to each other. The best fits obtained for these cumulative distributions (displayed as a continuous line in Fig. 1) correspond to cumulative normal distributions for a normal random variable with mean $\bar{s}$ and variance σ^2:

$$P(s, \bar{s}, \sigma^2) = \frac{1}{2}\left[1 + erf\left(\frac{s - \bar{s}}{\sigma\sqrt{2}}\right)\right], \tag{9}$$

where

$$erf(s) = \frac{2}{\sqrt{\pi}} \int_0^{\pi} e^{-t^2} dt. \tag{10}$$

The fits in Fig. 1 were obtained by using the Distribution Fitting Tool of Matlab which uses the Maximum Likelihood Estimates (MLE) method to estimate the best parameters of a distribution for a given data. The parameters for the best fits of all distributions in Fig. 1 are given in Table 1.

The situation is more complex when we consider the cumulative distance-sum distributions of some real-world networks. For the sake of illustration we show in Fig. 2 the CDSD for the networks representing the food web of *Benguela*, a social network of injecting drug users in *Colorado Spring*, the food web of *Skipwith* pond and the protein–protein interaction of *Drosophila melanogaster* (see further). The best fits found for such distributions are given in Fig. 2 as solid lines. In no one case the best fit corresponds to the normal distribution but to Log-Logistic, Generalized extreme value, Weibull and Log-Normal distributions.

The expressions for these cumulative distributions are given by the following expressions from left to right and up to bottom:

$P(s) = \frac{1}{1+(s/\alpha)^{-\beta}}$	$P(s) = e^{-t(x)}, t(x) = \begin{cases} \left[1 + \xi\left(\frac{s-\bar{s}}{\sigma}\right)\right]^{-1/\xi} \\ e^{-(s-\bar{s})/\sigma} \end{cases}, \xi \in \mathbb{R}$
$P(s) = 1 - e^{-(s/\lambda)^k}$	$P(s) = \frac{1}{2} + \frac{1}{2} erf\left[\frac{\ln s - \bar{s}}{\sqrt{2\sigma^2}}\right]$

The analysis of the real-world networks provides a very good example of the difficulties that arise when statistical distributions are used as a way to quantify the distance-sum heterogeneity in networks. For instance, how can we compare distributions so mismatched as the ones found for only four real-world networks? This difficulty together with those previously

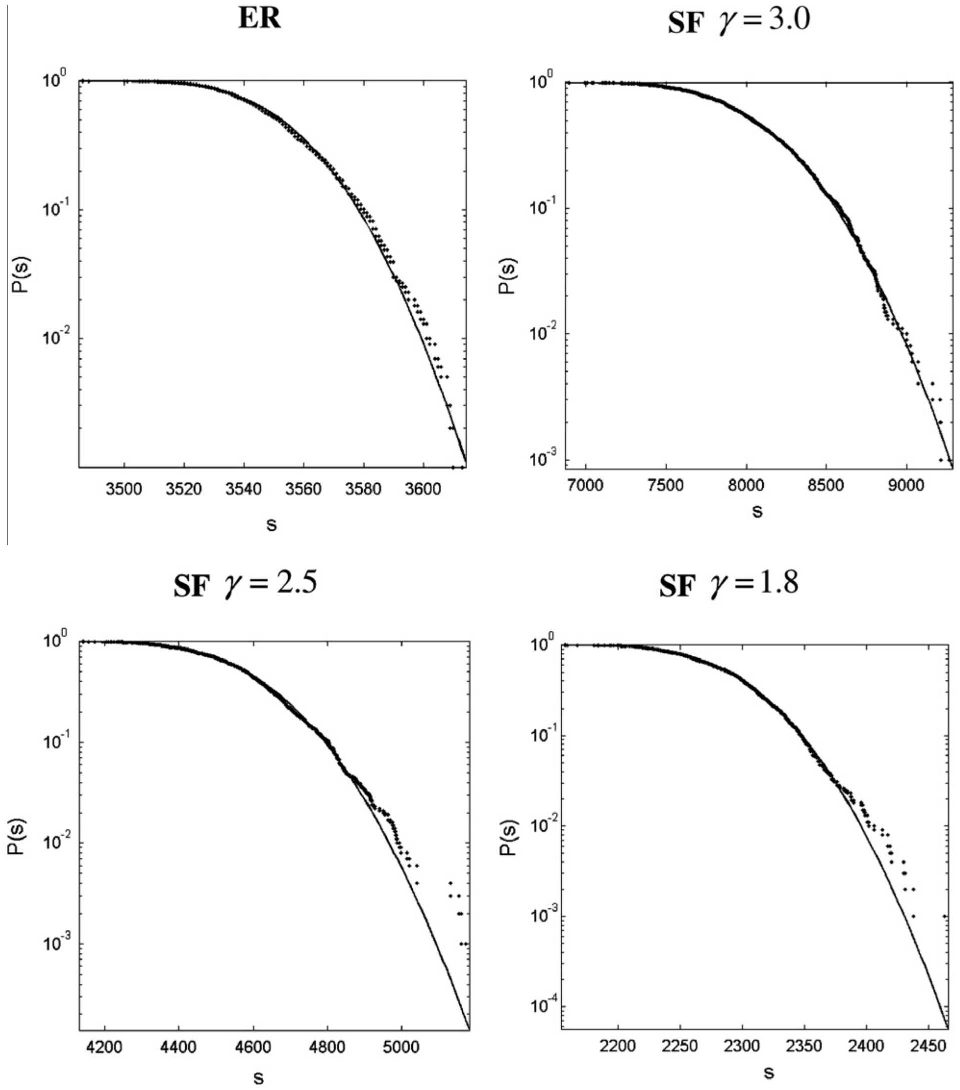

Fig. 1. Cumulative distance-sum distributions (CDSD) for random networks with different topologies: ER (top left), scale-free with exponent 3.0 (top right), scale-free with exponent 2.5 (bottom left) and scale-free with exponent 1.8 (bottom right). The best fit for normal CDF are displayed as solid lines.

Table 1
Fitting parameters for all CDSD of random networks and real networks represented in Fig. 1.

Network	μ	σ
ER ($\bar{k} = 8$)	3552.34	20.19
SF (γ = 3.0)	8053.17	393.96
SF (γ = 2.5)	4580.65	165.75
SF (γ = 1.8)	2288.79	45.89

mentioned in the Introduction points out to the necessity of defining an index of distance-sum heterogeneity. In the next section we propose a new approach to quantify the distance-sum heterogeneity of networks and we will show that there are some important differences in the distance-sum heterogeneity of random and real-world networks.

4. Distance-sum heterogeneity index

In order to introduce the distance-sum heterogeneity index we start by considering a hypothetical process in which the nodes of a given network reach a consensus about their distance-sums. For an excellent review on consensus models in

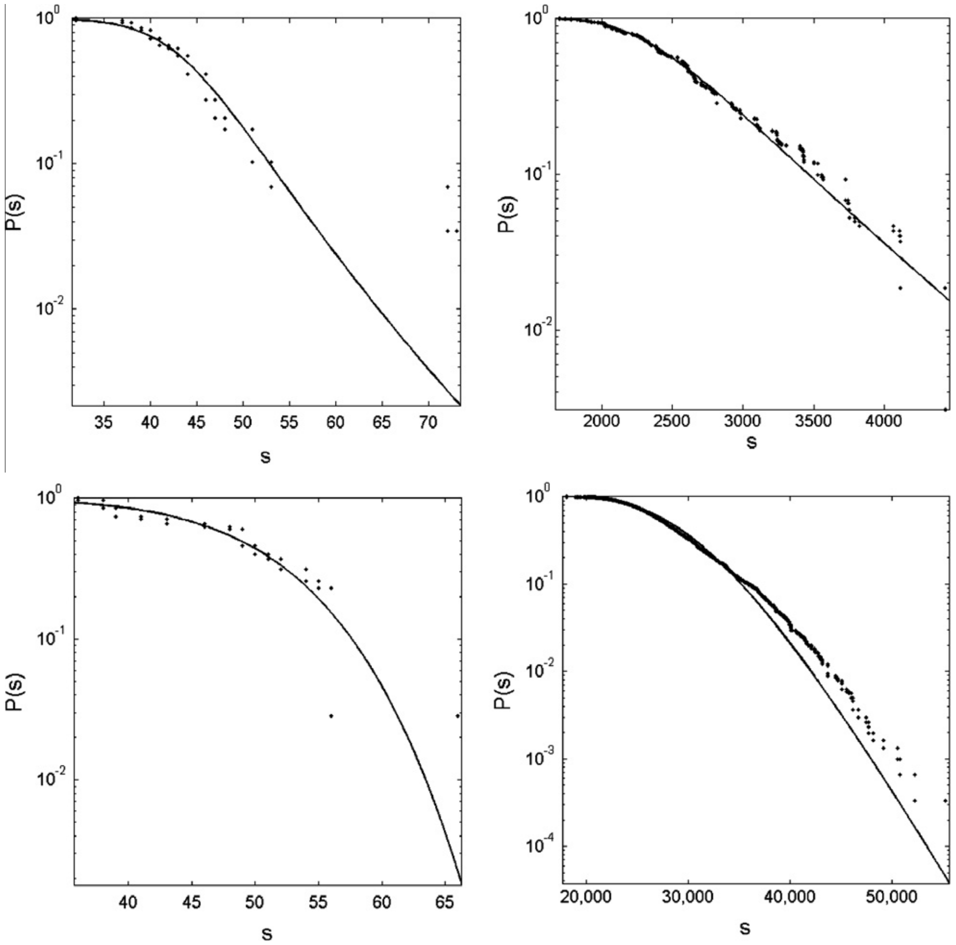

Fig. 2. Cumulative distance-sum distributions (CDSD) for some real networks: (top left) *Benguela* food web (top right) social networks of injecting drug users *at Colorado Springs*, USA, (bottom left) food web of *Skipwith* pond and (bottom right) protein–protein interaction networks of *Drosophila melanogaster*.

networks the reader is referred to [24]. That is, let $G = (V, E)$ be a simple, undirected and unweighted graph with distance-sum of the nodes given by the vector **s**. Let $f(s_i)$ be a function of the distance-sum of node i. In the hypothetical consensus process every pair of connected nodes tries to 'equalize' their functions $f = f(s_i)$ of distance-sums by a consensus process. The final consensus state is reached if, for all $f_i(0)$ and all $i, j = 1, \ldots n$, $|f_i(t) - f_j(t)| \to 0$ as $t \to \infty$ [24]. The consensus model has the form

$$d\mathbf{f}/dt = -\mathbf{L}\mathbf{f}, \mathbf{f}(0) = \mathbf{f}_0 \tag{11}$$

where **L** is the Laplacian matrix of the network. In order to control the evolution of the consensus process in the network a disagreement function $\varphi(f)$ is defined as [24]

$$\varphi(f) = \frac{1}{2}\mathbf{f}^T\mathbf{L}\mathbf{f} = \frac{1}{2}\langle f|\mathbf{L}|f\rangle \tag{12}$$

such as that the consensus model can be written as the gradient-descent algorithm [24]

$$d\mathbf{f}/dt = -\nabla\boldsymbol{\varphi}(f), \quad \mathbf{f}(0) = \mathbf{f}_0 \tag{13}$$

Now, returning to the quadratic form (12), we remark that it can be written as

$$\varphi(f) = \frac{1}{2}\sum_{(i,j)\in E}(f_i - f_j)^2 \tag{14}$$

indicating that $\varphi(f)$ measures the 'heterogeneity' in the distance-sum function f in every time-step of the consensus process.

Here we are not concerned with the time evolution of the 'heterogeneity' function in the consensus process, but mainly on how much heterogeneity a given graph has. That is, we are interested in finding $\varphi(f)$ only for time zero of the consensus process. For the sake of convenience we select our function f to be a power of the distance-sum, i.e., $f = f(s_i) = s_i^{\alpha}$. This is a general form that can embrace such indices like the Wiener index and closeness centrality ($\alpha = 1$), as well as the Balaban index ($\alpha = -1/2$) one. In closing, the distance-sum heterogeneity of a graph is given by the following formula:

$$\varphi(G) = \frac{1}{2}\sum_{(i,j)\in E}(s_i^{\alpha} - s_j^{\alpha})^2 = \frac{1}{2}(s^{\alpha})^T \mathbf{L} s^{\alpha} \tag{15}$$

In the rest of the paper we will consider only the case $\alpha = -1/2$, which relates the heterogeneity index with the Balaban index for a given graph.

5. Properties of the distance-sum heterogeneity index

Let $\varphi(G)$ be the heterogeneity index of a simple, undirected, unweighted graph G and let $\alpha = -1/2$. Then, it can be easily shown that

$$\varphi(G) = \sum_{i=1}^{n}\frac{\delta_i}{s_i} - 2\sum_{(i,j)\in E}(s_i s_j)^{-1/2}, \tag{16}$$

where δ_i is the degree of the node i. Note that the term in the second part of the right-hand side of (16) corresponds to the Balaban index except for the correction factor $m/(C+1)$.

The term δ_i/s_i in the expression (16) has the following interpretation. Let us consider a walker living at node i who can visit every node j of the connected graph by using the shortest paths from i to j. Let us consider a discrete-time process in which the time needed by the walker for going from one node to a nearest neighbor is $t = 1$. Here we consider independent visits to the nodes of the graph. That is, if a walker at node i visits the node j at distance d_{ij} it is assumed that the walker returns to i before he visits another node k. Thus, the total time needed by a walker living at node i for independently visiting every node of the network is $t_T(i) = 2s_i$. On the other hand, the time needed for independently visiting every nearest neighbor of node i is given by $t_{nn}(i) = 2\delta_i$. Consequently, the fraction of the total time needed by the walker to independently visiting all his nearest neighbors is given by:

$$t_R(i) = t_{nn}(i)/t_T(i) = \delta_i/s_i, \tag{17}$$

which defines a new centrality index for the nodes of a network.

If we consider n walkers living at the n nodes of a network, the average time needed by them to independently visiting their nearest neighbors is $t_R = \frac{1}{n}\sum_{i=1}^{n} t_R(i)$. Using these expressions we can rewrite the Balaban $J(G)$ index [14] in terms of the average time t_R and the distance-sum heterogeneity index as

$$J(G) = \gamma[nt_R - \varphi(G)], \tag{18}$$

where $\gamma = 2m/(C+1)$.

It is evident from (15) that the lower bound for the distance-sum heterogeneity index is zero, which is reached when the graph has the value of s_i for every node. In order to search for the maximum of this index we searched computationally all connected graphs with 3–8 nodes (~12,000 graphs). For graphs with $n = 3,4,5$ the maximum of the distance-sum heterogeneity index is reached for the star graph. For graphs with $n = 6,7,8$ the maximum is always reached for the graphs having the structures illustrated in Fig. 3. These graphs are easily constructed from a star graph $S_{1,n+1}$ by making a duplicate copy of the node having degree $n-1$. By obvious reasons we call these graphs 'agave' in allusion to the plant from which Tequila is produced. We note in passing that the clustering coefficients of agave graphs were previously studied by Bollobás [25]. We conjecture here that the agave graph always maximizes the distance-sum heterogeneity index for graphs having $n \geqslant 6$ nodes:

Conjecture. *Among the graphs having $n \geqslant 6$, the agave graph has the maximum distance-sum heterogeneity index.*

The distance-sum heterogeneity index for an agave graph with n nodes is given by

$$\varphi(agave) = (2n-4)\left(\frac{1}{\sqrt{n-1}} - \frac{1}{\sqrt{2n-4}}\right)^2 = \frac{(3n-5)}{n-1} - 2\left(2 - \frac{2}{n-1}\right)^{1/2}. \tag{19}$$

The agave graph with n nodes has $m = 2n-3$ links. However, most real-world networks have different densities than an agave graph with the same number of nodes. Consequently, it is more interesting to find the graphs that maximize the distance-sum heterogeneity index for a given number of nodes and links. We explored computationally the structure of these graphs by studying all connected graphs having $n = 4$–8. In this search we found a remarkable regularity in the structure of the graphs maximizing $\varphi(G)$. Among all trees with n nodes the star graph is always found as the one having the largest value of the distance-sum heterogeneity index. When the number of links is $n-1 < m \leqslant 2n-3$, the graphs maximizing $\varphi(G)$ have structures that point out to an iterative process in which a star graph is transformed into an agave one (see first line in Fig. 4). The iterative process continues up to the complete graph. Thus we propose an algorithm that

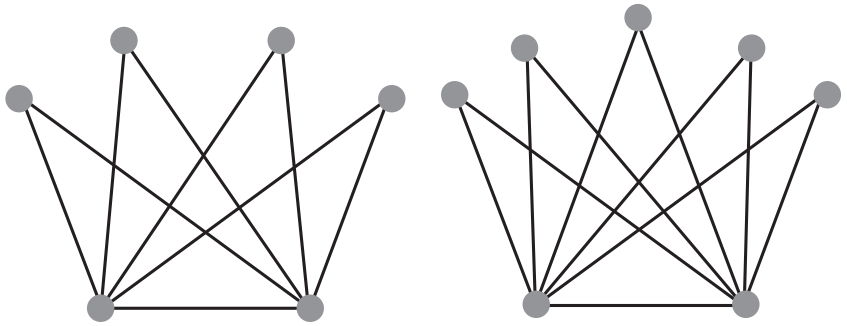

Fig. 3. Illustration of the agave graphs with 6 nodes and 7 nodes.

allows the construction of a graph with what we conjecture has the maximum value of the distance-sum heterogeneity index for a given number of nodes and edges. This algorithm is described below given a graph with n nodes and $m \geqslant n-1$ links.

Algorithm 1. Construction of the graph with the conjectured maximum value of distance-sum heterogeneity for a given number of nodes and edges.

```
(1) With $m = n-1$ links construct a star graph;
(2) Select a link $(i,j)$ of the star, such as $\delta_i = n-1$ and $\delta_i = 1$;
(3) Starting in a counterclockwise way (the same is obtained for a clockwise way) connect every node different from $i$ and $j$ to the node $j$;
(4) When $\delta_i = n-1$ (i.e., the agave graph) select a link $(i,k)$ where $\delta_k = 2$;
(5) Starting in a counterclockwise way (the same is obtained for a clockwise way) connect every node different from $i, j$ and $k$ to the node $k$;
(6) Repeat the process until all links of the graph have been used.
```

For instance, in Fig. 4 we illustrate the process for graphs having $n = 7$. In the first row of the figure the process starts by building a star graph in which the link (i,j) is marked as a bold line. Then, every node with degree 1 starting from the right is linked to the node j. The first line finishes when the agave graph with $n = 7$ and $m = 2n-3 = 11$ links is obtained. The second line starts by selecting the link (i,k) and connecting every node with degree 2 from the right with the node k. The rest of the process is self-explained in the figure.

The distance-sum heterogeneity index can be expressed in terms of the 'optimal' values of the index obtained from the algorithm given before. The following formula expresses the distance-sum heterogeneity as a percentage of the conjectured maximum possible value of distance-sum heterogeneity:

$$\varphi_{rel}(G) = \frac{100 \cdot \varphi(G)}{\varphi_{opt}(G)} \tag{20}$$

6. Spectral representation of the distance-sum heterogeneity index

Here we consider a spectral representation of the distance-sum heterogeneity index. We start by considering the $\mathbf{u}_j$ orthonormal eigenvector of the Laplacian matrix associated with the μ_j eigenvalue. The cosine of the angle formed between this eigenvector and the vector of distance-sum $\mathbf{s}^{-1/2}$ for a given network is expressed as

$$\cos\theta_j = \frac{\mathbf{s}^{-1/2} \cdot \mathbf{u}_j}{\|\mathbf{s}^{-1/2}\|}, \tag{21}$$

where $\|\mathbf{s}^{-1/2}\|$ is the Euclidean norm that can be written as $\|\mathbf{s}^{-1/2}\| = \sqrt{\sum_i s_i^{-1}}$. Let ${}^0J_{-1} = \sum_{i=1}^n s_i^{-1}$. Then, using the Euler theorem (see p. 457 of Ref. [26]) we can represent the distance-sum heterogeneity index in terms of the eigenvalues of the Laplacian and the cosines θ_j as follows

$$\varphi(G) = \frac{1}{{}^0J_{-1}} \sum_{j=2}^{n} \mu_j \cos^2\theta_j \tag{22}$$

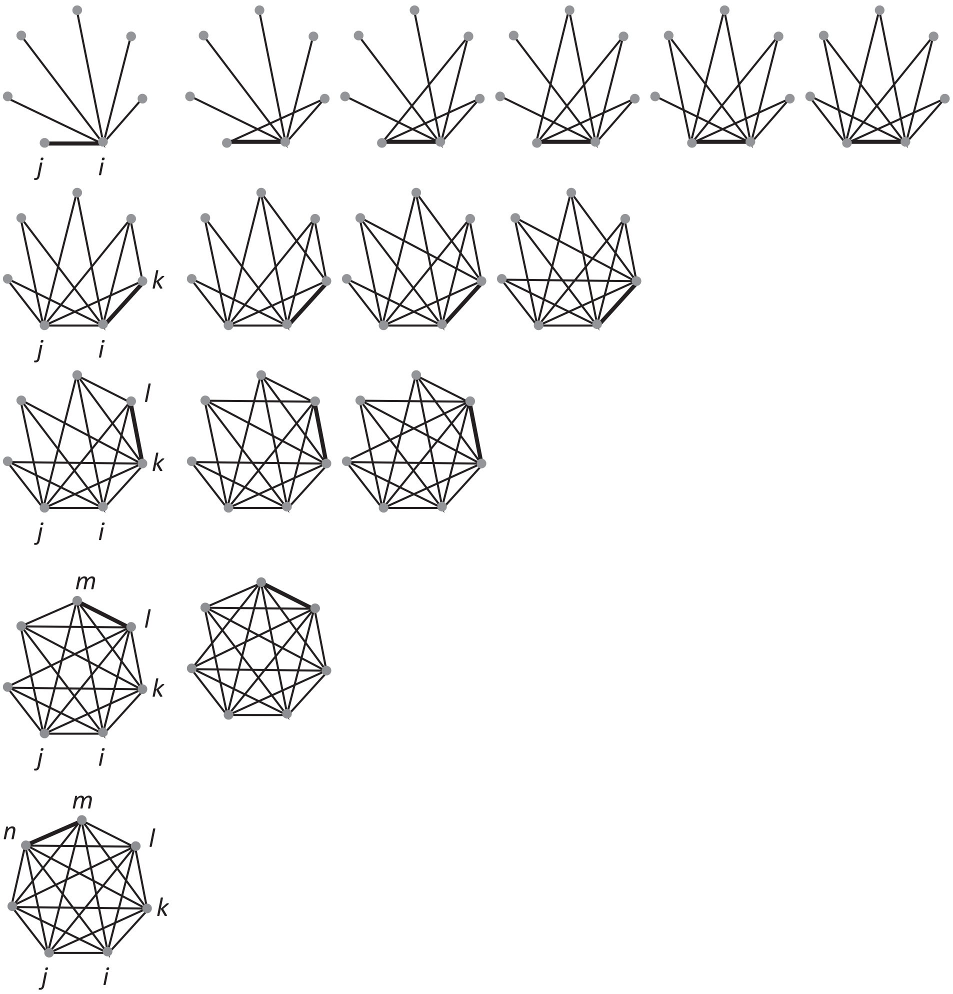

Fig. 4. Illustration of the process for constructing a graph with the conjectured maximum value of distance-sum heterogeneity for graphs with 7 nodes (see text for explanations).

The term $\cos^2\theta_j$ represents the similarity between the normalized distance-sum and the corresponding eigenvector (or vice versa). For instance, $\cos^2\theta_j = 0$ means that the vector $\mathbf{s}^{-1/2}$ is perpendicular to the Laplacian eigenvector $\mathbf{u}_j$, and no "duplicated" information is contained in both vectors, which means that they are dissimilar.

Now we can consider a graphical representation of the distance-sum heterogeneity of a network if we take a coordinate system with origin at $\mu_1 = 0$. We can represent the other eigenvalues of the Laplacian for a given network as points in this system in the following way. We consider that the eigenvector $\mu_{j>1}$ is represented by a point whose distance from the origin of coordinates O is given by $B = \sqrt{\mu_{j>1}}$. The segment OB forms an angle θ_j with the y axis of coordinates, which determines the full position of the point in the coordinate system. It can be seen that the projection of $\sqrt{\mu_{j>1}}$ on the x axis is given by $x_j = \sqrt{\mu_{j>1}}\cos\theta_j$, and the projection of $\sqrt{\mu_{j>1}}$ on the y axis is given by $y_j = \sqrt{\mu_{j>1}}\sin\theta_j$. This means that the distance-sum heterogeneity index $\varphi(G)$ can be written as

$$\varphi(G) = \left({}^{0}J_{-1}\right)^{-1}\sum_{j=1}^{n} x_j^2. \tag{23}$$

We can use this kind of plot to represent the distance-sum heterogeneity of a network in a graphical form by plotting x_j vs. y_j for all values of j. Thus the distance-sum heterogeneity index is given by the sum of the squares of the projections of all these points on the abscissa. Obviously, all projections on y-axis are positive but those on x-axis can have positive and negative signs. We will call these plots distance-sum heterogeneity plots or simply S-plots.

7. Distance-sum heterogeneity in random networks

As we have seen in a previous section the cumulative distance-sum distributions for random graphs do not display any significant difference in the distance-sum heterogeneity of networks with quite different topologies. We have calculated the distance-sum heterogeneity index for the same random graphs displayed in Fig. 1 as well as their values relative to the conjectured maximum heterogeneities and the results are given in Table 2.

As can be seen in Table 2 the degree distribution induces distance-sum heterogeneity in the random networks. The ER network displays the lowest distance-sum heterogeneity with a value of $\varphi_{rel}(G)$ close to zero. This indicates that most of the nodes in an ER network have approximately the same distance-sum displaying a remarkable regularity. As soon as

Table 2
Distance-sum heterogeneity index and their relative values for different random graphs.

Random network	$\varphi(G)$	$\varphi_{rel}(G)$
SF γ = 1.8	0.05487	13.95
SF γ = 2.5	0.00684	7.53
SF γ = 3.0	0.00295	3.44
ER	0.00103	0.30

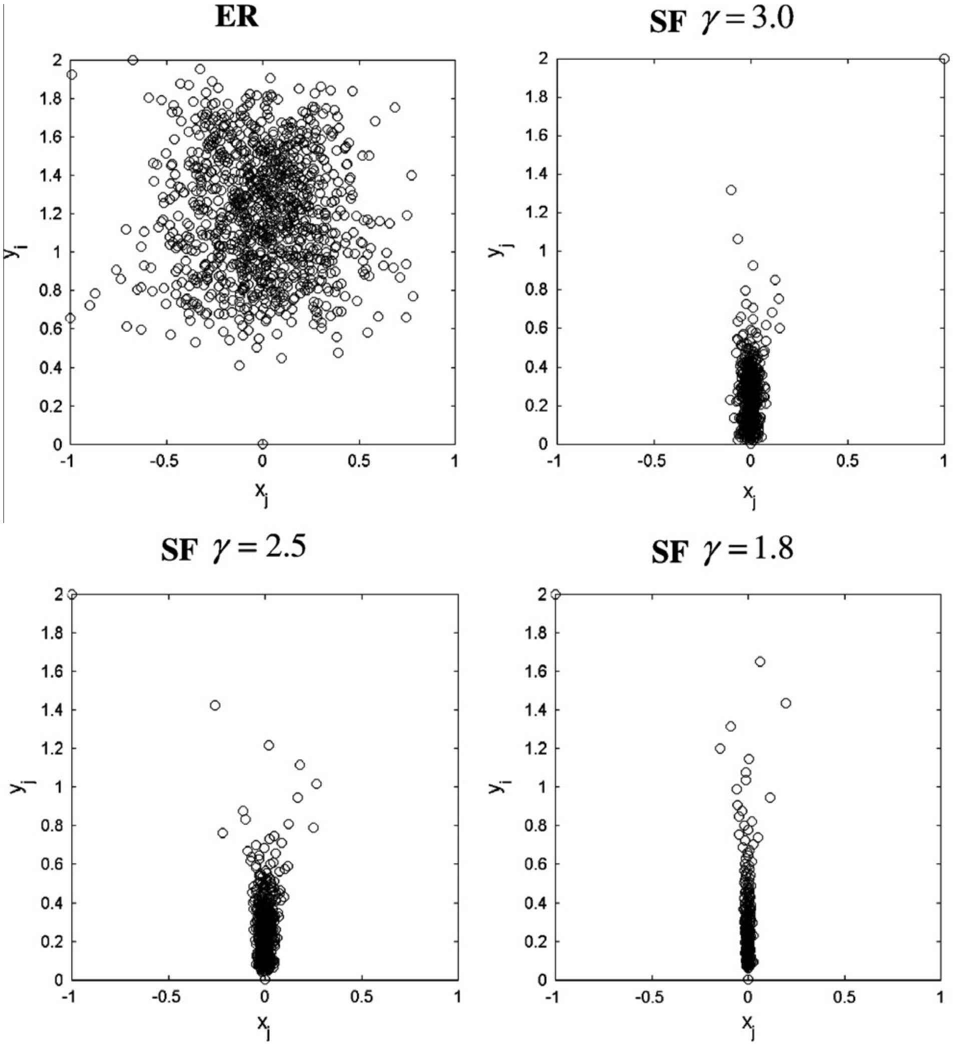

Fig. 5. *S*-plots for different random networks: ER (top left), scale-free with exponent 3.0 (top right), scale-free with exponent 2.5 (bottom left) and scale-free with exponent 1.8 (bottom right).

the degree distribution becomes more skewed there are some nodes that concentrate much more links than the rest, i.e., the hubs of the networks. As a consequence, the hubs have a larger number of small shortest paths than the poorly-connected nodes. This unbalance makes that the distance-sum heterogeneity increases in these networks. Graphically, these heterogeneities can be better observed by using the *S*-plots for these networks. In Fig. 5 we illustrate the *S*-plots for the random networks studied and it can be seen that the ER network has a very homogeneous *S*-plot, which covers practically all the values of x_j in the interval $[-1,1]$. The networks with SF topologies display very narrow *S*-plots in which most of the x_j values are concentrated around the zero value. Obviously, a further characterization of these plots would add more value to the analysis of distance-sum heterogeneity in networks. However, we will not consider such kinds of quantitative characterizations in this work.

8. Distance-sum heterogeneity in real-world networks

In this section we study the distance-sum heterogeneity of 57 real-world networks representing biological (B), ecological (E), informational (I), social (S) and technological (T) systems. The description of all these networks as well as the references for the original sources can be found in the Appendix of book [3]. Biological networks include: the neural network of *C. elegans*; the transcription networks of yeast, *E. coli* and urchins; the PPI networks of *D. melanogaster*, *H. pylori*, *A. fulgidus*, *B. subtilus*, *E. coli*, *malaria parasite*, *Kaposi sarcoma herpes* virus, *human* and *yeast*. Ecological networks represent the following food webs: *Benguela*, *Coachella Valley*, *Reef Small*, *Shelf*, *Skipwith* pond, *St. Marks* seagrass, *Stony* stream, *Bridge Brook*, *Canton Creek*, *Chesapeake* Bay, *El Verde* rainforest, *Scotch Broom*, Grassland *Little Rock*, *St. Martin* and *Ythan* estuary with and without parasites. Informational networks represent the following systems: a network of the *Roget thesaurus*; a citation network consisting of papers published in the *Proceedings of Graph Drawing* in the period 1994–2000; a semantic network of the Online Dictionary of Library and Information Science (*ODLIS*); a citation network in the field of "small-world". Social networks considered are: the social networks of corporate elite in USA, inmates in prison, the friendship network between physicians (*Galesburg*), the friendship ties among the employees in a small hi-tech computer firm which sells, installs, and maintains computer systems (*high-tech*), and a sawmill communication network; the social networks of injecting drug users, a social network among college students in a course about leadership and the *Zachary* karate club; persons with HIV infection during its early epidemic phase in *Colorado Springs*, a scientific collaboration network in the field of computational geometry, and two sexual networks, one consisting of heterosexual relations only and the other including both heterosexual and homosexual relationships. Finally, technological networks include: three electronic sequential logic circuits parsed from the ISCAS89 benchmark set, the western USA power grid; the software network of *Abi, Digital, MySQL, VTK* and *XMMS*; the USA airport transportation network of 1997; two versions of Internet at autonomous system of 1997 and 1998.

According to our calculations, real-world networks do not display very large distance-sum heterogeneity indices. The average value of the relative distance-sum heterogeneity is about 5%. However, there are significant variations of this index for individual networks. For instance, the food webs of Skipwith and Bridge Brooks have relative distance-sum heterogeneities of 32% and 20%, respectively, and the citation network of 'small-world' has a value of near 12%. On the other side of the coin there are 7 networks with relative distance-sum heterogeneities smaller than 1%, which are accordingly very close to those observed for random networks with Poisson degree distributions. In Fig. 6 we illustrate the values of the relative distance-sum heterogeneity indices for all the real-world networks studied here.

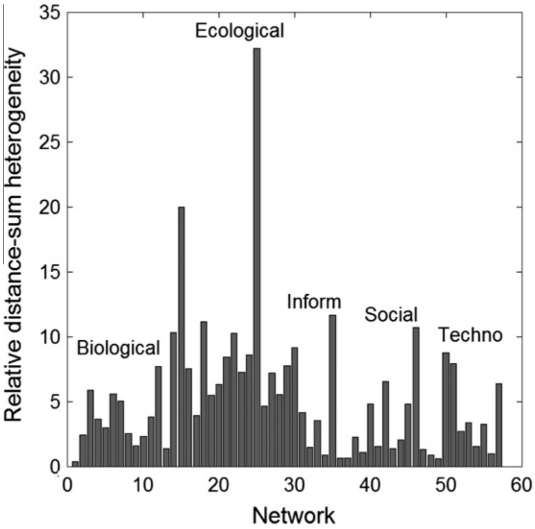

Fig. 6. Values of the relative distance-sum heterogeneity indices for the 57 real-world networks studied here.

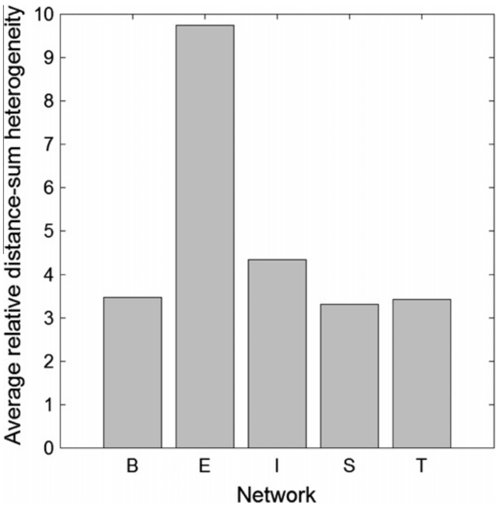

Fig. 7. Average relative distance-sum heterogeneity for all networks grouped into different functional classes: Biological (B), Ecological (E), Informational (I), Social (S) and Technological (T).

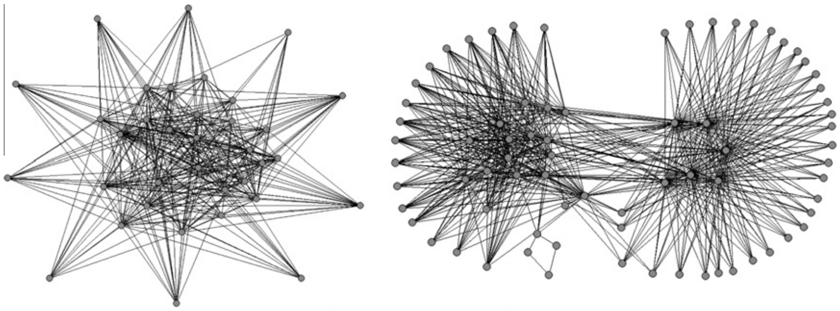

Fig. 8. Illustrations of the two food webs with the largest relative distance-sum heterogeneities: *Skipwith* (left) and *Bridge Brooks* (right). Nodes represent species and links represent trophic interactions (who-eat-who) in the ecosystem.

When the average relative distance-sum heterogeneity is considered for all networks in the different functional classes, i.e., B, E, I, S and T, we find some interesting observations. First, the largest distance-sum heterogeneity is observed for the ecological networks, which display an average of about 10% of the conjectured maximum value for this index. The removal of the two networks with the largest distance-sum heterogeneity does not change very much this situation. For instance, after removing the food webs of Skipwith and Bridge Brooks the remaining food webs have an average of 8% of relative distance-sum heterogeneity, which is still the double of those observed for the networks in the other groups. Informational networks have average relative distance-sum heterogeneity of about 4% and the other three groups are very close to each other with percentages between 3.31% (S) and 3.47% (B) (see Fig. 7).

The reason why food webs display more relative distance-sum heterogeneity is not clear. These are networks with higher densities than the rest of the networks studied here. For instance, the densities of the food webs analyzed here are about 6 times larger than the average of the rest of the networks. As a consequence, there is a linear correlation between the density and the relative distance-sum heterogeneity of the networks studied here. It displays a Pearson correlation coefficient of 0.71 indicating that the denser networks are also the most distance-sum heterogeneous. However, it is easy to be fooled by this kind of correlations as we can build networks with high density and very poor distance-sum heterogeneity (think for instance about the complete graph). In fact, if we remove all food webs from the previous correlation, the correlation coefficients drops to 0.51, indicating that there is no such kind of dependence and that the previous observation appears to be biased by the presence of food webs. Thus, it is plausible that there is some kind of functional cause for the appearance of distance-sum heterogeneity in food webs. A view of the two networks with the largest relative distance-sum heterogeneities gives some important hints (see Fig. 8). It is evident from this Figure that the food webs of *Skipwith* and *Bridge Brooks* resemble very much the type of graphs we have conjectured to display the largest values of distance-sum heterogeneity. This type of structure can appears naturally in the evolution of food webs, where there could be a central core of species with trophic relations among them, surrounded by one or more layers of species that have trophic relations with the central core but not among them. This could be the case, for instance, of parasites that have trophic interactions with other species but

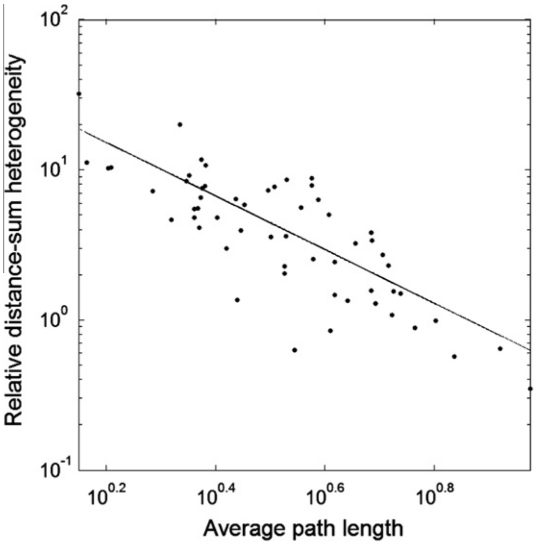

Fig. 9. Relation between the relative distance-sum heterogeneity index and average shortest path distance for the 57 real-world networks studied. The plot is in log–log scale to illustrate the power-law relationships existing between both parameters.

not among them. In closing, we have found that the type of topological structure that maximizes the distance-sum heterogeneity of any graph can appear naturally in ecological food webs, where they can explain some of the structural and dynamical properties of such ecological systems.

We have also explored the relationship between the relative distance-sum heterogeneity index and other invariants for networks, such as the network size, average degree, average clustering coefficient (relative proportion of triangles in which a node participates), and the average path length. We have found that the relative distance-sum heterogeneity index decays as a power-law with the average distance. Accordingly, $\varphi_{rel}(G) = 34.88 \cdot \bar{l}^{-1.79}$ with correlation coefficient equal to 0.75 (see Fig. 9). This relationship indicates that the networks with large relative distance-sum heterogeneity have small average path length. According to the Algorithm 1 (see also Fig. 4) we can infer that among all graphs with n nodes which are conjectured to maximize the distance-sum heterogeneity, the star graph has the largest average shortest path distance. In the Fig. 4 we can see that the star is the initial stage for the generation of graphs with maximum $\varphi(G)$ and that in every further step we are adding links in a way that decreases the average distance among nodes. For instance, the average path length in the star graph with n nodes is given by

$$\bar{l}(S_n) = 2 - \frac{2}{n}, \tag{24}$$

which is reduced for the agave graph with n nodes to:

$$\bar{l} = 2 - \frac{4}{n}. \tag{25}$$

It is straightforward to realize that $\bar{l}(S_n) \to 2$ as $n \to \infty$, and as the density of the graphs increases the average path length drops quickly. Consequently, the graphs which have high relative distance-sum heterogeneity necessarily have small average path length.

9. Conclusions

We have analyzed here the distance-sum heterogeneity of artificial and real-world networks. The distance-sum of nodes appears in many different graph-theoretic invariants used for studying graphs and networks in different fields. We first have analyzed the distance-sum cumulative distribution as a natural extension as what have been extensively done for the degree of the nodes in the network science literature. We have shown here that distance-sum distributions do not account for the heterogeneity in the distance-sums of random and real-world networks. Then, we have motivated and introduced a new index of distance-sum heterogeneity, which is shown to be a quadratic form of the Laplacian matrix of the graph. The index allows an interpretation of the Balaban index of a graph to be the contribution of the average time needed by n walkers to independently visiting their nearest neighbors minus the contribution of the distance-sum heterogeneity of the graph.

We have conjectured here that the maximum value of the distance-sum heterogeneity index for a graph with n nodes is obtained for the agave graph. We also conjectured that the graphs with n nodes and m links that maximize this index are related to the agave graph and can be generated by a simple iterative algorithm. Using the distance-sum heterogeneity values for these graphs we have proposed an index of relative distance-sum heterogeneity. We have shown that this index differentiates very well random graphs with different degree distributions as well as real-world networks with different functions and topologies.

Finally, we have given evidence that the new index of relative distance-sum heterogeneity of a graph is an important addition for the structural analysis of complex networks. In particular, the analysis of methods and algorithms explaining why food webs and possibly other ecological networks display larger distance-sum heterogeneity than networks in other fields appears to be a promising avenue.

Acknowledgments

This work is dedicated to Prof. A.T. Balaban (Texas A&M) for his many outstanding contributions to graph theory and its applications. EE thanks partial financial support from New Professor's Fund, University of Strathclyde and from the project *Mathematics of Large Technological Evolving Networks* (*MOLTEN*) supported by the Engineering and Physical Sciences Research Council and the Research Councils UK Digital Economy programme, with grant ref. EP/I016058/1. EVE thanks CONACYT, Mexico for support through a scholarship for doctoral studies at the University of Strathclyde.

References

[1] S. Boccaletti, V. Latora, Y. Moreno, M. Chavez, D.-U. Hwang, Complex networks: structure and dynamics, Phys. Rep. 424 (2006) 175.
[2] L. da Fontoura Costa, O.N. Oliveira Jr., G. Travieso, F. Aparicio Ridrigues, P.R. Villas Boas, M.P. Viana, L.E. Correa Rocha, Analyzing and modeling real-world phenomena with complex networks: a survey of applications, Adv. Phys. 60 (2011) 329–412.
[3] E. Estrada, The Structure of Complex Networks. Theory and Applications, Oxford University Press, 2011.
[4] E. Estrada, Quantifying network heterogeneity, Phys. Rev. E 82 (2010) 066102.
[5] R. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) 47–97.
[6] T.A.B. Snijder, Accounting for degree distributions in empirical analysis of network dynamics, in: R. Breiger, K. Carley, P. Pattison (Eds.), Dynamic Social Network Modeling and Analysis: Workshop Summary and Papers, National Research Council of the National Academies, The National Academies Press, Washington, DC, 2003, pp. 146–161.
[7] M.P.H. Stumpf, P.J. Ingram, Probability models for degree distributions of protein interaction networks, Europhys. Lett. 71 (2005) 152–158.
[8] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, Comp. Comm. Rev. 29 (1999) 251–262.
[9] S.N. Dorogovtsev, J.F.F. Mendes, A.N. Samukhin, Metric structure of random networks, Nucl. Phys. B 653 (2003) 307–338.
[10] K. Malarz, J. Karpinska, A. Kardas, K. Kulalowski, Node-node distance distribution for growing networks, arXiv:cond-mat/0309255v2.
[11] V.D. Blondell, J.-L. Guillaume, J.M. Hendrickx, R.M. Jungers, Distance distribution in random graphs and application to network exploration, Phys. Rev. E 76 (2007) 066101.
[12] H. Wiener, Structural determination of paraffin boiling points, J. Amer. Chem. Soc. 69 (1947) 17–20.
[13] A.T. Balaban, Highly discriminating distance-based topological index, Chem. Phys. Lett. 89 (1982) 399–404.
[14] J. Devillers, A.T. Balaban (Eds.), Topological Indices and Related Descriptors in QSAR and QSPR, Gordon & Breach, Amsterdam, 1999.
[15] D.J. Watts, S.H. Strogatz, Collective dynamic of "small-world" networks, Nature 393 (1998) 440–442.
[16] L.C. Freeman, Centrality in networks: I. Conceptual clarification, Social Networks 1 (1979) 215–239.
[17] F. Buckley, F. Harary, Distance in Graphs, Addison-Wesley, Redwood, 1990.
[18] I. Gutman, Y.N. Yeh, S.L. Lee, Y.L. Luo, Some recent results in the theory of the Wiener number, Indian J. Chem. 32A (1993) 651–661.
[19] S.H. Strogatz, Exploring complex networks, Nature 410 (2001) 268–276.
[20] S. Wasserman, K. Faust, Social Network Analysis, Cambridge University Press, Cambridge, 1994.
[21] P. Erdös, A. Rényi, On random graphs I, Publ. Math. Debrecen 5 (1959) 290–297.
[22] A.A. Hagberg,, D.A. Schult, P.J. Swart, in: G. Varoquaux, T. Vaught, J. Millman (Eds.), Proceedings of the 7th Python in Science Conference (SciPy2008), Pasadena, CA, USA, 2008, pp. 11–15.
[23] A.-L. Barabási, R. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[24] R. Olfati-Saber, J.A. Fax, R.M. Murray, Consensus and cooperation in networked multi-agent systems, Proc. IEEE 95 (2007) 215–233.
[25] B. Bollobás, Mathematical results on scale-free random graphs, in: S. Bornholdt, H.G. Schuster (Eds.), Handbook of Graph and Networks: From the genome to the internet, Wiley-VCH, Weinheim, 2003, pp. 1–32.
[26] F.E. Hohn, Elementary Matrix Algebra, Dover, New York, 1973.