Physics Letters A 381 (2017) 1659–1664

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Minimal partition coverings and generalized dimensions of a complex network

Eric Rosenberg

AT&T Labs, Middletown, NJ 07748, United States

ARTICLE INFO

Article history:
Received 4 January 2017
Received in revised form 3 March 2017
Accepted 4 March 2017
Available online 21 March 2017
Communicated by C.R. Doering

Keywords:
Complex networks
Generalized dimensions
Multifractals
Box counting
Partition function

ABSTRACT

Computing the generalized dimensions D_q of a complex network requires covering the network by a minimal number of "boxes" of size s. We show that the current definition of D_q is ambiguous, since there are in general multiple minimal coverings of size s. We resolve the ambiguity by first computing, for each s, the minimal covering that is summarized by the lexicographically minimal vector $x(s)$. We show that $x(s)$ is unique and easily obtained from any box counting method. The $x(s)$ vectors can then be used to unambiguously compute D_q. Moreover, $x(s)$ is related to the partition function, and the first component of $x(s)$ can be used to compute D_∞ without any partition function evaluations. We compare the box counting dimension and D_∞ for three networks.

1. Introduction

A network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$ is a set $\mathcal{N}$ of nodes connected by a set $\mathcal{A}$ of arcs. For example, in a friendship social network [21], a node might represent a person and an arc indicates that two people are friends. In a co-authorship network, a node represents an author, and an arc connecting two authors means that they co-authored (possibly with other authors) at least one paper. In a communications network [14], a node might represent a router, and an arc might represent a physical connection between two routers. Many applications of network models are discussed in [2]. We use the term "complex network" to mean an arbitrary network without special structure (as opposed to, e.g., a regular lattice), for which all arcs have unit cost (so the length of a shortest path between two nodes is the number of arcs in that path), and all arcs are undirected (so the arc between nodes i and j can be traversed in either direction).

There are many measures used to characterize complex networks. The degree of a node is the number of arcs having that node as one of its endpoints, and one of the most studied measures is the average node degree [11]. The clustering coefficient quantifies, in social networking terms, the extent to which my friends are friends with each other. The diameter Δ is defined by $\Delta \equiv \max\{dist(x, y) \,|\, x, y \in \mathcal{N}\}$, where $dist(x, y)$ is the length of the shortest path between nodes x and y. (We use "$\equiv$" to denote a definition.) Other network measures include the average path length [1], the box counting dimension d_B ([8,19]), the information dimension d_I ([17,22]), and the correlation dimension d_C ([9,16,18]).

In [17], Rosenberg showed that the definition proposed in [22] of the information dimension d_I of a complex network $\mathcal{G}$ is ambiguous, since d_I is computed from a minimal covering of $\mathcal{G}$ by "boxes" of size s, and there are in general different minimal coverings of $\mathcal{G}$ by boxes of size s, yielding different values of d_I. Using the maximal entropy principle of Jaynes [7], the ambiguity is resolved for each s by maximizing the entropy over the set of minimal coverings by boxes of size s. We face the same ambiguity when using box counting to compute D_q (as in the method proposed in [20]), since the different minimal coverings by boxes of size s can yield different values of D_q. We illustrate this indeterminacy, for a very simple network, in Section 3. Thus different researchers applying different methods for computing a minimal covering of the same network might compute very different values of D_q. The solution to this indeterminacy is to select, for each s, the minimal covering of $\mathcal{G}$ satisfying some appropriate criterion that guarantees uniqueness. Moreover, the method used to select the unique minimal covering should require only negligible additional computation beyond what is required to compute a minimal covering.

A natural way to obtain a unique minimal covering for a given s and q is to compute a minimal covering that minimizes the partition function; we call such a covering an "(s, q) minimal covering". We show that for $q > 1$ an (s, q) minimal covering will try to

E-mail address: ericr@att.com.

http://dx.doi.org/10.1016/j.physleta.2017.03.004
0375-9601/

equalize the number of nodes over all boxes in a minimal covering. An (s, q) minimal covering can be computed by a minor modification of whatever method is used to compute a minimal covering.

Although the new notion of (s, q) minimal coverings removes the ambiguity in the calculation of D_q, it is chiefly of theoretical interest, since we do not want to compute an (s, q) minimal covering for each s and q. Rather, we want to compute D_q using only a *single* minimal covering for each s. To this end, we introduce the new notion of a lexico (short for *lexicographically*) minimal summary vector $x(s)$, which summarizes a minimal covering of size s. The value $x_j(s)$ is the number of nodes in box B_j of a minimal covering, and $x_j(s)$ is non-increasing in j. We prove that $x(s)$ is unique for each s and that $x(s)$ summarizes an (s, q) minimal covering for all sufficiently large q. Computing $x(s)$ requires essentially no extra computation beyond what is required to compute a minimal covering. Since for each s there is a unique lexico minimal vector $x(s)$, and $x(s)$ summarizes a minimal covering, we can use the $x(s)$ vectors to unambiguously compute D_q.

We also show that $D_\infty \equiv \lim_{q\to\infty} D_q$ can be computed from the $x_1(s)$ values, where $x_1(s)$ is the first component of $x(s)$, without any partition function evaluations. We illustrate this by computing D_∞ for three networks, and comparing D_∞ to the box counting dimension d_B. For two of the three networks $d_B > D_\infty$ and for the third network $d_B \approx D_\infty$.

We emphasize that this paper does not propose a new box counting method for computing the generalized dimensions D_q of a network. Nor is our goal to compare D_q with other network dimensions such as d_B, d_C, or d_I. Rather, our intent is to introduce the $x(s)$ summary vectors, describe their interesting properties, and show how any box counting method can easily be modified to compute the $x(s)$ vectors, which can then be used to unambiguously compute D_q.

2. Preliminary definitions

Throughout this paper, $\mathcal{G}$ will refer to a complex network with node set $\mathcal{N}$ and arc set $\mathcal{A}$. We assume that $\mathcal{G}$ is connected, meaning there is a path of arcs in $\mathcal{A}$ connecting any two nodes. Let $N \equiv |\mathcal{N}|$ be the number of nodes. The network B is a subnetwork of $\mathcal{G}$ if B is connected and B can be obtained from $\mathcal{G}$ by deleting nodes and arcs. For each positive integer s such that $s \geq 2$, let $\mathcal{B}(s)$ be a collection of subnetworks (called *boxes*) of $\mathcal{G}$ satisfying two conditions: (i) each node in $\mathcal{N}$ belongs to exactly one subnetwork (i.e., to one box) in $\mathcal{B}(s)$, and (ii) the diameter of each box in $\mathcal{B}(s)$ is at most $s-1$. We call $\mathcal{B}(s)$ a *covering* of $\mathcal{G}$ of size s, or more simply, an s-covering. We do not consider $\mathcal{B}(s)$ for $s = 1$, since a box of diameter 0 contains only a single node. Define $B(s) = |\mathcal{B}(s)|$, so $B(s)$ is the number of boxes in $\mathcal{B}(s)$. The s-covering $\mathcal{B}(s)$ is *minimal* if $B(s)$ is less than or equal to the number of boxes in any other s-covering. For $s > \Delta$, the minimal s-covering consists of a single box, which is $\mathcal{G}$ itself. The term "box counting" refers to computing a minimal s-covering of $\mathcal{G}$ for a range of values of s. In general, we cannot easily compute a minimal s-covering, but good heuristics are known (e.g., [3,8,15,19,23]).

The next set of definitions concern the generalized dimensions of a geometric object. Consider a dynamical system in which motion is confined to some bounded set $\Omega \subset \mathbb{R}^E$ (E-dimensional Euclidean space) equipped with a natural invariant measure σ. Define a "box" to be a neighborhood (centered at some point) of Ω. We cover Ω with a set $\mathcal{B}(s)$ of boxes of diameter s such that $\sigma(B_j) > 0$ for each box $B_j \in \mathcal{B}(s)$ and such that for any two boxes $B_i, B_j \in \mathcal{B}(s)$ we have $\sigma(B_i \cap B_j) = 0$ (i.e., boxes may overlap, but the intersection of each pair of boxes has measure zero). Define the probability $p_j(s)$ of B_j by $p_j(s) \equiv \sigma(B_j)/\sigma(\Omega)$. In practice, $p_j(s)$ is approximated by $N_j(s)/N$, where N is the total number of observed points and $N_j(s)$ is the number of points in box B_j [13]. For $q \in \mathbb{R}$, define

$$Z_q\big(\mathcal{B}(s)\big) \equiv \sum_{B_j \in \mathcal{B}(s)} [p_j(s)]^q \,. \tag{1}$$

For $q > 0$ and $q \neq 1$, the generalized dimension D_q was defined in 1983 by Grassberger [5] and by Hentschel and Procaccia [6] as

$$D_q \equiv \frac{1}{q-1} \lim_{s\to 0} \frac{\log Z_q\big(\mathcal{B}(s)\big)}{\log s} \,. \tag{2}$$

Since definition (1) was presented only in the context of a geometric object, we extend the definition to a complex network. Let $\mathcal{B}(s)$ be an s-covering of $\mathcal{G}$. For $B_j \in \mathcal{B}(s)$, define $p_j(s) \equiv N_j(s)/N$, where $N_j(s)$ is the number of nodes in B_j. For $q \in \mathbb{R}$, we use (1) to define $Z_q\big(\mathcal{B}(s)\big)$, and we call $Z_q\big(\mathcal{B}(s)\big)$ the *partition function value* for $\mathcal{B}(s)$.

For convenience, the symbols used in this paper are summarized in Table 1.

Table 1
Symbols and their definitions.

Symbol	Definition
Δ	network diameter
$\mathcal{B}(s)$	covering of $\mathcal{G}$ by boxes of size s
$B(s)$	cardinality of $\mathcal{B}(s)$
B_j	box in $\mathcal{B}(s)$
d_B	box counting dimension
D_q	generalized dimension
d_I	information dimension
$\mathcal{G}$	complex network
N	number of nodes in $\mathcal{G}$
$N_j(s)$	number of nodes in box $B_j \in \mathcal{B}(s)$
$p_j(s)$	probability of box $B_j \in \mathcal{B}(s)$
$\mathbb{R}^E$	E-dimensional Euclidean space
$x(s)$	vector summarizing the covering $\mathcal{B}(s)$
$Z_q\big(\mathcal{B}(s)\big)$	partition function value for the covering $\mathcal{B}(s)$
$Z(x, q)$	partition function value for the summary vector x

3. Minimizing the partition function

The method of [20] for computing D_q for $\mathcal{G}$ is the following. For each s, compute a minimal s-covering $\mathcal{B}(s)$ and $Z_q\big(\mathcal{B}(s)\big)$. (In practice, if using a randomized box counting heuristic, $Z_q\big(\mathcal{B}(s)\big)$ is the average partition function value, averaged over some number of executions of the heuristic.) Then $\mathcal{G}$ has the generalized dimension D_q (for $q \neq 1$) if over some range of s and for some constant c

$$\log Z_q\big(\mathcal{B}(s)\big) \approx (q-1) D_q \log(s/\Delta) + c \,. \tag{3}$$

This definition is ambiguous, since different minimal s-coverings can yield different values of $Z_q\big(\mathcal{B}(s)\big)$. In particular, [17] showed that the value of d_I for $\mathcal{G}$ depends on the particular minimal s-coverings of $\mathcal{G}$ selected, and proposed the notion of a maximal entropy minimal covering for use in computing d_I. A similar ambiguity arises in defining D_q for $\mathcal{G}$, since different minimal s-coverings can yield different box probabilities $p_j(s)$ and hence different values of D_q.

Example 1. Consider the "chair" network of Fig. 1, which shows two minimal 3-coverings and a minimal 2-covering. Choosing $q = 2$, for the covering $\widetilde{\mathcal{B}}(3)$ from (1) we have $Z_2\big(\widetilde{\mathcal{B}}(3)\big) = (\frac{3}{5})^2 + (\frac{2}{5})^2 = \frac{13}{25}$, while for $\widehat{\mathcal{B}}(3)$ we have $Z_2\big(\widehat{\mathcal{B}}(3)\big) = (\frac{4}{5})^2 + (\frac{1}{5})^2 = \frac{17}{25}$. For $\mathcal{B}(2)$ we have $Z_2\big(\mathcal{B}(2)\big) = 2(\frac{2}{5})^2 + (\frac{1}{5})^2 = \frac{9}{25}$. If we use $\widetilde{\mathcal{B}}(3)$ then from (3) and the range $s \in [2, 3]$ we obtain $D_2 = \big(\log\frac{13}{25} - \log\frac{9}{25}\big)/(\log 3 - \log 2) = 0.907$. If instead we use $\widehat{\mathcal{B}}(3)$ and the same range of s we obtain $D_2 = \big(\log\frac{17}{25} - \log\frac{9}{25}\big)/(\log 3 - \log 2) = 1.569$.