Physics Letters A ••• (••••) •••–•••

Contents lists available at ScienceDirect

Physics Letters A

www.elsevier.com/locate/pla

Maximal entropy coverings and the information dimension of a complex network

Eric Rosenberg

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

ARTICLE INFO

Article history:
Received 1 October 2016
Received in revised form 5 December 2016
Accepted 7 December 2016
Available online xxxx
Communicated by C.R. Doering

Keywords:
Information dimension
Box counting dimension
Entropy
Networks
Fractals
Multifractal analysis

ABSTRACT

Computing the information dimension d_I of a complex network $\mathcal{G}$ requires covering $\mathcal{G}$ by a minimal collection of "boxes" of size s to obtain a set of probabilities, computing the entropy $H(s)$, and quantifying how $H(s)$ scales with $\log s$. We show that to determine whether $d_I \leq d_B$ holds for $\mathcal{G}$, where d_B is the box counting dimension, it is not sufficient to determine a minimal covering for each s. We introduce the new notion of a maximal entropy minimal covering of $\mathcal{G}$, and a corresponding new definition of d_I. The use of maximal entropy minimal coverings in many cases enhances the ability to compute d_I.

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. Networks are used to model a wide range of systems. For example, in a friendship social network [23], 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 [17,18], a node might represent a router, and an arc might represent a physical connection between two routers. In manufacturing, a node might represent a station in an assembly line, and an arc might represent the logical flow of a product being assembled as it moves from one station to the subsequent station. Other applications of network models include human brain function [12] and public health [14]. The term "complex network" generally refers to an arbitrary network without special structure, as opposed to, e.g., a regular lattice or grid network. Typically, a complex network also refers to a network in 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 [15]. The clustering coefficient measures, in social networking terms, the extent to which my friends are friends with each other. The diameter Δ of a network 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 [4] and the box counting dimension d_B [21]. There are some results relating these measures, e.g., for a scale-free network (for which the degree distribution p_k satisfies $p_k \propto k^{-\lambda}$), the average path length scales as $\log N$, the diameter Δ scales as $\log\log N$ for $2 < \lambda < 3$, and d_B and λ are not independent [3,6].

In this paper we extend Wei et al.'s analysis [24] of the information dimension d_I of a complex network and compare d_I to d_B. The study of d_I for a network is, compared to the study of d_B, quite recent, and specific applications of d_I to real-world problems have not yet appeared in the literature. Since the information dimension d_I of a network is a natural extension of d_I of a probability distribution [1,5,20], we begin by reviewing d_I of a distribution. 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. 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 intersec-

E-mail address: ericr@att.com.

http://dx.doi.org/10.1016/j.physleta.2016.12.015
0375-9601/

tion of each pair of boxes has measure zero). Define the probability $p_j(s)$ of box 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 [16]. Define the entropy $H(s)$ by

$$H(s) \equiv -\sum_{B_j \in \mathcal{B}(s)} p_j(s) \log p_j(s) . \quad (1)$$

Then the information dimension d_I of σ is given by $d_I \equiv -\lim_{s\to 0} H(s)/\log s$, assuming the limit exists [2].

Now consider an undirected, unweighted network $\mathcal{G} = (\mathcal{N}, \mathcal{A})$. To define d_I for $\mathcal{G}$, we require a few definitions. 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 in $\mathcal{G}$. 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. We do not consider $\mathcal{B}(s)$ for $s=1$, since a box of diameter 0 contains only a single node. For $s \geq \Delta + 1$, the covering $\mathcal{B}(s)$ consists of a single box which is $\mathcal{G}$ itself. It is convenient to define $\mathcal{S} \equiv \{s \mid s \text{ is integer and } 2 \leq s \leq \Delta\}$. For $s \in \mathcal{S}$, define $B(s) = |\mathcal{B}(s)|$, so $B(s)$ is the number of boxes in $\mathcal{B}(s)$. Finally, the covering $\mathcal{B}(s)$ is *minimal* if $B(s)$ is less than or equal to the number of boxes in any other covering of $\mathcal{G}$ of size s. In general, for $s \in \mathcal{S}$ we cannot easily compute a minimal covering of size s, but good heuristics are known, e.g., the method of Song et al. [21], which was utilized in [24]; see also [19], which provides a method for computing a lower bound on $B(s)$. Our computational results employ the heuristic of [13], described in Section 5.

With these definitions, we can now explain how d_I is computed for $\mathcal{G}$. For $s \in \mathcal{S}$, let $\mathcal{B}(s)$ be a minimal covering of $\mathcal{G}$. Let $N_j(s)$ be the number of nodes of $\mathcal{G}$ contained in box $B_j \in \mathcal{B}(s)$. We obtain a set of box probabilities $\{p_j\}$ from $\mathcal{B}(s)$ by defining $p_j(s) \equiv N_j(s)/N$. We then use (1) to compute the entropy $H(s)$. Roughly speaking, $\mathcal{G}$ has the information dimension d_I if $H(s) \sim -d_I \log(s/\Delta)$.

Just as we can define d_I for a complex network $\mathcal{G}$, we can also define d_B for $\mathcal{G}$ ([13], [21]). Roughly speaking, $\mathcal{G}$ has the box counting dimension d_B if the minimal covering $\mathcal{B}(s)$ follows the scaling law $B(s) \sim s^{-d_B}$ over some range of s. For geometric objects, the relationship between d_B and d_I is part of the theory of multifractals. A geometric multifractal is an object that cannot be completely described by a single fractal dimension, and instead is characterized by a family $\{D_q\}$, $q \in \mathbb{R}$ of generalized dimensions [16]. It is known that $D_0 = d_B$, that $D_1 = d_I$, and that $d_I \leq d_B$; more generally, D_q is nonincreasing in q for $q \geq 0$ [8]. Generalized dimensions have also been considered for complex networks [22]. The proof that the inequality $d_I \leq d_B$ holds for a probability distribution and its support does not extend to unweighted networks, since (as we discuss in Section 3), for unweighted networks we cannot take a limit as the box size approaches zero. Moreover, Wei et al. [24] compute d_B and d_I for four networks and report that $d_I > d_B$ for all four networks.

In this paper, we consider the definition given in [24] of d_I for a complex network $\mathcal{G}$, and recast this definition in a computationally useful form. We exhibit a small network $\widetilde{\mathcal{G}}$ for which $d_I > d_B$. We show that by using a different minimal covering of $\widetilde{\mathcal{G}}$ we now obtain $d_I < d_B$. Thus to determine whether $d_I \leq d_B$ holds for $\mathcal{G}$, it is not sufficient to determine a minimal covering for each box size s. A new framework is needed, and accordingly we propose the new notion of a maximal entropy minimal covering of $\mathcal{G}$, and a new definition of d_I for $\mathcal{G}$ based on maximal entropy minimal coverings. We examine four larger networks and find that, with this new definition of d_I, for three of them we have $d_I < d_B$; for the fourth, by a very narrow margin we have $d_I > d_B$. Moreover, the use of maximal entropy minimal coverings in many cases enhances the ability to compute d_I.

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 $\mathcal{G}$ of size s
$B(s)$	Cardinality of $\mathcal{B}(s)$
$B_j(s)$	Box in $\mathcal{B}(s)$
d_B	Box counting dimension
d_I	Information dimension
$\mathcal{G}$	Complex network
$\mathcal{G}(n,r)$	Subnetwork of $\mathcal{G}$ with center n and radius r
$H(s)$	Entropy computed from $\mathcal{B}(s)$
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 $B_j \in \mathcal{B}(s)$
$\mathcal{S}$	$\{s \mid s \text{ is integer and } 2 \leq s \leq \Delta\}$

2. The information dimension of a network

The information dimension d_I defined in [24, eq. (9)], for an unweighted, undirected network $\mathcal{G}$ is

$$d_I \equiv -\lim_{s\to 0} \frac{H(s)}{\log s} = \lim_{s\to 0} \frac{\sum_{B_j \in \mathcal{B}(s)} p_j(s) \log p_j(s)}{\log s} . \quad (2)$$

However, (2) is not computationally useful, since the distance between each pair of nodes is at least 1. Moreover, we cannot use the value $s=1$ in (2), since then the denominator of the fraction is zero, while for $s \geq 2$ we have $H(s) > 0$ and $\log s > 0$, which implies $-H(s)/\log s < 0$ and thus, from (2), $d_I < 0$. The fact that s cannot become arbitrarily small was recognized by Wei et al. [24], who call (2) a “theoretic formulation”. Instead of (2), we propose in Definition 1 below a computationally useful definition of d_I. By quantifying how $H(s)$ scales with $\log(s/\Delta)$, rather than how $H(s)$ scales with $\log s$, our definition has the same functional form as the definition of d_I in [9] and [10]. We require, as is typical in studying fractal dimensions, the existence of a “regime” or “plateau” over which the desired scaling relation holds approximately.

Definition 1. The network $\mathcal{G}$ has the information dimension d_I if for some constant c, for some positive integers L_I and U_I satisfying $2 \leq L_I < U_I \leq \Delta$, and for each integer $s \in [L_I, U_I]$,

$$H(s) \approx -d_I \log\left(\frac{s}{\Delta}\right) + c , \quad (3)$$

where $H(s)$ is defined by (1), $p_j(s) = N_j(s)/N$ for $B_j \in \mathcal{B}(s)$, and $\mathcal{B}(s)$ is a minimal covering of size s. □

The following example shows that $d_I = 3$ for a 3-dimensional cubic rectilinear lattice. The analysis has the obvious extension to a square E-dimensional rectilinear lattice, for any positive integer E.

Example 1. Let $\mathcal{G}(L)$ be a 3-dimensional cubic rectilinear lattice of L^3 nodes, where L is the number of nodes on an edge of the square (this is just the familiar primitive Bravais lattice in three dimensions). The diameter Δ of $\mathcal{G}(L)$ is $3(L-1)$. Suppose $L = 2^K$ for a given positive integer K. For $M = 1, 2, \cdots, K-1$, we can cover $\mathcal{G}(2^K)$ using copies of $\mathcal{G}(2^M)$. The number of copies required