Physica A 456 (2016) 157–166

Contents lists available at ScienceDirect

Physica A

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

On the likelihood of forests

Yilun Shang

Department of Mathematics, Tongji University, Shanghai 200092, China

HIGHLIGHTS

- We introduce forest likelihood to assess complexity of sparse networks without loops.
- Forest likelihood involves non-preferential attachments to construct a forest.
- Properties and examples for forest likelihood are presented.
- Analytical formula and algorithms to compute forest likelihood are given.
- Forest likelihood of synthetic networks and real networks are studied.

ARTICLE INFO

Article history:
Received 3 January 2016
Received in revised form 14 February 2016
Available online 29 March 2016

Keywords:
Likelihood
Growing networks
Forest
Graph complexity

ABSTRACT

How complex a network is crucially impacts its function and performance. In many modern applications, the networks involved have a growth property and sparse structures, which pose challenges to physicists and applied mathematicians. In this paper, we introduce the forest likelihood as a plausible measure to gauge how difficult it is to construct a forest in a non-preferential attachment way. Based on the notions of admittable labeling and path construction, we propose algorithms for computing the forest likelihood of a given forest. Concrete examples as well as the distributions of forest likelihoods for all forests with some fixed numbers of nodes are presented. Moreover, we illustrate the ideas on real-life networks, including a benzenoid tree, a mathematical family tree, and a peer-to-peer network.

1. Introduction

A large variety of real-life complex systems in nature, technology and society can be epitomized by networks with nodes interconnected by edges [1,2]. The dynamics and function of complex systems are intimately related to the structural complications of these underlying networks. Although complexity in networks, like *life* and *consciousness*, does not have a rigorous definition, it may be viewed, from an abstract perspective, as a function of state and rate of change over the components of a network. It often goes side by side with notions like connectivity, self-organization, nonlinearity, etc. Many graph-theoretic measures, based on different heuristics, have been proposed to gauge the *complexity* of networks. For instance, the complexity of a network has been defined to be the number of its spanning trees [3,4]. Linear complexity of a network is the smallest number of arithmetic operations required to compute Ax, where A is the adjacency matrix of the network [5]. Complexity can also be defined as the number of Boolean operations, such as union and intersection, required to generate the network starting from a star [6]. On top of that, motivated by applications in biology and chemistry, the notion of graph entropy was used to measure network complexity [7].

Nowadays, it is well-known that numerous real systems have a common property of growth, where the numbers of nodes and edges evolve over time [8–10]. The above mentioned complexity measures, based on either combinatorial or

E-mail address: shyl@tongji.edu.cn.

http://dx.doi.org/10.1016/j.physa.2016.03.021
0378-4371/

algebraic operations, unfortunately shed little light on the natural growth of networks. As a result, they seem to be not useful in quantifying topological complications within the context of network theory. Recently, Banerji et al. [11] proposed the notion of *graph likelihood* to characterize the complexity of growing networks. They grow graphs by attaching nodes one by one starting from a single node. At each subsequent step, the neighbors of the newly added node are chosen uniformly at random, and the number is a uniform random variable within the range defined by the number of nodes already present. The graph likelihood is then defined to be the probability that a given graph on N nodes is obtained in this manner after N steps. Clearly, every graph can be constructed in this way (an analogue of the so-called infinite monkey theorem [12]). Using a rooted tree decomposition algorithm, the authors in Ref. [11] obtained a closed formula for the graph likelihood. They explicitly computed the graph likelihood for some special graphs including complete graphs, paths, and cycles.

In the meantime, thanks to the progress of data science, many real-world networks are found to be sparse (see Table II in Ref. [13] and Table 1 in Ref. [14]), namely, the number of edges in the network of size N is about $O(N)$. Moreover, in most random network models the structure is locally tree-like [15–17]. It is imperative, therefore, to have complexity measures tailored for sparse networks.

Motivated by the above considerations, we in this paper develop a complexity measure accommodating sparse graphs, or more precisely, for forests, which are disjoint unions of tree components. This measure, which we referred to as *forest likelihood*, involves a natural non-preferential attachment similar to Ref. [11], but we are able to focus only on sparse graphs (see Remark 1). In Section 2, we define the forest likelihood and present some concrete examples. In Section 3, we describe algorithms to calculate the forest likelihood of a given forest, as well as the upper and lower bounds involving its automorphism group. By virtue of the notions of admittable labeling and path construction, we give a closed formula for the forest likelihood. In Section 4, we show the feasibility of our algorithms by computing the forest likelihoods for all forests with some fixed numbers of nodes. Applications to a benzenoid tree, a mathematical family tree, and a peer-to-peer network are provided. Finally, we conclude the paper in Section 5 with further remarks.

2. Forest likelihood and some examples

Throughout the paper, we use 'graph' and 'network' interchangeably. Let $G = (V(G), E(G))$ denote a graph with the node set $V(G)$ and the edge set $E(G)$. We construct the random graph sequence $\{G_t\}_{t\geq 1}$ as follows.

- Let $G_1 = K_1$, the graph on a single node.
- For $t \geq 2$, let $d(t)$ be a random variable with Bernoulli distribution Ber(1/2) and construct G_t by adding a new node to G_{t-1} and connecting it to a uniformly random subset of size $d(t)$ of $V(G_{t-1})$.

In other words, if we let $V(G_{t-1}) = \{v_1, v_2, \ldots, v_{t-1}\}$, then at the tth step we select a number $d(t) \in \{0, 1\}$ with equal probability. We then add a new node v_t and connect it to $d(t)$ randomly chosen node in $V(G_{t-1})$ with equal probability. (If $d(t) = 0$, v_t is isolated in G_t.) Obviously, no cycle appears in G_t, i.e., G_t forms a forest for $t \geq 1$.

The above construction reflects an interesting non-preferential attachment mechanism, where no local knowledge of the network is assumed except the labels of the nodes. To our knowledge, such non-preferential growing processes were first formulated by Janson and Severini [18], which constitute a natural counterpart of the Barabási–Albert model [9].

Definition 1. Let G be a forest on t nodes. The forest likelihood of G, denoted by $\mathcal{FL}(G)$, is defined as the probability that $G_t = G$, where G_t is obtained by the above construction. Namely, $\mathcal{FL}(G) = \mathbb{P}(G_t = G)$.

Remark 1. Clearly, if we relax $d(t) \in \{0, 1\}$ to $d(t) \in \{0, 1, \ldots, t-1\}$, we readily recover the (*graph*) *likelihood* studied in Ref. [11]. However, we clarify that the forest likelihood is different from conditional likelihood given that G is a forest. The reason is that a forest can also be generated by attaching two or more edges at a time. A simple example consists of adding an isolated node in the 2nd step, and adding a node with degree 2 at the 3rd step. The idea of conditional likelihood is natural but its computation seems to be a formidable task. This partially motivated our consideration of forest likelihood.

Proposition 1. *Let $K_{1,t-1}$ be the star graph on t nodes. Then $\mathcal{FL}(K_{1,t-1}) = 1/2$ for $t = 2$ and $\mathcal{FL}(K_{1,t-1}) = \frac{1}{2^{t-2}(t-1)!}$ for $t \geq 3$.*

Proof. For $t = 2$, $\mathcal{FL}(K_{1,1}) = \mathbb{P}(d(2) = 1) \cdot \mathbb{P}(\{v_1, v_2\} \in E(G_2)) = 1/2 \cdot 1 = 1/2$.

For $t \geq 3$, we divide it into two cases (see Fig. 1).

(i) Since v_1 is the internal node, we have $\mathbb{P}(d(2) = 1) \cdot \mathbb{P}(\{v_1, v_2\} \in E(G_t)) = 1/2$, and similarly for $i \geq 3$, $\mathbb{P}(d(i) = 1) \cdot \mathbb{P}(\{v_1, v_i\} \in E(G_t)) = \frac{1}{2} \cdot \frac{1}{i-1}$. Therefore,

$$\mathcal{FL}(K_{1,t-1} \text{ in (i)}) = \prod_{i=2}^{t} \left(\frac{1}{2} \cdot \frac{1}{i-1}\right) = \frac{1}{2^{t-1}(t-1)!}.$$

(ii) Reasoning similarly as above, we have

$$\mathcal{FL}(K_{1,t-1} \text{ in (ii)}) = \prod_{i=2}^{t} \left(\frac{1}{2} \cdot \frac{1}{i-1}\right) = \frac{1}{2^{t-1}(t-1)!}.$$

Combining (i) and (ii), we obtain the desired result. □