Physica A 509 (2018) 588–598

Contents lists available at ScienceDirect

Physica A

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

Master equation for the degree distribution of a Duplication and Divergence network

Vítor Sudbrack, Leonardo G. Brunnet, Rita M.C. de Almeida, Ricardo M. Ferreira, Daniel Gamermann*

Instituto de Física - Universidade Federal do Rio Grande do Sul (UFRGS), Av. Bento Gonçalves, 9500, Brazil

HIGHLIGHTS

- Modeling of network evolution dynamics through Markov chains.
- Determination of the model's parameter range for which the network reaches a steady state.
- Evaluation of the asymptotic network degree distribution.
- Original and copied nodes mutation rates are treated independently.

ARTICLE INFO

Article history:
Received 7 December 2017
Received in revised form 10 April 2018
Available online xxxx

Keywords:
Duplication and divergence
Markov process
Network evolution

ABSTRACT

Network growth as described by the Duplication–Divergence model proposes a simple general idea for the evolution dynamics of natural networks. In particular it is an alternative to the well known Barabási–Albert model when applied to protein–protein interaction networks. In this work we derive a master equation for the node degree distribution of networks growing via Duplication and Divergence and we obtain an expression for the total number of links and for the degree distribution as a function of the number of nodes. Using algebra tools we investigate the degree distribution asymptotic behavior. Analytic results show that the network nodes average degree converges if the total mutation rate is greater than 0.5 and diverges otherwise. Treating original and duplicated node mutation rates as independent parameters has no effect on this result. However, difference in these parameters results in a slower rate of convergence and in different degree distributions. The more different these parameters are, the denser the tail of the distribution. We compare the solutions obtained with simulated networks. These results are in good agreement with the expected values from the derived expressions. The method developed is a robust tool to investigate other models for network growing dynamics.

1. Introduction

Physicists are required to build models that extract the essence of observable phenomena seen in nature in order to understand and describe them. As physics endeavors in studying complex phenomena in distinct fields such as biology or social sciences, a commonly applied paradigm is the use of graph theory. In this approach the system under study is described as a network consistent of a set of nodes and a set of links among them.

Examples of systems studied within this approach are social networks [1], author citations [2], flights connections [3], metabolic models [4,5], protein–protein interactions [6,7], electrical grids [8,9], among others.

* Corresponding author.
E-mail address: danielg@if.ufrgs.br (D. Gamermann).

https://doi.org/10.1016/j.physa.2018.06.066
0378-4371/

In the present work we are interested in *Network dynamics*, i.e. the study of the dynamics behind a growing graph. The motivation lies in the dynamics of biological networks, given that many such systems find a natural description within those models. Understanding the evolutionary selection rules resulting in networks with similar topological characteristics as the real observed ones, may give insights about the underlying biological processes (natural selection) behind these structures. For example, the protein association networks can help understand the evolution of species' genomes [10–13].

In the present work, the network dynamics is described as a Markovian process. Within this approach, the network state at a given time depends only on its configuration on the previous moment. Given a set of rules that describe how the network changes in each time step, we construct the corresponding master equation representing the evolution of the system's configurations. A similar methodology was used in the work by Ferreira et al. [14]. With this approach, they presented analytical results and simulations of networks growing according to the Barábasi–Albert rule. Here we focus on a complementary approach to model protein–protein network dynamics [15] and explore the evolution of an adapted version of the Duplication Divergence model [16]. This model is usually applied to study the evolution of proteins related networks [17–20]. The importance in describing the average behavior of stochastic processes in this manner is to know the network behavior for different values of the parameters without the need of long, time consuming, numerical simulations to obtain statistically relevant information.

This article is organized as follows. In the next section we focus in explaining the Duplication and Divergence model. Following the model explanation, we derive an expression for the total number of links as a function of the number of nodes, which gives us a straight forward way to obtain the mean degree of the graph. Then we study the graph growth as a Markovian chain, in which the next degree distribution of the network is a function of the current degree distribution, pondered by the probabilities of all possible occurrences in each time step. Finally, we study the asymptotic limit of the degree distribution.

2. The model

Given an initial small network (three nodes connected to each other forming a triangle[1]) we study the Markov process where, in each time step a node of the network is randomly chosen to be copied i.e. a new node is created with exactly the same neighbors as the chosen one. In what follows we refer to the copied node as original and its copy as duplicated. After duplication, original and duplicated nodes may diverge, meaning that each link of the original node is lost with probability m_o and each link in the duplicated node is lost with probability m_d. Also, a link between original and the duplicated nodes is always added.

This model is an adaptation of the Duplication and Divergence model, originally developed by Vázquez et al. [16,21]. In the model proposed by Vázquez a new node is also added by copying an existing node and all its links. New node and ancestor are linked with a probability p. Also, either the link between new node and a third neighbor or the link between ancestor and this neighbor is lost with probability q [21]. Our model treats independently the loss of a link by original and duplicated nodes and sets $p = 1$.

In Vázquez's model the network growth is based on local rules, that is, rules that require only information on one node instead of rules that require information over all the network. This model of network dynamics applied to protein–protein interaction networks allows all proteins to evolve from a common ancestor through gene copies (represented by duplications) and mutations (divergence). Therefore, it would mimic the entire history of a genome evolution [21].

3. Mean degree and total number of links

First, let us evaluate the behavior of the network average node degree $\bar{k}$ as a function of the number of nodes in the network, t. Given that a node with degree k is chosen to be duplicated, the number of links in the next step changes. The mean change in the number of links is given by[2]:

$$\begin{aligned} L_{t+1} &= L_t + \overline{(k - m_o k - m_d k + 1)} \\ &= L_t + (1 - m_o - m_d)\bar{k} + 1 \end{aligned} \tag{1}$$

Terms in the right-hand side of Eq. (1) represent, from left to right: existing links, links added due to the duplication of a k-degree node, mean number of lost links of the original node, mean number of lost links for the duplicated node and creation of the original–duplicated link.

Let us define the total mutation parameter M as the sum of the independent mutation parameters m_o and m_d: $M = m_o + m_d$. Now, given the relation between node's degree distribution and number of links in a network, known as the

1 As will be discussed latter on, the asymptotic behavior of the graph's properties is independent of this choice of seed (initial) graph.

2 The bar over a term in the equation indicates that the average of the process will be considered.