Theoretical Computer Science 476 (2013) 94–108

Contents lists available at SciVerse ScienceDirect

Theoretical Computer Science

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

Degree distribution of large networks generated by the partial duplication model

Si Li [a], Kwok Pui Choi [a,b], Taoyang Wu [a,*]

[a] Department of Mathematics, National University of Singapore, Singapore 119076, Singapore

[b] Department of Statistics and Applied Probability, National University of Singapore, Singapore 119076, Singapore

ARTICLE INFO

Article history:
Received 25 September 2012
Accepted 25 December 2012
Communicated by P. Spirakis

Keywords:
Random graph
Power law
Limiting behavior
Degree distribution
Computational proteomics

ABSTRACT

In this paper, we present a rigorous analysis on the limiting behavior of the degree distribution of the partial duplication model, a random network growth model in the duplication and divergence family that is popular in the study of biological networks. We show that for each non-negative integer k, the expected proportion of nodes of degree k approaches a limit as the network becomes large. This fills in a gap in previous studies. In addition, we prove that $p = 1/2$, where p is the selection probability of the model, is the phase transition for the expected proportion of isolated nodes converging to 1, and hence answer a question raised in Bebek et al. [G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, S.C. Sahinalp, The degree distribution of the generalized duplication model, Theoret. Comput. Sci. 369 (2006) 239–249]. We also obtain asymptotic bounds on the convergence rates of degree distribution. Since the observed networks typically do not contain isolated nodes, we study the subgraph consisting of all non-isolated nodes contained in the networks generated by the partial duplication model, and show that $p = 1/2$ is again a phase transition for the limiting behavior of its degree distribution.

1. Introduction

Over the past decade, networks have been used to elucidate many complex systems in different disciplines, including computer science, biology, technology and social science. In biology, a network provides a useful tool to represent and study interaction data of different types in cellular systems, such as protein–protein interaction, metabolic and gene regulation [1]. By investigating the interactions at a network level, new insights into the molecular mechanisms behind these systems can be discovered [2]. For example, a protein–protein interaction (PPI) network of the plant *Arabidopsis thaliana* containing about 6200 physical interactions between about 2700 proteins was constructed recently by [3], and a study based on it by [4] indicated how pathogens may exploit protein interactions to manipulate a plant's cellular machinery.

Since cellular networks in biology are often huge and complex, they are typically modeled in the framework of random networks, which enables simulation, inference and prediction to be made. The most studied random network model in mathematics is the Erdös–Rényi (ER) model proposed by [5], in which each pair of nodes is connected independently with a probability specified by the model. However, the degree distributions obtained from this model approximately follow Poisson distributions, which do not exhibit the heavy-tailed phenomenon commonly observed in the empirical degree distributions of many real networks, such as the World Wide Web, PPI network of budding yeast, and metabolic network of *Escherichia coli*. To capture such heavy-tailed phenomena, the preferential attachment (PA) model was popularized by [6], where a new node is added at each step and connected to a fixed number of nodes that are chosen with probabilities

* Correspondence to: School of Computing Sciences, University of East Anglia, Norwich NR4 7TJ, UK. Tel.: +44 1603 59 2954; fax: +44 1603 593345.
E-mail addresses: g0800874@nus.edu.sg (S. Li), stackp@nus.edu.sg (K.P. Choi), taoyang.wu@gmail.com (T. Wu).

0304-3975/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2012.12.045

proportional to their degrees. A characteristic feature of the PA model is that it produces networks with degree distributions following power laws. This fits well with many observed networks, although the universality of power laws is being questioned [7].

In modeling biological networks, one should not simply focus on whether the summary statistics produced by the random networks fit well with those of the biological networks, but whether the modeling is biologically supported. In proteome, gene duplication followed by mutation is believed to be the main driving force behind protein evolution [8,9]. The effect of this mechanism on PPI network evolution was described by [10]: immediately after a gene is duplicated, the new node representing this gene copies all the interactions of the duplicated gene, and then the edges adjacent to the duplicated node or new node are randomly lost. This mechanism was formulated into a network growth model by [11]. Since then, many variations and extensions have been proposed and studied [12–18]. We refer to this family of network growth models as duplication and divergence (DD) models.

Arguably, one of the most fundamental models in the DD family is the partial duplication (PD) model studied by [15]. In this model, at each step an anchor node is chosen uniformly from the current network and a new node is added and independently this new node is connected to each neighbor of the anchor node with selection probability p (see Section 2 for more details). This model is particularly attractive for two reasons: it captures the basic principles behind PPI evolution, and its simplicity enables us to conduct rigorous mathematical analysis. By studying this model we can gain insights into other more sophisticated DD models.

Here we focus on the degree distribution of the PD model. By degree distribution we mean the sequence $\{f_t(k)\}_{k\geq 0}$, where $f_t(k)$ denotes the *expected proportion* of nodes of degree k at time t. Note that the PD model is studied at the ensemble level in this paper, that is, we are mainly interested in the average behavior over many different realizations. One general tool to study the degree distribution of random networks is the master equation of $f_t(k)$ (see [19] and the references therein). However, despite the simplicity of the PD model, its master equation is still too complicated to be solved analytically and no analytic solution is known yet, except for the full duplication model, the special case when $p = 1$ [20]. Instead, the attention has been centered on the limiting degree distribution, which provides valuable information on the long run behavior of the model [15,16,21].

Prior to studying limiting degree distribution, we need to establish its existence, that is, whether the limit of $f_t(k)$ for a given k exists as t approaches infinity. For the special case $k = 0$, the existence of $f(0) = \lim_{t\to\infty} f_t(0)$ was proved by [16] by showing $\{f_t(0)\}_{t\geq 0}$ is indeed a non-decreasing sequence. However, the other cases remained open and it was often *assumed* that they do exist in previous studies. For example, Lemma 2 in [16] states that for $k \geq 1$, if $f_t(k)$ tends to a limit, then this limit must be 0. In this paper, we close this gap by showing that the limit of $f_t(k)$ *does* exist for each $k \geq 0$, and hence the sequence $(f_t(0), f_t(1), f_t(2), \ldots)$ converges pointwise to $(f(0), 0, 0, \ldots)$ as t approaches infinity.

An important property of the PD model is that it may produce graphs containing a large proportion of isolated nodes, that is, $f(0)$ is typically large when p is small. Therefore, it is of interest to know the behavior of $f(0)$ relative to selection probability p. Indeed, one central problem for the PD model, as stated in [16, Section 3.1], is to characterize the values of p for which $f_t(0)$ tends to 1. Here we answer this question by showing that $p = 1/2$ is the phase transition for the expected proportion of isolated nodes converging to 1. More precisely, we prove that $f(0) < 1$ for $1/2 \leq p \leq 1$, and $f(0) = 1$ for $0 < p < 1/2$. In addition, we also obtain upper and lower asymptotic bounds on the convergence rate of $\{f_t(0)\}_{t\geq 0}$, as well as a uniform upper bound on the convergence rate of $\{f_t(k)\}_{t\geq 0}$ for all $k \geq 1$.

Since isolated nodes are generally irrelevant to the observed PPI networks, here we also study the subgraph consisting of all non-isolated nodes in the PD model. Interestingly, $p = 1/2$ turns out again to be a phase transition for the limiting degree distribution. When $1/2 \leq p \leq 1$, the limiting degree distribution does exist and it is $(0, 0, \ldots)$, that is, the expected fraction of degree k in this subgraph tends to 0 for all $k \geq 1$. Therefore, the limiting degree distribution does not follow a power law in this region. However, the case when $0 < p < 1/2$ is more delicate. With the assumption that the limiting degree distribution exists, we prove that the entries in this limiting distribution must be strictly positive, and they satisfy a system of equations. In addition, the limiting degree distribution in this region also follows a power law. Our results are then applied to three real PPI networks to obtain the power law exponent and selection probability for each network.

The structure of the rest of this paper is as follows. In the next section, we describe the PD model and the master equation for the expected degree sequence. In Section 3, we present some preliminary results. In Section 4, we establish the existence of limiting degree distribution and show that $p = 1/2$ is the phase transition for the expected fraction of isolated nodes converging to 1. Section 5 is devoted to the bounds on rates of convergence. In Section 6 we study the limiting degree distribution of the subgraph with all isolated nodes removed, and apply the results to three real PPI networks. Finally, we end with Section 7 for some concluding comments and possible directions for further study.

2. The model

All networks studied in this paper are undirected; they are also referred to as graphs. In the partial duplication model $\mathcal{M}(G_{t_0}, p)$, where G_{t_0} is the seed graph and $0 < p \leq 1$ is the *selection probability* of the model, we start with G_{t_0} and at each time step t, the graph G_t is obtained from G_{t-1} by the following procedures: A node u_t is chosen uniformly from the set of nodes in G_{t-1}, and a new node v_t is added and independently connected to each neighbor of u_t with probability p (see Fig. 1 for an illustration). Here u_t and v_t are often referred to as the anchor node and new node at step t, respectively. Throughout