Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 369 (2006) 234–249
Theoretical Computer Science
www.elsevier.com/locate/tcs

The degree distribution of the generalized duplication model

G. Bebek[a], P. Berenbrink[b], C. Cooper[c], T. Friedetzky[d], J. Nadeau[e], S.C. Sahinalp[b,*]

[a] *Department of EECS, CWRU, USA*
[b] *School of Computing Science, SFU, Burnaby, BC, Canada*
[c] *Department of Computer Science, King's College, London, UK*
[d] *Department of Computer Science, University of Durham, UK*
[e] *Department of Genetics, CWRU, USA*

Received 21 June 2005; received in revised form 18 January 2006; accepted 30 August 2006

Communicated by A. Apostolico

Abstract

We study and generalize the duplication model of Pastor-Satorras et al. [Evolving protein interaction networks through gene duplication, J. Theor. Biol. 222 (2003) 199–210]. This model generates a graph by iteratively "duplicating" a randomly chosen node as follows: we start at t_0 with a fixed graph $G(t_0)$ of size t_0. At each step $t > t_0$ a new node v_t is added. The node v_t selects an existing node u from $V(G(t-1)) = \{v_1, \ldots, v_{t-1}\}$ uniformly at random (uar). The node v_t then connects to each neighbor of the node u in $G(t-1)$ independently with probability p. Additionally, v_t connects uar to every node of $V(G(t-1))$ independently with probability r/t, and parallel edges are merged. Unlike other copy-based models, the degree of the node v_t in this model is not fixed in advance; rather it depends strongly on the degree of the original node u it selected.

Our main contributions are as follows: we show that (1) the duplication model of Pastor-Satorras et al. does not generate a truncated power-law degree distribution as stated in Pastor-Satorras et al. [Evolving protein interaction networks through gene duplication, J. Theor. Biol. 222 (2003) 199–210]. (2) The special case where $r = 0$ does not give a power-law degree distribution as stated in Chung et al. [Duplication models for biological networks, J. Comput. Biol. 10 (2003) 677–687]. (3) We generalize the Pastor-Satorras et al. duplication process to ensure (if required) that the minimum degree of all vertices is positive. We prove that this generalized model has a power-law degree distribution.

Keywords: Random graph generation; Power law; Computational proteomics

1. Introduction

A proteome network of an organism is a graph in which each node represents a protein and each edge represents an interaction between a pair of proteins. Recent studies on the proteome network of the yeast *Saccharomyces Cerevisiae* [28,33] suggests that the degree distribution is in the form of a *power-law* [19,30]. Power-law degree distributions have previously been observed in a number of naturally occurring graphs such as the internet graph, the web graph, peer-to-peer networks, etc. [1,5,11,13,17,21,22].

* Corresponding author. Tel.: +1 604 291 5415.
E-mail addresses: gurkan@case.edu (G. Bebek), petra@cs.sfu.ca (P. Berenbrink), ccooper@dcs.kcl.ac.uk (C. Cooper), tom@friedetzky.org (T. Friedetzky), jhn4@po.cwru.edu (J. Nadeau), cenk@cs.sfu.ca (S.C. Sahinalp).

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2006.08.045

The classical random graph models studied by Erdös and Rényi [16] (in which edges between pairs of nodes are determined independently) do not have a power-law degree distribution. However, there are a number of recently developed alternative random graph models which *do* generate power-law degree distributions; see for example Barabási and Albert [5], Bollobás et al. [11], Watts [31], or Aiello et al. [1,2]. (See also the surveys by Albert and Barabási [4], Bollobás and Riordan [9], Hayes [18], and Mitzenmacher [23] for more models and details.)

Among these models, many [11,8,13,21,22] are based on an iterative random graph generation process which adds new nodes and/or edges to the graph in each iteration. The new node is connected to ℓ of the existing nodes where ℓ is a fixed constant or an independent random variable. The way the number and the endpoints of these ℓ edges are chosen determines the specific graph generation model. For example, in the *preferential attachment model* the probability that an existing node is connected to the newly created node increases with the degree of the node. Another example is the *uniform model* in which the newly created node is connected to other nodes that are simply picked uniformly at random.

The preferential attachment model dates back to Yule [34] and Simon [27]. It was proposed as a random graph model for the web by Barabási and Albert [5], and their description was elaborated by Bollobás and Riordan [10] who showed that with high probability the diameter of a graph constructed in this way was $\sim \log t / \log\log t$—here t stands for the time step and thus (is approximately) the number of nodes. Subsequently, Bollobás et al. [11] proved that the degree sequence of such graphs does follow a power-law distribution. Attention has also been given to models where the attractiveness of vertices fades over time, for example [14]. More recently Cooper and Frieze [13] gave a general analysis of random graph processes revealing that many graphs generated by preferential attachment exhibit power-law degree distributions. This analysis, and those of [3,15,22]; obtained graphs with a power-law parameter larger than 2 but smaller than 3 by using a graph generation model that allows edge insertion between existing nodes.

In a number of naturally occurring graphs with power-law degree distributions such as proteome networks, peer-to-peer networks and in a limited way the web graph, the mechanism underlying the growth process seems to be different from that in preferential attachment models. Rather, these networks seem to grow via node duplications. For example, Ohno's theory [24] of genome evolution states that the main driving force behind proteome growth is gene duplication followed by point mutations. [1] In peer-to-peer networks, a new user typically chooses the servers used by an existing node; similarly, new web pages tend to share content with related web pages. Possibly the first analysis of a duplication-based random graph model is given in [22]; this model generates directed graphs with constant outdegree. A more general version of this directed model was later analyzed in [13]. In both of these models the (out)degree of the newly generated node is bounded by a constant and does not depend on the degree of the duplicated node.

A less constrained duplication-based random graph model where the degree of the newly generated node is not bounded by a constant was recently introduced in [6,25,29]. In this model, at each iteration t, one existing node is chosen uniformly at random and is "duplicated" with all its edges. Then, in a "divergence" move, (i) each existing edge of the new node is deleted with probability q and (ii) a new edge is generated between the new node and every other node with probability r/t. This last step is referred to as "mutation" by some authors (e.g. [20]).

The first analytical work on the above model was by Pastor-Satorras et al. [25] (henceforth the *Pastor-Satorras et al. duplication model*) which suggested that its degree distribution is a "power law with exponential cut-off". This means that f_k, the fraction of nodes with degree k among all nodes, is independent of time and is approximated by $f_k = ck^{-b} \cdot a^{-k}$; here a, b, c are constants. However, the analysis in [25] makes a number of simplifying assumptions in to get this result. For instance, it approximates the probability of generating a node with degree k by the probability of duplicating a node with degree $k+1$ only and subsequently deleting one of its edges.

An analysis of the degree distribution of the Pastor-Satorras et al. duplication model, for the special case that $r = 0$, is given by Chung et al. [12]. Following [12], we will refer to this special case as the *pure duplication model*. This model creates many *singleton* nodes, i.e. nodes that are not connected to any other node of the graph. Since a node can get a new edge only if one of its neighbors is duplicated, a singleton will remain a singleton during the whole graph generation process. The pure duplication model creates a network in which all non-singleton nodes form a single connected component. In contrast to [25], Chung et al. suggest that the fraction of nodes with degree k is independent

[1] There are graph generation models which are not based on node duplications that seem to better capture certain other properties of the yeast proteome network; e.g. geometric graph generators [26] seem to better approximate the distribution of subgraphs with $\leqslant 5$ nodes. The focus of this paper is the degree distribution of the yeast proteome network and we only consider duplication-based network generators that aim to emulate Ohno's model of genome evolution.

of time and is a power-law distribution of the form $f_k = ck^{-b}$; here b is a function of q given by Eq. (1); values of $b \leqslant 2$ are possible for some q.

A general limiting analysis of the Pastor-Satorras et al. duplication model using generating functions is given by Kim et al. [20]. This paper derives (1) and obtains the expected number of edges (Lemma 7). It also studies the mean component size and threshold value of r for a giant component in the limit as $q \to 1$. It finds that a giant component exists for $r > \frac{1}{4}$ and predicts a power-law component size distribution for the subcritical case $r < \frac{1}{4}$. The precise power-law parameter is $\tau = 1 + 2/(1 - \sqrt{1 - 4r})$.

1.1. Summary of our contributions

(1) We show that the degree distribution of the Pastor-Satorras et al. duplication model cannot be a power law with exponential cut-off as stated in [25]; rather, it is a (regular) power law, provided $r > 0$ and $1 - q \leqslant 0.58$. (2) We show that, for the pure duplication model ($r = 0$) the fraction of nodes with degree k cannot be independent of time and cannot be a power-law distribution of the form $f_k = ck^{-b}$ as stated in [12]. This is due to the fact that the fraction of singletons increases with time in the pure duplication model. (3) We finally show that it is possible to slightly modify the pure duplication model so that it does not generate any singletons and achieves a power-law degree distribution consistent with the work of [12]. These are first results that establish power-law degree distributions for graph models where the degree of a copied node is determined strongly by the degree of the original node.

1.2. Details of our results and the organization of the paper

We first show in Section 3 that the (expected) fraction of singletons generated by the pure duplication model ($r = 0$) grows in time. In fact, the only limiting (time independent) solution is $f_0 = 1$ and $f_k = 0$ for all $k > 0$. Note that for the case $q = 0.5$ the average degree of nodes in the pure duplication model does not change over time (see Lemma 1). Together with the fact that the fraction of singletons increases in time, this implies that (i) the average degree of non-singletons must increase in time and (ii) there is a single connected component of size $o(t)$ with increasing average degree. It is quite possible that this connected component of the network generated by the pure duplication model exhibits a power law with parameter $b \leqslant 2$, however this is difficult to establish.

In the rest of Section 3, we show that the degree distribution of the generalized duplication model (in fact, any random model based on duplications) is not a "power law with exponential cut-off" as stated in [25]. We achieve this by showing a bound for the maximum degree of the generalized duplication model and contrasting it with that of a network which exhibits power law with exponential cut-off.

We also generalize the Pastor-Satorras et al. duplication model (in Section 2) so that each iteration has an additional (optional) edge generation step. For this generalized duplication model we show in Section 4 that: (i) not too many singletons are generated; (ii) the degree distribution of the nodes exhibit a power law, i.e. is of the form $f_k = ck^{-b}$.

Here, for $p = 1 - q$, the power-law parameter b is given by

$$1 = bp - p + p^{b-1}. \tag{1}$$

This equation holds irrespective of the model variant or the value of r. Similar results (with varying notation) are given in [12] for the pure duplication model, and in [20] for the Pastor-Satorras et al. model. The equation is problematic, in that $b = 1$ is always a solution, and for $p \geqslant p^* \sim 0.58$ it is the only solution [12]. We interpret this to mean that a power-law degree distribution no longer applies for $p \geqslant p^*$. For $p < p^*$ there are two other solutions $b_1 > 1$, $b_2 < 1$. Choosing the larger value b_1, the power laws predicted by (1) take values less than 3 for $p > \sqrt{2} - 1$ and less than 2 for $\frac{1}{2} < p < p^*$.

1.3. Summary of notation

We use t for the discrete time step, $G(t)$ for the graph at step t, and $C(t)$ for the largest component (where appropriate). The vertex set $V(t)$ is of size t, and v_t the vertex added at step t. We use k for vertex degree, and f_k for the expected limiting proportion of vertices of degree k. The power-law parameter is b. The probability that an edge is retained on duplication is p (i.e. $(1 - \delta)$ in the Pastor-Satorras et al. model [25]) and $q = 1 - p$. The parameter r is from r/t, the probability that a vertex of $G(t)$ is chosen uniformly at random (uar) at step $t + 1$.

2. Formal description of the duplication models

The focus of this paper is the duplication model considered in [12, 25], which grow iteratively in discrete time steps. The model starts with an arbitrary connected network $G(t_0)$, of size t_0. For $t > t_0$, let $G(t-1)$ be the network at the end of time step $t-1$. At iteration t, exactly one new node, denoted as v_t, is added to $G(t-1)$ as follows:

A node w is picked uniformly at random from $G(t-1)$, and w is "duplicated" to create the new node v_t which is initially connected to all the neighbors $N_{t-1}(w)$ of w, but not to w itself. The edges initially incident to v_t are then updated in the following way:

Step 1. Duplication: Each edge $e = (v_t, u)$, $u \in N_{t-1}(w)$ is independently deleted with probability q or retained with probability $p = 1 - q$.

Step 2. Uar edge addition: Each node u of $G(t-1)$ is independently connected to v_t with probability $r/(t-1)$, where r is a non-negative constant of the process, and any parallel edges created are merged.

In the Pastor-Satorras et al. model, iteration t is completed at this point. In this paper we consider a possible further uar edge addition step. The purpose of this step is to maintain connectivity of $G(t)$, and optionally, to restrict the number of the edges added during the uar step (put $r = 0$).

Version 1: *Step* 3. If v_t has become a singleton at the end of the duplication move, it is connected to $a_1 \geqslant 1$ uniformly chosen random nodes.

Version 2: *Step* 3. The vertex v_t is connected to $a_2 \geqslant 1$ additional nodes chosen uniformly at random. This occurs even if v_t did not become a singleton at the end of the duplication move.

Thus in either version $i = 1, 2$, the minimum degree is at least a_i. We remark that these additional edge insertions (Step 3) are made after duplication (Step 1) and (in the case that $r > 0$) without regard to the number of edges inserted by uar edge addition (Step 2). This allows us to choose the parameter $r = 0$ if we so wish, and yet maintain connectivity of the graph $G(t)$. Let δ_i, $i = 1, 2$ be the indicator for model Version i. We refer to the special case $r + \delta_1 + \delta_2 = 0$ as the *pure duplication model*, and the case where $r + \delta_1 + \delta_2 > 0$ as a *generalized duplication model*.

We now give a number of definitions relating to vertex degree which we use in our analysis. For the node v_s, added at step s, denote its degree (or expected degree if the context is clear) at time step $t \geqslant s$ by $d_s(t)$. Let $\boldsymbol{F}_k(t)$ denote the number of nodes of degree k at the end of step t and let $\boldsymbol{F}(t) = (\boldsymbol{F}_0(t), \boldsymbol{F}_1(t), \ldots)$ be the degree sequence. Also let $F_k(t) = \mathbf{E}\boldsymbol{F}_k(t)$ be the expected value, and $f_k(t) = F_k(t)/t$ the expected fraction of nodes of degree k. We say a model has a power law degree sequence if we can find constants $b, c > 0$ such that $f_k(t) \to f_k$ as $t \to \infty$ where $f_k = (1 + \mathrm{O}(1/k))ck^{-b}$. Finally let $\boldsymbol{e}(t) = |E(G(t))|$ be the number of edges in $G(t)$ and $e(t) = \mathbf{E}\boldsymbol{e}(t)$.

3. A discussion on the properties of the Pastor-Satorras et al. duplication model

In what follows we assume that $f_k(t) \to f_k$ as $t \to \infty$, i.e. there is a meaningful limiting distribution of the proportional degree sequence. Given this assumption there are two further possibilities namely $\sum f_k = 1$ and $\sum f_k < 1$. The second case, corresponds to the case where the limiting distribution is defective ($f_\infty > 0$) which is usually identified with the existence of a giant component. This occurs for example when $p = 1$ where the minimum vertex degree grows linearly with t. It is easily shown (see Lemma 1) that the expected average degree in (e.g.) the pure duplication model is of order t^{2p-1}, so it is certainly the case that the solution is not defective for $p < \frac{1}{2}$. It is unknown at what value of p the limiting distribution becomes defective.

We start by showing in Section 3.1 that the fraction of singletons in the pure duplication model grows with time in such a way that $f_k(t) \to 0$ for $k \geqslant 1$ and thus $f_0(t) \to 1$, is the only limiting solution compatible with $\sum_k f_k = 1$. For the particularly interesting case that $p = q = \frac{1}{2}$, we show that the expected number of non-singletons at time step t is between $\mathrm{O}(\sqrt{t})$ and $\mathrm{O}(t/\log\log t)$. Thus, without some modification, the pure duplication model cannot have a power-law degree distribution in the form $F_k(t) \sim ctk^{-b}$ for any constants c, b.

Section 3.2 is on the analysis Pastor-Satorras et al. model. In [25], it is stated that the generalized duplication model has a degree distribution following a "power law with exponential cut-off"; i.e. there exists constants a, b, c such that, as $t \to \infty$, we have $f_k(t) \sim ck^{-b}a^{-k}$ for $k \to \infty$. We show that this cannot be true by demonstrating that the expected maximum degree of a graph with degree distribution in the form of a power law with exponential cut-off is $\mathrm{O}(\log t)$, whereas the generalized duplication model has an expected maximum degree of $\Omega(t^p)$ for any combination of $r + \delta_1 + \delta_2 > 0$.

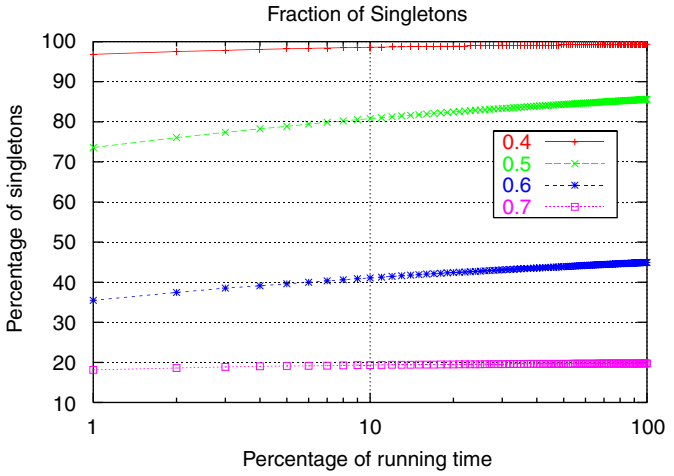

Fig. 1. Percentage of singletons in the pure duplication model as function of time (each curve is for a different value of p).

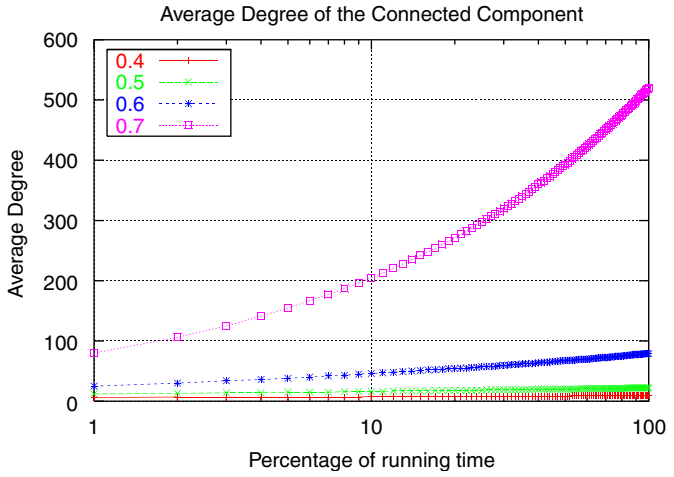

Fig. 2. Average degree of non-singleton nodes in the pure duplication model as function of time (each curve is for a different value of p).

3.1. Properties of the pure duplication model

Let $0 < p < 1$ be fixed. Assuming that the initial subgraph $G(t_0)$ is connected, then $G(t)$ consists of a unique connected component $C(t)$, and isolated vertices (singletons). Initially, $C(t_0) = G(t_0)$. Inductively, at step t, either v_t is a singleton; or, if vertex v_t retains at least one edge during duplication, it will have chosen a vertex in $C(t-1)$ to duplicate, and hence be connected to that component. The central question for this model are does $f_0(t)$ tend to 1, i.e. does $|C(t)|/t \to 0$, and if so, for which values of p?

As an illustration of the behavior of this model we first give some simulation results. In both cases below, the model was run until 1,000,000 non-singleton nodes were created. It can be seen that convergence to the steady state is very slow (if at all). The figures described below are at the end of the paper.

Fig. 1 shows the percentage of the singletons in the network over the time for different values of p. The plot uses a linear scale on the y-axis (percentage of singletons) and a logarithmic scale on the x-axis (running time).

Fig. 2 depicts the average degree over time for different values of p. The average degree of the network increases with time, and the larger the value of p, the larger is the increase.

We first prove a lemma giving the expected number of edges in the pure duplication model.

Lemma 1. *The expected total number of edges at step t satisfies*

$$e(t) \sim e(0)t^{2p}.$$

Proof. The number of edges at time $t+1$ in terms of the number of edges at time t is

$$\mathbf{E}(\boldsymbol{e}(t+1) \mid \boldsymbol{e}(t)) = \boldsymbol{e}(t) + \frac{1}{t}\sum_{s \leqslant t} p d_s(t).$$

The second term is obtained by considering the possibility that each given node v_s is duplicated at time t; then $pd_s(t)$ would be the expected number of its edges retained. Because the sum of the degrees of all nodes is twice the number of edges, we have, taking expectations again, that

$$e(t+1) = \left(1 + 2\,\frac{p}{t}\right) e(t)$$

which has a solution $e(t) \sim e(0)t^{2p}$ on iterating the recurrence. □

Lemma 2. *In the pure duplication model, the expected proportion of singletons, $f_0(t)$, is a non-decreasing function of t and tends to a limit $f_0 \leqslant 1$. For all $k \geqslant 1$, constant, if $f_k(t)$ tends to a limit f_k, then $f_k = 0$.*

Proof. We have the following recurrence for singletons in the pure duplication model:

$$F_0(t+1) = F_0(t) + \sum_{k \geqslant 0} \frac{F_k(t) q^k}{t}.$$

Thus, writing $F_k(t) = t f_k(t)$ we have

$$(t+1)(f_0(t+1) - f_0(t)) = \sum_{k \geqslant 1} f_k(t) q^k \geqslant 0, \tag{2}$$

and we see that $f_0(t+1) \geqslant f_0(t)$. As $f_0(t) \leqslant 1$ it follows that $f_0(t) \to f_0 \leqslant 1$ from below as $t \to \infty$.

Suppose next that for some $k \geqslant 1$, constant, $f_k(t) \to f_k > 0$. Then $\sum_{j \geqslant 1} f_j q^j \geqslant f_k q^k = c > 0$. Thus there exists T such that for $t \geqslant T$, $\sum_{k \geqslant 1} f_k(t) q^k \geqslant c/2 > 0$ and using (2)

$$f_0(t+1) \geqslant f_0(t) + \frac{c}{2(t+1)}.$$

Iterating this we get

$$f_0(t) \geqslant \frac{c}{2} \log t/T + \mathrm{O}(1/T) + f_0(T),$$

i.e. $f_0(t) > 1$ for t large enough, which is impossible. □

For $k \geqslant 1$ constant, or tending to infinity more slowly than $(\log\log t/\log(1/q))$, this lemma preludes the existence of power-law solutions $f_k \sim ck^{-b}$, as suggested in [12]; or indeed any limiting solution other than $f_k = 0$. We cannot exclude non-limiting degree distributions by this argument. For example, it is possible that for some values of $p > \frac{1}{2}$ the connected component $C(t)$ is a giant component of order t whose minimum degree $k_{\min} \to \infty$, or that the proportion of vertices of any fixed degree $1 \leqslant k \leqslant t$ tends to 0. From Lemma 1, the expected number of edges in $G(t)$ is of order t^{2p}. Suppose that for some $p > \frac{1}{2}$, $|C(t)|$ is of order t. Then the average degree of the connected component $C(t)$ is $t^{2p-1} \to \infty$.

For $p < \frac{1}{2}$ on the other hand the average degree of the connected component $C(t)$ is $t^{2p-1} \to 0$. Combining this with the previous lemma, it is clear that $f_0(t) \to 1$.

For the interesting case that $p = \frac{1}{2}$, it is possible to obtain a tighter estimate on $|C(t)|$, and thus the proportion of singletons in $G(t)$. Let $\boldsymbol{F}^+(t) = |C(t)|$ be the number of non-singleton nodes at time t and let $F^+ = \mathbf{E}\boldsymbol{F}^+$.

Lemma 3. *For $p = \frac{1}{2}$, there are constants $c_1, c_2 > 0$ such that $c_1\sqrt{t} \leqslant F^+(t) \leqslant c_2 t/\log\log t$.*

Proof. We have the following recurrence:

$$F^+(t+1) = F^+(t) + \frac{1}{t}\sum_{k\geqslant 0} F_k(t)(1-(1/2)^k). \tag{3}$$

Thus

$$F^+(t+1) = F^+(t) + \frac{F^+(t)}{t} - \frac{F^+(t)}{t}\sum_{k\geqslant 1}\frac{F_k(t)}{F^+(t)}\frac{1}{2^k}. \tag{4}$$

As $F_1(t)\leqslant F^+(t)$, one can easily check $F^+(t)\geqslant F^+(0)\sqrt{t}$ giving the lower bound.

Now let $g(k) = 1/2^k$, which is convex and thus for any set of λ_k for which $\sum\lambda_k = 1$, we must have $\sum\lambda_k g(k)\geqslant g(\sum k\lambda_k)$. Now pick $\lambda_k = F_k(t)/F^+(t)$. We have $\sum kF_k(t) = 2e(t) = 2e(0)t$. Thus

$$\sum_{k\geqslant 1}\frac{F_k(t)}{F^+(t)}\left(\frac{1}{2}\right)^k \geqslant \left(\frac{1}{2}\right)^{2e(t)/F^+(t)}. \tag{5}$$

By substituting (3) into (2) and using $e(t) = e(0)t$ we get

$$F^+(t+1)\leqslant F^+(t) + \frac{F^+(t)}{t}\left(1-\left(\frac{1}{2}\right)^{2e(0)t/F^+(t)}\right).$$

This is only satisfied if $F^+(t)\leqslant c_2 t/\log\log t$. This can be verified as follows. Let $c_2 = 4e(0)\log 2$. Either $F^+(t)\leqslant c_2 t/\log\log t$ or if not we can substitute this lower bound into the exponent on the right-hand side and iterate the recurrence on t to obtain a contradiction. □

As previously mentioned, Lemma 1 proves that the expected number of edges $e(t)$ is $e(t) = ct^{2p}$ and consequently the expected average degree $e(t)/t = 2ct^{2p-1}$. Thus, for $p < 0.5$ the average degree decreases over time and for $p > 0.5$ it increases. Only for $p = 0.5$ the average degree remains constant; however, the proportion of singletons is $\geqslant 1 - \mathrm{O}(1/\log\log t)$ by Lemma 3.

Observation 1. *The power-law exponent b in* (1) *is given by the solution of* $1 = bp - p + p^{b-1}$ *and has the value* 2 *when* $p = \frac{1}{2}$. *A power-law degree sequence with parameter* $b = 2$ *is incompatible with* $e(t) = e(0)t$.

Under the assumption of a power-law degree distribution at $p = \frac{1}{2}$, we have $F_k(t)\sim ck^{-2}t$ and

$$e(t) = \frac{ct}{2}\sum_{k\geqslant 1}\left(1+\mathrm{O}\left(\frac{1}{k}\right)\right)k^{-1},$$

which diverges unless the maximum degree k^* is a constant, contradicting the assumption of a power-law degree sequence.

It is however possible that a power law with exponent $b = 2$ holds for the connected component $C(t)$ when $p = \frac{1}{2}$. We see that $\sum_{k=1}^{t} k^{-1} = \mathrm{O}(\log t)$ which is compatible with $e(t) = e(0)t$ provided $|C(t)| = \mathrm{O}(t/\log t)$, which is in general accordance with the results of Lemma 3. It is plausible therefore that for $p > \frac{1}{2}$ the results of [12] hold for the component $C(t)$, although we cannot establish this at present.

3.2. Properties of the Pastor-Satorras et al. duplication model

We next consider the degree sequence of the Pastor-Satorras et al. duplication model [25]. A definition of this model is given in Section 2 (Steps 1,2). The next lemma shows that the degree sequence of this model cannot be a power law with exponential cut-off as was suggested in [25].

Lemma 4. *Let $a, b, c > 0$ be constants. The degree distribution of the Pastor-Satorras et al. duplication model cannot be in the form $F_k(t)\sim ctk^{-b}a^{-k}$ as stated in* [25].

Proof. Denote by $k_{\max}$, the expected maximum degree in $G(t)$. Assume an exponential cut-off i.e. $F_k(t) \sim tck^{-b}a^{-k}$. Then $\sum_{k \geqslant k_0} F_k(t) = o(1)$ for $k_0 > \log t / \log a$, and so $k_{\max} = O(\log t / \log a)$.

On the other hand, consider the expected degree of the node v_s at time $t+1$, which is a non-decreasing function of t. Even in the worst case situation ($r = 0$) we have

$$d_s(t+1) = d_s(t) + \frac{d_s(t)}{t}\, p \tag{6}$$

as the degree of v_s can only increase if one of its neighbors is picked at time t and the edge is retained. Thus

$$d_s(t+1) = d_s(t)\left(1 + \frac{p}{t}\right) = d_s(s)\left(1 + \frac{p}{s}\right) \cdot \left(1 + \frac{p}{s+1}\right) \ldots \left(1 + \frac{p}{t}\right).$$

Since $\log(1+x) = x - O(x^2)$ we have

$$\exp\left(\sum_{\tau=s}^{t} \log(1 + p/\tau)\right) \sim \exp\left(p \sum_{\tau=s}^{t} 1/\tau\right) = e^{p \log(t/s)}$$

which implies that $d_s(t+1) = \Omega(d_s(s)(t/s)^p)$ and that $k_{\max} = \Omega(t^p)$ contradicting the claim. □

The question of the correct power-law degree distribution for the Pastor-Satorras et al. model is resolved in Section 4 of this paper. Before considering this further, we need to prove that for $r > 0$ there are no degenerate limiting solutions of the form $f_0 = 1$, $f_k = 0, k \geqslant 1$ for the Pastor-Satorras et al. model.

Lemma 5. *Assuming* $\sum f_k = 1$, *for any* $r > 0$ *constant, the Pastor-Satorras et al. model does not have a degenerate limiting solution of the form* $f_0 = 1$, $f_k = 0, k \geqslant 1$.

Proof. We have the following recurrence for the expected number of singletons:

$$F_0(t+1) = F_0(t) + \sum_{k \geqslant 0} \frac{F_k(t)}{t} q^k \left(1 - \frac{r}{t}\right)^t - \frac{r}{t} F_0(t).$$

Assuming the existence of a limiting solution $F_k(t) = f_k t$ (after taking limits) and noting that $1 + r - e^{-r} > 0$ for $r > 0$, we have

$$f_0 = \frac{e^{-r}}{1 + r - e^{-r}} \sum_{k \geqslant 1} f_k q^k,$$

and thus $f_0 > 0$. If $f_0 = 1$ then $\sum_{k \geqslant 1} f_k q^k = 0$, giving a contradiction. □

4. The degree distribution of the generalized duplication model

In this section we show that the degree distribution of the generalized duplication model is a power law. We start with stating the expected maximum degree in the generalized duplication model.

Lemma 6. *The expected maximum degree of generalized duplication model at time t is* $\Omega(t^p)$.

It was proved in Lemma 4 that the expected maximum degree in the pure model is $\Omega(t^p)$. The maximum degree in the generalized model stochastically dominates the maximum degree in the pure duplication model. The formal coupling is to separate the edges $E_1(v)$, $E_2(v)$ at any node v into those derived entirely by duplication, and those arising generally in the graph by a u.a.r. (uniform at random) edge addition (possibly at some ancestor node). The expected maximum degree for $E_1(v)$ is that of the pure duplication model, which is $\Omega(t^p)$.

We start with the recurrence relation that governs the degree distribution in the pure duplication model.

$$F_k(t+1) = \left(F_k(t) - \frac{pkF_k(t)}{t}\right) + \frac{p(k-1)F_{k-1}(t)}{t} + \sum_{j \geqslant k} \frac{F_j(t)}{t} \binom{j}{k} p^k q^{j-k}. \tag{7}$$

The first term stands for the expected number of nodes with degree k at time t which still have degree k at time $t+1$. The second term stands for those nodes with degree $k-1$ at time t which will have degree k in time $t+1$ due to the duplication of one of the neighbors. The third term gives the probability that the degree of the duplicated node is k.

The standard analysis of these models in e.g. [12, 20] writes $F_k(t) = f_k(t)t$ and assumes that $f_k(t) \to f_k$, the limiting solution. This is problematic as we cannot offer a formal proof that such convergence occurs. However, on the assumption that the limit exists we provide an analysis.

The generalized duplication model fixes the problem in pure duplication model, that the proportion of singletons $f_0(t)$ can tend to 1. This is achieved by inserting a random edge to each new node which becomes a singleton after the deletion process. Indeed we now obtain the required power-law degree sequence. The Pastor-Satorras et al. model obtains similar results by choosing $r > 0$ to ensure that $f_0 < 1$ (see Lemma 5).

Theorem 1. *Provided $r + \delta_1 + \delta_2 > 0$, the generalized duplication model has a solution $f_k, k \geqslant 1$ of the form $f_k = (1 + \mathrm{O}(1/k))ck^{-b}$. The power-law parameter b is the largest solution of $1 = pb - p + p^{b-1}$. It is independent of the value of r, δ_1, δ_2 in the model.*

Proof. Let δ_i be the indicator for version $i = 1, 2$ (the enhanced versions of our model), so that, as $r + \delta_1 + \delta_2 > 0$ we have $\delta_1 + \delta_2 \leqslant 1$ for $r > 0$ and $\delta_1 + \delta_2 = 1$ for $r = 0$. Let $a_i \geqslant 1$ be the number of uar edges added in version $i = 1, 2$. Let $B(t, r/t; j) = \binom{t}{j}(r/t)^j(1 - r/t)^{t-j}$. The recurrence for $F_k(t)$ can be written as follows:

$$\begin{aligned} F_k(t+1) = F_k(t) &+ \frac{p(k-1)}{t}F_{k-1}(t) - \frac{pk}{t}F_k(t) + \left(\frac{r}{t} + \frac{a_2\delta_2}{t}\right)(F_{k-1}(t) - F_k(t)) \\ &+ \frac{a_1\delta_1}{t}(F_{k-1}(t) - F_k(t)) \sum_{j \geqslant 1} \frac{F_j(t)}{t} q^j + \sum_{L \geqslant k-j-a_2\delta_2} \sum_{j \geqslant 0} \frac{F_L(t)}{t} \binom{L}{k-j-a_2\delta_2} \\ &\times p^{k-j-a_2\delta_2} q^{L-(k-j-a_2\delta_2)} B(t, r/t; j). \end{aligned}$$

The first line of this recurrence equation is identical to the first few terms of the recurrence Eq. (7) for the pure duplication model. The second line gives the expected changes deriving from u.a.r. edge insertion. This occurs with probability r/t at each node in the generalized duplication model. Similarly, the expected number of edges at a node is $a_2\delta_2/t$ in Version 2. The third line is for Version 1, and the fourth line is the degree of the duplicated node. The number of u.a.r. edges at the new node arising from the r/t effect is $B(t, r/t; j)$.

Replacing $F_k(t)$ by $f_k t$, writing $\Psi = \sum_{j \geqslant 1} f_j q^j$ we find

$$\begin{aligned} 0 = f_k(-1 - kp - r - a_1\delta_1\Psi - a_2\delta_2) &+ f_{k-1}((k-1)p + r + a_1\delta_1\Psi + a_2\delta_2) \\ &+ \sum_{L \geqslant k-j-a_2\delta_2} \sum_{j \geqslant 0} f_L \binom{L}{k-j-a_2\delta_2} p^{k-j-a_2\delta_2} q^{L-(k-j-a_2\delta_2)} B(t, r/t; j). \end{aligned}$$

Substituting $f_j = (1 + \mathrm{O}(1/j))cj^{-b}$, multiplying through by k^b we obtain

$$0 = (-1 - kp - r - a_1\delta_1\Psi - a_2\delta_2) + \frac{k^b}{(k-1)^b}((k-1)p + r + a_1\delta_1\Psi + a_2\delta_2) + \mathrm{O}(1/k) \tag{8}$$

$$+ \sum_{L \geqslant k-j-a_2\delta_2} \sum_{j \geqslant 0} \frac{k^b}{L^b} \binom{L}{k-j-a_2\delta_2} p^{k-j-a_2\delta_2} q^{L-(k-j-a_2\delta_2)} B(t, r/t; j). \tag{9}$$

Note first that

$$\left(\frac{k}{k-1}\right)^b = 1 + \frac{b}{k} + \mathrm{O}\left(\frac{1}{k^2}\right),$$

so that the right-hand side of (8) evaluates to $-1 - p + bp + \mathrm{O}(1/k)$.

For any constant $b > 0$, and any J, K we have

$$\binom{J}{J-K}\left(\frac{K}{J}\right)^b = \left(1 + \mathrm{O}\left(\frac{1}{K+1}\right)\right)\binom{J-b}{J-K},$$

see e.g. [12] for details. Thus

$$\left(\frac{k}{L}\right)^b \binom{L}{k-j-a_2\delta_2} = \left(\frac{k}{k-j-a_2\delta_2}\right)^b \left(\frac{k-j-a_2\delta_2}{L}\right)^b \binom{L}{L-(k-j-\delta_2)}$$
$$= \left(1+\mathrm{O}\left(\frac{j}{k}\right)+\mathrm{O}\left(\frac{1}{k-j-a_2\delta_2+1}\right)\right)\binom{L-b}{L-(k-j-a_2\delta_2)}.$$

Fix $k-j-a_2\delta_2 \geqslant 0$, and let $l = L-(k-j-a_2\delta_2)$. Thus

$$\sum_{l\geqslant 0}\binom{l+k-j-a_2\delta_2-b}{l}q^l = \frac{1}{(1-q)^{k-j-a_2\delta_2-b+1}}.$$

Summing over $j \geqslant 0$ we have $\sum B(t, r/t; j) = 1$ so the term (9) is $(1+\mathrm{O}(1/k))p^{b-1}$ and we find that b is the solution of

$$bp - p + p^{b-1} = 1. \qquad \square$$

The theorem is true irrespective of the version selected (if either) and the value of r, δ_1, δ_2 provided that $r+\delta_1+\delta_2 > 0$. We remark that choosing $r > 0, \delta_1, \delta_2 = 0$ gives the degree distribution for the Pastor-Satorras et al. duplication model, and that this is independent of the value of r. This result was obtained in [20], as was the equation $bp - p + p^{b-1} = 1$ for b.

The next lemma gives the expected number of edges in the generalized duplication model. It is similar to Lemma 1 for the pure duplication model for $p > \frac{1}{2}$ but differs for $p \leqslant \frac{1}{2}$ as the uar step of the process ensures that the expected number of vertices with positive degree is linear. As usual, the results for the Pastor-Satorras et al. duplication model are obtained as a special case (set $\delta_1, \delta_2 = 0$).

Lemma 7. *Let $e(t)$ be the expected number of edges at step t. Let $\lambda = r + a_1\delta_1\Psi + a_2\delta_2$, then provided $\lambda > 0$*

$$e(t) \sim \begin{cases} \dfrac{\lambda}{1-2p}\,t, & p < 1/2, \\ \lambda\, t\log t, & p = 1/2, \\ \left(e(0)+\dfrac{\lambda}{2p-1}\right)t^{2p}, & p > 1/2. \end{cases}$$

Proof. We have

$$e(t+1) = e(t) + r + a_1\delta_1\Psi + a_2\delta_2 + \sum_k \frac{pkF_k(t)}{t},$$

where $\sum kF_k(t) = 2e(t)$. The simplest approach is to approximate the recurrence by the differential equation $e'(t) = 2pe(t)/t + \lambda$, obtain the solution, and then check the validity by direct substitution. $\square$

References

[1] W. Aiello, F. Chung, L. Lu, A random graph model for power law graphs, Proc. ACM STOC, 2000, pp. 171–180.
[2] W. Aiello, F. Chung, L. Lu, Random evolution in massive graphs, Proc. FOCS, 2001, pp. 510–519.
[3] R.A. Albert, A.-L. Barabási, Topology of evolving networks: local events and universality, Phys. Rev. Lett. 85 (2000) 5234.
[4] R.A. Albert, A.-L. Barabási, Statistical mechanics of complex networks, Rev. Modern Phys. 74 (2002) 47–94.
[5] A.-L. Barabási, R.A. Albert, Emergence of scaling in random networks, Science 286 (1999) 509–512.
[6] A. Bhan, D.J. Galas, T.G. Dewey, A duplication growth model of gene expression networks, Bioinformatics 18 (2002) 1486–1493.
[7] B. Bollobás, Modern Graph Theory, Springer, New York, 1998.
[8] B. Bollobás, C. Borgs, J. Chayes, O. Riordan, Directed scale-free graphs, Proc. ACM-SIAM SODA, 2003, pp. 132–139.
[9] B. Bollobás, O. Riordan, Handbook of Graphs and Networks, Wiley-VCH, Berlin, 2002.
[10] B. Bollobás, O. Riordan, The diameter of a scale-free random graph, Combinatorica 24 (2004) 5–34.
[11] B. Bollobás, O. Riordan, J. Spencer, G. Tusanády, The degree sequence of a scale-free random graph process, Random Structures Algorithms 18 (2001) 279–290.
[12] F. Chung, L. Lu, T.G. Dewey, D.J. Galas, Duplication models for biological networks, J. Comput. Biol. 10 (2003) 677–687.

[13] C. Cooper, A. Frieze, A general model of webgraphs, Random Structures Algorithms 22 (3) (2003) 311–335.
[14] S.N. Dorogovstev, J.F.F. Mendes, Evolution of networks with aging of sites, Phys. Rev. E 62 (2000) 1842.
[15] S.N. Dorogovstev, J.F.F. Mendes, A.N. Samukhin, Structure of growing networks with preferential linking, Phys. Rev. Lett. 85 (2000) 4633.
[16] P. Erdös, A. Rényi, On random graphs I, Publ. Math. Debrecen 6 (1959) 290–297.
[17] M. Faloutsos, P. Faloutsos, C. Faloutsos, On power-law relationships of the internet topology, SIGCOMM, 1999.
[18] B. Hayes, Graph theory in practice: Part II, Amer. Sci. 88 (2000) 104–109.
[19] H. Jeong, S. Mason, A.-L. Barabasi, Z.N. Oltvai, Lethality and centrality in protein networks, Nature 411 (2001) 41.
[20] J. Kim, P.L. Krapivsky, B. Kahng, S. Redner, Infinite-order percolation and giant fluctuations in a protein interaction network, Phys. Rev. E 66 (2002) 055101(R).
[21] J. Kleinberg, R. Kumar, P.P. Raphavan, S. Rajagopalan, A. Tomkins, The Web as a graph: measurements, models and methods, Proc. COCOON, Tokyo, Japan, 1999, pp. 1–17.
[22] R. Kumar, P. Raghavan, S. Rajagopalan, D. Sivakumar, A. Tomkins, E. Upfal, Stochastic models for the web graph, FOCS, 2000, pp. 57–65.
[23] M. Mitzenmacher, A brief history of generative models for power law and lognormal distributions, in: Proc. 39th Annu. Allerton Conf. on Communication, Control, and Computing, 2001, pp. 182–191.
[24] S. Ohno, Evolution by Gene Duplication, Springer, Berlin, 1970.
[25] R. Pastor-Satorras, E. Smith, R.V. Sole, Evolving protein interaction networks through gene duplication, J. Theor. Biol. 222 (2003) 199–210.
[26] N. Przulj, D.G. Corneil, I. Jurisica, Modeling interactome: scale-free or geometric? Bioinformatics 20 (18) (2004) 3508–3515.
[27] H.A. Simon, On a class of skew distribution functions, Biometrika 42 (1955) 425–440.
[28] P.L. Uetz, et al., A comprehensive analysis of protein–protein interactions in *S. Cerevisiae*, Nature 403 (2000) 623–627.
[29] A. Vázquez, A. Flammini, A. Maritan, A. Vespignani, Modelling of protein interaction networks, Complexus 1 (2003) 38–44.
[30] A. Wagner, The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, Mol. Biol. Evol. 18 (2001) 1283–1292.
[31] D.J. Watts, Small Worlds: The Dynamics of Networks between Order and Randomness, Princeton University Press, Princeton, 1999.
[32] D.J. Watts, S.H. Strogatz, Collective dynamics of small-world networks, Nature 393 (1998) 440–442.
[33] I. Xenarios, et al., DIP, the database of interacting proteins: a research tool for studying cellular networks of protein interactions, Nucleic Acids Res. 30 (2002) 303–305.
[34] G. Yule, A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, Philos. Trans. Roy. Soc. London (Ser. B) 213 (1925) 21–87.