The degree distribution of the generalized duplication model

G.Bebek[1], P.Berenbrink[2], C.Cooper[3], T.Friedetzky[4], J.H.Nadeau[5], and S.C.Sahinalp[6]

[1] Department of EECS, CWRU, Cleveland, OH, USA; gurkan@case.edu
[2] School of Computing Science, SFU, Vancouver, Canada; petra@cs.sfu.ca
[3] Department of Computer Science, King's College, London, UK; ccooper@dcs.kcl.ac.uk
[4] Department of Computer Science, University of Durham, UK; tom@friedetzky.org
[5] Department of Genetics, CWRU, Cleveland, OH, USA; jhn4@case.edu
[6] School of Computing Science, SFU, Vancouver, Canada; cenk@cs.sfu.ca

Abstract. We study the generalized duplication model of Pastor-Satorras *et al.* which generates a graph by iteratively "duplicating" a randomly chosen node. The duplicate node v is connected to each neighbor of the original node with probability p; additionally v is connected to every other node with probability r/t where t is the iteration number. Unlike other copy based models, the degree of the duplicate node in this model is not fixed in advance; rather it strongly depends on the degree of the original node and thus the degree distribution of the generated network.

Our main contributions are as follows. We show that (1) the generalized duplication model does not generate a truncated power law degree distribution as stated in [29] and (2) the special case where $r = 0$ does not generate a power law degree distribution as stated in [13]. In fact we show exactly the opposite: (3) apart from the special case considered in [13] the degree distribution of the generalized duplication model is a power law for $p \leq 0.58$. (4) We also modify the model so that it achieves a power law degree distribution even for $r = 0$.

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 *S.Cerevisiae* [35] (the only completely known proteome network) suggests that the degree distribution is in the form of a *power-law* [23, 34]. Power-law degree distributions have previously been observed in a number of naturally occurring graphs such as communication networks [18], web graphs [3, 5, 12, 14, 24, 25, 17], research citation networks [30], human language graphs [19] and neural nets [36].

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 Bollobás and Riordan [10], Hayes [22], Watts [37], or Aiello, Chung and Lu [3, 4]. Among these models, some of the most interesting ones (see [3, 4, 6, 9, 12, 14, 24, 25] and the survey by Mitzenmacher [26]) are based on an iterative random graph generation process which adds one new node to the graph in each iteration. The new node is then 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 [38] and Simon [31]. 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 [11] who showed that with high probability the diameter of a graph constructed in this way was $\sim \frac{\log t}{\log \log t}$ - here t stands for the time step and thus (is approximately) the number of nodes. Subsequently, Bollobás, Riordan, Spencer and Tusnády [12] proved that the degree sequence of such graphs does follow a power law distribution. More recently [14] introduced a very general analysis of random graphs revealing that many graphs generated through preferential attachment exhibit power law degree distributions. This was the first result to obtain graphs with a power law parameter smaller than three by using a graph generation model that allows edge insertion between existing nodes. More recently, in [20, 15], the authors consider iterative graph generation models where edges can be deleted subsequently. Further interesting results on preferential attachment models include [1], which studies the spread of viruses in the internet and [2], which studies the influence of search engines on preferential attachment models.

Motivated by the above examples of success, a number of random graph models were recently developed for emulating the growth of proteome networks; among them the one suggested independently in [29, 33, 7] received particular attention. A random graph model that aims to emulate proteome network growth must capture the essence of genome evolution process. The two underlying mechanisms for genome evolution is gene duplication and point mutations [28]. The model of [29, 33, 7] emulates these mechanisms through an iterative process. In each iteration t, one existing node (representing a gene or an associated protein) is chosen uniformly at random and is "duplicated" with all its edges; this emulates the gene duplication process. Then, (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 emulates the mutation process. We call models such as the one described here, where in addition to the duplication move, edges are added uniformly at random (uar) as a *generalized duplication models*.

Previous work on random graphs that grow by copying nodes have mostly focused on models where the duplicate node retains only a fixed number of edges of the original node. Such copying based models were analyzed in [25] and later in [14]. The result in [14] shows that the degree distribution of these models are similar to that of the preferential attachment models. The generalized duplication model is quite different from these copying based models, as in the former, the duplicate node's degree depends on that of the original node. As a consequence the generalized duplication model is harder to analyze.

Previous work on the generalized duplication model. It was observed in [29] that the generalized duplication model, after appropriate selection of the parameters q and r, leads to a random graph which has a degree distribution similar to that of the yeast proteome network. The first analytical work on the degree distribution of the generalized duplication model was [29] which suggests that the degree distribution of both the yeast proteome network and the generalized duplication model 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, this paper makes a number of simplifying assumptions in the analysis 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. The paper further approximates this probability with a function linear in k.

A more recent analysis of the degree distribution of the generalized duplication model, for the special case that $r = 0$, is given by Chung et al. [13]. Following [13], we will refer to this special case as the *pure duplication model*. This model creates *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 [29], Chung et al. suggest that the fraction of nodes with degree k is independent of time and is a power law distribution of the form $f_k = ck^{-b}$; here b is a function of q; values of $b \leq 2$ are possible for some q.

Summary of our contributions. (1) We show that the degree distribution of the generalized duplication model can not be a power law with exponential cut-off as stated in [29]; rather, it is a (regular) power law, provided $r > 0$ and $1 - q \leq 0.58$. (2) We show that, for the pure duplication model ($r = 0$) the fraction of nodes with degree k can not be independent of time and can not be a power law distribution of the form $f_k = ck^{-b}$ as stated in [13]. 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 [13]. 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.

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 3). 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 $b \leq 2$, 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 [29]. 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 study modification of the pure duplication model (Section 2), in which each iteration has an additional edge generation step. For the generalized duplication model with $r > 0$ and the modified pure duplication model, we show in Section 4 that our model guarantees the following: (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 parameter b is given by the equation in [13] irrespective of the model variant or the value of r, namely

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

2 Definition of the Generalized Duplication Model

The focus of this paper is the generalized duplication model [29, 13], which grows iteratively in discrete time steps. We start an arbitrary, but constant size, connected network. Let $G(t-1)$ be the network at the end of time step $t-1$. In time step t exactly one new node is generated and will be denoted as v_t . For any node v_s, we will denote its degree (or expected degree if the context is clear) at time step $t \geq s$ by $d_s(t)$.

The following steps are performed at each time step t.
First, a node w is picked uniformly at random and then it is "duplicated" to create the new node v_t which is initially connected to all the neighbors of w.
The edges initially incident to v_t are then updated through the following way:

Duplication: Each edge e is considered independently and is deleted with probability q and retained with probability $p = 1 - q$.
Uar edge addition: Each node u which is not connected to w is considered independently and an edge between u and v_t is created with probability r/t.
There are two possible enhanced versions of this uar edge addition step:
Version 1: If v_t has become a singleton at the end of the duplication move, it is connected to $a_1 \geq 1$ uniformly chosen random nodes. As a result the minimum degree will be a_1.
Version 2: v_t is always connected to $a_2 \geq 1$ additional nodes chosen uniformly at random, even if it did not become a singleton at the end of the duplication move.

We remark that these additional edge insertions are made after duplication, and without regard to the number of edges inserted by uar edge addition. This allows us to choose the parameter $r = 0$ if we so wish, and yet maintain connectivity of the graph $G(t)$. 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 the *generalized duplication model.*

We now give a number of definitions that we will use in our analysis. Let $\boldsymbol{F}_k(t)$ denote the number of nodes of degree k at the end of step t and let $\boldsymbol{F}(t) = (\boldsymbol{F}_0(t), \boldsymbol{F}_1(t), \cdots)$ 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. Finally let $\boldsymbol{e}(t)$ be the number of edges in $G(t)$ and $e(t) = \mathbf{E}\boldsymbol{e}(t)$; similarly let $\boldsymbol{h}(t)$ be the average degree of a node (averaged over all nodes) in $G(t)$, and $h(t) = \mathbf{E}\boldsymbol{h}(t)$. 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 $f_k = (1 + O(1/k))ck^{-b}$.

3 A Discussion on the Properties of the generalized Duplication Model

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 $F_0(t) \to t$ is the only consistent limiting solution. This implies that, unless $f_k = 0$ for $k \geq 1$, $F_k(t) \neq tf_k$; here f_k is a time independent solution for the limiting proportion of nodes of degree k. In fact, for the particularly interesting case that $p = q = 1/2$, we show that the expected number of non singletons at time step t is between $O(\sqrt{t})$ and $O(t/\log\log t)$. This contradicts the assumption in Eqn(6) of [13]. Thus, without some modification, the pure duplication model of [13] 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 in [29] which states that the generalized duplication model has a degree distribution of the form '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 cutoff is $O(\log t)$, whereas the generalized duplication model has an expected maximum degree of $\Omega(t^p)$.

In what follows we assume that $f_k(t) \to f_k$ as $t \to \infty$, ie. 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 ($f_\infty > 0$). This occurs for example when $p = 1$ where the minimum vertex degree grows linearly with t. In this paper we consider the existence of solutions to the limiting degree sequence in the case where $\sum f_k = 1$. It is easily shown (see Lemma 3) that the expected average degree in (eg.) the pure duplication model is $2e(0)t^{2p-1}$, so it is certainly the case that the solution is not defective for $p \leq 1/2$. It is unknown at what value of $p \leq 1$ the limiting distribution becomes defective.

3.1 Properties of the pure duplication model

Lemma 1. *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 $f_0 \leq 1$. If also we have that $\sum f_k = 1$, then $f_0 = 1$ and $f_k = 0$ for $k \geq 1$.*

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

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

Thus writing $F_k(t) = tf_k(t)$ we have

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

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

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

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

Iterating this we get

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

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

□

This lemma excludes the existence of power law solutions $f_k \sim ck^{-b}$ for finite $k \geq 1$ (which are suggested in [13]), but we cannot exclude non-limiting degree distributions by this argument.

It is possible to obtain a tighter estimate on the proportion of singletons in the network for the particularly interesting case that $p = q = 1/2$. We will see in Lemma 3 that this case preserves the (expected) average degree of the nodes throughout the generation of $G(t)$. Thus, $e(t) = e(0) \cdot t$ (where $e(0)$ is the number of edges of $G(0)$).

Lemma 2. *Consider the case $q = 1/2$. Let $\boldsymbol{F}^+(t) = t - \boldsymbol{F}_0(t)$ be the number of non-singleton nodes at time t and $F^+ = \mathbf{E}\boldsymbol{F}^+$. Then there are constants $c_1, c_2 > 0$ such that $c_1\sqrt{t} \leq F^+(t) \leq c_2 t/\log\log t$.*

Proof. We have the following recurrence:

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

Thus:

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

As $F_1(t) \leq F^+(t)$, one can easily check $F^+(t) \geq F^+(0)\sqrt{t}$ 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 $\sum \lambda_k g(k) \geq g(\sum k\lambda_k)$. Now pick $\lambda_k = \frac{F_k(t)}{F^+(t)}$. We have $\sum kF_k(t) = 2e(t) = 2e(0)t$. Thus:

$$\sum_{k\geq 1} \frac{F_k(t)}{F^+(t)} \left(\frac{1}{2}\right)^k \geq \left(\frac{1}{2}\right)^{2e(t)/F^+(t)} \tag{3}$$

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

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

This is only satisfied if $F^+(t) \leq c_2 t/\log\log t$. This can be verified as follows. Let $c_2 = 4e(0)\log 2$. Either $F^+(t) \leq c_2 t/\log\log t$, 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. □

Lemma 3 (below) states that the expected number of edges is $e(t) = ct^{2p}$ and consequently the expected average degree is $h(t) = 2ct^{2p-1}$. 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 as the proportion of singletons is $\geq 1 - O\left(\frac{1}{\log\log t}\right)$ due to Lemma 2, the average degree of non-singletons (which all form a single connected component) is $\geq c\log\log t$.

Observation 1. *The power law exponent* b *in [13] is given by the solution of* $1 = bp - p + p^{b-1}$ *and has the value 2 when* $p = 1/2$*. This is incompatible with* $e(t) = e(0)t$ *unless the connected component is of size* $o(t)$*.*

To see this, recall that $\sum kF_k(t) = 2e(t)$. Under the assumption that we have a power law degree distribution at $p = 1/2$, we have $F_k(t) \sim ck^{-2}t$ and

$$e(t) = \frac{ct}{2}\sum_{k\geq 1}\left(1 + O\left(\tfrac{1}{k}\right)\right)k^{-1}.$$

However, $\sum_{k=1}^{k^*} k^{-1}$ diverges as $k^* \to \infty$, and we cannot have $e(t) = 2e(0)t$, unless we truncate k^* at a finite value. Lemma 4 (below) sets the expected maximum degree in the pure model at $\Omega(t^p)$, and the power law assumption itself is not compatible with k^* being finite.

It is however still possible that a power law with exponent $b = 2$ holds for the connected component C. Putting $k^* = O(t^{1/2})$ we see that $\sum k^{-1} = O(\log t)$ which gives $e(t) = e(0)t$ provided $|C| = O(t/\log t)$, in accordance with the results of Lemma 2.

Lemma 3. *The expected total number of edges and the expected average degree of nodes at step* t *satisfy*

$$e(t) \sim e(0)t^{2p} \ \textit{and} \ h(t) \sim h(0)t^{2p-1}$$

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

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

The first term is trivial; 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 $e(t) \sim e(0)t^{2p}$. □

Figure 1 (see the end of the paper) depicts the percentage of the singletons in the network over the time for different values of p. The model was run until 1000000 non-singleton nodes were created. The plot uses a linear scale on the y-axis (percentage of singletons) and a logarithmic scale on the x-axis (running time).

Figure 2 depicts the average degree over time for different values of p. Again, the model was run until 1000000 non-singleton nodes were created. The average degree of the network increases with time and the larger the value of p is, the larger is the increases of the average degree.

3.2 Properties of the generalized duplication model

The next lemma shows that the degree distribution of the generalized duplication model can not be a power law with exponential cutoff as suggested in [29].

Lemma 4. *Let $a, b, c > 0$ be constants. The degree distribution of the generalized duplication model cannot be in the form $F_k(t) \sim ctk^{-b}a^{-k}$ as stated in [29].*

Proof. Denote by k_{max}, the expected maximum degree in $G(t)$. Assume an exponential cutoff i. e. $F_k(t) \sim tck^{-b}a^{-k}$. Then $\sum_{k \geq k_0} F_k(t) = o(1)$ 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:

$$d_s(t+1) = d_s(t) + \frac{d_s(t)}{t} p \qquad (4)$$

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:

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

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 generalized model of [29] 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 $f_0 = 1, f_k = 0, k \geq 1$ for the generalized model of [29].

Lemma 5. *Assuming $\sum f_k = 1$, for any $r > 0$ constant, the generalized model does not have a degenerate limiting solution of the form $f_0 = 1, f_k = 0, k \geq 1$.*

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

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

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

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

and thus $f_0 > 0$. If $f_0 = 1$ then $\sum_{k \geq 1} f_k q^k = 0$, 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 (as well as the modified pure 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.

$$F_k(t+1) = \left(F_k(t) - \frac{pkF_k(t)}{t}\right) + \frac{p(k-1)F_{k-1}(t)}{t} + \sum_{j \geq k} \frac{F_j(t)}{t} \binom{j}{k} p^k q^{j-k}. \quad (5)$$

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 change to $F_k(t)$ arising from the degree of the duplicated node.

The analysis provided in [13] replaces $F_k(t)$ with the time independent solution $f_k t$ for all k and substitutes $f_k = ck^{-b}$. As we have shown, this is problematic due to the fact that $F_0(t)/t$ grows with t and thus the only time independent solution is $f_0 = 1, f_k = 0, k \geq 1$. The modified pure duplication model fixes this problem by inserting a random edge to each new node which becomes a singleton after the deletion process.

Theorem 1. *The generalized duplication model and the modified pure duplication model have a solution $f_k, k \geq 1$ of the form $f_k = (1+O(1/k))ck^{-b}$. The power law parameter b is the solution of $1 = pb - p + p^{b-1}$ and it is independent of the value of $r \geq 0$ for the generalized duplication model and the version of the modified pure duplication 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 $\delta_1 + \delta_2 \leq 1$ for $r > 0$ and $\delta_1 + \delta_2 = 1$ for $r = 0$. Let $a_i \geq 1$ 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:

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

The first line of this recurrence equation is identical to the first few terms of the recurrence equation 5 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 $\Psi = \sum_{j\geq 1} f_j q^j$ we find

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

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

$$0 = (-1 - kp - r - a_1\delta_1\Psi - a_2\delta_2) + \frac{k^b}{(k-1)^b}((k-1)p + r + a_1\delta_1\Psi + a_2\delta_2) + O(1/k) \tag{6}$$

$$+ \sum_{L \geq k-j-a_2\delta_2} \sum_{j \geq 0} \frac{k^b}{L^b} \binom{L}{k-j-a_2\delta_2} p^{k-j-a_2\delta_2} q^{L-(k-j-a_2\delta_2)} B(t, r/t; j) \tag{7}$$

Note first that

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

so that the right hand side of (6) evaluates to $-1 - p + bp + O(1/k)$.

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

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

see e.g. [13] for details. Thus

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

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

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

Summing over $j \geq 0$ we have $\sum B(t, r/t; j) = 1$ so the term (7) is $(1 + O(1/k))p^{b-1}$ and we find that b is the solution of

$$1 = bp - p + p^{b-1},$$

as given in [13]. This is irrespective of the version selected and the value of $r \geq 0$. It is equally valid for the generalized duplication model [29] (i. e. $\delta_1, \delta_2 = 0$) provided we choose $r > 0$ to ensure that $f_0 < 1$ (see Lemma 5). □

The next lemma gives the expected number of edges in the generalized duplication model. This is similar to the pure duplication model for $p > 1/2$ but differs for $p \leq 1/2$.

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 $\lambda > 0$*

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

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. □

The results for the generalized duplication model are obtained as a special case $(\delta_1, \delta_2 = 0)$.

References

1. Noam Berger, Christian Borgs, Jennifer T. Chayes, and Armin Saberi, On the Spread of viruses on the Internet, *Proc. SIAM SODA*, 2005.
2. Soumen Chakrabarti, Alan Frieze, and Juan Vera, The Influence of Search Engines on Preferential Attachment, *Proc. SIAM SODA*, 2005.
3. Aiello W., Chung F., Lu L., A random graph model for power law graphs, *Proc. ACM STOC*, pp 171-180, 2000.
4. Aiello W., Chung F., Lu L., Random evolution in massive graphs, *Proc. FOCS*, pp 510-519, 2001.
5. Barabási, A.-L., Albert, R. A., Emergence of scaling in random networks, *Science* **286**, pp 509-512, 1999.
6. Berger N., Bollobás, B., Borgs C., Chayes J., Riordan O., Degree distribution of the FKP network model, *Proc. ICALP*, LNCS 2719, pp 725-738, 2003.
7. Bhan A., Galas D. J., & Dewey T. G., A duplication growth model of gene expression networks, *Bioinformatics*, **18**, pp 1486-1493, 2002.
8. Bollobás, B., Modern Graph Theory, Springer-Verlag, New York, 1998.
9. Bollobás, B., Borgs C., Chayes J., Riordan O., Directed scale-free graphs, *Proc. ACM-SIAM SODA*, pp 132-139, 2003.
10. Bollobás, B., Riordan, O., Handbook of Graphs and Networks, Wiley-VCH, Berlin, 2002.
11. Bollobás, B., Riordan, O. The diameter of a scale-free random graph, *Combinatorica* **24**, pp 5-34, 2004.
12. Bollobás, B., Riordan, O., Spencer, J., and Tusanády, G., The degree sequence of a scale-free random graph process, *Random Structures and Algorithms*, **18**, pp 279-290, 2001.
13. Chung, F., Lu L., Dewey T.G., Galas D.J., Duplication models for biological networks, *Journal of Computational Biology*, **10**, pp 677-687, 2003.
14. Cooper C., Frieze A., A general model of webgraphs, *Random Structures and Algorithms*, **22(3):** pp 311-335, 2003.
15. Colin Cooper, Alan Frieze, and Juan Vera, Random Deletion in a Scale free Random Graph, Internet mathematics, to appear.
16. Erdös, P., Rényi, A., On random graphs I, *Publicationes Mathematicae Debrecen*, **6**, pp 290-297, 1959.
17. Fabrikant, A., Koutsoupias, E., Papadimitriou, C. B., Heuristically Optimized Trade-offs: A New Paradigm for Power Laws in the Internet, *In Proc. of 2002 ICALP*, pages 110-122.
18. Faloutsos M., Faloutsos P., Faloutsos C., On Power-Law Relationships of the Internet Topology, *SIGCOMM*, 1999.

19. Ferrer i Cancho, R., Janssen, C., The small world of human language, *Procs. Roy. Soc. London B*, **268**, pp 2261-2266, 2001.
20. Abraham D. Flaxmann, Alan M. Frieze, and Juan Vera, Adversarial Deletion in a Scale free Random Graph Process, *Proc. SIAM SODA*, 2005.
21. Force A., Lynch M., Pickett F.B., Amores A., Yan Y., Postlethwait J., Preservation of duplicate genes by complementary degenerative mutations. *Genetics*, **151**, pp 1531-1545, 1999.
22. Hayes, B., Graph theory in practice: Part II, *American Scientist*, **88**, pp 104-109, 2000.
23. Jeong, H., Mason, S., Barabasi, A.-L. & Oltvai, Z. N., Lethality and centrality in protein networks, *Nature*, **411**, pp 41, 2001.
24. Kleinberg, J., Kumar, R., Raphavan, PP, Rajagopalan, S. and Tomkins A., The Web as a graph: Measurements, models and methods, *Proc. COCOON*, Tokyo, Japan, pp 1-17, 1999.
25. Kumar R., Raghavan P., Rajagopalan S., Sivakumar D., Tomkins A., Upfal E., Stochastic models for the web graph, *FOCS* pp 57-65, 2000.
26. Mitzenmacher, M., A brief history of generative models for power law and lognormal distributions, *Proc. of the 39th Annual Allerton Conf. on Communication, Control, and Computing*, pp 182-191, 2001.
27. Nadeau, J.H., Sankoff D., Comparable Rates of Gene Loss and Functional Divergence After Genome Duplications Early in Vertebrate Evolution, *Genetics*, **147**, pp 1259, 1997.
28. Ohno, S., Evolution by gene duplication. Berlin: Springer, 1970.
29. Pastor-Satorras, R., Smith, E., and Sole, R.V., Evolving protein interaction networks through gene duplication, *J. Theor. Biol.* **222**, pp 199-210, 2003.
30. Redner, S., How Popular is Your Paper? An Empirical Study of the Citation Distribution, *Eur. Phys. Jour.* **B 4**, pp 131-134, 1998.
31. Simon, H. A., On a class of skew distribution functions, *Biometrika*, **42**, pp 425-440, 1955.
32. Uetz, P. L. *et. al.*, A comprehensive analysis of protein-protein interactions in S.Cerevisiae, *Nature*, **403**, pp 623-7, 2000.
33. Vázquez, A., Flammini, A., Maritan, A. & Vespignani, A., Modelling of protein interaction networks, *Complexus* **1**, 38-44, 2003.
34. Wagner, A., The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes, *Mol. Biol. Evol.* **18**, pp 1283-1292, 2001.
35. Xenarios, I. *et. al.*, DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions, *Nucleic Acids Res.* **30**, pp 303-305, 2002.
36. Watts, D. J. & Strogatz, S. H., Colective dynamics of small-world networks, *Nature*, **393**, pp 440-442, 1998.
37. Watts, D. J.. *Small Worlds: The Dynamics of Networks between Order and Randomness*, Princeton: Princeton University Press, 1999.
38. Yule, G., A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, *Philosophical Transactions of the Royal Society of London (Series B)*, **213**, pp 21-87, 1925.

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).