J. Appl. Prob. **52,** 375–390 (2015)
Printed in England
© *Applied Probability Trust* 2015

ASYMPTOTIC PROPERTIES OF A RANDOM GRAPH WITH DUPLICATIONS

ÁGNES BACKHAUSZ * AND
TAMÁS F. MÓRI,** *Eötvös Loránd University*

Abstract

We deal with a random graph model evolving in discrete time steps by duplicating and deleting the edges of randomly chosen vertices. We prove the existence of an almost surely asymptotic degree distribution, with stretched exponential decay; more precisely, the proportion of vertices of degree d tends to some positive number $c_d > 0$ almost surely as the number of steps goes to ∞, and $c_d \sim (\mathrm{e}\pi)^{1/2} d^{1/4} \mathrm{e}^{-2\sqrt{d}}$ holds as $d \to \infty$.

Keywords: Scale-free; duplication; deletion; random graph; martingale

2010 Mathematics Subject Classification: Primary 60G42
Secondary 05C80

1. Introduction

In the last decades, inspired by the examination of large real networks, various types of random graph models with preferential attachment dynamics (meaning that vertices with larger degree have grreater chance to add new edges as the graph evolves randomly) were introduced and analysed. After some early work (see [23], [17], [8], and [19]) this area of research started with the seminal papers of Barabási and Albert [2] and Bollobás *et al.* [5]. Among many others, we mention the model of Cooper and Frieze for the Internet [3] or the model for social networks of Sridharan *et al.* [18].

An important feature of these graph sequences is the scale-free property: the proportion of vertices of degree d tends to some positive number c_d almost surely (a.s.) as the number of steps goes to ∞, and $c_d \sim K d^{-\gamma}$ holds as $d \to \infty$, where K and γ are some positive constants (throughout this paper, $a_d \sim b_d$ means that $a_d / b_d \to 1$ as $d \to \infty$). To put it in another way, the asymptotic degree distribution (c_d) is polynomially decaying. See [2], [10], and [22] and the references therein regarding the scale-free property of the internet.

However, the scale-free property captures only the behaviour of the degrees of the vertices, and does not examine other kinds of structures. In biological networks, for example, proteomes (as protein-protein interaction networks, i.e. the nodes are proteins, and two of them are connected if they interact in a natural biological processes), we can find groups of vertices having a similar neighbourhood, that is, most of their neighbours are the same. One can say that these networks are highly clustered; loosely speaking, there are large cliques, in which almost every vertex is connected to almost every other one, and there are only a few edges available between cliques.

Received 11 September 2013; revision received 20 June 2014.

* Postal address: MTA Alfréd Rényi Institute of Mathematics, Pázmány P. s. 1/C, H-1117, Budapest, Hungary.
Email address: agnes@math.elte.hu
** Postal address: Department of Probability Theory and Statistics, Pázmány P. s. 1/C, H-1117, Budapest, Hungary.
Email address: mori@math.elte.hu

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

A simple process to generate cliques is duplication: when a new vertex is added, we choose an old vertex randomly, and connect the new vertex to the neighbours of the old vertex. In other words, the new vertex becomes a copy of the old vertex. Note that if the old vertex is chosen uniformly at random, then the probability that a vertex of degree d gets a new edge is just the probability that one of its neighbours is chosen, which is proportional to its actual degree. Hence, this model is also driven by a kind of preferential attachment dynamics.

After the duplication, we can add some extra edges randomly, or we can delete some of them to guarantee that the network remains sparse. The graph may still have some large cliques due to the duplication.

Duplication is not only a technical step that proved to be useful: it is inherent. To quote Chung [6] 'This may be because duplication of the information in the genome is a dominant evolutionary force in shaping biological networks (like gene regulatory networks and protein–protein interaction networks)' .

These types of models – where the duplicated vertex is chosen uniformly at random – were examined, for example, by Kim *et al.* [13]. In their model the new vertex is connected to each neighbour of the chosen vertex with probability $1 - \delta$, independently. In addition, the new vertex is connected to each old vertex independently with probability β/n at the nth step (δ, β are the parameters of the model). The Scale-free property is claimed for this model. However, Pastor-Satorras *et al.* [14] stated that, instead of polynomial decay, for the limit c_d of the expected value of the proportion of vertices of degree d, the degree distribution has a polynomial decay with exponential cut-off $c_d \sim K d^{-\gamma} \mathrm{e}^{-\lambda d}$ with some positive constant λ. However, Chung *et al.* [6] claimed that for $\beta = 0$, when we do not have any extra edges, the asymptotic degree distribution exists, and (c_d) is decaying polynomially. None of these papers contained a mathematically rigorous proof.

However, Bebek *et al.* [4] disclaimed the results of [14] and [6]. In the latter case, it was shown that the fraction of isolated vertices (that have no edges) increased with time in the pure duplication model, where $\beta = 0$. The model was modified to avoid singletons by adding a fixed number of edges to the new vertex, chosen uniformly at random. They assumed without any proof that the asymptotic degree distribution exists, and they claimed that it is decaying polynomially.

Hamdi *et al.* [11] presented a model where the probabilities of adding a duplicated edge depends on the state of a hidden Markov chain. Polynomial decay is stated for the limit of the mean of degree distribution. We also mention the somewhat different model of Jordan [12], and the duplication model of Cohen *et al.* [7], where the duplicated vertex is chosen not uniformly, but with probabilities proportional to the actual degrees.

In our paper we present a simple random graph model based on the duplication of a vertex chosen uniformly at random, and the erasure of the edges of another vertex also chosen uniformly at random. We prove that for all d, the proportion of vertices of degree d tends to some c_d with probability 1 as the number of steps goes to ∞. Here c_d is a positive number; we will determine the asymptotics of the sequence (c_d) as $d \to \infty$, showing that it has a stretched exponential decay. Hence, this model does not have the scale-free property. We use the methods of martingale theory for proving a.s. convergence, and generating function and Taylor series techniques for deriving the integral representation and the asymptotics of the sequence (c_d).

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

2. Definition of the model and main results

Our model starts with a single vertex. The graph evolves in discrete time steps; each step has a duplication and an erasure part. At each step a new vertex will be born; therefore the number of vertices after n steps is $n + 1$. The graph is always a simple graph; it has neither multiple edges nor loops. At each step we do the following.

Version 1. We choose two (not necessarily different) old vertices independently, uniformly at random. Then the new vertex is added to the graph; we connect it to the first vertex and to all its neighbours. After that we delete all edges emanating from the second old vertex we have selected, with the possible exception that edges of the new vertex cannot be deleted.

Our main results are about the model above. However, in the proofs, we will use a helpful simplification of this model which is defined as follows.

Version 2. We choose two (not necessarily different) old vertices independently, uniformly at random. The new vertex is connected to the first one and to its neighbours. Then, we delete all edges of the second vertex without any exceptions.

That is, the new edges are protected in the erasure part of the same step in version 1, but they might be deleted immediately in version 2. We will see that the version 2 graph has a simple structure that enables us to describe its asymptotical degree distribution. Then, using this and a coupling of the two models, we can prove similar results for version 1.

Let us remark that the presence of deletion makes the analysis more difficult than in the usual recursive graph models, since it causes intensive fluctuation in the behaviour of the model.

Our model is a type of coagulation–fragmentation model: the effect of duplication is coagulation, and deletion results in fragmentation. Coagulation–fragmentation models are frequently used in several areas; see, e.g. [9]. These models have been applied to random graph models [15], namely, for the Erdős–Rényi model, which is completely different from ours.

The basic property of version 2 is that the evolving graph always consists of separated complete graphs. That is, it is a disjoint union of cliques. Within a component, every pair of vertices is connected, and there are no edges between the components. Indeed, we start from a single vertex, which is a clique of size one, and both duplication and erasure make cliques from cliques. Moreover, it is easy to see that if we start the model with an arbitrary graph, all edges of the initial configuration are deleted after a while, and after that the graph will consist of separated cliques. So the initial configuration does not make any difference asymptotically.

We may formulate the second version as follows. At each step we choose two components independenty such that the probability that a given clique is chosen is proportional to its size. The new vertex is attached to the first clique, so its size is increased by 1; the size of the secondly chosen clique is decreased by 1, and an isolated vertex (the deleted one) comes into existence. Note that if we choose an isolated vertex to be deleted, then it remains isolated.

This structure of version 2 makes it easier to handle, as the number of d-cliques does not vary so vehemently as the number of degree d vertices; the fluctuation is bounded by 2. This will lead to the description of the asymptotic degree distribution of version 1 in an a.s. sense. Our main results are the following.

Theorem 1. *Denote by $X[n, d]$ the number of vertices of degree d after n steps in version 1. Then*

$$\frac{X[n, d]}{n+1} \to c_d$$

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

holds a.s. as $n \to \infty$, where (c_d) is a sequence of positive numbers satisfying

$$c_0 = \frac{1+c_1}{3}, \qquad c_d = \frac{d+1}{2d+3}(c_{d-1}+c_{d+1}) \quad \textit{for } d \geq 2. \tag{1}$$

For the asymptotic analysis, we first present an integral representation for the limiting sequence (c_d). As a corollary, it follows that the sum of this sequence is 1; it is really a probability distribution.

Theorem 2. *For the sequence (c_d) of Theorem 1, we have*

$$c_d = (d+1)\int_0^\infty \frac{y^d \mathrm{e}^{-y}}{(1+y)^{d+2}}\,\mathrm{d}y \quad \textit{for } d \geq 0$$

and $\sum_{d=0}^{\infty} c_d = 1$.

Using this equation we can derive the asymptotics of c_d.

Theorem 3. *For the sequence (c_d) of Theorem 1, we have*

$$c_d \sim (\mathrm{e}\pi)^{1/2} d^{1/4} \mathrm{e}^{-2\sqrt{d}} \quad \textit{as } d \to \infty.$$

Our model was devised to ensure high degree clustering. Finally, let us quantify this property.

The local clustering coefficient of a vertex of degree d is defined to be the fraction of connections that exist between the $\binom{d}{2}$ pairs of neighbours (which means 0 when $d < 2$). Watts and Strogatz [21] defined the clustering coefficient of the whole graph as the average of the local clustering coefficients of all the vertices. Let us call this quantity the average clustering coefficient. Another possibility for a such a measure is the ratio of three times the number of triangles divided by the number of connected triplets (paths of length 2); see [20]. This version is sometimes called transitivity; we will refer to it as the global clustering coefficient.

Since the graph in version 2 consists of disjoint cliques, its global clustering coefficient is obviously 1, while the average clustering coefficient is equal to the proportion of vertices with degree at least 2. By Theorem 1 it converges to $1 - c_0 - c_1 = 2 - 4c_0$ a.s. as $n \to \infty$. We note that the limit is equal to $0.385\,38\ldots$ by Theorem 2. These results can be transferred to version 1.

Theorem 4. *In version 1, the global clustering coefficient converges to 1, and the average clustering coefficient converges to $1 - c_0 - c_1$ a.s. as $n \to \infty$.*

The high clustering property of our model shows that it is a so-called small-world graph [21].

3. Proofs

3.1. Preliminaries

First, we formulate the lemma from martingale theory that we will use several times and whose proof can be found in [1].

Lemma 1. *Let $(\mathcal{F}_n)$ be a filtration, (ξ_n) a nonnegative adapted process. Suppose that the following holds with some $\delta > 0$,*

$$\mathbb{E}\{(\xi_n - \xi_{n-1})^2 \mid \mathcal{F}_{n-1}\} = O(n^{1-\delta}). \tag{2}$$

Let (u_n), (v_n) be nonnegative predictable processes such that $u_n < n$ for all $n \geq 1$. Finally, let (w_n) be a regularly varying sequence of positive numbers with exponent $\mu \geq -1$.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

(i) *Suppose that*

$$\mathbb{E}\{\xi_n \mid \mathcal{F}_{n-1}\} \le \Big(1 - \frac{u_n}{n}\Big)\xi_{n-1} + v_n,$$

and $\lim_{n\to\infty} u_n = u$, $\limsup_{n\to\infty} v_n/w_n \le v$ *with some random variables* $u > 0$, $v \ge 0$. *Then*

$$\limsup_{n\to\infty} \frac{\xi_n}{n w_n} \le \frac{v}{u+\mu+1} \quad \textit{a.s.}$$

(ii) *Suppose that*

$$\mathbb{E}\{\xi_n \mid \mathcal{F}_{n-1}\} \ge \Big(1 - \frac{u_n}{n}\Big)\xi_{n-1} + v_n$$

and $\lim_{n\to\infty} u_n = u$, $\liminf_{n\to\infty} v_n/w_n \ge v$ *with some random variables* $u > 0$, $v \ge 0$. *Then*

$$\liminf_{n\to\infty} \frac{\xi_n}{n w_n} \ge \frac{v}{u+\mu+1} \quad \textit{a.s.}$$

3.2. Asymptotic degree distribution in version 2

Recall that in this case the graph is always a disjoint union of complete graphs.

First, we prove the following analogue of Theorem 1.

Proposition 1. *Denote by* $Y[n,k]$ *the number of cliques of size* k *after* n *steps in version 2. Then for all positive integers* k, *we have*

$$\frac{Y[n,k]}{n} \to y_k \quad \textit{a.s. as } n \to \infty,$$

where (y_k) *is a sequence of positive numbers satisfying*

$$y_1 = \frac{1+2y_2}{3}, \qquad y_k = \frac{(k-1)y_{k-1} + (k+1)y_{k+1}}{2k+1} \quad \textit{for } k \ge 2. \tag{3}$$

Note that (3) (as well as (1)) is not a recursion. This prevents us proceeding simply in the usual, direct way, with induction over k.

Proof. For $n = 0$, we have $Y[0,1] = 1$ and all others equal to 0. The total number of vertices is n after $n-1$ steps. Let $\mathcal{F}_n$ denote the σ-field generated by the first n steps.

We enumerate the events that can happen to the cliques of different sizes during one step. At the nth step an isolated vertex may become

- a clique of size 2 (increased but not decreased) with probability $(1/n)(1-(1/n))$,
- an isolated vertex (any other cases).

A clique of size $k \ge 2$ may become a clique of size

- $k-1$ (not increased but decreased) with probability $(k/n)(1-(k/n))$,
- $k+1$ (increased but not decreased) with probability $(k/n)(1-(k/n))$,
- k (any other cases).

The deleted vertex will be a new isolated vertex unless one of them is chosen for erasure but not for duplication, which has probability $(1/n)(1-(1/n))$ for each vertex.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Putting this together with the fact that the random choices are independent and the probabilities are proportional to clique sizes, we can compute the conditional expectation of $Y[n,k]$ with respect to $\mathcal{F}_{n-1}$, which is the σ-field generated by the first $n-1$ steps. Thus,

$$\begin{aligned}\mathbb{E}\{Y[n,1]\mid\mathcal{F}_{n-1}\} &= Y[n-1,1]\biggl[1-\frac{1}{n}\biggl(1-\frac{1}{n}\biggr)-\frac{1}{n}\biggl(1-\frac{1}{n}\biggr)\biggr]\\ &\quad + 1 + Y[n-1,2]\frac{2}{n}\biggl(1-\frac{2}{n}\biggr),\\ \mathbb{E}\{Y[n,k]\mid\mathcal{F}_{n-1}\} &= Y[n-1,k]\biggl[1-2\frac{k}{n}\biggl(1-\frac{k}{n}\biggr)\biggr] + Y[n-1,k-1]\frac{k-1}{n}\biggl(1-\frac{k-1}{n}\biggr)\\ &\quad + Y[n-1,k+1]\frac{k+1}{n}\biggl(1-\frac{k+1}{n}\biggr) \qquad \text{for } k\ge 2.\end{aligned}$$

Let $A_k = \liminf_{n\to\infty}(Y[n,k]/n)$ and $B_k = \limsup_{n\to\infty}(Y[n,k]/n)$ for $k\ge 1$. It is clear that $0\le A_k\le B_k\le 1$ holds for these random variables.

We will derive a sequence of lower bounds for (A_k), and, similarly, a sequence of upper bounds for (B_k); then we will show that their limits are equal to each other. First, let $a_k^{(0)} = 0$ for $k\ge 1$. Having constructed the sequence $(a_k^{(j)})_{k\ge 1}$, we define

$$a_1^{(j+1)} = \frac{1+2a_2^{(j)}}{3}, \qquad a_k^{(j+1)} = \frac{(k-1)a_{k-1}^{(j)} + (k+1)a_{k+1}^{(j)}}{2k+1} \quad \text{for } k\ge 2. \tag{4}$$

We obtain $a_k^{(j)}$ recursively for every $k\ge 1$ and $j\ge 1$.

We prove by induction on j that $a_k^{(j)}\le A_k$ for $k\ge 1$. Since $Y[n,k]\ge 0$, this is clear for $j=0$. Suppose that this is satisfied for some j for every k. For $k=1$, we apply Lemma 1 with

$$\xi_n = Y[n,1], \qquad u_n = 2-\frac{2}{n}\longrightarrow 2, \qquad v_n = 1+Y[n-1,2]\frac{2}{n}\biggl(1-\frac{2}{n}\biggr).$$

Now, (ξ_n) is nonnegative adapted. We see that (u_n) and (v_n) are clearly nonnegative predictable sequences; we can choose $w_n=1$, $\mu=0$, $u=2>0$, and, finally, $v=1+2a_2^{(j)}\ge 0$ due to the induction hypothesis. Note that at each step at most one of the isolated points vanishes and at most two may appear. Thus, (2) is clearly satisfied. Lemma 1 implies that

$$A_1 = \liminf_{n\to\infty}\frac{Y[n,1]}{n} = \liminf_{n\to\infty}\frac{\xi_n}{n}\ge\frac{v}{u+1} = \frac{1+2a_2^{(j)}}{3} = a_1^{(j+1)} \quad \text{a.s.}$$

Similarly, for $k\ge 2$ if we have $A_k\ge a_k^{(j)}$ for some $j\ge 1$, we can choose

$$\xi_n = Y[n,k], \qquad u_n = 2k-\frac{2k^2}{n}\longrightarrow 2k, \qquad v = (k-1)a_{k-1}^{(j)} + (k+1)a_{k+1}^{(j)},$$
$$v_n = Y[n-1,k-1]\frac{k-1}{n}\biggl(1-\frac{k-1}{n}\biggr) + Y[n-1,k+1]\frac{k+1}{n}\biggl(1-\frac{k+1}{n}\biggr).$$

At each step at most three cliques are changed, which implies that (2) holds. Thus, in this case from Lemma 1, we obtain

$$A_k = \liminf_{n\to\infty}\frac{Y[n,k]}{n}\ge\frac{v}{u+1} = \frac{(k-1)a_{k-1}^{(j)} + (k+1)a_{k+1}^{(j)}}{2k+1} = a_k^{(j+1)} \quad \text{a.s.}$$

By induction on j it follows that $A_k\ge a_k^{(j)}$ holds a.s. for $k\ge 1$ and $j\ge 0$.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Now, we verify that for fixed k the sequence $(a_k^{(j)})$ is monotone increasing in j. Since $a_k^{(0)} = 0$ for every k, from (4) it is clear that $a_k^{(1)} \geq a_k^{(0)}$. Suppose that for some $j \geq 1$, we have $a_k^{(j)} \geq a_k^{(j-1)}$ for every k. From (4), it follows that

$$a_1^{(j+1)} = \frac{1 + 2a_2^{(j)}}{3} \geq \frac{1 + 2a_2^{(j-1)}}{3} = a_1^{(j)},$$

$$a_k^{(j+1)} = \frac{(k-1)a_{k-1}^{(j)} + (k+1)a_{k+1}^{(j)}}{2k+1} \geq \frac{(k-1)a_{k-1}^{(j-1)} + (k+1)a_{k+1}^{(j-1)}}{2k+1} = a_k^{(j)}.$$

Thus, by induction on j it follows that that $a_k^{(j)} \geq a_k^{(j-1)}$ for $k, j \geq 1$.

The sequence $(a_k^{(j)})_{j\geq 0}$ is uniformly bounded from above by 1, because A_k is at most 1 for all k (being the the limit inferior of a sequence of certain proportions), and we have proved that $a_k^{(j)} \leq A_k$ holds for all j. Using monotonicity, we define

$$a_k = \lim_{j\to\infty} a_k^{(j)} \quad \text{for } k \geq 1.$$

From (4), it follows that (a_k) satisfies (3), i.e.

$$a_1 = \frac{1 + 2a_2}{3}, \qquad a_k = \frac{(k-1)a_{k-1} + (k+1)a_{k+1}}{2k+1} \quad \text{for } k \geq 2.$$

On the other hand, since $A_k \geq a_k^{(j)}$ for $k \geq 1$ and $j \geq 0$, we have $A_k \geq a_k$ a.s.

Similarly, we define $b_k^{(0)} = 1$ for every k, and then

$$b_1^{(j+1)} = \frac{1 + 2b_2^{(j)}}{3}, \qquad b_k^{(j+1)} = \frac{(k-1)b_{k-1}^{(j)} + (k+1)b_{k+1}^{(j)}}{2k+1} \quad \text{for } k \geq 2.$$

Using part (a) of Lemma 1 it follows by induction on j that $B_k \leq b_k^{(j)}$ holds a.s.

In this case, for fixed k the sequence $b_k^{(j)}$ is decreasing, and for the limits $b_k = \lim_{j\to\infty} b_k^{(j)}$, we also have

$$b_1 = \frac{1 + 2b_2}{3}, \qquad b_k = \frac{(k-1)b_{k-1} + (k+1)b_{k+1}}{2k+1} \quad \text{for } k \geq 2.$$

In addition, $B_k \leq b_k$ a.s. Since $b_k^{(0)} = 1$, and the sequence $(b_k^{(j)})$ is decreasing for fixed k, it follows that $b_k \leq b_k^{(j)} \leq 1$ for every k.

By definition, $0 \leq A_k \leq B_k \leq 1$ and $0 \leq a_k \leq b_k \leq 1$ hold. Let $d_k = b_k - a_k \geq 0$ for all k. We have the same equations for (a_k) and (b_k). This yields

$$d_1 = \frac{2d_2}{3}, \qquad d_k = \frac{(k-1)d_{k-1} + (k+1)d_{k+1}}{2k+1} \quad \text{for } k \geq 2.$$

By rearranging, we obtain

$$d_2 = \frac{3}{2}d_1, \qquad d_{k+1} = \frac{(2k+1)d_k - (k-1)d_{k-1}}{k+1} \quad \text{for } k \geq 2. \tag{5}$$

Suppose that $d_k \geq ((k+1)/k)d_{k-1}$ holds for some $k \geq 2$. (For $k = 2$ this is true with equality.) Since d_{k-1} is nonnegative, $d_k \geq d_{k-1}$ follows also from this assumption. From (5), we obtain

$$d_{k+1} \geq \frac{(k+2)d_k}{k+1}.$$

Therefore, this inequality holds for every k.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

This implies that $d_k \geq (k+1)d_1$ for every k. Since $0 \leq d_k = b_k - a_k \leq 1$, it follows that $d_1 = 0$.

From (5), we obtain $d_k = 0$ for all k, which implies that $a_k = b_k$. Since these were the lower and upper bounds for the limit inferior and limit superior of $Y[n,k]/n$, it follws that the latter must converge a.s. as $n \to \infty$, and the limits satisfy (3).

Corollary 1. *In version 2, the proportion of vertices of degree d tends to c_d, satisfying (1) a.s. as $n \to \infty$.*

Proof. For a fixed d, we have $d+1$ vertices of degree d in each clique of size $k = d+1$. Therefore, the proportion of vertices of degree d tends to $(d+1)y_{d+1}$ by Proposition 1. From (3), we obtain

$$c_0 = y_1 = \frac{1+2y_2}{3} = \frac{1+c_1}{3} \qquad c_d = (d+1)y_{d+1} = \frac{d+1}{2d+3}\big(c_{d-1} + c_{d+1}\big) \quad \text{for } d \geq 2.$$

3.3. Asymptotic degree distribution in version 1

When proving the results for version 2, we essentially used the property that the graph consists of a disjoint union of cliques: at most three of the cliques may change at any one step, but the number of vertices whose degree is changed is not bounded uniformly. However, we can advance the results by a kind of coupling of versions 1 and 2.

Proof of Theorem 1. In versions 1 and 2, two old vertices were selected with replacement, independently and uniformly at random. Thus, we can couple the models such that the selected vertices are the same in all steps. The duplication part is the same for both versions. The difference is in the deletion: in version 1, the edges of the new vertex cannot be deleted. So in version 1, we do the following. In the deletion part, we colour an edge red if it is saved in version 2. That is, if it connects the new vertex with the old vertex to be deleted. In the duplication part, copies of red edges are also red: if there is a red edge between the duplicated vertex and one of its neighbours then the new edge connecting this neighbour to the new vertex is also red. All other new edges are originally black, but they may turn red in the deletion part of the same step.

The colouring is defined in such a way that the graph sequence of the black edges is a realization of version 2. Indeed, edges turning red are deleted and, hence, the copies of them do not appear in this model, but all other edges are black.

Our goal is to prove that the number of vertices having red edges divided by n tends to 0 a.s. This implies that the results of Corollary 1 hold for version 1 as well.

First, we need an upper bound for the total number of edges.

Lemma 2. *Denote by S_n the number of edges (both black and red) after n steps in version 1. Then for all $\varepsilon > 0$, we have $S_n = O\big(n \log^{1+\varepsilon} n\big)$ with probability* 1.

Proof. Let $\delta_n = S_n - S_{n-1}$. As before, $\mathcal{F}_n$ denotes the σ-field generated by the first n steps, and $X[n,d]$ is the number of vertices of degree d after n steps. Let U_n, respectively V_n, denote the degree of the old vertex selected for duplication, respectively deletion, at step n. The new vertex is connected to the duplicated vertex with an edge that cannot be deleted; this increases the number of edges by 1 for sure. Thus, $\delta_n = U_n - V_n + 1$. Clearly, U_n and V_n are conditionally independent and indentically distributed (i.i.d.) with respect to $\mathcal{F}_{n-1}$, hence, $S_n - n = \sum_{j=1}^{n}(\delta_j - 1)$ is a zero mean martingale. Consequently, $\mathbb{E}S_n = n$ for every n.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Clearly,

$$\mathbb{E}\{|\delta_n - 1| \mid \mathcal{F}_{n-1}\} \le 2\mathbb{E}\{U_n \mid \mathcal{F}_{n-1}\} = \sum_{d=0}^{n} \frac{X[n-1,d]}{n} d = \frac{2S_{n-1}}{n}.$$

Hence,

$$\mathbb{E}\left\{\sum_{n=2}^{\infty} \frac{|\delta_n - 1|}{n \log^{1+\varepsilon} n}\right\} < \infty,$$

therefore, the series

$$\sum_{n=2}^{\infty} \frac{\delta_n - 1}{n \log^{1+\varepsilon} n}$$

is convergent with probability 1. Then Kronecker's lemma [16, Lemma IV.3.2] implies that

$$\frac{S_n - n}{n \log^{1+\varepsilon} n} \to 0 \quad \text{a.s. as } n \to \infty.$$

Now we will colour some of the vertices red in such a way that the remaining black vertices cannot have any red edges. We will be able to provide an upper bound for the number of red vertices.

At the duplication step the new vertex becomes red if and only if the duplicated vertex is red. If this old vertex is black and has no red edges then the same holds for the new vertex at that moment. After that if there is an edge between the new vertex and the deleted vertex this edge may turn red, as we defined above. We colour both endpoints of this new red edge red. On the other hand, if the old vertex chosen for deletion loses all its edges then its new colour will be black. Note that black vertices still have only black edges, but it may happen that an old vertex has only one red edge which is deleted, because its other endpoint is chosen for deletion; in this case the vertex stays red without having any red edges.

The proof continues with giving an upper bound for the number of red vertices.

Lemma 3. *Denote by Z_n the number of red vertices after n steps. Then for all $\varepsilon > 0$, we have $Z_n = O(\log^{2+\varepsilon} n)$ a.s.*

Proof. At each step, every old vertex has the same probability to be duplicated or deleted. If a red vertex is duplicated, then the new vertex becomes red; if it is deleted, then Z_n decreases by 1 unless the deleted vertex is connected to the new one which turns this edge red. Therefore, without the exceptional new red edge, the conditional expectation of Z_n with respect to $\mathcal{F}_{n-1}$ would equal Z_{n-1}. The deleted vertex and the new vertex are connected if and only if the deleted and duplicated vertices are identical or they are connected to each other. Since we carried out sampling with replacement, the probability of the first event is $1/n$; while the probability of the second event is $2S_{n-1}/n^2$. In the first case, the new vertex is red originally, but the other vertex stays red instead of turning back to black when deleted; Z_n is increased by an extra 1. In the other case, both endpoints of the edge turning red may be also red vertices. To sum up, we obtain

$$\mathbb{E}\{Z_n \mid \mathcal{F}_{n-1}\} \le Z_{n-1} + \frac{1}{n} + 4\frac{S_{n-1}}{n^2}.$$

Set $\eta_n = Z_n - Z_{n-1}$. With this notation

$$\mathbb{E}\{\eta_n \mid \mathcal{F}_{n-1}\} \le \frac{1}{n} + 4\frac{S_{n-1}}{n^2}. \tag{6}$$

We have already shown that $\mathbb{E}S_{n-1} = n-1$, hence, $\mathbb{E}\eta_n \le 5/n$, and $\mathbb{E}Z_n = O(\log n)$.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Note that the number of red vertices cannot change by more than three at a single step, because if an old vertex is neither deleted, nor duplicated, it cannot be coloured red. Hence, $|\eta_n| \le 3$ for all n. Moreover, we can provide an upper bound on the probability that the number of red vertices changes at step n. Namely, it can change only if

- we duplicate and delete the same vertex; this has (conditional) probability $1/n$,
- the duplicated and the deleted vertices are connected to each other; this has probability $2S_{n-1}/n^2$, because there are S_{n-1} edges,
- a red vertex is duplicated; this has probability Z_{n-1}/n,
- a red vertex is deleted; this has probability Z_{n-1}/n.

Thus,

$$\mathbb{P}\{Z_n \ne Z_{n-1} \mid \mathcal{F}_{n-1}\} \le \frac{1}{n} + 2\frac{S_{n-1}}{n^2} + 2\frac{Z_{n-1}}{n}. \tag{7}$$

Therefore,

$$\mathbb{E}|\eta_n| \le 3\mathbb{P}\{Z_n \ne Z_{n-1}\} = O\Big(\frac{\log n}{n}\Big),$$

which implies that

$$\mathbb{E}\Bigg\{\sum_{n=2}^{\infty} \frac{|\eta_n|}{\log^{2+\varepsilon} n}\Bigg\} < \infty.$$

The proof can be completed with the help of Kronecker's lemma, as in the proof of Lemma 2.

Now we can complete the proof of Theorem 1.

Proof of Theorem 1. The total number of vertices is $n+1$ after n steps, hence the proportion of red vertices converges to 0 a.s. as $n \to \infty$. Since we defined the colours in such a way that red edges are exactly the edges that are present in version 1 but not present in version 2, and only red vertices may have red edges, it follows that the proportion of vertices having different degree in the two versions converges to 0. Corollary 1 states that for every d the proportion of vertices of degree d in version 2 converges a.s. to c_d. Now the same follows for version 1, which is the statement of Theorem 1.

Remark 1. We could have provided an upper bound for the conditional expectation of the number of red edges. The advantage of using red vertices is the uniform bound on the total change in their number; there is no such bound for the change in the number of red edges.

Remark 2. It follows that version 1 has a quite specific structure: it consists of cliques that are connected with relatively few edges (those are coloured red). An edge can be red only if both its endpoints are red, hence Lemma 3 provides an $O(\log^{4+\varepsilon} n)$ bound for the number of red edges.

This is not sharp; however, the estimates of Lemmas 2 and 3 can be further improved, which might be, as pointed out above, of independent interest. Thus, before turning to the proof of Theorem 2, we present the following improvement.

Proposition 2. *We have $S_n \sim n$ a.s. In addition, $Z_n = O(\log^{1+\varepsilon} n)$ for every $\varepsilon > 0$ a.s. as $n \to \infty$.*

Proof. First, we provide a crude bound for the maximal degree $M_n = \max\{d : X[n,d] > 0\}$. According to Lemma 2, $S_n = O(n \log^{1+\varepsilon} n)$ holds also for the number of edges in version 2.

Since a clique of size k contains $\binom{k}{2}$ edges it follows that the size of the maximal clique is $O(n^{1/2+\varepsilon})$. The same holds for the maximal degree in version 2; and, by Lemma 3, in version 1. Thus, $M_n = O(n^{1/2+\varepsilon})$ for every $\varepsilon > 0$.

Next, consider the martingale $S_n - n = \sum_{j=1}^{n}(\delta_j - 1)$ from the proof of Lemma 2. In order to prove that $S_n - n = o(\gamma_n)$ for a positive increasing predictable sequence (γ_n) it is sufficient to show that

$$\sum_{n=1}^{\infty} \gamma_n^{-2}\mathbb{E}\{(\delta_n - 1)^2 \mid \mathcal{F}_{n-1}\} < \infty$$

with probability 1 (see [16, Theorem VII.5.4]). To this end, we need to estimate the conditional variance of the martingale differences.

$$\begin{aligned}\operatorname{var}(\delta_n - 1 \mid \mathcal{F}_{n-1}) &= 2\operatorname{var}(U_n \mid \mathcal{F}_{n-1}) \\ &\le 2\mathbb{E}\{U_n^2 \mid \mathcal{F}_{n-1}\} \\ &= 2\sum_{d=1}^{n} \frac{X[n-1,d]}{n} d^2 \\ &\le \frac{2}{n} M_{n-1} \sum_{d=1}^{n} X[n-1,d]d \\ &= \frac{2}{n} M_{n-1} S_{n-1} \\ &= O(n^{1/2+\varepsilon})\end{aligned}$$

for every positive ε. Hence,

$$\sum_{n=1}^{\infty} \frac{\mathbb{E}\{(\delta_n - 1)^2 \mid \mathcal{F}_{n-1}\}}{n^{3/2+\varepsilon}} < \infty,$$

implying that

$$S_n - n = o(n^{3/4+\varepsilon}).$$

Thus, $S_n \sim n$ a.s.

Finally, let us consider the martingale $\zeta_n = \sum_{j=1}^{n}(\eta_j - \mathbb{E}\{\eta_j \mid \mathcal{F}_{j-1}\})$, where $\eta_n = Z_n - Z_{n-1}$, and derive an upper bound for the conditional variance of the differences. Keeping in mind that $|\eta_n| \le 3$ and using (7), we have

$$\begin{aligned}\mathbb{E}\{(\zeta_n - \zeta_{n-1})^2 \mid \mathcal{F}_{n-1}\} &= \operatorname{var}(\eta_n \mid \mathcal{F}_{n-1}) \\ &\le \mathbb{E}\{(Z_n - Z_{n-1}\}^2 \mid \mathcal{F}_{n-1}) \\ &\le 9\mathbb{P}\{Z_n \ne Z_{n-1} \mid \mathcal{F}_{n-1}\} \\ &\le 9\biggl(\frac{1}{n} + 2\frac{S_{n-1}}{n^2} + 2\frac{Z_{n-1}}{n}\biggr) \\ &= O\biggl(\frac{1 + Z_{n-1}}{n}\biggr).\end{aligned}$$

Now suppose that $Z_n = O(\log^{\alpha} n)$ is satisfied for some $\alpha > 0$. Then

$$\mathbb{E}\{(\zeta_n - \zeta_{n-1})^2 \mid \mathcal{F}_{n-1}\} = O\biggl(\frac{\log^{\alpha} n}{n}\biggr).$$

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Hence,

$$\sum_{n=2}^{\infty} \frac{\mathbb{E}\{(\zeta_n - \zeta_{n-1})^2 \mid \mathcal{F}_{n-1}\}}{\log^{\alpha+1+\varepsilon} n} < \infty$$

with probability 1. Again, by [16, Theorem VII.5.4], we have

$$\zeta_n = o(\log^{(\alpha+1)/2+\varepsilon}) \quad \text{a.s.} \tag{8}$$

for every positive ε.

Clearly,

$$Z_n = \sum_{j=1}^{n} \eta_j = \zeta_n + \sum_{j=1}^{n} \mathbb{E}\{\eta_j \mid \mathcal{F}_{j-1}\},$$

where the last sum can be estimated with the help of (6) in the following way. Since $S_{n-1} \sim n$, we have $\mathbb{E}\{\eta_n \mid \mathcal{F}_{n-1}\} = O(1/n)$. Hence,

$$\sum_{j=1}^{n} \mathbb{E}\{\eta_j \mid \mathcal{F}_{j-1}\} = O(\log n).$$

This, combined with (8) proves that $Z_n = O(\log^{(\alpha+1)/2+\varepsilon})$ holds a.s. for all $\varepsilon > 0$. Using Lemma 3, we can start from $\alpha = 2+\varepsilon$ and repeating the argument we arrive at the a.s. estimation $Z_n = O(\log^{1+\varepsilon})$ for all $\varepsilon > 0$.

Proof of Theorem 2. Let $G(z)$ denote the generating function of the sequence (c_d), i.e.

$$G(z) = \sum_{d=0}^{\infty} c_d z^d, \qquad |z| \le 1.$$

Multiplying $(d+1)(c_{d-1} + c_{d+1}) = (2d+3)c_d$ by z^d then summing from $d = 1$ to ∞ and using the fact that $c_0 = (1 + c_1)/3$, we obtain an inhomogeneous linear differential equation for $G(z)$.

$$(1-z)^2 G'(z) = (3-2z)G(z) - 1, \qquad G(0) = c_0.$$

Solving this equation we obtain the following expression:

$$G(z) = \frac{c(z)}{(1-z)^2} \exp\biggl(\frac{z}{1-z}\biggr),$$

where

$$c(z) = c_0 - \int_0^z \exp\biggl(-\frac{y}{1-y}\biggr)\,\mathrm{d}y.$$

By definition, c_d is the a.s. limit of $X[n,d]/(n+1)$, which is the proportion of vertices of degree d after n steps. Hence, for each fixed D, we have $\sum_{d=0}^{D} c_d = \lim_{n\to\infty} \sum_{d=0}^{D} X[n,d]/(n+1) \le 1$. This implies that $G(1) = \sum_{d=0}^{\infty} c_d \le 1$. It follows that

$$c_0 = \int_0^1 \exp\biggl(-\frac{y}{1-y}\biggr)\,\mathrm{d}y,$$

hence, via the substitution $x = 1 - y$,

$$c(z) = \int_z^1 \exp\biggl(-\frac{y}{1-y}\biggr)\,\mathrm{d}y = \int_0^{1-z} \exp\biggl(1 - \frac{1}{x}\biggr)\,\mathrm{d}x.$$

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Thus, we have

$$G(z) = \int_0^{1-z} \exp\biggl(1 - \frac{1}{x}\biggr)\,\mathrm{d}x\; \frac{1}{(1-z)^2} \exp\biggl(\frac{z}{1-z}\biggr),$$

from which, by substituting $y = 1/x - 1/(1-z)$, we obtain

$$\begin{aligned} G(z) &= \int_0^\infty \frac{\mathrm{e}^{-y}}{(1+(1-z)y)^2}\,\mathrm{d}y \\ &= \int_0^\infty \frac{\mathrm{e}^{-y}}{(1+y)^2(1-z(y/(1+y)))^2}\,\mathrm{d}y \\ &= \int_0^\infty \sum_{d=0}^\infty (d+1)\frac{z^d y^d\,\mathrm{e}^{-y}}{(1+y)^{d+2}}\,\mathrm{d}y \\ &= \sum_{d=0}^\infty z^d\,(d+1) \int_0^\infty \frac{y^d \mathrm{e}^{-y}}{(1+y)^{d+2}}\,\mathrm{d}y, \end{aligned} \tag{9}$$

completing the proof of the first statement of the theorem.

In addition, note that the first equality of (9) immediately implies that $\sum_{d=0}^{\infty} c_d = G(1) = 1$.

Proof of Theorem 3. In order to approximate the integral of Theorem 2, we first analyse the behaviour of the integrand around the point where it attains its maximum. Let

$$y_d = \arg\max \frac{y^d \mathrm{e}^{-y}}{(1+y)^{d+2}} = \arg\max f(y),$$

where

$$f(y) = d \log y - (d+2)\log(1+y) - y.$$

Clearly,

$$f'(y) = \frac{d}{y} - \frac{d+2}{y+1} - 1 = -\frac{y^2+3y-d}{y(y+1)},$$

$$f''(y) = -\frac{d}{y^2} + \frac{d+2}{(y+1)^2} = \frac{2y^2 - 2dy - d}{y^2(y+1)^2}, \qquad f'''(y) = \frac{2d}{y^3} - \frac{2(d+2)}{(y+1)^3}.$$

Since y_d satisfies $f'(y_d) = 0$, we obtain

$$y_d = -\tfrac{3}{2} + \sqrt{d + \tfrac{9}{4}} = \sqrt{d} - \tfrac{3}{2} + o(1).$$

We introduce the new variable $t = t(y) = y_d^{-1/2}(y - y_d)$, i.e. $y = y_d + y_d^{1/2}t$. This will turn out to be useful, because the function f is concentrated around y_d of the order of $y_d^{-1/2}$ (which is just $d^{-1/4}$); i.e. $f(y_d + y_d^{1/2}t)$ converges as d goes to ∞. Then

$$g(t) := f(y) - f(y_d) = \frac{y_d}{2} f''(y_d + \theta y_d^{1/2}t)t^2,$$

where $\theta = \theta(d,t)$ belongs to the interval $[0; 1]$. For every fixed t,

$$f''(y_d + \theta y_d^{1/2}t) \sim -2y_d^{-1},$$

thus, $g(t) \to -t^2$ as $d \to \infty$. Moreover, for $y \le y_d$, i.e. for $y_d^{1/2} \le t \le 0$, we have $f'(y) \ge 0$.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

Thus, $d/y-(d+2)/(y+1) > 0$ holds, and after rearranging, we obtain $(d+2)/d < (y+1)/y$. This yields $(d+2)/d < (y+1)^3/y^3$ is satisfied, which implies that $f'''(y) \geq 0$. Hence,

$$g(t) \leq \frac{y_d}{2} f''(y_d)t^2 = a_d t^2,$$

where $a_d \to -1$, as $d \to \infty$. On the other hand, let $y_d \leq y \leq (3/2)y_d$, i.e. $0 \leq t \leq (1/2)y_d^{1/2}$. In this domain f''' is increasing, hence, $f'''(y) \leq f'''(y_d) \sim 6dy_d^{-4} \sim 6d^{-1}$. Thus,

$$g(t) \leq \frac{y_d}{2} f''(y_d)t^2 + \frac{1}{6} y_d^{3/2} f'''(y_d)t^3 \leq \biggl(\frac{y_d}{2} f''(y_d) + \frac{y_d^2}{12} f'''(y_d)\biggr)t^2 = b_d t^2,$$

where $b_d \to -\frac{1}{2}$ as $d \to \infty$.

Thus, by the dominated convergence theorem,

$$\begin{aligned}\int_0^{3y_d/2} \mathrm{e}^{f(y)}\,\mathrm{d}y &= y_d^{1/2} \int_{-y_d^{1/2}}^{(1/2)y_d^{1/2}} \exp(f(y_d)+g(t))\,\mathrm{d}t \\ &\sim y_d^{1/2} \exp(f(y_d)) \int_{-\infty}^{+\infty} \exp(-t^2)\,\mathrm{d}t \\ &= \sqrt{\pi}\, y_d^{1/2} \exp(f(y_d)).\end{aligned}$$

We have

$$f(y_d) = -2\log y_d - (d+2)\log\biggl(1+\frac{1}{y_d}\biggr) - y_d,$$

and

$$\begin{aligned}(d+2)\log\biggl(1+\frac{1}{y_d}\biggr) &= (d+2)\biggl(\frac{1}{y_d} - \frac{1}{2y_d^2}\biggr) + o(1) \\ &= y_d + \frac{(d+2)(2y_d-1) - 2y_d^3}{2y_d^2} + o(1) \\ &= y_d + \frac{(y_d^2+3y_d+2)(2y_d-1) - 2y_d^3}{2y_d^2} + o(1) \\ &= y_d + \frac{5y_d^2+y_d-2}{2y_d^2} + o(1),\end{aligned}$$

where we used the fact that $y_d^2 + 3y_d = d$. Thus,

$$f(y_d) = -2\log y_d - 2y_d - \tfrac{5}{2} + o(1) = -2\log y_d - 2\sqrt{d} + \tfrac{1}{2} + o(1).$$

Finally,

$$\begin{aligned}\int_{3y_d/2}^{\infty} \mathrm{e}^{f(y)}\,\mathrm{d}y &\leq (2y_d)^{-2} \int_{3y_d/2}^{\infty} \biggl(1-\frac{1}{1+y}\biggr)^d \mathrm{e}^{-y}\underset{.}{\mathrm{y}} \\ &\leq (2y_d)^{-2} \int_{3y_d/2}^{\infty} \exp\biggl(-\frac{d}{y+1} - y\biggr)\,\mathrm{d}y.\end{aligned}$$

The exponent on the right-hand side can be estimated with the help of the inequality of arithmetic and geometric means as follows:

$$-\frac{d}{y+1}-y=-\frac{d}{y+1}-\frac{y+1}{2}-\frac{y-1}{2}\le-\sqrt{2d}-\frac{y-1}{2}.$$

Hence,

$$\int_{3y_d/2}^{\infty}\mathrm{e}^{f(y)}\,\mathrm{d}y\le(2y_d)^{-2}\exp\bigl(-\sqrt{2d}+\tfrac{1}{2}-\tfrac{3}{4}y_d\bigr)=o\bigl(y_d^{-2}\exp\bigl(-2\sqrt{d}\bigr)\bigr).$$

From all these, we obtain

$$c_d=(d+1)\int_0^{\infty}\mathrm{e}^{f(y)}\,\mathrm{d}y\sim(\mathrm{e}\pi)^{1/2}d^{1/4}\mathrm{e}^{-2\sqrt{d}}$$

as claimed.

Proof of Theorem 4. Black vertices have the same local clustering coefficient in both versions. Since the proportion of red vertices tends to be negligible as $n\to\infty$, the limit of the average clustering coefficient is also the same in both versions. The global clustering coefficient of version 2 is identically equal to 1. In its defining fraction the numerator and the denominator are proportional to n. When turning to version 1 the denominator has to be increased by the number of triplets containing at least one red edge. Such a triplet must have a red central vertex and at least one more red vertex. Hence, the increment of the denominator cannot exceed $M_nZ_n^2$, where M_n denotes the maximal degree, and Z_n the number of red vertices. In the proof of Proposition 2, we have shown that $M_n=O(n^{1/2+\varepsilon})$ and $Z_n=O(\log^{1+\varepsilon}n)$, thus, the increment of the denominator is asymptotically negligible with respect to n. Hence, the global clustering coefficient of version 1 must converge to 1.

Acknowledgement

This research was supported by the Hungarian Scientific Research Fund (OTKA) (grant no. K 108615).

References

[1] Backhausz, Á. and Móri, T. F. (2014). A random model of publication activity. *Discrete Appl. Math.* **162,** 78–89.

[2] Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science* **286,** 509–512.

[3] Cooper, C. and Frieze, A. (2003). A general model of web graphs. *Random Structures Algorithms* **22,** 311–335.

[4] Bebek, G. *et al.* (2006). The degree distribution of the generalized duplication model. *Theoret. Comput. Sci.* **369,** 239–249.

[5] Bollobás, B., Riordan, O., Spencer, J. and Tusnády, G. (2001). The degree sequence of a scale-free random graph process. *Random Structures Algorithms* **18,** 279–290.

[6] Chung, F., Lu, L., Dewey, T. G. and Galas, D. J. (2003). Duplication models for biological networks. *J. Comput. Biol.* **10,** 677–687.

[7] Cohen, N., Jordan, J. and Voliotis, M. (2010). Preferential duplication graphs. *J. Appl. Prob.* **47**, 572–585.

[8] de Solla Price, D. (1976). A general theory of bibliometric and other cumulative advantage processes. *J. Amer. Soc. Inf. Sci.* **27,** 292–306.

[9] Dong, R., Goldschmidt, C. and Martin, J. B. (2006). Coagulation-fragmentation duality, Poisson–Dirichlet distributions and random recursive trees. *Ann. Appl. Prob.* **16,** 1733–1750.

[10] Faloutsos, M., Faloutsos, P. and Faloutsos, C. (1999). On power-law relationships of the internet topology. In *Proceedings of the Conference on Applications, Technologies, Architectures, and Protocols for Computer Communication, SIGCOMM '99*, ACM, New York, pp. 251–262.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press

[11] Hamdi, M., Krishnamurthy, V. and Yin, G. (2014). Tracking a Markov- modulated stationary degree distribution of a dynamic random graph, *IEEE Trans. Inf. Theory* **60,** 6609–6625.

[12] Jordan, J. (2011). Randomised reproducing graphs. *Electron. J. Prob.* **16,** 1549–1562.

[13] Kim, J., Krapivsky, P. L., Kahng, B. and Redner, S. (2002). Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev. E* **66,** 055101(R).

[14] Pastor-Satorras, R., Smith, E. and Solé, R. V. (2003). Evolving protein interaction networks through gene duplication. *J. Theoret. Biol.* **222,** 199–210.

[15] Ráth, B. and Tóth, B. (2009). Erdős–Rényi random graphs + forest fires = self-organized criticality. *Electron. J. Prob.* **14,** 1290–1327.

[16] Shiryaev, A. N. (1996). *Probability,* 2nd edn. Springer, New York.

[17] Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika* **42,** 425–440.

[18] Sridharan, A., Gao, Y., Wu, K. and Nastos, J. (2011). Statistical behavior of embeddedness and communities of overlapping cliques in online social networks. In *Proc. IEEE INFOCOM 2011,* IEEE, New York, pp. 546–550.

[19] Szymański, J. (1987). On a nonuniform random recursive tree, In *Random Graphs '85* (Ann. Discrete Math. **33,** North-Holland, Amsterdam, pp. 297–306.

[20] van der Hofstad, R. *Random graphs and complex networks.* Preprint, Eindhoven University of Technology. Available at http://www.win.tue.nl/˜rhofstad/NotesRGCN.pdf.

[21] Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature* **393,** 440–442.

[22] Willinger, W., Alderson, D. and Doyle, J. C. (2009). Mathematics and the Internet: a source of enormous confusion and great potential. *Notices Amer. Math. Soc.* **56,** 586–599.

[23] Yule, G. U. (1925). A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis, F.R.S. *Philos. Trans. R. Soc. London B.* **213,** 21–87.

https://doi.org/10.1239/jap/1437658604 Published online by Cambridge University Press