Taylor & Francis
Taylor & Francis Group

ASYMPTOTIC DEGREE DISTRIBUTION OF A DUPLICATION–DELETION RANDOM GRAPH MODEL

Erik Thörnblad

Department of Mathematics, University of Uppsala, Uppsala, Sweden

Abstract *We study a discrete-time duplication-deletion random graph model and analyze its asymptotic degree distribution. The random graphs consist of disjoint cliques. In each time step, either a new vertex is brought in with probability $0 < p < 1$ and attached to an existing clique, chosen with probability proportional to the clique size, or all the edges of a random vertex are deleted with probability $1 - p$. We prove almost sure convergence of the asymptotic degree distribution and find its exact values in terms of a hypergeometric integral, expressed in terms of the parameter p. In the regime $0 < p < \frac{1}{2}$, we show that the degree sequence decays exponentially at rate $\frac{p}{1-p}$, whereas it satisfies a power law with exponent $\frac{p}{2p-1}$ if $\frac{1}{2} < p < 1$. At the threshold $p = \frac{1}{2}$, the degree sequence lies between a power law and exponential decay.*

1. INTRODUCTION

Over the past few decades, dynamic random graph models have been studied extensively. These are random graphs that evolve over time; for instance, by introducing new vertices and edges (and possibly deleting them). Perhaps the most well-known example of a dynamic random graph is the preferential attachment model [4]. In this model, new vertices are brought in one at a time, and are connected to a single old vertex chosen with probability proportional to the degree of the target vertex. Subsequently, this was analyzed in, among others [5] (which also allowed for new vertices to attach to several old vertices), and [13]. One of the main results of the preferential attachment model [4] is that the degree distribution satisfies a power law a.s.; that is, that the asymptotic proportion of vertices of degree k decays as $k^{-\alpha}$ for some constant $\alpha > 0$. Many other models incorporating some sort of preferential attachment rule also exhibit power-law behavior. Furthermore, it is possible to allow for more complicated dynamics without destroying the power-law property. One possible variation is to also allow for deletion of vertices or edges. A dynamic random graph model that allowed for addition of new vertices and edges (with endpoints chosen proportionally to degree) and deletion of randomly chosen vertices or edges was considered [8]. A number of properties of this random graph were determined, in particular, that the degree distribution satisfied a power law a.s., with the coefficient being a function of the addition and deletion probabilities [8].

However, in several models allowing for edge or vertex deletion, a phase transition is seen for the asymptotic degree distribution. Typically, this phase transition occurs when the

Address correspondence to Erik Thörnblad, Department of Mathematics, University of Uppsala, Box 480, S-75106 Uppsala, Sweden. E-mail: erik.thornblad@math.uu.se

probability of deletion is too high. For instance, Deijfen and Lindholm [10] considered a model for which the asymptotic expected degree distribution satisfied a power law behavior in the regime below some threshold deletion probability, while it decayed exponentially above this threshold value. However, there was no comment on the behavior at the threshold value. Similar results were achieved, and the asymptotic degree distribution at the threshold probability for another random graph model was also found [18]. The work in [10, 18] was subsequently extended in [17], which study showed that a similar phenomenon occurs in a random graph model considered by Cooper et al. [9], who, in their original work had restricted themselves to the power-law regime. We shall see a similar phase transition for the random graph model considered here.

In this article, we study the following random graph process. Let $0 < p < 1$ be fixed. Start at time $m = 0$ with a graph G_0 consisting of a single isolated vertex. At integer times $m \geq 1$ generate G_m from G_{m-1} by doing one of the following steps.

1. With probability p, do a *duplication step*. Choose uniformly at random a vertex v from G_{m-1}. Introduce a new vertex w and form edges between w and all the neighbors of v. Also form an edge between v and w.
2. With probability $1 - p$, do a *deletion step*. Choose uniformly at random a vertex from G_{m-1} and make it isolated by deleting all its incident edges.

The results in this study concern the limiting degree distribution of the vertices in the graph, that is, the limiting proportion of vertices of degree k. We will refer to the three regimes $0 < p < \frac{1}{2}$, $p = \frac{1}{2}$, and $\frac{1}{2} < p < 1$ as the subcritical case, the critical case, and the supercritical case, respectively. Since duplication and deletion steps affect a whole clique, vertices in large cliques are more likely to have edges added and edges removed. In this sense, edges are attached and deleted preferentially. The subcritical, critical, and supercritical cases will require separate analyse, and we will see that a phase transition in the asymptotic degree distribution occurs, passing from exponential decay to power law at the critical value $p = \frac{1}{2}$.

This model is motivated by the recent work [2], in which the random graph process introduced has the same deletion and duplication steps, but duplication and deletion steps are carried out alternatingly. Hence, there is no parameter p and no randomness in the step type. We note that, in the critical case $p = \frac{1}{2}$, on average we do as many duplication steps as deletion steps. One might, therefore, expect that the model [2] has similar properties to our model in the critical case. We will see later that the asymptotic degree distribution in the critical case agrees with the one found in [2].

In a biological context, it is more natural to study the corresponding continuous-time version of the model. Such studies were done recently [6, 7] and earlier [15]. In this framework, one would attach to each clique two exponential clocks, ringing at rate $k\lambda$ and $k\mu$, respectively, where k is the size of the clique and $\lambda, \mu > 0$ two parameters. If the first clock rings, a new vertex is added to the clique. If the second clock rings, a vertex is removed from the clique and made isolated. Indeed, the findings in the present study agree with the results given in [6] in the critical and supercritical cases. However, no corresponding result for the subcritical case is given in [6].

The rest of the article is outlined as follows. Let $D_{m,k}$ denote the number of vertices of degree $k - 1$ (that is, the number of vertices in cliques of size k) at time m, and let N_m denote the number of vertices at time m. In Section 2, we find a positive bounded sequence $(d_k)_{k=1}^{\infty}$ such that $\liminf_{m \to \infty} D_{m,k}/N_m \geq d_k$ and $(d_k)_{k=1}^{\infty}$ satisfies a recurrence relation. In Section

3, we prove that the limit $\lim_{m\to\infty} D_{m,k}/N_m = d_k$ exists a.s. The critical case has been analyzed [2], but in Section 4 we consider the limit sequence $(d_k)_{k=1}^{\infty}$ for the subcritical and supercritical cases. Using Laplace's method, we derive its exact values expressed in terms of certain hypergeometric integrals. We also consider the asymptotics of these integrals and show that the supercritical case gives rise to a power law with exponent $\beta = \frac{p}{2p-1}$. For the subcritical case, we obtain that d_k decays exponentially as γ^{-k} where $\gamma = \frac{1-p}{p}$. Finally, we show that this sequence indeed defines a probability distribution for all $0 < p < 1$.

2. A BOUND ON THE LIMIT INFERIOR

This section is devoted to proving the following theorem.

Theorem 2.1. *Let $D_{m,k}$ denote the number of vertices of degree $k-1$ (that is, the number of vertices in k-cliques) at time m. Denote by N_m the total number of vertices at time m. Then, for each $k \geq 1$ we have that*

$$\liminf_{m\to\infty} \frac{D_{m,k}}{N_m} \geq d_k \tag{2.1}$$

a.s., where $(d_k)_{k=0}^{\infty}$ is the unique positive bounded sequence satisfying

$$d_0 = \frac{1-p}{p}, \tag{2.2}$$

$$d_k = \frac{pkd_{k-1} + (1-p)kd_{k+1}}{k+p}, \quad k \geq 1. \tag{2.3}$$

It will later be shown that the sequence $(d_k)_{k=1}^{\infty}$ indeed is the asymptotic degree distribution. We will prove Theorem 2.1 by proving a similar result on the clique sizes. First, we find expressions for the expected number of k-cliques at time m, conditional on the graph at time $m-1$. Let $C_{m,k}$ denote the number of k–cliques at time m. Denote by $\mathcal{F}_m$ the σ-field generated by the random graph process up to and including time m. Since we start with a single isolated vertex at time 0, we have the boundary conditions $C_{0,1} = 1$ and $C_{0,k} = 0$ for $k \geq 2$.

The 1-cliques at time m originate from three sources.

1. A 1-clique at time $m-1$ not chosen for duplication (recall that a 1-clique selected for deletion is left unaffected). This happens for each 1-clique with probability $1 - p/N_{m-1}$.
2. A k-clique ($k > 2$) was selected for deletion. This happens with probability $(1-p)k/N_{m-1}$ for each k-clique and gives rise to one new 1-clique.
3. A 2-clique was selected for deletion. This happens with probability $(1-p)2/N_{m-1}$ for each 2-clique, and this gives rise to two new 1-cliques.

This leads to the following conditional expectation of the number of 1-cliques.

$$\mathbb{E}[C_{m,1} \mid \mathcal{F}_{m-1}] = C_{m-1,1}\left(1 - p\frac{1}{N_{m-1}}\right) + 4(1-p)\frac{C_{m-1,2}}{N_{m-1}}$$

$$+ (1-p)\sum_{i=3}^{m}\frac{iC_{m-1,i}}{N_{m-1}}. \tag{2.4}$$

Note that $\sum_{i=3}^{m} iC_{m-1,i} = N_{m-1} - C_{m-1,1} - 2C_{m-1,2}$. Thus,

$$\mathbb{E}[C_{m,1} \mid \mathcal{F}_{m-1}] = C_{m-1,1}\left(1 - \frac{1}{N_{m-1}}\right) + (1-p) + 2(1-p)\frac{C_{m-1,2}}{N_{m-1}}. \tag{2.5}$$

We do a similar analysis for the number of k-cliques ($k \geq 2$) at time m. These come from three sources.

1. The k-cliques at time $m - 1$ that were not selected for duplication or deletion. For each k-clique, this happens with probability $1 - k/N_{m-1}$.
2. A $(k - 1)$-clique at time $m - 1$ was selected for duplication. This happens with probability $p(k - 1)/N_{m-1}$ and gives rise to one new k-clique.
3. A $(k + 1)$-clique at time $m - 1$ was selected for deletion. This happens with probability $(1 - p)(k + 1)/N_{m-1}$ and gives rise to one new k-clique.

This gives us the following conditional expectation of the number of k-cliques at time m:

$$\mathbb{E}[C_{m,k} \mid \mathcal{F}_{m-1}] = C_{m-1,k}\left(1 - \frac{k}{N_{m-1}}\right) + p(k-1)\frac{C_{m-1,k-1}}{N_{m-1}}$$

$$+ (1-p)(k+1)\frac{C_{m-1,k+1}}{N_{m-1}}. \tag{2.6}$$

We now quote an essential lemma [3]. The proof uses martingale techniques. The following version is a slightly less general version, adapted to our purposes. There is a corresponding result for the limit superior that we shall not use, and therefore, do not quote here. Furthermore, although not immediately applicable to our purposes because it would require us to know a priori that the stochastic processes involved are a.s. convergent, it seems likely that a modification of Lemma 3.1 in [11] should give a result very similar to the following lemma.

Lemma 2.2. [3] *Let $(\mathcal{F}_m)_{m=0}^{\infty}$ be a filtration. Let $(\xi_m)_{m=0}^{\infty}$ be a nonnegative process adapted to $(\mathcal{F}_m)_{m=0}^{\infty}$, and let $(u_m)_{m=1}^{\infty}$, $(v_m)_{m=1}^{\infty}$ be nonnegative predictable processes such that $u_m < m$ for all $m \geq 1$ and $\lim_{m\to\infty} u_m = u > 0$ exists a.s. Let w be a positive constant. Suppose that there exists $\delta > 0$ such that $\mathbb{E}[(\xi_m - \xi_{m-1})^2 \mid \mathcal{F}_m] = O(m^{1-\delta})$.*
If

$$\liminf_{m\to\infty}\frac{v_m}{w} \geq v \tag{2.7}$$

for some constant $v \geq 0$ and

$$\mathbb{E}[\xi_m \mid \mathcal{F}_{m-1}] \geq \left(1 - \frac{u_m}{m}\right)\xi_{m-1} + v_m, \qquad m \geq 1, \tag{2.8}$$

then

$$\liminf_{m\to\infty} \frac{\xi_m}{mw} \geq \frac{v}{u+1} \quad a.s. \tag{2.9}$$

In order to prove Theorem 2.1, we will prove the following theorem.

Theorem 2.3. *Let $C_{m,k}$ denote the number of k-cliques at time m. Denote by N_m the total number of vertices in the graph at time m. Then, for all $k \geq 1$ we have that*

$$\liminf_{m\to\infty} \frac{C_{m,k}}{N_m} \geq c_k \tag{2.10}$$

a.s., where $(c_k)_{k=1}^{\infty}$ is a positive sequence satisfying

$$c_1 = \frac{(1-p)(1+2c_2)}{1+p}, \tag{2.11}$$

$$c_k = \frac{p(k-1)c_{k-1} + (1-p)(k+1)c_{k+1}}{k+p}, \quad k \geq 2. \tag{2.12}$$

We outline the idea behind the proof of Theorem 2.3.

1. Let $A_k = \liminf_{m\to\infty} \frac{C_{m,k}}{N_m}$ for each $k \geq 1$. For each $k \geq 1$, we find a sequence $(a_k^{(j)})_{j=0}^{\infty}$ such that $a_k^{(j)} \leq A_k$ for all $j \geq 0$. This sequence is defined recursively and the inequalities are proved by induction.
2. For each $k \geq 1$, we prove that the sequence $(a_k^{(j)})_{j=0}^{\infty}$ is monotonically increasing (in j). Since it lies in the bounded set $[0, 1]$ it must be convergent to some a_k. By construction we have that $a_k \leq A_k$ for all $k \geq 1$.

The proof is similar to the first half of the proof of Proposition 5 in [2]; in that study, however, the limit superior was also bound from above, and it was shown that the limit superior and limit inferior are equal. This would work in this case as well (at least after one passes from the cliques to the individual vertices, as we do in the next section), but we prefer another method.

Proof of Theorem 2.3.

Step 1. Define $A_k = \liminf_{m\to\infty} \frac{C_{m,k}}{N_m}$. Since the number of vertices increases by 1 in a duplication step and is left unaffected in a deletion step, the strong law of large numbers implies that

$$\lim_{m\to\infty} \frac{N_m}{pm} = 1, \tag{2.13}$$

which means that

$$A_k = \liminf_{m\to\infty} \frac{C_{m,k}}{pm}. \tag{2.14}$$

This representation will allow us to invoke Lemma 2.2, so we will use this from now on.

For each $k \geq 1$, we now construct a sequence $(a_k^{(j)})_{j=0}^{\infty}$ and prove by induction that it satisfies $a_k^{(j)} \leq A_k$ for all $j \geq 0$. Let $a_k^{(0)} = 0$ for all $k \geq 1$. Define recursively the sequence $(a_k^{(j)})_{j=0}^{\infty}$ by

$$a_1^{(j+1)} = \frac{(1-p)(1+2a_2^{(j)})}{1+p} \tag{2.15}$$

and

$$a_k^{(j+1)} = \frac{p(k-1)a_{k-1}^{(j)} + (1-p)(k+1)a_{k+1}^{(j)}}{k+p}, \qquad k \geq 2. \tag{2.16}$$

The induction statement is true for $j = 0$, by definition. Suppose that there exists some j such that $a_k^{(j)} \leq A_k$ for all $k \geq 1$. For $k = 1$ define the following variables, where we use the notation in Lemma 2.2.

$$\begin{cases} \xi_m = C_{m,1}, \\ w = p, \\ u_m = \frac{m}{N_{m-1}}, \\ v_m = (1-p) + 2(1-p)\frac{C_{m-1,2}}{N_{m-1}}. \end{cases} \tag{2.17}$$

Note that u_m and v_m are both positive predictable sequences, that ξ_m is a non-negative adapted sequence, and that $u_m \to \frac{1}{p} =: u$ a.s. By the induction hypothesis, $\liminf_{m\to\infty} \frac{C_{m-1,2}}{N_{m-1}} \geq a_2^{(j)}$. Choosing $v = \frac{(1-p)(1+2a_2^{(j)})}{p}$, we ensure that $\liminf_{m\to\infty} \frac{v_m}{w} \geq v$. The technical condition $\mathbb{E}[(\xi_m - \xi_{m-1})^2|\mathcal{F}_{m-1}] = O(m^{1-\delta})$ is satisfied because the maximum change in the number of 1-cliques is 2. By Lemma 2.2 and (2.5), we have that

$$
\begin{aligned}
A_1 = \liminf_{m\to\infty} \frac{C_{m,1}}{N_m} = \liminf_{m\to\infty} \frac{C_{m,1}}{pm} &\geq \frac{v}{u+1} \\
&= \frac{(1-p)(1+2a_2^{(j)})}{p(\frac{1}{p}+1)} \\
&= \frac{(1-p)(1+2a_2^{(j)})}{1+p} \\
&= a_1^{(j+1)}.
\end{aligned} \tag{2.18}
$$

For $k > 1$ we define the following:

$$\begin{cases} \xi_m = C_{m,k}, \\ w = p, \\ u_m = \frac{km}{N_{m-1}}, \\ v_m = p(k-1)\frac{C_{m-1,k-1}}{N_{m-1}} + (1-p)(k+1)\frac{C_{m-1,k+1}}{N_{m-1}}. \end{cases} \tag{2.19}$$

Note that $u_m \to \frac{k}{p}$ a.s. and that

$$\liminf_{m\to\infty} \frac{v_m}{p} \geq \frac{p(k-1)a_{k-1}^{(j)} + (1-p)(k+1)a_{k+1}^{(j)}}{p} =: v \qquad (2.20)$$

by the induction hypothesis. The technical condition $\mathbb{E}[(\xi_m - \xi_{m-1})^2|\mathcal{F}_{m-1}] = O(m^{1-\delta})$ is satisfied since the number of k-cliques changes by, at most, one. By Lemma 2.2 and (2.6), we have that

$$\begin{aligned}
A_k = \liminf_{m\to\infty} \frac{C_{m,k}}{N_m} = \liminf_{m\to\infty} \frac{C_{m,k}}{pm} &\geq \frac{v}{u+1} \\
&= \frac{p(k-1)a_{k-1}^{(j)} + (1-p)(k+1)a_{k+1}^{(j)}}{p(k/p+1)} \\
&= \frac{p(k-1)a_{k-1}^{(j)} + (1-p)(k+1)a_{k+1}^{(j)}}{k+p} \\
&= a_k^{(j+1)}.
\end{aligned} \qquad (2.21)$$

This completes the proof that $a_k^{(j)} \leq A_k$ for all $j \geq 0$ and all $k \geq 1$.

Step 2. We now prove that the sequence $(a_k^{(j)})_{j=0}^{\infty}$ is increasing for any fixed k. We prove this by induction. Since $a_k^{(0)} = 0$ for all k, it is clear that $a_k^{(1)} \geq a_k^{(0)}$. Suppose that the induction statement is true for some j. For $k = 1$ we have that

$$a_1^{(j+1)} = \frac{(1-p)(1+2a_2^{(j)})}{1+p} \geq \frac{(1-p)(1+2a_2^{(j-1)})}{1+p} = a_1^{(j)}. \qquad (2.22)$$

For $k > 1$ we have that

$$\begin{aligned}
a_k^{(j+1)} &= \frac{p(k-1)a_{k-1}^{(j)} + (1-p)(k+1)a_{k+1}^{(j)}}{k+p} \\
&\geq \frac{p(k-1)a_{k-1}^{(j-1)} + (1-p)(k+1)a_{k+1}^{(j-1)}}{k+p} = a_k^{(j)}.
\end{aligned} \qquad (2.23)$$

Hence, the sequence $(a_k^{(j)})_{j=0}^{\infty}$ is an increasing sequence. Since $0 \leq a_k^{(j)} \leq A_k \leq 1$, we have that the sequence is bounded above by 1, and so it must be convergent and have a limit c_k. The limit sequence $(c_k)_{k=1}^{\infty}$ satisfies the recurrence

$$c_k = \frac{p(k-1)c_{k-1} + (1-p)(k+1)c_{k+1}}{k+p} \qquad k \geq 2, \qquad (2.24)$$

$$c_1 = \frac{(1-p)(1+2c_2)}{1+p}, \qquad (2.25)$$

which is seen by taking limits in (2.15) and (2.16).

$\square$

Let us now, instead, consider the degree distribution and use Theorem 2.3 to prove Theorem 2.1. The first part of the proof follows straightforwardly from Theorem 2.3.

Proof of Theorem 2.1. There are k vertices in each k-clique, so $D_{m,k} = kC_{m,k}$. By Theorem 2.3, we have that $D_{m,k}/N_m \to kc_k$ a.s. Defining $d_0 = \frac{1-p}{p}$ and putting $d_k = kc_k$ in (2.11) and (2.12), we retrieve (2.3) for $k \geq 1$.

Since $0 \leq \frac{D_{m,k}}{N_m} \leq 1$ and the sequence $(d_k)_{k=1}^{\infty}$ is positive, we have that $(d_k)_{k=1}^{\infty}$ is a bounded sequence in $(0, 1]$.

For uniqueness, suppose for contradiction that there are two distinct bounded sequences satisfying (2.2) and (2.3), $(d_k)_{k=0}^{\infty}$ and $(\hat{d}_k)_{k=0}^{\infty}$, say. Without loss of generality, suppose that $d_1 > \hat{d}_1$. Define $\theta_k = d_k - \hat{d}_k$ for all $k \geq 0$. This is a bounded sequence satisfying the homogeneous system of equations

$$\theta_0 = 0, \tag{2.26}$$

$$\theta_{k+1} = \frac{(k+p)\theta_k - pk\theta_{k-1}}{(1-p)k}, \quad k \geq 1. \tag{2.27}$$

We prove by induction that

$$\theta_{k+1} \geq \left(1 + \frac{p}{(1-p)k + \mathbb{1}_{\{k=0\}}}\right)\theta_k \tag{2.28}$$

for all $k \geq 0$. Since $\theta_0 = 0$ and $\theta_1 = d_1 - \hat{d}_1 > 0$, this is certainly true for $k = 0$. Suppose it holds true for $k - 1$. This implies trivially that $\theta_k \geq \theta_{k-1}$. But then,

$$\theta_{k+1} = \frac{(k+p)\theta_k - pk\theta_{k-1}}{(1-p)k} \geq \frac{k+p-kp}{(1-p)k}\theta_k = \frac{(1-p)k+p}{(1-p)k}\theta_k$$

$$= \left(1 + \frac{p}{(1-p)k}\right)\theta_k. \tag{2.29}$$

This proves the statement by induction. Inductively we obtain that

$$\theta_k \geq \prod_{i=1}^{k-1}\left(1 + \frac{p}{(1-p)i}\right)\theta_1. \tag{2.30}$$

If $\theta_1 > 0$, then this product is strictly larger than the sum $\frac{p}{1-p}\theta_1 \sum_{i=1}^{k}\frac{1}{i}$, which diverges as $k \to \infty$. This implies that $\theta_k \to \infty$, contradicting the boundedness of the sequence $(\theta_k)_{k=0}^{\infty}$. Hence, $\theta_1 = 0$ and $d_1 = \hat{d}_1$. But this implies that $d_k = \hat{d}_k$ for all $k \geq 1$. Hence, there is a unique bounded sequence satisfying (2.2) and (2.3). $\qquad\square$

We remark that although d_0 does not have any probabilistic interpretation, it will simplify the analysis later to have defined this.

3. ALMOST SURE CONVERGENCE OF THE DEGREE DISTRIBUTION

We now proceed to prove that we in fact have

$$\lim_{m \to \infty}\frac{D_{m,k}}{N_m} = d_k \tag{3.1}$$

a.s. for all $k \geq 1$. We will need the following auxiliary lemma. In order to apply this lemma, we need to show that $\sum_{k=1}^{\infty} d_k = 1$, the proof of which relies on asymptotic analysis of the sequence $(d_k)_{k=1}^{\infty}$ and will be postponed until Lemma 4.6.

Lemma 3.1. *Suppose we have a countable family $(Y_{m,k})_{k=1}^{\infty}$ of nonnegative random variables indexed by discrete time m, with the property that $\sum_{k=1}^{\infty} Y_{m,k} = 1$ for all m. Let $(b_k)_{k=1}^{\infty}$ be a nonnegative sequence with $\sum_{k=1}^{\infty} b_k = 1$ and such that*

$$\liminf_{m\to\infty} Y_{m,k} \geq b_k \tag{3.2}$$

a.s. Then,

$$\lim_{t\to\infty} Y_{m,k} = b_k \tag{3.3}$$

exists a.s.

Proof. The idea is to prove that $\limsup_{m\to\infty} Y_{m,k} \leq b_k$ for all $k \geq 1$. The following calculation is routine and uses only Fatou's lemma and well-known facts about the limit inferior and limit superior. For $k = 1$ we have that

$$
\begin{aligned}
\limsup_{m\to\infty} Y_{m,1} = \limsup_{m\to\infty} \left(1 - \sum_{k=2}^{\infty} Y_{m,k}\right) &= 1 + \limsup_{m\to\infty} \left(-\sum_{k=2}^{\infty} Y_{m,k}\right) \\
&= 1 - \liminf_{m\to\infty} \sum_{k=2}^{\infty} Y_{m,k} \\
&\leq 1 - \sum_{k=2}^{\infty} \liminf_{m\to\infty} Y_{m,k} \\
&\leq 1 - \sum_{k=2}^{\infty} b_k \\
&= 1 - (1 - b_1) \\
&= b_1.
\end{aligned}
\tag{3.4}
$$

A similar proof works for any $k \geq 1$, so we have that

$$\limsup_{m\to\infty} Y_{m,k} \leq b_k \tag{3.5}$$

a.s. for all $k \geq 1$. This implies that

$$\lim_{m\to\infty} Y_{m,k} = b_k \tag{3.6}$$

a.s. $\square$

Aided by this lemma, we can now show that $\lim_{m\to\infty} \frac{D_{m,k}}{N_m} = d_k$ a.s. We note that the proof of this theorem relies on Lemma 4.6, which we prove later.

Theorem 3.2. *For each $k \geq 1$ we have that*

$$\lim_{m\to\infty} \frac{D_{m,k}}{N_m} = d_k \tag{3.7}$$

a.s., where d_k is the unique bounded positive sequence defined in Theorem 2.1.

Proof. By Theorem 2.1 the sequence $(d_k)_{k=1}^{\infty}$ is unique, bounded, and positive. Furthermore, $\liminf_{m\to\infty} \frac{D_{m,k}}{N_m} \geq d_k$ holds a.s. Since $\sum_{k=1}^{\infty} D_{m,k} = N_m$ for all $m \geq 1$, it holds that $\sum_{k=1}^{\infty} \frac{D_{m,k}}{N_m} = 1$. In Section 4 we will determine the sequence $(d_k)_{k=1}^{\infty}$ explicitly and show that it has the property $\sum_{k=1}^{\infty} d_k = 1$ for all $0 < p < 1$. Invoking Lemma 3.1 with

$Y_{m,k} = \frac{D_{m,k}}{N_m}$, we find that

$$\lim_{m \to \infty} \frac{D_{m,k}}{N_m} = d_k \qquad\qquad (3.8)$$

exists a.s.   □

4. ANALYSIS OF THE PROBABILITY DISTRIBUTION

In this section we analyze the sequence $(d_k)_{k=1}^{\infty}$ and complete the proof of Theorem 3.2 by showing that the sequence sums to 1. We find exact expressions in terms of the hypergeometric function and later derive asymptotics. For the subcritical and supercritical cases, we use Laplace's solution method of recursions with polynomial coefficients. The critical case is covered in [2], so we mention it only briefly below.

4.1. Exact Expressions of the Asymptotic Degree Distribution

4.1.1. Supercritical Case. Suppose that $\frac{1}{2} < p < 1$. Using Laplace's method, we can determine the exact solution to the sequence $(d_k)_{k=0}^{\infty}$ defined in (2.2) and (2.3). For details about Laplace's method, we refer the reader to [12, 14]. It is used in a similar vein in among others [9, 10, 16, 17].

Theorem 4.1. *Let* $\frac{1}{2} < p < 1$. *Then,*

$$d_k = \gamma \int_0^1 \frac{t^k(1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}} dt, \qquad k \geq 0, \qquad\qquad (4.1)$$

where $\beta = \frac{p}{2p-1}$ *and* $\gamma = \frac{1-p}{p}$.

Proof. First we put (2.2) in the form

$$k(1-p)d_{k+1} - (k+p)d_k + kpd_{k-1} = 0, \qquad k \geq 1, \qquad\qquad (4.2)$$

and identify constants Φ_0, Φ_1, Φ_2, Ψ_0, Ψ_1 and Ψ_2, so that it can be written in the form

$$(\Phi_2(k+1) + \Psi_2)\, d_{k+1} + (\Phi_1 k + \Psi_1)\, d_k + (\Phi_0(k-1) + \Psi_0)\, d_{k-1} = 0. \qquad (4.3)$$

Comparing coefficients we see that

$$\begin{cases} \Phi_2 = (1-p) \\ \Phi_1 = -1 \\ \Phi_0 = p \end{cases} \qquad\qquad (4.4)$$

and

$$\begin{cases} \Psi_2 = -(1-p) \\ \Psi_1 = -p \\ \Psi_0 = p. \end{cases} \qquad\qquad (4.5)$$

Now define $\hat{\Phi}(t) = \Phi_2 t^2 + \Phi_1 t + \Phi_0$ and $\hat{\Psi}(t) = \Psi_2 t^2 + \Psi_1 t + \Psi_0$, that is

$$\begin{cases} \hat{\Phi}(t) = (1-p)t^2 - t + p = (t-1)((1-p)t - p), \\ \hat{\Psi}(t) = -(1-p)t^2 - pt + p. \end{cases} \tag{4.6}$$

If we put

$$d_k = \int_{t_0}^{t_1} t^{k-1} v(t)dt, \qquad k \geq 0, \tag{4.7}$$

we need only determine $t_0 < t_1$ and a function $v : [t_0, t_1] \to \mathbb{R}$ such that

$$\frac{v'(t)}{v(t)} = \frac{\hat{\Psi}(t)}{t\hat{\Phi}(t)} = \frac{-(1-p)t^2 - pt + p}{t(t-1)((1-p)t - p)}, \tag{4.8}$$

and

$$\left[t^k v(t)\hat{\Phi}(t) \right]_{t_0}^{t_1} = 0 \tag{4.9}$$

in order to have a solution to the recursion. Define $\gamma = \frac{1-p}{p}$ so that (4.8) can be written in the form

$$\frac{v'(t)}{v(t)} = \frac{\gamma t^2 + t - 1}{t(t-1)(1-\gamma t)}. \tag{4.10}$$

Integrating, we find that

$$\log v(t) = \frac{\gamma \log(1-t) + (1-\gamma)\log(t) + (\gamma - 2)\log(1-\gamma t)}{1-\gamma} + \text{const.} \tag{4.11}$$

Exponentiating and defining $\beta = \frac{p}{2p-1}$, we finally obtain

$$\begin{aligned} v(t) &= Rt(1-t)^{-\frac{\gamma}{\gamma-1}}(1-\gamma t)^{\frac{1}{\gamma-1}-1} \\ &= Rt(1-t)^{\beta-1}(1-\gamma t)^{-(\beta+1)}, \end{aligned} \tag{4.12}$$

where R is some positive constant, to be determined. By (4.6) and (4.12), it follows that

$$t^k v(t)\hat{\Phi}(t) = \text{const} \cdot t^{k+1}(1-t)^\beta (1-\gamma t)^{-\beta}. \tag{4.13}$$

For $\frac{1}{2} < p < 1$ we have that $\beta > 0$ and $\gamma < 1$, so the function $t^k v(t)\hat{\Phi}(t)$ is continuous on $[0, 1]$ and zero at the endpoints. Choosing $t_0 = 0$ and $t_1 = 1$, we ensure that (4.9) is satisfied. Note also that the singularities of the right-hand side in (4.10) occur at $t = 0, t = 1$, and $t = \gamma^{-1} > 1$.

We thus obtain

$$d_k = \int_0^1 t^{k-1} v(t)dt = R \int_0^1 \frac{t^k(1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}}dt. \tag{4.14}$$

The integral in (4.14) is a hypergeometric integral [1], and can be expressed in terms of the hypergeometric function $_2F_1$. To be precise, we have that

$$d_k = R\frac{\Gamma(k+1)\Gamma(\beta)}{\Gamma(\beta+k+1)} \, _2F_1(\beta+1, k+1, \beta+k+1; \gamma). \tag{4.15}$$

We now determine the constant R. Recall the Pochhammer symbol $(x)_k = \prod_{i=0}^{k-1}(x + i)$. Using a standard series expansion of the hypergeometric function [1], we find for $k = 0$ that

$$
\begin{aligned}
d_0 &= R \frac{\Gamma(1)\Gamma(\beta)}{\Gamma(\beta + 1)} \, {}_2F_1(\beta + 1, 1, \beta + 1; \gamma) \\
&= R\beta^{-1} \sum_{n=0}^{\infty} \frac{(\beta + 1)_n (1)_n}{(\beta + 1)_n n!} \gamma^n \\
&= R\beta^{-1} \sum_{n=0}^{\infty} \gamma^n \\
&= R\beta^{-1}(1 - \gamma)^{-1} \\
&= R.
\end{aligned}
\tag{4.16}
$$

Since $d_0 = \gamma$, we find that $R = \gamma$. An alternative approach to evaluating R is to compute the integral for d_0 directly.

Having determined R, we know that

$$
d_k = \gamma \int_0^1 \frac{t^k (1 - t)^{\beta - 1}}{(1 - \gamma t)^{\beta + 1}} \, dt
\tag{4.17}
$$

for all $k \geq 0$. $\square$

4.1.2. Subcritical Case.

Theorem 4.2. *Let* $0 < p < \frac{1}{2}$. *Then,*

$$
d_k = \gamma^{-k} \int_0^1 \frac{t^k (1 - t)^{-1 - \beta}}{(1 - \gamma^{-1} t)^{1 - \beta}} \, dt
\tag{4.18}
$$

where $\beta = \frac{p}{2p - 1}$ *and* $\gamma = \frac{1 - p}{p}$.

Proof. The subcritical case is similar to the supercritical case, so we give less detail here. Indeed, we can follow the analysis for the supercritical case up until (4.13). When $0 < p < \frac{1}{2}$, we have that $\beta < 0$ and $\gamma > 1$, so we need to make the choices $t_0 = 0$ and $t_1 = \gamma^{-1}$, instead. This implies that

$$
d_k = R \int_0^{\gamma^{-1}} \frac{s^k (1 - s)^{\beta - 1}}{(1 - \gamma s)^{\beta + 1}} \, ds = R\gamma^{-k-1} \int_0^1 \frac{t^k (1 - t)^{-1 - \beta}}{(1 - \gamma^{-1} t)^{1 - \beta}} \, dt,
\tag{4.19}
$$

where we used the change of variables $t = \gamma s$. Again, using the fact that $d_0 = \gamma$ and the interpretation of this integral as the hypergeometric function, we find that $R = \gamma$. Hence,

$$
d_k = \gamma^{-k} \int_0^1 \frac{t^k (1 - t)^{-1 - \beta}}{(1 - \gamma^{-1} t)^{1 - \beta}} \, dt.
\tag{4.20}
$$

$\square$

4.1.3. Critical Case. In the critical case $p = \frac{1}{2}$, we need to analyze the sequence defined by

$$d_0 = 1, \tag{4.21}$$

$$d_k = \frac{kd_{k-1} + kd_{k+1}}{2k + 1}, \qquad k \geq 1. \tag{4.22}$$

As noted in the introduction, in [2], a dynamic random graph model very similar to ours was introduced, with the difference that every other step was a duplication step and the alternate step was a deletion step. For $p = \frac{1}{2}$, this is in some sense true, on average, in our model. Therefore, it is not surprising that the sequence given by (4.21) and (4.22) appears as the asymptotic degree distribution in [2], as well, wherein was found the exact solution

$$d_k = k \int_0^\infty \frac{t^{k-1}e^{-t}}{(1+t)^{k+1}} dt \tag{4.23}$$

and the asymptotic result

$$d_k \sim (e\pi)^{1/2} k^{1/4} e^{-2\sqrt{k}}, \tag{4.24}$$

where $x_k \sim y_k$ means that $\frac{x_k}{y_k} \to 1$ as $k \to \infty$.
This expression decreases slower than any exponential function and decreases faster than any polynomial, which demonstrates the fact that the asymptotic degree distribution in the critical case $p = \frac{1}{2}$ lies between a power law and exponential decay.

It is worth noting that the integral in (4.23) appears as the limiting case as $p \to \frac{1}{2}$ in Theorem 4.1 and Theorem 4.2; that is,

$$\lim_{p \downarrow \frac{1}{2}} \gamma \int_0^1 \frac{t^k(1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}} dt = k \int_0^\infty \frac{t^{k-1}e^{-t}}{(1+t)^{k+1}} dt$$

$$= \lim_{p \uparrow \frac{1}{2}} \gamma^{-k} \int_0^1 \frac{t^k(1-t)^{-1-\beta}}{(1-\gamma^{-1}t)^{1-\beta}} dt. \tag{4.25}$$

This is justified by noting that the left most and right most integrands both can be dominated by an integrable function, so the dominated convergence theorem allows us to interchange the limit and integral signs. The integrands both converge pointwise to $\frac{t^k}{(1-t)^2} e^{-\frac{t}{1-t}}$ as $p \to \frac{1}{2}$. The middle integral then follows by a change of variables.

4.2. Asymptotic Results

In this section, we analyze the exact expressions and show asymptotic results as $k \to \infty$. We leave out the critical case and refer the reader to (4.24).

4.2.1. Supercritical Case.

Theorem 4.3. *For $\frac{1}{2} < p < 1$, asymptotically as $k \to \infty$ the degree sequence $(d_k)_{k=1}^\infty$ satisfies*

$$d_k \sim \gamma\beta^\beta \Gamma(\beta + 1)k^{-\beta}, \tag{4.26}$$

where $\beta = \frac{p}{2p-1}$ and $\gamma = \frac{1-p}{p}$.

Proof. Recall first the beta integral

$$\int_0^1 t^k (1-t)^{\beta-1} dt = \frac{\Gamma(k+1)\Gamma(\beta)}{\Gamma(k+\beta+1)}. \tag{4.27}$$

We bound the integral in Theorem 4.1 from above and below. Estimating from above, we find that

$$\gamma \int_0^1 \frac{t^k(1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}} dt \leq \frac{\gamma}{(1-\gamma)^{\beta+1}} \int_0^1 t^k(1-t)^{\beta-1} dt$$

$$= \frac{\gamma}{(1-\gamma)^{\beta+1}} \frac{\Gamma(k+1)\Gamma(\beta)}{\Gamma(k+\beta+1)}$$

$$\sim \frac{\gamma}{(1-\gamma)^{\beta+1}} \Gamma(\beta) k^{-\beta}. \tag{4.28}$$

In order to simplify notation, define

$$A_k = \frac{\gamma}{\left(1 - \gamma\left(1 - 1/\sqrt{k}\right)\right)^{\beta+1}}, \tag{4.29}$$

and note that $A_k \to \frac{\gamma}{(1-\gamma)^{\beta+1}}$ as $k \to \infty$. We obtain the lower bound

$$\gamma \int_0^1 \frac{t^k(1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}} dt \geq \gamma \int_{1-1/\sqrt{k}}^1 \frac{t^k(1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}} dt$$

$$\geq A_k \int_{1-1/\sqrt{k}}^1 t^k(1-t)^{\beta-1} dt$$

$$= A_k \left[\int_0^1 t^k(1-t)^{\beta-1} dt - \int_0^{1-1/\sqrt{k}} t^k(1-t)^{\beta-1} dt \right]$$

$$= A_k \left[\frac{\Gamma(k+1)\Gamma(\beta)}{\Gamma(k+1+\beta)} - O\left(\int_0^{1-1/\sqrt{k}} t^k dt \right) \right]$$

$$= A_k \left[\frac{\Gamma(k+1)\Gamma(\beta)}{\Gamma(k+1+\beta)} - O\left(e^{-\sqrt{k}} \right) \right]$$

$$\sim \frac{\gamma}{(1-\gamma)^{\beta+1}} \Gamma(\beta) k^{-\beta}. \tag{4.30}$$

The theorem is proved once we note that

$$\frac{\gamma}{(1-\gamma)^{\beta+1}} \Gamma(\beta) k^{-\beta} = \gamma \beta^\beta \Gamma(\beta+1) k^{-\beta}. \tag{4.31}$$

$\square$

Alternatively, one could use a series expansion of the $\frac{1}{(1-\gamma t)^{\beta+1}}$–term and then apply Stirling's formula to achieve the same result. Estimations like these were done in [9].

Since $\beta = \frac{p}{2p-1} \in (1, \infty)$ when $\frac{1}{2} < p < 1$, the supercritical case gives rise to a power law with exponent greater than 1.

4.2.2. Subcritical Case. The proof of the following theorem is similar to Theorem 4.3, and the proof is omitted.

Theorem 4.4. *For $0 < p < \frac{1}{2}$, asymptotically as $k \to \infty$, we have that*

$$d_k \sim (-\beta)^{-1}(1 - \beta)^{1-\beta}\Gamma(1 - \beta)\gamma^{-k}k^{\beta}, \tag{4.32}$$

where $\beta = \frac{p}{2p-1}$ and $\gamma = \frac{1-p}{p}$.

Remark 4.5. If $0 < p < \frac{1}{2}$, we have that $\gamma > 1$, so the factor of γ^{-k} forces most vertices to have very low degree. This is what one would expect, because, on average, there are more deletion steps (which reduce the size of a clique and create new isolated vertices) than duplication steps (which increase the size of a clique).

4.3. Proof that $(d_k)_{k=1}^{\infty}$ Defines a Probability Distribution

Finally, we prove that the sequence $(d_k)_{k=1}^{\infty}$ defines a probability distribution, which is the final step in the proof of Theorem 3.2. The following proof relies on Theorems 4.3 and 4.4, the proofs of which do not rely on Theorem 3.2.

Lemma 4.6. *Let $0 < p < 1$. The unique bounded sequence $(d_k)_{k=1}^{\infty}$ given in Theorem 2.1 as*

$$d_0 = \frac{1 - p}{p}, \tag{4.33}$$

$$d_k = \frac{pkd_{k-1} + (1 - p)kd_{k+1}}{k + p}, \quad k \geq 1 \tag{4.34}$$

defines a probability distribution, i.e.,

$$\sum_{k=1}^{\infty} d_k = 1. \tag{4.35}$$

Proof. Summing over $k = 1, \ldots n$ we find that

$$
\begin{aligned}
\sum_{k=1}^{n}(k + p)d_k &= (1 - p)(1 + d_2) + \sum_{k=2}^{n} k(pd_{k-1} + (1 - p)d_{k+1}) \\
&= (1 - p)(1 + d_2) + p\sum_{k=1}^{n}(k + 1)d_k + (1 - p)\sum_{k=1}^{n}(k - 1)d_k \\
&\quad -(1 - p)d_2 - p(n + 1)d_n + (1 - p)nd_{n+1} \\
&= (1 - p) + \sum_{k=1}^{n} kd_k + (2p - 1)\sum_{k=1}^{n} d_k - p(n + 1)d_n + (1 - p)nd_{n+1}.
\end{aligned}
\tag{4.36}
$$

Hence,

$$\sum_{k=1}^{n} d_k = 1 + \frac{1}{1-p} \left(-p(n+1)d_n + (1-p)nd_{n+1} \right). \tag{4.37}$$

The asymptotics in (4.24), (4.26), and (4.32) imply that $d_n = o(n^{-1})$ for any $0 < p < 1$. Hence, the sum in (4.32) tends to 1 as $n \to \infty$. $\qquad\square$

Remark 4.7. The above proof shows that any sequence satisfying (4.33) and (4.34) will sum to 1 under the assumption that $d_n = o(n^{-1})$. Alternatively, one could prove this directly from the exact expressions. For instance, in the supercritical case, it is not difficult to show that

$$\sum_{k=1}^{\infty} d_k = \sum_{k=0}^{\infty} d_k - d_0 = \sum_{k=0}^{\infty} \gamma \int_0^1 \frac{t^k (1-t)^{\beta-1}}{(1-\gamma t)^{\beta+1}} dt - \gamma = 1 \tag{4.38}$$

by interchanging the summation sign and the integral sign, and using the geometric series.

ACKNOWLEDGMENTS

REFERENCES

[1] G. E. Andrews, R. Askey, and R. Roy. *Special Functions*. Cambridge, New York: Cambridge University Press, 1999.

[2] Á. Backhausz and T. Móri. "Asymptotic Properties of a Random Graph with Duplications." Available at http://arXiv:1308.1506v2. To appear in *Journal of Applied Probability*, 2015.

[3] Á. Backhausz and T. Móri. "A Random Model of Publication Activity." *Discrete Appl. Math.* 162 (2014), 78–89.

[4] A.-L. Barabási and R. Albert. "Emergence of Scaling in Random Networks." *Science* 286:5439 (1999), 509–512.

[5] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. "The Degree Sequence of a Scale-Free Random Graph Process." *Random Structures and Algorithms* 18 (2001), 279–290.

[6] N. Champagnat and A. Lambert. "Splitting Trees with Neutral Poissonian Mutations in Small Families." *Stochastic Processes and their Applications* 122:3 (2012), 1003–1033.

[7] N. Champagnat, A. Lambert, and M. Richard. "Birth and Death Processes with Neutral Mutations." *International Journal of Stochastic Analysis* 2012(2012), 20 pp.

[8] F. Chung and L. Lu. "Coupling Online and Offline Analyses for Random Power Law Graphs." *Internet Math.* 1:4 (2004), 409–461.

[9] C. Cooper, A. Frieze, and J. Vera. "Random Deletion in a Scale Free Random Graph Process." *Internet Math.* 124:4 (2004), 463–483.

[10] M. Deijfen and M. Lindholm. "Growing Networks with Preferential Deletion and Addition of Edges." *Physica A: Statistical Mechanics and its Applications* 388:19 (2009), 4297–4303.

[11] S. Dereich and M. Ortgiese. "Robust Analysis of Preferential Attachment Models with Fitness." *Combinatorics, Probability and Computing* 23:03 (2014), 386–411.

[12] C. Jordan. *Calculus of Finite Differences*. Budapest: Rottig and Romwalter, 1939.

[13] T. Móri. "On Random Trees." *Studia Sci. Math. Hungar.* 39:1–2 (2002), 143–155.

[14] N.-E. Nörlund. *Differenzenrechnung*. Berlin, Heidelberg: Springer-Verlag, 1924.

[15] A. G. Pakes. "An Infinite Alleles Version of the Markov Branching Process." *J. Australian. Math. Soc.* 46 (1989), 146–170.

[16] P. Pralat and C.Wang. "An Edge Deletion Model for Complex Networks." *Theoretical Computer Science* 412:39 (2011), 5111–5120.

[17] T. Vallier. "Transition of the Degree Sequence in the Random Graph Model of Cooper, Frieze and Vera." *Stochastic Models* 29:3 (2013), 341–352.

[18] X.-Y. Wu, Z. Dong, K. Liu, and K.-Y. Cai. "On the Degree Sequence of an Evolving Random Graph Process and Its Critical Phenomenon." *Journal of Applied Probability* 46:4 (2009), 1213–1220.