Towards Degree Distribution of a Duplication-Divergence Graph Model

Krzysztof Turowski
Theoretical Computer Science Department
Jagiellonian University
Kraków, Poland
krzysztof.szymon.turowski@gmail.com

Wojciech Szpankowski*
Center for Science of Information
Purdue University
West Lafayette, IN, U.S.A.
szpan@purdue.edu

Submitted: Dec 31, 2019; Accepted: Jan 13, 2021; Published: Jan 29, 2021

Abstract

We present a rigorous and precise analysis of degree distribution in a dynamic graph model introduced by Solé, Pastor-Satorras et al. in which nodes are added according to a duplication-divergence mechanism. This model is discussed in numerous publications with only very few recent rigorous results, especially for the degree distribution. In this paper we focus on two related problems: the expected value and variance of the degree of a given node over the evolution of the graph and the expected value and variance of the average degree over all nodes. We present exact and precise asymptotic results showing that both quantities may decrease or increase over time depending on the model parameters. Our findings are a step towards a better understanding of the graph behaviors such as degree distributions, symmetry, power law, and structural compression.

Mathematics Subject Classifications: 05C07, 05C80, 05C82

1 Introduction

Many real-world networks, such as protein-protein and citation networks, are widely viewed as driven by an internal evolution mechanism based on duplication and mutation [19]. New nodes are added to the network as copies of existing nodes together with some random divergence, resulting in differences among the original nodes and their

*This work was partially supported by the NSF Center for Science of Information (CSoI) Grant CCF-0939370, and in addition by NSF Grants CCF-1524312, CCF-2006440, CCF-2007238, Polish National Science Center 2018/31/B/ST6/01294 and Google Research Award.

copies. It has been claimed that graphs generated from these models exhibit many properties characteristic to real-world networks such as power-law degree distribution, the large clustering coefficient, and a large amount of symmetry [3, 11]. However, some of these results turned out not to be correct (e.g., power-law degree distribution was disproved in [6]) or not proved rigorously. In this paper we focus on presenting exact and precise asymptotic results for the expected degree of a given node over time and the average degree in the graph. We show that these two quantities exhibit phase transitions over the parameter space.

The widest known duplication divergence model was introduced by Solé, Pastor-Satorras et al. [14], denoted here as $\mathtt{DD}(t, p, r)$. It is defined as follows: starting from a given graph G_{t_0} on t_0 vertices (labeled from 1 to t_0) we repeat the following procedure until we get a graph on t vertices: (i) *Duplication*: Select a node u from a current graph G (on k vertices) uniformly at random. Add node v (with label $k + 1$) to the graph and add edges between v and all neighbors of u; (ii) *Divergence*: connections from v are randomly retained with probability p (otherwise they are deleted). Furthermore, for all nodes w not adjacent to u we add an edge between v and w, independently at random with probability r/k. Note that nodes in the graph are labeled by the numbers from 1 to t, according to their order of appearance in the graph.

This model is a generalization of the *pure duplication model*, where no edges are added in the divergence step ($r = 0$). It is also similar to the Chung-Lu model, in which the duplication step and the first part of the divergence step (retaining p-edges) is exactly the same as in our model, however instead of adding r-edges we only add a single edge between v and u with some fixed probability q [3].

Additionally note that the case $p = 0$ of our model somewhat resembles Erdős-Renyí graphs in the sense that all pairs of vertices correspond to independent Bernoulli variables. However, their respective probabilities are different, as at k-th step we add a vertex with $k - 1$ possible incident edges with probabilities $Ber(r/k)$.

It has been shown that graphs generated by this model for a set of parameters fit very well into the structure of some real-world networks (e.g., protein-protein and citation networks) in terms of the degree distribution [4] and small subgraphs (graphlets) count [13]. It was also shown that this model may exhibit a large amount of symmetry (measured by the number of automorphisms) [15], and this distinguishes it from other graph models such as Erdős-Renyi and preferential attachment [10]. We formally showed in [17] that for the special case of $p = 1, r = 0$ the expected logarithm of the number of automorphisms for graphs on t vertices is asymptotically $\Theta(t \log t)$, which indicates a lot of symmetry. However, the extension of it for all p and r is a difficult open problem.

The most interesting open problem in the duplication-divergence model $\mathtt{DD}(t, p, r)$ is the quest to uncover the behavior of the degree distribution, that is, the number of nodes of a given degree. For $r = 0$ it was recently proved by Hermann and Pfaffelhuber in [6] that depending on a value of p either there exists a limiting distribution with almost all vertices isolated or there is no limiting distribution as $t \to \infty$. It should be mentioned that still the most interesting problem, namely the rate of convergence is open. They also asserted (but without proof) that this holds also for $r > 0$. Our findings in this paper indicate that this is not the case as the size of the graph grows to infinity. Moreover, it is shown in [9] that the number of vertices of degree one is $\Omega(\ln t)$ but again the precise rates of growth of the number of vertices with degrees $k > 0$ are yet unknown. Recently, also for $r = 0$, Jordan [8] showed that the power-law of the degree distribution exists for the *connected component* for $p < e^{-1}$. In this case the exponent is equal to γ which is the solution of $3 = \gamma + p^{\gamma-2}$. The case $p \geqslant e^{-1}$ still remains an open problem.

In this paper we approach the problem of the degree distribution from a different perspective. We investigate the behavior of two closely related variables: the degree of a vertex s in G_t denoted $\deg_t(s)$ and the average degree $D(G_t)$ in G_t. We present in Theorems 5–17 exact and precise asymptotics of the expected values and variances of the parameters under investigation when $t \to \infty$. We show that all parameters exhibit phase transition as a function of p. In particular, we find that $\mathbb{E}[\deg_t(s)]$ grows respectively like $\left(\frac{t}{s}\right)^p$, $\sqrt{\frac{t}{s}}\log s$ or $\left(\frac{t}{s}\right)^p s^{2p-1}$, depending whether $p < 1/2$, $p = 1/2$ or $p > 1/2$. Furthermore, $\mathbb{E}[D(G_t)]$ is either $\Theta(1)$, $\Theta(\log t)$ or $\Theta(t^{2p-1})$ for the same ranges of p. We also determine the exact constants for the leading terms that strictly depend on p, r, t_0 and the structure of the seed graph G_{t_0}. This confirms the empirical findings of [7] regarding the seed graph influence on the structure of G_t.

These findings allow us to better understand why the $\mathtt{DD}(t, p, r)$ model differs quite substantially from other graph models such as preferential attachment model [18]. In particular, we observe that the expected degree behaves differently as $t \to \infty$ for various values of s and p. For example, if $p > 1/2$, then for $s = O(1)$ (that is, for very old nodes) we observe that $\mathbb{E}[\deg_t(s)] = O(t^p)$ while for $s = \Theta(t)$ (i.e., very young nodes) we have $\mathbb{E}[\deg_t(s)] = O(t^{2p-1})$. This behavior is very different than the degree distribution for, say, preferential attachment model, for which the expected degree of a vertex s in a graph on t vertices is of order $\sqrt{t/s}$ [10] and may lead to our better understanding of some graph behaviors such as degree distribution, existence of power-law phenomenon, symmetry, and structural compression.

2 Main results

In this section we present our main results with proofs and auxiliary lemmas delayed until the next section.

We use the standard graph notation, e.g. from [5]: $V(G)$ denotes the set of vertices of graph G, $\mathcal{N}_G(u)$ – the set of neighbors of vertex u in G, $\deg_G(u) = |\mathcal{N}_G(u)|$ – the degree of u in G. For brevity we use the abbreviations for G_t, e.g. $\deg_t(u)$ instead of $\deg_{G_t}(u)$. All graphs are simple. Let us also introduce the *average degree* $D(G_t)$ of G_t as

$$D(G) = \frac{1}{|V(G)|} \sum_{v \in V(G)} \deg_G(u)$$

, and *average degree squared* $D(G_t)$ of G_t as

$$D_2(G) = \frac{1}{|V(G)|} \sum_{v \in V(G)} \deg^2_G(u)$$

. They are also known in the literature as the first and second moment of the degree distribution, respectively.

Formally, we define the $\texttt{DD}(t, p, r)$ model as follows: let $0 \leqslant p \leqslant 1$ and $0 \leqslant r \leqslant t_0$ be the parameters of the model. Let also G_{t_0} be a graph on t_0 vertices, with $V(G_{t_0}) = \{1, \ldots, t_0\}$. Now, for every $t = t_0, t_0 + 1, \ldots$ we create G_{t+1} from G_t according to the following rules:

1. add a new vertex $t + 1$ to the graph,
2. pick vertex u from $V(G_t) = \{1, \ldots, t\}$ uniformly at random – and denote u as $\mathit{parent}(t + 1)$,
3. for every vertex $i \in V(G_t)$:
   (a) if $i \in \mathcal{N}_t(\mathit{parent}(t + 1))$, then add an edge between i and $t + 1$ with probability p,
   (b) if $i \notin \mathcal{N}_t(\mathit{parent}(t + 1))$, then add an edge between i and $t + 1$ with probability $\frac{r}{t}$.

2.1 Expected value

First, we derive the exact and the asymptotic expression for the expected values of the average degree, the degree of the last node, and the degree of a given node. We start our considerations by deriving a recurrence expression for $\mathbb{E}[\deg_t(s)]$ from the definition of the model.

Lemma 1. *For all $t \geqslant t_0$ it holds that*

$$\mathbb{E}[\deg_{t+1}(s)] = \mathbb{E}[\deg_t(s)] \left(1 + \frac{p}{t} - \frac{r}{t^2}\right) + \frac{r}{t}.$$

Note that in order to solve this recurrence we need to know the behavior of $\mathbb{E}[\deg_s(s)]$ for $s \geqslant t_0$.

Next, we establish a relation between the expected degree of the newest vertex and the expected average degree of the previous graph.

Lemma 2. *For any $t \geqslant t_0$ it holds that*

$$\mathbb{E}[\deg_{t+1}(t + 1)] = \left(p - \frac{r}{t}\right) \mathbb{E}[D(G_t)] + r.$$

Now, we use the model definition again to find the following recurrence for $\mathbb{E}[D(G_t)]$:

Lemma 3. *For any $t \geqslant t_0$ it is true that*

$$\mathbb{E}[D(G_{t+1})] = \mathbb{E}[D(G_t)] \left(1 + \frac{2p - 1}{t + 1} - \frac{2r}{t(t + 1)}\right) + \frac{2r}{t + 1}.$$

This leads us to the exact and asymptotic expressions for $\mathbb{E}[D(G_t)]$.

Lemma 4. *For all $t \geqslant t_0$ we have*

$$\mathbb{E}[D(G_t)] = \frac{\Gamma(t+c_3)\Gamma(t+c_4)}{\Gamma(t)\Gamma(t+1)}$$
$$\left(D(G_{t_0})\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} + 2r\sum_{j=t_0}^{t-1}\frac{\Gamma(j+1)^2}{\Gamma(j+c_3+1)\Gamma(j+c_4+1)}\right),$$

where $c_3 = p + \sqrt{p^2+2r}$, $c_4 = p - \sqrt{p^2+2r}$, and $\Gamma(z)$ is the Euler gamma function.

Theorem 5. *Asymptotically as $t \to \infty$ we find*

$$\mathbb{E}[D(G_t)] = \begin{cases} t^{2p-1}\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}D(G_{t_0})(1+o(1)) & \text{if } p \leqslant \frac{1}{2},\ r=0,\\ \frac{2r}{1-2p}(1+o(1)) & \text{if } p < \frac{1}{2},\ r>0,\\ 2r\ln t\,(1+o(1)) & \text{if } p=\frac{1}{2},\ r>0,\\ t^{2p-1}\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}(1+o(1)) & \text{if } p>\frac{1}{2},\\ \quad\left(D(G_{t_0}) + \frac{2rt_0}{t_0^2+2pt_0-2r}\,{}_3F_2\left[{t_0+1,t_0+1,1 \atop t_0+c_3+1,t_0+c_4+1};1\right]\right) & \end{cases}$$

where $D(G_{t_0})$ is the average degree of the initial graph G_{t_0} and

$${}_3F_2\left[{a_1,a_2,a_3 \atop b_1,b_2};z\right] = \sum_{l=0}^{\infty}\frac{(a_1)_l(a_2)_l(a_3)_l}{(b_1)_l(b_2)_l}\frac{z^l}{l!}$$

is the generalized hypergeometric function with $(a)_l = a(a+1)\dots(a+l-1)$, $(a)_0 = 1$ the rising factorial (see [1] for details).

Combining Lemmas 2 and 4 we obtain the following result:

Lemma 6. *For all $t > t_0$ it is true that*

$$\mathbb{E}[\deg_t(t)] = (pt-p-r)\,\frac{\Gamma(t+c_3-1)\Gamma(t+c_4-1)}{\Gamma(t)^2}$$
$$\left(D(G_{t_0})\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} + 2r\sum_{j=t_0}^{t-2}\frac{\Gamma(j+1)^2}{\Gamma(j+c_3+1)\Gamma(j+c_4+1)}\right) + r,$$

where c_3, c_4 are as above.

Similarly from Lemma 4 and Theorem 5 we get the final formula for $\deg_t(t)$.

Theorem 7. *Asymptotically as $t \to \infty$ it holds that*

$$\mathbb{E}[\deg_t(t)] = \begin{cases} pt^{2p-1}\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}D(G_{t_0})(1+o(1)) & \text{if } p \leqslant \frac{1}{2},\ r=0,\\ \frac{r}{1-2p}(1+o(1)) & \text{if } p < \frac{1}{2},\ r>0,\\ 2rp\ln t\,(1+o(1)) & \text{if } p=\frac{1}{2},\ r>0,\\ pt^{2p-1}\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}(1+o(1)) & \text{if } p>\frac{1}{2},\\ \quad\left(D(G_{t_0}) + \frac{2rt_0}{t_0^2+2pt_0-2r}\,{}_3F_2\left[{t_0+1,t_0+1,1 \atop t_0+c_3+1,t_0+c_4+1};1\right]\right) & \end{cases}$$

with the same notation as in Theorem 5.

Now we are in the position to state the exact and asymptotic expressions for $\mathbb{E}[\deg_t(s)]$.

Lemma 8. *For all $t > s > t_0$ it is true that*

$$\begin{aligned}\mathbb{E}[\deg_t(s)] = &\frac{\Gamma(t+c_1)\Gamma(t+c_2)}{\Gamma(t)^2}\\ &\Bigg[(ps-p-r)\,\frac{\Gamma(s+c_3-1)\Gamma(s+c_4-1)}{\Gamma(s+c_1)\Gamma(s+c_2)}\\ &\quad\left(D(G_{t_0})\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} + 2r\sum_{j=t_0}^{s-2}\frac{\Gamma(j+1)^2}{\Gamma(j+c_3+1)\Gamma(j+c_4+1)}\right)\\ &\quad+\frac{r\Gamma(s)^2}{\Gamma(s+c_1)\Gamma(s+c_2)} + r\sum_{j=s}^{t-1}\frac{\Gamma(j)\Gamma(j+1)}{\Gamma(j+c_1+1)\Gamma(j+c_2+1)}\Bigg],\end{aligned}$$

where $c_1 = \frac{p+\sqrt{p^2+4r}}{2}$, $c_2 = \frac{p-\sqrt{p^2+4r}}{2}$, c_3 and c_4 as above.

Theorem 9. *Asymptotically as $t \to \infty$ it holds that:*
(i) *for $s = O(1)$*

$$\begin{aligned}\mathbb{E}[\deg_t(s)] = &\,t^p(1+o(1))\\ &\Bigg[(ps-p-r)\,\frac{\Gamma(s+c_3-1)\Gamma(s+c_4-1)}{\Gamma(s+c_1)\Gamma(s+c_2)}\\ &\quad\left(D(G_{t_0})\frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} + 2r\sum_{j=t_0}^{s-2}\frac{\Gamma(j+1)^2}{\Gamma(j+c_3+1)\Gamma(j+c_4+1)}\right)\\ &\quad+\frac{r\Gamma(s)^2}{\Gamma(s+c_1)\Gamma(s+c_2)}\left(1+{}_3F_2\left[\genfrac{}{}{0pt}{}{s,s+1,1}{s+c_1+1,s+c_2+1};1\right]\frac{s}{s^2+ps-r}\right)\Bigg].\end{aligned}$$

(ii) *for $s = \omega(1)$ and $s = o(t)$*

$$\mathbb{E}[\deg_t(s)] = \begin{cases} D(G_{t_0})\frac{p\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}\left(\frac{t}{s}\right)^p s^{2p-1}(1+o(1)) & \textit{if } p \leqslant \frac{1}{2},\ r=0,\\ r\log\left(\frac{t}{s}\right)(1+o(1)) & \textit{if } p=0,\ r>0,\\ \frac{r(1-p)}{p(1-2p)}\left(\frac{t}{s}\right)^p(1+o(1)) & \textit{if } 0<p<\frac{1}{2},\ r>0,\\ r\sqrt{\frac{t}{s}}\log s\,(1+o(1)) & \textit{if } p=\frac{1}{2},\ r>0,\\ \left(D(G_{t_0})+\frac{2rt_0}{t_0^2+2pt_0-2r}\,{}_3F_2\left[\genfrac{}{}{0pt}{}{t_0+1,t_0+1,1}{t_0+c_3+1,t_0+c_4+1};1\right]\right) & \\ \quad\frac{p\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}\left(\frac{t}{s}\right)^p s^{2p-1}(1+o(1)) & \textit{if } p>\frac{1}{2}. \end{cases}$$

(iii) *for $s = ct - o(t)$, $0 < c \leqslant 1$,*

$$\mathbb{E}[\deg_t(s)] = \begin{cases} D(G_{t_0})\frac{p\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}t^{2p-1}c^{p-1}(1+o(1)) & \textit{if } p \leqslant \frac{1}{2},\ r=0, \\ r\,(1-\log c)\,(1+o(1)) & \textit{if } p=0,\ r>0, \\ \left(\frac{r(1-p)}{p(1-2p)c^p} - \frac{r}{p}\right)(1+o(1)) & \textit{if } 0<p<\frac{1}{2},\ r>0, \\ \frac{r}{\sqrt{c}}\log t\,(1+o(1)) & \textit{if } p=\frac{1}{2},\ r>0, \\ \left(D(G_{t_0}) + \frac{2rt_0}{t_0^2+2pt_0-2r}\,{}_3F_2\left[{t_0+1,t_0+1,1 \atop t_0+c_3+1,t_0+c_4+1};1\right]\right) & \\ \qquad \frac{p\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}t^{2p-1}c^{p-1}(1+o(1)) & \textit{if } p>\frac{1}{2}. \end{cases}$$

Thus we have presented a complete characterization of the expected average degree and the expected degree of a vertex s at time t.

2.2 Variance

The procedure described above can be extended to find also the second moment of the degree distribution.

First, the analogous reasoning as in the previous subsection leads to a recurrence

Lemma 10. *For all $t \geqslant s$ it holds that*

$$\mathbb{E}[\deg^2_{t+1}(s)] = \mathbb{E}[\deg^2_t(s)]\left(1 + \frac{2p}{t} - \frac{2r}{t^2}\right) + \mathbb{E}[\deg_t(s)]\left(\frac{p+2r}{t} - \frac{r}{t^2}\right) + \frac{r}{t}.$$

To solve this recurrence we need to know the behavior of $\mathbb{E}[\deg^2_s(s)]$ for all $s \geqslant t_0$. To do this we will need the following lemma connecting $\mathbb{E}[\deg^2_t(t)]$ and the first two moments of the degree distribution.

Lemma 11. *For any $t \geqslant t_0$ it holds that*

$$\begin{aligned}\mathbb{E}[\deg^2_{t+1}(t+1)] = {} & \left(p^2 - \frac{2pr}{t} + \frac{r^2}{t^2}\right)\mathbb{E}[D_2(G_t)] \\ & + \left(p - p^2 + 2pr - \frac{r+2r^2}{t} + \frac{r^2}{t^2}\right)\mathbb{E}[D(G_t)] + r^2 + r - \frac{r^2}{t}.\end{aligned}$$

Similarly as before, here we need to find $\mathbb{E}[D_2(G_t)]$, that is, the average degree squared of G_t. We find the following recurrence:

Lemma 12. *For all $t \geqslant t_0$ it is true that*

$$\begin{aligned}\mathbb{E}[D_2(G_{t+1})] = {} & \mathbb{E}[D_2(G_t)]\left(1 + \frac{2p+p^2-1}{t+1} - \frac{2r(1+p)}{t(t+1)} + \frac{r^2}{t^2(t+1)}\right) \\ & + \mathbb{E}[D(G_t)]\left(\frac{2p-p^2+2pr+2r}{t+1} - \frac{2r+2r^2}{t(t+1)} + \frac{r^2}{t^2(t+1)}\right) \\ & + \frac{2r^2+2r}{t+1} - \frac{r^2}{t(t+1)}.\end{aligned}$$

Here $\mathbb{E}[D_2(G_{t+1})]$ should not be confused with $\mathbb{E}[D^2(G_{t+1})]$, however the latter one may be derived in similar fashion:

Lemma 13. *For all $t \geqslant t_0$ it is true that*

$$\begin{aligned}
&\mathbb{E}[D^2(G_{t+1}) \mid G_t] = \\
&\quad D^2(G_t)\frac{t^2+4tp-4r}{(t+1)^2} + D_2(G_t)\frac{4}{(t+1)^2}\left(p^2 - \frac{2pr}{t} + \frac{r^2}{t^2}\right) \\
&\quad + D(G_t)\frac{4}{(t+1)^2}\left(tr + p - p^2 + 2pr - \frac{r+2r^2}{t} + \frac{r^2}{t^2}\right) \\
&\quad + \frac{4}{(t+1)^2}\left(r^2 + r - \frac{r^2}{t}\right).
\end{aligned}$$

Finally, we may obtain both exact and asymptotic formulas for the variances of investigated random variables. However, due to the complicated form of the constants in exact formulas, we drop the exact formulas from the paper and present only the asymptotic rates of growth.

Theorem 14. *The following holds*

$$\mathbb{E}[D_2(G_t)] = \begin{cases} \Theta(1) & \text{if } p < \sqrt{2}-1, \\ \Theta(\log t) & \text{if } p = \sqrt{2}-1, \\ \Theta(t^{p^2+2p-1}) & \text{if } p > \sqrt{2}-1. \end{cases}$$

Applying Theorems 5 and 14 to Lemma 11 we observe $\mathbb{E}[D_2(G_t)]$ asymptotically dominates $\mathbb{E}[D(G_t)]$ and therefore

Theorem 15. *It holds that*

$$\mathrm{Var}[\deg_t(t)] = \begin{cases} \Theta(1) & \text{if } p < \sqrt{2}-1, \\ \Theta(\log t) & \text{if } p = \sqrt{2}-1, \\ \Theta(t^{p^2+2p-1}) & \text{if } p > \sqrt{2}-1. \end{cases}$$

Now we present the asymptotic expressions for $\mathrm{Var}[\deg_t(s)]$. Here we have two cases with different regimes – however it should be noted that the leading terms may have different exact leading constants for different ranges of p and r, the same as we have in Theorem 9.

Theorem 16. *Asymptotically as $t \to \infty$ it holds that:*
(i) *for $s = O(1)$*

$$\mathrm{Var}[\deg_t(s)] = \begin{cases} \Theta(\log t) & \text{if } p = 0, \\ \Theta(t^{2p}) & \text{if } p > 0. \end{cases}$$

(ii) *for $s = \omega(1)$*

$$\mathrm{Var}[\deg_t(s)] = \begin{cases} \Theta\big(\log\left(\frac{t}{s}\right)\big) & \text{if } p = 0 \\ \Theta\big(\left(\frac{t}{s}\right)^{2p}\big) & \text{if } 0 < p < \sqrt{2}-1 \\ \Theta\big(\left(\frac{t}{s}\right)^{2p}\log s\big) & \text{if } p = \sqrt{2}-1 \\ \Theta\big(\left(\frac{t}{s}\right)^{2p} s^{p^2+2p-1}\big) & \text{if } p > \sqrt{2}-1. \end{cases}$$

We conclude by stating the formula for $\mathrm{Var}[D(G_t)]$.

Theorem 17. *It is true that asymptotically as $t \to \infty$*

$$\mathrm{Var}[D(G_t)] = \begin{cases} \Theta(1) & \text{if } p < \frac{1}{2}, \\ \Theta(\log^2 t) & \text{if } p = \frac{1}{2}, \\ \Theta(t^{4p-2}) & \text{if } p > \frac{1}{2}. \end{cases}$$

It is worth noting that for both $\deg_t(t)$ and $\deg_t(s)$ the order of growth of variance is completely dominated by the second moment (unless it's $O(1)$). However, with $D(G_t)$ the situation is different: both $\mathbb{E}[D^2(G_t)]$ and $(\mathbb{E}[D(G_t)])^2$ have the same order – although with different leading constants.

3 Analysis

In this section we provide proofs of our main results. We start with a sequence of lemmas that allow us to solve a particular type of recurrence encountered in this analysis, and then we extract asymptotics using analytic tools.

3.1 Useful lemmas

We begin our analysis by deriving a series of lemmas useful for the analysis of the following type of recurrence

$$\mathbb{E}[f(G_{n+1}) \mid G_n] = f(G_n)g_1(n) + g_2(n) \tag{1}$$

for some nonnegative functions $g_1(n)$, $g_2(n)$ and a Markov process G_n. It should be noted that our recurrences for $\mathbb{E}[\deg_t(s)]$ and $\mathbb{E}[D(G_t)]$ (e.g. see Lemmas 1 and 3) fall under this pattern.

Next lemma is a generalization of a result obtained in [6], where only the case $g_1(n) = 1 + \frac{a}{n}$, $a > 0$, was analyzed.

Lemma 18. *Let $(G_n)_{n=n_0}^{\infty}$ be a Markov process for which $\mathbb{E}f(G_{n_0}) > 0$ and (1) holds with $g_1(n) > 0$, $g_2(n) \geqslant 0$ for all $n = n_0, n_0+1, \ldots$. Then*
(ii) *The process $(M_n)_{n=n_0}^{\infty}$ defined by $M_{n_0} = f(G_{n_0})$ and*

$$M_n = f(G_n)\prod_{k=n_0}^{n-1}\frac{1}{g_1(k)} - \sum_{j=n_0}^{n-1} g_2(j)\prod_{k=n_0}^{j}\frac{1}{g_1(k)} \tag{2}$$

is a martingale.
(ii) *For all $n \geqslant n_0$*

$$\begin{aligned}\mathbb{E}f(G_n) &= f(G_{n_0}) \prod_{k=n_0}^{n-1} g_1(k) + \sum_{j=n_0}^{n-1} g_2(j) \prod_{k=j+1}^{n-1} g_1(k) \\ &= \prod_{k=n_0}^{n-1} g_1(k) \left(f(G_{n_0}) + \sum_{j=n_0}^{n-1} g_2(j) \prod_{k=n_0}^{j} \frac{1}{g_1(k)} \right).\end{aligned}$$

Proof. Observe that

$$\begin{aligned}\mathbb{E}[M_{n+1} \mid G_n] &= \mathbb{E}[f(G_{n+1}) \mid G_n] \prod_{k=n_0}^{n} \frac{1}{g_1(k)} - \sum_{j=n_0}^{n} g_2(j) \prod_{k=n_0}^{j} \frac{1}{g_1(k)} \\ &= f(G_n) \prod_{k=n_0}^{n-1} \frac{1}{g_1(k)} - \sum_{j=n_0}^{n-1} g_2(j) \prod_{k=n_0}^{j} \frac{1}{g_1(k)} = M_n\end{aligned}$$

which proves (i). Furthermore, after some algebra and taking expectation with respect to G_n we arrive at

$$\begin{aligned}\mathbb{E}f(G_n) &= \mathbb{E}[M_n] \prod_{k=n_0}^{n-1} g_1(k) + \sum_{j=n_0}^{n-1} g_2(j) \prod_{k=n_0}^{j} \frac{1}{g_1(k)} \prod_{k=n_0}^{n-1} g_1(k) \\ &= f(G_{n_0}) \prod_{k=n_0}^{n-1} g_1(k) + \sum_{j=n_0}^{n-1} g_2(j) \prod_{k=j+1}^{n-1} g_1(k)\end{aligned}$$

which completes the proof. □

We now observe that (2) as a solution of recurrences of type (1) contains sophisticated products and the sum of products with which we must deal to find asymptotics. The next lemma shows how to handle such products.

Lemma 19. *Let $W_1(k)$, $W_2(k)$ be polynomials of degree d with respective roots a_i, b_i ($i = 1, \ldots, d$), that is, $W_1(k) = \prod_{i=1}^{d}(k - a_i)$ and $W_2(k) = \prod_{j=1}^{d}(k - b_j)$. Then*

$$\prod_{k=n_0}^{n-1} \frac{W_1(k)}{W_2(k)} = \prod_{i=1}^{d} \frac{\Gamma(n-a_i)}{\Gamma(n-b_i)} \frac{\Gamma(n_0-b_i)}{\Gamma(n_0-a_i)}.$$

Proof. We have

$$\prod_{k=n_0}^{n-1} \frac{W_1(k)}{W_2(k)} = \prod_{k=n_0}^{n-1} \prod_{i=1}^{d} \frac{k-a_i}{k-b_i} = \prod_{i=1}^{d} \prod_{k=n_0}^{n-1} \frac{k-a_i}{k-b_i} = \prod_{i=1}^{d} \frac{\Gamma(n-a_i)}{\Gamma(n-b_i)} \frac{\Gamma(n_0-b_i)}{\Gamma(n_0-a_i)}$$

which completes the proof. □

The next lemma presents the well-known asymptotic expansion of the gamma function but we include it here for the sake of completeness.

Lemma 20 (Abramowitz, Stegun [1])**.** *For any $a, b \in \mathbb{R}$ if $n \to \infty$, then*

$$\begin{aligned}\frac{\Gamma(n+a)}{\Gamma(n+b)} &= n^{a-b} \sum_{k=0}^{\infty} \binom{a-b}{k} B_k^{(a-b+1)}(a) \cdot n^{-k} \\ &= n^{a-b}\left(1 + \frac{(a-b)(a+b-1)}{2n} + O\left(\frac{1}{n^2}\right)\right),\end{aligned}$$

where $B_k^{(l)}(x)$ are the generalized Bernoulli polynomials.

Now we deal with sum of products as seen in (2). In particular, we are interested in the following sum of products

$$\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)}$$

with $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$. In the next three lemmas we consider three cases: $a+1 > b$, $a+1 = b$ and $a+1 < b$.

Lemma 21. *Let $a_i, b_i \in \mathbb{R}$ ($k \in \mathbb{N}$) with $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$ such that $a+1 > b$. Then it holds asymptotically for $n \to \infty$ that*

$$\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} = \frac{1}{a-b+1} n^{a-b+1} + O\left(n^{\max\{a-b,0\}}\right)$$

Proof. We estimate the sum using Lemma 20 and the Euler-Maclaurin formula [16, p. 294]

$$\begin{aligned}\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} &= \sum_{j=n_0}^{n} j^{a-b}\left(1 + O\left(\frac{1}{j}\right)\right) = \int_{n_0}^{n} j^{a-b}\left(1 + O\left(\frac{1}{j}\right)\right) \mathrm{d}j \\ &= n^{a-b+1}\left(\frac{1}{a-b+1} + O\left(\frac{1}{n}\right)\right) + O(1)\end{aligned}$$

which completes the proof. □

Lemma 22. *Let $a_i, b_i \in \mathbb{R}$ ($k \in \mathbb{N}$) with $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$ such that $a+1 = b$. Then asymptotically*

$$\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} = \ln n + O(1)$$

Proof. We proceed as before

$$\begin{aligned}\sum_{j=n_0}^{n} \frac{\prod_{i=1}^{k} \Gamma(j+a_i)}{\prod_{i=1}^{k} \Gamma(j+b_i)} &= \sum_{j=n_0}^{n} \frac{1}{j}\left(1 + O\left(\frac{1}{j}\right)\right) = \int_{n_0}^{n} \frac{1}{j}\left(1 + O\left(\frac{1}{j}\right)\right) \mathrm{d}j \\ &= \left[\ln j + O(1)\right]_{n_0}^{n} = \ln n + O(1)\end{aligned}$$

which completes the proof. □

Lemma 23. *Let $a_i, b_i \in \mathbb{R}$ $(i = 1, \ldots, k,\ k \in \mathbb{N})$ with $a = \sum_{i=1}^{k} a_i$, $b = \sum_{i=1}^{k} b_i$ such that $a + 1 < b$. Then it holds for every $n \in \mathbb{N}_+$ that*

$$\sum_{j=n}^{\infty} \frac{\prod_{i=1}^{k} \Gamma(j + a_i)}{\prod_{i=1}^{k} \Gamma(j + b_i)} = \frac{\prod_{i=1}^{k} \Gamma(n + a_i)}{\prod_{i=1}^{k} \Gamma(n + b_i)} \ {}_{k+1}F_k\left[{n+a_1,\ldots,n+a_k,1 \atop n+b_1,\ldots,n+b_k} ; 1 \right]$$

where ${}_pF_q\left[{\mathbf{a} \atop \mathbf{b}} ; z\right]$ is the generalized hypergeometric function. Moreover it is true that asymptotically

$$\sum_{j=n}^{\infty} \frac{\prod_{i=1}^{k} \Gamma(j + a_i)}{\prod_{i=1}^{k} \Gamma(j + b_i)} = n^{a-b+1} \left(\frac{1}{b - a - 1} + O\left(\frac{1}{n}\right) \right).$$

Proof. The proof of the first formula follows directly from the definition of the generalized hypergeometric function. Second formula follows from Lemma 20, as we know that for $n \to \infty$:

$$\begin{aligned}\sum_{j=n}^{\infty} \frac{\prod_{i=1}^{k} \Gamma(j + a_i)}{\prod_{i=1}^{k} \Gamma(j + b_i)} &= \sum_{j=n}^{\infty} j^{a-b} \left(1 + O\left(\frac{1}{j}\right)\right) = \int_{n}^{\infty} j^{a-b} \left(1 + O\left(\frac{1}{j}\right)\right) \mathrm{d}j \\ &= n^{a-b+1} \left(\frac{1}{b - a - 1} + O\left(\frac{1}{n}\right)\right)\end{aligned}$$

as desired. □

3.2 Proofs for expected values

Proof of Lemma 1. Observe that for any $t \geqslant s$ we know that vertex s may be connected to vertex $t + 1$ in one of the following two cases:

- either $s \in \mathcal{N}_t(\mathit{parent}(t + 1))$ (which holds with probability $\frac{\deg_t(s)}{t}$) and we add an edge between s and $t + 1$ (with probability p),
- or $s \notin \mathcal{N}_t(\mathit{parent}(t + 1))$ (with probability $\frac{t-\deg_t(s)}{t}$) and we an add edge between s and $t + 1$ (with probability $\frac{r}{t}$).

Therefore, we obtain the following recurrence for $\mathbb{E}[\deg_t(s)]$:

$$\begin{aligned}\mathbb{E}[\deg_{t+1}(s) \mid G_t] &= \left(\frac{\deg_t(s)}{t} p + \frac{t - \deg_t(s)}{t} \frac{r}{t} \right) (\deg_t(s) + 1) \\ &\quad + \left(\frac{\deg_t(s)}{t} (1 - p) + \frac{t - \deg_t(s)}{t} \left(1 - \frac{r}{t}\right) \right) \deg_t(s) \\ &= \deg_t(s) \left(1 + \frac{p}{t} - \frac{r}{t^2}\right) + \frac{r}{t}.\end{aligned}$$

After applying the law of total expectation we get the desired result. □

Proof of Lemma 2. We first observe that it follows from the definition of the model that the degree of the new vertex $t+1$ is the total number of edges from $t+1$ to $N_t(\mathit{parent}(t+1))$ (chosen independently with probability p) and to all other vertices (chosen independently with probability $\frac{r}{t}$). Note that it can be expressed as a sum of two binomial variables

$$\begin{aligned}&\deg_{t+1}(t+1)\\ &\qquad \sim \mathrm{Bin}\left(\deg_t(\mathit{parent}(t+1)), p\right) + \mathrm{Bin}\left(t - \deg_t(\mathit{parent}(t+1)), \frac{r}{t}\right).\end{aligned}$$

These variables are independent conditional on the choice of parent, hence

$$\begin{aligned}\mathbb{E}[\deg_{t+1}(t+1) \mid G_t] &= \sum_{k=0}^{t} \Pr(\deg_t(\mathit{parent}(t+1)) = k) \left(pk + \frac{r}{t}(t-k)\right)\\ &= \left(p - \frac{r}{t}\right) \sum_{k=0}^{t} k \Pr(\deg_t(\mathit{parent}(t+1)) = k) + r.\end{aligned}$$

Since parent sampling is uniform, we know that $\Pr(\mathit{parent}(t+1) = i) = \frac{1}{t}$ and therefore

$$D(G_t) = \sum_{k=0}^{t} k \Pr(\deg_t(\mathit{parent}(t+1)) = k).$$

By combining the two equations above with the law of total expectation we finally establish the lemma. □

Proof of Lemma 3. Again we turn directly to the definition of the model, from which we derive the following recurrence for the average degree of G_{t+1}:

$$\begin{aligned}\mathbb{E}[D(G_{t+1}) \mid G_t] &= \frac{1}{t+1} \left(\sum_{i=1}^{t} \deg_t(i) + 2\mathbb{E}\left[\deg_{t+1}(t+1) \mid G_t\right] \right)\\ &= \frac{1}{t+1} \left(tD(G_t) + 2\mathbb{E}[\deg_{t+1}(t+1) \mid G_t]\right)\\ &= D(G_t) \left(1 + \frac{2p-1}{t+1} - \frac{2r}{t(t+1)}\right) + \frac{2r}{t+1}.\end{aligned}$$

where the last equality follows from Lemma 2.

The final result is a direct consequence of applying the law of total expectation. □

Proof of Lemma 4. It is sufficient to apply Lemmas 18 and 19 to Lemma 3 with $g_1(t) = 1 + \frac{2p-1}{t+1} - \frac{2r}{t(t+1)}$ and $g_2(t) = \frac{2r}{t+1}$. □

Proof of Theorem 5. For convenience, we introduce

$$A_1(t) := \frac{\Gamma(t+c_3)\Gamma(t+c_4)}{\Gamma(t)\Gamma(t+1)} D(G_{t_0}) \frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}$$

$$A_2(t) := \frac{\Gamma(t+c_3)\Gamma(t+c_4)}{\Gamma(t)\Gamma(t+1)} \sum_{j=t_0}^{t-1} \frac{\Gamma(j+1)^2}{\Gamma(j+c_3+1)\Gamma(j+c_4+1)},$$

so that

$$\mathbb{E}[D(G_t)] = A_1(t) + 2rA_2(t).$$

Then, the result follows directly from the equations above combined with Lemmas 21 to 23 for the respective ranges of p:

$$A_1(t) = t^{2p-1} D(G_{t_0}) \frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)}(1+o(1))$$

$$A_2(t) = \begin{cases} \frac{1}{1-2p}(1+o(1)) & \text{if } p < \frac{1}{2}, \\ \ln t\,(1+o(1)) & \text{if } p = \frac{1}{2}, \\ t^{2p-1} \frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} \frac{t_0}{t_0^2+2pt_0-2r} & \\ \quad {}_3F_2\left[{t_0+1, t_0+1, 1 \atop t_0+c_3+1, t_0+c_4+1} ; 1 \right](1+o(1)) & \text{if } p > \frac{1}{2}. \end{cases}$$

□

Proof of Lemma 8. By applying Lemma 18 to Lemma 1 with $g_1(t) = 1 + \frac{p}{t} - \frac{r}{t^2}$ and $g_2(t) = \frac{r}{t}$ we obtain

$$\mathbb{E}[\deg_t(s)] = \mathbb{E}[\deg_s(s)] \prod_{k=s}^{t-1} \left(1 + \frac{p}{k} - \frac{r}{k^2}\right) + \sum_{j=s}^{t-1} \frac{r}{j} \prod_{k=j+1}^{t-1} \left(1 + \frac{p}{k} - \frac{r}{k^2}\right).$$

Now we combine this result with Lemma 19 and find that

$$\mathbb{E}[\deg_t(s)] = \frac{\Gamma(t+c_1)\Gamma(t+c_2)}{\Gamma(t)^2} \left(\mathbb{E}[\deg_s(s)] \frac{\Gamma(s)^2}{\Gamma(s+c_1)\Gamma(s+c_2)} + r \sum_{j=s}^{t-1} \frac{\Gamma(j)\Gamma(j+1)}{\Gamma(j+c_1+1)\Gamma(j+c_2+1)} \right)$$

where $c_1 = \frac{p+\sqrt{p^2+4r}}{2}$, $c_2 = \frac{p-\sqrt{p^2+4r}}{2}$.

Now it is sufficient to apply Lemma 6 to this equation to get the final result. □

Proof of Theorem 9. All parts of this theorem – as it was in the case of $\mathbb{E}[D(G_t)]$ above – follow as consequences of Lemma 8 combined with Lemmas 21 to 23 for the respective ranges of p and r. □

3.3 Proofs for second moments and variances

Proof of Lemma 10. The formula follows directly from the model itself:

$$\begin{aligned}\mathbb{E}[\deg_{t+1}^2(s) \mid G_t] &= \left(\frac{\deg_t(s)}{t}p + \frac{t-\deg_t(s)}{t}\frac{r}{t}\right)(\deg_t(s)+1)^2 \\ &\quad + \left(\frac{\deg_t(s)}{t}(1-p) + \frac{t-\deg_t(s)}{t}\left(1-\frac{r}{t}\right)\right)\deg_t^2(s),\end{aligned}$$

so we need only to rearrange the terms and apply the law of total expectation. □

Proof of Lemma 11. It may be observed that

$$\begin{aligned}&\mathbb{E}[\deg_{t+1}^2(t+1) \mid G_t] = \\ &= \sum_{k=0}^{t} \Pr(\deg_t(\mathit{parent}(t+1)) = k) \sum_{a=0}^{k} \binom{k}{a} p^a (1-p)^{k-a} \\ &\quad \sum_{b=0}^{t-k} \binom{t-k}{b} \left(\frac{r}{t}\right)^b \left(1-\frac{r}{t}\right)^{t-k-b} (a+b)^2 \\ &= \sum_{k=0}^{t} \Pr(\deg_t(\mathit{parent}(t+1)) = k) \\ &\quad \left(k^2\left(p^2 - \frac{2pr}{t} + \frac{r^2}{t^2}\right) + k\left(p - p^2 + 2pr - \frac{r+2r^2}{t}\right) + r^2 + r - \frac{r^2}{t}\right) \\ &= D_2(G_t)\left(p^2 - \frac{2pr}{t} + \frac{r^2}{t^2}\right) + D(G_t)\left(p - p^2 + 2pr - \frac{r+2r^2}{t} + \frac{r^2}{t^2}\right) \\ &\quad + r^2 + r - \frac{r^2}{t},\end{aligned}$$

since we have similarly as before in the case of $D(G_t)$:

$$\begin{aligned}D_2(G_t) &= \sum_{i=1}^{t} \Pr(\mathit{parent}(t+1) = i) \deg_t^2(i) \\ &= \sum_{k=0}^{t} \Pr(\deg_t(\mathit{parent}(t+1)) = k) k^2.\end{aligned}$$

This, after applying the law of total expectation, establishes the lemma. □

Proof of Lemma 12. We start from the definition of the second moment of the degree distribution of G_t:

$$\mathbb{E}[D_2(G_{t+1}) \mid G_t] = \frac{1}{t+1}\mathbb{E}\left[\sum_{i=1}^{t+1} \deg_{t+1}^2(i) \mid G_t\right]$$

$$
\begin{aligned}
&= \frac{1}{t+1}\mathbb{E}\left[\sum_{i=1}^{t}\left(\deg_t(i)+I_{t+1}(i)\right)^2+\deg_{t+1}^2(t+1)\;\Big|\; G_t\right] \\
&= \frac{1}{t+1}\mathbb{E}\left[\sum_{i=1}^{t}\deg_t^2(i)+2\sum_{i=1}^{t}\deg_t(i)I_{t+1}(i)\right. \\
&\qquad\left.+\sum_{i=1}^{t}I_{t+1}^2(i)+\deg_{t+1}^2(t+1)\;\Big|\; G_t\right]
\end{aligned}
$$

where $I_{t+1}(i)$ is an indicator variable denoting whether there is an edge between vertices $t+1$ and i.

Now we use the following simple facts

$$
\begin{aligned}
\sum_{i=1}^{t}I_{t+1}^2(i) &= \sum_{i=1}^{t}I_{t+1}(i)=\deg_{t+1}(t+1), \\
\mathbb{E}\left[\sum_{i=1}^{t}\deg_t(i)I_{t+1}(i)\;\Big|\; G_t\right] &= \sum_{i=1}^{t}\deg_t(i)\mathbb{E}[I_{t+1}(i)\mid G_t] \\
&= \sum_{i=1}^{t}\deg_t(i)\left(\frac{\deg_t(i)}{t}p+\frac{t-\deg_t(i)}{t}\frac{r}{t}\right) \\
&= \left(p-\frac{r}{t}\right)D_2(G_t)+rD(G_t).
\end{aligned}
$$

This lead us to

$$
\begin{aligned}
\mathbb{E}[D_2(G_{t+1})\mid G_t] = D_2(G_t)&\left(1+\frac{2p+p^2-1}{t+1}-\frac{2r(1+p)}{t(t+1)}+\frac{r^2}{t^2(t+1)}\right) \\
&+D(G_t)\left(\frac{2p-p^2+2pr+2r}{t+1}-\frac{2r+2r^2}{t(t+1)}+\frac{r^2}{t^2(t+1)}\right) \\
&+\frac{r^2+2r}{t+1}-\frac{r^2}{t(t+1)}.
\end{aligned}
$$

To finish the proof, we again apply the law of total expectation. □

Proof of Lemma 13. From the model definition we get

$$
\begin{aligned}
\mathbb{E}[D^2(G_{t+1})\mid G_t] &= \frac{1}{(t+1)^2}\mathbb{E}\left[\left(\sum_{i=1}^{t+1}\deg_{t+1}(i)\right)^2\;\Big|\; G_t\right] \\
&= \frac{1}{(t+1)^2}\left(\left(\sum_{i=1}^{t}\deg_t(i)\right)^2+4\mathbb{E}\left[\deg_{t+1}(t+1)\mid G_t\right]\sum_{i=1}^{t}\deg_t(i)\right. \\
&\qquad\left.+4\mathbb{E}\left[\deg_{t+1}^2(t+1)\mid G_t\right]\right).
\end{aligned}
$$

Now it is sufficient to substitute the respective parts of the formula by $D^2(G_t)$ and $D(G_t)$ and once again apply the law of total expectation. □

Proof of Theorem 14. We split the formula

$$\begin{aligned}
B_1(t) &:= \frac{\Gamma(t+c_5)\Gamma(t+c_6)\Gamma(t+c_7)}{\Gamma(t)^2\Gamma(t+1)} D_2(G_{t_0}) \frac{\Gamma(t_0)^2\Gamma(t_0+1)}{\Gamma(t_0+c_5)\Gamma(t_0+c_6)\Gamma(t_0+c_7)}, \\
B_2(t) &:= \frac{\Gamma(t+c_5)\Gamma(t+c_6)\Gamma(t+c_7)}{\Gamma(t)^2\Gamma(t+1)}(1+o(1)) \\
&\qquad \sum_{j=t_0}^{t-1} \mathbb{E}[D(G_j)] \frac{1}{j+1} \frac{\Gamma(j+1)^2\Gamma(j+2)}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)}, \\
B_3(t) &:= \frac{\Gamma(t+c_5)\Gamma(t+c_6)\Gamma(t+c_7)}{\Gamma(t)^2\Gamma(t+1)}(1+o(1)) \\
&\qquad \sum_{j=t_0}^{t-1} \frac{1}{j+1} \frac{\Gamma(j+1)^2\Gamma(j+2)}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)},
\end{aligned}$$

so that

$$\mathbb{E}[D_2(G_t)] = B_1(t) + (2p - p^2 + 2pr + 2r)B_2(t) + (r^2 + 2r)B_3(t),$$

where c_5, c_6, c_7 are the roots of equation $t^3 - (2p + p^2)t^2 - 2r(1+p)t - r^2 = 0$.

Now we may show that asymptotically:

$$\begin{aligned}
B_1(t) &= t^{p^2+2p-1} D_2(G_{t_0}) \frac{\Gamma(t_0)^2\Gamma(t_0+1)}{\Gamma(t_0+c_5)\Gamma(t_0+c_6)\Gamma(t_0+c_7)}(1+o(1)), \\
B_2(t) &= t^{p^2+2p-1}(1+o(1)) \\
&\qquad \sum_{j=t_0}^{t-1} \mathbb{E}[D(G_j)] \frac{\Gamma(j+1)^3}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)}, \\
B_3(t) &= t^{p^2+2p-1}(1+o(1)) \\
&\qquad \sum_{j=t_0}^{t-1} \frac{\Gamma(j+1)^3}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)}.
\end{aligned}$$

The rate of growth of $B_1(t)$ is of course $\Theta(t^{p^2+2p-1})$. The rates of growth of $B_3(t)$ can be found by applying Lemmas 21 to 23 to the respective cases: for example, when $p > \sqrt{2}-1$ we have $B_3(t) = \Theta(t^{p^2+2p-1}\sum j^{-p^2-2p}) = \Theta(1)$.

Finding the asymptotics of $B_2(t)$ is more complicated. To solve it we substitute $\mathbb{E}[D(G_j)]$ using Theorem 5 and note that

$$B_{21}(t) = \sum_{j=t_0}^{t-1} \frac{\Gamma(j+c_3)\Gamma(j+c_4)\Gamma(j+1)^2}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)\Gamma(j)},$$

$$B_{22}(t) = \sum_{j=t_0}^{t-1} \sum_{k=t_0}^{j-1} \frac{\Gamma(k+1)^2}{\Gamma(k+c_3+1)\Gamma(k+c_4+1)} \frac{\Gamma(j+c_3)\Gamma(j+c_4)\Gamma(j+1)^2}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)\Gamma(j)},$$

so that

$$B_2(t) = t^{p^2+2p-1}(1+o(1)) \left(D(G_{t_0}) \frac{\Gamma(t_0)\Gamma(t_0+1)}{\Gamma(t_0+c_3)\Gamma(t_0+c_4)} B_{21}(t) + 2rB_{22}(t) \right).$$

Here $B_{21}(t)$ poses no problem, as it can be analyzed similarly as $B_3(t)$. Moreover, we bound

$$B_{22}(t) \leqslant \sum_{k=t_0}^{t-1} \frac{\Gamma(k+1)^2}{\Gamma(k+c_3+1)\Gamma(k+c_4+1)} \sum_{j=t_0}^{t-1} \frac{\Gamma(j+c_3)\Gamma(j+c_4)\Gamma(j+1)^2}{\Gamma(j+c_5+1)\Gamma(j+c_6+1)\Gamma(j+c_7+1)\Gamma(j)}.$$

Now if $p > \sqrt{2}-1$ the right hand side is upper bounded by a constant, as the first sum is of order $\Theta(t^{1-2p})$ and the second sum is of order $\Theta(t^{2p-2p-p^2})$ – and therefore in total $B_{22}(t) = O(t^{1-2p-p^2})$, so it's bounded from above by a constant.

All other cases are treated in exactly analogous way. By putting them all together in the way presented above we obtain the final result. □

Proof of Theorem 16. Lemma 10 combined with Lemma 18 lead us to

$$\mathbb{E}[\deg_t^2(s)] = \mathbb{E}[\deg_s^2(s)] \prod_{k=s}^{t-1} \left(1 + \frac{2p}{k} - \frac{2r}{k^2}\right) + \sum_{j=s}^{t-1} \left[\mathbb{E}[\deg_j(s)] \left(\frac{p+2r}{j} - \frac{r}{j^2} \right) + \frac{r}{j} \right] \prod_{k=j+1}^{t-1} \left(1 + \frac{2p}{k} - \frac{2r}{k^2}\right).$$

From Lemmas 19 and 20 we find that

$$\prod_{k=j}^{t-1} \frac{k^2+2kp-2r}{k^2} = \Theta\left(\frac{t^{2p}}{j^{2p}}\right).$$

To find the asymptotic expression for $\mathbb{E}[\deg_t^2(s)]$ it is sufficient to combine this with Lemma 8 and Theorem 15 (substituting $\mathbb{E}[\deg_j(s)]$ and $\mathbb{E}[\deg_s^2(s)]$, respectively) and apply Lemmas 21 to 23 for the proper ranges of p. For example, for $s = O(1)$ it holds that

- the first part grows like $\Theta(\log t)$ for $p = 0$ and $\Theta(t^{2p})$ otherwise,
- the second part grows like $\Theta(1)$, $\Theta(\log t)$ or $\Theta(t^{2p-1})$ for p less, equal to or greater than $\frac{1}{2}$,

so the first part is always asymptotically greater than the second part.

The variance is obtained in the usual way, by subtracting the square of the respective formulas from Theorem 5. Note that here, in contrast to Theorem 9 we do not need to distinguish the case $s = ct - o(t)$, as it differs only in the leading constant, but not in the rate of growth. □

Proof of Theorem 17. The product form of $\mathbb{E}[D^2(G_t)]$ may be derived from Lemmas 13 and 18 as

$$\begin{aligned}\mathbb{E}[D^2(G_t)] = D^2(G_{t_0}) &\prod_{j=t_0}^{t-1} \frac{j^2+4jp-4r}{(j+1)^2} \\ &+ \sum_{j=t_0}^{t-1} \mathbb{E}[D_2(G_j)] \frac{4p^2}{(j+1)^2}(1+o(1)) \prod_{k=j+1}^{t-1} \frac{k^2+4kp-4r}{(k+1)^2} \\ &+ \sum_{j=t_0}^{t-1} \mathbb{E}[D(G_j)] \frac{4jr}{(j+1)^2}(1+o(1)) \prod_{k=j+1}^{t-1} \frac{k^2+4kp-4r}{(k+1)^2} \\ &+ \sum_{j=t_0}^{t-1} \frac{4r^2}{(j+1)^2}(1+o(1)) \prod_{k=j+1}^{t-1} \frac{k^2+4kp-4r}{(k+1)^2}.\end{aligned}$$

Now use Lemmas 19 and 20 to prove that for any j it holds that

$$\prod_{k=j}^{t-1} \frac{k^2+4kp-4r}{(k+1)^2} = \Theta\left(\frac{t^{4p-2}}{j^{4p-2}}\right).$$

Finally, we may consider each term of the sum separately and find – using Lemmas 21 to 23 – that asymptotically:

- the first term grows like $\Theta(t^{4p-2})$,
- the second term grows like $\Theta(t^{4p-2}\sum \mathbb{E}[D_2(G_j)j^{-4p}])$, which is $\Theta(t^{-1})$, $\Theta(t^{-1}\log t)$ or $\Theta(t^{p^2+2p-2})$ for $p < \sqrt{2}-1$, $p = \sqrt{2}-1$ and $p > \sqrt{2}-1$ respectively – and in each case dominated by the first term,
- the third term grows like $\Theta(t^{4p-2}\sum \mathbb{E}[D(G_j)j^{1-4p}])$, which is $\Theta(1)$, $\Theta(\log t)$ or $\Theta(t^{2p-1})$ for $p < \frac{1}{2}$, $p = \frac{1}{2}$ and $p > \frac{1}{2}$ respectively,
- the fourth term grows like $O(t^{4p-2}\sum j^{-4p})$, which is always asymptotically dominated by the second term.

This establishes the behavior of $\mathbb{E}[D^2(G_t)]$ – and the only remaining part is to combine this with Theorem 5 to find $\mathrm{Var}[D(G_t)]$. □

4 Discussion

In this paper we focus on rigorous and precise analysis of the expected average degree and variance of a given node in the network as well as the average degree over all nodes. We presented exact and asymptotic results showing phase transitions of these quantities as a function of p.

It is worth noting that the parameter p solely drives the rate of growth of both first and second moments of variables $D(G_t)]$, $\deg_t(t)$ and $\deg_t(s)$. The parameter r impacts only the leading constant and lower order terms. The proposed methodology can be easily extended to obtain higher moments of the above quantities, if needed.

The future work may include investigations both the large deviation of the degree distribution as well as the complete spectrum of the degree distribution (i.e., the number of nodes of degree k) as a function of k, t, G_{t_0}, p and r.

References

[1] M. Abramowitz and I. Stegun. *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*, volume 55. Dover Publications, 1972.

[2] G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau, S. C. Sahinalp. The degree distribution of the generalized duplication model. *Theo. Comp. Sci.*, 369(1–3):239–249, 2006.

[3] F. Chung, L. Lu, T. G. Dewey, D. Galas. Duplication models for biological networks. *J. of Comp. Biology*, 10(5):677–687, 2003.

[4] R. Colak, F. Hormozdiari, F. Moser, A. Schönhuth, J. Holman, M. Ester, S. C. Sahinalp. Dense graphlet statistics of protein interaction and random networks. In *Biocomputing 2009*, pages 178–189. World Scientific Publishing, 2009.

[5] R. Diestel. *Graph Theory.* Springer, 2005.

[6] F. Hermann, P. Pfaffelhuber. Large-scale behavior of the partial duplication random graph. *ALEA*, 13:687–710, 2016.

[7] F. Hormozdiari, P. Berenbrink, N. Pržulj, S. C. Sahinalp. Not all scale-free networks are born equal: the role of the seed graph in PPI network evolution. *PLoS Comp. Biology*, 3(7):e118, 2007.

[8] J. Jordan. The connected component of the partial duplication graph. *ALEA*, 15:1431–1445, 2018.

[9] S. Li, K. P. Choi, T. Wu. Degree distribution of large networks generated by the partial duplication model. *Theo. Comp. Sci.*, 476:94–108, 2013.

[10] T. Łuczak, A. Magner, W. Szpankowski. Asymmetry and structural information in preferential attachment graphs. arXiv:1607.04102, 2016.

[11] M. Newman. *Networks: An Introduction.* Oxford University Press, 2010.

[12] R. Pastor-Satorras, E. Smith, R. Solé. Evolving protein interaction networks through gene duplication. *J. of Theo. Biology*, 222(2):199–210, 2003.

[13] M. Shao, Y. Yang, J. Guan, S. Zhou. Choosing appropriate models for protein–protein interaction networks: a comparison study. *Briefings in Bioinformatics*, 15(5):823–838, 2013.

[14] R. Solé, R. Pastor-Satorras, E. Smith, T. Kepler. A model of large-scale proteome evolution. *Advances in Complex Systems*, 5(1):43–54, 2002.

[15] J. Sreedharan, K. Turowski, W. Szpankowski. Revisiting Parameter Estimation in Biological Networks: Influence of Symmetries. *IEEE/ACM Trans. on Comp. Biology and Bioinf.*, 2020.

[16] W. Szpankowski. *Average case analysis of algorithms on sequences.* John Wiley & Sons, 2011.

[17] K. Turowski, A. Magner, W. Szpankowski. Compression of Dynamic Graphs Generated by a Duplication Model. *Algorithmica*, 82:2687–2707, 2020.

[18] R. Van Der Hofstad. *Random graphs and complex networks.* Cambridge University Press, 2016.

[19] J. Zhang. Evolution by gene duplication: an update. *Trends in Ecology & Evolution*, 18(6):292–298, 2003.