ALEA, Lat. Am. J. Probab. Math. Stat. **13**, 687–710 (2016)
DOI: 10.30757/ALEA.v13-27

Large-scale behavior of the partial duplication random graph

Felix Hermann and Peter Pfaffelhuber

Abteilung für mathematische Stochastik,
Albert-Ludwigs University of Freiburg,
Eckerstraße 1, 79104 Freiburg, Germany
E-mail address: { felix.hermann , p.p } @stochastik.uni-freiburg.de
URL: http://www.stochastik.uni-freiburg.de/homepages/pfaffelh/

Abstract. The following random graph model was introduced for the evolution of protein-protein interaction networks: Let $\mathcal{G} = (G_n)_{n=n_0,n_0+1,...}$ be a sequence of random graphs, where $G_n = (V_n, E_n)$ is a graph with $|V_n| = n$ vertices, $n = n_0, n_0 + 1, ...$ In state $G_n = (V_n, E_n)$, a vertex $v \in V_n$ is chosen from V_n uniformly at random and is partially duplicated. Upon such an event, a new vertex $v' \notin V_n$ is created and every edge $\{v, w\} \in E_n$ is copied with probability p, i.e. E_{n+1} has an edge $\{v', w\}$ with probability p, independently of all other edges.

Within this graph, we study several aspects for large n. (i) The frequency of isolated vertices converges to 1 if $p \leq p^* \approx 0.567143$, the unique solution of $pe^p = 1$. (ii) The number C_k of k-cliques behaves like $n^{kp^{k-1}}$ in the sense that $n^{-kp^{k-1}}C_k$ converges against a non-trivial limit, if the starting graph has at least one k-clique. In particular, the average degree of a vertex (which equals the number of edges – or 2-cliques – divided by the size of the graph) converges to 0 iff $p < 0.5$ and we obtain that the transitivity ratio of the random graph is of the order $n^{-2p(1-p)}$. (iii) The evolution of the degrees of the vertices in the initial graph can be described explicitly. Here, we obtain the full distribution as well as convergence results.

1. Introduction

Random graph models are a topic of active research in probability theory. Since the introduction of the first models, like the models of Erdős and Rényi (1959) and Gilbert (1959), several classes of models for the evolution of networks have been introduced. Frequently, such models try to mimic the behavior of social networks like the internet; see Cooper and Frieze (2003) and Barabási et al. (2002). For

Received by the editors June 29, 2015; accepted July 1st, 2016.
2010 *Mathematics Subject Classification.* 05C80, 60K35.
Key words and phrases. Random graph; degree distribution; duality of Markov processes.

a general introduction to random graphs see the monographs Durrett (2010) and van der Hofstad (2016) and references therein.

Another set of models aim at modeling (micro-)biological networks, such as protein-protein interaction networks (see e.g. Wagner, 2001, and Albert, 2005 for a specific application to yeast) or metabolic networks (Jeong et al., 2000). In this paper, we study a model introduced in Bhan et al. (2002), Pastor-Satorras et al. (2003), Chung et al. (2004) and Bebek et al. (2006a). Here, a vertex models a protein and an edge denotes some form of interaction (e.g. one protein that inhibits the expression of the second protein). Within the genome, the DNA encoding for a protein can be duplicated (which in fact is a long evolutionary process), such that the interactions of the copied protein are partially inherited to the copy; see Ohno (1970). In the model we study, every edge is copied with the same, independent, probability p.

Our analysis extends previous work of Chung et al. (2004), Bebek et al. (2006a) and Bebek et al. (2006b) in various directions. We obtain results for the limit of the (expected) degree distribution for the partial duplication model. Precisely, we are able to determine a critical parameter $p \approx 0.567143$, the unique solution of $pe^p = 1$, below which approximately all vertices are isolated; see Theorem 2.7. Moreover, we are able to obtain almost sure limiting results for the number of k-cliques and k-stars in the random graph; see Theorem 2.9. This entails precise asymptotics of the transitivity ratio of the partial duplication random graph; see Remark 2.13. Lastly, we study the distribution and the large-scale behavior of the degrees of fixed vertices; see Theorem 2.14.

2. Model and results

2.1. *Model.* Let us introduce some notation for (undirected) graphs. Afterwards, we will define the random graph model we will study in the sequel.

Definition 2.1 (Graph, degree, clique)**.**

(1) A(n *undirected*) *graph* (without loops) is a tuple $G = (V, E)$, where V is the set of vertices and $E \subseteq \{\{v, w\} : v, w \in V, v \neq w\}$ is the set of edges.

(2) A k-*clique* within $G = (V, E)$ is a subset $V' \subseteq V$ with $|V'| = k$ and $\{\{v, w\} : v, w \in V', v \neq w\} \subseteq E$ (i.e. all vertices in V' are connected). We denote by $C_k(G)$ the number of k-cliques in G and by $C_k^\circ(G) := C_k(G)/|V|$ the relative frequency of k-cliques.

(3) For a graph $G = (V, E)$ and $v \in V$, we define the *degree* of v by

$$D_v := D_v(G) := |\{w : \{v, w\} \in E\}|.$$

Moreover, the absolute and relative *degree distribution* is given by $(F_k(G))_{k=0,1,2,\ldots}$ and $(F_k^\circ(G))_{k=0,1,2,\ldots}$ through

$$F_k(G) := |\{v : D_v(G) = k\}|, \qquad F_k^\circ(G) := \frac{1}{|V|} F_k(G).$$

We also define their probability generating functions as

$$H_q(G) := \sum_{k=0}^{\infty} F_k(G) q^k, \qquad H_q^\circ(G) := \sum_{k=0}^{\infty} F_k^\circ(G) q^k \qquad \text{for } q \in [0, 1].$$

(4) A *k-star* within $G = (V, E)$ with center v is a vector $(v, v_1, ..., v_k)$ with $v, v_1, ..., v_k \in V$ and $\{v, v_i\} \in E, i = 1, ..., k$, (i.e. every v_i is connected to v). We denote by

$$S_k(G) = \sum_{\ell=k}^{\infty} \ell \cdots (\ell - k + 1) F_\ell(G)$$

the number of k-stars in G and by $S_k^\circ(G) := S_k(G)/|V|$ the relative frequency of k-stars.

Remark 2.2 (Relationships). The quantities we just defined are intertwined by some relationships. For example, since the S_k-values equal the kth factorial moments of the degree distributions, we have

$$S_k(G) = \frac{d^k}{dq^k} H_q(G)\Big|_{q=1}.$$

In particular, note that $S_1(G) = \sum_\ell \ell F_\ell(G) = 2C_2(G)$. This is clear, since every 1-star counts an edge twice, having two possibilities of its center, while each edge corresponds to a 2-clique. However, $C_k(G)$ cannot be obtained from the degree distribution, if $k \geq 3$.

We start with a basic definition of the model; see also Figure 2.1.

Definition 2.3 (Partial duplication random graph)**.** Let $p \in [0, 1]$. We define the following random graph process – called *partial duplication random graph* or PDn graph – $\mathcal{G} = (G_n)_{n=n_0,n_0+1,...}$ with $G_n = (V_n, E_n)$, where G_n is the graph at time $n = n_0, n_0+1, ...$ with vertex set V_n and (undirected) edge set $E_n \subseteq \{\{v, w\} : v, w \in V_n, v \neq w\}$. Starting in some $G_{n_0} = (V_{n_0}, E_{n_0})$ with $|V_{n_0}| = n_0$, the dynamics at time n is as follows: A vertex v is picked uniformly at random from V_n. Upon such an event, a new node $v' \notin V_n$ is created and every edge connected to v (i.e. every $e \in E_n$ with $e = \{v, w\}$ for some $w \in V_n$) is copied with probability p, i.e. $\{v', w\} \in E_{n+1}$ with probability p, independently of all other edges.

We define by $C_k(n) := C_k(G_n)$ and $C_k^\circ(n) := C_k^\circ(G_n)$ the number of k-cliques in G_n and the average number of cliques a vertex is involved in, respectively. Similarly, we define by $S_k(n) := S_k(G_n)$ and $S_k^\circ(n) := S_k^\circ(G_n)$ the number of k-stars in G_n and the average number of k-stars a vertex is centered in, respectively. Moreover, define $F_k(n) := F_k(G_n), F_k^\circ(n) := F_k^\circ(G_n), k = 0, 1, 2, ...$ the degree distribution of G_n and its probability generating function by $H_q(n) := H_q(G_n), H_q^\circ(n) := H_q^\circ(G_n)$.

Throughout the manuscript, we will assume that the initial graph G_{n_0} is connected and deterministic.

Remark 2.4 (Basic observations). (1) Since we assume that the initial graph G_{n_0} is connected, G_n consists of one connected component and singleton nodes which arise if a vertex is copied but none of its edges (unless $p = 1$ where all vertices are connected), $n = n_0 + 1, n_0 + 2, ...$ In Theorem 2.7, we will study the expected proportion of singleton vertices.

(2) Let G_{n_0} be an m-partite graph for some $m \leq n_0$, i.e. there is a partition of V_{n_0} into sets $W_1(n_0), ..., W_m(n_0)$ such that $E_{n_0} \subseteq \big\{\{v, w\} : v \in W_i(n_0), w \in W_j(n_0) \text{ for some } i \neq j\}\big\}$. This means that vertices in $W_i(n_0)$ are only connected to vertices outside $W_i(n_0)$, $i = 1, ..., m$. Then, G_n is m-partite for all $n \geq n_0$.
Indeed, if a vertex $v \in W_i(n_0)$ is copied, it is connected only to vertices outside

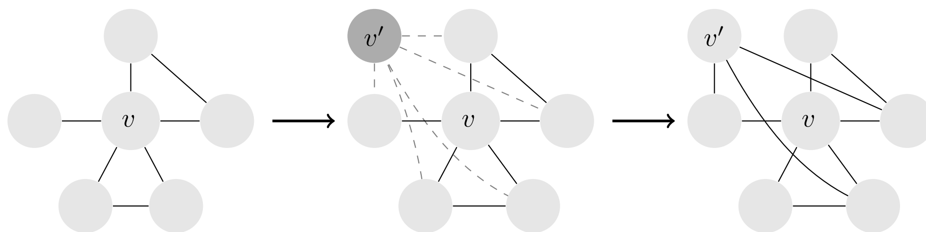

Figure 2.1. Illustration of one step in the PDn random graph; see also Definition 2.3. At time $n = 6$ (since there are 6 vertices in the graph on the left), the vertex v is picked uniformly at random. It is copied, giving rise to the new vertex v', together with all potential edges to neighbors of v (see the dashed lines in the middle). Then, every dashed line is kept independently of the others with probability p. The result is the random graph with $n = 7$ vertices on the right.

$W_i(n_0)$, and so is the copied vertex. Iterating this argument shows that G_n is m-partite, as well. Moreover, we see that the sizes $(W_1(n), ..., W_m(n))_{n=n_0,n_0+1,...}$ of the partition elements, follow Pólya's urn dynamics.

Remark 2.5 (Related random graph models).

(1) In Pastor-Satorras et al. (2003), an extension of the PDn-model was introduced. After partially (with probability p per edge) duplicating a vertex $v \in V_n$, giving rise to the new vertex v', every vertex $w \in V_n$ additionally is connected to v' with probability r/n for some constant $r > 0$. This simple modification is said to induce the scale-free property (Kim et al., 2002; Bebek et al., 2006a), but, as we will see in Remark 2.8.3, this does not hold for at least some values of p.

(2) As stated by Ispolatov et al. (2005) the famous preferential attachment model also shows up in a special limiting case of the PDn-model. Assume the case of small p, which implies that at most one edge is copied upon a duplication event, while the probability that at a time n a fixed node v_k (with degree $D_k(n)$) becomes connected to the new node conditioned on the event that at least one edge is retained equals

$$\pi_k(n) := \frac{\frac{D_k(n)}{n} \cdot p}{\sum_{k \geq 1} F_k^\circ(n)(1-(1-p)^k)}.$$

Using $1-(1-p)^k \stackrel{p\to 0}{\sim} pk$ and that $S_1(n) = 2C_2(n)$, we obtain that $\pi_k(n) \xrightarrow{p\to 0} D_k(n)/2C_2(n)$ for each n. So, when conditioning the PDn-model to have no isolated vertices, the preferential attachment model arises in the limit $p \to 0$.

(3) Another duplication model was recently introduced by Thörnblad (2015) and further analyzed by Backhausz and Móri (2015, 2016). Here, the random graphs consist of disjoint cliques, almost surely. In each time step, a vertex v is chosen uniformly at random and duplicated with probability θ. Upon such an event, a new vertex v' is created and connected to all neighbors of v and to v itself. With probability $1-\theta$, all edges connecting v to its neighbors are deleted. For this model, the degree distribution was analyzed and a phase transition at $\theta = 1/2$ was discovered in Thörnblad (2015). The limiting case $\theta = 1/2$ and the maximal degree was studied in Backhausz and Móri (2015, 2016). The PDn-model is related in the

case $p = \theta = 1$, although the new vertex is not connected to the copied vertex in the PDn-model.

2.2. *Results.* Let us now come to the main conclusions about the PDn model we have derived. First, Theorem 2.7 states a critical value $p \approx 0.567143$, below which almost all vertices have degree 0, i.e. are isolated. Its proof, which is based on a time-continuous version of PDn and a duality argument with a piece-wise deterministic Markov process is found in Section 4. Second, Theorem 2.9 studies the occurrences of certain subgraphs in the PDn: k-cliques, which deliver an understanding of topological properties of the initial graph retained during the process, and k-stars, which give insights in the degree distribution, since they describe its factorial moments. Here, we are able to obtain almost sure limit results using Martingale theory. Third, Theorem 2.14 deals with the evolution of the degrees of fixed vertices in the initial graph. Here, we obtain almost sure as well as $\mathcal{L}^r$-limit results. The proofs of Theorems 2.9 and 2.14 are found in Section 5.

Remark 2.6 (Notation). In our Theorems, for sequences $a_1, a_2, ...$ and $b_1, b_2, ...$, we will write $a_n \overset{n\to\infty}{\sim} b_n$ iff $a_n/b_n \xrightarrow{n\to\infty} 1$. The Gamma-function is denoted $t \mapsto \Gamma(t) := \int\limits_0^\infty x^{t-1}e^{-x}dx$. Empty products, i.e. products of the form $\prod_{i=1}^0 f(i)$, are defined to be 1.

Theorem 2.7 (Frequency of isolated vertices)**.** *Let p^* be the (unique) solution of $pe^p = 1$ (or $p + \log p = 0$). Then, the following dichotomy holds:*

(1) For $p \le p^*$, it holds that $\sup_{q\in[0,1]} |H_q^\circ(n) - 1| \xrightarrow{n\to\infty} 0$ almost surely. In particular, for $q = 0$, we have that $F_0^\circ(n) \xrightarrow{n\to\infty} 1$, i.e. the proportion of isolated vertices converges to 1.

(2) For $p > p^*$, it holds that $\mathbf{E}[H_q^\circ(n)] \xrightarrow{n\to\infty} x_\infty < 1$ for all $q \in [0, 1)$ (and in particular $\mathbf{E}[F_0^\circ(n)] \xrightarrow{n\to\infty} x_\infty$) with

$$x_\infty := 1 - \Big(1 - \tfrac{1}{p}\log\big(\tfrac{1}{p}\big)\Big) \cdot \sum_{k=1}^\infty \frac{S_k^\circ(n_0)}{k!}(-1)^{k-1} \prod_{\ell=1}^{k-1}\Big(1 - \frac{1-p^\ell}{p\ell}\Big).$$

Remark 2.8 (Connections to work by Bebek et al (2006)).

(1) Previously, it has been known that $F_0^\circ(n) \xrightarrow{n\to\infty} 1$ for $p < 0.5$; see e.g. Lemma 2 in Bebek et al. (2006b). More precisely, as explained in the same paper, and as a consequence of Theorem 2.9 below, the expected number of neighbors of a randomly chosen node converges to 0 for $p < 0.5$. Theorem 2.7 extends the range for which $F_0^\circ(n)$ converges to 1 to the range $p \le p^*$. Interestingly, this number already appeared in the analysis of PDn in a different context; see below Theorem 1 in Chung et al. (2004) (which we recall in Remark 3.2).

(2) Consider the expected degree distribution $(\mathbf{E}[F_k^\circ(n)])_{k=0,1,2,...}$ for large n. It is a well-known consequence of a fact (usually attributed to Paul Lévy) that a weak limit for such a sequence of distributions as $n \to \infty$ exists if and only if the probability generating functions $x \mapsto \mathbf{E}[H_x^\circ(n)]$ converge to a function h which is continuous at $x = 1$. The limiting distribution then has h as its probability generating function. As the Theorem shows, only for $p \le p^*$ such a convergence

holds and $h = 1$. This implies that the degree distribution converges to δ_0 for $p \le p^*$ and there is no limiting degree distribution for $p > p^*$.

Bebek et al. (2006a) call a distribution $(f_k^\circ)_{k=0,1,2,\dots}$ defective if $f_0^\circ + f_1^\circ + \dots < 1$ and non-defective if $f_0^\circ + f_1^\circ + \dots = 1$. They also raise the question of a critical value for p which separates defective from non-defective limits of $(\mathbf{E}[F_k^\circ(n)])_{k=0,1,2,\dots}$. As Theorem 2.7 shows, the (vague) limit of $(\mathbf{E}[F_k^\circ(n)])_{k=0,1,2,\dots}$ is non-defective only for $p \le p^*$ and defective otherwise. In particular, we have resolved a question raised in Bebek et al. (2006b), since we have in fact shown that there is no limiting (probability) distribution for $(\mathbf{E}[F_k^\circ(n)])_{k=0,1,2,\dots}$ in the case $p > p^*$.

(3) Furthermore, considering any generalization of the partial duplication model producing additional edges (e.g. the Pastor–Satorras et al model), by a suitable coupling argument and Theorem 2.7 it follows immediately that the limiting degree distribution is bound to be defective if the edge retaining probability is greater than p^*. Fueling the duplication mechanism with more edges, might also generate a defective limit for even lower values than p^*.

(4) Frequently in the literature on several random graph models, power laws for the (expected) degree distributions are found; see e.g. Albert and Barabási (2002). In mathematical terms, let $(F_k^\circ(n))_{k=0,1,2,\dots}$ be the degree distribution for some random graph at time n. We say that a power-law for $n \to \infty$ with exponent b holds, if for some $c > 0$,

$$\lim_{k\to\infty} k^b \lim_{n\to\infty} \mathbf{E}[F_k^\circ(n)] = c.$$

In this sense, a power law does not exist for the PDn model since for all $k > 0$ we have shown that $\lim_{n\to\infty} \mathbf{E}[F_k^\circ(n)] = 0$. This observation was also made by Bebek et al. (2006b), who argue that the proof for the power law behavior of PDn given in Chung et al. (2004) is false. We come back to this proof in Remark 3.2. Note, however, that for $p \le p^*$ it is still possible that the connected component of G_n satisfies a power law, i.e. there are $b, c > 0$ with

$$\lim_{k\to\infty} k^b \lim_{n\to\infty} \mathbf{E}\left[\frac{F_k(n)}{\sum_{\ell=1}^\infty F_\ell(n)}\right] = c.$$

Actually, the preferential attachment model arising for small p as explained in Remark 2.5.2, and that model being known to satisfy a power-law (see e.g. Albert and Barabási, 2002), supports this conjecture. Although this power law behavior of the connected component has already been discussed in Ispolatov et al. (2005), care must be taken in order to provide a rigorous result. We defer the deeper analysis of the connected component to future research.

Theorem 2.9 (Cliques, stars). *(1) Let $C_k(n_0) > 0$ and $\mathcal{F}_\infty := \sigma(G_n; n \ge n_0)$. Then, there is an $\mathcal{F}_\infty$-measurable random variable $C_k(\infty)$ with $\mathbf{P}(C_k(\infty) > 0) > 0$, such that*

$$n^{-kp^{k-1}} C_k(n) \xrightarrow{n\to\infty} C_k(\infty), \tag{2.1}$$

almost surely and in $\mathcal{L}^2$ for each k. Moreover,

$$\mathbf{E}[C_k(n)] = C_k(n_0) \cdot \prod_{m=n_0}^{n-1} \frac{m + kp^{k-1}}{m} \overset{n\to\infty}{\sim} \frac{C_k(n_0)\cdot\Gamma(n_0)}{\Gamma(n_0 + kp^{k-1})} n^{kp^{k-1}}. \tag{2.2}$$

(2) For $S_k(n)$, note that $S_1(n) = 2C_2(n)$, such that the asymptotics for S_1 can be read off from the asymptotics of C_2. In addition, for each $k \geq 2$ there is an $\mathcal{F}_\infty$-measurable random variable $S_k(\infty)$, such that

$$n^{-(kp+p^k)} S_k(n) \xrightarrow{n\to\infty} S_k(\infty) \tag{2.3}$$

almost surely. Moreover,

$$\begin{aligned} \mathbf{E}[S_2(n)] &= \Big(S_2(n_0) + \frac{2}{p} S_1(n_0)\Big) \prod_{k=n_0}^{n-1} \frac{k+2p+p^2}{k} - S_1(n_0) \frac{2}{p} \prod_{k=n_0}^{n-1} \frac{k+2p}{k} \\ &\overset{n\to\infty}{\sim} \Big(S_2(n_0) + \frac{2}{p} S_1(n_0)\Big) \frac{\Gamma(n_0)}{\Gamma(n_0+2p+p^2)} n^{2p+p^2}. \end{aligned} \tag{2.4}$$

Remark 2.10 (Dependence on the initial graph). The $C_k(n)$ demonstrate a topological discrepancy between duplication based and preferential attachment models: In preferential attachment often the number of edges added to the graph in one time step is bounded by some constant m. Thus, the formation of new $(m+2)$-cliques is impossible, while the emergence of k-cliques with $k \leq m+1$ is not – both independent of the initial graph. In contrast to this, in the PDn new k-cliques occur if and only if there is at least one k-clique in the initial graph.

Also, Theorem 2.7.2 shows that the limits of the degree distribution for $p > p^*$ very well depend on the initial values $S_k^\circ(n_0)$ as opposed to many limiting results for preferential attachment.

Remark 2.11 (Moments of the degree distribution). Using $\left\{ {n \atop k} \right\}$, the number of partitions of $\{1, ..., n\}$ into k nonempty sets, also known as the Stirling numbers of the second kind, we can write the moments of the degree distribution as

$$M_\ell(n) := \sum_{k\geq 0} k^\ell F_k^\circ(n) = \sum_{k\geq 0} F_k^\circ(n) \sum_{m=0}^{\ell} \left\{ {\ell \atop m} \right\} k_{\downarrow m} = \sum_{m=0}^{\ell} \left\{ {\ell \atop m} \right\} S_m^\circ(n),$$

where $\left\{ {\ell \atop \ell} \right\} = 1$. By (2.3) we obtain $M_\ell(n) \sim S_\ell^\circ(n)$ almost surely and thus immediate limiting results for the moments.

Remark 2.12 (Critical values and $\mathcal{L}^1$-convergence).

(1) It is a simple consequence of Theorem 2.9, that several critical values exist which distinguish cases for $C_k^\circ(n)$, the relative frequencies of k-cliques, and $S_k^\circ(n)$, the relative frequencies of k-stars, converging to 0 or diverging. Precisely, we obtain for C_k°

$$C_k^\circ(n) \xrightarrow{n\to\infty} \begin{cases} \infty, & \text{if } p > k^{-1/(k-1)}, \\ 0, & \text{if } p < k^{-1/(k-1)}. \end{cases}$$

For the limiting case $p := k^{-1/(k-1)}$, we obtain that $C_k^\circ(n) \xrightarrow{n\to\infty} C_k(\infty)$. Analogously, for S_k°, and for the (unique) solution p_k (in $[0,1]$) of $pk + p^k = 1$,

$$S_k^\circ(n) \xrightarrow{n\to\infty} \begin{cases} \infty, & \text{if } p > p_k, \\ 0, & \text{if } p < p_k. \end{cases}$$

Surprisingly, none of those critical values equal p^* from Theorem 2.7.

There might be a connection to the p_k though: Assume that the moments $M_x(n) := \sum k^x F_k^\circ(n)$ satisfy $M_x(n) \sim n^{px+p^x-1}$ not only for $x \in \mathbb{N}$ (as follows from Remark 2.11), but also for each $x \in \mathbb{R}^+$. For $p \geq p^*$, we see that the inequality $px + p^x \geq px + e^{-px} > 1$ holds for all $x > 0$, while for $p < p^*$ there are $x > 0$ with $px + p^x < 1$. Thus, p^* is the smallest value of $p \in [0,1]$ such that there is no positive solution x of $px + p^x = 1$. Hence, all $M_x(n)$ tend to ∞ if and only if $p > p^*$.

(2) Unfortunately, we are not able to show that the convergence in (2.3) also holds in $\mathcal{L}^1(\mathcal{F}_\infty)$ if $k \geq 2$ and thus we cannot rule out the possibility that $S_k(\infty)$ is trivial, i.e. we cannot rule out $S_k(\infty) = 0$. It should be possible to use a technique similar to the proof of the $\mathcal{L}^2$ convergence of the $C_k(n)$, but it is much more difficult since additionally there are three possible relations of the centers c_1, c_2 of two k-stars s_1 and s_2:

$$a)\ c_1 = c_2, \qquad b)\ (c_1, c_2) \in E_n, \qquad c) \text{ neither } a) \text{ nor } b),$$

where each of those as well as the number of shared nodes or edges influence the evolutions of k-star pairs in a different way. Since the structures of pairs of k-stars are so complex, another approach might be more suitable in order to show non-triviality of $S_k(\infty)$.

Remark 2.13 (Transitivity ratio). The transitivity ratio $Tr(G)$ of a graph $G = (V, E)$ is defined via $C_3(G)$ and $S_2(G)$ by

$$Tr(G) := \frac{6C_3(G)}{S_2(G)}.$$

(Precisely, it is defined by the quotient of three times the number of triangles $C_3(G)$ and the number of connected triples, i.e. the number of triples $v, w, u \in V$ with $\{v, w\}, \{w, u\} \in E$. Each connected triple is counted twice by $S_2(G)$ upon summing over vertex w.) Hence, we find that

$$\frac{\mathbf{E}[6C_3(n)]}{\mathbf{E}[S_2(n)]} \overset{n\to\infty}{\sim} \frac{6C_3(n_0)}{S_2(n_0) + \frac{2}{p}S_1(n_0)} \frac{\Gamma(n_0 + 2p + p^2)}{\Gamma(n_0 + 3p^2)} n^{-2p(1-p)}.$$

Moreover, $n^{2p(1-p)}Tr(n)$ converges (at least on the set $S_2(\infty) \neq 0$) to some integrable random variable by Theorem 2.9.

Theorem 2.14 (Degree evolution of the initial vertices)**.** *Let $V_{n_0} = \{1, ..., n_0\}$, i.e. we number the initial vertices by $1, ..., n_0$. In addition, let $D_k(n) > 0$ be the degree of vertex $k \leq n_0$ at time n. Then, for $n \geq n_0$ and $\ell \geq a$,*

$$\mathbf{P}(D_k(n) = \ell | D_k(n_0) = a) = \sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell - 1}{m - 1} \binom{m - 1}{a - 1} \prod_{j=n_0}^{n-1} \left(1 - \frac{pm}{j}\right). \tag{2.5}$$

Moreover, there is an almost surely positive random variable $D_k(\infty)$, such that

$$n^{-p} D_k(n) \xrightarrow{n\to\infty} D_k(\infty) \tag{2.6}$$

almost surely and in $\mathcal{L}^r$ for each $r \geq 1$ and, using $\ell_{\uparrow m} := \ell \cdots (\ell + m - 1)$ for $m = 1, 2, \ldots$,

$$\mathbf{E}[n^{-mp} D_k(n)_{\uparrow m}] = \frac{D_k(n_0)_{\uparrow m}}{n^{mp}} \prod_{\ell=n_0}^{n-1} \tfrac{\ell+mp}{\ell} \xrightarrow{n\to\infty} \frac{D_k(n_0)_{\uparrow m} \cdot \Gamma(n_0)}{\Gamma(n_0 + mp)} = \mathbf{E}[D_k(\infty)^m]. \tag{2.7}$$

Remark 2.15 (Degree evolution of arbitrary vertices). In the case of $k > n_0$ we can obtain results for the behavior of $D_k(n)$ by conditioning on the graph at time k, i.e. considering G_k as initial graph.

Remark 2.16 (Connection to the Pólya urn). In the special case $p = 1$, the sequence $(D_k(n))_{n=n_0,n_0+1,\ldots}$ is connected to Pólya's urn. Note that the degree of vertex v_k increases by one at time n iff one of the neighbors of v_k is copied. In Pólya's urn, start with n_0 balls, where $D_k(n_0) = a$ balls are red and all others are black. Then, as usual, pick a ball from the urn at random, and put it back together with a second ball of the same color. From this construction, $D_k(n)$ is equal in distribution to the number of red balls when there are n balls in the urn for all n. Of course, it is well-known that in this case, the probability that there are ℓ red balls in the urn at the time when there are n balls in total equals

$$\binom{n-n_0}{\ell-a}\frac{a_{\uparrow(\ell-a)}\cdot(n_0-a)_{\uparrow(n-n_0-\ell+a)}}{n_{0\uparrow(n-n_0)}}, \tag{2.8}$$

with $m_{\uparrow k} := m\cdot(m+1)\cdots(m+k-1)$ (e.g. (4.2) in Johnson and Kotz, 1977). Using (5) of Chapter 1 of Riordan (1979) we obtain

$$\begin{aligned}\binom{n-n_0}{\ell-a} &= \binom{(n-a-1)-(n_0-a-1)}{\ell-a}\\ &= \sum_{m=0}^{\ell-a}(-1)^m\binom{n-a-1-m}{\ell-a-m}\binom{n_0-a-1}{m}.\end{aligned}$$

From this it follows, that (2.8) equals the right hand side of (2.5) if $p = 1$.
Also section 6.3.3 in Johnson and Kotz (1977) shows that the proportion of red balls in the urn converges to a β-distributed random variable with parameters a and $n_0 - a$. Therefore it is not surprising that for $p = 1$ the moments of $D_k(\infty)$ as given in (2.7) match those of the $\beta(a, n_0 - a)$-distribution.
The connection of (2.5) to an extension of Pólya's urn would look as follows: Consider an urn, starting with n_0 balls, a of which are red and $n_0 - a$ of which are black. In each step, choose a ball at random from the urn. If the ball is black, put it back to the urn together with another black ball. If the ball is red, put it back to the urn together with another ball. The color of the additional ball is red with probability p and black with probability $1 - p$. Then, the chance that there are ℓ red balls in the urn at the time when there are a total of n balls in the urn equals the right hand side of (2.5).

Remark 2.17 (The limiting distribution). The limiting distribution with moments given by the right hand side of (2.7) seems to be not well–known. Thus far, we deduced, using Stirling's formula, that for $p < 1$

$$\|D_k(\infty)\|_{\mathcal{L}^\infty} = \lim_{m\to\infty}\|D_k(\infty)\|_{\mathcal{L}^m} = \lim_{m\to\infty}\left(\frac{\Gamma(D_k(n_0)+m)}{\Gamma(n_0+pm)}\right)^{1/m} = \infty,$$

which shows that (in contrast to the special case $p = 1$) $D_k(\infty)$ is not bounded. Also, Stirling's formula shows that the moments satisfy

$$\sum_{m=1}^{\infty}\mathbf{E}[D_k(\infty)^m]^{-1/(2m)} = \infty,$$

which is Carleman's condition for the determinacy of the corresponding Stieltjes moment problem (e.g. Shohat and Tamarkin, 1943, Theorem 1.11). Thus, the limiting distribution is defined by its moments.

3. Preparation

3.1. *Some recursions.* We collect some simple calculations in this section. Throughout, we denote by $(\mathcal{F}_n)_{n=n_0,n_0+1,\dots}$ the filtration generated by $\mathcal{G} = (G_n)_{n=n_0,n_0+1,\dots}$.

Proposition 3.1 (Evolution of F, H, S, C)**.** *It holds that*

$$\begin{aligned}\mathbf{E}[F_k(n+1)|\mathcal{F}_n] &= F_k(n) + p(k-1)F_{k-1}^\circ(n) - pkF_k^\circ(n) \\ &\qquad + \sum_{\ell\geq k} F_\ell^\circ(n)\binom{\ell}{k}p^k(1-p)^{\ell-k}, \end{aligned} \tag{3.1}$$

$$\mathbf{E}[H_q(n+1)|\mathcal{F}_n] = H_q(n) - pq(1-q)\frac{d}{ds}H_s^\circ(n)\Big|_{s=q} + H_{1-p+pq}^\circ(n), \tag{3.2}$$

$$\mathbf{E}[S_k(n+1)|\mathcal{F}_n] = \Big(1+\frac{pk+p^k}{n}\Big)S_k(n) + \frac{pk(k-1)}{n}S_{k-1}(n), \tag{3.3}$$

$$\mathbf{E}[C_k(n+1)|\mathcal{F}_n] = C_k(n)\Big(1+\frac{k}{n}p^{k-1}\Big). \tag{3.4}$$

Proof: Let us start with (3.1). The quantity F_k increases in two cases: either, a vertex of degree $\ell \geq k$ is copied, together with k edges (which has probability $\binom{\ell}{k}p^k(1-p)^{\ell-k}$), or one of the neighbors of a vertex of degree $k-1$ is copied together with the connecting edge. On the other hand, F_k decreases by one, if one of the neighbors of a vertex of degree k is copied together with the connecting edge. These three cases make up the right hand side of (3.1).
For (3.2), recall the definition of H_q. We multiply (3.1) by q^k and sum in order to obtain

$$\begin{aligned}\mathbf{E}[H_q(n+1)-H_q(n)|\mathcal{F}_n] &= pq^2\sum_{k=1}^\infty (k-1)F_{k-1}^\circ(n)q^{k-2} - pq\sum_{k=0}^\infty kF_k^\circ(n)q^{k-1} \\ &\qquad + \sum_{k=0}^\infty\sum_{\ell=k}^\infty F_\ell^\circ(n)\binom{\ell}{k}p^k(1-p)^{\ell-k}q^k \\ &= -pq(1-q)\frac{d}{ds}\Big(\sum_{k=0}^\infty F_k^\circ(n)s^k\Big)\Big|_{s=q} + \sum_{\ell=0}^\infty F_\ell^\circ(n)(1-p+pq)^\ell \\ &= -pq(1-q)\frac{d}{ds}H_s^\circ(n)\Big|_{s=q} + H_{1-p+pq}^\circ(n).\end{aligned}$$

We now turn to (3.3). Again, use (3.1), multiply by $k\cdots(k-m+1) =: k_{\downarrow m}$ and sum for

$$\begin{aligned}&\mathbf{E}[S_m(n+1)-S_m(n)|\mathcal{F}_n] \\ &= \sum_{k=m}^\infty\Big(pk_{\downarrow m}(k-1)F_{k-1}^\circ(n) - pk_{\downarrow m}kF_k^\circ(n) + \sum_{\ell=k}^\infty F_\ell^\circ(n)k_{\downarrow m}\binom{\ell}{k}p^k(1-p)^{\ell-k}\Big) \\ &= p\sum_{k=m}^\infty((k+1)_{\downarrow m} - k_{\downarrow m})kF_k^\circ(n)\end{aligned}$$

$$
\begin{aligned}
&+ p^m \sum_{\ell=m}^{\infty} \sum_{k=m}^{\ell} F_\ell^\circ(n) \ell_{\downarrow m} \binom{\ell - m}{k - m} p^{k-m} (1-p)^{\ell-k} \\
&= p \sum_k \big(m k_{\downarrow m} + m(m-1) k_{\downarrow(m-1)} \big) F_k^\circ(n) + p^m \sum_\ell F_\ell^\circ(n) \ell_{\downarrow m} \\
&= \frac{pm + p^m}{n} S_m(n) + \frac{pm(m-1)}{n} S_{m-1}(n),
\end{aligned}
$$

where we have used that

$$
\begin{aligned}
k \cdot &\big((k+1)_{\downarrow m} - k_{\downarrow m} \big) \\
&= \big(k - m + 1 + (m-1) \big) \cdot k_{\downarrow(m-1)} \big(k + 1 - (k - m + 1) \big) \\
&= m k_{\downarrow m} + m(m-1) k_{\downarrow(m-1)}.
\end{aligned}
$$

For (3.4), a k-clique arises if a vertex v which is member of a k-clique is copied, together with all $k-1$ edges connecting v to the other members of the clique. Hence,

$$
\mathbf{E}[C_k(n+1) | \mathcal{F}_n] = C_k(n) \Big(1 + \frac{k}{n} p^{k-1} \Big).
$$

□

Remark 3.2 (Scale-free property). (1) In Chung et al. (2004), the authors show the following: If for some $b > 0$ (necessarily we will have $b > 1$) and $c > 0$ it holds that

$$
\lim_{k \to \infty} k^b \lim_{n \to \infty} \mathbf{E}[F_k^\circ(n)] = c, \tag{3.5}
$$

then, b must satisfy $p(b-1) = 1 - p^{b-1}$.

Let us briefly recall the arguments leading to this power-law behavior of the (expected) degree distribution. Starting off with (3.1), taking expectations on both sides, and setting $\mathbf{E}[F_k(n)] = ck^{-b} n + o(n)$ for some c, b, we see that for $n \to \infty$, if a stationary state for $\mathbf{E}[F_k^\circ]$ is reached,

$$
ck^{-b} \overset{k \to \infty}{\sim} p(k-1)c(k-1)^{-b} - pkck^{-b} + c \sum_{\ell \geq k} \ell^{-b} \binom{\ell}{k} p^k (1-p)^{\ell-k}.
$$

Note that $k(1 - 1/k)^{-b} - k \overset{k \to \infty}{\sim} b$, and (see Lemma 2 in Chung et al., 2004)

$$
\sum_{\ell \geq k} \ell^{-b} \binom{\ell}{k} p^k (1-p)^{\ell-k} \overset{k \to \infty}{\sim} k^{-b} p^{b-1}.
$$

Therefore, by dividing by ck^{-b}, the parameter b must satisfy

$$
1 = pb - p + p^{b-1}, \tag{3.6}
$$

the desired relationship.

Of course, with this proof Chung et al. (2004) only show an assertion about the scaling exponent b in the case that the limiting distribution of $(\mathbf{E}[F_k^\circ(n)])_{k=0,1,2,\ldots}$ satisfies a power law. No assertion is made if such a power law exists. In order to resolve this, consider $p \geq p^*$ as given in Theorem 2.7, that is $p \geq e^{-p}$. For such p the inequality $px + p^x \geq px + e^{-px} > 1$ holds for all $x > 0$. More precisely, the desired relationship has a solution $b > 1$ if and only if $p < p^*$. However, in this case we have seen that $(\mathbf{E}[F_k^\circ(n)]) \xrightarrow{n\to\infty} \delta_{k0}$ and so (3.5) cannot hold. We conclude that no power–law behavior is possible.

(2) The works on the Pastor–Satorras et al modification we mentioned in Remark 2.5 claim the same power law (3.6) to hold, again with $p(b-1) = 1 - p^{b-1}$, irrespective of r as long as $r > 0$ (see Theorem 1 in Bebek et al., 2006b using arguments similar to those of Chung et al., 2004 and (18) in Kim et al., 2002, where the connection gets clear by multiplying $1-\delta$ and substituting $p = 1-\delta$). It is also recognized that such a power law can only exist for $p \le p^*$. However, the proofs of this power law only show stationarity of a degree distribution with power law. The question of convergence in the sense of (3.5) still is an open problem.

3.2. *An auxiliary process.* In the proof of Theorem 2.7, we will need a piece-wise deterministic process which we introduce here. There, we will obtain and use a *duality* (see Subsection 4.2), i.e. a relationship of the form

$$\mathbf{E}[H_{1-x}(t)] = \mathbf{E}[H_{1-X_t}(0)|X_0 = x]$$

for continuous-time versions of the probability generating functions of the degree distributions H and a $[0,1]$-valued process $\mathcal{X} = (X_t)_{t\ge0}$, which jumps from x to px at rate 1 and in between jumps follows the logistic equation $\dot{X} = pX(1-X)$. Recall that such piece-wise deterministic processes have been studied recently in more detail; see e.g. Davis (1984), Costa and Dufour (2008), Azaïs et al. (2014).

Lemma 3.3 (The auxiliary process $\mathcal{X}$)**.** *Let $p \in [0,1]$ and $\mathcal{X} = (X_t)_{t\ge0}$ be a Markov process with state space $[0,1]$ and generator*

$$G_{\mathcal{X}} f(x) = px(1-x)f'(x) + (f(px) - f(x)) \tag{3.7}$$

for $f \in \mathcal{C}_b^1([0,1])$ and $X_0 \in [0,1]$. In addition, let $p^* \approx 0.567143$ be the unique solution of $pe^p = 1$ (or $p + \log p = 0$).

Then, if $p \le p^*$, it holds that $X_t \xrightarrow{t\to\infty} 0$ almost surely, whereas if $p > p^*$ it holds that $\mathcal{X}$ is ergodic and $X_t \xRightarrow{t\to\infty} X_\infty$ for some $[0,1]$-valued random variable X_∞ with $\mathbf{P}(X_\infty > 0) = 1$ and

$$\mathbf{E}[X_\infty^k] = \Big(1 - \tfrac{1}{p}\log\big(\tfrac{1}{p}\big)\Big) \cdot \prod_{\ell=1}^{k-1}\Big(1 - \frac{1-p^\ell}{p\ell}\Big).$$

Proof: We consider the process $-\log\mathcal{X} = (-\log X_t)_{t\ge0}$ with state space $[0,\infty)$. From (3.7), we read off that this process has the generator

$$G_{-\log\mathcal{X}} g(y) = -p(1-e^{-y})g'(y) + g(y + \log(1/p)) - g(y).$$

In other words, $-\log\mathcal{X}$ decreases at rate $p(1-e^{-y})$ at time t if $-\log X_t$ equals y and increases by $\log(1/p)$ at the times of a Poisson process. Note that $X_t \xrightarrow{t\to\infty} 0$ iff $-\log X_t \xrightarrow{t\to\infty} \infty$.

We start with the case $p < p^*$. Here, we can couple the process $-\log\mathcal{X}$ with a process $\mathcal{U} = (U_t)_{t\ge0}$ with generator

$$G_{\mathcal{U}} g(y) = -pg'(y) + g(y + \log(1/p)) - g(y)$$

by using the same Poisson processes for $-\log\mathcal{X}$ and $\mathcal{U}$. Since $1 - e^{-y} \le 1$, we have that $U_t \le -\log X_t$. However, we can write $\mathcal{U}$ as

$$U_t = U_0 - pt + \log(1/p)P_t$$

for some unit-rate Poisson process $\mathcal{P} = (P_t)_{t\geq 0}$ and by the law of large numbers for Poisson processes (i.e. $\frac{P_t}{t} \xrightarrow{t\to\infty} 1$ almost surely), we see that $U_t \xrightarrow{t\to\infty} \infty$ almost surely, if $\log(1/p) > p$ or $p < p^*$. Since $U_t \leq -\log X_t$, this implies $-\log X_t \xrightarrow{t\to\infty} \infty$ or $X_t \xrightarrow{t\to\infty} 0$, as claimed.

Now, we turn to the case $p > p^*$. First, we have to prove ergodicity of $-\log \mathcal{X}$ (which is equal to ergodicity of $\mathcal{X}$). Let $T_z := T_z^{-\log \mathcal{X}} := \inf\{t \geq 0 : -\log X_t = z\}$. According to Davis (1993), Theorem 3.10, we have to show that (i) there is $z \geq 0$ such that $\mathbf{E}[T_z | -\log X_0 = z] < \infty$ and (ii) $\mathbf{P}(T_z < \infty | -\log X_0 = x) = 1$ for all $x \geq 0$.

Let z be large enough such that

$$p_z := p(1 - e^{-z}) > \log(1/p).$$

We define

$$S_{(z,z+\log(1/p)]} := \inf\{t : z < -\log X_t \leq z + \log(1/p)\}.$$

Then, $\mathbf{E}[S_{(z,z+\log(1/p)]} | -\log X_0 = x] < \infty$ for all $x \leq z$. Indeed, the probability for at least $z/\log(1/p)$ jumps in some small time interval of length $\varepsilon > 0$ is positive. After the first such time interval we can be sure that $S_{(z,z+\log(1/p)]}$ has occurred. By finitness of first moments of geometric distributions, $\mathbf{E}[S_{(z,z+\log(1/p)]} | -\log X_0 = x] < \infty$ follows. By a restart argument, we have to show that $\mathbf{E}[T_z | -\log X_0 = x] < \infty$ for all $z < x \leq z + \log(1/p)$, which will be done by using a comparison argument. For this, let $\mathcal{R} = (R_t)_{t\geq 0}$ be a process with generator

$$G_{\mathcal{R}} g(y) = -p_z g'(y) + g(y + \log(1/p)) - g(y).$$

If $z < R_0 = -\log X_0 \leq z + \log(1/p)$, then – using the same Poisson processes for $-\log \mathcal{X}$ and $\mathcal{R}$ – we have that $T_z \leq T_z^{\mathcal{R}} := \inf\{t \geq 0 : R_t = z\}$ since $p(1 - e^{-y}) \geq p_z$ for $y \geq z$. Since $(R_t - R_0 + t(p_z - \log(1/p)))_{t\geq 0}$ is a martingale and $T_z^{\mathcal{R}} < \infty$ almost surely, we have by optional stopping that $\mathbf{E}[R_0 - R_{T_z^{\mathcal{R}}}] = R_0 - z = (p_z - \log(1/p))\mathbf{E}[T_z^{\mathcal{R}}]$, hence $\mathbf{E}[T_z | -\log X_0 = x] \leq \log(1/p)/(p_z - \log(1/p)) < \infty$. It is now straight-forward to obtain the properties (i) and (ii) and we see that $-\log \mathcal{X}$ is ergodic. In particular, $-\log X_\infty < \infty$, i.e. $X_\infty > 0$ almost surely.

By the ergodic Theorem, we have that $\frac{1}{t}\int_0^t 1 - X_s ds \xrightarrow{t\to\infty} 1 - \mathbf{E}[X_\infty]$. This can be used when we study the martingale $\big(-\log X_t + \log X_0 - \int_0^t \log(1/p) - p(1 - X_s)ds\big)_{t\geq 0}$. By dividing by t and ergodicity, we see that

$$\begin{aligned} 0 &= \lim_{t\to\infty} \frac{1}{t}\log(X_0/X_t) - \frac{1}{t}\int_0^t \log(1/p) - p(1 - X_s)ds \\ &= -\log(1/p) + p(1 - \mathbf{E}[X_\infty]), \end{aligned}$$

i.e.

$$\mathbf{E}[X_\infty] = 1 - \tfrac{1}{p}\log(1/p).$$

Now, since $\mathbf{E}[G_{-\log \mathcal{X}} f(X_\infty)] = 0$, we find that for $f(x) = e^{-kx}$

$$\begin{aligned} 0 &= \mathbf{E}[kp(1 - X_\infty)X_\infty^k + X_\infty^k(p^k - 1)] \\ &= -pk\mathbf{E}[X_\infty^{k+1}] + (pk + p^k - 1)\mathbf{E}[X_\infty^k] \end{aligned}$$

or

$$\mathbf{E}[X_\infty^{k+1}] = \frac{pk + p^k - 1}{pk}\mathbf{E}[X_\infty^k] = \Big(1 - \frac{1 - p^k}{pk}\Big)\mathbf{E}[X_\infty^k].$$

By induction, we see that

$$\mathbf{E}[X_\infty^k] = \Big(1 - \tfrac{1}{p}\log\big(\tfrac{1}{p}\big)\Big) \cdot \prod_{\ell=1}^{k-1}\Big(1 - \frac{1-p^\ell}{p\ell}\Big).$$

Last, we consider the case $p = p^*$. Let $X_t^{(p)}$ be the Markov process with generator (3.7) for a specific value of p. If $p \mapsto X_0^{(p)}$ is constant, we can couple these processes by using the same jump times such that $X_t^{(p)} \le X_t^{(p')}$ for $p < p'$. Therefore,

$$\begin{aligned} 0 &\le \mathbf{E}[\limsup_{t\to\infty} X_t^{(p^*)}] \le \inf_{p>p^*} \mathbf{E}[\limsup_{t\to\infty} X_t^{(p)}] = \inf_{p>p^*} \mathbf{E}[X_\infty^{(p)}] \\ &= \inf_{p>p^*}\Big(1 - \tfrac{1}{p}\log\Big(\tfrac{1}{p}\Big)\Big) = 0. \end{aligned}$$

Hence, $\limsup_{t\to\infty} X_t^{(p^*)} = \lim_{t\to\infty} X_t^{(p^*)} = 0$, almost surely. □

3.3. *Martingales, the Gamma function and a recursion.* We prepare some facts needed in the proofs of Theorems 2.9 and 2.14.

Lemma 3.4 (Asymptotics for the Gamma function)**.** *Let $n_0 \ge 0$ and $a > -t_0$. Then,*

$$\prod_{k=n_0}^{n-1} \frac{k+a}{k} = \frac{\Gamma(n+a)}{\Gamma(n)} \cdot \frac{\Gamma(n_0)}{\Gamma(n_0+a)} \overset{n\to\infty}{\sim} \frac{n^a\Gamma(n_0)}{\Gamma(n_0+a)}.$$

Proof: The first identity follows by iterating the functional equation $x\Gamma(x) = \Gamma(x+1)$ and for the asymptotics see e.g. Abramowitz and Stegun (1992), 6.1.46. □

Lemma 3.5 (Martingale estimates)**.** *Let $\mathcal{X} = (X_n)_{n=n_1,n_0+1,\dots}$ be a non-negative, integrable stochastic process, adapted to a filtration $\mathcal{F} := (\mathcal{F}_n)_{n=n_0,n_0+1,\dots}$, $\mathcal{F}_\infty := \sigma\Big(\bigcup_{n=n_0}^{\infty} \mathcal{F}_n\Big)$ and $x_0 := \mathbf{E}[X_{n_0}] > 0$. Moreover, let $a > -n_0$ and assume that*

$$\mathbf{E}\left[X_{n+1}|\mathcal{F}_n\right] = \Big(1 + \frac{a}{n}\Big)X_n$$

for all $n = n_0, n_0+1, \dots$ Then, the following holds:

(1) The process $\mathcal{M} = (M_n)_{n\ge n_0}$ defined by $M_{n_0} = X_{n_0}$ and

$$M_n = X_n \cdot \prod_{k=n_0}^{n-1} \frac{k}{k+a}$$

is an $\mathcal{F}$-martingale and the expectations of the X_n hold

$$\mathbf{E}[X_n] = x_0 \cdot \prod_{k=n_0}^{n-1} \frac{k+a}{k} \overset{n\to\infty}{\sim} \frac{x_0\Gamma(n_0)}{\Gamma(n_0+a)} \cdot n^a. \tag{3.8}$$

(2) There is a non-negative random variable $X_\infty \in \mathcal{L}^1(\mathcal{F}_\infty)$ with $\mathbf{E}[X_\infty] \le x_0\Gamma(n_0)/\Gamma(n_0+a)$ such that

$$n^{-a}X_n \xrightarrow{n\to\infty} X_\infty \text{ almost surely.}$$

(3) If, in addition to 2., $\mathbf{E}[X_n^r] = O(n^{ar})$ for some $r > 1$, then the convergence also holds in $\mathcal{L}^r$ and thus $\mathbf{P}(X_\infty > 0) > 0$.

Proof: By the assumptions and its definition it is easy to see that $\mathcal{M}$ is a martingale. Thus, the equality in (3.8) follows by induction and the asymptotic expansion is a consequence of Lemma 3.4.

Since $\mathcal{M}$ is non-negative, it converges almost surely to some random variable $M_\infty \in \mathcal{L}^1(\mathcal{F}_\infty)$ with $\mathbf{E}[M_\infty] \le \mathbf{E}[M_{n_0}]$ and hence,

$$n^{-a}X_n = n^{-a}M_n \cdot \prod_{k=n_0}^{n-1} \frac{k+a}{k} \xrightarrow{n\to\infty} M_\infty \cdot \frac{\Gamma(n_0)}{\Gamma(n_0+a)} =: X_\infty$$

almost surely. The convergence is also in $\mathcal{L}^r$ if $\mathcal{M}$ is $\mathcal{L}^r$-bounded. We compute, using $\mathbf{E}[X_n^r] \le cn^{ar}$ and Lemma 3.4,

$$\sup_n \mathbf{E}[M_n^r] = \sup_n \mathbf{E}[X_n^r]\Big(\prod_{k=n_0}^{n-1} \frac{k}{k+a}\Big)^r \le \sup_n \frac{\Gamma(n_0+a)^r cn^{ar}}{\Gamma(n_0)^r c' n^{ar}} < \infty,$$

which shows the assertion. In particular, $\mathcal{M}$ converges in $\mathcal{L}^1$ which gives us $\mathbf{E}[M_\infty] > 0$ concluding the proof. □

Lemma 3.6 (Recursions)**.** *Let $n_0 > 0, a > -n_0, \varepsilon > 0$ and $f, g : \{n_0, n_0+1, \dots\} \to (0,\infty)$, satisfying*

$$f(n+1) = \Big(1 + \frac{a}{n}\Big) f(n) + \frac{g(n)}{n} \tag{3.9}$$

for all $n \ge n_0$. Then, for $n \to \infty$

(1) If $g = O(n^{a-\varepsilon})$, then $f = \Theta(n^a)$.

(2) If $g = O(n^a)$, then $f = O(n^a \log n)$, and if $g = \Omega(n^a)$, then $f = \Omega(n^a \log n)$.

(3) If $g = O(n^{a+\varepsilon})$, then $f = O(n^{a+\varepsilon})$, and if $g = \Omega(n^{a+\varepsilon})$, then $f = \Omega(n^{a+\varepsilon})$.

Proof: At first note that from Lemma 3.4 and the positivity of g we easily obtain $f = \Omega(n^a)$ in any case. Iteration of (3.9) gives us

$$\begin{aligned} f(n) &= f(n_0) \prod_{k=n_0}^{n-1} \Big(1 + \frac{a}{k}\Big) + \sum_{k=n_0}^{n-1} \frac{g(k)}{k} \prod_{\ell=k+1}^{n-1} \Big(1 + \frac{a}{\ell}\Big) \\ &= \prod_{k=n_0}^{n-1} \frac{k+a}{k} \cdot \Bigg(f(n_0) + \sum_{k=n_0}^{n-1} \frac{g(k)}{k} \prod_{m=n_0}^{k} \frac{m}{m+a} \Bigg). \end{aligned} \tag{3.10}$$

Lemma 3.4 provides constants $c_0, c_1, c_2 > 0$ which hold

$$(3.10) \le c_0 n^a \Big(f(n_0) + \sum_{k=n_0}^{n-1} \frac{g(k)}{k} c_1 k^{-a} \Big) \le c_2 n^a \sum_{k=n_0}^{n-1} \frac{g(k)}{k^{a+1}}. \tag{3.11}$$

Now 1. and the first parts of 2. and 3. follow immediately by considering a suitable integral as upper bound for the sum. Lastly, note that Lemma 3.4 also provides constants c_0, c_1, c_2 which satisfy the respective lower bounds in (3.11), such that the remaining claims follow analogously. □

Corollary 3.7. *If, in Lemma 3.6, g is $O(n^b)$, then $f = o\big(n^{\max\{a,b\}+\varepsilon}\big)$ for all $\varepsilon > 0$.*

4. Proof of Theorem 2.7

4.1. *A time-continuous partial duplication graph.* It will be helpful to have a time-continuous version of $\mathcal{G}$.

Definition 4.1 (Partial duplication random graph PDt)**.** Let $p \in [0,1]$. We define the following random graph process – called *time-continuous partial duplication random graph* or PDt graph – $\mathcal{G} = (G_t)_t \geq 0$ with $G_t = (V_t, E_t)$, where G_t is the graph at time t with vertex set V_t and (undirected) edge set $E_t \subseteq \{\{v,w\} : v, w \in V_t, v \neq w\}$. Starting in some $G_0 = (V_0, E_0)$, every $v \in V_t$ gives rise at rate $1 + 1/|V_t|$ to a duplication event. Upon such an event, a new node $v' \notin V_{t-}$ is created and every edge connected to v (i.e. every $e \in E_{t-}$ with $e = \{v, w\}$ for some $w \in V_{t-}$) is copied at time t with probability p, i.e. $\{v', w\} \in E_t$ with probability p, independently of all other edges.

We define as in Definition 2.3 the degree distribution $F_k(t) := F_k(G_t)$ and $F_k^\circ(t) := F_k^\circ(G_t)$ and its probability generating function $H_q(t) := H_q(G_t)$ and $H_q^\circ(t) := H_q^\circ(G_t)$.

Remark 4.2 (Connection between PDn and PDt). (1) We abuse notation here and use $(G_t)_{t\geq 0}$ for the time-continuous PDt graph while $(G_n)_{n=n_0,n_0+1,\dots}$ is the time-discrete PDn graph. Of course, these two processes are closely connected. Let $\tau_n := \inf\{t \geq 0 : |V_t| = n\}$ Then, $(G_{\tau_n})_{n=n_0,n_0+1,\dots} \sim (G_n)_{n=n_0,n_0+1,\dots}$

(2) The choice of the rate $1 + 1/|V_t|$ for initiating a duplication event seems unnatural. It will however turn out that this choice simplifies our line of argument; see the next proposition.

We now derive an important relationship for $H_q^\circ(t)$.

Proposition 4.3 (Evolution of $H_q^\circ(t)$)**.** *For $\mathcal{G} = (G_t)_{t\geq 0}$ and $H^\circ(t)$ as above,*

$$\frac{d}{dt}\mathbf{E}[H_q^\circ(t)] = \mathbf{E}\Big[- pq(1-q)\frac{d}{ds}H_s^\circ(t)\Big|_{s=q} + H_{1-p+pq}^\circ(t) - H_q^\circ(t)\Big].$$

Remark 4.4. Later, it will be useful to define $x := 1 - q$ and $\widetilde{H}_x(t) := H_q(t)$ in order to obtain

$$\frac{d}{dt}\mathbf{E}[\widetilde{H}_x^\circ(t)] = \mathbf{E}\Big[px(1-x)\frac{d}{dx}\widetilde{H}_x^\circ(t) + \widetilde{H}_{px}^\circ(t) - \widetilde{H}_x^\circ(t)\Big]. \tag{4.1}$$

In particular, note that the right hand side is reminiscent of (3.7).

Proof: We have already seen the evolution of $n \mapsto \mathbf{E}[H_q(n)]$ in Proposition 3.1. From this, we derive, since the total rate for a duplication event at time t is $|V_t|+1$,

$$\begin{aligned}\mathbf{E}[H_q(t+dt)] = \mathbf{E}\Big[&H_q(t)(1-(|V_t|+1)dt) \\ &+ dt\cdot(|V_t|+1)\Big(H_q(t) - pq(1-q)\frac{d}{ds}H_s^\circ(t)\Big|_{s=q} + H_{1-q+pq}^\circ(t)\Big)\Big].\end{aligned}$$

From this, we obtain

$$\begin{aligned}\mathbf{E}[H_q^\circ(t+dt)] =&\mathbf{E}\Big[H_q^\circ(t)(1-(|V_t|+1)\cdot dt) + dt\cdot(|V_t|+1)\Big(\frac{|V_t|}{|V_t|+1}H_q^\circ(t) \\ &- \frac{1}{|V_t|+1}pq(1-q)\frac{d}{ds}H_s^\circ(t)\Big|_{s=q} + \frac{1}{|V_t|+1}H_{1-q+pq}^\circ(t)\Big)\Big]\end{aligned}$$

$$= \mathbf{E}\Big[H_q^\circ(t) + dt \cdot \Big(- pq(1-q)\frac{d}{ds}H_s^\circ(t)\Big|_{s=q} + H_{1-q+pq}^\circ - H_q^\circ(t)\Big)\Big].$$

□

4.2. *A duality relationship between $\mathcal{X}$ and $\mathcal{G}$.* Now, we make clear why we need the auxiliary process $\mathcal{X}$ from Subsection 3.2. Here, we borrow ideas from the notion of duality of Markov processes; see Chapter 4.4 in Ethier and Kurtz (1986).

Recall that two Markov processes $\mathcal{G} = (G_t)_{t\geq 0}$ (which will be the PDt-graph below) and $\mathcal{X} = (X_t)_{t\geq 0}$ (which will be the piecewise-deterministic process from Subsection 3.2) with state spaces E and E' are called dual with respect to the function $H : E \times E' \to \mathbb{R}$ if

$$\mathbf{E}[H(G_t, x)|G_0 = g] = \mathbf{E}[H(g, X_t)|X_0 = x] \tag{4.2}$$

for all $g \in E, x \in E'$. (In our application, H will be the moment generating function of the degree distribution of the PDt-graph evaluated at $1 - x$.) When one is interested in the process $\mathcal{G}$, this relationship is most helpful if the process $\mathcal{X}$ is easier to analyse than the process $\mathcal{G}$. Moreover, frequently, the set of functions $\{H(., x) : x \in E'\}$ is separating on E such that the left hand side of (4.2) determines the distribution of G_t. In this case, the distribution of the simpler process $\mathcal{X}$ determines via (4.2) the distribution of $\mathcal{G}$, so analysing $\mathcal{G}$ becomes feasible. (In our application, however, $\{H(., x) : x \in E'\}$ is only separating on the space of degree distributions and hence (4.2) will determine the degree distribution of the PDt-graph.)

There is no straight-forward way how to find dual processes, but they arise frequently in the literature; see Jansen and Kurt (2014) for a survey. Examples span reflected and absorbed Brownian motion, interacting particle models such as the voter model and the contact process, as well as branching processes.

Proposition 4.5 (Duality)**.** *Let $\mathcal{X} = (X_t)_{t\geq 0}$ be a Markov process with state space $[0, 1]$ with generator as given in (3.7) and $\widetilde{H}_x^\circ(t) := \sum_{k=0}^\infty F_k^\circ(t)(1-x)^k$ as above. Then,*

$$\mathbf{E}[\widetilde{H}_x^\circ(t)|G_0] = \mathbf{E}[\widetilde{H}_{X_t}^\circ(0)|X_0 = x].$$

Proof: On a probability space where $\mathcal{X}$ and the PDt-graph are independent, combining (4.1) and (3.7),

$$\frac{d}{ds}\mathbf{E}[\widetilde{H}_{X_{t-s}}^\circ(s)|G_0, X_0 = x] = 0.$$

The result then follows since $s \mapsto \mathbf{E}[\widetilde{H}_{X_{t-s}}^\circ(s)]$ is constant. □

4.3. *Proof of Theorem 2.7.* We start with the case $p \leq p^*$. Here, we know from Lemma 3.3 that $X_t \xrightarrow{t\to\infty} 0$ almost surely. Hence, using Proposition 4.5, for $q \in [0, 1)$ and $x := 1 - q$,

$$\begin{aligned}\lim_{n\to\infty}\mathbf{E}[H_q^\circ(n)] = \lim_{t\to\infty}\mathbf{E}[H_q^\circ(t)] = \lim_{t\to\infty}\mathbf{E}[\widetilde{H}_x^\circ(t)] = \lim_{t\to\infty}\mathbf{E}[\widetilde{H}_{X_t}^\circ(0)|X_0 = x]\\ = \mathbf{E}[\widetilde{H}_0^\circ(0)] = \mathbf{E}[H_1^\circ(0)] = 1.\end{aligned}$$

In particular, since by (3.1)

$$\mathbf{E}[H_0^\circ(n+1)|\mathcal{F}_n] = \mathbf{E}[F_0^\circ(n+1)|\mathcal{F}_n] = \frac{n}{n+1}F_0^\circ(n) + \frac{1}{n+1}\sum_{\ell\geq 0}F_\ell^\circ(n)(1-p)^\ell$$

$$\geq F_0^\circ(n) = H_0^\circ(n),$$

$(H_0^\circ(n))_{n=0,1,2,\dots}$ is a bounded sub-martingale and thus converges almost surely to 1. By the monotonicity of the probability generating function, we also obtain the stated uniform convergence result.

The case $p > p^*$ can be treated similarly, but X_t does not converge almost surely to a constant. Hence, in this case with X_∞ from Lemma 3.3 we can compute

$$\begin{aligned}\lim_{n\to\infty} \mathbf{E}[H_q^\circ(n)] &= \lim_{t\to\infty} \mathbf{E}[\widetilde{H}_{X_t}^\circ(0)|X_0 = x] = \sum_{k=0}^{\infty} F_k^\circ(n_0)\mathbf{E}[(1-X_\infty)^k] \\ &= \sum_{k=0}^{\infty} F_k^\circ(n_0) \sum_{\ell=0}^{k} \binom{k}{\ell} (-1)^\ell \mathbf{E}[X_\infty^\ell] \\ &= \sum_{\ell=0}^{\infty} (-1)^\ell \mathbf{E}[X_\infty^\ell] \sum_{k=\ell}^{\infty} \binom{k}{\ell} F_k^\circ(n_0) \\ &= 1 - \sum_{\ell=1}^{\infty} \frac{S_\ell^\circ(n_0)}{\ell!}(-1)^{\ell-1}\mathbf{E}[X_\infty^\ell].\end{aligned}$$

Now by Lemma 3.3, the result follows.

5. Proof of Theorems 2.9 and 2.14

Proof of Theorem 2.9: We start with 1. where we make use of Proposition 3.1 and Lemma 3.5. For the almost sure convergence in (2.1), we use Lemma 3.5.2 with $a = kp^{k-1}$, and for (2.2), we use (3.8). We will set aside the $\mathcal{L}^2$ convergence for now.

The proof of 2. is a bit more involved since the recursions from Proposition 3.1 for S_k involve both, S_k and S_{k-1}. But considering the quantity

$$Q_k(n) := \sum_{\ell=1}^{k} a_\ell S_\ell(n)$$

where, recalling that the empty product is 1,

$$a_\ell := \prod_{m=\ell}^{k-1} \frac{m(m+1)}{k - m + p^{k-1} - p^{m-1}},$$

we obtain a fitting recursion as follows:

$$\begin{aligned}\mathbf{E}[Q_k(n+1)|\mathcal{F}_n] &= \sum_{\ell=1}^{k} a_\ell \Big(\Big(1 + \frac{p\ell + p^\ell}{n}\Big) S_\ell(n) + \frac{p\ell(\ell-1)}{n} S_{\ell-1}(n)\Big) \\ &= \Big(1 + \frac{pk + p^k}{n}\Big) S_k(n) + \sum_{\ell=1}^{k-1} S_\ell(n) \Big(\Big(1 + \frac{p\ell + p^\ell}{n}\Big) a_\ell + \frac{p\ell(\ell+1)}{n} a_{\ell+1}\Big) \\ &= \Big(1 + \frac{pk + p^k}{n}\Big) S_k(n) + \sum_{\ell=1}^{k-1} a_\ell S_\ell(n) \Big(1 + \frac{p\ell + p^\ell}{n} + \frac{pk - p\ell + p^k - p^\ell}{n}\Big) \\ &= \Big(1 + \frac{pk + p^k}{n}\Big) Q_k(n).\end{aligned}$$

Thus, by Lemma 3.5.2 there are random variables $S_k(\infty)$ satisfying $n^{-(kp+p^k)}Q_k(n) \xrightarrow{n\to\infty} S_k(\infty)$ almost surely. Since $n^{\ell p+p^\ell} = o\big(n^{kp+p^k}\big)$ for all $\ell < k$ and (2.2) provides the asymptotics of $S_1(n) = 2C_2(n)$, inductively the almost sure convergence in (2.3) follows.

Now, writing $Q_1(n) = S_1(n)$ as well as $Q_2(n) = S_2(n) + \frac{2}{p}S_1(n)$, we have from Lemma 3.5

$$\begin{aligned}\mathbf{E}[S_2(n)] &= \mathbf{E}[Q_2(n)] - \frac{2}{p}\mathbf{E}[Q_1(n)] \\ &= Q_2(n_0)\prod_{k=n_0}^{n-1}\frac{k+2p+p^2}{k} - \frac{2}{p}Q_1(n_0)\prod_{k=n_0}^{n-1}\frac{k+2p}{k}\end{aligned}$$

and (2.4) follows.

For the $\mathcal{L}^2$ convergece in (2.1) first consider the number of pairs of k-cliques at time n, $\frac{1}{2}C_k(n)_{\downarrow 2}$. Now let $C_{k,\ell}(n)$ be the number of *ℓ-pairs* at time n, that is pairs of k-cliques which share exactly ℓ nodes. (e.g. two disjoint cliques form a 0-pair and a $(k-1)$-pair of k-cliques is a $(k+1)$-clique with an edge missing.) Thus, we obtain

$$\tfrac{1}{2}C_k(n)_{\downarrow 2} = \binom{C_k(n)}{2} = \sum_{\ell=0}^{k-1} C_{k,\ell}(n). \tag{5.1}$$

Supposing there is an ℓ-pair of k-cliques at time n, there are four ways for new pairs to arise during the next time step:

1) First of all, every new clique forms a $(k-1)$-pair with the clique it was duplicated from, since they only do not share the new node. As (3.4) shows, this happens $\frac{kp^{k-1}}{n}C_k(n)$ times on average during the next time step. In the next 3 cases we will ignore those events.

2) One of the $2(k-\ell)$ not-shared nodes is chosen and the one clique of the pair it is contained in is duplicated. Then, since the new clique retains the ℓ nodes which are part of the non-duplicated clique of the pair, a new ℓ-pair is formed.

   Corresponding probability: $\frac{2(k-\ell)}{n}p^{k-1}$

3) One of the ℓ shared nodes is chosen and both cliques of the pair are duplicated. Obviously, this way a new ℓ-pair arises. Additionally the other two new pairs (one original and the copy of the other original respectively) are $(\ell-1)$-pairs, since those cliques do not share the new node.

   Corresponding probability: $\frac{\ell}{n}p^{2k-\ell-1}$

4) One of the ℓ shared nodes is chosen, but only one of the cliques is duplicated. Similarly to 3), a new $(\ell-1)$-pair arises. (Since the duplication of one clique fails, so does the creation of the new ℓ-pair and one of the $(\ell-1)$-pairs.)

   Corresponding probability: $\frac{\ell}{n}\cdot 2p^{k-1}(1-p^{k-\ell}) = \frac{\ell}{n}2p^{k-1} - \frac{\ell}{n}2p^{2k-\ell-1}$

Following this, for $\ell \le k-2$ we obtain

$$\mathbf{E}[C_{k,\ell}(n+1) - C_{k,\ell}(n) \mid \mathcal{F}_n] =$$

$$= \frac{2(k-\ell)p^{k-1} + \ell p^{2k-\ell-1}}{n} C_{k,\ell}(n) + \frac{2(\ell+1)p^{k-1}}{n} C_{k,\ell+1}(n) \tag{5.2}$$

and

$$\begin{aligned}&\mathbf{E}[C_{k,k-1}(n+1) - C_{k,k-1}(n) \mid \mathcal{F}_n] \\ &\quad= \frac{2p^{k-1} + (k-1)p^k}{n} C_{k,k-1}(n) + \frac{kp^{k-1}}{n} C_k(n).\end{aligned} \tag{5.3}$$

Using (5.1) we compute

$$\begin{aligned}\mathbf{E}[C_k(n+1)_{\downarrow 2} &- C_k(n)_{\downarrow 2} \mid \mathcal{F}_n] = 2\sum_{\ell=0}^{k-1} \mathbf{E}[C_{k,\ell}(n+1) - C_{k,\ell}(n) \mid \mathcal{F}_n] \\ &= \sum_{\ell=0}^{k-1} \frac{2(k-\ell)p^{k-1} + \ell p^{2k-\ell-1}}{n} 2C_{k,\ell}(n) + \sum_{\ell=1}^{k-1} \frac{2\ell p^{k-1}}{n} 2C_{k,\ell}(n) + \frac{2kp^{k-1}}{n} C_k(n) \\ &= \frac{2kp^{k-1}}{n} C_k(n)_{\downarrow 2} + \frac{2}{n} \sum_{\ell=0}^{k-1} \ell p^{2k-\ell-1} C_{k,\ell}(n) + \frac{2kp^{k-1}}{n} C_k(n) \\ &= \frac{2kp^{k-1}}{n} C_k(n)^2 + \frac{2p^k}{n} \sum_{\ell=1}^{k-1} \ell p^{k-1-\ell} C_{k,\ell}(n)\end{aligned}$$

and thus

$$\begin{aligned}&\mathbf{E}[C_k(n+1)^2] \\ &\quad= \Big(1 + \frac{2kp^{k-1}}{n}\Big)\mathbf{E}[C_k(n)^2] + \frac{2p^k}{n} \sum_{\ell=1}^{k-1} \ell p^{k-1-\ell} \mathbf{E}[C_{k,\ell}(n)] + \frac{kp^{k-1}}{n} \mathbf{E}[C_k(n)].\end{aligned}$$

Since $\mathbf{E}[C_k(n)] = O\big(n^{kp^{k-1}}\big) = O\big(n^{2kp^{k-1} - kp^{k-1}}\big)$, for the use of Lemma 3.6 it suffices to show the existence of a $\delta > 0$ holding $\sum_\ell \mathbf{E}[C_{k,\ell}(n)] = O\big(n^{2kp^{k-1}-\delta}\big)$. From (5.3) it follows, that

$$\begin{aligned}\mathbf{E}[C_{k,k-1}(n+1)] &= \Big(1 + \frac{2p^{k-1} + (k-1)p^k}{n}\Big)\mathbf{E}[C_{k,k-1}(n)] + \frac{1}{n} O\big(n^{kp^{k-1}}\big) \\ &\overset{3.7}{=} O\Big(n^{\max\{2p^{k-1} + (k-1)p^k, kp^{k-1}\} + \varepsilon}\Big)\end{aligned}$$

for arbitrarily small $\varepsilon > 0$. Using Corollary 3.7 again, inductively, (5.2) implies

$$\mathbf{E}[C_{k,\ell}(n)] = O\Big(n^{\max_{\ell \le m \le k} \big(2(k-m)p^{k-1} + mp^{2k-m-1}\big) + \bar{\varepsilon}}\Big)$$

and hence

$$\sum_{\ell=1}^{k-1} \mathbf{E}[C_{k,\ell}(n)] = O\Big(n^{\max_{1 \le m \le k} \big(2(k-m)p^{k-1} + mp^{2k-m-1}\big) + \tilde{\varepsilon}}\Big) \tag{5.4}$$

for arbitrarily small $\tilde{\varepsilon} > 0$. Since

$$\max_{1 \le m \le k} \Big(2(k-m)p^{k-1} + mp^{2k-m-1}\Big) \le \max_{1 \le m \le k} (2k-m)p^{k-1} = 2kp^{k-1} - p^{k-1},$$

letting $\tilde{\varepsilon} = p^{k-1}/2$, (5.4) satisfies the conditions of Lemma 3.6.1, we finally obtain $\mathbf{E}[C_k(n)^2] = O\big(n^{2kp^{k-1}}\big)$ and Lemma 3.5.3 applies. □

Proof of Theorem 2.14: For (2.5), we will show that

$$\mathbf{P}(D_k(n) \le \ell | D_k(n_0) = a) = \sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell}{m} \binom{m-1}{a-1} \prod_{j=n_0}^{n-1} \left(1 - \frac{pm}{j}\right) \quad (5.5)$$

which implies (2.5).

We fix n_0, k and a and set

$$\Phi_\ell(n) := \mathbf{P}(D_k(n) \le \ell | D_k(n_0) = a).$$

We will prove (5.5) by induction over n. For $n = n_0$, we have that $\Phi_\ell(n_0) = 1_{\ell \ge a}$. In addition, the right hand side of (5.5) gives for $n = n_0$

$$\begin{aligned}\sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell}{m} \binom{m-1}{a-1} &= \sum_{m} (-1)^{m-a} \binom{(\ell-a)-(-a)}{m} \binom{-1+m}{-a+m} \\ &= (-1)^{\ell-a} \binom{-1}{\ell-a} = \binom{\ell-a}{\ell-a} = 1_{\ell \ge a}\end{aligned}$$

according to Riordan (1979), (8) and (ii) in Chapter 1. This shows that (5.5) holds for $n = n_0$ and all ℓ. In order to apply induction, we get the recursion

$$\begin{aligned}\Phi_\ell(n+1) &= \Phi_\ell(n) - \frac{p\ell}{n} \mathbb{P}(D_k(n) = \ell | D_k(n_0) = a) \\ &= \Phi_\ell(n) - \frac{p\ell}{n} \cdot \Big(\Phi_\ell(n) - \Phi_{\ell-1}(n)\Big)\end{aligned}$$

since D_k increases by at most one in every time step.

Assume that (5.5) holds for an n for all ℓ. Then, using the recursion, and the assumption for n,

$$\begin{aligned}\Phi_\ell(n+1) = &\sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell}{m} \binom{m-1}{a-1} \prod_{j=n_0}^{n-1} \left(1 - \frac{pm}{j}\right) \\ &- \sum_{m=a}^{\ell} (-1)^{m-a} \frac{p\ell}{n} \left(\binom{\ell}{m} - \binom{\ell-1}{m}\right) \binom{m-1}{a-1} \prod_{j=n_0}^{n-1} \left(1 - \frac{pm}{j}\right) \\ = &\sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell}{m} \binom{m-1}{a-1} \prod_{j=n_0}^{n-1} \left(1 - \frac{pm}{j}\right) \\ &- \frac{pm}{n} \cdot \sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell}{m} \binom{m-1}{a-1} \prod_{j=n_0}^{n-1} \left(1 - \frac{pm}{j}\right) \\ = &\sum_{m=a}^{\ell} (-1)^{m-a} \binom{\ell}{m} \binom{m-1}{a-1} \prod_{j=n_0}^{n} \left(1 - \frac{pm}{j}\right)\end{aligned}$$

and we are done.

For (2.6), we will use Lemma 3.5. We have that

$$D_k(n+1) - D_k(n) = \begin{cases} 1, & \text{with probability } p\frac{D_k(n)}{n}, \\ 0, & \text{with probability } 1 - p\frac{D_k(n)}{n} \end{cases}$$

since D_k increases by one iff one neighbor of v_k and the respective edge are copied. Using Lemma 3.4 and that $D_k(n) \xrightarrow{n\to\infty} \infty$ we obtain for $r > -1 \geq -D_k(n_0)$ that

$$D_k(n)^r \sim \frac{\Gamma(D_k(n)+r)}{\Gamma(D_k(n)}$$

almost surely, where the right hand side satisfies

$$\begin{aligned}
&\mathbf{E}\Big[\frac{\Gamma(D_k(n+1)+r)}{\Gamma(D_k(n+1))}\Big|\mathcal{F}_n\Big] \\
&\quad = \frac{pD_k(n)}{n} \cdot \frac{\Gamma(D_k(n)+1+r)}{\Gamma(D_k(n)+1)} + \Big(1 - \frac{pD_k(n)}{n}\Big) \cdot \frac{\Gamma(D_k(n)+r)}{\Gamma(D_k(n))} \\
&\quad = \frac{\Gamma(D_k(n)+r)}{\Gamma(D_k(n))} \cdot \Big(\frac{pD_k(n)}{n} \cdot \frac{D_k(n)+r}{D_k(n)} + 1 - \frac{pD_k(n)}{n}\Big) \\
&\quad = \frac{\Gamma(D_k(n)+r)}{\Gamma(D_k(n))} \cdot \Big(1 + \frac{pr}{n}\Big).
\end{aligned}$$

Thus, Lemma 3.5.2 shows

$$n^{-rp}\frac{\Gamma(D_k(n)+r)}{\Gamma(D_k(n))} \sim \Big(n^{-p}D_k(n)\Big)^r \xrightarrow{n\to\infty} D_k(\infty)^r$$

almost surely. Furthermore, Lemma 3.5.1 gives us the $\mathcal{L}^r$-boundedness for $r > 1$ we need for Lemma 3.5.3. Hence, we obtain the $\mathcal{L}^r$-convergence of $n^{-p}D_k(n)$ and (2.7). Lastly, Lemma 3.5.1 also shows the convergence of $(n^{-p}D_k(n))^{-\frac{1}{2}}$ to an integrable and thus finite random variable which delivers the almost sure positivity of $D_k(\infty)$. □

References

M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables.* Dover Publications, Inc., New York (1992). ISBN 0-486-61272-4. Reprint of the 1972 edition. MR1225604.

R. Albert. Scale-free networks in cell biology. *Journal of Cell Science* **118** (21), 4947–4957 (2005). DOI: 10.1242/jcs.02714.

R. Albert and A.-L. Barabási. Statistical mechanics of complex networks. *Rev. Modern Phys.* **74** (1), 47–97 (2002). MR1895096.

R. Azaïs, J.-B. Bardet, A. Génadot, N. Krell and P.-A. Zitt. Piecewise deterministic Markov process—recent results. In *Journées MAS 2012*, volume 44 of *ESAIM Proc.*, pages 276–290. EDP Sci., Les Ulis (2014). MR3178622.

Á. Backhausz and T. F. Móri. Asymptotic properties of a random graph with duplications. *J. Appl. Probab.* **52** (2), 375–390 (2015). MR3372081.

Á. Backhausz and T. F. Móri. Further properties of a random graph with duplications and deletions. *Stoch. Models* **32** (1), 99–120 (2016). MR3457123.

A.-L. Barabási, H. Jeong, Z. Néda, E. Ravasz, A. Schubert and T. Vicsek. Evolution of the social network of scientific collaborations. *Phys. A* **311** (3-4), 590–614 (2002). MR1943379.

G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. Nadeau and S. C. Sahinalp. The degree distribution of the generalized duplication model. *Theoret. Comput. Sci.* **369** (1-3), 239–249 (2006a). MR2277572.

G. Bebek, P. Berenbrink, C. Cooper, T. Friedetzky, J. H. Nadeau and S. C. Sahinalp. *Improved Duplication Models for Proteome Network Evolution*, pages 119–137. Springer Berlin Heidelberg, Berlin, Heidelberg (2006b). ISBN 978-3-540-48540-7. DOI: 10.1007/978-3-540-48540-7_11.

A. Bhan, D. J. Galas and T. G. Dewey. A duplication growth model of gene expression networks. *Bioinformatics* **18** (11), 1486–1493 (2002). DOI: 10.1093/bioinformatics/18.11.1486.

F. Chung, L. Lu, T. Dewey and D. Galas. Duplication models for biological networks. *J. Comp. Biol.* **10** (5), 667–687 (2004). DOI: 10.1089/106652703322539024.

C. Cooper and A. Frieze. A general model of web graphs. *Random Structures Algorithms* **22** (3), 311–335 (2003). MR1966545.

O. L. V. Costa and F. Dufour. Stability and ergodicity of piecewise deterministic Markov processes. *SIAM J. Control Optim.* **47** (2), 1053–1077 (2008). MR2385873.

M. H. A. Davis. Piecewise-deterministic Markov processes: a general class of non-diffusion stochastic models. *J. Roy. Statist. Soc. Ser. B* **46** (3), 353–388 (1984). MR790622.

M. H. A. Davis. *Markov models and optimization*, volume 49 of *Monographs on Statistics and Applied Probability*. Chapman & Hall, London (1993). ISBN 0-412-31410-X. MR1283589.

R. Durrett. *Random graph dynamics*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge (2010). ISBN 978-0-521-15016-3. MR2656427.

P. Erdős and A. Rényi. On random graphs. I. *Publ. Math. Debrecen* **6**, 290–297 (1959). MR0120167.

S. N. Ethier and T. G. Kurtz. *Markov processes*. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York (1986). ISBN 0-471-08186-8. MR838085.

E. N. Gilbert. Random graphs. *Ann. Math. Statist.* **30**, 1141–1144 (1959). MR0108839.

R. van der Hofstad. Random graphs and complex networks (2016). Lecture notes, Preprint.

I. Ispolatov, P. L. Krapivsky and A. Yuryev. Duplication-divergence model of protein interaction network. *Phys. Rev. E* **71**, 061911 (2005). DOI: 10.1103/PhysRevE.71.061911.

S. Jansen and N. Kurt. On the notion(s) of duality for Markov processes. *Probab. Surv.* **11**, 59–120 (2014). MR3201861.

H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai and A.-L. Barabási. The large-scale organization of metabolic networks. *Nature* **407** (6804), 651–654 (2000). DOI: 10.1038/35036627.

N. L. Johnson and S. Kotz. *Urn models and their application*. John Wiley & Sons, New York-London-Sydney (1977). An approach to modern discrete probability theory, Wiley Series in Probability and Mathematical Statistics. MR0488211.

J. Kim, P. L. Krapivsky, B. Kahng and S. Redner. Infinite-order percolation and giant fluctuations in a protein interaction network. *Phys. Rev. E* **66**, 055101 (2002). DOI: 10.1103/PhysRevE.66.055101.

S. Ohno. *Evolution by gene duplications*. Springer Berlin Heidelberg (1970). ISBN 978-3-642-86661-6. DOI: 10.1007/978-3-642-86659-3.

R. Pastor-Satorras, E. Smith and R. V. Solé. Evolving protein interaction networks through gene duplication. *J. Theoret. Biol.* **222** (2), 199–210 (2003). MR2070214.

J. Riordan. *Combinatorial Identities*. Robert E. Krieger Publishing Co., Huntington, N.Y. (1979). ISBN 0-88275-829-2. Reprint of the 1968 original. MR554488.

J. A. Shohat and J. D. Tamarkin. *The Problem of Moments*. American Mathematical Society Mathematical surveys, vol. I. American Mathematical Society, New York (1943). MR0008438.

E. Thörnblad. Asymptotic degree distribution of a duplication-deletion random graph model. *Internet Math.* **11** (3), 289–305 (2015). MR3344243.

A. Wagner. The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol. Biol. Evol.* **18** (7), 1283–1292 (2001).