AN EDGE DELETION MODEL FOR COMPLEX NETWORKS

PAWEŁ PRAŁAT AND CHANGPING WANG

Abstract. We propose a new random graph model—Edge Popularity—for the web graph and other complex networks, where edges are deleted over time and an edge is chosen to be deleted with probability inversely proportional to the in-degree of the destination. We show that with probability tending to one as time tends to infinity, the model generates graphs whose degree distribution follows a power law. Depending on the parameters of the model, the exponent of the power law can be any number in $(2, \infty)$.

1. Introduction

Complex networks arise a large number of settings and disciplines, ranging from the web graph, networks of social interactions, and to protein interaction networks in biology. One of the central properties of complex networks are power law degree distributions. Several models—such as preferential attachment—were proposed which simulate power laws and other emergent properties of complex networks. For an overview of such models, see the books [1, 2].

In most models, edges and vertices are added but never deleted. This is less realistic, since complex networks often both add and delete edges over time. A natural mechanism underlying such deletions is based on *popularity* measured by in-degree: edges pointing to nodes with higher in-degree are less likely to be deleted. We propose a new model where directed edges are deleted with probability inversely proportional to the in-degree of the destination. We note that other random graph models which incorporate deletion were considered independently by [3, 4].

We formally introduce the *Edge Popularity (EP)* model. Let α and β be two nonnegative real numbers satisfying $\alpha + 2\beta < 1$. (In order to get nontrivial sequence of graphs it is required to assume that $\alpha + \beta < 1$, and $\beta < 1/2$. We assume a little bit more here.) We consider a random graph process which generates a sequence of digraphs $G_t, t \in \mathbb{N}$. The graph $G_t = (V_t, E_t)$ will have n_t vertices and e_t edges. Note that n_t and e_t are themselves random variables. At $t = 0$ we start with any fixed initial digraph G_0 with n_0 vertices and m_0 edges. At time t, with probability $1 - \alpha - \beta$ we add a new vertex v_t to G_{t-1}, with a directed loop. With probability α, if $e_{t-1} > 0$, then we add a new directed edge uv to the existing vertices, where the origin is chosen with probability proportional to its out-degree and the destination is chosen with probability proportional to its in-degree; if $e_{t-1} = 0$, then we add a new directed edge to the existing vertices uniformly at random. With probability β, if $e_{t-1} > 0$, then we delete a directed edge, where an edge is chosen inversely proportional to the in-degree of the destination; if $e_{t-1} = 0$, then we do nothing.

We say that an event holds *asymptotically almost surely* (*aas*), if it holds with probability tending to one as $t \to \infty$. We will use the stronger notion of *wep* in favour of the more commonly used *aas*, since it simplifies some of our proofs. We say that an event holds *with* *extreme probability* (*wep*), if it holds with probability at least $1-\exp(-\Theta(\log^2 t))$ as $t \to \infty$. Thus, if we consider a polynomial number of events that each holds *wep*, then *wep* all events hold. To combine this notion with asymptotic notations such as $O()$ and $o()$, we follow the conventions in [8].

Finally, we will make use of the following standard result about the sum of independent random variables, known as the *Chernoff's inequality*:

Theorem 1 (Chernoff bound, see for example Theorem 2.8 [6])**.** *Let X be a random variable that can be expressed as a sum $X = \sum_{i=1}^n X_i$ of independent random indicator variables where $X_i \in \mathrm{Be}(p_i)$ with (possibly) different $p_i = \mathbb{P}(X_i = 1) = \mathbb{E}X_i$. Then the following holds for $t \ge 0$:*

$$\begin{aligned} \mathbb{P}(X \ge \mathbb{E}X + t) &\le \exp\left(-\frac{t^2}{2(\mathbb{E}X + t/3)}\right), \\ \mathbb{P}(X \le \mathbb{E}X - t) &\le \exp\left(-\frac{t^2}{2\mathbb{E}X}\right). \end{aligned}$$

In particular, if $\varepsilon \le 3/2$, then

$$\mathbb{P}(|X - \mathbb{E}X| \ge \varepsilon\mathbb{E}x) \le 2\exp\left(-\frac{\varepsilon^2\mathbb{E}X}{3}\right).$$

Moreover, if $\mathbb{E}X \le \log^2 n$, then wep $X = O(\log^2 n)$.

2. Expected degree distribution

Before we analyze the degree distribution, let us present a few simple properties of random variables n_t, e_t.

Lemma 2. *For the EP model we have that*

$$\mathbb{E}n_t = n_0 + (1-\alpha-\beta)t.$$

Moreover, wep

$$n_t = (1-\alpha-\beta)t + O\left(\sqrt{t}\log t\right).$$

Proof. It is clear that the expected number of vertices added at a given time-step is $1-\alpha-\beta$. The concentration follows directly from the Chernoff's inequality. □

Moreover, we can show that and the concentration follows from martingale method. Alternatively we repeatedly use the Chernoff's inequalities in the proof of the following lemma.

Lemma 3. *For the EP model we have that*

$$\mathbb{E}m_t = m_0 + (1-2\beta)t + O(1).$$

Moreover, wep

$$e_t = (1-2\beta)t + O\left(\sqrt{t}\log t\right).$$

Proof. Not that at time-step i an edge is added with probability $\alpha + (1-\alpha-\beta) = 1-\beta$ (with probability α an edge between two vertices is added; with probability $1-\alpha-\beta$ a loop is added) and we are trying to delete an edge with probability β.

Let $\{Z_i\}$ be a sequence of t independent random variables each of which is equal to 1 with probability $1-\beta$ and -1 with probability β. Then

$$e_t = m_0 + \sum_{i=1}^{t} Z_i + f\Big(m_0, \{Z_i\}\Big) \ ,$$

where $f = f(m_0, \{Z_i\})$ is a deterministic function arising from the fact that an edge is not deleted if e_i is equal to 0. It is clear that $\mathbb{E}f = \sum_{j=1}^{t} \beta \mathbb{P}(e_j = 0)$. From the Chernoff's inequality it follows that

$$\mathbb{P}(e_j = 0) \le \mathbb{P}\Big(\sum_{i=1}^{j} Z_i \le 0\Big) \le \exp\Big(-\frac{(1-2\beta)j}{2}\Big),$$

so

$$\mathbb{E}f \le \sum_{j=1}^{t} \beta \exp\Big(-\frac{(1-2\beta)j}{2}\Big) = O(1).$$

Now, we will show the concentration for e_t. Since f is nonnegative, the random variable e_t is stochastically bounded from below by $m_0 + \sum_{i=1}^{t} Z_i$. The lower tail of this variable has the claimed sharp concentration, by the Chernoff's inequality. Thus, for every $\varepsilon = \Theta\big(\log t/\sqrt{t}\big)$,

$$\begin{aligned}
&\mathbb{P}\Big(e_t < m_0 + (1-\varepsilon)(1-2\beta)t\Big) \\
&\qquad \le \mathbb{P}\Big(\sum_{i=1}^{t} Z_i < (1-\varepsilon)\mathbb{E}\sum_{i=1}^{t} Z_i\Big) \\
&\qquad \le 2\exp\Big(-\frac{\varepsilon^2}{3}\mathbb{E}\sum_{i=1}^{t} Z_i\Big) = \exp\Big(-\Theta(\log^2 t)\Big) \ .
\end{aligned}$$

For the upper tail, we note first (again using Chernoff) that *wep* the random variable $Z(k) = \sum_{i=1}^{k} Z_i$ is positive for every k in the range $t^{1/4} \le k \le t$. Hence, *wep* $f < t^{1/4}$. The upper tail bound again follows from Chernoff's inequality. For every $\varepsilon = \Theta\big(\log t/\sqrt{t}\big)$

$$\begin{aligned}
&\mathbb{P}\Big(e_t > m_0 + (1+\varepsilon)(1-2\beta)t\Big) \\
&\qquad = \mathbb{P}\Big(\sum_{i=1}^{t} Z_i > (1+\varepsilon)\ \mathbb{E}\sum_{i=1}^{t} Z_i - f\Big) \\
&\qquad \le \mathbb{P}\Big(\sum_{i=1}^{t} Z_i > (1+\frac{\varepsilon}{2})\ \mathbb{E}\sum_{i=1}^{t} Z_i\Big) \\
&\qquad \le 2\exp\Big(-\frac{\varepsilon^2}{12}\mathbb{E}\sum_{i=1}^{t} Z_i\Big) = \exp\Big(-\Theta(\log^2 t)\Big) \ .
\end{aligned}$$

□

We also need the following lemma on real sequences. The proof of this lemma is analogous to Lemma 3.1 of [2], and so is omitted.

Lemma 4. *If a sequence $\{a_t\}$ satisfies the following recursive formula*

$$a_{t+1} \leq \left(1 - \frac{p_t}{t}\right) a_t + x_t, \tag{2.1}$$

and

$$a_{t+1} \geq \left(1 - \frac{q_t}{t}\right) a_t + y_t, \tag{2.2}$$

where $\{p_t\}, \{q_t\}, \{x_t\}$ and $\{y_t\}$ are real sequences satisfying that $\lim_{t\to\infty} p_t = \lim_{t\to\infty} q_t = b \geq 0$ and $\lim_{t\to\infty} x_t = \lim_{t\to\infty} y_t = c$, then

$$\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}.$$

We now state and prove the main result of this section. Let $N_{k,t}$ be the number of vertices with in-degree k at time t in the EP model. In the following theorem, we will show that the expectation of $N_{k,t}$ follows a power law. Note that we add an assumption that the expectation has a linear behaviour. This is not justifiable (at least, at this point) but the results presented in the next section imply that this is the right approach.

Theorem 5. *We assume that $\mathbb{E}(N_{k,t}) = b_k t + o(t)$. Then the expected in-degree distribution follows a power law with exponent $\eta = 1 + \frac{1-2\beta}{\alpha} \in (2, \infty)$. More precisely, we have that*

$$b_k = C_1(\alpha, \beta) k^{-\eta} (1 + O(k^{-1})),$$

and $C_1(\alpha, \beta)$ is a constant.

Proof. Assume that there are e_t edges, n_t vertices, and $N_t = n_t - N_{0,t}$ vertices with in-degree at least 1 at time t, for $t \geq 0$. We abbreviate "with probability" by "w.p.". It is not hard to see that

$$N_{0,t+1} = \begin{cases} N_{0,t} + 1 & \text{w.p. } \beta \frac{N_{1,t}}{N_t}; \\ N_{0,t} & \text{otherwise,} \end{cases}$$

and

$$N_{1,t+1} = \begin{cases} N_{1,t} + 1 & \text{w.p. } 1 - \alpha - \beta + \beta \frac{N_{2,t}}{N_t}; \\ N_{1,t} - 1 & \text{w.p. } \alpha \frac{N_{1,t}}{e_t} + \beta \frac{N_{1,t}}{N_t}; \\ N_{1,t} & \text{otherwise.} \end{cases}$$

In general, for $k > 1$, we have that

$$N_{k,t+1} = \begin{cases} N_{k,t} + 1 & \text{w.p. } \alpha \frac{(k-1)N_{k-1,t}}{e_t} + \beta \frac{N_{k+1,t}}{N_t}; \\ N_{k,t} - 1 & \text{w.p. } \alpha \frac{kN_{k,t}}{e_t} + \beta \frac{N_{k,t}}{N_t}; \\ N_{k,t} & \text{otherwise.} \end{cases}$$

Hence,

$$\begin{aligned}
\mathbb{E}(N_{0,t+1} \mid G_t) &= N_{0,t} + \beta \frac{N_{1,t}}{N_t}, && (2.3)\\
\mathbb{E}(N_{1,t+1} \mid G_t) &= N_{1,t} \left(1 - \frac{\alpha}{e_t} - \frac{\beta}{N_t}\right) + \beta \frac{N_{2,t}}{N_t} + 1 - \alpha - \beta, && (2.4)\\
\mathbb{E}(N_{k,t+1} \mid G_t) &= N_{k,t} \left(1 - \frac{k\alpha}{e_t} - \frac{\beta}{N_t}\right) + \alpha \frac{(k-1)N_{k-1,t}}{e_t} + \beta \frac{N_{k+1,t}}{N_t}, && (2.5)
\end{aligned}$$

for $k > 1$.

Define $\widehat{e}_t = m_0 + (1-2\beta)t$ and $\widehat{N}_t = (1-\alpha-\beta)t - \mathbb{E}(N_{0,t})$. For a fixed real number $\varepsilon > 0$, let A_t be the event that $|e_t - \widehat{e}_t| \le \varepsilon t^{\frac{2}{3}}$, and let B_t be the event that $|N_t - \widehat{N}_t| \le \varepsilon t^{\frac{2}{3}}$. Under an assumption that

$$\mathbb{P}\left(|N_t - \mathbb{E}(N_t)| > \varepsilon t^{\frac{2}{3}}\right) < 2\exp\left(-\frac{\varepsilon^2 t^{1/3}}{2}\right),$$

by Chernoff bound, we have that

$$\mathbb{P}(A_t) \ge 1 - 2\exp\left(-\frac{\varepsilon^2 t^{1/3}}{2}\right) - 2\exp\left(-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t\right), \tag{2.6}$$

and

$$\mathbb{P}(B_t) \ge 1 - 2\exp\left(-\frac{\varepsilon^2 t^{1/3}}{2}\right). \tag{2.7}$$

By (2.6) and (2.7) we have that

$$\begin{aligned}\mathbb{P}\left(A_t \cap B_t\right) &= \mathbb{P}\left(A_t\right) + \mathbb{P}\left(B_t\right) - \mathbb{P}\left(A_t \cup B_t\right)\\ &\ge 1 - 4e^{-\frac{\varepsilon^2 t^{\frac{1}{3}}}{2}} - 2e^{-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t}. \qquad (2.8)\end{aligned}$$

By (2.5) and (2.8), we know that if A_t and B_t hold, then

$$\mathbb{E}(N_{k,t+1}|G_t) \ge N_{k,t}\left(1 - \frac{k\alpha}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}} - \frac{\beta}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}}\right) + \alpha\frac{(k-1)N_{k-1,t}}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}} + \beta\frac{N_{k+1,t}}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}}.$$

This occurs with probability $\mathbb{P}\left(A_t \cap B_t\right) \ge 1 - 4e^{-\frac{\varepsilon^2 t^{\frac{1}{3}}}{2}} - 2e^{-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t}$. It always holds that

$$\mathbb{E}(N_{k,t+1}|G_t) \ge N_{k,t}\left(1 - \frac{k\alpha}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}} - \frac{\beta}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}}\right) + \alpha\frac{(k-1)N_{k-1,t}}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}} + \beta\frac{N_{k+1,t}}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}} - M_{k1}t,$$

where $M_{k1} > 0$ is a constant. Thus,

$$\begin{aligned}\mathbb{E}(N_{k,t+1}) \ge\ & \mathbb{E}(N_{k,t})\left(1 - \frac{k\alpha}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}} - \frac{\beta}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}}\right) + \alpha\frac{(k-1)\mathbb{E}(N_{k-1,t})}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}}\\ & + \beta\frac{\mathbb{E}(N_{k+1,t})}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}} - M_{k1}t\mathbb{P}\left(A_t \cap B_t\right)\\ \ge\ & \mathbb{E}(N_{k,t})\left(1 - \frac{k\alpha}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}} - \frac{\beta}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}}\right) + \alpha\frac{(k-1)\mathbb{E}(N_{k-1,t})}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}}\\ & + \beta\frac{\mathbb{E}(N_{k+1,t})}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}} - M_{k1}t\left(4e^{-\frac{\varepsilon^2 t^{\frac{1}{3}}}{2}} + 2e^{-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t}\right). \qquad (2.9)\end{aligned}$$

Similarly, we have that

$$\begin{aligned}\mathbb{E}(N_{k,t+1}) \le\ & \mathbb{E}(N_{k,t})\left(1 - \frac{k\alpha}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}} - \frac{\beta}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}}\right) + \alpha\frac{(k-1)\mathbb{E}(N_{k-1,t})}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}}\\ & + \beta\frac{\mathbb{E}(N_{k+1,t})}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}} + M'_{k1}t\left(4e^{-\frac{\varepsilon^2 t^{\frac{1}{3}}}{2}} + 2e^{-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t}\right). \qquad (2.10)\end{aligned}$$

Let $a_t = \mathbb{E}(N_{k,t})$,

$$p_t = \left(\frac{k\alpha}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}} + \frac{\beta}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}} \right) t,$$

$$q_t = \left(\frac{k\alpha}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}} + \frac{\beta}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}} \right) t,$$

$$x_t = \alpha \frac{(k-1)\mathbb{E}(N_{k-1,t})}{\widehat{e}_t - \varepsilon t^{\frac{2}{3}}} + \beta \frac{\mathbb{E}(N_{k+1,t})}{\widehat{N}_t - \varepsilon t^{\frac{2}{3}}} + M'_{k1} t \left(4e^{-\frac{\varepsilon^2 t^{\frac{1}{3}}}{2}} + 2e^{-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t} \right),$$

and

$$y_t = \alpha \frac{(k-1)\mathbb{E}(N_{k-1,t})}{\widehat{e}_t + \varepsilon t^{\frac{2}{3}}} + \beta \frac{\mathbb{E}(N_{k+1,t})}{\widehat{N}_t + \varepsilon t^{\frac{2}{3}}} - M_{k1} t \left(4e^{-\frac{\varepsilon^2 t^{\frac{1}{3}}}{2}} + 2e^{-\frac{1}{2}\left(\frac{1}{2}-\beta\right)^2 t} \right).$$

By hypothesis that the limit $\lim_{t\to\infty} \frac{\mathbb{E}(N_{k,t})}{t} = b_k$ exists for all $k \geq 0$, (2.9), (2.10) and Lemma 4, we obtain that, for $k > 1$,

$$\begin{aligned}(\beta - 2\beta^2) b_{k+1} + [(\alpha + b_0 - 1)(1-2\beta) - \alpha(1-\alpha-\beta-b_0)k] b_k & \\ +\alpha(1-\alpha-\beta-b_0)(k-1) b_{k-1} &= 0.\end{aligned}$$

In the following we will solve (2.11) by using the Laplace Method. This method was first used in the study of the web graph models by [4]. Replacing k by $k+1$ in (2.11), we have that

$$\begin{aligned}(\beta - 2\beta^2) b_{k+2} + [(\alpha + b_0 - 1)(1-2\beta) - \alpha(1-\alpha-\beta-b_0)(k+1)] b_{k+1} & \\ +\alpha(1-\alpha-\beta-b_0) k b_k &= 0,\end{aligned}$$

which is of the form

$$(A_2(k+2) + B_2)\, b_{k+2} + (A_1(k+1) + B_1)\, b_{k+1} + (A_0 k + B_0)\, b_k = 0, \tag{2.11}$$

where $A_2 = 0, B_2 = \beta - 2\beta^2, A_1 = -\alpha(1-\alpha-\beta-b_0), B_1 = (\alpha + b_0 - 1)(1-2\beta), A_0 = \alpha(1-\alpha-\beta-b_0)$ and $B_0 = 0$. We make the substitution

$$b_k = \int_a^b t^{k-1} v(t) dt, \tag{2.12}$$

where a, b are constants, and $v(t)$ is a function of t to be determined.

Integrating by parts, we obtain that

$$k b_k = [t^k v(t)]_a^b - \int_a^b t^k v'(t) dt.$$

Let $\phi_1(t) = A_2 t^2 + A_1 t + A_0$ and $\phi_0(t) = B_2 t^2 + B_1 t + B_0$. Substituting (2.12) into (2.11), we obtain that

$$[t^k \phi_1(t) v(t)]_a^b - \int_a^b t^k \phi_1(t) v'(t) dt + \int_a^b t^{k-1} \phi_0(t) v(t) dt = 0.$$

If we ensure that

$$\frac{v'(t)}{v(t)} = \frac{\phi_0(t)}{t\phi_1(t)},$$

and

$$[t^k v(t) \phi_1(t)]_a^b = 0, \tag{2.13}$$

then (2.11) will be satisfied. Now (2.13) can be satisfied by choosing $a = 0$ and b equal to a root of $v(t)\phi_1(t) = 0$. Moreover, since $A_2 = 0, B_2 = \beta - 2\beta^2, A_1 = -\alpha(1-\alpha-\beta-b_0), B_1 = (\alpha + b_0 - 1)(1-2\beta), A_0 = \alpha(1-\alpha-\beta-b_0)$ and $B_0 = 0$, we can obtain that

$$\phi_1(t) = A_2t^2 + A_1t + A_0 = \alpha(1-\alpha-\beta-b_0)(1-t),$$

and

$$\phi_0(t) = B_2t^2 + B_1t + B_0 = (1-2\beta)\left(\beta t^2 + (\alpha + b_0 - 1)t\right).$$

Thus, we have the following differential equation

$$\frac{v'(t)}{v(t)} = \frac{\phi_0(t)}{t\phi_1(t)} = \frac{(1-2\beta)(\beta t + \alpha + b_0 - 1)}{\alpha(1-\alpha-\beta-b_0)(1-t)}. \tag{2.14}$$

Integrating (2.14), we obtain that

$$v(t) = Ce^{-\rho t}(1-t)^{\gamma},$$

where $\rho = \frac{\beta(1-2\beta)}{\alpha(1-\alpha-\beta-b_0)}$, $\gamma = \frac{1-2\beta}{\alpha}$ and C is a constant. For convenience, we choose $C = 1$. With this choice of $v(t)$, we can choose $b = 1$ and (2.13) is satisfied. So, we have $a = 0, b = 1$ and $v(t) = e^{-\rho t}(1-t)^{\gamma}$.

Now we go back to (2.12) and determine b_k as follows.

$$\begin{aligned} b_k &= \int_0^1 t^{k-1}v(t)dt \\ &= \int_0^1 t^{k-1}e^{-\rho t}(1-t)^{\gamma}dt \\ &= \int_0^1 t^{k-1}(1-t)^{\gamma}\sum_{j=0}^{\infty}\frac{(-\rho t)^j}{j!}dt \\ &= \sum_{j=0}^{\infty}\frac{(-\rho)^j\Gamma(\gamma+1)}{j!}\frac{\Gamma(k+j)}{\Gamma(k+j+\gamma+1)}. \end{aligned}$$

Using Stirling's formula for $\Gamma(k+j)$ and $\Gamma(k+j+\gamma+1)$, and assuming that k is large, we have that

$$\begin{aligned} b_k &= \left(1 + O(k^{-1})\right)\sum_{j=0}^{\infty}\frac{e^{2+\gamma}(-\rho)^j\Gamma(\gamma+1)}{j!}(k+\gamma+j+1)^{-\gamma-1} \\ &= C_1(\alpha,\beta)k^{-\gamma-1}\left(1 + O(k^{-1})\right), \end{aligned}$$

where $C_1(\alpha,\beta)$ is a constant. □

3. Concentration

One may attempt to use the differential equation method [9] to show the concentration for $N_{k,t}$, the number of vertices of degree k at time t. It provides some insight if we define real function $z_k(x)$ to model the behaviour of the scaled random variable $\frac{1}{n}N_{k,xt}$, $n(x)$ to model $\frac{1}{n}n_{xt}$, and $e(x)$ to model $\frac{1}{n}e_{xt}$. If we presume that the changes in the function correspond to the expected changes of the random variable (see (2.3, 2.4, 2.5)), then we obtain the following set of differential equations:

$$\begin{aligned} e'(x) &= 1-2\beta \\ n'(x) &= 1-\alpha-\beta \\ z_0'(x) &= \beta\frac{z_1(x)}{n(x)-z_0(x)} \qquad (3.1)\\ z_1'(x) &= 1-\alpha-\beta+\beta\frac{z_2(x)-z_1(x)}{n(x)-z_0(x)}-\alpha\frac{z_1(x)}{e(x)} \qquad (3.2)\\ z_k'(x) &= \beta\frac{z_{k+1}(x)-z_k(x)}{n(x)-z_0(x)}-\alpha\frac{kz_k(x)-(k-1)z_{k-1}(x)}{e(x)} \qquad (3.3)\end{aligned}$$

One particular solution is the following:

$$e(x)=(1-2\beta)x, n(x)=(1-\alpha-\beta)x, z_i(x)=b_ix,$$

where b_i is defined recursively as in the previous section. The general solution, and so the behaviour of the process, might be slightly different but the numerical solutions suggest that the process is self-correcting, and the process do converge to the stationary distribution, that is, *aas* $N_{k,t}=(1+o(1))b_kt$ for any $k\in\mathbb{N}\cup\{0\}$. Before we discuss the concentration issues, let us look at the number of isolated vertices.

3.1. **The number of isolated vertices.** As before, we assume that $N_{k,t}=(1+o(1))b_kt$. This is a reasonable assumption based on the discussion in the next subsection. It seems that it should be easy to find a sequence $(b_i)_{i\geq 0}$ explicitly, that is, one can get from (3.1) that

$$b_1=b_0\frac{1-\alpha-\beta-b_0}{\beta},$$

then find b_2 as a function of b_0 (from (3.2)), and finally find b_k's as functions of b_0 (one by one, from (3.3)). Unfortunately, this approach does not work (with tedious details omitted). However, for given α and β, one can solve this numerically by calculating b_k for a relatively large value of k (for example, $k=5$) as a function $f(b_0)$ (polynomial of order $k+1$). The solutions of $f(b_1)=0$ and $f(b_2)=1$ should give a very good approximation (upper and lower bound) for b_0.

In order to illustrate this technique, let us consider the following example: let $\alpha=0.4$ and $\beta=0.05$. We get that

$$\begin{aligned} b_5 \approx\ & -1.07070758847736662551\cdot 10^5+4.00603411979881115 68\cdot 10^6\cdot b_0 \\ & -3.0761538911751257430\cdot 10^7\cdot b_0^2+1.0418253955189757659\cdot 10^8\cdot b_0^3 \\ & -1.8135354366712391403\cdot 10^8\cdot b_0^4+1.5969397347965249199\cdot 10^8\cdot b_0^5 \\ & -5.6588934613625971649\cdot 10^7\cdot b_0^6,\end{aligned}$$

so

$$0.035154140737645644868\leq b_0\leq 0.035154595454535434252$$

(the length of the interval is $4.54716889789384\cdot 10^{-7}$). After performing a few more steps, one can get much better precision (for example, $1.77168067\cdot 10^{-13}$ for $k=9$). Below (Figure 2), we present a graph of $b_i, i\in[5]$ as a function of b_0.

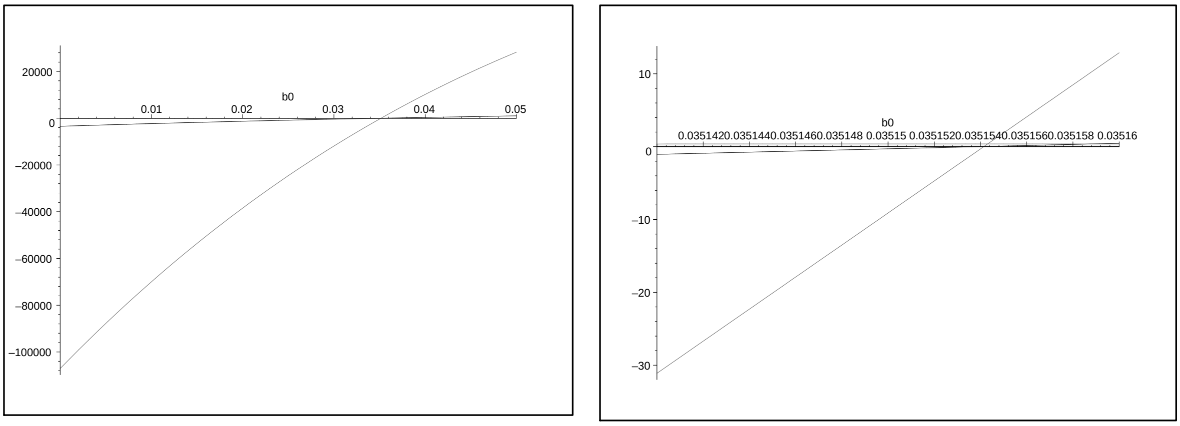


FIGURE 1. $b_i = b_i(b_0)$ as a function of b_0.

Let us also mention that the sequence $(b_k)_{k\geq 0}$ is not monotonic. Since we do have power law distribution, it is the case for k sufficiently large but certainly not for small values of k. For example, if $\alpha = 0.4$ and $\beta = 0.05$ as before, then we get the following:

$k =$	$b_k \approx$
0	0.0351541443
1	0.3619793105
2	0.0825107242
3	0.0307573540
4	0.0145324331
5	0.0079224760

This behaviour is what one should expect. Since with probability $1-\alpha-\beta$ we add a new vertex with a directed loop, a large fraction of vertices have in-degree 1.

3.2. **Concentration.** In order to show the concentration for random variables $N_{k,t}$'s, one can use the differential equations method introduced by Wormald [9]. Unfortunately, since we delete edges in our model, this method cannot be directly applied. We do have a few issues that prevent us from doing this. We discuss them independently pointing out the solution.

3.2.1. *Issue I (infinite number of variables).* We would like to claim something for an infinite number of variables (the number of variables being a function of n). The general theorem presented in [9] does not apply to this situation. However, the proof method can also work for infinitely many variables but one has to go through the proof for the specific case: smaller and smaller error bounds on the variables are required, as the degree increases. This approach would work, but we do the following instead in order to be able to use the general purpose theorem. This gives us slightly weaker result but with much simpler argument.

We consider random variables up to degree K (very large, but constant) and use a trivial upper/lower bound for $N_{K,t}$, namely,

$$0 \leq N_{K,t} \leq n_t - \sum_{i=0}^{K-1} N_{i,t}.$$

An upper bound for will yield an upper bounds for $N_{i,t}$'s (say, coefficients c_i's). A lower bound will yield a lower bounds as well (say, coefficients a_i's). The DEs method can be used to show that a.a.s. for any $i, 0 \le i \le K-1$,

$$a_i t(1+o(1)) \le N_{i,t} \le c_i t(1+o(1)).$$

The recurrence relations for all three sequences are the same (of course, a_0, b_0, c_0 are slightly different). Since a_i and c_i can be as close to b_i as we want by taking $K = K(i)$ sufficiently large, we finally get that a.a.s.

$$N_{i,t} = b_i t(1+o(1))$$

for any $i \in \mathbb{N} \cup \{0\}$.

3.2.2. *Issue II (early phase of the process).* The standard approach to show the concentration of $N_{k,t}$ would be to consider the whole process (up to time t) and re-scale all random variables to get the system of differential equations (3.1, 3.2, 3.3). The solution to the system describes the behaviour of the process; we get that a.a.s.

$$N_{k,i} = tz_k(i/t) + o(t)$$

for $0 \le i \le t$ so, in particular, that a.a.s $N_{k,t} = (1+o(1))tz_k(1)$.

Unfortunately, after scaling the denominators in the system of DEs we have can be equal to zero, and this prevents us to use the method from the very beginning of the process. To overcome this problem we can start using it from time T (say, $T = \sqrt{t}$), for which we are sure that a.a.s. $n_T > (1+o(1))(1-\alpha-\beta)T$ and $N_{0,T} < (1+o(1))\beta T$ (see Lemmas 2 and 3). Now if we scale all random variables by T, we get $\Omega(x)$ in the denominator. The only problem is that we do not know the initial values of random variables we deal with, and so we do not know the initial value problem we should consider. However, if we suppose that the initial graph (that is, graph at time T) is known, then we can run the process up to time $2T$, solve corresponding initial value problem to get that a.a.s. $N_{k,i} = (1+o(1))Tz_k(i/T)$ for $T \le i \le 2T$. In particular, we get that a.a.s $N_{k,2T} = (1+o(1))Tz_k(2)$. We repeat the argument to cover the time interval from $2T$ to $4T$, using the final values from the phase one as the initial ones for the phase two. Next, we consider the time interval from $4T$ to $8T$, etc. We get a concentration for every single phase, to discover a self-correcting property based on the solutions of the DEs. In other words, all initial conditions for the first phase should lead to approximately the same solutions once we have chained together arbitrarily many time intervals.

3.2.3. *Issue III (exact solution).* We have solved two issues discussed before but, unfortunately, the third one remains still open. We still do not know how to prove the self-correcting property of the general solution to the system of DEs associated with the problem we consider. However, we tested a number of different initial graphs (which generated a number of different initial value problems to consider), including some extreme cases, to convince ourselves that the general solution tends to the equilibrium point regardless of the initial value vector. Numerical results support our intuition based on the observation that we are getting a new vertices of degree one at a constant rate during the whole process. So if we have too many vertices of degree one, the probability that we remove (or add) an arc to a vertex of degree one is higher comparing to the corresponding value at the equilibrium point. This, we will see the relative number of such vertices dropping down. Similarly, if the number of vertices of degree one is much smaller comparing to the expected value, the probability of this event is smaller comparing to its equilibrium counterpart and, since we get new vertices of degree one at a constant rate, this random variable is going to be corrected. When the number of vertices of degree one is close to its equilibrium, then degree zero vertices as well as degree two vertices are stabilizing as well, etc. We do converge to the stationary distribution.

In order to illustrate this technique, we consider the system of 6 differential equations. For the last equation (involving $z_5'(x)$) we consider two possibilities to get lower and upper bounds for coefficients: a_i, c_i (see *Issue I*). We test the behaviour of the process for different (in some sense, extreme) initial conditions. We start the process with 2-regular graph (collection of cycles), 5-regular graph, and the graph with the degree distribution we expect from the process, that is, the number of vertices of degree k is proportional to b_k. Clearly, for the last case, the property is preserved, and the number of vertices of degree k will remain proportional to b_k. However, it seems that if we start from any other initial value, the distribution stabilizes quickly, and after a while we obtain the desired property. It is certainly the case for a few cases we investigated. The conclusion is that no nice behaviour can be expected at the beginning of the process but, at some point when the number of vertices and edges are well concentrated around its expectations, the process becomes self-correcting and the initial condition does not matter.

The bounds we get are surprisingly close to each other, even for small number of equations. For $K = 6$, we get the following bounds, regardless of which initial values are used.

k	a_k	b_k	c_k
b_0	0.0351541442	0.0351541443	0.351541445
b_1	0.3619793098	0.3619793105	0.361979313
b_2	0.0825107143	0.0825107242	0.082510750
b_3	0.0307571553	0.0307573540	0.030757856
b_4	0.0145275514	0.0145324331	0.014544765
b_5	0.0077806939	0.0079224760	0.008280641

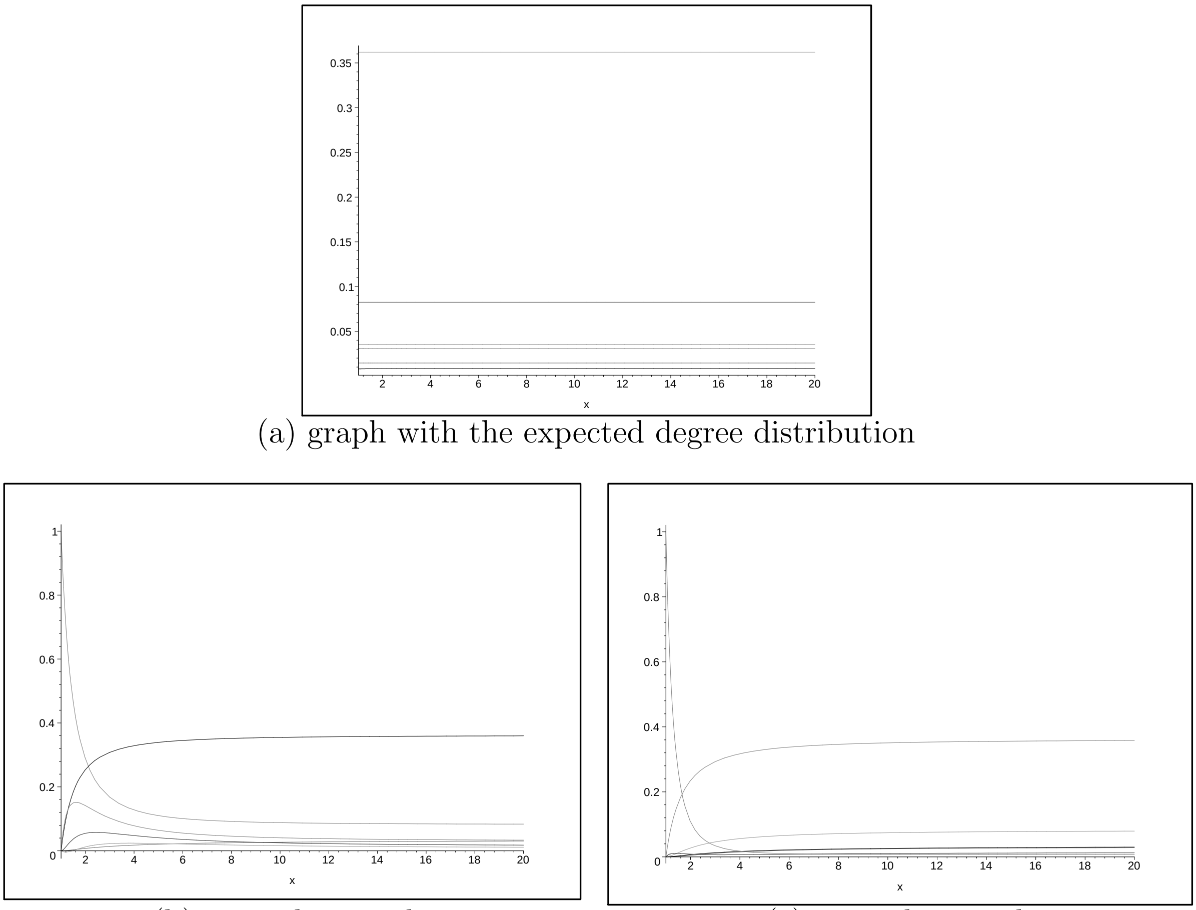
The computations presented in the paper were performed by using Maple$^{\text{TM}}$ [7]. The worksheets can be found at the following address: "`http://www.math.wvu.edu/~pralat/`".

Acknowledgment.

We would like to thank Anthony Bonato and Jeannette Janssen for discussion and for comments on the early version of this paper.

References

[1] A. Bonato, *A Course on the Web Graph*, American Mathematical Society Graduate Studies Series in Mathematics, Providence, Rhode Island, 2008.
[2] F. Chung, L. Lu, *Complex graphs and networks*, American Mathematical Society, U.S.A., 2006.
[3] F. Chung, L. Lu, Coupling on-line and off-line analyses for random power law graphs, *Internet Mathematics*, **1** (2004), 409–461.
[4] C. Cooper, A. Frieze , J. Vera, Random deletion in a scale free random graph process, *Internet Mathematics*, **1** (2004), 463-483.
[5] G.R. Grimmett, D.R. Stirzaker, *Probability and Random Processes*, Clarendon Press, Oxford, 1992.
[6] S. Janson, T. Łuczak, and A. Ruciński, *Random Graphs*, Wiley, New York, 2000.
[7] M.B. Monagan, K.O. Geddes, K.M. Heal, G. Labahn, S.M. Vorkoetter, J. McCarron, and P. DeMarco, Maple 10 Programming Guide, Maplesoft, Waterloo ON, Canada, 2005.



(a) graph with the expected degree distribution

(b) 2-regular graph

(c) 5-regular graph

FIGURE 2. Numerical solution to the DEs for a given initial graph.

[8] N.C. Wormald, Random graphs and asymptotics. Section 8.2 in *Handbook of Graph Theory*, J.L. Gross and J. Yellen (eds), pp. 817–836. CRC, Boca Raton, 2004.
[9] N.C. Wormald, The differential equation method for random graph processes and greedy algorithms. Lectures on Approximation and Randomized Algorithms, eds. M. Karoński and H. J. Prömel, PWN, Warsaw, pp. 73-155, 1999.

DEPARTMENT OF MATHEMATICS, WEST VIRGINIA UNIVERSITY, MORGANTOWN, WV 26506-6310, USA
E-mail address: pralat@math.wvu.edu

DEPARTMENT OF MATHEMATICS, RYERSON UNIVERSITY, TORONTO, ON, CANADA, M5B 2K3
E-mail address: cpwang@ryerson.ca