Preferential Attachment Trees with Fitness

Nandan Malhotra (Mathematical Institute, Leiden University)

Joint work with Konstantin Avrachenkov (INRIA Sophia Antipolis-Méditerranée) and Neeraja Sahasrabudhe (IISER Mohali).

September 8, 2021

Introduction

Setting

What are PA graphs?

Preferential Attachment graphs are growing random graphs that are used to model real world networks. Introduced by Barabási and Albert and formalised by Bollobás et.al. in 2001, they were shown to exhibit a scale free nature.

PA graphs evolve in discrete time, where at every time step, a newly added vertex attaches m edges to the graph with probability of attaching one edge to a vertex being proportional to its degree.

Formally, the model is defined as follows.

PA law

At $t = 0$, the graph consists of a root vertex labelled '0'. At time $t + 1$, an incoming vertex '$t + 1$' attaches a directed edge from itself to an existing vertex chosen according to the law of attachment given by

$$\mathbb{P}((t+1) \to v | \mathcal{G}_t) = \frac{g(d_v(t)) + f(v)}{\sum\limits_{u=0}^{t} (g(d_u(t)) + f(u))} \tag{1}$$

where $\mathcal{G}_t$ is the graph realization of time t, $d_u(t)$ is the indegree of vertex u, $f(u)$ is the fitness of u and is positive and real valued, and g is a linear function.

Note that for g of the form $g(x) = ax + b$, $\sum_{u=0}^{t} g(d_u(t)) = at + b(t + 1)$.

We restrict ourselves to *directed Preferential Attachment trees*, that is, at every time step, a new vertex attaches exactly one directed edge ($m = 1$) from itself to an existing vertex chosen according to a preferential attachment law.

Figure 1: Illustrating the process

We restrict ourselves to *directed Preferential Attachment trees*, that is, at every time step, a new vertex attaches exactly one directed edge ($m = 1$) from itself to an existing vertex chosen according to a preferential attachment law.

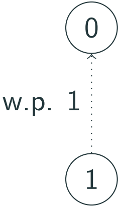
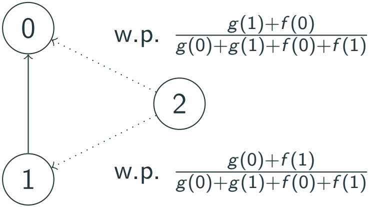


Figure 1: Illustrating the process

We restrict ourselves to *directed Preferential Attachment trees*, that is, at every time step, a new vertex attaches exactly one directed edge ($m = 1$) from itself to an existing vertex chosen according to a preferential attachment law.



w.p. $\frac{g(1)+f(0)}{g(0)+g(1)+f(0)+f(1)}$

w.p. $\frac{g(0)+f(1)}{g(0)+g(1)+f(0)+f(1)}$

Figure 1: Illustrating the process

Background

Existing literature on $f \equiv 0$ (no fitness):

Our model with $f \equiv 0$ becomes a directed version of standard Barabási-Albert model, which has been well studied and analysed in Hofstad's "Random Graphs and Complex Networks".

The work "Random Networks with Sublinear Attachment" of Dereich and Morters (2008) considers a graph where a new vertex connects to a random number of vertices with probability $\frac{g(i)}{n}$ for the i^{th} vertex (g is sublinear). They arrive at the expression

$$p(k) = \frac{1}{1+g(k)} \prod_{i=0}^{k-1} \frac{g(i)}{1+g(i)} \tag{2}$$

where $p(k)$ is the asymptotic degree distribution.

In the work of Troillet, Girorie and Pérennes (2020) for (1) with $f \equiv 0$, the following relation between g and $p(.)$ is derived

$$g(k) = \frac{1}{p(k)} \sum_{i>k} p(i) \tag{3}$$

We borrow ideas from this for the proofs of our main results.

Krapivsky et.al. study nonlinear preferential attachment. They study sublinear attachment (2000) for which they show the existence of a power law, whereas they predict the existence of a single dominant vertex attracting all future edges for the superlinear case (2001).

Literature on Prefential Attachment with fitness:

Why incorporate fitness?

Fitness preserves the identity of the vertex itself, since the degree of the vertex may not be unique

Chapter 8, section 8.9 of Hofstad's text briefly defines PA with fitness. Fitness is usually classified into two broad categories: additive fitness and multiplicative fitness. It may or may not be random.

The work of Garavaglia, Hofstad and Woeginger (2017) considers aging as well as multiplicative fitness. Using the theory of aging birth processes, they arrive at an expression very similar to the one derived by Dereich and Morters. For comparison, the two equations (as in the papers) are

$$p(k) = \frac{\alpha^*}{\alpha^* + f(k)\hat{\mathcal{L}}^g(k, \alpha^*)} \prod_{i=0}^{k-1} \frac{f(i)\hat{\mathcal{L}}^g(i, \alpha^*)}{\alpha^* + f(i)\hat{\mathcal{L}}^g(i, \alpha^*)}$$

$$p(k) = \frac{1}{1 + f(k)} \prod_{i=0}^{k-1} \frac{f(i)}{1 + f(i)}$$

Note: f here is not the fitness function.

The focus of our work will be on additive fitness. A model similar to this has been studied by Lodewijks and Ortgiese (2020). For their model, they consider random additive i.i.d. fitness, for which they arrive at results for degree distributions and maximal degree.

Fitness Regimes

We study three regimes for f, defined by $\Phi(t) = \sum_{u=0}^{t} f(u)$.

- Sublinear: $\lim_{t\to\infty} \frac{\Phi(t)}{t} = 0$
- Linear: $\lim_{t\to\infty} \frac{\Phi(t)}{t} = C_L > 0$
- Superlinear: $\lim_{t\to\infty} \frac{\Phi(t)}{t} = \infty$, $\lim_{t\to\infty} \frac{\Phi(t)}{t^n} = C_S > 0$, $n > 1$

The fitness function $f(.)$ in our work is a general function, and can be either deterministic or random. Moreover, our methods are based on more explicit computations using concentration inequalities and standard recursions.

Main Results

Main Results

We state our main results together for the sublinear and linear regimes, and separately for the superlinear regime. Recall that our Preferential attachment law in (1) was given by

$$\mathbb{P}((t+1) \to v|\mathcal{G}_t) = \frac{g(d_v(t)) + f(v)}{\sum\limits_{u=0}^{t}(g(d_u(t)) + f(u))}$$

Define $\mu_g(t) = \sum_{u=0}^{t} g(d_u(t))$ and $\mu := \lim_{t\to\infty} \frac{\mu_g(t)}{t}$. Let $p(.)$ denote the asymptotic degree distribution, that is, $p(k)$ is the probability that a chosen vertex will asymptotically have indegree k.

Sublinear and Linear Regime

Theorem

Consider a Preferential attachment process with the attachment law as in (1), with g linear. Let $\Phi(t)$ be linear, that is, $\Phi(t)/t \to C_L$ for large t and μ is as defined. Then,

$$g(k) = \frac{\mu + C_L}{p(k)} \sum_{i>k} p(i) - C_L$$

Equivalently,

$$p(k) = \frac{\mu + C_L}{\mu + 2C_L + g(k)} \prod_{i=0}^{k} \frac{g(i) + C_L}{g(i) + \mu + 2C_L}$$

For $C_L = 0$, the above becomes a statement for the sublinear regime.

Remarks

- Given a distribution $p(.)$, one can derive the expression for $g(.)$ required to obtain the desired distribution from the first relation.
- When $C_L = 0$, our relations become

  $$g(k) = \frac{\mu}{p(k)} \sum_{i>k} p(i)$$

  and

  $$p(k) = \frac{\mu}{\mu + g(k)} \prod_{i=0}^{k} \frac{g(i)}{g(i) + \mu}$$

  We will show that this is in line with the work of Dereich and Morters, and will give an explicit computation for μ when g is linear.

Superlinear Regime and remarks

Theorem

Consider a Preferential attachment process with the attachment law as in (1)*, with g linear. Let* $\Phi(t)$ *be superlinear, that is,* $\Phi(t)/t \to \infty$ *and* $\Phi(t)/t^n \to C_S$ *for some* $n > 1$ *for large* t*. Then,*

$$p(k) = \frac{1}{2^{k+1}}$$

- $p(.)$ has no dependence on f and g.
- The choice of fitness function does not matter as long as it lies in the superlinear regime.
- $p(.)$ does not follow a power law.

Concentration Results and Proofs

Ideas and results from Hofstad's text

Let $P_k(t)$ be the empirical degree distribution, that is, $P_k(t) = \frac{N_k(t)}{t+1}$.

Let $M_n = \mathbb{E}[N_k(t)|\mathcal{G}_n]$. Then, lemmas 8.5 and 8.6 from the text show that M_n is a martingale with bounded differences. Since $M_t = N_k(t) = (t+1)P_k(t)$ and $M_0 = \mathbb{E}N_k(t)$, one can use Azuma-Hoeffding to show

$$\mathbb{P}\left(|(t+1)P_k(t) - \mathbb{E}N_k(t)| \geq \sqrt{2}\sqrt{t \log t}\right) = \mathcal{O}(1/t)$$

Since intuitively we can define $p(k) = \lim_{t\to\infty} \frac{\mathbb{E}N_k(t)}{t}$, one can see that $|P_k(t) - p(k)| \leq \varepsilon_t \to 0$ for large t. This can be made more precise by working along the lines of proposition 8.7 from the text to show that $|\mathbb{E}N_k(t) - (t+1)p(k)| < C$ where $p(k)$ satisfies a certain recursion.

General Recursion

Let $N_k(t)$ denote the number of vertices of indegree k at time t. Then, for $k = 0$ and $k > 0$, we can write

$$N_0(t+1) = \begin{cases} N_0(t)+1, \text{ w.p.} \quad 1 - \frac{\sum_{u=0}^{t}(g(0)+f(u))\mathbb{1}_{\{d_u(t)=0\}}}{\sum_{u'=0}^{t} g(d_{u'}(t))+f(u')} \\ N_0(t), \text{ w.p} \quad \frac{\sum_{u=0}^{t}(g(0)+f(u))\mathbb{1}_{\{d_u(t)=0\}}}{\sum_{u'=0}^{t} g(d_{u'}(t))+f(u')} \end{cases} \tag{4}$$

$$N_k(t+1) = \begin{cases} N_k(t)+1, \text{ w.p.} \quad \frac{\sum_{u=0}^{t}(g(k-1)+f(u))\mathbb{1}_{\{d_u(t)=k-1\}}}{\sum_{u'=0}^{t} g(d_{u'}(t))+f(u')} \\ N_k(t)-1, \text{ w.p} \quad \frac{\sum_{u=0}^{t}(g(k)+f(u))\mathbb{1}_{\{d_u(t)=k\}}}{\sum_{u'=0}^{t} g(d_{u'}(t))+f(u')} \\ N_k(t), \text{ w.p } 1 - \frac{\sum_{u=0}^{t}(g(k-1)+f(u))\mathbb{1}_{\{d_u(t)=k-1\}}}{\sum_{u'=0}^{t} g(d_{u'}(t))+f(u')} - \frac{\sum_{u=0}^{t}(g(k)+f(u))\mathbb{1}_{\{d_u(t)=k\}}}{\sum_{u'=0}^{t} g(d_{u'}(t))+f(u')} \end{cases} \tag{5}$$

Taking expectation of the recurrences (4) and (5), we get

$$\mathbb{E}N_k(t+1) = \mathbb{E}N_k(t) + \mathbb{1}_{\{k=0\}} + \frac{g(k-1)\mathbb{E}N_{k-1}(t) - g(k)\mathbb{E}N_k(t)}{\mu_g(t) + \Phi(t)} + \frac{\sum_{u=0}^{t} f(u)\left(\mathbb{P}(d_u(t) = k-1) - \mathbb{P}(d_u(t) = k)\right)}{\mu_g(t) + \Phi(t)}$$

with the convention that $g(-1) = 0$. We can rewrite the above as

$$\mathbb{E}N_k(t+1) = \mathbb{E}N_k(t)\left(1 - \frac{1}{t}\frac{tg(k)}{\mu_g(t) + \Phi(t)}\right) + \mathbb{1}_{\{k=0\}} + \frac{g(k-1)\mathbb{E}N_{k-1}(t)}{\mu_g(t) + \Phi(t)} + \frac{\sum_{u=0}^{t} f(u)\left(\mathbb{P}(d_u(t) = k-1) - \mathbb{P}(d_u(t) = k)\right)}{\mu_g(t) + \Phi(t)} \tag{6}$$

Proof for the linear regime

Let $h(k) = \frac{g(k)}{\mu + C_L}$ where $\mu = \lim_{t \to \infty} \frac{\mu_g(t)}{t}$, and $\beta_k = \lim_{t \to \infty} \frac{1}{t} \sum_{u=0}^{t} f(u) \mathbb{P}(d_u(t) = k)$. For $k = 0$, define b_t and c_t as

$$b_t = \frac{t g(0)}{\mu_g(t) + \Phi(t)}$$

$$c_t = 1 - \frac{\sum\limits_{u=0}^{t} f(u) \mathbb{P}(d_u(t) = 0)}{\mu_g(t) + \Phi(t)}$$

Then, $b = \lim_{t \to \infty} b_t = h(0)$ and $c = \lim_{t \to \infty} c_t = 1 - \frac{\beta_0}{\mu + C_L}$.

One can see that (6) is of the form

$$a_{t+1} = a_t \left(1 - \frac{b_t}{t}\right) + c_t$$

where $a_t = \mathbb{E}N_k(t)$, $b_t = \frac{tg(k)}{\mu_g(t)+\Phi(t)}$ and

$c_t = \mathbb{1}_{\{k=0\}} + \frac{g(k-1)\mathbb{E}N_{k-1}(t)}{\mu_g(t)+\Phi(t)} + \frac{\sum_{u=0}^{t} f(u)\big(\mathbb{P}(d_u(t)=k-1)-\mathbb{P}(d_u(t)=k)\big)}{\mu_g(t)+\Phi(t)}$.

So, we use the following lemma

Lemma

Let $\{a_t\}_{t\geq 0}, \{b_t\}_{t\geq 0}$ and $\{c_t\}_{t\geq 0}$ be three real sequences such that $a_{t+1} = a_t\left(1 - \frac{b_t}{t}\right) + c_t$ with $b_t < t$ such that $\lim_{t\to\infty} b_t = b \geq 0$, and $\lim_{t\to\infty} c_t = c$. Then, $\lim_{t\to\infty} \frac{a_t}{t} = \frac{c}{1+b}$.

Thus, we get $p(0) := \lim_{t\to\infty} \frac{\mathbb{E}N_0(t)}{t} = \frac{1-\beta_0/(\mu+C_L)}{h(0)}$. Similarly, we get $p(k) = \frac{h(k-1)p(k-1)+\frac{\beta_{k-1}-\beta_k}{\mu+C_L}}{1+h(k)}$ for $k > 0$.

This implies,

$$\begin{aligned} h(k)p(k) &= h(k-1)p(k-1) - p(k) + \frac{\beta_{k-1} - \beta_k}{\mu + C_L} \\ &= \sum_{i>k} p(i) - \frac{\beta_k}{\mu + C_L} \end{aligned}$$

We know that $\beta_k = \lim_{t\to\infty} \frac{1}{t} \sum\limits_{u=0}^{t} f(u)\mathbb{P}(d_u(t) = k)$. Recall that $P_k(t)$ is the probability that a chosen vertex at time t has indegree k, and is precisely $\frac{N_k(t)}{t+1}$. So, using the fact that for large t, $|P_k(t) - p(k)| \to 0|$, we have

$$\beta_k = \lim_{t\to\infty} \frac{1}{t} \sum_{u=0}^{t} f(u)\mathbb{P}(d_u(t) = k) = \lim_{t\to\infty} \frac{P_k(t)}{t} \sum_{u=0}^{t} f(u)$$

$$= p(k). \lim_{t\to\infty} \frac{\Phi(t)}{t} = p(k)C_L$$

Substituting this yields us our desired expressions. For $C_L = 0$, we obtain the result for the sublinear case.

Superlinear Regime

We use a similar argument for the superlinear regime. Take

$$b_t = \frac{tg(0)}{\mu_g(t) + \Phi(t)}$$

Then, $b = 0$. Take

$$c_t = 1 - \frac{\sum_{u=0}^{t} f(u)\mathbb{P}(d_u(t) = 0)}{\mu_g(t) + \Phi(t)}$$

which gives us that $c = 1 - \frac{\beta_0}{C_S}$ where $\beta_k = \lim_{t\to\infty} \frac{1}{t^n} \sum_{u=0}^{t} f(u)\mathbb{P}(d_u(t) = k)$. So, we get $\beta_k = p(k)C_S$. Using lemma 3 by taking $a_t = \mathbb{E}N_0(t)$, we get $p(0) = 1/2$ and $p(k) = \frac{p(k-1)}{2}$.

Comparing to the work of Dereich and Morters

Take $g(x) = ax + b$. We know that $\mu = \lim_{t\to\infty} \mu_g(t)/t = a + b$. Let $a' = a/(a+b)$ and $b' = b/(a+b)$. Taking $g'(k) = g(k)/\mu$, we have

$$p(k) = \frac{1}{1+g'(k)} \prod_{i=0}^{k} \frac{g'(i)}{1+g'(i)} = \frac{1}{a'} \frac{\prod_{i=0}^{k-1} i + b'/a'}{\prod_{i=0}^{k} i + (b'+1)/a'}$$

$$= \frac{1}{a'} \frac{\Gamma\left(k + \frac{b'}{a'}\right)\Gamma\left(\frac{1+b'}{a'}\right)}{\Gamma\left(\frac{b'}{a'}\right)\Gamma\left(k + \frac{1+b'+a'}{a'}\right)} \xrightarrow[\text{Stirling's approx.}]{k\to\infty} \frac{1}{a'} \frac{\Gamma\left(\frac{1+b'}{a'}\right)}{\Gamma\left(\frac{b'}{a'}\right)} k^{-\left(1+\frac{1}{a'}\right)}$$

where $a', b' \in (0, 1]$, which is in line with example 1 from the paper. Taking $a' = 1$ gives us the power law exponent of '-2'.

What happens when f is random?

Suppose $f(u)$ are independent random variable such that $\mathbb{E}[f(u)] = \gamma_u$ and $\mathbb{P}(f(u) \in [a, b]) = 1 \ \forall i$ for some $a, b \in \mathbb{R}$. Using Chernoff bound, we get

$$\mathbb{P}\left(|\sum_{u=0}^{t} f(u) - \gamma| \geq (b-a)\sqrt{t \log t} \right) \leq \mathcal{O}\left(\frac{1}{t}\right),$$

where $\gamma = \sum_{u=0}^{t} \gamma_u$. That is, for large t,
$\Phi(t) = \sum_{u=0}^{t} f(u) = \gamma + \mathcal{O}\left(\sqrt{t \log t}\right)$.

Let $f(u) \sim Poi(\lambda_u)$ be independent random variables for all $u \geq 0$ and $g(k) = ak + b$. We have $\Phi(t) = \sum\limits_{i=1}^{t} f(i) \sim Poi(\sum\limits_{i=1}^{t} \lambda_i)$. Then, we get

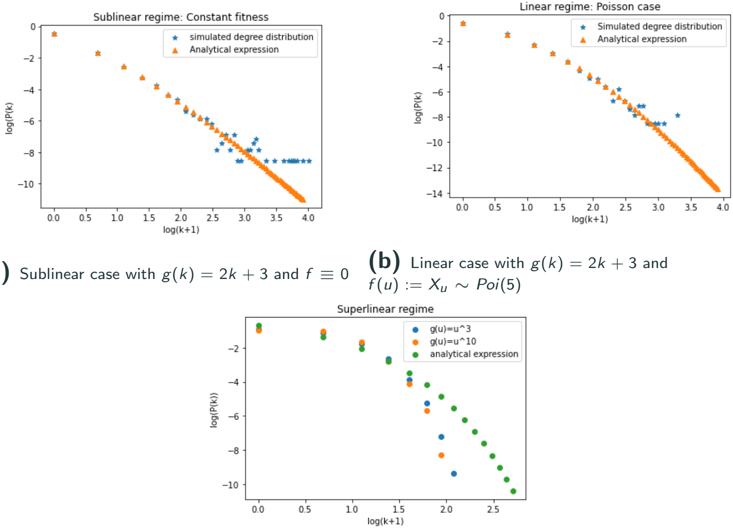
$$\mathbb{P}\left(\left|\Phi(t) - \sum_{i=0}^{t} \lambda_i\right| \geq \delta\right) \leq 2\exp\left\{-\frac{c\delta^2}{\sum\limits_{i=1}^{t} \lambda_i}\right\}$$

Let $\Lambda = \sup_i \lambda_i$. Then, $\sum\limits_{i=1}^{t} \lambda_i \leq t\Lambda$. That is, $-\frac{1}{\sum\limits_{i=1}^{t} \lambda_i} \leq -\frac{1}{t\Lambda}$.

Taking $\delta = \sqrt{\frac{\Lambda}{c} t \log t}$, we get

$$\mathbb{P}\left(\left|\Phi(t) - \sum_{i=1}^{t} \lambda_i\right| \geq \sqrt{\frac{\Lambda}{c} t \log t}\right) = \mathcal{O}\left(\frac{1}{t}\right)$$

and thus we can write $\sum\limits_{u=0}^{t} f(u) = \sum\limits_{u=0}^{t} \lambda_u + \mathcal{O}\left(\sqrt{t \log t}\right)$



(a) Sublinear case with $g(k) = 2k + 3$ and $f \equiv 0$

(b) Linear case with $g(k) = 2k + 3$ and $f(u) := X_u \sim Poi(5)$

(c) Superlinear cases with $g(k) = 2k + 3$ and $f(u) = u^3$ and $f(u) = u^{10}$

Figure 2: A comparison of simulated and analytical degree distributions

Working with a general g

The idea is to assume the following:

- g has bounded differences, that is, $|g(i+1) - g(i)| < K$
- $\mu_g(t) := \sum_j g(j)\mathbb{E}N_j(t)$ is such that $\mu_g(t)/t < \infty$.

Lemma 4 from [**?**] shows the following:

Lemma

Suppose the assumption that $|g(i+1) - g(i)| < K$ *is true. Then,*

$$\mathbb{P}\left(\left|\sum_{u=0}^{t} g(d_u(t)) - \mu_g(t)\right| \geq \sqrt{32K^2 t \log t}\right) = \mathcal{O}(1/t^4)$$

The idea is to use this in recursions (4) and (5). However, computing $\mu := \lim_{t\to\infty} \mu_g(t)/t$ is not trivial.

Thank You!