Learning growth mechanisms of tail realistic preferential attachment models from network degree distributions

Thomas William Boughen[1], Clement Lee[1], Vianey Palacios Ramirez[1]

[1]School of Mathematics, Statistics and Physics, Newcastle University.

Abstract

Devising the underlying generating mechanism of a real-life network is difficult as, more often than not, only its snapshots are available, but not its full evolution. One candidate for the generating mechanism is preferential attachment which, in its simplest form, results in a degree distribution that follows the power law. Consequently, the growth of real-life networks that roughly display such power-law behaviour is commonly modelled by preferential attachment. However, the validity of the power law has been challenged by the presence of alternatives with comparable performance, as well as the recent findings that the right tail of the degree distribution is often lighter than implied by the body, whilst still being heavy. In this paper, we study a modified version of the model with a flexible preference function that allows super/sub-linear behaviour whilst also guaranteeing that the limiting degree distribution has a heavy tail. We relate the distributions tail index directly to the model parameters, allowing direct inference of the parameters from the degree distribution alone.

Keywords: networks, discrete extremes, power law, preferential attachment

1 Introduction

Networks have become very powerful tools for representing and analysing complex systems, with uses in a large array of fields. In network science and statistics, they have been studies by various families of models, from stochastic block models for detecting communities online (**?**), to exponential random graph models (ERGMs) for analysing the global trade network (**?**), and mechanistic models for investigating patterns in neural systems (**?**).

Amid the recent rise of interest in networks, there has been a debate on whether most real networks are scale-free. Claiming a real network is scale-free is equivalent to saying that its degree distribution follows a power law, that is the fraction of nodes with degree k is proportional to $k^{-\alpha}$, and therefore has a regularly varying tail with tail index α. On the side against the claim is **?** compared the fits of a power law model against that of several non-scale-free models to nearly a thousand networks, only to find strong evidence for scale-freeness in four percent and weak evidence in over half of the networks, thus claiming that scale-free networks do not make up a majority in real networks. They compare the fits of a power law model against that of several non-scale-free models only finding strong evidence for scale-freeness in four percent and weak evidence in fifty-two percent of the networks. On the other side of this debate is **?** who disagrees and claims that these networks are not nearly as rare and only appear to so as a result of an unrealistic expectation of a power law without deviations or noise. Additional evidence of these deviations from a power law is shown by **?** who demonstrate that a lot of networks are partially scale-free, in that the body of the degree distribution is often modelled well by a power law, while the tail is lighter than what is implied by the body, albeit still regularly varying..

Nevertheless, most studies into the appropriateness of a power law for the degrees of real networks, the aforementioned references included, have been largely descriptive in the sense that no information about the growth of the networks is revealed.

The popularity of using the power law for network degrees can be traced back to the preferential attachment (PA) model popularised by **?**. In the general model, as new nodes join the network, an existing node with degree k gains edges at a rate proportional to $b(k)$, where b is a non-negative preference function. **?** showed that, when $b(k) = k + 1$, in the limit the resulting degree distribution is regularly varying with index 3. Subsequently, if a real network is shown to be scale-free, one can loosely justify preferential attachment as the underlying mechanism of its growth.

The model from **?** provided the foundations for various generalisations.

? considered $b(k) = (k+1)^\alpha$, and showed that the degree distribution is not regularly varying (and therefore not following the power law) when $0 < \alpha < 1$, and when $\alpha > 1$ a finite number of nodes end up with all edges after a certain point resulting in a degenerate degree distribution. **?** returns to a linear preference function of the form $b(k) = k + \varepsilon$ but adds the possibility for reciprocal edges to be sent, resulting in the joint distribution of in-degree and out-degree being multivariate regularly varying and having the property of hidden regular variation. **?** follows in the footsteps of

? , by considering a preferential attachment tree and using theory from continuous branching processes, derives a limiting degree distribution in terms of the preference function $b(\cdot)$. Nevertheless, research in this area tends to only focus on the theoretical asymptotic results of network growth models with little analysis of real networks.

This paper aims to address the gap between the applied and theoretical works, by asking if a network is assumed to come from a preferential attachment model, can we use the degree distribution alone to directly infer the parameters of the the preference function and learn about the growth mechanisms? Moreover, proper consideration is given to the tail of the degree distribution, because otherwise the effects of the largest degrees, which correspond to the most influential nodes, deviating from the power law will be discounted.

It is important to think about how we intend to consider the tail of the degree distributions, because if not done correctly we may end up essentially discounting the effects of the largest degrees (often the most influential nodes) deviating from a power law.

For the above reason, we will use methods from discrete extreme value theory, in particular those by **?**, who provided theoretical guarantees for a discrete distribution to be regularly varying or not. Using these results, we demonstrate how the tail of the degree distribution is affected by $b(\cdot)$, and subsequently propose a class of preference functions that is tail realistic for real networks. These analytical results enables the likelihood of the degree distribution to be expressed in terms of the parameters of $b(\cdot)$, which in turn allows the underlying mechanism of the network, assumed to grow according to preferential attachment, to be inferred directly.

The remainder of the paper is as follows: Section **??** gives a detailed description of the preferential attachment model alongside the theoretical results for the survival of the limiting degrees, with a focus on the tail behaviour in terms of the preference function $b(\cdot)$. A class of asymptotically linear preference functions will be introduced and shown to guarantee regular variation in the degrees while remaining flexible up until a threshold. Section **??** utilises the proposed preference function and illustrates numerically how the tail index of the degree distribution varies with the model parameters. Section **??** consists of a simulation study where networks are simulated from the proposed model for various parameter combinations, demonstrating that the parameters can be recovered from fitting the model to only the degree distribution. Section **??** fits the model to some real data and provides posterior estimates for the preference function. Section **??** concludes the article.

2 Tail Behaviour of Preferential Attachment Model

The model that we will focus on in this paper is the General Preferential Attachment model in **?** and is defined as follows:

Starting at time $t = 0$ with an initial network of m vertices that each have no edges, at times $t = 1, 2, ...$ a new vertex is added to the network bringing with it m directed edges from the new vertex; the target for each of these edges are selected from the vertices already in the network with weights proportional to some non-decreasing preference function $b(\cdot)$ of their degree, where $b : \mathbb{N} \mapsto \mathbb{R}^+ \setminus \{0\}$ is such that:

$$\sum_{k=0}^{\infty} \frac{1}{b(k)} = \infty. \tag{1}$$

Special cases of this model include the Barabási-Albert (BA) model when $b(k) = k+\varepsilon$, which leads to a power-law degree distribution with tail index 2, and the Uniform Attachment (UA) model where $b(k) = c$ leading to a degree distribution that is not regularly varying.

The survival function of the limiting degree distribution, called the limiting survival hereafter, under condition **??** can be analytically derived in the case where $m = 1$, this is presented below.

Consider a continuous time branching process $\zeta(t)$ driven by a Markovian pure birth process, with $\zeta(0) = 0$ and birth rates depending on a non-negative function $b(\cdot)$:

$$\Pr(\zeta(t + dt) = k + 1 | \zeta(t) = k) = b(k)dt + o(dt).$$

Now let $\Upsilon(t)$ be the tree determined by $\zeta(t)$ as follows: $\Upsilon(t) = \{\emptyset\}$ and $\Upsilon(t) = G$ where each existing node x in $\Upsilon(t)$ gives birth to a child with rate $b(\deg(x, \Upsilon(t))$ independently of the other nodes where $\deg(x, \Upsilon(t))$ is the degree of node x in the tree $\Upsilon(t)$ at time t.

Theorem 1 from **?** states that for the tree $\Upsilon(t)$ at time t and a characteristic function of the tree $\varphi(\cdot)$:

$$\lim_{t\to\infty} \frac{1}{|\Upsilon(t)|} \sum_{x\in\Upsilon(t)} \varphi(\Upsilon(t)_{\downarrow x}) = \lambda^* \int_0^{\infty} e^{-\lambda^* t} \mathbb{E}\left[\varphi(\Upsilon(t))\right] dt \tag{2}$$

where λ^* satisfies $\hat{\rho}(\lambda^*) = 1$ and $\hat{\rho}$ is the Laplace transform of the density of the point process associated with the pure birth process that corresponds to the growth of an individual node, that is $\hat{\rho}(\lambda) := \int_0^{\infty} e^{-\lambda t} \rho(t) \mathrm{d}t$

The limiting survival can be viewed as the limit of the empirical proportion of vertices with degree over a threshold $k \in \mathbb{N}$, that is:

$$\bar{F}(k) = \lim_{t\to\infty} \frac{\sum_{x\in\Upsilon(t)} \mathbb{I}\left\{\deg(x, \Upsilon(t)_{\downarrow x}) > k\right\}}{\sum_{x\in\Upsilon(t)} 1}$$

which can also be written using Equation **??** as:

$$\bar{F}(k) = \frac{\int_0^\infty e^{-\lambda^* t}\mathbb{E}\left[\mathbb{I}\left\{\deg(x, \Upsilon(t)) > k\right\}\right] dt}{\int_0^\infty e^{-\lambda^* t} dt} = \prod_{i=0}^{k} \frac{b(i)}{\lambda^* + b(i)}. \tag{3}$$

Therefore, the corresponding probability mass function of the degree distribution $f(k) = \bar{F}(k-1) - \bar{F}(k)$ is

$$f(k) = \frac{\lambda^*}{\lambda^* + b(k)} \prod_{i=0}^{k-1} \frac{b(i)}{\lambda^* + b(i)}. \tag{4}$$

We are interested in how the tail behaviour of the discrete limiting degree distribution is affected by the preference function b.

For a distribution F with survival function $\bar{F}$ and $k \in \mathbb{Z}^+$ let

$$\Omega(F, k) = \left(\log \frac{\bar{F}(k+1)}{\bar{F}(k+2)}\right)^{-1} - \left(\log \frac{\bar{F}(k)}{\bar{F}(k+1)}\right)^{-1}.$$

This allows us to show the following:

Proposition 2.1. *If* $\bar{F}(k) = \prod_{i=0}^{k} \frac{b(i)}{\lambda^* + b(i)}$ *and* $b(k) \to \infty$ *as* $k \to \infty$*, then*

$$\lim_{k\to\infty} \Omega(F, k) = \lim_{k\to\infty} \frac{b(k+1) - b(k)}{\lambda^*}.$$

See Appendix **??** for the details of the proof.

? states that if $\lim_{k\to\infty} \Omega(F, k) = 1/\alpha$ $(\alpha > 0)$, then F is regularly varying with $\bar{F}(k) \sim k^{-\alpha}$. On the other hand, if $\lim_{k\to\infty} \Omega(F, k) = 0$ then we will refer to the distribution as light-tailed.

Proposition **??** aligns with the result from **?** demonstrating that a sub-linear preference function will lead to a light-tailed distribution, as $\lim_{k\to\infty} b(k+1) - b(k) = 0$ if $b(k) = k^\alpha$ where $\alpha < 1$. Proposition 2.1 also aligns with the fact that BA model produces a regularly varying degree distribution, with tail index 2, by considering the preference function $b(k) = k + \varepsilon$, as $\lim_{k\to\infty} b(k+1) - b(k) = 1$ leaving the tail index to be $1/\lambda^*$ which using $\hat{\rho}$ can be found to be $1/2$. So, in order for the degree distribution to be regularly varying we need that the limit $\lim_{k\to\infty} b(k+1) - b(k)$ exists and is positive. The following proposition determines the class of functions that will result in regular varying degree distributions.

Proposition 2.2. *Consider a GPA model with preference function* $b(\cdot)$ *satisfying Proposition* **??**. *Then the limiting degree distribution is regularly varying with tail index* λ^*/c *if and only if* $\lim_{k\to\infty} b(k)/k = c > 0$.

Proof

From Proposition **??**, *we have that:*

$$\lim_{k\to\infty}[b(k+1)-b(k)] = c > 0.$$

Now, setting $b_k = b(k)$ *and* $a_k = k$*:*

$$\lim_{k\to\infty}[b(k+1)-b(k)] = \lim_{k\to\infty}\frac{b_{k+1}-b_k}{a_{k+1}-a_k} = c > 0,$$

Using the Stolz-Cesaro Theorem @thm-stolz,

$$\lim_{k\to\infty}\frac{b_k}{a_k} = \lim_{k\to\infty}\frac{b(k)}{k} = c > 0 \qquad \square.$$

Using this result we can understand how the preference function is directly connected with the tail behaviour of the degree distribution. Specifically, regular variation is achieved if and only if $b(k)$ is asymptotically linear with k. We use this result to create a preference function that guarantees regular variation in the tail of the degree distribution, aligning with analysis of real networks, whilst allowing for the tail to deviate from the shape implied by the body. This gives the model the capability to produce realistic behaviour in the degrees like what was in **?** by using a piecewise function inspired by the observed deviation from the power law after a certain threshold:

$$b(k) = \begin{cases} k^\alpha + \varepsilon, & k < k_0, \\ k_0^\alpha + \varepsilon + \beta(k-k_0), & k \geq k_0 \end{cases} \tag{5}$$

for $\alpha, \beta, \varepsilon > 0$ and $k_0 \in \mathbb{N}$.

Using Proposition **??**, we can show that $\lim_{n\to\infty}\Omega(F,k) = \lambda^*/\beta$, meaning the degree distribution obtained using this preference funnction is regularly varying with tail index λ^*/β.

3 A Preferential Attachment Model with Flexible Regular Variation

In the previous section, we found that using an asymptotically linear preference function allows for the inclusion of sub/super-linear behaviour below the threshold, while simultaneously guaranteeing regular varation of the degrees. In this section, we demonstrate the flexibility of the preference function in Equation **??**, with regard to the tail behaviour of the limiting degree distribution. Using Equation **??**, the limiting survival is

$$\bar{F}(k) = \begin{cases} \prod_{i=0}^{k} \frac{i^\alpha+\varepsilon}{\lambda^*+i^\alpha+\varepsilon}, & k < k_0, \\ \left(\prod_{i=0}^{k_0-1} \frac{i^\alpha+\varepsilon}{\lambda^*+i^\alpha+\varepsilon}\right) \frac{\Gamma(\lambda^*+k_0^\alpha+\varepsilon)/\beta)}{\Gamma((k_0^\alpha+\varepsilon)/\beta)} \frac{\Gamma\left(k-k_0+1+\frac{k_0^\alpha+\varepsilon}{\beta}\right)}{\Gamma\left(k-k_0+1+\frac{\lambda^*+k_0^\alpha+\varepsilon}{\beta}\right)}, & k \geq k_0, \end{cases} \tag{6}$$

with λ^* satisfying $\hat{\rho}(\lambda^*) = 1$ where

$$\hat{\rho}(\lambda) = \sum_{n=0}^{k_0} \prod_{i=0}^{n-1} \frac{i^\alpha+\varepsilon}{\lambda+i^\alpha+\varepsilon} + \left(\frac{k_0^\alpha+\varepsilon}{\lambda-\beta}\right) \prod_{i=0}^{k_0-1} \frac{i^\alpha+\varepsilon}{\lambda+i^\alpha+\varepsilon} \tag{7}$$

which has to be solved numerically for most parameter choices. Also, note that $\lambda > \beta$.

For some parameter combinations, the limiting survival $\bar{F}(k)$ is shown on log-log scale in Figure **??**:

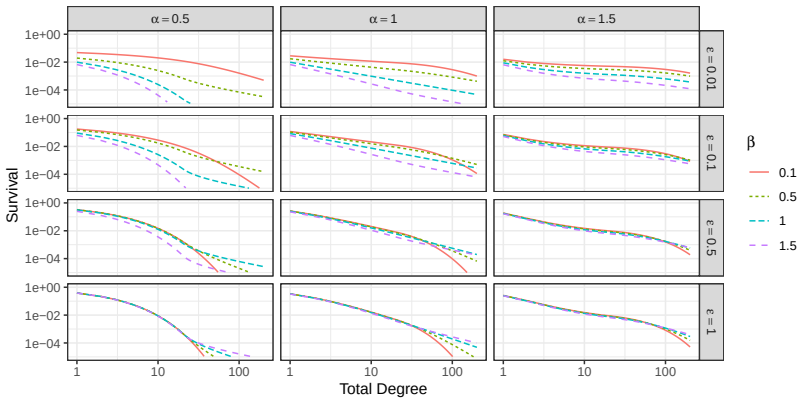

Figure 1: The limiting survival, according to various combinations of $(\alpha, \beta, \varepsilon)$ and $k_0 = 20$ of the proposed preferential attachment model.

Figure **??** demonstrates that this model can capture a range of tail behaviour, including a large range of possible tail indices ranging from 0.035 ($\alpha = 1.5, \beta = 0.1, \varepsilon = 1$) to 0.999 ($\alpha = 0.5, \beta = 1.5, \varepsilon = 0.01$).