INFERENCE FOR PREFERENTIAL ATTACHMENT MODELS AND RELATED TOPICS

by

Daniel Saxton

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
In Partial fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

______________________________ Dr. Anand Vidyashankar, Dissertation Director

______________________________ Dr. Daniel Carr, Committee Member

______________________________ Dr. Guoqing Diao, Committee Member

______________________________ Dr. Yunpeng Zhao, Committee Member

______________________________ Dr. Huzefa Rangwala, Committee Member

______________________________ Dr. William F. Rosenberger, Department Chair

______________________________ Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering

Date: _________________________ Spring Semester 2014
George Mason University
Fairfax, VA

Inference for Preferential Attachment Models and Related Topics

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Daniel Saxton
Master of Science
George Mason University, 2011
Bachelor of Science
George Mason University, 2007

Director: Dr. Anand Vidyashankar, Professor
Department of Statistics

Spring Semester 2014
George Mason University
Fairfax, VA

Acknowledgments

I would like to express my appreciation to my dissertation advisor Dr. Anand Vidyashankar for his continued support, extensive knowledge, and generosity with his time. I would also like to thank all members of my doctoral committee, Dr. Daniel Carr, Dr. Guoqing Diao, Dr. Jacqueline Hughes-Oliver, Dr. Huzefa Rangwala and Dr. Yunpeng Zhao, as well as our department chair Dr. William Rosenberger, Dr. Clifton Sutton and Dr. Linda Davis, for giving me the opportunity to study in this department.

Table of Contents

List of Tables

List of Figures

Abstract

INFERENCE FOR PREFERENTIAL ATTACHMENT MODELS AND RELATED TOPICS

Daniel Saxton, PhD

George Mason University, 2014

Dissertation Director: Dr. Anand Vidyashankar

Preferential attachment models arise in several areas of mathematics and scientific applications such as in the analysis of social, financial, and gene regulatory networks. However, inferential questions related to such models are challenging and have so far not been addressed. In this dissertation, we provide a framework using branching processes within which to investigate these issues. In particular, we develop theory that may be employed to extract information about the strength of preferential attachment from graph data, as well as information about degree asymptotics. We also study an extension of the model incorporating random effects which helps to introduce added heterogeneity into the process which is not represented in existing models. Questions concerning cascades on trees are also studied.

Chapter 1: Introduction

1.1 Preferential attachment

1.1.1 Basic setting and problem

Stochastic processes with positive reinforcement have been studied in great detail over the past century. Generally speaking, these are processes where the occurrence of some event renders the event more likely to occur in the future, the classic example of which being the Pólya urn scheme [12]. Over the past ten years, there has been renewed interest in these processes with the introduction by Barabasi and Albert [1] of what is now known as the preferential attachment model. This model is a random graph process which begins with some number of vertices connected by edges, where at each discrete time step a new vertex joins the graph and connects randomly to one of the existing vertices. The key feature of the preferential attachment model is that the newly joining vertex is more likely to connect to vertices that are already well-connected, and Barabasi and Albert showed through simulation and heuristic arguments that this simple mechanism was enough to produce the kinds of power-law behavior that are ubiquitous in real world networks. (Figure 1.1 shows the first several steps in the evolution of a hypothetical preferential attachment tree.)

While the existence of preferential attachment has been postulated as an explanation for several properties of real world networks, there currently exists no statistical framework for detecting or measuring the strength of such a mechanism. The purpose of this dissertation is to begin to develop such theory through estimation of a parameter that arises naturally in these models, one which provides insight into the nature of the preferential attachment mechanism at work in a given graph process.

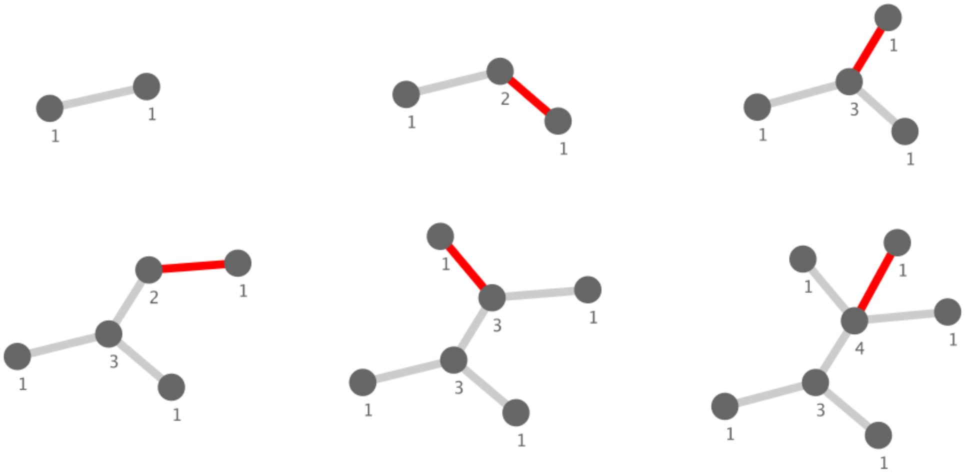

Figure 1.1: Evolution of a preferential attachment graph. The degrees are indicated underneath each node.

Another drawback of the standard model is that it implicitly assumes that two vertices with equal degree are indistinguishable. However, this is clearly an oversimplification of networks in nature. To overcome this, we introduce an extension of the existing model called preferential attachment with random effects which tries to incorporate other vertex features aside from connectivity.

Possible applications of this research would be in settings where a linear preferential attachment has been assumed and one wishes to test the strength of this mechanism. Or, if one wished to improve the fit of a preferential attachment model by the inclusion of random effects to account for other vertex-specific traits.

1.1.2 Previous work

Though Barabasi and Albert were the first to study the preferential attachment model, it was not well-defined until the work of Bollobas and Riordin, who also established the first rigorous results for the model. For instance, they were able to formally derive the power-law asymptotics for the degree distribution (the degree of a vertex being the number of edges incident to it), and also show that the diameter of the graph, defined as the maximum distance between vertices, grows logarithmically in the limit.

Since then many variants of the preferential attachment model have been proposed including one by Athreya, Sethuraman and Ghosh [3] which is constructed as follows. At time $n = 0$ initialize a graph consisting of two vertices connected by a single edge, denoted G_0. Then, given the graph G_n at time $n \geq 0$, add a new vertex and have it connect X_{n+1} times to the i^{th} vertex, $1 \leq i \leq n+2$, with probability equal to,

$$\frac{d_n(i) + \beta}{\sum_{j=1}^{n+2}(d_n(j) + \beta)},$$

where $d_i(n)$ is the degree of vertex i at time n, and $\beta \geq 0$ is a constant. The $\{X_i\}_{i=1}^{\infty}$ are taken to be i.i.d. positive integer-valued random variables with distribution $\{p_j\}_{j=1}^{\infty}$, and $m \equiv \sum_{j=1}^{\infty} j p_j < \infty$.

Their main result is the following.

Theorem 1.1. *Set $\theta \equiv m/(2m + \beta)$ and suppose that $\sum_{j=1}^{\infty} j \log(j)\, p_j < \infty$. Then,*

(i) For each $i \geq 1$ there exists a random variable $\gamma_i \in (0, \infty)$ such that,

$$\lim_{n \to \infty} \frac{d_n(i)}{n^{\theta}} \overset{\text{a.s.}}{=} \gamma_i.$$

(ii) If we denote by M_n the maximal degree at time n, and I_n the index at which this maximum is attained (i.e., $d_n(I_n) = M_n$), then so long as $\sum_{j=1}^{\infty} j^r p_j$ for some $r > 1/\theta$ we have $I_n \xrightarrow{\text{a.s.}} I \in \mathbb{N}_+$ as $n \to \infty$ and,

$$\lim_{n \to \infty} \frac{M_n}{n^{\theta}} \overset{\text{a.s.}}{=} \max_{i \geq 1} \gamma_i. \ \square$$

This states that the degree sequence of each vertex grows at an asymptotic rate of n^{θ} so long as $\sum_{j=1}^{\infty} j \log(j) p_j < \infty$ (if $\sum_{j=1}^{\infty} j \log(j) p_j = \infty$ then $\gamma_i \overset{\text{a.s.}}{=} 0$ for every $i \geq 1$), and that under a slightly stronger moment condition the same holds true of the maximal degree. In particular, under the pure preferential attachment regime where $\beta = 0$ the degree sequences grow at a rate of $\sqrt{n}$. In addition, they show that the parameter θ is related to the scaling exponent of the asymptotic power law distribution, with larger values corresponding to degree distributions with heavier tails. Hence, θ contains significant amounts of information about the nature of the process, and an ability to estimate this parameter is of much value.

Athreya [2] also considered what happens when the attachment function is not necessarily a simple linear function of the degree, but rather that the probability of attachment is proportional to some more general function $f : \mathbb{N}_+ \mapsto \mathbb{R}_+$ applied to the degree of a given vertex. He showed that certain features of the process such as polynomial growth of the degrees and power law asymptotics for the degree distribution no longer hold when the attachment function deviates from the simple linear setting. Specifically, when the attachment is asymptotically sublinear, defined as $\lim_{n\to\infty} f(n)/cn^p \to 1$ for some $c > 0$ and $1/2 < p < 1$, the degrees now grow roughly at a rate of $\log(n)^q$, where $q \equiv 1/(1-p)$. In the asymptotically superlinear case, defined as $\sum_{n=1}^{\infty} f(n)^{-1} < \infty$, then either all degree sequences converge almost surely if the preference function is not too strong, or each vertex will with positive probability eventually receive all edges from newly arriving vertices. However, in the latter case he leaves open the question as to whether or not the probability that this happens for some vertex is one.

The primary tool used to analyze the model with linear weight function is that of continuous time Markov branching processes, and pure birth Markov processes in the case of preferential attachment with general weight function. These tools are the topics of the next two sections.

1.2 Branching processes

A Markov branching process $\{Z(t) : t \geq 0\}$ is a continuous time Markov chain with state space $\mathbb{S} = \{0, 1, 2, \ldots\}$, lifetime parameter $0 < \alpha < \infty$, and sojourn time parameter $\alpha_i = i\alpha$. The process can be imagined as a population initialized by a single founder at time $t = 0$ who lives for an exponential(α) length of time and at death produces a number of descendants according to the distribution $\{p_j\}_{j\geq 0}$, typically referred to as the offspring distribution. After splitting, each descendent is an i.i.d. copy of the parent, and the process continues as $t \to \infty$. If we set $\lambda \equiv \alpha(m-1)$, then it can be shown that $\mathrm{E}[Z(t)] = e^{\lambda t}$ [6] and one immediately gets the following theorem.

Theorem 1.2. *$\{Z(t)e^{-\lambda t} : t \geq 0\}$ is a non-negative martingale and hence $\lim_{t\to\infty} Z(t)e^{-\lambda t} \equiv W$ exists and is finite almost surely.*

Proof: *First for any $t \geq 0$, $\mathrm{E}[|Z(t)e^{-\lambda t}|] = \mathrm{E}[Z(t)e^{-\lambda t}] = 1 < \infty$. Now letting $\{Z^{(i)}\}$ denote i.i.d. copies of Z we have by the branching property [6] that for $0 < s < t$,*

$$\begin{aligned}
\mathrm{E}\left[Z(t)e^{-\lambda t} \,|\, Z(s)\right] &= e^{-\lambda t}\mathrm{E}\left[\sum_{i=1}^{Z(s)} Z^{(i)}(t-s) \,|\, Z(s)\right] \\
&= e^{-\lambda t} Z(s)\mathrm{E}\left[Z^{(1)}(t-s)\right] \\
&= e^{-\lambda t} Z(s) e^{\lambda(t-s)} \\
&= Z(s)e^{-\lambda s}, \;\; w.p.\, 1.
\end{aligned}$$

That $\lim_{t\to\infty} Z(t)e^{\lambda t} \equiv W$ exists and is finite a.s. follows from Doob's first convergence theorem [13] and the non-negativity of $Z(t)$ ($Z(t) \geq 0$ implies $\mathrm{E}[|Z(t)e^{-\lambda t}|]$ is constant, and hence $\sup_{t\geq 0} \mathrm{E}[|Z(t)e^{-\lambda t}|] < \infty$). □

While the previous theorem establishes the existence of W, it may be that $W \overset{\text{a.s.}}{=} 0$. This issue is addressed by the following theorem (found in [6]), which provides a necessary and sufficient condition for the uniform integrability of $\{Z(t)e^{-\lambda t}\}_{t\geq 0}$.

Theorem 1.3. *Let W be as above and assume $\lambda > 0$. If $\sum_{j=1}^{\infty} j \log(j)\, p_j < \infty$ then $\mathrm{E}(W) = 1$, otherwise $\mathrm{E}(W) = 0$.* □

Notice that it can occur that $Z(t') = 0$ for some $t' \in \mathbb{R}_+$, and if this is the case then $Z(t) = 0$ for all $t \geq t'$. This is referred to as the extinction of the process, and whenever $p_0 > 0$ this event has positive probability. To deal with this issue, one may choose to introduce an auxiliary immigration component into the process. That is, we imagine that there exists an infinite collection $\{Z_i(t)\}_{i=0}^{\infty}$ of i.i.d. copies of $Z(t)$ which are combined with the original population according to a Poisson process with rate $0 < \beta < \infty$. A Markov branching process with immigration $\{D(t) : t \geq 0\}$ can then be written as $D(t) = \sum_{i=0}^{\infty} Z_i(t - T_i)I(T_i \leq t)$, where $\{T_i\}_{i=1}^{\infty}$ are given by the sums $T_i = \sum_{j=1}^{i} L_j$, the $\{L_i\}_{i=1}^{\infty}$ being i.i.d. exponential(β), and $T_0 := 0$.

The idea behind the proofs in [3] is to construct a collection of branching processes which have the same distribution as the collection of degree sequences within a preferential attachment tree, and thus many results which hold for the former will also hold for the latter. The embedding works by first initializing two i.i.d. Markov branching processes $D_1(t)$ and $D_2(t)$ with $D_1(0), D_2(0) := 1$. These processes have lifetime and immigration time parameters $\lambda = 1$ and $\beta \in [0, \infty)$ respectively, and immigration and offspring particle distributions $\{p_j\}_{j=1}^{\infty}$ and $\{p'_j\}_{j=2}^{\infty}$ where $p'_j = p_{j-1}$ for $j \geq 2$. (That is, regardless of whether the event is immigration or the death of a particle, the net addition to the process where the event occurs is distributed according to $\{p_j\}_{j=1}^{\infty}$.) Next we wait for an event to happen in $D_1(t)$ or $D_2(t)$ (immigration or the death of a particle), let τ_1 denote the random time at which this event occurs, and start the new process $D_3(t)$ with $D_3(0) = X_1$ at time τ_1, where X_1 is the net addition of particles to the process within which the event occurred. Continue in this manner for $n \geq 2$ and construct the collection $\{D_k(t)\}_{k=1}^{\infty}$ along with the increasing sequence of event times $\{\tau_n\}_{n=1}^{\infty}$. (It will also be notationally convenient to set $\tau_{-1} = \tau_0 = 0$.)

Before reproducing the embedding theorem, we first prove two useful properties of the exponential distribution on which this result will depend.

Lemma 1.1. *Let $\{Y_i\}_{i=1}^{n}$ be independent exponential random variables with respective rates $\{\lambda_i\}_{i=1}^{n}$. Then $\min\{Y_1, Y_2, \ldots, Y_n\}$ is distributed exponential with rate $\sum_{i=1}^{n} \lambda_i$.*

Proof: *We can assume without loss of generality that $n = 2$ since the general case follows from this by repeatedly taking pair-wise minima; i.e.,* $\min\{Y_1, \ldots, Y_n\} = \min\{\min\{Y_1, Y_2\}, \ldots, Y_n\}$*, etc. Now for $t \geq 0$,*

$$\begin{aligned} P(\min\{Y_1, Y_2\} \leq t) &= 1 - P(\min\{Y_1, Y_2\} > t) \\ &= 1 - P(Y_1 > t)P(Y_2 > t) \\ &= 1 - [1 - P(Y_1 \leq t)][1 - P(Y_2 \leq t)] \\ &= 1 - e^{-\lambda_1 t}e^{-\lambda_2 t} \\ &= 1 - e^{-(\lambda_1+\lambda_2)t}. \ \square \end{aligned}$$

Lemma 1.2. *Let $\{Y_i\}_{i=1}^n$ be independent exponential random variables with rates $\{\lambda_i\}_{i=1}^n$. Then, for $1 \leq i \leq n$ $P(\min\{Y_1, \ldots, Y_n\} = Y_i\} = \lambda_i / \sum_{j=1}^n \lambda_j$.*

Proof: *Since* $\min\{Y_1, \ldots, Y_n\} = \min\{\min\{Y_j\}_{j\neq i}, Y_i\}$ *and* $\min\{Y_j\}_{j\neq i} \sim$ *exponential*$(\sum_{j\neq i} \lambda_j)$ *we again only need to consider the case $n = 2$.*

$$\begin{aligned} P(\min\{Y_1, Y_2\} = Y_1) &= P(Y_1 \leq Y_2) \\ &= \mathrm{E}[P(Y_1 \leq Y_2 \,|\, Y_2)] \\ &= \mathrm{E}\left(1 - e^{-\lambda_1 Y_2}\right) \\ &= 1 - \int_0^\infty e^{-\lambda_1 x}\lambda_2 e^{-\lambda_2 x}\, dx \\ &= 1 - \frac{\lambda_2}{\lambda_1+\lambda_2}\int_0^\infty (\lambda_1+\lambda_2)e^{-(\lambda_1+\lambda_2)x}\, dx \\ &= \frac{\lambda_1}{\lambda_1+\lambda_2}. \ \square \end{aligned}$$

Theorem 1.4. *(Embedding theorem 1.) For all $n \in \mathbb{N}_+$ the collections $\{d_j(n) : 1 \leq j \leq$*

$n+2\}$ and $\{D_j(\tau_n - \tau_{j-2}) : 1 \leq j \leq n+2\}$ have the same distribution.

Proof: *First we note that both collections have the Markov property and initial values so it will be sufficient to show that the conditional transition probabilities are the same. Therefore, for $k \geq 1$ consider the collection $\{D_j(\tau_k - \tau_{j-2}) : 1 \leq j \leq k+2\}$ and note that when the $(k+1)^{th}$ event occurs $D_{k+3}(0)$ is initialized at X_{k+1} and the process in which the event occurred is incremented by X_{k+1}. (This is analogous to a new vertex joining a connecting X_{k+1} times to its chosen neighbor.) The probability that this event occurs within process $i \in \{1, 2, \ldots, k+2\}$ is the probability that the minimum of $\sum_{j=1}^{k+2} D_j(\tau_k - \tau_{j-2})$ exponential($\lambda = 1$) and $k+2$ exponential(β) random variables was among one of the $D_j(\tau_k - \tau_{j-2})$ exponential($\lambda = 1$) random variables or the single exponential(β) random variable corresponding to process j. By lemmas 1.1 and 1.2 this is equal to,*

$$\frac{D_i(\tau_k - \tau_{i-2}) + \beta}{\sum_{j=1}^{k+2}(D_j(\tau_k - \tau_{j-2}) + \beta)},$$

and these are the transition probabilities corresponding to $\{d_j(k) : 1 \leq j \leq k+2\}$. □

1.3 Pure birth Markov chains

In the general weight function model the construction of the tree is similar to that under the model with linear weights, with the exception that each new vertex connects exactly once to its neighbor, and the weight of the i^{th} vertex at time n is given by $f[d_n(i)]$, where $d_n(i)$ denotes the degree, and $f : \mathbb{N}_+ \mapsto \mathbb{R}_+$ is increasing. In this model, the proofs rely on an embedding in general pure birth Markov chains.

Specifically, initialize pure birth processes $Z_1(t)$ and $Z_2(t)$ with $Z_1(0), Z_2(0) := 1$, and exponential sojourn times with rate function $f(i)$, $i \in \mathbb{N}_+$. Wait for a birth to occur in either $Z_1(t)$ or $Z_2(t)$, let τ_1 denote the random time of this event, and simultaneously initialize $Z_3(t)$ with $Z_3(0) := 1$. Continue this process exactly as before and one gets the next theorem [2].

Theorem 1.5. *(Embedding theorem 2.) The collections* $\{d_n(i) : 1 \leq i \leq n+2\}$ *and* $\{Z_i(\tau_n - \tau_{i-2}) : 1 \leq i \leq n+2\}$ *have the same distribution.*

Proof: *As in the case of the previous embedding theorem, we only need to show that the conditional transition probabilities are equal. Now, when the* $(k+1)^{th}$ *process is started, one of the existing processes* $\{Z_i(\tau_k - \tau_{i-2} : 1 \leq i \leq k+2)\}$ *will be incremented by one, and the probability that this process is the* i^{th} *for* $1 \leq i \leq k+2$ *is equal to the probability that the minimum of* $k+2$ *exponential random variables with rates* $\{f[Z_j(\tau_k - \tau_{j-2})]\}_{j=1}^{k+2}$ *corresponds to the one with rate* $f[Z_i(\tau_k - \tau_{i-2})]$ *and by Lemma 1.2 this is equal to,*

$$\frac{f[Z_i(\tau_k - \tau_{i-2})]}{\sum_{j=1}^{k+2} f[Z_i(\tau_k - \tau_{j-2})]}.$$

As before, these are the transition probabilities associated with the collection $\{d_n(i) : 1 \leq i \leq n+2\}$*.* □

Using this embedding, Athreya [2] established the following results.

Theorem 1.6. *(Superlinear case.) Let,*

$$\sum_{n=1}^{\infty} \frac{1}{f(n)} < \infty.$$

(a) If in addition,

$$\sum_{n=1}^{\infty} \frac{n}{n+f(n)} = \infty,$$

then for all $i \geq 1$, $\lim_{n\to\infty} d_n(i) \equiv \xi_i < \infty$ *exists almost surely.*

(b) If on the other hand,

$$\sum_{n=1}^{\infty} \frac{n}{n+f(n)} < \infty,$$

then $\forall\, i \geq 1$, the i^{th} vertex will with positive probability be the recipient of all but finitely-many edges from newly joining vertices as $n \to \infty$. (c) Let $\pi_j(n)$ denote the proportion of vertices with degree j at time n. Then whenever $\sum_{n=1}^{\infty} f(n)^{-1} < \infty$ holds, $\pi_n(1) \xrightarrow{\text{a.s.}} 1$, and $\pi_j(n) \xrightarrow{\text{a.s.}} 0$ for $j > 1$ as $n \to \infty$. □

The last part of this theorem says that the asymptotic degree distribution is degenerate at one.

Theorem 1.7. *(Sublinear case.) Suppose that,*

$$\lim_{n\to\infty} \frac{f(n)}{cn^p} \to 1,$$

for some $c > 0$ and $1/2 < p < 1$. Then there is a deterministic sequence $\{c(n)\}_{n=1}^{\infty}$ and a constant $0 < \alpha < \infty$ such that for every $i \geq 1$, $d_n(i)/c(n)^q \to \alpha$ w.p. 1, where $q \equiv 1/(1-p)$. In addition,

$$0 < \liminf_{n\to\infty} \frac{c(n)}{\log n} \leq \limsup_{n\to\infty} \frac{c(n)}{\log n} < \infty. \ \square$$

Theorem 1.8 characterizes the so-called non-explosion criterion for the process $Z(t)$, which helps to illuminate the relevance of the conditions in the above theorems (we assume for convenience that $Z(0) = 1$, but an identical argument works if the process begins in any of the other states). Before proving this theorem, we first prove a useful lemma which will also be used in subsequent sections.

Lemma 1.3. *Let $\{x_n\}_{n=1}^{\infty}$ be a sequence of numbers such that $x_n \in (0,1)$ for each n, $x_n \to 0$ as $n \to \infty$, and let $\{y_n\}_{n=1}^{\infty}$ be a sequence of positive numbers. Then $\prod_{n=1}^{\infty}(1-x_n)^{y_n} > 0$ if and only if $\sum_{n=1}^{\infty} y_n x_n < \infty$.*

Proof: *Note that $\prod_{n=1}^{\infty}(1-x_n)^{y_n} = e^{\sum_{n=1}^{\infty} y_n \log(1-x_n)}$ and hence $\prod_{n=1}^{\infty}(1-x_n)^{y_n} > 0$ if and only if $-\sum_{n=1}^{\infty} y_n \log(1-x_n) < \infty$. However, since $x_n \to 0$, $-\log(1-x_n)/x_n \to 1$ as $n \to \infty$ and thus $-\sum_{n=1}^{\infty} y_n \log(1-x_n) < \infty$ if and only if $\sum_{n=1}^{\infty} y_n x_n < \infty$ as claimed.* □

Theorem 1.8. *Let $\{L_i\}_{i=1}^{\infty}$ be independent random variables representing the sojourn times for states $i \in \{1, 2, \ldots\}$, with L_i distributed exponential with rate $f(i)$. Also denote $T_n \equiv \sum_{i=1}^{n} L_i$, the total time spent in the first n states, and $T_\infty \equiv \lim_{n\to\infty} T_n$, the explosion time of the process. Then if $\sum_{n=1}^{\infty} f(n)^{-1} = \infty$, $P(T_\infty = \infty) = 1$, and if $\sum_{n=0}^{\infty} f(n)^{-1} < \infty$, $P(T_\infty = \infty) = 0$.*

Proof: *For $\gamma > 0$,*

$$\begin{aligned} \mathrm{E}(e^{-\gamma T_\infty}) &= \lim_{n\to\infty} \mathrm{E}(e^{-\gamma T_n}) \\ &= \lim_{n\to\infty} \prod_{i=1}^{n} \mathrm{E}(e^{-\gamma L_i}) \\ &= \prod_{i=1}^{\infty} \frac{f(i)}{f(i)+\gamma} \\ &= \prod_{i=1}^{\infty} \left(1 - \frac{\gamma}{f(i)+\gamma}\right). \end{aligned}$$

Now, if $\sum_{n=1}^{\infty} f(n)^{-1} = \infty$, then by the above lemma this product is zero for all $\gamma > 0$, implying $T_\infty \overset{\text{a.s.}}{=} \infty$. If $\sum_{n=1}^{\infty} f(n)^{-1} < \infty$, then by an application of the dominated convergence theorem we can conclude that the limit of the right hand side is one as $\gamma \to 0$, and since $\lim_{\gamma\downarrow 0} \mathrm{E}(e^{-\gamma T_\infty}) = \lim_{\gamma\downarrow 0} \mathrm{E}(e^{-\gamma T_\infty}; T_\infty < \infty) = P(T_\infty < \infty)$ we have that $P(T_\infty < \infty) = 1$. □

1.4 Outline

The remainder of the dissertation is structured as follows. In Chapter 2 we will develop our framework for conducting inference on preferential attachment graphs, which will depend on the embedding of Athreya, Sethuraman and Ghosh described above. Then in Chapter 3 we will describe the extension of our model to random effects and prove our main results. Finally, in Chapter 4 we give results concerning cascades on preferential attachment trees.

Chapter 2: Inference for preferential attachment models

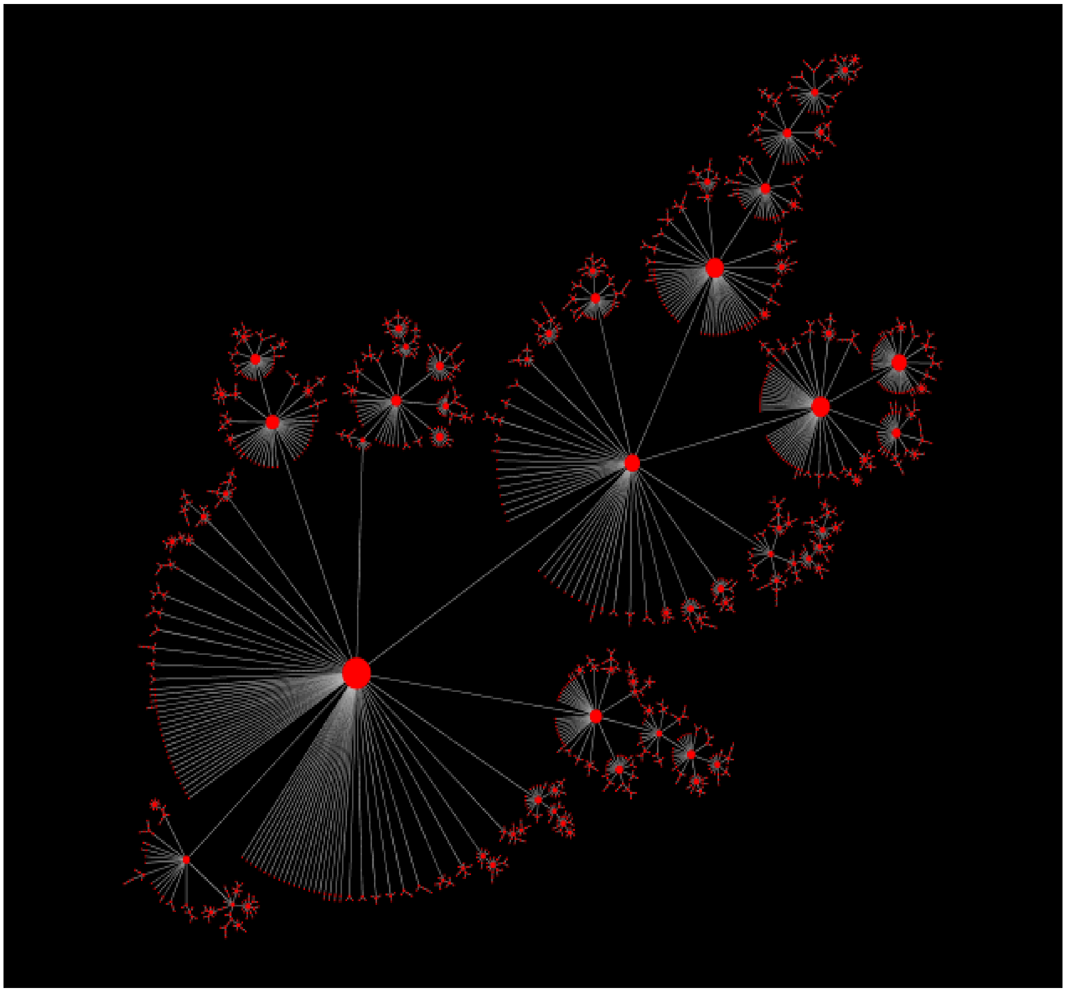

Figure 2.1: Preferential attachment tree with $\theta = 1/2$.

2.1 Introduction

Determining the precise nature of the attachment mechanism in preferential attachment models is an important question in network inference. We will explore the question of how

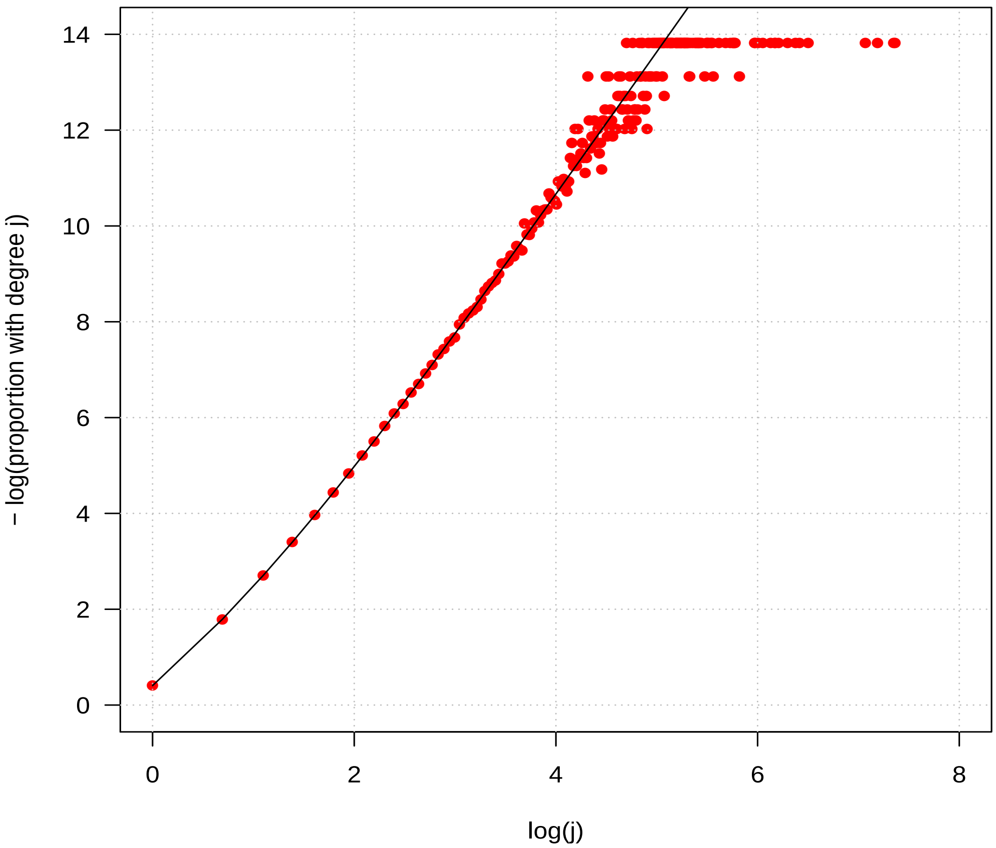

Figure 2.2: Log-log plot of the empirical degree distribution of a preferential attachment tree. $\theta = 1/2$, $n = 1,000,000$. The solid line is the true asymptotic distribution.

to measure the strength of this mechanism through estimation of the parameter θ. We begin by proposing two strongly consistent estimators and looking at their performance through simulation. Then we will prove the asymptotic normality for a special case of our primary estimator, and use this result to construct confidence intervals and perform hypothesis tests.

Figures 2.1 and 2.3 show how the features of the graph change for different values of $\theta \equiv m/(2m + \beta)$. In particular, values close to $1/2$ (which correspond to more "pure" preferential attachment) lead to heavier tail degree distributions, and values closer to zero correspond to lighter tails and fewer hubs. Figures 2.2 and 2.4 show log-log plots of the empirical degree distribution of preferential attachment trees with 1,000,000 vertices for

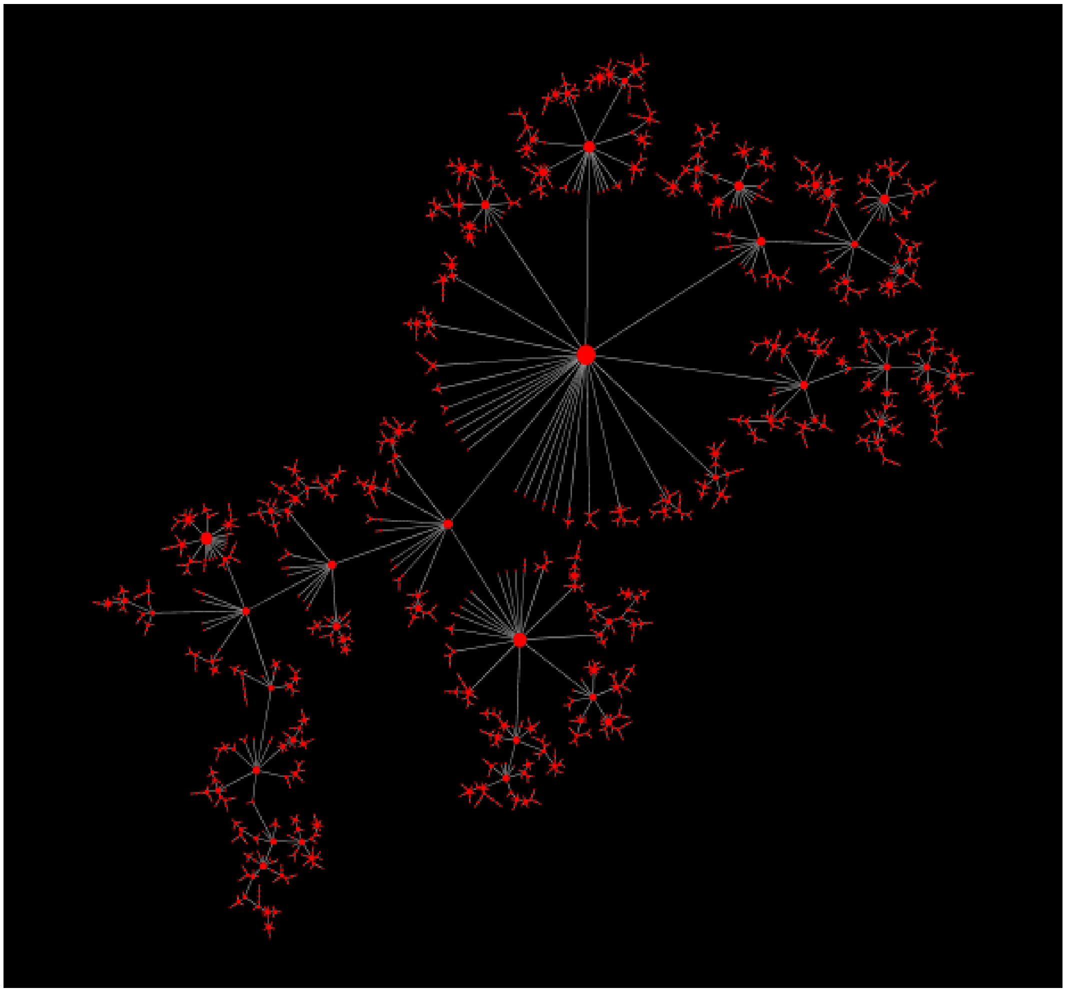

Figure 2.3: Preferential attachment tree with $\theta = 1/5$.

$\theta = 1/3$ and $\theta = 1/5$. These are helpful in visualizing power laws since if a random variable D follows a power law with scaling exponent η, then $-\log[P(D = j)] = -\log(c) + \eta \log(j)$, for some constant c, and so the scaling exponent is simply the slope of the plot. We see from these plots that not only does the graph with $\theta = 1/2$ have vertices with significantly higher degree, but both clearly have power law tails with a larger scaling exponent in the case $\theta = 1/5$.

Since we are making inferences about the evolution of the process, the statistics that we propose will require information about the state of the tree at more than one point in time. The first requires only knowledge of the state of the tree at two time points, whereas the

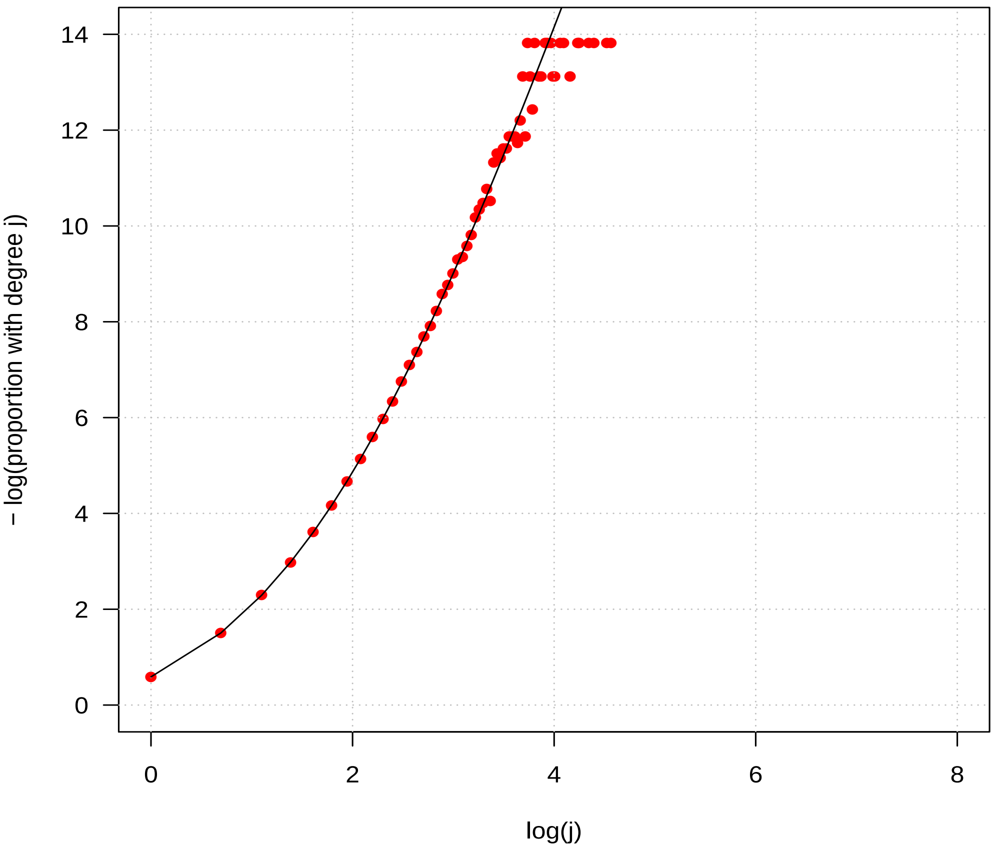

Figure 2.4: Log-log plot of the empirical degree distribution of a preferential attachment tree. $\theta = 1/5$, $n = 1,000,000$. The solid line is the true asymptotic distribution.

second takes an average over the state of the tree at many time points in the past. Before introducing and proving the consistency of our estimators we state an analysis result found in the paper of Athreya, et al. [3] which we will extend and use in our first proof.

Lemma 2.1. *Let* $\{a_{n,i} : 1 \leq i \leq n, n \geq 1\}$ *be a double array of nonnegative numbers such that (i) for all* $i \geq 1$, $\lim_{n\to\infty} a_{n,i} = a_i$*, and (ii) denoting* $b_i \equiv \sup_{n\geq 1} a_{n,i}$*,* $b_i \to 0$ *as* $i \to \infty$*. Then* $\max_{1\leq i\leq n} a_{n,i} \to \max_{i\geq 1} a_i$ *as* $n \to \infty$*, and if in addition the* a_i *are all distinct, then* $\max_{1\leq i\leq n} a_i = \max_{i\geq 1} a_i$ *for almost all* n*.* □

This lemma is useful since a Borel-Cantelli argument shows that under a mild moment condition the double array of appropriately-scaled degree sequences will satisfy the above conditions almost surely. Since our estimator will be based on degree maxima, this fact will help us to guarantee consistency.

2.2 Log max estimator

Our first estimator involves looking at the logarithm of the empirical growth rates of vertices with maximal degree viewed across a particular time interval. It is intuitive to look at logarithms of empirical growth rates since we know from previous results that the degree sequences grow at an asymptotic rate that is exponential in θ. The reason for looking at maxima is not so obvious, and is based primarily on having obtained good simulation results compared to other estimators. One plausible explanation for this phenomenon is that θ represents asymptotic growth rates of degree sequences, and it may be that the vertex with maximal degree has the greatest chance of its empirical growth rate being well-approximated by its asymptotic rate. (In what follows we use the notation that τ_k is the time when the k^{th} process is initialized with $\tau_1, \tau_2 := 0$.)

Theorem 2.1. *Let $r \in \mathbb{N}_+$, and let $I_{n,r}$ denote the r-vector whose elements $\{I_{n,r}^{(i)}\}_{1 \le i \le r}$ are the indices corresponding to the r most connected vertices of the graph at time n. Then so long as $\mathrm{E}(X^p) < \infty$ for some $p > 1/\theta$ we have $\forall\, k \in \{2, 3, 4, \ldots\}$ that,*

$$\frac{1}{r}\sum_{i=1}^{r} \min\left(\log_k \left[\frac{d_n(I_{\lfloor n/k \rfloor, r}^{(i)})}{d_{\lfloor n/k \rfloor}(I_{\lfloor n/k \rfloor, r}^{(i)})}\right], \frac{1}{2}\right) \xrightarrow{\text{a.s.}} \theta.$$

Proof: *It will be argued that for a triangular array of real numbers $\{a_{n,i} : 1 \le i \le n\}_{n \ge 1}$ satisfying all the hypotheses of the above proposition, the vector comprised of the indices corresponding to the r maximal entries of the n^{th} row of this array converges to some vector* *in* $\mathbb{N}_+^r$*. This will be used to establish that,*

$$\frac{1}{r}\sum_{i=1}^{r}\log_k\left[\frac{d_n(I^{(i)}_{\lfloor n/k\rfloor,r})}{d_{\lfloor n/k\rfloor}(I^{(i)}_{\lfloor n/k\rfloor,r})}\right]\xrightarrow{\text{a.s.}}\theta,$$

and from this the claim will follow since $\theta \in (0, 1/2]$*. To establish the first claim, note that* $\max_{1\le i\le n} a_{n,i} \to \max_{i\ge1} a_i$*, denote the index of this maximal column* i^**, and note also that the index of* $\max_{1\le i\le n} a_{n,i}$ *equals* i^* *for almost all* n*. Now consider the reduced array* $\{a_{n,i} : 1 \le i \le n,\, i \ne i^*\}_{n\ge1}$*. (This is simply the original array with column* i^* *deleted.) Since we have for large* n *that* $a_{n,i^*} > a_{n,i}, \forall\, i \ne i^*$*, finding the second largest entry in this row is the same as finding the largest entry of* $\{a_{n,i}\}_{1\le i\le n, i\ne i^*}$ *for all such* n*. But the array* $\{a_{n,i} : 1 \le i \le n,\, i \ne i^*\}_{n\ge1}$ *satisfies the same conditions on* $\{a_{n,i} : 1 \le i \le n\}_{n\ge1}$ *which ensure the convergence of the index of the maximal element, and so this sequence converges as well to* $\max_{i\ge1, i\ne i^*} a_i$*. The convergence of the remaining* $r-2$ *terms then follows in the same fashion.*

Now if $\mathrm{E}(X^p) < \infty$ *for some* $p > 1/\theta$*, the array of random variables* $\{D_i(\tau_n - \tau_i)/n^\theta : 1 \le i \le n\}_{n\ge1}$ *satisfies the hypotheses of the proposition a.s., and hence the* r*-vector of maximal indices freezes for large* n*. Further, for any fixed* j*,*

$$\frac{D_j(\tau_n-\tau_j)}{n^\theta}\xrightarrow{\text{a.s.}}\gamma_j\in(0,\infty),$$

and thus $\forall\, i \in \mathbb{N}_+$*,*

$$\log_k\left[\frac{D_i(\tau_n-\tau_i)}{D_i(\tau_{\lfloor n/k\rfloor}-\tau_i)}\right]=\log_k\left[\frac{D_i(\tau_n-\tau_i)}{n^\theta}\cdot\frac{(n/k)^\theta}{D_i(\tau_{\lfloor n/k\rfloor}-\tau_i)}\right]+\theta$$

$$\xrightarrow{\text{a.s.}}\theta,$$

and by the embedding theorem the same holds for the scaled degree sequences. That the truncated and untruncated estimators are asymptotically equivalent is evident since $\theta \in$

$(0, 1/2]$. □

The reason for truncating is simply to force estimates to lie within the parameter space $(0, 1/2]$.

2.3 Mean log max estimator

Our next estimator involves looking at the tree at several snapshots in the past, and can be viewed as a kind of Cesàro average of the log max estimator with $r = 1$.

Theorem 2.2. *Let I_n be the index of the maximal vertex at time n and suppose $\mathrm{E}(X^p) < \infty$ for some $p > 1/\theta$. Then for any $k \in \{2, 3, \ldots\}$,*

$$\frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \min\left(\log_k \left[\frac{d_{kj}(I_j)}{d_j(I_j)} \right], \frac{1}{2} \right) \xrightarrow{\text{a.s.}} \theta.$$

Proof: *Since $\mathrm{E}(X^p) < \infty$ for some $p > 1/\theta$ we have that $\log_k[d_{kj}(I_j)/d_j(I_j)] \xrightarrow{\text{a.s.}} \theta$ as $j \to \infty$ for all $k \in \{2, 3, \ldots\}$. Thus for any $\epsilon > 0$ there a.s. exists a random $j^* \in \mathbb{N}_+$ such that $|\log_k[d_{kj}(I_j)/d_j(I_j)] - \theta| \le \epsilon$ for all $j > j^*$. Therefore with probability 1,*

$$\left| \frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \log_k \left[\frac{d_{kj}(I_j)}{d_j(I_j)} \right] - \theta \right| \le \frac{1}{\lfloor n/k \rfloor} \left| \sum_{j=1}^{j^*} \left(\log_k \left[\frac{d_{kj}(I_j)}{d_j(I_j)} \right] - \theta \right) \right| + \epsilon$$

$$\to \epsilon,$$

and hence,

$$\frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \log_k \left[\frac{d_{kj}(I_j)}{d_j(I_j)} \right] \xrightarrow{\text{a.s.}} \theta.$$

Now, since again $\log_k[d_{kj}(I_j)/d_j(I_j)] \xrightarrow{\text{a.s.}} \theta \in (0, 1/2]$,

$$\left| \frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \log_k \left[\frac{d_{kj}(I_j)}{d_j(I_j)} \right] - \frac{1}{\lfloor n/k \rfloor} \sum_{j=1}^{\lfloor n/k \rfloor} \min \left(\log_k \left[\frac{d_{kj}(I_j)}{d_j(I_j)} \right], \frac{1}{2} \right) \right| \xrightarrow{\text{a.s.}} 0. \ \square$$

Of course, one could combine the ideas behind both of these estimators and compute averages across time for several maxima, and this would likely improve efficiency. It is also natural to ask why we aren't averaging over all vertices. The reason is that this estimator actually performs poorly in practice, which is likely due to the fact that it involves taking an average over new vertices whose empirical growth rates are not yet well-approximated by their asymptotic growth rates, which is also the reason why the above proof breaks down in this setting. (Specifically, consider a double array $\{c_{ij} : i \geq 1, j \geq i\}$ where for each i, $c_{ij} \to c$ as $j \to \infty$. Then although one can take an average of the c_{ij}'s over finitely-many i's, let $j \to \infty$ and the quantity will still converge to c, this need not be the case for the overall average $n^{-1} \sum_{i=1}^{n} c_{in}$, and this is why the argument fails.)

2.4 Distributional results

In addition to guaranteeing consistency, it is important that we also characterize the speed of convergence and be able to assess the quality of our estimator. The following theorems address this issue, establishing that the rate of convergence is of order $n^{-\theta/2}$, while also providing us with a means through which to construct confidence intervals and perform hypothesis tests.

It is useful to first cast the problem in terms of estimation of k^θ, and then apply the delta method to obtain central limit theorems for estimators of θ.

Theorem 2.3. *Consider the preferential attachment graph where each new vertex connects once to its chosen neighbor, and let* $d_n(i)$ *denote the degree of the* i^{th} *vertex at time* n*. Then* *for any $i, k \in \mathbb{N}_+$, $k > 1$,*

$$\sqrt{d_{\lfloor n/k \rfloor}(i)} \left(\frac{d_n(i)}{d_{\lfloor n/k \rfloor}(i)} - k^{\theta} \right) \xrightarrow{\mathrm{d}} \textit{normal} \left(0, k^{\theta}(k^{\theta} - 1) \right).$$

Proof: *As before, we will prove the corresponding result for $\{D_i(\tau_n - \tau_i) : n \geq 1, i \leq n\}$ and then the theorem will follow by an appeal to the embedding theorem. (For convenience we take n/k to equal $\lfloor n/k \rfloor$ whenever n/k is not an integer.)*

First, note by the branching property that $D_i(\tau_n - \tau_i) \overset{\mathrm{d}}{=} \sum_{j=1}^{D_i(\tau_{n/k} - \tau_i)} D_j^{(n)}(\tau_n - \tau_{n/k})$, where the $\left\{ D_j^{(n)}(\tau_n - \tau_{n/k}) \right\}_{j=1}^{D_i(\tau_{n/k} - \tau_i)}$ are i.i.d. Markov branching processes with $D_j^{(n)}(0) := 1$, unit splitting rate and immigration rate $\beta / D_i(\tau_{n/k} - \tau_i)$. Thus,

$$\sqrt{D_i(\tau_{n/k} - \tau_i)} \left(\frac{D_i(\tau_n - \tau_i)}{D_i(\tau_{n/k} - \tau_i)} - k^{\theta} \right) \overset{\mathrm{d}}{=} \frac{\sum_{j=1}^{D_i(\tau_{n/k} - \tau_i)} \left\{ D_j^{(n)}(\tau_n - \tau_{n/k}) - \mu_{n,k} \right\}}{\sqrt{D_i(\tau_{n/k} - \tau_i))}}$$

$$+ \sqrt{D_i(\tau_{n/k} - \tau_i)} \left(\mu_{n,k} - k^{\theta} \right),$$

where $\mu_{n,k} \equiv \mathrm{E}[D^{(n)}(\tau_n - \tau_{n/k})]$. Next, note that $\tau_n - \tau_{n/k} \xrightarrow{\mathrm{a.s.}} \alpha \log(k), D_i(\tau_{n/k} - \tau_i) \xrightarrow{\mathrm{a.s.}} \infty$ as $n \to \infty$ and so $D_j^{(n)}(\tau_n - \tau_{n/k}) \xrightarrow{\mathrm{d}} Z(\alpha \log(k))$, where $Z(t)$ is a Yule process (i.e., a Markov branching process with offspring distribution degenerate at 2) with unit lifetime parameter and $Z(0) := 1$. By the Lindeberg-Feller central limit theorem the first term converges to a normal$(0, \sigma_k^2)$ where $\sigma_k^2 \equiv \mathrm{Var}[Z(\alpha \log(k))] = k^{\theta}(k^{\theta} - 1)$, and so we only need to show that $\sqrt{D_i(\tau_{n/k} - \tau_i)} \left(\mu_{n,k} - k^{\theta} \right) \xrightarrow{\mathrm{p}} 0$, which holds if $n^{\theta/2} \left(\mu_{n,k} - k^{\theta} \right) \to 0$ since $D_i(\tau_{n/k} - \tau_i)$ is almost surely of order n^{θ}. Now write,

$$n^{\theta/2} \left(\mu_{n,k} - k^{\theta} \right) = n^{\theta/2} \left(\mu_{n,k} - \mathrm{E} \left(e^{\tau_n - \tau_{n/k}} \right) \right) + n^{\theta/2} \left(\mathrm{E} \left(e^{\tau_n - \tau_{n/k}} \right) - k^{\theta} \right),$$

and note that,

$$\begin{aligned}\mathrm{E}\left(e^{\tau_n-\tau_{n/k}}\right) &= \mathrm{E}\left(e^{\sum_{j=n/k}^{n-1}(\tau_{j+1}-\tau_j)}\right)\\ &= \prod_{j=n/k}^{n-1}\left(\frac{j}{j-\theta}\right)\\ &= \frac{\Gamma(n)\,\Gamma({}^{n}\!/\!{}_{k}-\theta)}{\Gamma(n-\theta)\,\Gamma({}^{n}\!/\!{}_{k})}.\end{aligned}$$

Since $\Gamma(t+a)/\Gamma(t) = t^a(1+O({}^{1}\!/\!{}_{t}))$ *as* $t\to\infty$ *[10] we see after a few manipulations that,*

$$\begin{aligned}n^{\theta/2}\left(\frac{\Gamma(n)\,\Gamma({}^{n}\!/\!{}_{k}-\theta)}{\Gamma(n-\theta)\,\Gamma({}^{n}\!/\!{}_{k})}-k^\theta\right) &= n^{\theta/2}k^\theta\left(\frac{1+O({}^{1}\!/\!{}_{n})}{1+O({}^{1}\!/\!{}_{n})}-1\right)\\ &= k^\theta\left(\frac{O({}^{1}\!/\!{}_{n^{1-\theta/2}})-O({}^{1}\!/\!{}_{n^{1-\theta/2}})}{1+O({}^{1}\!/\!{}_{n})}\right)\\ &\to 0.\end{aligned}$$

Also, letting $\{Z_j(t)\}_{j=0}^{\infty}$ *denote i.i.d. Yule processes initialized by a single particle and* $\{T_j^{(n)}\}_{j=1}^{\infty}$ *the jump times of a Poisson process with rate* $\beta_n \equiv \beta/D_i(\tau_{n/k}-\tau_i)$ *and* $T_0 := 0$ *we have,*

$$\begin{aligned}\mu_{n,k} &= \mathrm{E}\left[\sum_{j=0}^{\infty} Z_j(\tau_n-\tau_{n/k}-T_j^{(n)})I(T_j^{(n)}\le\tau_n-\tau_{n/k})\right]\\ &= \mathrm{E}\left[\sum_{j=0}^{\infty} e^{\tau_n-\tau_{n/k}}\left(\frac{\beta_n}{\beta_n+1}\right)^j P(Y_n\ge j\,|\,\tau_n-\tau_{n/k},\,\beta_n)\right],\end{aligned}$$

where $Y_n|\beta_n \sim Poisson(\lambda_n \equiv \beta(\tau_n - \tau_{n/k})/D_i(\tau_{n/k} - \tau_i))$. This yields,

$$\begin{aligned}
n^{\theta/2}\left|\mu_{n,k} - \mathrm{E}\left(e^{\tau_n - \tau_{n/k}}\right)\right| &= n^{\theta/2}\mathrm{E}\left[e^{\tau_n-\tau_{n/k}}\sum_{j=1}^{\infty}\left(\frac{\beta_n}{\beta_n+1}\right)^j P(Y_n \geq j \,|\, \tau_n - \tau_{n/k},\, \beta_n)\right] \\
&\leq n^{\theta/2}\mathrm{E}\left[e^{\tau_n-\tau_{n/k}}\sum_{j=1}^{\infty} P(Y_n \geq j \,|\, \tau_n - \tau_{n/k},\, \beta_n)\right] \\
&= n^{\theta/2}\mathrm{E}\left[e^{\tau_n-\tau_{n/k}}\beta\frac{\tau_n - \tau_{n/k}}{D_i(\tau_{n/k} - \tau_i)}\right] \\
&= \beta\, n^{-\theta/2}\mathrm{E}\left[e^{\tau_n-\tau_{n/k}}(\tau_n - \tau_{n/k})\frac{n^{\theta}}{D_i(\tau_{n/k} - \tau_i)}\right].
\end{aligned}$$

Now, using results from [5] and [11] it can be shown that $\mathrm{E}[\sup_{n\geq 1} e^{2(\tau_n - \tau_{n/k})}(\tau_n - \tau_{n/k})^2] < \infty$ and $\sup_{n\geq 1}\mathrm{E}[n^{2\theta}/D_i(\tau_{n/k} - \tau_i)^2] < \infty$, and so after an application of the Cauchy-Schwarz inequality it follows that the above bound converges to zero as $n \to \infty$. □

We can now establish the limiting distribution of the log max estimator for the special case of an edge distribution degenerate at one.

Theorem 2.4. *Let $d_n(i)$ denote the degree of the i^{th} vertex of a preferential attachment graph at time n. Then for any $i, k \in \mathbb{N}_+$, $k > 1$,*

$$\sqrt{d_{\lfloor n/k \rfloor}(i)}\left(\log_k\left[\frac{d_n(i)}{d_{\lfloor n/k\rfloor}(i)}\right] - \theta\right) \xrightarrow{\mathrm{d}} normal\left(0, \frac{1-k^{-\theta}}{\log^2(k)}\right).$$

Proof: *By expanding $\log_k(x)$ in a Taylor series about k^θ, rearranging terms and setting $x := d_n(i)/d_{\lfloor n/k\rfloor}(i)$ we have,*

$$\log_k\left[\frac{d_n(i)}{d_{\lfloor n/k\rfloor}(i)}\right] - \theta = \frac{d_n(i)/d_{\lfloor n/k\rfloor}(i) - k^\theta}{k^\theta \log(k)} + \delta_n,$$

where $\delta_n = o_{\text{a.s.}}(d_n(i)/d_{\lfloor n/k \rfloor}(i) - k^\theta)$ as $n \to \infty$. Hence,

$$\sqrt{d_{\lfloor n/k \rfloor}(i)}\left(\log_k\left[\frac{d_n(i)}{d_{\lfloor n/k \rfloor}(i)}\right] - \theta\right) = \sqrt{d_{\lfloor n/k \rfloor}(i)}\left(\frac{d_n(i)/d_{\lfloor n/k \rfloor}(i) - k^\theta}{k^\theta \log(k)}\right) + \delta_n\sqrt{d_{\lfloor n/k \rfloor}(i)}$$

$$\xrightarrow{\text{d}} \textit{normal}\left(0, \frac{1-k^{-\theta}}{\log^2(k)}\right),$$

by Theorem 2.3 above. □

Theorem 2.5. *Let I_n denote the index of the vertex with maximal degree at time $n \in \mathbb{N}_+$. Then,*

$$\sqrt{d_{\lfloor n/k \rfloor}(I_{\lfloor n/k \rfloor})}\left(\log_k\left[\frac{d_n(I_{\lfloor n/k \rfloor})}{d_{\lfloor n/k \rfloor}(I_{\lfloor n/k \rfloor})}\right] - \theta\right) \xrightarrow{\text{d}} \textit{normal}\left(0, \frac{1-k^{-\theta}}{\log^2(k)}\right).$$

Proof: *Since* $\text{E}(X^r) < \infty$ *for an $r > 1/\theta$ holds trivially under this model we have that $I_n \xrightarrow{\text{a.s.}} I \in \mathbb{N}_+$, and the result now follows from Theorem 2.4.* □

2.5 Interval estimation

Using the previous results one can also obtain interval estimates for θ. One small difficulty that arises is that the asymptotic variance depends on the parameter being estimated, but this can be addressed by simply plugging in an estimate of θ. Setting $\hat{\theta}_{n,k} \equiv \log_k[d_n(I_{\lfloor n/k \rfloor})/d_{\lfloor n/k \rfloor}(I_{\lfloor n/k \rfloor})]$ our confidence interval then takes the form,

$$\hat{\theta}_{n,k} \pm z_{\alpha/2}\frac{1}{\log(k)}\sqrt{\frac{1-k^{-\hat{\theta}_{n,k}}}{d_{\lfloor n/k \rfloor}(I_{\lfloor n/k \rfloor})}},$$

where $z_{\alpha/2}$ denotes the $\alpha/2$ quantile of the standard normal distribution. Tables 2.1 and 2.2 show the performance of this interval (with confidence level 0.95) using various graph sizes for $\theta = 1/3, 1/4$. (Monte Carlo sample sizes of 10,000 were used.)

Number of vertices	Coverage proportion	Average length
1,000,000	0.9470	0.1614
500,000	0.9432	0.0.1812
250,000	0.9464	0.2040
100,000	0.9414	0.2372
50,000	0.9458	0.2669
25,000	0.9453	0.3004
10,000	0.9365	0.3520
5,000	0.9345	0.3961

Table 2.1: Confidence interval performance. $\theta = 1/3$.

Number of vertices	Coverage proportion	Average length
1,000,000	0.9405	0.2057
500,000	0.9400	0.2252
250,000	0.9409	0.2464
100,000	0.9363	0.2775
50,000	0.9324	0.3037
25,000	0.9299	0.3333
10,000	0.9269	0.3772
5,000	0.9212	0.4148

Table 2.2: Confidence interval performance. $\theta = 1/4$.

We see that the interval is not quite covering with proper frequency, but nearly so. The expected length appears to be somewhat high, but this could be corrected by including more vertices in the calculation of $\hat{\theta}_{n,k}$. (Of course, this would require an extension of the existing proof.)

2.6 Simulations

Figures 2.5-2.7 show the results of a simulation study of our estimators with $k = 2$ for various edge distributions and values for θ. In these simulations our estimators are based on only a single maximum. Monte Carlo sample sizes of 10,000 were used, and the yellow, tan and red density estimates use graph sizes of 250, 10,000 and 1,000,000 respectively. The edge distributions were taken to be Poisson, geometric and discrete Pareto; that is, with probability mass function,

$$P(X = j) = \frac{1}{j^\tau} - \frac{1}{(j+1)^\tau},$$

$\tau \geq 1$, $j \in \mathbb{N}_+$.

We see that the convergence rate is directly related to the value of θ, with smaller values corresponding to slower rates of convergence in the Poisson and geometric cases (this was also verified for the case of an edge distribution degenerate at one), but the Pareto case seems to be an anomaly, with smaller values of θ corresponding to more rapid convergence. The fact that it converges at all is interesting in itself, since in our experiments the Pareto edge distribution has infinite variance, even though our proofs require at least finite variance to guarantee consistency. This suggests that the condition in the theorems could be relaxed somewhat.

2.6.1 Computational challenges

The simulation of very large preferential attachment trees involves some computational difficulties since it requires sampling from a multinomial distribution whose probabilities are being iteratively updated. The usual method of doing so is to generate a uniform$(0,1)$ variate u and taking $F_M^{-1}(u) \equiv \min\{c : F_M(c) \geq u\}$, where F_M is the distribution function of a multinomial random variable M assuming values in $\{1, 2, \ldots, r\}$ for an integer r. To apply this method here we must iteratively update the cumulative distribution function corresponding to the preferential attachment tree, or equivalently a cumulative weight function $u \mapsto \min\{c : \sum_{i=1}^{c}(d_n(i) + \beta) \geq u\}$, which becomes very computationally expensive if we wish to simulate a large graph.

Instead, we work around this by still using a probability transform as above, only based on a function H_n which is initialized at,

$$H_2(s) = \begin{cases} 1 & \text{if} \quad 0 < s < 1+\beta \\ 2 & \text{if} \quad 1+\beta \leq s < 2+2\beta \end{cases},$$

and is undefined elsewhere. To construct H_3 one generates a uniform$(0, 2+2\beta)$ variate u, a positive integer-valued variate x_3 and sets $H_3(s) := H_2(s)$ for $s < 2+2\beta$, $H_3(s) = 3$ for $2+2\beta \leq s < 2+x_3+3\beta$ and $H_3(s) = H_2(u)$ for $2+x_3+3\beta \leq s < 2+2x_3+3\beta$. This gives us the updated function,

$$H_3(s) = \begin{cases} 1 & \text{if} \quad 0 < s < 1+\beta \\ 2 & \text{if} \quad 1+\beta \leq s < 2+2\beta \\ 3 & \text{if} \quad 2+2\beta \leq s < 2+x_3+3\beta \\ H_2(u) & \text{if} \quad 2+x_3+3\beta \leq s < 2+2x_3+3\beta \end{cases},$$

and is similarly undefined for $s \notin (0, 2+2x_3+3\beta)$. This is a function that "jumps down" to the index of the neighbor chosen by the newly-joining vertex rather than extending the length of that interval and shifting the positions of the intervals that follow, as would be necessary in order to calculate the inverse distribution function. If we continue this process, the total lengths of the intervals over which $H_n(s) = i$ for $i \in \{1, 2, \ldots, n\}$ will be the total weight assigned to vertex i, and we can then apply the probability transform to the function H_n, thereby avoiding a significant amount of unnecessary calculation involved in the iterative updates of the cumulative distribution function.

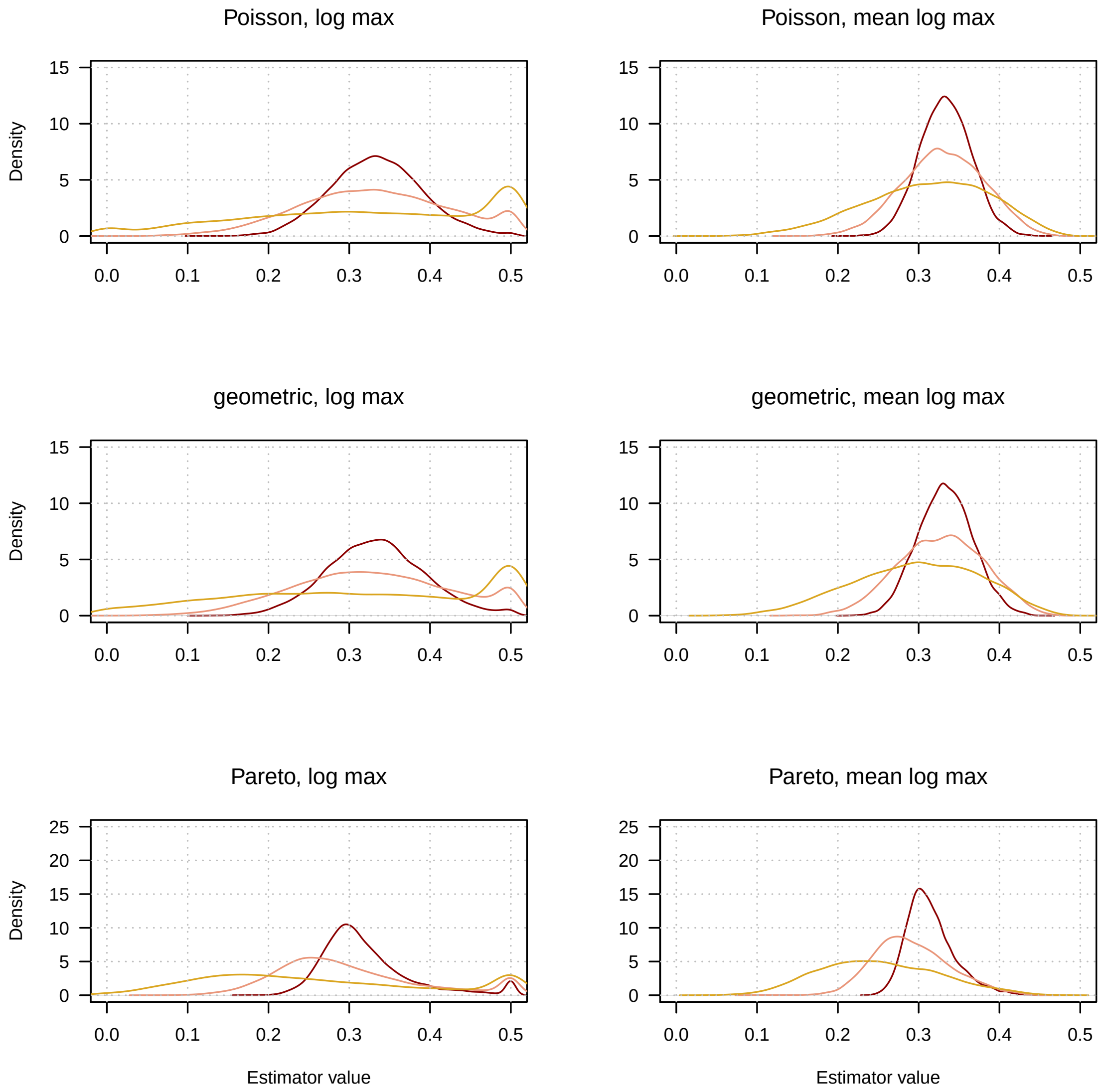

Figure 2.5: Performance of estimators for various edge distributions. $\theta = {}^1\!/_3$. (Note the change of scale in the Pareto density plots.)

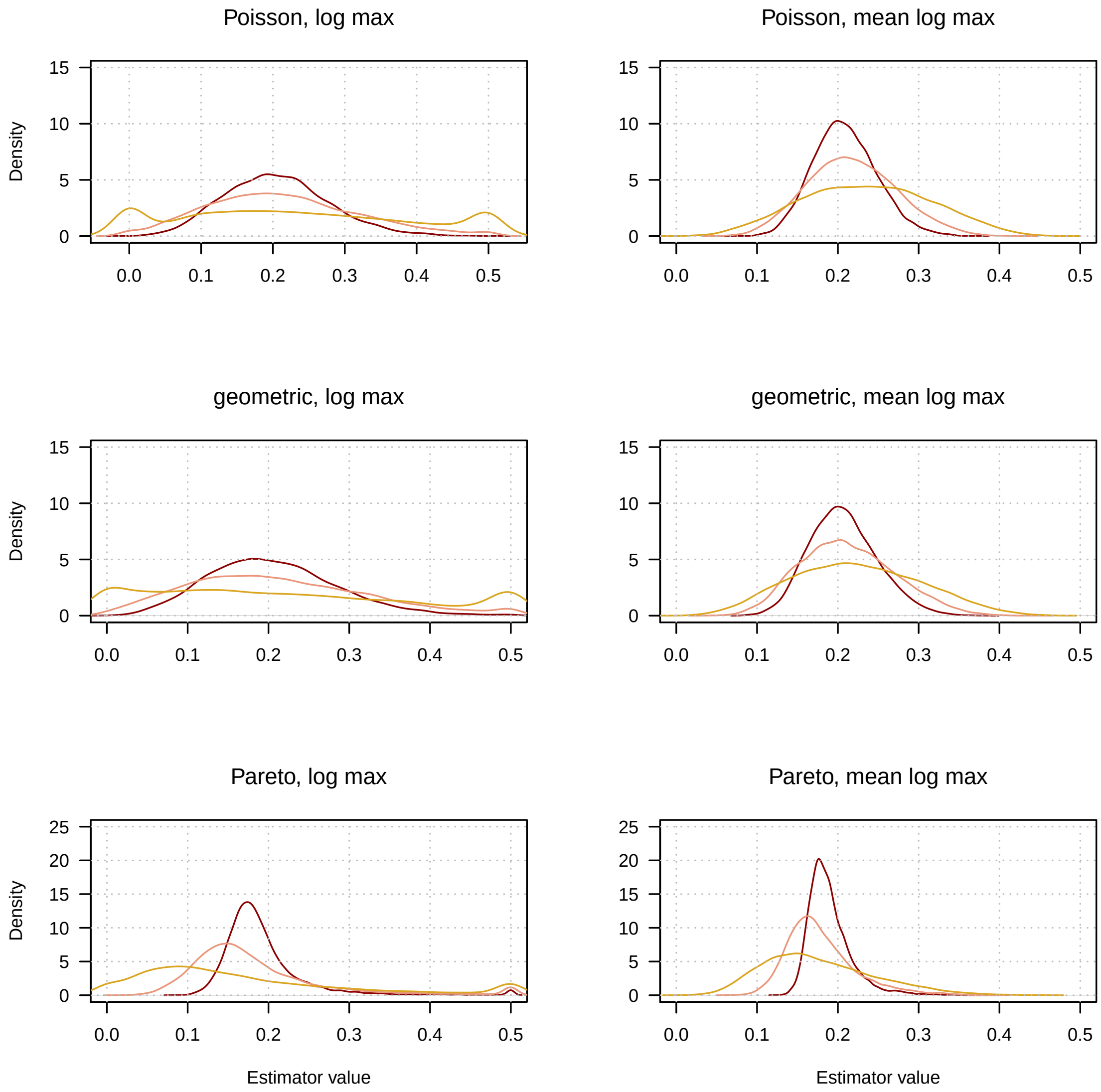

Figure 2.6: Performance of estimators for various edge distributions. $\theta = 1/5$. (Note the change of scale in the Pareto density plots.)

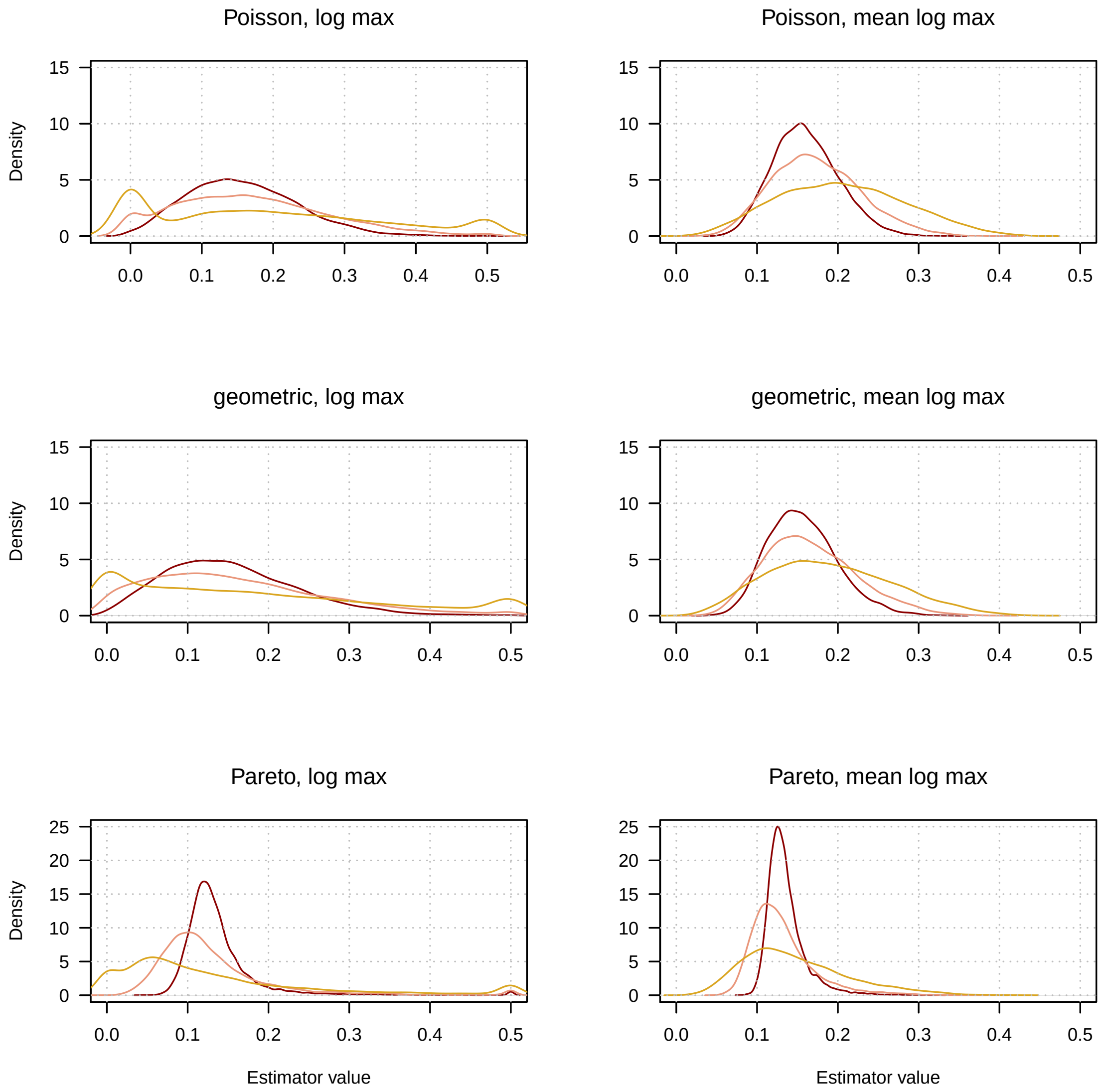

Figure 2.7: Performance of estimators for various edge distributions. $\theta = 1/7$. (Note the change of scale in the Pareto density plots.)

Chapter 3: Preferential attachment with random effects

3.1 Introduction

The standard model of preferential attachment takes the connection probability to be proportional to a linear function of the degree, but this function does not take into account any other features that vertices might possess. In reality this is a fairly unreasonable assumption, since generally nodes in real world networks with the same degree are not indistinguishable in terms of attractiveness. For instance, in a social network an individual's personality might render that individual more appealing independent of their popularity. In this chapter we will consider one way to introduce this heterogeneity into the process, which is to consider the model of Athreya, Sethuraman and Ghosh and allow the additive component β to be random and differ across vertices. We investigate how properties of the graph generalize to this setting.

3.2 Main results

The embedding works as in Athreya, Sethuraman and Ghosh [3] only now we associate with each process a random immigration parameter which is initialized at the start of the process. That is to say, at time $n = 2$ (starting at this time is for notational convenience) begin with two processes $D_1(0), D_2(0) := 1$ with respective immigration parameters β_1 and β_2 (the collection $\{\beta_i\}_{i=1}^{\infty}$ are taken to be i.i.d. nonnegative random variables) and unit lifetime parameters. Set $\tau_1, \tau_2 := 0$ and let τ_3 denote the random time at which the first event (either immigration or the splitting of a particle) occurs in either of the two processes. Once this occurs we initialize a third process with $D_3(0) := X_3$, where X_3 is the net addition of particles in the process where the event occurred, this time with unit lifetime parameter and immigration parameter β_3, and so on. We then have a generalized form of the embedding theorem which we can use to extend the results of Athreya, Sethuraman and Ghosh.

Theorem 3.1. *The collections $\{D_i(\tau_n - \tau_i) : i \leq n, n \geq 1\}$ and $\{d_n(i) : i \leq n, n \geq 1\}$ have the same distribution.*

Proof: *Since by construction both collections are Markov with the same state space, it will be sufficient to show that the transition probabilities are the same. Within the preferential attachment random graph, given G_k and $(\beta_1, \beta_2, \ldots, \beta_k)$ exactly one vertex will have its degree incremented by X_{k+1} and all others will remain constant at time $k+1$. The probability that this vertex is the i^{th} for $1 \leq i \leq k$ is $(d_k(i) + \beta_i)/\sum_{j=1}^{k}(d_k(j) + \beta_j)$.*

Similarly, if we consider the collection $\{D_i(\tau_n - \tau_i) : i \leq n, n \geq 1\}$ and condition on the states of all processes at time k as well as $(\beta_1, \beta_2, \ldots, \beta_k)$, then at time $k+1$ again only one process will have its value incremented by X_{k+1}, and the probability that this process is the i^{th} one is the probability that the minimum of $\sum_{j=1}^{k} D_i(\tau_k - \tau_j)$ unit exponential and k exponential random variables with rates $\{\beta_j\}_{j=1}^{k}$ is among the $D_k(\tau_n - \tau_i)$ unit exponential and exponential(β_i) random variables associated with process i, and this probability is $(D_i(\tau_n - \tau_i) + \beta_i)/\sum_{j=1}^{k}(D_j(\tau_n - \tau_j) + \beta_j)$. □

To investigate this more general model we start by studying the asymptotics of the random sequence $\{\tau_n\}_{n=1}^{\infty}$.

Theorem 3.2. *Let $S_n \equiv 2+\beta_1+\beta_2+\sum_{j=3}^{n}(2X_j+\beta_j)$ and $\alpha \equiv 1/(2m+\mu)$, where $\mu \equiv \mathrm{E}(\beta)$. Then $\sum_{j=1}^{\infty}(1/S_j - \alpha/j)$ converges almost surely if and only if $\mathrm{E}[(X+\beta)\log(X+\beta)] < \infty$.*

Proof: *Set $S_n^* \equiv S_n - \beta_1 - \beta_2$ and write,*

$$\sum_{j=1}^{n}\left(\frac{1}{S_j} - \frac{\alpha}{j}\right) = \sum_{j=1}^{n}\left(\frac{1}{S_j} - \frac{1}{S_j^*}\right) + \sum_{j=1}^{n}\left(\frac{1}{S_j^*} - \frac{\alpha}{j}\right).$$

For the first sum note that the inequality $S_j^* \geq j$ holds for all j and since $x \mapsto 1/x - 1/(x+c)$

is decreasing for $c > 0$ we have,

$$\sup_n \left| \sum_{j=1}^{n} \left(\frac{1}{S_j} - \frac{1}{S_j^*} \right) \right| = \sum_{j=1}^{\infty} \left(\frac{1}{S_j^*} - \frac{1}{S_j^* + \beta_1 + \beta_2} \right)$$

$$\leq \sum_{j=1}^{\infty} \left(\frac{1}{j} - \frac{1}{j + \beta_1 + \beta_2} \right)$$

$$< \infty,$$

for any $\beta_1, \beta_2 \in [0, \infty)$. Therefore, being bounded and monotone, the sum converges as $n \to \infty$. As for the second, Theorem III.9.4 on reciprocal sums from "Branching Processes" by Athreya and Ney states that for sums of the form $T_n \equiv t + U_1 + U_2 + \ldots + U_n$ where t is a non-negative constant and the $\{U_i\}_{i=1}^{\infty}$ are i.i.d. non-negative random variables with $\eta \equiv \mathrm{E}(U)$, $\lim_{n\to\infty} \sum_{i=1}^{n} (1/T_i - 1/\eta i)$ exists and is finite almost surely if and only if $\mathrm{E}(U \log |U|) < \infty$, provided that $t > 0$ or $\mathrm{E}(1/U) < \infty$. In our case, $U_i = 2X_i + \beta_i$, $c = 2$, $\eta = 2m + \mu$, and hence $\sum_{j=1}^{\infty} (1/S_j^* - \alpha/j)$ converges almost surely if and only if $\mathrm{E}[(2X + \beta)\log(2X + \beta)] < \infty \Leftrightarrow \mathrm{E}[(X + \beta)\log(X + \beta)] < \infty$. □

We also consider the sequence $\{\tau_n - \sum_{j=1}^{n} 1/S_{j-1}\}_{n=3}^{\infty}$ and letting $\mathcal{F}_n \equiv \sigma(\{X_i\}_{i=3}^{n}, \{\beta_i\}_{i=1}^{n})$ note that for $n \geq 2$ and given $\mathcal{F}_n$ the random variable $\tau_{n+1} - \tau_n$ is exponentially distributed with rate S_n. In addition, we have the following theorem.

Theorem 3.3. *The family $\{\tau_n - \sum_{j=1}^{n} 1/S_{j-1}; \mathcal{F}_{n-1}\}_{n=3}^{\infty}$ is a martingale uniformly bounded in L^2 and hence converges almost surely and in L^2.*

Proof: *First, for $n \geq 2$ we have,*

$$\begin{aligned}
\mathrm{E}\left[\tau_{n+1} - \sum_{j=1}^{n+1} \frac{1}{S_{j-1}} \;\Big|\; \mathcal{F}_n\right] &= \mathrm{E}\left[\sum_{j=1}^{n+1}\left(\tau_j - \tau_{j-1} - \frac{1}{S_{j-1}}\right) \;\Big|\; \mathcal{F}_n\right] \\
&= \tau_n - \sum_{j=1}^{n} \frac{1}{S_{j-1}} + \mathrm{E}\left[\tau_{n+1} - \tau_n - \frac{1}{S_n} \;\Big|\; \mathcal{F}_n\right] \\
&= \tau_n - \sum_{j=1}^{n} \frac{1}{S_{j-1}},
\end{aligned}$$

almost surely. Also,

$$\begin{aligned}
\sup_n \mathrm{E}\left[\left(\tau_n - \sum_{j=1}^{n} \frac{1}{S_{j-1}}\right)^2\right] &= \sup_n \mathrm{E}\left\{\left[\sum_{j=1}^{n}\left(\tau_j - \tau_{j-1} - \frac{1}{S_{j-1}}\right)\right]^2\right\} \\
&= \sup_n \sum_{j=1}^{n} \mathrm{E}\left[\left(\tau_j - \tau_{j-1} - \frac{1}{S_{j-1}}\right)^2\right] \\
&= \sum_{j=0}^{\infty} \mathrm{E}\left(\frac{1}{S_j^2}\right) \\
&\leq \frac{1}{4} + \sum_{j=1}^{\infty} \frac{1}{j^2} \quad (S_0 \geq 2 \text{ and } S_j \geq j \text{ for } j \geq 1.) \\
&< \infty,
\end{aligned}$$

and hence the martingale is uniformly bounded in L^2. The almost sure and L^2 convergence follow from Doob's martingale convergence theorems. □

The next result establishes the almost sure growth rates of the degree sequences within this model. Interestingly, the rate is similar to the original model only with β replaced by its expectation. This is due to a law of large numbers for reciprocal sums related to the sequence $\{\tau_n\}_{n=1}^{\infty}$ that occurs in the embedding.

Theorem 3.4. *Suppose that* $\mathrm{E}[(X+\beta)\log(X+\beta)] < \infty$. *Then for any* $i \in \mathbb{N}_+$, $\lim_{n\to\infty} d_n(i)n^{-\theta}$ *exists, is finite and positive almost surely, where* $\theta \equiv m/(2m+\mu)$, $m \equiv \mathrm{E}(X)$, *and* $\mu \equiv \mathrm{E}(\beta)$.

Proof: *It is known under the hypothesis* $\mathrm{E}(X\log(X)) < \infty$ *that there exists a random variable* $\zeta_i \in (0,\infty)$ *for which* $P(\lim_{t\to\infty} D_i(t)e^{-mt} = \zeta_i) = 1$. *Note also that* $\sum_{j=1}^{n} 1/S_{j-1} \xrightarrow{\text{a.s.}} \infty$ *as* $n \to \infty$ *by the strong law of large numbers and therefore* $\tau_n \xrightarrow{\text{a.s.}} \infty$ *as well by Theorem 3.3 above.*

Hence, setting $\theta \equiv m/(2m+\mu)$ *we have* $\forall\, i \geq 1$,

$$\frac{D_i(\tau_n - \tau_i)}{n^\theta} = \frac{D_i(\tau_n - \tau_i)}{e^{m(\tau_n-\tau_i)}} \times$$

$$\exp\left\{ m\left[\tau_n - \sum_{j=1}^{n}\frac{1}{S_{j-1}} + \sum_{j=1}^{n}\left(\frac{1}{S_{j-1}} - \frac{\alpha}{j}\right) + \alpha\left(\sum_{j=1}^{n}\frac{1}{j} - \log(n)\right) - \tau_i\right]\right\}$$

$$\xrightarrow{\text{a.s.}} \zeta_i e^{m(Y''+Y'+\alpha\gamma-\tau_i)} \in (0,\infty),$$

where γ *is the Euler-Mascheroni constant, and* Y'' *and* Y' *are random variables. Hence by the embedding theorem the same holds for the sequence* $\{d_n(i)n^{-\theta}\}_{n=1}^{\infty}$. □

Next we shift our focus to the vertex with maximal degree and study its growth rate. We will require the use of three lemmas in order to prove our main results.

Lemma 3.1. *For any* $c > 0$, $p > 1$ *and* $\epsilon \in (0,1)$, $\sum_{i=1}^{\infty}(\frac{i^{1-\epsilon}}{i^{1-\epsilon}+c})^{i/p} < \infty$.

Proof: *Take* $0 < \eta < c$ *and write,*

$$\sum_{i=1}^{\infty}\left(\frac{i^{1-\epsilon}}{i^{1-\epsilon}+c}\right)^{i/p} = \sum_{i=1}^{\infty}\left[\frac{\left(1-\frac{c}{i^{1-\epsilon}+c}\right)^{i^{1-\epsilon}}}{e^{-\eta}}\right]^{i^\epsilon/p} e^{-i^\epsilon\eta/p}.$$

Now $(1-\frac{c}{i^{1-\epsilon}+c})^{i^{1-\epsilon}} \downarrow e^{-c}$ *as* $i \to \infty$ *and hence* $(1-\frac{c}{i^{1-\epsilon}+c})^{i^{1-\epsilon}} \leq e^{-\eta}$ *for almost all* i, *ensuring that the first factor of each summand remains bounded for large* i. *In addition, for* *any $\alpha > 0$, $\epsilon \in (0,1)$ we have,*

$$\begin{aligned}\int_1^\infty e^{-\alpha x^\epsilon}\,dx &= \frac{1}{\alpha^{1/\epsilon}\epsilon}\int_\alpha^\infty u^{1/\epsilon - 1}e^{-u}\,du \\ &< \frac{\Gamma(1/\epsilon)}{\alpha^{1/\epsilon}\epsilon} \\ &< \infty,\end{aligned}$$

and thus $\sum_{i=1}^\infty e^{-i^\epsilon \eta/p} < \infty$. □

Lemma 3.2. *Let β be a nonnegative random variable, $p > 1$, and suppose $\mathrm{E}(\beta^{p+\delta}) < \infty$ for some $\delta > 0$. Then for any $c > 0$,*

$$\sum_{i=1}^\infty \mathrm{E}^{1/p}\left[\left(\frac{\beta}{\beta + c}\right)^i\right] < \infty.$$

Proof: *Let $\delta > 0$ be such that $\mathrm{E}(\beta^{p+\delta}) < \infty$ and take $\epsilon > 0$ so small that $(p+\delta)(1-\epsilon)/p > 1$. Then,*

$$\begin{aligned}\sum_{i=1}^\infty \mathrm{E}^{1/p}\left[\left(\frac{\beta}{\beta + c}\right)^i\right] &= \sum_{i=1}^\infty \left\{\mathrm{E}\left[\left(\frac{\beta}{\beta+c}\right)^i ; \beta \le i^{1-\epsilon}\right] + \mathrm{E}\left[\left(\frac{\beta}{\beta+c}\right)^i ; \beta > i^{1-\epsilon}\right]\right\}^{1/p} \\ &\le \sum_{i=1}^\infty \left(\frac{i^{1-\epsilon}}{i^{1-\epsilon}+c}\right)^{i/p} + \sum_{i=1}^\infty P(\beta > i^{1-\epsilon})^{1/p} \\ &\le \sum_{i=1}^\infty \left(\frac{i^{1-\epsilon}}{i^{1-\epsilon}+c}\right)^{i/p} + \mathrm{E}^{1/p}(\beta^{p+\delta})\sum_{i=1}^\infty \frac{1}{i^{(p+\delta)(1-\epsilon)/p}} \\ &< \infty,\end{aligned}$$

by Lemma 3.1 along with the assumption $\mathrm{E}(\beta^{p+\delta}) < \infty$. □

Lemma 3.3. *Let $\{Y_i\}_{i=1}^\infty$ be a collection of independent random variables with $\phi_i \equiv \mathrm{E}(Y_i) <$*

∞. *Then* $n^{-1}\sum_{i=1}^n (Y_i - \phi_i) \xrightarrow{\text{a.s.}} 0$ *whenever* $\sup_{i\geq 1} \mathrm{E}[(Y_i - \phi_i)^4] < \infty$.

Proof: *Let* $\epsilon > 0$. *By Markov's inequality we have,*

$$P\left(\frac{|\sum_{i=1}^n (Y_i - \phi_i)|}{n} \geq \epsilon\right) \leq \frac{\mathrm{E}\left\{[\sum_{i=1}^n (Y_i - \phi_i)]^4\right\}}{n^4\epsilon^4}.$$

Setting $c \equiv \sup_{i\geq 1} \mathrm{E}[(Y_i - \phi_i)^4] < \infty$,

$$\begin{aligned}
\mathrm{E}\left\{\left[\sum_{i=1}^n (Y_i - \phi_i)\right]^4\right\} &= \sum\sum\sum\sum_{1\leq i,j,k,l\leq n} \mathrm{E}\left[(Y_i - \phi_i)(Y_j - \phi_j)(Y_k - \phi_k)(Y_l - \phi_l)\right] \\
&= \sum_{1\leq i\leq n} \mathrm{E}\left[(Y_i - \phi_i)^4\right] + \binom{4}{2}\sum\sum_{1\leq i<j\leq n} \mathrm{E}\left[(Y_i - \phi_i)^2(Y_j - \phi_j)^2\right] \\
&\leq nc + 3n(n-1)c \\
&< 4n^2c,
\end{aligned}$$

where $\mathrm{E}\left[(Y_i - \phi_i)^2(Y_j - \phi_j)^2\right] \leq c$ *holds by the Cauchy-Schwarz inequality. Hence,*

$$\begin{aligned}
\sum_{n=1}^\infty P\left(\frac{|\sum_{i=1}^n (Y_i - \phi_i)|}{n} \geq \epsilon\right) &< \sum_{n=1}^\infty \frac{4c}{n^2\epsilon^4} \\
&< \infty,
\end{aligned}$$

so that $P(\{n^{-1}|\sum_{i=1}^n (Y_i - \phi_i)| \geq \epsilon \text{ i.o.}\}) = 0$ *and* $n^{-1}\sum_{i=1}^n (Y_i - \phi_i) \xrightarrow{\text{a.s.}} 0$ *since* ϵ *was arbitrary.* □

We can now prove the main theorem of this section.

Theorem 3.5. *If for an integer* $r > \theta^{-1}$ *we have* $\mathrm{E}(X^r) < \infty$ *and* $\mathrm{E}(\beta^{r+\delta}) < \infty$ *for some* $\delta > 0$, *then* $I_n \xrightarrow{\text{a.s.}} I \in \mathbb{N}_+$ *and* $M_n/n^\theta \xrightarrow{\text{a.s.}} \max_{i\geq 1}\gamma_i$.

Proof: *Consider the double array of random variables* $\{D_i(\tau_n - \tau_i)e^{-m\tau_n} : i \geq 1, n \geq i\}$

and note that $D_i(\tau_n - \tau_i)e^{-m\tau_n} \xrightarrow{\text{a.s.}} \zeta_i e^{-m\tau_i}$ *and* $\sup_{n\geq i} D_i(\tau_n - \tau_i)e^{-m\tau_n} \leq \widetilde{D}_i e^{-m\tau_i}$*, where* $\widetilde{D}_i \equiv \sup_{t\geq 0} D_i(t)e^{-mt}$*. If we can show that* $\widetilde{D}_i e^{-m\tau_i} \xrightarrow{\text{a.s.}} 0$ *as* $i \to \infty$*, then the hypotheses of Lemma 2.1 would be satisfied with probability one for this particular array, and hence* $\text{argmax}_{i\leq n} D_i(\tau_n - \tau_i)e^{-m\tau_n}$ *would almost surely be constant for large* n*. Now for* $\epsilon > 0$ *and* $r > \theta^{-1}$,

$$\sum_{i=1}^{\infty} P\left(\frac{\widetilde{D}_i}{i^\theta} \geq \epsilon\right) \leq \frac{\mathrm{E}(\widetilde{D}^r)}{\epsilon^r}\sum_{i=1}^{\infty}\frac{1}{i^{r\theta}}$$

$$< \infty,$$

so long as $\mathrm{E}(\widetilde{D}^r) < \infty$*, and note that this implies* $\widetilde{D}_i e^{-m\tau_i} \xrightarrow{\text{a.s.}} 0$ *since* $i^\theta e^{-m\tau_i} = e^{-m(\tau_i - \alpha\log(i))}$ *converges almost surely. Also, from Jensen's inequality we can deduce that,*

$$\mathrm{E}(\widetilde{D}^r) \leq \mathrm{E}(\widetilde{W^r})\sum_{i=0}^{\infty}\mathrm{E}\left[e^{-mT_i}\left(\sum_{j=0}^{\infty}e^{-mT_j}\right)^{r-1}\right].$$

Since $\mathrm{E}(X^r) < \infty$ *it follows that* $\mathrm{E}(\widetilde{W^r}) < \infty$*. In addition, by Hölder's inequality we have for* $p, q > 1$ *with* $1/p + 1/q = 1$ *that,*

$$\mathrm{E}\left[e^{-mT_i}\left(\sum_{j=0}^{\infty}e^{-mT_j}\right)^{r-1}\right] \leq \mathrm{E}^{1/p}(e^{-pmT_i})\,\mathrm{E}^{1/q}\left[\left(\sum_{j=0}^{\infty}e^{-mT_j}\right)^{q(r-1)}\right].$$

Now setting $q := r/(r-1)$ *we have,*

$$\mathrm{E}\left[\left(\sum_{j=0}^{\infty}e^{-mT_j}\right)^{q(r-1)}\right] = \mathrm{E}\left\{\mathrm{E}\left[\left(\sum_{j=0}^{\infty}e^{-mT_j}\right)^{r}\,\Big|\,\beta\right]\right\},$$

and note that conditional on β,

$$\begin{aligned}
\mathrm{E}\left[\left(\sum_{j=0}^{\infty} e^{-mT_j}\right)^r\right] &= r! \sum_{0\le j_1\le\cdots\le j_r} \mathrm{E}\left[\prod_{i=1}^{r} e^{-mT_{j_i}}\right] \\
&= r! \sum_{0\le j_1\le\cdots\le j_{r-1}} \mathrm{E}\left[\prod_{i=1}^{r-2} e^{-mT_{j_i}} e^{-2mT_{j_{r-1}}}\right] \left(\frac{\beta}{m+\beta}\right)^{j_r-j_{r-1}} \\
&\le r! \left(\frac{m+\beta}{m}\right) \sum_{0\le j_1\le\cdots\le j_{r-1}} \mathrm{E}\left[\prod_{i=1}^{r-2} e^{-mT_{j_i}}\right] \\
&\le r! \left(\frac{m+\beta}{m}\right)^r .
\end{aligned}$$

Therefore,

$$\begin{aligned}
\mathrm{E}\left[\left(\sum_{j=0}^{\infty} e^{-mT_j}\right)^{q(r-1)}\right] &\le r!\, \mathrm{E}\left[\left(\frac{m+\beta}{m}\right)^r\right] \\
&< \infty,
\end{aligned}$$

since $\mathrm{E}(\beta^r) < \infty$. *Also,* $q = r/(r-1) \Rightarrow p = r$ *and so,*

$$\begin{aligned}
\sum_{i=0}^{\infty} \mathrm{E}^{1/p}(e^{-pmT_i}) &= \sum_{i=0}^{\infty} \mathrm{E}^{1/r}\left[\mathrm{E}(e^{-pmT_i}) \,|\, \beta\right] \\
&= \sum_{i=0}^{\infty} \mathrm{E}^{1/r}\left[\left(\frac{\beta}{\beta+mp}\right)^i\right] \\
&< \infty,
\end{aligned}$$

by Lemma 3.2 and the hypothesis that $\mathrm{E}(\beta^{r+\delta}) < \infty$, *and hence* $\mathrm{E}(\widetilde{D}^r) < \infty$. $\square$

Next we consider the asymptotics of the degree distribution. For this we will use another lemma.

Lemma 3.4. *Let X be a nonnegative random variable and $c > 0$. Then,*

$$\mathrm{E}\left[\frac{\Gamma(X+c)}{\Gamma(X)}\right] < \infty,$$

if and only if $\mathrm{E}(X^c) < \infty$.

Proof: *Letting $\epsilon > 0$ we have by Stirling's formula that there is an M_ϵ such that $x^{-c}\Gamma(x+c)/\Gamma(x) \in (1-\epsilon, 1+\epsilon)$ whenever $x \geq M_\epsilon$. Therefore, setting $a_\epsilon \equiv \mathrm{E}[\Gamma(X+c)/\Gamma(X)\,;\, X < M_\epsilon] < \infty$ we have,*

$$a_\epsilon + (1-\epsilon)\mathrm{E}(X^c; X > M_\epsilon) \leq \mathrm{E}\left[\frac{\Gamma(X+c)}{\Gamma(X)}\right] \leq a_\epsilon + (1+\epsilon)\mathrm{E}(X^c; X > M_\epsilon),$$

and from this the claim follows since $\mathrm{E}(X^c; X > M_\epsilon) < \infty \Leftrightarrow \mathrm{E}(X^c) < \infty$ *for any* $\epsilon > 0$. □

Theorem 3.6. *Consider the random effects preferential attachment model where each newly-joining vertex connects once to its chosen neighbor, suppose* $\mathrm{E}(\beta\log(\beta)) < \infty$*, set $\alpha \equiv 1/(2+\mu)$, and let $\pi_j(n)$ denote the proportion of vertices with degree $j \geq 1$ at time n. Then $\pi_j(n) \xrightarrow{\mathrm{p}} \pi_j$ as $n \to \infty$ where,*

$$\pi_j \equiv \mathrm{E}\left[\left(\frac{1/\alpha}{j+\beta+1/\alpha}\right)\frac{\Gamma(j+\beta)}{\Gamma(j+\beta+1/\alpha)}\frac{\Gamma(1+\beta+1/\alpha)}{\Gamma(1+\beta)}\right].$$

Proof: *Let $p_j(t) \equiv P(D(t) = j)$ and write,*

$$\pi_j(n) \overset{\mathrm{d}}{=} \frac{1}{n}\sum_{i=1}^{n} I[D_i(\tau_n - \tau_i) = j]$$

$$
\begin{aligned}
&= \frac{1}{n}\sum_{i=1}^{n}\{I[D_i(\tau_n-\tau_i)=j]-I[D_i(\alpha\log(n/i))=j]\}\\
&+\frac{1}{n}\sum_{i=1}^{n}\{I[D_i(\alpha\log(n/i))=j]-p_j(\alpha\log(n/i))\}\\
&+\frac{1}{n}\sum_{i=1}^{n}p_j(-\alpha\log(i/n))-\int_0^1 p_j(-\alpha\log(x))\,dx\\
&+\int_0^1 p_j(-\alpha\log(x))\,dx\\
&\equiv T_1(n)+T_2(n)+T_3(n)+\int_0^1 p_j(-\alpha\log(x))\,dx.
\end{aligned}
$$

Now through a change of variables we note that,

$$\int_0^1 p_j(-\alpha\log(x))\,dx=\frac{1}{\alpha}\int_0^\infty p_j(t)e^{-t/\alpha}\,dt,$$

and so it will be enough to show that the integral above is equal to π_j as defined in the statement of the theorem, and that $T_1(n), T_2(n) \xrightarrow{\text{p}} 0$ and $T_3(n)\to 0$.

Now,

$$
\begin{aligned}
\mathrm{E}(|T_1(n)|) &\le \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}\left(|I[D_i(\tau_n-\tau_i)=j]-I[D_i(\alpha\log(n/i))=j]|\right)\\
&=\frac{1}{n}\sum_{i=1}^{n}P(|I[D_i(\tau_n-\tau_i)=j]-I[D_i(\alpha\log(n/i))=j]|=1)\\
&\le\frac{1}{n}\sum_{i=1}^{n}P(|I[D_i(\tau_n-\tau_i)=j]-I[D_i(\alpha\log(n/i))=j]|\ge 1).
\end{aligned}
$$

Since $\sup_{i\ge n}(\tau_n-\tau_i-\alpha\log({}^{n}\!/\!{}_{i})) \xrightarrow{\text{a.s.}} 0$ as $n\to\infty$ we conclude that this bound converges to zero as $n\to\infty$, and thus $T_1(n)\xrightarrow{\text{p}} 0$ by Markov's inequality. Next we note that the *hypotheses of Lemma 3.3 hold for the summands in* $T_2(n)$*, and hence* $T_2(n) \xrightarrow{\text{a.s.}} 0$*. Finally, since* $p_j(t)$ *is bounded and continuous and therefore Riemann integrable,* $T_3(n) \to 0$*.*

Turning our attention to π_j*, let* A_j *denote the time at which* $D(t)$ *enters state* j*,* B_j *the total time spent in state* j *and write,*

$$\begin{aligned}\int_0^\infty \frac{1}{\alpha} p_j(t) e^{-t/\alpha}\, dt &= \int_0^\infty \frac{1}{\alpha}\,\mathrm{E}[P(D(t)=j|A_j,B_j)]e^{-t/\alpha}\, dt \\ &= \mathrm{E}\left[\int_{A_j}^{A_j+B_j} \frac{1}{\alpha}\, e^{-t/\alpha}\, dt\right] \\ &= \mathrm{E}\left[e^{-A_j/\alpha}\left(1-e^{-B_j/\alpha}\right)\right] \\ &= \mathrm{E}\left\{\mathrm{E}\left[e^{-A_j/\alpha}\left(1-e^{-B_j/\alpha}\right) \mid \beta\right]\right\} \\ &= \mathrm{E}\left\{\mathrm{E}\left[e^{-A_j/\alpha} \mid \beta\right]\mathrm{E}\left[\left(1-e^{-B_j/\alpha}\right) \mid \beta\right]\right\}.\end{aligned}$$

Now using that (conditional on β*)* B_j *is exponential with rate* $j+\beta$ *and* $A_j \overset{\mathrm{d}}{=} \sum_{i=1}^{j-1} C_i$ *where the* C_i *are each independently distributed as exponential with respective rates* $\{i+\beta\}_{i=1}^{j-1}$ *we have,*

$$\mathrm{E}\left[\left(1-e^{-B_j/\alpha}\right) \mid \beta\right] = \frac{1/\alpha}{j+\beta+1/\alpha},$$

and,

$$\begin{aligned}\mathrm{E}\left[e^{-A_j/\alpha} \mid \beta\right] &= \prod_{i=1}^{j-1}\left(\frac{i+\beta}{i+\beta+1/\alpha}\right) \\ &= \frac{\Gamma(j+\beta)}{\Gamma(j+\beta+1/\alpha)}\frac{\Gamma(1+\beta+1/\alpha)}{\Gamma(1+\beta)},\end{aligned}$$

which together yield the theorem. □

Theorem 3.7. *Let* π_j *be defined as above. If* $\mathrm{E}(\beta^{1/\alpha}) < \infty$*, then* $\lim_{j\to\infty} j^{1/\alpha+1}\pi_j$ *exists, is* *finite and positive.*

Proof: *By Lemma 3.4 and the dominated convergence theorem,*

$$j^{1/\alpha+1}\pi_j = \mathrm{E}\left[\left(\frac{j/\alpha}{j+\beta+1/\alpha}\right) j^{1/\alpha}\frac{\Gamma(j+\beta)}{\Gamma(j+\beta+1/\alpha)}\frac{\Gamma(1+\beta+1/\alpha)}{\Gamma(1+\beta)}\right]$$

$$\to \frac{1}{\alpha}\mathrm{E}\left[\frac{\Gamma(1+\beta+1/\alpha)}{\Gamma(1+\beta)}\right] \in (0,\infty),$$

as $j \to \infty$. □

This last theorem provides a sufficient condition for $\{\pi_j\}_{j=1}^{\infty}$ to exhibit power law behavior. It is interesting perhaps to note that the condition $\mathrm{E}(\beta^{1/\alpha}) < \infty$ becomes stronger as the expected value μ increases, and one could speculate as to possible reasons for this. However, it should also be kept in mind that this may not actually be a property of the model itself since the condition could turn out to be unnecessary.

Finally, as a consequence of Theorem 3.6 we see that if the random variables $\{\beta_i\}_{i=1}^{\infty}$ are taken to be degenerate at $b \in [0,\infty)$ then the probabilities become,

$$\pi_j = \left(\frac{2+b}{j+2+2b}\right)\frac{\Gamma(j+b)}{\Gamma(j+2+2b)}\frac{\Gamma(3+2b)}{\Gamma(1+b)},$$

which for $b = 0$ reduces to,

$$\pi_j = \frac{4}{(j+2)(j+1)j},$$

and this is the corresponding result for the standard Barabasi-Albert model.

Chapter 4: Cascades on trees

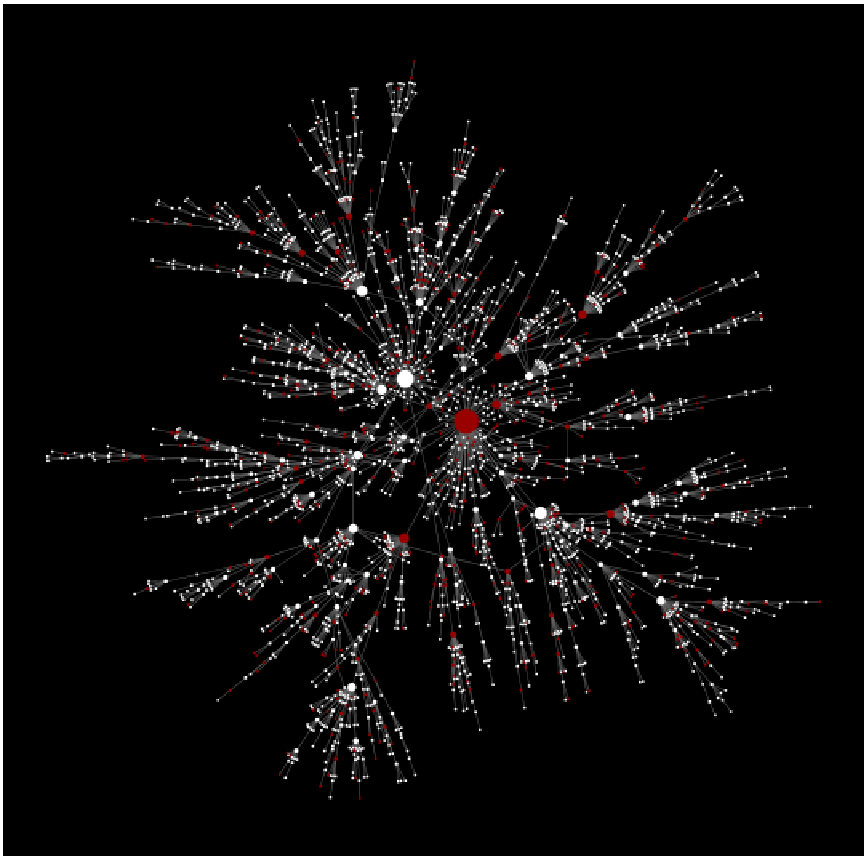

Figure 4.1: Preferential attachment tree with cascade.

4.1 Introduction

Often in real applications we are not concerned only with the evolution of the graph itself, but also with how information or activity spreads over that graph. The spread of ideas or diseases amongst people are both examples of such processes. In this chapter we will consider preferential attachment models which form a stage for cascading activity. The basic model we will consider is as follows. Begin with a simple Barabasi-Albert model initialized by two nodes, one active and one inactive. When a new node connects to an active node, it activates with probability p, otherwise it becomes inactive, and it becomes inactive whenever it connects to an inactive node. We could view these events as either the transmission of an idea or an infection event. One question we might ask about such a process is, what happens to the proportion of active nodes in the tree as the size of the tree tends to infinity? To address this question it will be useful to make use of a certain equivalence between preferential attachment trees and Pólya urns.

We first state a special case of an important theorem concerning exchangeable random variables (for a proof see Hofstad [10]):

Theorem 4.1. *(De Finetti's theorem.) Let $\{X_i\}_{i=1}^{\infty}$ be an infinite sequence of exchangeable Bernoulli random variables. Then there exists a random variable $U \in [0,1]$ such that for all $1 \leq k \leq n$,*

$$P(X_1 = \ldots = X_k = 1, X_{k+1} = \ldots = X_n = 0) = \mathrm{E}[U^k(1-U)^{n-k}]. \ \square$$

This states that an infinite sequence of exchangeable Bernoulli random variables has the same distribution as an i.i.d. sequence with a random success probability U. An immediate consequence of this is that if we set $S_n \equiv \sum_{i=1}^{n} X_i$, then $S_n/n \xrightarrow{\text{a.s.}} U$ by the strong law of large numbers. Note that it is crucial that the sequence be infinite, otherwise one has trivial counterexamples such as $X_1 \sim \text{Bernoulli}(1/2)$ and $X_2 \equiv 1 - X_1$. The pair (X_1, X_2) is exchangeable, but if we had a random variable U satisfying the above, then $0 = P(X_1 = 1, X_2 = 1) = \mathrm{E}(U^2)$, so that $U \overset{\text{a.s.}}{=} 0$ and hence $P(X_1 = 0, X_2 = 0) = 1$, which is obviously not true.

A famous model where we encounter an infinite sequence of exchangeable random variables is the Pólya urn process. This is a process where one begins at time $n = 1$ with an urn containing some number of white balls and some number of black balls, which we denote W_1 and B_1 respectively. Then at time $n \geq 1$, a ball is drawn from the urn where the probability that it is black is equal to $\omega_B(B_n)/[\omega_B(B_n)+\omega_W(W_n)]$, with $\omega_B(k)=k+a_B$, $\omega_W(k)=k+a_W\ a_B, a_W \geq 0$ constants, and W_n and B_n denoting the number of white and black balls in the urn before the n^{th} ball is selected. Once a ball is drawn, it is replaced along with another ball of the same color. It turns out one can show that the Bernoulli random variables $\{B_{n+1}-B_n\}_{n\geq 1}$ are indeed exchangeable, and we can then by an application of de Finetti's theorem obtain the next result (a proof of which is also found in Hofstad [10]):

Theorem 4.2. *Let $\{(B_n, W_n)\}_{n=1}^{\infty}$ be a Pólya urn process with weight functions $\omega_W(k)=k+a_W$ and $\omega_B(k)=k+a_B$. Then,*

$$\frac{B_n}{B_n+W_n}\xrightarrow{\text{a.s.}} U,$$

where $U \sim$ beta$(\alpha = B_1+a_B, \beta = W_1+a_W)$. □

4.2 Preferential attachment trees

To understand the proportion of active vertices in our model we use a decomposition of the tree into an infinite sequence of nested Pólya urns, a procedure which we will now describe.

Fix nonnegative integers $\{x_i\}_{i=1}^{\infty}$ and consider constructing a sequence of Pólya urns $\{\mathcal{U}_j\}_{j=1}^{\infty}$ (with arbitrary weight functions) as follows. Initialize $\mathcal{U}_1$ with one black and one white ball and allow it to evolve as usual. After the $(x_1+1)^{\text{th}}$ black ball has been added to the urn, initialize a second urn $\mathcal{U}_2$ with one white ball, and a number of black balls equal to the number of black balls in $\mathcal{U}_1$ minus one. The second urn now represents a decomposition of the black balls in the first urn into white and black balls (think of the last black ball added as being counted as white in the newly generated urn), and a ball is drawn from $\mathcal{U}_2$ when and only when a black ball is drawn from $\mathcal{U}_1$. Now after the $(x_2+1)^{\text{th}}$ black ball within $\mathcal{U}_2$ is selected, initialize $\mathcal{U}_3$ as in the case of $\mathcal{U}_2$, with one white ball and a number of black balls equal to the number of black balls in $\mathcal{U}_2$ minus one, and again drawing a ball from $\mathcal{U}_3$ only when a black ball is drawn from $\mathcal{U}_2$. Continue this process for $j \geq 4$ and in this way generate the collection $\{\mathcal{U}_j\}_{j=1}^{\infty}$. Denote by $\mathcal{B}_{j,n}$ the proportion of black balls in urn j at time n, and $\mathcal{B}_{j,\infty} \equiv \lim_{n\to\infty} \mathcal{B}_{j,n}$.

It turns out that the random variables $\{\mathcal{B}_{j,\infty}\}_{j=1}^{\infty}$ defined above are independent. (This may seem fairly obvious, but it is still nontrivial since for $j < k$ and each $n \in \mathbb{N}$ the random variables $\mathcal{B}_{j,n}$ and $\mathcal{B}_{k,n}$ are certainly not independent since a change in $\mathcal{B}_{k,n}$ is always accompanied by an increase in $\mathcal{B}_{j,n}$, and hence also a decrease in $\mathcal{B}_{j,n}$ implies that $\mathcal{B}_{k,n}$ remains constant.) To see this, consider generating an independent collection of Pólya urns $\{\mathcal{U}^*\}_{j=1}^{\infty}$ where the $\mathcal{U}^*$ have the same starting allocations as the $\{\mathcal{U}\}_{j=1}^{\infty}$, and so by construction the associated $\{\mathcal{B}_{j,\infty}^*\}_{j=1}^{\infty}$ are independent random variables. Now interleave the processes $\{\mathcal{U}_j^*\}_{j=1}^{\infty}$ in such a way that they obey the mechanism described above (which only requires adjusting event times) and the resulting collection will be a realization of $\{\mathcal{U}_j\}_{j=1}^{\infty}$ so that in particular $\mathcal{B}_{l,\infty} = \mathcal{B}_{l,\infty}^*$ for every $l \geq 1$, and hence $\{\mathcal{B}_{j,\infty}\}_{j=1}^{\infty} \overset{\mathrm{d}}{=} \{\mathcal{B}_{j,\infty}^*\}_{j=1}^{\infty}$. We can now prove the following.

Theorem 4.3. *Let A_n denote the number of active vertices within the tree at time n. If $p = 1$, $A_n/n \xrightarrow{\text{a.s.}} Y \sim beta(\alpha = 1/2, \beta = 1/2)$, and if $p < 1$, $A_n/n \xrightarrow{\text{a.s.}} 0$.*

Proof: *Let $\mathcal{T}_{0,n}$ be the tree rooted in the initial active vertex, and for $j \geq 1$ at the time of the j^{th} failed activation let $\mathcal{T}_{j,n}$ contain all those vertices within $\mathcal{T}_{j-1,n}$ which are not rooted in the unique inactive vertex, and denote $T_{j,n} \equiv |\mathcal{T}_{j,n}|$, the number of nodes in tree $T_{j,n}$. Then, after k failed activations we obtain a nested decreasing sequence of subtrees $\mathcal{T}_{0,n} \supset \mathcal{T}_{1,n} \supset \ldots \supset \mathcal{T}_{k,n}$ with $T_{k,n} = A_n$ and can write,*

$$\frac{A_n}{n} = \frac{T_{0,n}}{n} \cdot \frac{T_{1,n}}{T_{0,n}} \cdot \frac{T_{2,n}}{T_{1,n}} \cdots \frac{T_{k,n}}{T_{k-1,n}}.$$

Next we observe that the sequences $\{T_{0,n}, n - T_{0,n}\}$ and $\{T_{j,n}, T_{j-1,n} - T_{j,n}\}$ for $j \geq 1$ are in fact equivalent to a collection of nested Pólya urn processes as above. To see this, note that after the j^{th} failed activation of a vertex connecting to an active one, the total degree *of the active cluster is $2T_{j,n} + j - 1$ (each vertex other than the root contributes two to the degree while the initial root vertex contributes one, and each inactive vertex connected to the cluster contributes one as well) and increases by two with each addition of a new vertex. For the initial inactive vertex in this subtree, the total degree is one and increases by two with each additional vertex. Therefore, conditional on a vertex falling within $\mathcal{T}_{j-1,n}$, the probability that it connects to $\mathcal{T}_{j,n}$ is,*

$$\frac{2T_{j,n} + j - 1}{2T_{j,n} + j - 1 + 2(T_{j-1,n} - T_{j,n}) - 1} = \frac{T_{j,n} + (j-1)/2}{T_{j,n} + (j-1)/2 + T_{j-1,n} - T_{j,n} - 1/2}.$$

However, these are the transition probabilities of a Pólya urn process with weight functions $\omega_A(k,j) = k + (j-1)/2$ and $\omega_I(k) = k - 1/2$, and so the processes are equivalent since both have the Markov property. Using also that for $j \geq 0$ the initial allocation of the two corresponding urns is $1 + \sum_{i=1}^{j} X_i$ active balls, with $\{X_i\}_{i=1}^{\infty} \overset{\text{i.i.d.}}{\sim}$ geometric$(1-p)$, and one inactive ball, we have that $T_{j,n}/T_{j-1,n}$ (with the interpretation that $T_{-1,n} \equiv n$) converges a.s. to a random variable with distribution beta$(\alpha = 1 + \sum_{i=1}^{j} X_i + (j-1)/2, \beta = 1/2)$. We have now established the first claim since when $p = 1$ we have no failures and only one such process for which $j = 0$ and $\alpha = \beta = 1/2$. Returning to the product formula above and the case $p < 1$ we have,

$$\frac{A_n}{n} \xrightarrow{\text{a.s.}} \prod_{j=0}^{\infty} B_j,$$

where $B_j \sim$ beta$(\alpha = 1 + \sum_{i=1}^{j} X_i + (j-1)/2, \beta = 1/2)$. In addition, the B_j are conditionally *independent given* $\{X_i\}_{i=1}^{\infty}$ *and so we compute,*

$$\begin{aligned}
\mathrm{E}\left[\prod_{j=0}^{\infty} B_j\right] &= \mathrm{E}\left[\mathrm{E}\left(\prod_{j=0}^{\infty} B_j \,\middle|\, \{X_i\}_{i=1}^{\infty}\right)\right] \\
&= \mathrm{E}\left[\prod_{j=0}^{\infty} \mathrm{E}\left(B_j \mid \{X_i\}_{i=1}^{\infty}\right)\right] \\
&= \mathrm{E}\left[\prod_{j=0}^{\infty} \frac{1+j+2\sum_{i=1}^{j} X_i}{2+j+2\sum_{i=1}^{j} X_i}\right].
\end{aligned}$$

However, by the strong law of large numbers we have,

$$\frac{j}{2+j+2\sum_{i=1}^{j} X_i} \xrightarrow{\text{a.s.}} \frac{1}{1+2p/(1-p)} > 0,$$

and so we conclude that,

$$\begin{aligned}
\sum_{j=1}^{\infty} \frac{1}{2+j+2\sum_{i=1}^{j} X_i} &= \sum_{j=1}^{\infty} \frac{1}{j} \cdot \frac{j}{2+j+2\sum_{i=1}^{j} X_i} \\
&\overset{\text{a.s.}}{=} \infty,
\end{aligned}$$

which by Lemma 1.3 implies that,

$$\prod_{j=0}^{\infty} \frac{1+j+2\sum_{i=1}^{j} X_i}{2+j+2\sum_{i=1}^{j} X_i} \overset{\text{a.s.}}{=} 0.$$

Therefore $\mathrm{E}(\prod_{j=0}^{\infty} B_j) = 0$*, and it follows that* $\prod_{j=0}^{\infty} B_j \overset{\text{a.s.}}{=} 0$. □

Note that the distribution of X was irrelevant in the previous argument aside from the fact that it had finite first moment. This suggests the following model and theorem which can be proven in an identical manner, and of which the former result is a special case.

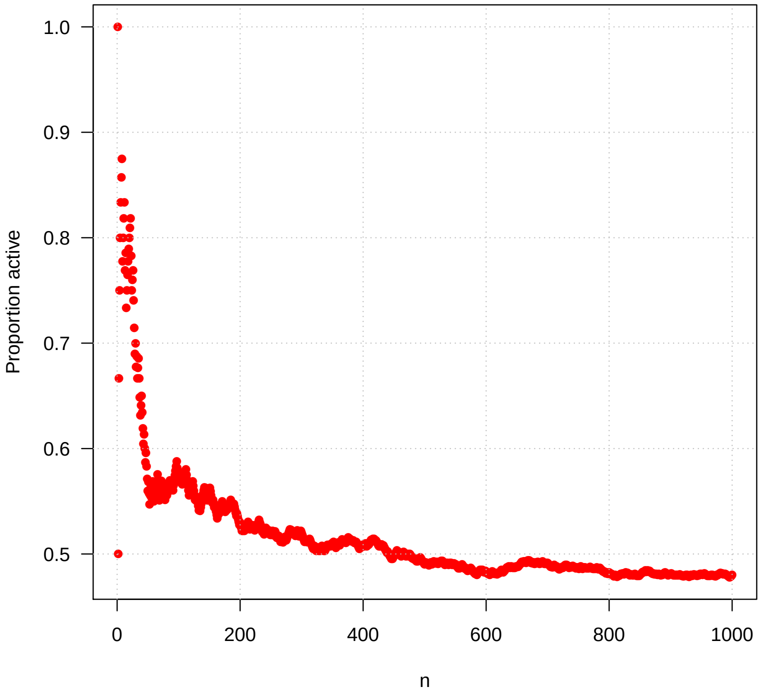

Figure 4.2: Proportion of active vertices in a Barabasi-Albert model with cascade and $p = {}^{9}/_{10}$. Even with a large probability of transmission the cascade eventually dies.

Theorem 4.4. *Let* $\{X_i\}_{i=1}^{\infty}$ *be i.i.d. nonnegative integer-valued random variables and consider the model where instead of vertices activating with probability* p *after connecting to an active vertex, for* $i \geq 1$ *a random number of vertices* X_i *are activated before the* i^{th} *failed activation, followed by* X_{i+1} *more activations and then the* $(i+1)^{th}$ *failure, and so on. Then* $A_n/n \xrightarrow{\text{a.s.}} 0$ *so long as* $\mathrm{E}(X) < \infty$. $\square$

This raises the question as to what happens when $\mathrm{E}(X) = \infty$. For instance, is this enough to give us a nonnegligible proportion of active vertices in the limit? Can anything else be said about the sequence $\{A_n\}_{n=1}^{\infty}$?

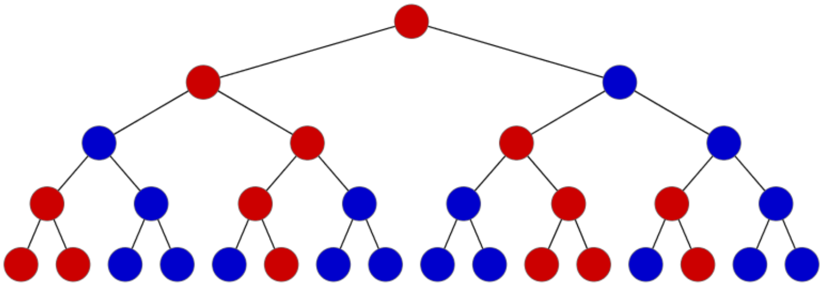

Figure 4.3: Several generations of a binary tree with cascade. (Red nodes are active, blue inactive.)

4.3 *b*-nary trees

Another cascade model we wish to study is as follows: consider a b-nary tree which at time zero consists of a single active root node, and at time $i \geq 1$, each of the leaf nodes at time $i - 1$ branches into b additional nodes, and the probability that a new leaf node becomes active is inversely proportional to some increasing function $\omega : \mathbb{N}_+ \mapsto (1, \infty)$ applied to that node's distance from it's most recent active ancestor, with all events being independent given these distances (hereafter referred to as a node's "depth"). We are interested in which functions ω will cause the number of active nodes to remain bounded with positive probability as the size of the tree tends to infinity. In particular, can this probability be one?

Let us denote by Λ the event that the cascade "dies out," and by Λ_0 the event that the last active vertex is the root. In addition, we note that $P(\Lambda) > 0$ if and only if $P(\Lambda_0) > 0$ since when $P(\Lambda_0) = 0$ it's almost surely the case that we have at least one activation in the tree, whereupon there will almost surely be another activation in the subtree rooted in the activated vertex, and so on ad infinitum, implying $P(\Lambda) = 0$ (that $P(\Lambda_0) > 0 \Rightarrow P(\Lambda) > 0$ is clear since $\Lambda_0 \subset \Lambda$). Further, $P(\Lambda_0)$ can be calculated directly as,

$$P(\Lambda_0) = \prod_{k=1}^{\infty} \left(1 - \frac{1}{\omega(k)}\right)^{b^k},$$

since Λ_0 occurs if and only if at time one each node (of which there are b, all with depth one) fails to become active, then at time two all b^2 vertices fail to activate given that the nodes at time one are inactive, and so forth. From this we obtain the following result.

Theorem 4.5. *$P(\Lambda) > 0$ if and only if,*

$$\sum_{k=1}^{\infty} \frac{b^k}{\omega(k)} < \infty.$$

Proof: *This follows from the fact that $P(\Lambda) > 0 \Leftrightarrow P(\Lambda_0) > 0$, and the product formula above is nonzero if and only if $\sum_{k=1}^{\infty} b^k/\omega(k) < \infty$.* $\square$

4.4 Galton-Watson trees

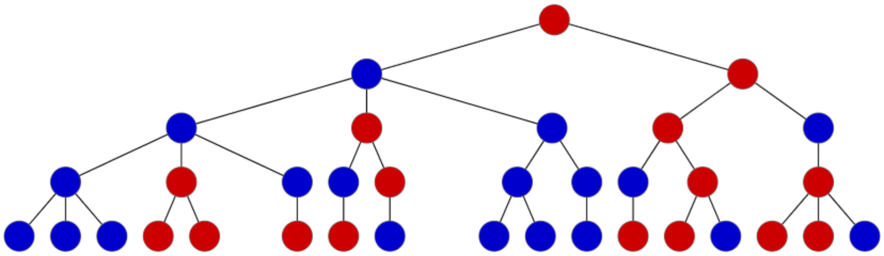

Figure 4.4: Several generations of a Galton-Watson tree with cascade.

A natural generalization of the previous development is to allow the evolution of the tree to be random rather than deterministic. Towards this end, set $Z_0 = 1$ and for $n \geq 1$ $Z_n = \sum_{j=1}^{Z_{n-1}} \zeta_{n-1,j}$, where $\{\zeta_{n,j} \,:\, n \geq 0\,,\, 1 \leq j \leq Z_n\}$ is a double array of i.i.d. strictly positive integer-valued random variables with $m \equiv \mathrm{E}(\zeta) < \infty$. Now consider the model which is as above only the node structure is that of a Galton-Watson tree, and note that the previous argument which established $P(\Lambda) > 0 \Leftrightarrow P(\Lambda_0) > 0$ applies to this model without modification. Similar to before we have,

$$\begin{aligned} P(\Lambda_0) &= \mathrm{E}\left[P(\Lambda_0 \,|\, \{Z_n\}_{n=1}^{\infty})\right] \\ &= \mathrm{E}\left[\prod_{k=1}^{\infty}\left(1-\frac{1}{\omega(k)}\right)^{Z_k}\right] \\ &> 0, \end{aligned}$$

if and only if $\prod_{k=1}^{\infty}(1-\omega(k)^{-1})^{Z_k}$ can be positive with positive probability, and note again that this product is positive or zero according to whether $\sum_{k=1}^{\infty} Z_k/\omega(k)$ converges or diverges. Now,

$$\sum_{k=1}^{\infty}\frac{Z_k}{\omega(k)} = \sum_{k=1}^{\infty}\frac{m^k}{\omega(k)}\cdot\frac{Z_k}{m^k},$$

and if $\mathrm{E}(\zeta\log(\zeta)) < \infty$ we have $Z_k/m^k \xrightarrow{\text{a.s.}} W \in (0,\infty)$ [6], so that the convergence of the above series coincides almost surely with that of $\sum_{k=1}^{\infty} m^k/\omega(k)$. If $\mathrm{E}(\zeta\log(\zeta)) = \infty$, then it is known that there is a sequence $\{c_k\}_{k=1}^{\infty}$ such that $c_{k+1}/c_k \to m$ and $c_k/m^k \to 0$ as $k\to\infty$ for which $Z_k/c_k \xrightarrow{\text{a.s.}} W' \in (0,\infty)$, and so in this case we obtain the same result as before with m^k replaced by c_k. Summarizing these ideas:

Theorem 4.6. *Set $m \equiv \mathrm{E}(\zeta) < \infty$. If $\mathrm{E}(\zeta\log\zeta) < \infty$, then $P(\Lambda) > 0$ if and only if,*

$$\sum_{k=1}^{\infty}\frac{m^k}{\omega(k)} < \infty.$$

If on the other hand $\mathrm{E}(\zeta\log\zeta) = \infty$, then there exists a sequence of constants $\{c_k\}_{k=1}^{\infty}$ satisfying $c_{k+1}/c_k \to m$ and $c_k/m^k \to 0$ as $k\to\infty$ such that $P(\Lambda) > 0$ if and only if,

$$\sum_{k=1}^{\infty}\frac{c_k}{\omega(k)} < \infty. \ \square$$

This subsumes the result for b-nary trees since the latter is a special case of the Galton-Watson tree model with ζ degenerate at b, so that $m = b$ and $\mathrm{E}(\zeta \log(\zeta)) = b \log b < \infty$.

We also have an interesting conjecture: consider the Galton-Watson tree model and assume that $\mathrm{E}(\zeta \log(\zeta)) < \infty$. Then $P(\Lambda)$ is zero or one according to whether $\sum_{k=1}^{\infty} m^k/\omega(k)$ diverges or converges. An intuitive argument for this is as follows. That $\sum_{k=1}^{\infty} m^k/\omega(k) = \infty \Rightarrow P(\Lambda) = 0$ under the assumption $\mathrm{E}(\zeta \log(\zeta)) < \infty$ has already been proven so we focus on the case $\sum_{k=1}^{\infty} m^k/\omega(k) < \infty$. Let A_n denote the number of active nodes in the tree other than the root at time n and $A_\infty \equiv \lim_{n\to\infty} A_n$, so that Λ occurs if and only if $A_\infty < \infty$. Now letting $I_k^{(j)}$ be a Bernoulli random variable equal to one if the j^{th} node within generation k is active we can write $A_n = \sum_{k=1}^{n} \sum_{j=1}^{Z_k} I_k^{(j)}$, and using also that the $I_k^{(j)}$ are identically distributed over j and independent of Z_k we have,

$$\begin{aligned}
\mathrm{E}(A_n) &= \mathrm{E}\left(\sum_{k=1}^{n}\sum_{j=1}^{Z_k} I_k^{(j)}\right) \\
&= \sum_{k=1}^{n} \mathrm{E}\left[\mathrm{E}\left(\sum_{j=1}^{Z_k} I_k^{(j)} \mid Z_k\right)\right] \\
&= \sum_{k=1}^{n} \mathrm{E}\left[Z_k \,\mathrm{E}\left(I_k^{(1)}\right)\right] \\
&= \sum_{k=1}^{n} m^k \,\mathrm{E}\left(I_k^{(1)}\right).
\end{aligned}$$

In addition, letting D_i denote the depth of the first node in generation i we arrive at,

$$\begin{aligned}
\mathrm{E}\left(I_k^{(1)}\right) &= \mathrm{E}\left[\mathrm{E}\left(I_k^{(1)} \mid D_{k-1}\right)\right] \\
&= \mathrm{E}\left(\frac{1}{\omega(D_{k-1}+1)}\right),
\end{aligned}$$

and thus by monotone convergence,

$$\begin{aligned}\mathrm{E}(A_\infty) &= \lim_{n\to\infty} \mathrm{E}(A_n) \\ &= \sum_{k=1}^{\infty} m^k \, \mathrm{E}\left(\frac{1}{\omega(D_{k-1}+1)}\right).\end{aligned}$$

Therefore, if it were the case that $\sum_{k=1}^{\infty} m^k/\omega(k) < \infty$ implies the convergence of the above series, then the conjecture would be proven since $\mathrm{E}(A_\infty) < \infty \Rightarrow P(A_\infty < \infty) = 1$. One would expect this to hold since it can shown that with probability one D_{k-1} increases by one at each step for large k, so that the behavior of $\omega(D_{k-1}+1)$ is eventually like that of $\omega(k)$. Unfortunately, calculating or even approximating this expectation is still not easy. One approach may be to use $T \equiv \max\{i : D_i = 0, i \geq 1\}$ (which is an almost surely finite random variable) by noting that $D_i = i - T$ whenever $T \leq i$ and thus,

$$\begin{aligned}\mathrm{E}\left(\frac{1}{\omega(D_{k-1}+1)}\right) &= \mathrm{E}\left(\frac{1}{\omega(D_{k-1}+1)}\,; T < k\right) + \mathrm{E}\left(\frac{1}{\omega(D_{k-1}+1)}\,; T \geq k\right) \\ &\leq \mathrm{E}\left(\frac{1}{\omega(k-T)}\,; T < k\right) + P(T \geq k).\end{aligned}$$

Therefore if we can understand the distribution of T (in particular obtaining tail bounds of order $\omega(k-l)^{-1}$ for some $l \in \mathbb{N}_+$ as $k \to \infty$), then it may be possible to prove the conjecture.

Chapter 5: Conclusion

In this dissertation we have proposed a new methodology for performing statistical inference on preferential attachment networks which begins to fill a gap in the existing literature. We made the following contributions: proposed two strongly consistent estimators which are capable of measuring the strength of preferential attachment based on graph data, and also established the asymptotic normality for a special case of these estimators. We generalized the model of Athreya, Sethuraman and Ghosh to one which allows for random effects, thereby including added heterogeneity into the process which is not represented in the existing model.

While our inferential approach has many attractive features, it has some drawbacks as well. First, it makes fairly stringent assumptions about the nature of the attachment mechanism. For instance, the probability of attachment need not be a linear function of the degree. Second, it ignores much of the information contained in the graph by only looking at small numbers of vertices. Finally, the proposed statistics require not only knowledge of the graph at the present time, but also information on past states as well. If this information is not available, then these methods cannot be used.

To address the first issue, we might instead wish to model the probability of attachment as being proportional to some increasing function $f : \mathbb{N}_+ \to \mathbb{R}_+$ of the degree and treat the problem as one of function estimation. Another area for future research is how one might test the assumption of preferential attachment. Might it be possible to devise a goodness of fit test for this purpose? Regarding the last question, it could be argued that without information from multiple time points these types of inferences could not be made at all, since they pertain to the evolution of the graph itself.

Bibliography

[1] Albert, R., Barabasi, A.L. (2002). "Statistical mechanics of complex networks." *Review of Modern Physics*, Vol. 74, 47-97.

[2] Athreya, K.B. (2007). "Preferential attachment random graphs with general weight function." *Internet Mathematics*, Vol. 4, No. 4, 401-418.

[3] Athreya, K.B., Ghosh, A., Sethuraman, S. (2008). "Growth of preferential attachment random graphs via continuous-time branching processes." *Mathematical Proceedings of the Indian Academy of Sciences*, 118:3, 473-494.

[4] Athreya, K.B., Karlin, S. (1968). "Embedding of urn schemes into continuous time Markov branching processes and related limit theorems." *Annals of Mathematical Statistics*, Vol. 39, No. 6, 1801-1817.

[5] Athreya, K.B., Karlin S. (1967) "Limit theorems for the split times of branching processes." *J. Math. Mech.* Vol. 17, 257-277.

[6] Athreya, K.B., Ney, P. (2004). *Branching Processes.* Dover, New York.

[7] Bollobas, B., Riordan, O., Spencer, J., Tusnady, G. (2001). "The degree sequence of a scale-free random graph process." *Random Structures and Algorithms*, Vol. 18, 279-290.

[8] Durrett, R. (2007). *Random graph dynamics.* Cambridge University Press, Cambridge.

[9] Eggenberger F., Pólya, G. (1923). "Über die statistik verketteter Vorgänge." *Zeitschrift für Angewandte Mathematik und Mechanik*, Vol. 3, Issue 4, 279-289.

[10] Hofstad, R. (2009). *Random graphs and complex networks.* Eindhoven. Lecture notes.

[11] Ney, P., Vidyashankar, A.N. (2003). "Harmonic moments and large deviation rates for supercritical branching processes." *Annals of Applied Probability*, Vol. 13, Issue 2, 475-489.

[12] Pólya, G. (1931). "Sur quelques points de la theorie des probabilites." *Annales de l'Institut Henri Poincare*, Vol. 1, Issue 2, 117-161.

[13] Williams, D. (1991). *Probability with Martingales.* Cambridge University Press.

Biography

Daniel Saxton grew up in Loudoun County, Virginia and graduated from Loudoun Valley High School in 2001. He attended George Mason University where he earned both his B.S. in Economics in 2007, and M.S. in Statistical Science in 2011.