Electronic Journal of Statistics
ISSN: 1935-7524

On the Consistency of Estimating General Preferential Attachment Functions

Fengnan Gao* **and Aad van der Vaart†**

Mathematical Institute
Leiden University
Niels Bohrweg 1
2333 CA Leiden, Netherlands

gaof@math.leidenuniv.nl avdvaart@math.leidenuniv.nl

Rui Castro and Remco van der Hofstad

Department of Mathematics
Eindhoven University of Technology
P.O. Box 513
5600 MB Eindhoven, Netherlands.

rmcastro@tue.nl rhofstad@win.tue.nl

Abstract: We propose an empirical estimator (EE) of general preferential attachment function f in the settings of the preferential attachment model and proceed to prove the consistency of the proposed estimator. Numerical illustrations are presented to demonstrate the performance of the EE.

MSC 2010 subject classifications: Primary 62G20.
Keywords and phrases: Preferential Attachment Model, Consistency, Branching Process.

1. Introduction

Introduced by [1] in 1999, preferential attachment model (PAM) has since gained popularity in the past decade. It has been known to be a candidate to model networks where the "scale-free" property appears. The idea is to build a *dynamic* network where "the rich get richer". Suppose one new vertex comes in and wants to connect to the existing vertices. We assume the new vertex is more likely to connect the vertices with high degrees in a certain way. Suppose we do this for the vertices that keep coming one by one recursively, then we get a preferential attachment model.

To be concrete, we formulate the model mathematically. We start with two vertices connecting to each other. Then the recursive scheme starts. When a new vertex comes in at time n, it looks at all the existing n vertices and makes

*Research supported by the Netherlands Organization for Scientific Research.

†The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

an independent multinomial choice to connect to only one existing vertex. The probability to connect a vertex with degree k is proportional to the associated "attractiveness" $f(k)$ where $f\colon \mathbb{N}_+ \to \mathbb{R}_+$. If f is affine as in [5], it is well known that the asymptotic degree sequence p_k as $t \to \infty$ follows a power-law. In fact [5] even showed a central limit theorem holds for the empirical degree distribution $P_k(n)$. Moreover if f is linear ([2, 5]), one can work out the exact asymptotic degree distribution $p_k = 4/(k(k+1)(k+2))$. [6] works out the asymptotic degree distribution for non-affine f. Though of much statistical interests, we know little about how to estimate the preferential attachment function f given the evolution of the model. We aim to fill in the blank with the paper.

This paper is organized as follows. In Section 2, we introduce the terminologies of branching processes and give a random tree model that is equivalent to the evolution of the PAM. In Section 3, we then proceed to introduce the empirical estimator and prove the almost sure consistency of the estimator within the framework of equivalent continuous time random tree model.

2. Borrowing strength from branching process

In this section, we introduce the terminologies needed to state the preferential attachment model in the language of branching process, similar to [6].

2.1. Rooted ordered tree

We consider the model as an evolving genealogical tree, where individuals (in the genealogical tree) are vertices (in the PAM) and parent-child relations are edges. It is of convenience to keep track of the birth orders of children of the same parent. Therefore we consider our PAMs as *rooted ordered tree*. As we later will see, a carefully designed rooted ordered random tree will coincide with the PAM in terms of certain quantities, such as vertices' degrees.

To label the vertices, we let $\emptyset$ the root of the tree and its children are labeled with (1), (2) ...and so on. In general an individual $x = (i_1, \dots, i_n)$ is the i_n-th child of the i_{n-1}-th child of ...of the i_1-th child of the root $\emptyset$. Suppose $\mathbb{N}_+$ is the set of positive natural numbers and $\mathscr{I}_n = \{(i_1, \dots, i_n) | i_j \in N_+\}$ is called the individual space, the set of all possible labels is the following

$$\mathscr{I} = \{\emptyset\} \cup \left(\bigcup_{n=1}^{\infty} \mathscr{I}_n \right).$$

If $x = (x_1, \dots, x_k)$ and $y = (y_1, \dots, y_l)$ we use the shorthand xy for the concatenation $(x_1, \dots, x_k, y_1, \dots, y_l)$. To simply notation we use xn to denote the concatenation of $(x_1, \dots, x_k, n)$.

Since the labelling of all vertices contains all the necessary information of edges regarding the tree, we can write down a rooted ordered tree as the collection of vertex labels. It is clear that $G \subset \mathscr{I}$ may be a rooted ordered tree as long as the labels satisfy the *compatibility condition*, ie. for every $(x_1, \dots, x_k) \in G$

we have $(x_1, \dots, x_{k-1}) \in G$ ("parent must be") as well as $(x_1, \dots, x_k - 1) \in G$ if $x_k \geq 2$ ("older sibling must be"). The set of all finite rooted ordered trees is denoted by $\mathcal{G}$. We think of $G \in \mathcal{G}$ as an oriented tree with edges pointing from parents to children. We define the *degree* of vertex $x \in G$ to be the number of its children in G plus 1

$$\deg(x, G) = |\{n \in N_+ | xn \in G\}| + 1.$$

Suppose $|x| = n$ if and only if $x \in \mathscr{I}_n$, the n-th ($n \geq 1$) generation of $G \in \mathcal{G}$ is

$$G_{[n]} = \{x \in G | |x| = n\}.$$

The subtree rooted at vertex $x \in G$ is defined as

$$G_{\downarrow x} = \{y | xy \in G\},$$

which is the progeny of x in G with x viewed as the root.

2.2. Branching process

We consider the pure birth process with a typical individual reproduces according to a random point process ξ on $[0, \infty)$. We denote the ξ-measure of $[0, t]$ by $\xi(t) = \xi([0, t])$. $\mu(t) = \mathbb{E}[\xi(t)]$ denotes the intensity measure of ξ, which we call the reproduction function.

Suppose we label the individuals as in the section above $x = (i_1, \dots, i_k) \in \mathscr{I}$. The basic probability space is

$$(\Omega, \mathcal{B}, P) = \prod_{x \in \mathscr{I}} (\Omega_x, \mathcal{B}_x, P_x)$$

with $(\Omega_x, \mathcal{B}_x, P_x)$ are identical spaces on which we define ξ_x distributed like ξ.

We define the birth time σ_x of the individuals to be 0 if $x = 0$ (the root) and if $y = xn$,

$$\sigma_y = \sigma_x + \inf\{u; \xi_x(u) \geq n\}.$$

We assume the existence of a product-measurable, separable, non-negative random process $\Phi(t)$, assigning some kind of score to the typical individual at time t. For simplicity we define $\Phi(t) = 0$ for all $t < 0$. This is often called a "random characteristic" of said individual. We then proceed to define

$$Z_t^\Phi = \sum_{x \in \mathscr{I}} \Phi_x(t - \sigma_x)$$

and say that $\{Z_t^\Phi\}$ is a general branching process counted with characteristic Φ. Throughout this paper, we only consider Φ that is equal to a bounded deterministic function of the rooted tree for $t \geq 0$. For instance, $\Phi(t) = 1_{\{t \geq 0\}}$ gives

$$Z_t^\Phi = \sum_{x \in \mathscr{I}} 1_{\{t \geq \sigma_x\}},$$

and this is the total number of individuals born up to time. We will see more examples of such characteristics in the proof of the main results.

What we consider here is *supercritical*, *Malthusian* processes, meaning the following two conditions hold.

(i) There exists a Malthusian parameter λ^* such that

$$\int_0^\infty e^{-\lambda^* t}\mu(dt) = 1.$$

(ii) The first moment of $e^{-\lambda t}\mu(dt)$ is finite, i.e.

$$\int_0^\infty u e^{-\lambda u}\mu(du) < \infty.$$

The first condition is the Malthusian assumption and the second the supercritical condition. Furthermore, we assume μ does not concentrate on any lattice $\{0, h, 2h, \dots\}$ for $h > 0$. We are ready to present Theorem A from [7], which is a weaker version of Theorem 6.3 of [4].

Proposition 1. *Consider a supercritical, Malthusian branching process with Malthusian parameter λ^*, counted by two random characteristics $\Phi(t)$ and $\Psi(t)$ that are 0 for $t < 0$ and deterministic bounded functions of the progeny tree for $t > 0$. Suppose that there exists a $\underline{\lambda} < \lambda^*$ such that*

$$\int_0^\infty e^{-\underline{\lambda} t}\mu(dt) < \infty.$$

Then we have almost surely

$$\frac{Z_t^\Phi}{Z_t^\Psi} \to \frac{\hat{\Phi}(\lambda^*)}{\hat{\Psi}(\lambda^*)}. \tag{1}$$

2.3. The discrete random tree model

For a preferential attachment function $f\colon \mathbb{N}_+ \to \mathbb{R}_+$, we define the *associated discrete random tree model* for the preferential attachment function f to be a Markov chain Υ_d on state space $\mathcal{G}$ with $\Upsilon_d(0) = \emptyset$. The transition probability is defined to be

$$P(\Upsilon_d(n+1) = G \cup \{xk\}) = \frac{f(\deg(x, G))}{\sum_{y\in G} f(\deg(y, G))},$$

for any $x \in G$ and $k = \deg(x, G)$. Therefore every time a new vertex appears, it is appended to one of the existing vertices, say x, with probability proportional to $f(\deg(x, G))$. We see that the associated discrete random tree model of the preferential attachment function f is equivalent to the PAM model.

2.4. The continuous random tree model

Given a preferential attachment function f, let $X(t)$ be a Markovian pure birth process with $X(0) = 1$ and birth rates

$$P(X(t + dt) = k + 1|X(t) = k) = f(k + 1)dt + o(dt).$$

Let $\rho\colon [0, \infty) \to (0, \infty]$ be the density of the point process corresponding to the pure birth process $X(t)$ and $\tilde{\rho}\colon [0, \infty) \to (0, \infty]$ the Laplace transform of ρ

$$\tilde{\rho}(\lambda) = \int_0^\infty e^{-\lambda t}\rho(t)dt = \sum_{n=1}^{\infty}\prod_{k=1}^{n}\frac{f(k)}{\lambda + f(k)}. \tag{2}$$

We define the random tree model $\Upsilon(t)$, which is a continuous time, time homogeneous Markov chain on the countable state space $\mathcal{G}$, with initial state $\Upsilon(0) = \{\emptyset\}$. The jump rate are the following: if for a $t \geq 0$ we have $\Upsilon(t) = G$, then the process may jump to $G \cup \{xk\}$ with rate $f(\deg(x, G))$, where $x \in G$ and $k = \deg(x, G)$. This means that each existing vertex $x \in \Upsilon(t)$ “gives birth to a child” with rate $f(\deg(x, \Upsilon(t))$ independent of others.

We define the *total preference* of a tree $G \in \mathcal{G}$ as

$$F(G) = \sum_{x\in G} f(\deg(x, G)).$$

The Markov chain $\Upsilon(t)$ evolves as follows: assuming $\Upsilon(t-) = G$, at time t a new vertex is added to it with total rate $F(G)$, which is attached with an oriented edge (pointing to the newly added vertex) to the already existing vertex $x \in G$ with probability

$$\frac{f(\deg(x, G)}{\sum_{y\in G} f(\deg(y, G))}.$$

If we define the stopping time

$$T_n = \inf\{t||\Upsilon(t)| = n + 1\},$$

and only look at the model at the stopping times, we get the discrete time model stated in Section 2.3.

3. Consistency of empirical estimator

For completeness of this article, we present a result without proof from [7] giving the limiting degree sequence for a class of preferential attachment function f. From this point on, without causing any confusion we write f_k as shorthand for $f(k)$.

Proposition 2. *Consider a preferential attachment function f that satisfying condition with λ^* being the Malthusian parameter for the associated continuous-time random tree model. Then as $N \to \infty$, the empirical degree distribution*

$P_k(N)$ converges almost surely for any k to some limit p_k specified by the equation below

$$P_k(N) \xrightarrow{a.s.} p_k = \frac{\lambda^*}{\lambda^* + f_k} \prod_{j=1}^{k-1} \frac{f_j}{\lambda^* + f_j}, \quad \forall k \in \mathbb{N}_+. \tag{3}$$

Note $p_1 = \lambda^*/(\lambda^* + f_1)$.

By the definition of the Malthusian parameter of the associated continuous time random tree model, λ^* is the solution to the equation (with respect to λ) Why the Malthusian parameter is defined this way needs more explanation.

$$\sum_{k=1}^{\infty} \prod_{i=1}^{k} \frac{f_k}{\lambda + f_k} = 1. \tag{4}$$

It will be useful later that (by the formula of p_k in (3))

$$\begin{aligned} \sum_{k=1}^{\infty} f_k p_k &= \sum_{k=1}^{\infty} \frac{\lambda^* f_k}{\lambda^* + f_k} \prod_{i=1}^{k-1} \frac{f_i}{\lambda^* + f_i} \\ &= \sum_{k=1}^{\infty} \lambda^* \prod_{i=1}^{k} \frac{f_i}{\lambda^* + f_i} = \lambda^*. \end{aligned} \tag{5}$$

3.1. Construction of the Empirical Estimator

Suppose we already have a preferential attachment graph, that is more or less not far off from the limiting distribution $(p_k)_{k=1}^{\infty}$ with the total number of vertices N. Suppose a new vertex comes in and decides to pick an existing vertex to attach to according to preferential attachment function f. Denoting the number of vertices of degree k is N_k ($\approx Np_k$ in the limiting regime), we can check that the probability of choosing an existing vertex of degree k is

$$\frac{f_k N_k}{\sum_{j=1}^{\infty} f_j N_j} \approx \frac{f_k p_k}{\sum_{j=1}^{\infty} f_j p_j}.$$

We are interested in the quantity f_k for each $k \geq 1$. However we note the denominator on the right hand side of the above display $\sum_{j=1}^{\infty} f_j p_j$ does not depend on k. Therefore it is sufficient to put an extra factor N/N_k on the above display then we obtain $r_k := f_k / \sum_{j=1}^{\infty} f_j p_j$ for each k

$$r_k \approx \frac{\text{Probability of choosing a vertex of degree } k}{N_k/N}. \tag{6}$$

We want to devise an estimator mimicking the above equation, which works also in the non-limiting regime. The probability of new vertex choosing an existing vertex of degree k can be estimated by counting the number of times when the incoming vertex chooses an existing vertex of degree k by looking up the history of the PAM. We denote the said number for the PAM at time N by $N_{\to k}(N)$. We denote the number of vertices of degree k in the PAM at time N by $N_k(N)$. Then we define the empirical estimator (EE) $\hat{r}_k(N)$

$$\hat{r}_k(N) = \frac{N_{\to k}(N)}{N_k(N)}. \tag{7}$$

Theorem 3. *The above constructed empirical estimator $\hat{r}_k(t)$ is consistent almost surely, i.e.*

$$\hat{r}_k(N) \xrightarrow{a.s.} r_k \quad \text{as } N \to \infty. \tag{8}$$

Suppose $N_{>k}(N)$ is the number fo vertices of degree strictly bigger than k at time N. For the PAM considered here, we have the following crucial observation.

Lemma 4. $N_{\to k}(N) = N_{>k}(N)$.

Proof. Indeed, $N_{\to k}(N)$ counts the number of times of incoming vertex choosing an existing vertex of degree k to connect to up to time N. However if a vertex was chosen to be connected to as a vertex of degree k at some point before time N, its degree at time N is at least $k+1$. On the other hand, we notice the vertex degree may only jump from 1 to 2, 2 to 3, …, k to $k+1$, etc., Therefore if a vertex has degree $> k$, it must have been chosen to be connected to as a vertex of degree k at some point. This gives the equality as in the statement of the lemma. ∎

In light of the above observation, we note (7) is equivalent to

$$\hat{r}_k(N) = \frac{N_{>k}(N)}{N_k(N)}. \tag{9}$$

3.2. Proof of Theorem 3

To prove Theorem 3, we need to apply Proposition 1 properly. We find appropriate characteristics Φ and Ψ so that on the left hand side of (1) we have the EE and the right hand side r_k.

We set $\Phi(t) = 1_{\{t\geq 0\}}1_{\{\deg(\emptyset,\Upsilon(t))>k\}}$ and $\Psi(t) = 1_{\{t\geq 0\}}1_{\{\deg(\emptyset,\Upsilon(t))=k\}}$, then

$$Z_t^{\Phi} = \sum_{x\in\mathscr{I}} 1_{\{t\geq\sigma_x\}}1_{\{\deg(x,\Upsilon(t)_{\downarrow x}>k\}},$$
$$Z_t^{\Psi} = \sum_{x\in\mathscr{I}} 1_{\{t\geq\sigma_x\}}1_{\{\deg(x,\Upsilon(t)_{\downarrow x}=k\}}.$$

Z_t^{Ψ} counts all the vertices who have born and have degree k up to time t and Z_t^{Φ} counts all the vertices who have been born and have degree strictly bigger than k at time t.

We apply Proposition 1 with the above defined Φ and Ψ and it gives us

$$\lim_{t\to\infty} \frac{|\{x\in\Upsilon(t)\colon \deg(x,\Upsilon(t))>k\}|}{|\{x\in\Upsilon(t)\colon \deg(x,\Upsilon(t))=k\}|} = \frac{\lambda^* \int_0^\infty e^{-\lambda^* t}\mathbb{P}(\deg(\emptyset,\Upsilon(t))>k)dt}{\lambda^* \int_0^\infty e^{-\lambda^* t}\mathbb{P}(\deg(\emptyset,\Upsilon(t))=k)dt}$$

holds almost surely. The denominator on the right hand side of the preceding display is identified as the (almost-surely) limit of degree distribution at k. Calculation here. By Fubini's theorem, the nominator is the sum of degree distribution from $k+1$. Therefore for $p_{>k} = \sum_{i=k+1}^\infty p_i$

$$\lim_{t\to\infty} \frac{|\{x\in\Upsilon(t)\colon \deg(x,\Upsilon(t))>k\}|}{|\{x\in\Upsilon(t)\colon \deg(x,\Upsilon(t))=k\}|} = \frac{p_{>k}}{p_k}$$

holds almost surely. It suffices to show that the right hand side of the preceding display is the same with $f_k/\sum_j f_j p_j$. We also define q_k as follows

$$q_k = \frac{f_k p_k}{\sum_{i=1}^\infty f_i p_i} = \frac{f_k p_k}{\lambda^*}.$$

to be the limiting preference towards degree k.

Lemma 5. *Suppose $(p_k)_{k=1}^\infty$ is the limiting degree distribution specified in (3) for the preferential attachment function f. The following equality holds*

$$\frac{f_k}{\sum_{j=1}^\infty p_j f_j} = \frac{p_{>k}}{p_k} \tag{10}$$

for $k\in\mathbb{N}_+$. The above display is the same with $q_k = p_{>k}$ for $k\in\mathbb{N}_+$.

Proof. For $k=1$ we note $p_1 = \lambda^*/(f_1+\lambda^*)$, $p_{>1} = 1-p_1 = f_1/(\lambda^*+f_1) = f_1p_1/\lambda^*$. Assuming $p_{>k}=q_k$ holds, consider the case of $k+1$. By (3), we have $p_{k+1} = f_k p_k/(\lambda^*+f_{k+1})$.

$$\begin{aligned} p_{>k+1} &= p_{>k} - p_{k+1} = q_k - p_{k+1} \\ &= f_k p_k \left(\frac{1}{\lambda^*} - \frac{f_{k+1}}{\lambda^*+f_{k+1}}\right) \\ &= \frac{f_{k+1}}{\lambda^*}\frac{f_k p_k}{\lambda^*+f_{k+1}} \\ &= \frac{f_{k+1}p_{k+1}}{\lambda^*} = q_{k+1}. \end{aligned}$$

Then by mathematical induction, $p_{>k}=q_k$ holds for all $k\in\mathbb{N}_+$. ∎

4. Numerical Studies

In this section we present the numerical illustration of the behavior of the empirical estimator.

We run the experiment for the following preferential attachment functions (after normalization such that $f(1) = 1$)

$$f^{(1)}(k) = (k+1/2)/(3/2), \qquad f^{(2)}(k) = \sqrt{k}, \qquad f^{(3)}(k) = \sqrt[4]{k+2}/\sqrt[4]{3}.$$

for $1,000$ times individually of a graph of $10,000$, $100,000$ and $1,000,000$ vertices respectively and calculate the sample mean, sample variance, sample median, sample first quartile and sample third quartile for all the values of at $k = 2,3,5,7,11$ and possibly $k = 23$ for realizations with for all $f^{(i)}$'s ($i = 1,2,3$) and plot them separately.

For example, we now plot the simulation study of f_2.

If we look at all these plots, we conclude the following observations.

- The estimator is consistent, as our theorem shows.
- The deterioration of the quality of the estimator rises fast over large k.
- The EE does not naturally imply the monotonicity, however we can slight modify the estimator so that it is still consistent but always gives monotone results.

Asymptotic normal distribution for each k?

A natural question to ask is for a particular k, what would be the (asymptotic) distribution of $\hat{r}_k$? It is natural to guess it is asymptotically normal and we study the simulation results here. We will study mainly $\hat{r}_2$ and $\hat{r}_3$, $\hat{r}_5$ of all three preferential attachment functions used before. The main point of this subsection is to show the asymptotic normality by simulation. The reason why this holds remains unclear.

Acknowledgement

This work is partially inspired by my master's project ([3]) with the supervision of Rui Castro and Remco van der Hofstad, whose great support then is graciously appreciated. In the master's project, the estimation problem of a preferential attachment function arose in the modeling of the movie-actor network but remained unaddressed. This paper aims to provide an (at least partial) answer to the question.

References

[1] A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

[2] B. Bollobás, O. Riordan, J. Spencer, G. Tusnády, et al. The degree sequence of a scale-free random graph process. *Random Structures & Algorithms*, 18(3):279–290, 2001.

[3] F. Gao. Modeling and interference of the internet movie database. Master's thesis, Eindhoven University of Technology, 9 2011.

[4] P. Jagers et al. *Branching processes with biological applications*. Wiley, 1975.
[5] T. Móri. On random trees. *Studia Scientiarum Mathematicarum Hungarica*, 39(1):143–155, 2002.
[6] A. Rudas, B. Tóth, and B. Valkó. Random trees and general branching processes. *Random Structures & Algorithms*, 31(2):186–202, 2007.
[7] R. van der Hofstad. *Random Graphs and Complex Networks*, volume I. 2014.