Topologica

Topologica
Vol. 3, 2014

Maximum Likelihood Estimation of Preferential Attachment in Growing Networks

Taku Onodera and Paul Sheridan

Human Genome Center, Institute of Medical Science, University of Tokyo 4-6-1 Shirokanedai, Minato-ku, Tokyo 108-8639, Japan
** E-mail: tk-ono@hgc.jp, sheridan@ims.u-tokyo.ac.jp*

Abstract
Preferential attachment is commonly invoked to explain the emergence of power law degree distributions in growing networks. The basic idea is that the preferential attachment is one of many ways to generate a scale-free network, i.e., a network enjoying a power law degree distribution, so that if preferential attachment is observed to occur during the growth of the network, then the scale-free property is explained by means of an appeal to this principle. Two related methods for estimating preferential attachment were advanced in the literature over a decade ago and they have been made use of down to the present day. These methods, while admittedly helpful, lack a sounds statistical foundation, sometimes making the estimates they produce confusing and open to misinterpretation. In this short work, we propose a maximum likelihood based fixes for the classical methods and demonstrate how they are better able to estimate preferential attachment in simulated examples.

Keywords: growing network, maximum likelihood, preferential attachment, power law, scale-free network

1. Introduction

Price's model [1] is a simple mathematical model of citation networks. In a modest generalization of the model, a directed network G with nodes $v_1, v_2, \ldots, v_n$ is generated over a series of time-steps via the successive addition of new nodes to the single starting node v_1. At time-step $0 \leq t < n$, the network begets the new node v_{t+1}, which is connected to exactly m of the nodes already present in the system, unless $m > t$ in which instance v_{t+1} connects to all preceding nodes. The probability that an edge from the node v_{t+1} connects to node v_i $(i = 1, 2, \ldots, t)$ is governed by the attachment rule

$$\pi\left(k_{i,t}\right) = \frac{k_{i,t}^{\alpha} + c}{\sum_{j=1}^{t}\left(k_{j,t}^{\alpha} + c\right)} = \frac{1}{\delta_t} \times \left(k_{i,t}^{\alpha} + c\right), \tag{1}$$

where $k_{i,t}$ is the in-degree of the node v_i at the onset of time-step t, $0 < c$ is a small constant, δ_t is a time-dependent normalising constant, and the parameter $\alpha \geq 0$ is known as the attachment exponent. It will be noted that α was taken to be unity the original formulation of the model, but this was later generalized to be any positive value [2]. The constant c may be considered as an "initial attractiveness" and is necessary to give in-degree zero nodes a fighting chance of acquiring new edges [3]. In the present work, we assume that c is small, i.e., not far from unity.

The in-degree distribution of a Price's model network generated under linear preferential attachment ($\alpha = 1$) on average follows a power-law with $\gamma = 2 + c/m$ [2].

Received
[DOI: 10.3731/topologica.3.xxx]

Topologica Vol. 3, 2014

In the sub-linear case ($\alpha < 1$), the degree distribution is the product of a power-law and a stretched exponential function. The absence of preferential attachment is attained in the limit $\alpha = 0$, when the attachment rule is independent of degree. Meanwhile in the super-linear case ($\alpha > 1$), there tends to emerge a handful of nodes to which almost all other nodes connect. Krapivsky et al. [3] provide demonstrations of these latter results.

In this short work, we address the problem of how to estimate α from a given G whose growth history $G_1, G_2, \ldots, G_n = G$ is at least partially known. Two methods for estimating the extent to which preferential attachment occurs during the growth of a network have been advanced in the literature: the snapshot method [4] and the filmstrip method [5].

The basic idea behind the snapshot method is to restrict focus on two non-overlapping stretches of nodes in time, or two "snapshots" of a growing network, denoted by $\mathcal{T}_{old}$ and $\mathcal{T}_{new}$. A tally of the degrees of the nodes in $\mathcal{T}_{old}$ acquiring edges from the $\mathcal{T}_{new}$ nodes is made, recording the degree of a node once per connection. From there an estimate for the value of α may be obtained by constructing a histogram of the degrees and working out the slope of the line-of-best-fit on a log-log scale. The problem of estimating δ_t in the attachment rule is sidestepped by resorting to the approximation that the degree of each node in $\mathcal{T}_{old}$ is fixed at a constant value. As a result, δ_t itself is rendered a constant and its contribution to the histogram can be ignored. However, to mitigate the effects of any bias incurred by this approximation, it is important that $\mathcal{T}_{new}$ is chosen to be small relative to the size of the growing network.

The filmstrip method is so-named to call attention to its relationship with the previous technique. Instead of relying on snapshots to estimate the attachment exponent in a growing network, however, the filmstrip method constructs a histogram by considering each time-step in sequence. More specifically, the probability that an edge added at time-step t connects to a node of in-degree k is given by $\pi(k) = (k^\alpha + c) \times n_k(t)/\delta_t$, where $n_k(t)$ is the number of nodes in the network with in-degree k at the onset of said time-step. In this approach, inference is achieved by making a histogram, as a function of k, of the $k^\alpha + c \propto \pi(k) \times t/n_k(t)$ values for each newly added edge. So when an edge is observed to connect to a node of in-degree k at time-step t, then $\pi(k) = 1$ and a contribution of $t/n_k(t)$ is added to the histogram at position k; otherwise $\pi(k) = 0$ and no contribution is sustained. As before, the estimated value of α is taken to be the slope of the line-of-best-fit on a log-log scale.

Based on the above descriptions, it may seem that estimating α using either of these methods is a straightforward task, but this is untrue. The main difficulty, although not the only one, lies in determining the line-of-best-fit. As we will see, this determination depends on the specification of a certain threshold value to which these methods are particularly sensitive. On the other hand, our maximum likelihood analogs for these methods require no such threshold to be specified. What is more, the maximum likelihood approach enable one to take the uncertainty of an estimate into account, i.e., confidence bounds, where as the classical methods are designed to produce point estimates.

2. Methodology

In this section, we describe a procedure in the spirit of the Filmstrip method for estimating the most likely α for a given growing network G. Note that we have a similar argument for the Snapshot methods, but we omit it owing to space considerations. It will also be noted that our approach ignores the effect of c on the ground that it is assumed to be a small constant. This is done primarily for ease of exposition of our method, but it could be incorporated into a future version of the method with little trouble, if desired. Returning now to the main argument, we first show that the likelihood function of α is concave. Then, we show how to

Topologica Vol. 3, 2014

calculate the differential of log-likelihood function in time linear to the size of the graph. Once these are done, we can find the most likely α by trivial binary search in $O(|G| \log(1/\epsilon))$ time where $|G|$ is the size of the graph and ϵ is the bound of the error.

Notation: As in the introduction, let G be a network with n vertices generated by Price's model for a given α. The process that generated G can be seen as a series of events $\{E_t\}_{1 \le t \le n}$ where E_t is the event that happened at time-step t, the addition of the t-th node and edges attached to it. We call the t-th node as v_t and the graph just after adding v_t as G_t. For a node v, let $N_v(t)$ and $d_v(t)$ denote v's neighbors and v's in-degree at time step t respectively. Also, let N_v and d_v denote $N_v(n)$ and $d_v(n)$ respectively.

Concavity of the Likelihood Function of α: Given a network $G = G_n$, the following term can be seen as the likelihood of α:

$$\mathrm{Li}(\alpha) = \prod_{1 \le t \le n} \Pr[G_t | G_{t-1}].$$

Let $N_{v_t}(t)$ be the set of vertices that receives edges at time step t. Conditioned on G_{t-1}, the probability that one of these nodes, say v_j, receives an edge is $d_{v_j}(t-1)^\alpha / \sum_{1 \le i < t} d_{v_i}(t-1)^\alpha$. The product of this terms for all j such that $v_j \in N_{v_t}(t)$ is the probability that nodes in $N_{v_t}(t)$ receive edges as actually happened, but with a certain order. Therefore, the following holds:

$$\Pr[G_t | G_{t-1}] \approx d_{v_t}(t)! \prod_{j : v_j \in N_{v_t}(t)} \frac{d_{v_j}(t-1)^\alpha}{\sum_{1 \le i < t} d_{v_i}(t-1)^\alpha}.$$

Here $d_{v_t}(t)$ denotes the number of incoming edges from node v_t. Note that c is may be dropped since it is assumed to be a small constant. Because the concavity of f and $\log \circ f$ are equivalent and concavity is closed under addition, to show $\mathrm{Li}(\alpha)$ is concave to α, it suffices to prove $\log \Pr[G_t | G_{t-1}]$ is concave to α. Moreover, because

$$\log \Pr[G_t | G_{t-1}] \approx \log d_{v_t}(t)! + \sum_{j : v_j \in N_{v_t}(t)} \left(\log d_{v_j}(t-1)^\alpha - \log \sum_{1 \le i < t} d_{v_i}(t-1)^\alpha \right),$$

it, in turn, reduces to the concavity of

$$\log d_{v_j}(t-1)^\alpha - \log \sum_{1 \le i < t} d_{v_i}(t-1)^\alpha \tag{2}$$

for any j s.t. $v_j \in N_{v_t}(t)$. By differentiating (2), we obtain

$$\log d_{v_j}(t-1) - \frac{\sum_{1 \le i < t} d_{v_i}(t-1)^\alpha \log d_{v_i}(t-1)}{\sum_{1 \le i < t} d_{v_i}(t-1)^\alpha}. \tag{3}$$

Differentiating (3), in turn, gives

$$\frac{\left(\sum d_{v_i}(t-1)^\alpha \log d_{v_i}(t-1)\right)^2 - \sum d_{v_i}(t-1)^\alpha \log^2 d_{v_i}(t-1) \sum d_{v_i}(t-1)^\alpha}{\left(\sum d_{v_i}(t-1)^\alpha\right)^2}$$

where all sumations are taken over $1 \le i < t$. By replacing $d_{v_i}(t-1)^\alpha$ by $X_i (> 0)$ and $\log d_{v_i}(t-1)$ by a_i, one can easily see that the numerator is non-positive, which

Topologica Vol. 3, 2014

means the desired concavity, as follows:

$$\begin{aligned}(\sum_i a_i X_i)^2 - \sum_i a_i^2 X_i \sum_j X_j &= \sum_{i\neq j}(2a_i a_j - a_i^2 - a_j^2)X_i X_j \\ &= -\sum_{i\neq j}(a_i - a_j)^2 X_i X_j \\ &\leq 0.\end{aligned}$$

Algorithm 1 Calculation of $\frac{d}{d\alpha}\log \mathrm{Li}(\alpha)$

Input: A graph with n vertices $\{v_1, v_2, \dots v_n\}$ and an order on its vertices. We assume that vertices are ordered as $v_1 < v_2 < \dots v_n$ w.l.o.g.

Output: $fracdd\alpha \log \mathrm{Li}(\alpha)$

$\mathsf{deg}[1,n] \leftarrow [1,1,0,0,\dots,0]$; $\mathsf{dLL} \leftarrow 0$; $\mathsf{sum1} \leftarrow 2$; $\mathsf{sum2} \leftarrow 0$; $\mathsf{tmp1} \leftarrow 0$; $\mathsf{tmp2} \leftarrow 0$

for $t = 3$ to n **do**

  for $j \in \{j : v_i \in N_{v_t}(t)\}$ **do**

    $\mathsf{dLL} \leftarrow \mathsf{dLL} + \log \mathsf{deg}[j] - \mathsf{sum2}/\mathsf{sum1}$

    $\mathsf{deg}[j] \leftarrow \mathsf{deg}[j] + 1$

    $\mathsf{tmp1} \leftarrow \mathsf{tmp1} + \mathsf{deg}[j]^\alpha - (\mathsf{deg}[j]-1)^\alpha$

    $\mathsf{tmp2} \leftarrow \mathsf{tmp2} + \mathsf{deg}[j]^\alpha \log \mathsf{deg}[j] - (\mathsf{deg}[j]-1)^\alpha \log(\mathsf{deg}[j]-1)$

    $\mathsf{deg}[t] \leftarrow \mathsf{deg}[t] + 1$

  end for

  $\mathsf{sum1} \leftarrow \mathsf{sum1} + \mathsf{tmp1} + \mathsf{deg}[t]^\alpha$; $\mathsf{tmp1} \leftarrow 0$

  $\mathsf{sum2} \leftarrow \mathsf{sum2} + \mathsf{tmp2} + \mathsf{deg}[t]^\alpha \log \mathsf{deg}[t]$; $\mathsf{tmp2} \leftarrow 0$

end for

return dLL

Calculation of the Differential of Log-Likelihood Function: The variable t represents the time-step and the algorithm proceeds by time step-wise. After time step t is completed, dLL holds $\sum_{1\leq s\leq t}\frac{d}{d\alpha}\log \Pr(G_s|G_{s-1})$, thus, after the outer for loop is completed, dLL holds $\frac{d}{d\alpha}\log \mathrm{Li}(\alpha)$. Obviously, the initialization of variables finishes in $O(n)$ time. Assuming basic arithmetic operations such as $+, -, *, /$ and somewhat more involved operations such as exponentiation and logarithm can be calculated in constant time, each execution of inner for loop finishes in constant time. Therefore, all but the enumeration of $\{j : v_i \in N_{v_t}(t)\}$, can be done within $O(n)$ time. To enumerate $\{j : v_i \in N_{v_t}(t)\}$ efficiently, we take the graph as an adjacency list. The total running time is bounded by $O(n+m)$ where m is the number of edges.

Remarks: It will be noted that the procedure described here essentially recapitulates the growth of a network, taking into account the contribution to the likelihood of α at each time-step. In this respect it is analogous to the Filmstrip method. There is one way, however, in which the proposed methodology is an clear improvement over the Filmstrip method. It is this: the Filmstrip method assumes that the normalising constant $\delta(t)$ is proportional to t, while our approach uses the actual observed value of $\delta(t)$. But it must be said that we are not the first to implement this improvement. That credit goes to Massen and Doye [6].

Topologica Vol. 3, 2014

3. Simulated Examples

For brevity the details of the simulation design and results are described in the table and figure captions. The main implication to be drawn from our simulated examples is that the maximum likelihood approach to estimating α is to be preferred over the classical methods because they give unambiguous estimates with confidence bounds.

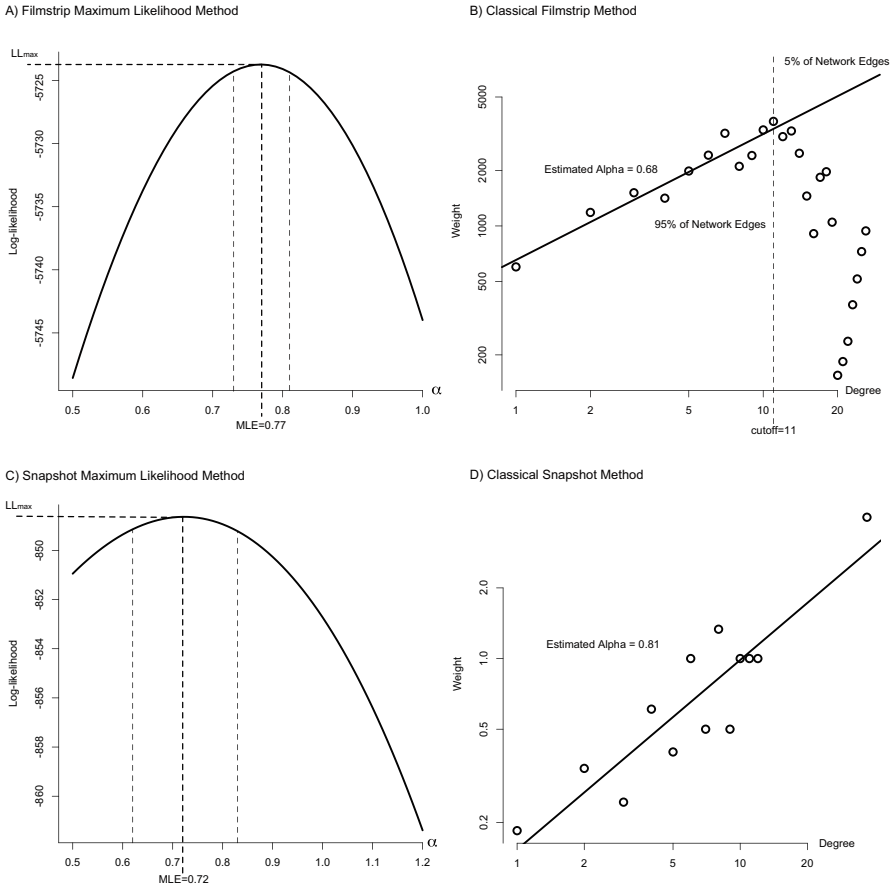

Fig. 1 **Estimation of α for a simulated Price model network.** A typical Price model network with $n = 1000$, $m = 1$, $c = 1$, and $\alpha = 0.75$ was generated. The value of alpha was estimated using A) the Filmstrip maximum likelihood method, and B) the classical Filmstrip method, C) the Snapshot maximum likelihood method, and D) the classical Snapshot method. In A) is plotted the log-likelihood, showing that the maximum LL_{max} corresponds to the MLE 0.77, and using the rule of thumb that the true value of α lies within $LL_{max} - 1/2$ with high probability, we have $\alpha \in (0.73, 0.81)$. In B) is plotted the weights obtained from the Filmstrip method algorithm. The cutoff of $k_{cut} = 11$ was selected so that the waterfall tail is omitted from the estimation and was selected by eye. The value of $\widehat{\alpha}$ is 0.68. Note that 95% of the edges were used to produce the estimate. In C) is plotted the log-likelihood, showing that the maximum LL_{max} corresponds to the MLE 0.72, and using the rule of thumb that the true value of α lies within $LL_{max} - 1/2$ with high probability, we have $\alpha \in (0.62, 0.83)$. In D) is plotted the weights obtained from the Snapshot method algorithm with estimated $\alpha = 0.81$.

Topologica Vol. 3, 2014

	Snapshot ML	Snapshot	Filmstrip ML	Filmstrip	
True α	$\widehat{\alpha} \pm s.d.$	$\widehat{\alpha} \pm s.d.$	$\widehat{\alpha} \pm s.d.$	$\widehat{\alpha} \pm s.d.$	$k_{cut} \pm s.d.$
0.00	0.27 ± 0.07	0.40 ± 0.30	0.02 ± 0.02	-0.05 ± 0.11	7.16 ± 0.61
0.50	0.46 ± 0.16	0.76 ± 0.24	0.50 ± 0.04	0.46 ± 0.08	7.41 ± 0.55
0.75	0.70 ± 0.13	0.93 ± 0.15	0.75 ± 0.04	0.67 ± 0.07	8.69 ± 0.76
1.00	1.00 ± 0.10	1.03 ± 0.12	1.00 ± 0.03	0.93 ± 0.06	7.23 ± 0.72
1.15	1.19 ± 0.06	1.20 ± 0.09	1.15 ± 0.02	1.08 ± 0.08	6.25 ± 0.94
1.25	1.37 ± 0.06	1.37 ± 0.06	1.25 ± 0.02	1.13 ± 0.10	7.38 ± 0.86
1.50	1.67 ± 0.08	n/a	1.49 ± 0.01	1.48 ± 0.34	1.93 ± 0.77

Table 1 **Classical Methods vs. Maximum Likelihood Estimation Methods** A hundred Price model networks were generated with $n = 1000$, $m = 1$, and $c = 1$ for each true value of α. The average estimated value of α is reported with standard deviation as obtained with maximum likelihood and the Filmstrip method. In the case of the Filmstrip method, the cutoff k_{cut} is also reported.

4. Discussion

In this short work we have proposed improvements on the classical methods for estimating α in growing networks. While the classical methods are somewhat ad hoc, our proposed fixes are based on the maximum likelihood principle, which rests on a sound statistical foundation. We demonstrated in simulated examples how our methods recover α with improved accuracy in smallish networks. What is more our approach makes giving confidence bounds on the estimated value of α straightforward. It is important to note that Price's model does not consider the edge additions, removals, and rewirings that are common in many non-citation growing networks, such as social networks. We therefore emphasize that the maximum likelihood approach presented here is best suited to citation networks. Extending this methodology to more complicated growing networks is left as a future work.

References

[1] D.J. de S. Price, Science **149**, 510–514 (1965).

[2] P.L. Krapivsky, S. Redner, and F. Leyvraz, Phys. Rev. Lett. **85**, 4629–4632 (2000). arXiv:cond-mat/0005139

[3] S.N. Dorogovtsev, J.F.F. Mendes, and A.N. Samukhin, Phys. Rev. Lett. **85**, 4633–4636 (2000). arXiv:cond-mat/0004434

[4] H. Jeong, Z. Néda, and A-L Barabási, Europhys. Lett. **61**, 567–572 (2003).

[5] M.E.J. Newman, Phys. Rev. E **64**, 025102 (2001).

[6] C.P. Massen, and J.P.K. Doye Physica A **377**, 351–362 (2007).