The Resolution of a Minor Preferential Attachment Paradox makes Major Confusions Plain

Taku Onodera and Paul Sheridan

Human Genome Center, Institute of Medical Science, University of Tokyo
Email: tk-ono@hgc.jp, sheridan@ims.u-tokyo.ac.jp

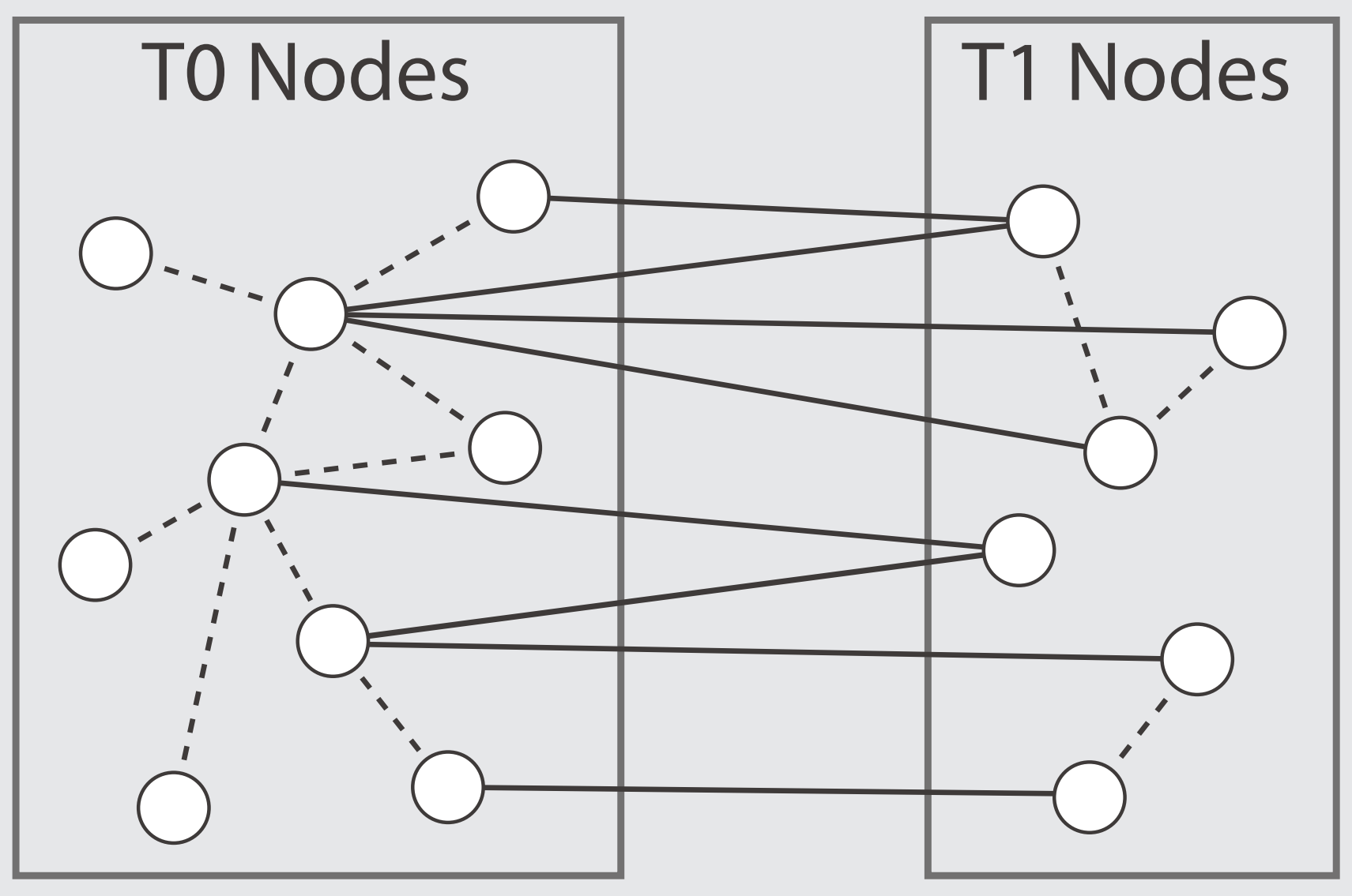
What is Preferential Attachment?

A simple definition of preferential attachment (**PA**) is furnished by **Price's model** [1]. In this model, an *n*-node directed network is generated over a series of time-steps via the successive addition of new nodes to a single starting node. At time-step $0 \le t < n$, node v_{t+1} is connected to *m* of the preceeding nodes such that the probability of v_{t+1} being connected to $v_i (i=1..t)$ is given by the **attachment function**

$$\pi(k_{i,t}) \propto k_{i,t}^{\theta} + a$$

where $k_{i,t}$ is the in-degree of v_i at time-step *t*, $a>0$ is a constant, and $\theta>0$ is the **attachment rate.** The in-degree distribution of a Price's model network generated under linear PA ($\theta=1$) follows a power-law with exponent $\gamma = 2 + a/m$. In the sub-linear case ($\theta<1$), the in-degree distribution is more light tailed. Indeed, the limiting case when $\theta=0$ is one way to define a random network. Lastly, in the super-linear case ($\theta>1$) there tends to emerge a handful of nodes to which almost all other nodes connect.

How is Preferential Attachment Estimated?



The **snapshot method** [2] is the most common procedure used to estimate the attachment rate in a given network. The idea behind the method is to focus on two non-overlapping stretches of nodes in time, i.e., snapshots, of the network, denoted by T0 and T1. To estimate the attachment rate, θ, first the in-degree of each T0 node within the confines of the T0 window is recorded (dashed edges in T0 window). Second, a tally of the in-degrees of the T0 nodes acquiring edges from the T1 nodes (solid edges from T1 to T0) is made, recording the in-degree of the T0 node once per connection. Finally, an estimate for θ is obtained by plotting the in-degree of the T0 nodes versus the counts and working out the line-of-best-fit on a log-log scale.

The Paradox

The phycisist Sidney Redner, in his most excellent analysis of the citation network of the Physical Review family of journals, articulated a certain paradox pertaining to preferential attachment [3].

Firstly, Professor Redner observed that the distribution of citations, i.e., the in-degree distribution, for all articles published from 1893 until June 2003 is better fit by a log-normal distribution, than by a power-law (see Figure 2, which is taken from his original manuscript).

Secondly, a log-normal distribution is known to arise from the nonlinear attachment function $k/(1 + c \log k)$ for a constant $c>0$.

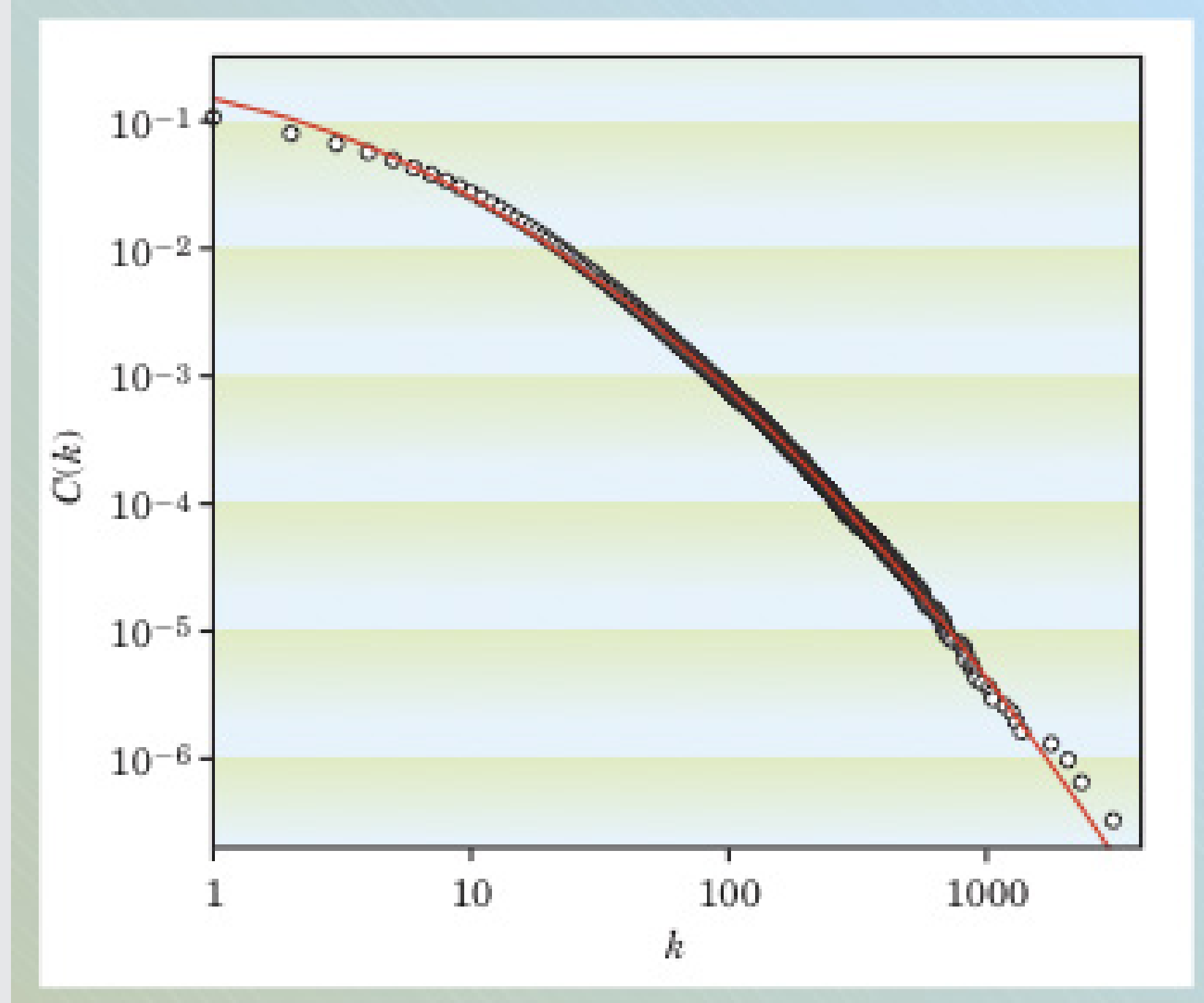


Figure 2. The cumulative citation distribution $C(k)$ versus the number of citations k for all papers published from July 1893 through June 2003 in the *Physical Review* journals. Circles indicate the data. The curve is the log-normal fit $C(k) = A \exp[-b \ln k - c(\ln k)^2]$, with $A = 0.15$, $b = 0.40$, and $c = 0.16$.

Thirdly, he found by using the snapshot method that the attachment function was linear in k with different attachment rates, θ, depending on the choice of time window T0 (see Redner's Figure 3).

The paradox is this: How can it be that the citation distribution is log-normal (Redner's Figure 2) and yet the various attachment functions (Redner's Figure 3) are linear?

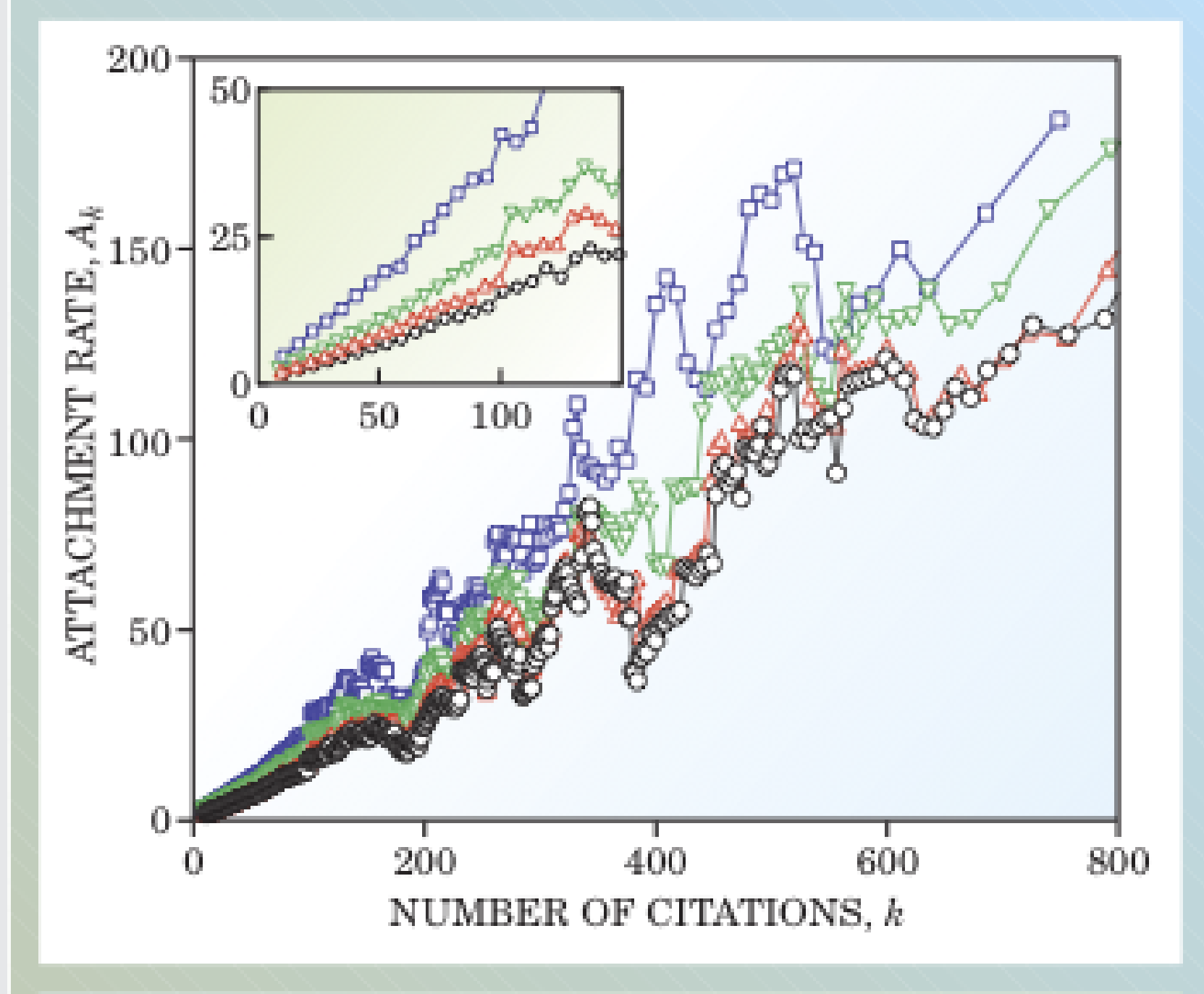


Figure 3. The attachment rate A_k is a nearly linear function of the number of citations k, especially for k less than 150 (inset). The different colors indicate different year ranges for establishing k: 1990–99 (purple), 1980–99 (green), 1970–99 (red), and 1893–1999 (black). The rate A_k is determined from citations in the year 2000. Data have been averaged over a range of ±2.5%. Other time ranges for existing and new citations give similar behavior.

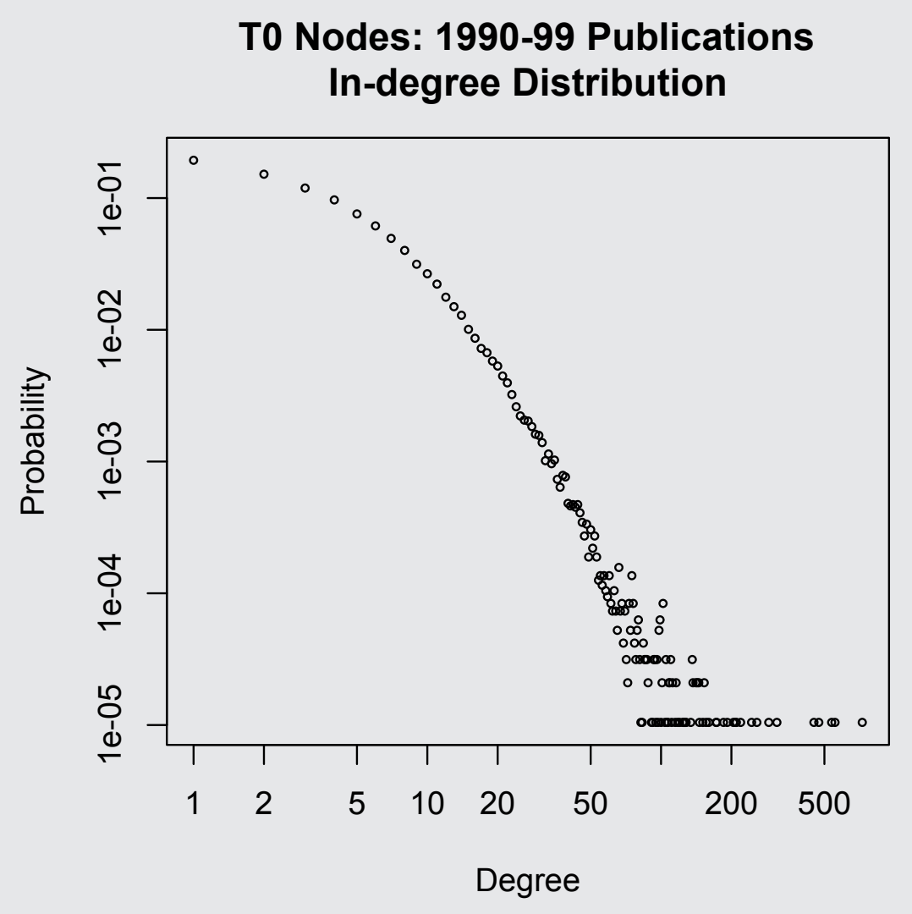
Resolution of the Paradox

The resolution of the paradox is lies in observing that the attachment rate, θ, as estimated by the snapshot method is functionally independent of the in-degree distribution of the T0 nodes, as formed from the dashed edges in the T0 window as depicted in the above figure. Rather, the estimate depends solely on the in-degree distribution of the T0 nodes as contributed by the T1 nodes, i.e., taking into account only contributions by the solid lines in the figure. While there are sound logical grounds for thinking this is true, for the purposes of this poster, empirical evidence in the form of simulated examples will need to suffice; see the adjacent table.

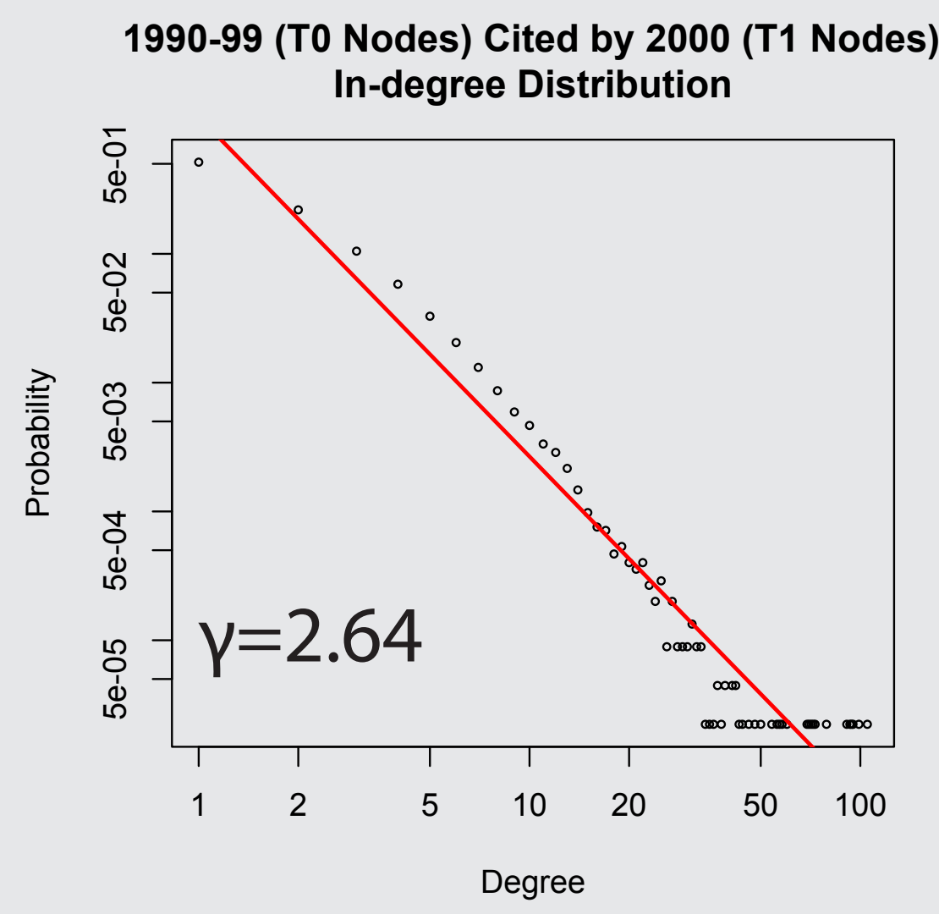
In-degree Distribution		Estimated Attachment Rate θ	
T0 Nodes	T1 to T0 Nodes	Snapshot	Snapshot ML
binomial: θ=0	power-law: θ=1	θ=0.94±0.10	θ=1.02±0.04
lognormal: c=2	power-law: θ=1	θ=0.98±0.08	θ=1.03±0.03
power-law: θ=1	power-law: θ=1	θ=0.94±0.06	θ=0.99±0.02

Table: Price's model simulated examples. We estimated the attachment rate, θ, for a trio of Price's model networks using the snapshot method to demonstrate that the estimated value of θ does not depend on the in-degree distribution of the T0 nodes. In the first row, for example, when the T0 nodes follow a binomial in-degree distribution, and the T1 nodes connect to the T0 nodes according to a power-law, the estimated θ is consistent with a power-law alone. All in all, we can see that the estimated θ does not depend on the in-degree distribution of the T0 nodes. Note that each estimate is actually an average taken from ten different simulated networks.

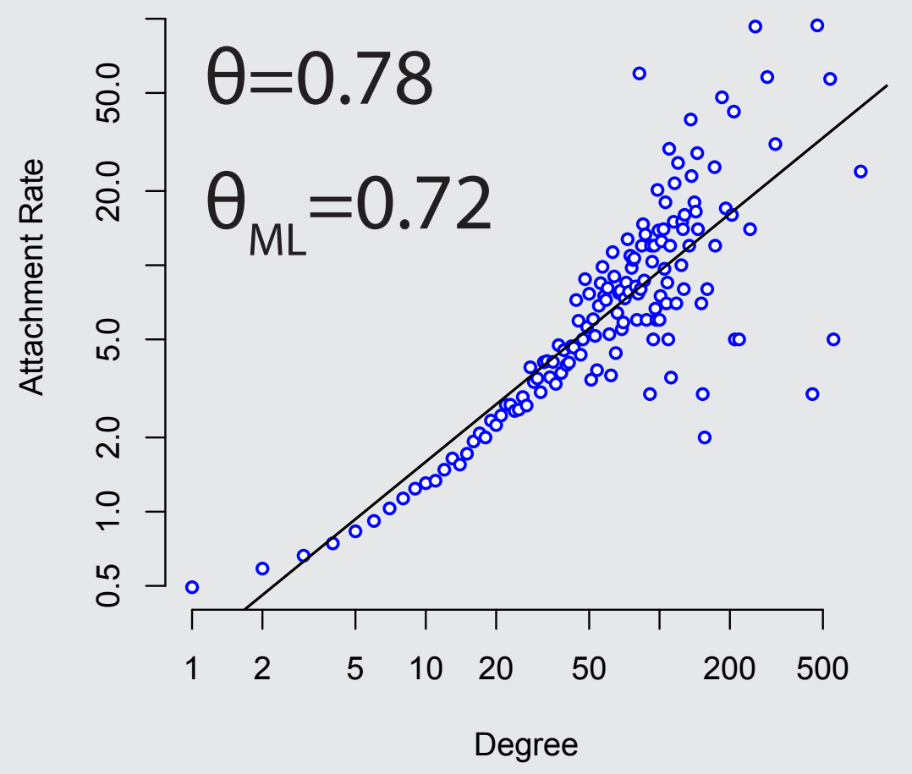
The Physics Review Citation Network with the T0 Nodes Running from 1990-99 and the T1 Nodes taken from year 2000



The T0 nodes citation distribution is in keeping with a log-normal distribution.



On the other hand this citation distribution closely follows a power-law.



As a result of the attachment function as constructed by the snapshot method necessarily follows a straight line.

It is plain that while the T0 papers follow a log-normal citation distribution, the attachment rate is linear, because the relevant citation distribution closely follows a power-law. Thus the paradox is explained.

Conclusions

1) PA and what is estimated by the snapshot method are two different things. On the one hand, PA is a property of a sequence of networks $\{G_1, G_2, \ldots, G_n\}$ defined by Price's model that culminates in an observed network $G = G_n$, while, on the other hand, the snapshot method estimate is obtained from knowledge of a certain portion of the degree distribution of G alone. Hence, thismethods tells us nothing that is not already apparent by examination of the degree distribution which determines its output.

2) The confusion lying at the heart of this paradox, we think, is a mistaken impression that it is intelligeable to speak of PA in absolute terms. PA, however, is always defined with respect to a model. As such, PA estimation is inherently a matter fitting a network model to an observed network. The filmstrip method [4] is one such approach, and we are presently in studying its working in detail.

References

[1] Networks of Scientific Papers, Science, Vol. 149, No. 3683, pp. 510-515 (1965) by Derek J. de Solla Price
[2] Measuring Preferential Attachment in Evolving Networks, Europhysics Letters, Vol. 64, No. 4, pp. 567-572 (2003) by Hawoong Jeong, Zoltan Néda, and A.-L. Barabási
[3] Citation Statistics from 110 Years of Physical Review, Physics Review, Vol. 58, No. 6, pp. 49-54 (2005) by Sidney Redner
[4] Clustering and Preferential Attachment in Growing Networks, Physical Review E, Vol. 64, No. 2, pp. 025102 (2001) by Mark Newman