SCIENTIFIC REPORTS

OPEN

Received: 8 September 2017
Accepted: 18 January 2018
Published online: 12 February 2018

A Preferential Attachment Paradox: How Preferential Attachment Combines with Growth to Produce Networks with Log-normal In-degree Distributions

Paul Sheridan[1] & Taku Onodera[2]

Every network scientist knows that preferential attachment combines with growth to produce networks with power-law in-degree distributions. How, then, is it possible for the network of American Physical Society journal collection citations to enjoy a log-normal citation distribution when it was found to have grown in accordance with preferential attachment? This anomalous result, which we exalt as the preferential attachment paradox, has remained unexplained since the physicist Sidney Redner first made light of it over a decade ago. Here we propose a resolution. The chief source of the mischief, we contend, lies in Redner having relied on a measurement procedure bereft of the accuracy required to distinguish preferential attachment from another form of attachment that is consistent with a log-normal in-degree distribution. There was a high-accuracy measurement procedure in use at the time, but it would have have been difficult to use it to shed light on the paradox, due to the presence of a systematic error inducing design flaw. In recent years the design flaw had been recognised and corrected. We show that the bringing of the newly corrected measurement procedure to bear on the data leads to a resolution of the paradox.

The physicist Sidney Redner reported a rather curious anomaly in a decade-old study on the citation statistics of the *American Physical Society* (APS) journal collection[1]. Redner effectively discovered that while the APS citation network had grown in accordance with a process commonly know as preferential attachment, the corresponding citation distribution closely follows a log-normal distribution on a double logarithmic scale. The network scientist will recognise preferential attachment as a process whereby the nodes of a network acquire new connections in proportion to the number of connections they already entertain. What makes his observations so puzzling is that growing network models based on preferential attachment have long been known to generate networks with power-law, as opposed to log-normal, in-degree distributions[2–6]. This anomaly, or paradox, as Redner referred to it, may be called for convenience the *preferential attachment paradox*.

In this paper we propose a resolution to the paradox. But we first take pains to reproduce the anomalous findings that Redner reported. In so doing we confirm that the APS citation distribution closely follows a log-normal distribution on a double logarithmic scale, and, moreover, that the associated APS citation network had grown in accordance with preferential attachment. Only then do we venture to resolve the paradox. The resolution we propose requires that two main obstacles be overcome. The first is to recognise that whether preferential attachment is observed in a growing network or not depends on the choice of measurement procedure. This insight will lead us to conclude that the preferential attachment observed by Redner amounts to an artefact of the procedure he used to measure the process over coarse time resolutions[7]. And, what is more, when we perform the measurements at comparatively fine time resolutions, the outcomes are found to be reconcilable with a form of attachment that is consistent with a log-normal citation distribution. The second obstacle is a purely technical matter related to the accurate measurement of preferential attachment at fine time resolutions. The fact that the paradox has

[1]Department of Active Life Promotion Science, Hirosaki University, Hirosaki, 036-8562, Japan. [2]The University of Tokyo, Institute of Medical Science, Human Genome Center, Tokyo, 108-8639, Japan. Correspondence and requests for materials should be addressed to P.S. (email: paul.sheridan.stats@gmail.com)

remained unresolved until now is explained in part by the presence of a design flaw in the standard fine time resolution measurement procedure[8]. The flaw, for which a correction has only recently been suggested[9], has been an obstacle to progress, because it functions to distort the measurements taken using the procedure, much as a crooked ruler distorts the true lengths of the objects that it measures. But once these obstacles to measuring preferential attachment have been realised, we will see how a little detective work is all that stands in the way of a definitive resolution to the paradox.

The Preferential Attachment Paradox

In the previous section we outlined the preferential attachment paradox. Informally it is this:

> Growing network models based on growth and preferential attachment are known to generate networks with power-law in-degree distributions. So how can preferential attachment combine with growth to generate networks with log-normally distributed in-degree distributions? In particular, how did preferential attachment give rise to a log-normally distributed citation distribution in the growth of the APS citation network?

In this section we formulate the paradox in technical terms and illustrate it on the APS journal collection citation network. The illustration is in point of fact a careful reproduction of those anomalous results reported by Redner in the very same paper in which he first called attention to the paradox[1]. For the sake of review they are: 1) the APS citation distribution closely follows a log-normal distribution on a double logarithmic scale, and 2) the associated APS citation network had grown in accordance with preferential attachment. We reproduce these results successively below. But we begin with an overview of the APS journal collection citation data.

The APS Journal Collection Citation Data. The APS ranks among the world's foremost learned societies for physicists. The society publishes a dozen research journals that span virtually all fields of modern physics. Its journal collection citation data from July 1893 through December 2009 is freely available for download upon request at the society website[10].

The dataset is comprised of just over 450,000 timestamped articles and 4,500,000 intra-APS journal citations. But in keeping with Redner, whose analysis we aim to reproduce, we restrict our attention to only those articles from July 1893 up to and including June 2003. According to our tally this 110 year stretch of data covers precisely 347,038 articles and 3,063,726 citations. The mean number of citations is 8.8 which agrees with Redner's reported value. The scrupulous reader may object that Redner reports 353,268 articles and 3,110,839 citations over the same time period. In other words, there is about a 1% shortfall on our part in both instances. Roughly 40% of the missing citations are accounted for by the fact that we filtered 12,425 duplicate citations and 115 self-citations in the course of processing the data. The remaining shortfalls are perhaps attributable to vigorous data cleansing efforts on the part of APS technicians over the years.

Any bibliographic dataset is readily conceptualised as a type of network called a citation network. The nodes of a citation network represent articles in such a manner that a node Y is connected to a node X if the article corresponding to X is cited by the article corresponding to Y in its references. The said connection, if it exists, is conferred with an orientation so as to point like an arrow from the citing article Y to the cited article X. Multiple connections from Y to X (i.e. duplicate citations) and self-loops from Y to Y (i.e. self-citations) are prohibited as a matter of convenience. Thus a citation network, at least in this paper, will be recognised by network aficionados as a simple directed network representation of bibliographic data. And when we speak of the APS citation network without qualification, we mean precisely this kind of representation of the APS citation data from July 1893 through June 2003 as related above.

The APS Citation Network Citation Distribution. The distribution of node degrees in a network is one of the most important network properties, and a defining characteristic of network structure. Many readers will already be familiar with the notion that the *in-degree* k of a node in a directed network is the number of incoming connections it shares with other nodes, and moreover that the *in-degree distribution*, $P(k)$, is an associated function which gives the proportion of nodes in the network with in-degree k. In the context of citation networks, a usual goal of the network scientists is to characterise the distribution of incoming citations. They give the name *citation distribution* to the in-degree distribution $P(k)$ of a citation network, which is at once seen to give the proportion of papers in the network cited k times. Network scientists have long explored fitting citation distributions by a variety of different functional forms; see Radicchi *et al.*[11] for a brief review. Suffice it to say here that the question of which functional form – if any – best characterises citation distributions remains a subject of ongoing research.

Redner appealed to the log-normal to describe the APS citation distribution in his study of citation statistics from the first 110 years of the APS journal collection[1]. Strictly speaking, he found that visual inspection reveals the (*complementary*) *cumulative in-degree distribution* $C(k)=\sum_{i\geq k}P(i)$ of the APS citation distribution $P_{\mathrm{APS}}(k)$ is well-fitted by the so-called log-normal form $\mathcal{L}(k;\,\beta_0,\,\beta_1,\,\beta_2)\,=\,\beta_0\,\exp[-\beta_1\log(k)\,-\,\beta_2\log^2(k)]$ over a substantial range of incoming citations when $\beta_0=0.15$, $\beta_1=0.40$, and $\beta_2=0.16$. In the context of a citation network, we refer to $C(k)$, which gives the proportion of papers cited at least k times, as a *cumulative citation distribution*. Redner concluded on the basis of the above outcome that the form of $P_{\mathrm{APS}}(k)$ is inconsistent with a power-law for reasons described in Supplementary Note 1.

The remainder of this section is devoted to showing that $P_{\mathrm{APS}}(k)$ is better described by a discretisation of the log-normal distribution

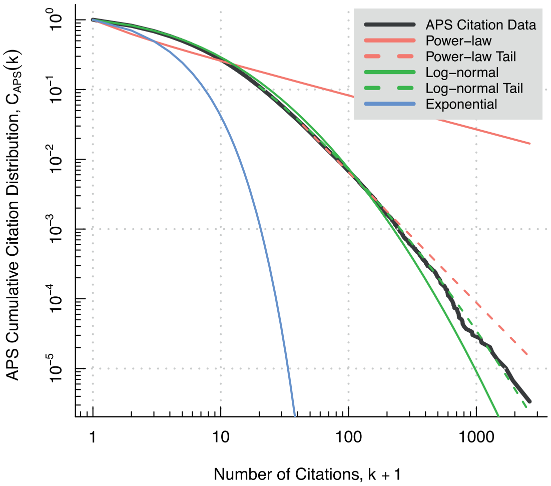

Figure 1. The APS cumulative citation distribution $C(k)$ for all publications dating from July 1893 up to and including June 2003. The log-normal cumulative distribution (green) fits the observed data better than either a power-law cumulative (red) or exponential cumulative distribution (blue). This holds true regardless of whether the data is fit over its full domain (solid lines) or merely in the tail region of k (dashed lines). The cutoff values defining the tail regions (dashed lines) are calculated using the maximum likelihood estimation method of Clauset *et al.*[13] described in the main text.

$$\log\mathcal{N}(k;\,\mu,\,\sigma)\;=\;\frac{1}{(k\,+\,1)\sqrt{2\pi\sigma^{2}}}\;\times\;\exp\left[-\frac{(\log(k\,+\,1)\;-\;\mu)^{2}}{2\sigma^{2}}\right] \tag{1}$$

for $k \geq 0$ with location parameter μ and scale parameter $\sigma > 0$, than by either of a corresponding discretised power-law $(k+1)^{-\gamma}$ or discretised exponential distribution $\lambda e^{-\lambda(k+1)}$ with rate parameter $\lambda > 0$. In addition, we investigate the extent to which the log-normal can be said to plausibly model $P_{\mathrm{APS}}(k)$ in absolute terms. We use $k+1$ instead of k in Eq. (1) on the one hand so as to include papers with zero citations, and on the other because doing so dovetails with the modelling framework that we will develop in a later section. The exponential distribution we include in our analysis as a token light-tailed alternative to the heavy-tailed log-normal and power-law distributions.

We used the poweRlaw R package v.0.60.3[12] to fit our trio of functional forms to the APS citation distribution. The package implements the maximum likelihood estimation methods and goodness-of-fit tests, based on the Kolmogorov-Smirnov (KS) test and likelihood ratios, for fitting heavy-tailed distributions to observed data described in Clauset *et al.*[13]. Their approach may be summed up as follows: Candidate functional forms are separately fitted to an observed in-degree distribution over the domain $k+1 \geq k_{min}$ by maximum likelihood, conditional on the choice of lower cutoff k_{min}. An optimal k_{min} is estimated for each candidate using a goodness-of-fit based testing approach. Alternately, the value of k_{min} may be set to 1 to fit the data over its full domain. The goodness-of-fit of each functional form is assessed using a KS based hypothesis testing approach.

The APS cumulative citation distribution is plotted in Fig. 1 on a double logarithmic scale together with a medley of fitted functional form cumulatives. The associated cumulative distributions are plotted in Fig. 1 solely on aesthetic grounds. That said, a cursory visual inspection will satisfy even the most quantitatively minded reader that the log-normal better describes the entire APS citation distribution ($k_{min} = 1$), than either a power-law or exponential distribution. And just like that our modest claim is proved correct.

This is by no means to say that the log-normal plausibly describes the APS citation distribution. On the contrary, we found that the goodness-of-fit of log-normal to the data ($H_0 =$ log-normal with $\mu = 1.41$, $\sigma = 1.27$, $k_{min} = 1$: KS = 0.01 & P = 0.00) was poor insofar as rigorous statistical testing is concerned. Clauset *et al.*[13] recommend a significance level of 0.10 as a conservative choice for pronouncing a given functional form a plausible fit to the data–for what it is worth. Thus the hypothesis that the log-normal plausibly fits the APS citation distribution over its full domain is to be rejected.

But proponents of the log-normal will be heartened to learn that this unhappy state of affairs is entirely reversed once we confine our attention to either the body or tail region of the distribution. Let us take the APS citation distribution body and tail to correspond to the region $0 \leq k \leq 150$ and $k \geq 20$, respectively. Visual inspection of the Fig. 1 plot alone is enough to conclude that the body is plausibly fit by the log-normal distribution. We visually selected $k \leq 150$ as a conservative choice of cutoff point on account that the poweRlaw R package cannot be used to assess the goodness-of-fit of a log-normal to the body of a distribution. On the other hand, we found that the APS citation distribution tail is plausibly fit by the log-normal distribution at significance level 0.10 ($H_0 =$ log-normal with $\mu = -1.00$, $\sigma = 1.76$, $k_{min} = 20$: KS = 0.00 & P = 0.33). Note the cutoff $k_{min} = 20$ is the minimum k yielding a plausible fit of the log-normal to the data at significance level 0.10. For this reason, we use $k \geq 20$ to define the tail of the distribution. The same, however, cannot be said for a power-law ($H_0 =$ power-law with $\gamma = 2.87$, $k_{min} = 44$: KS = 0.01 & P = 0.01). The plot of Fig. 1 serves to visually reinforce these conclusion.

Let us conclude by taking stock of our findings. The log-normal distribution, we found, provides an incontrovertibly better fit to the APS citation distribution over its full domain, than does a power-law. Moreover, the log-normal looks to fit the APS citation distribution pretty nearly as judged by visual inspection, but this is not supported by rigorous statistical testing. This is because the log-normal undershoots the target in the tail of the distribution. We may nevertheless speak informally of the APS citation distribution as "closely" following a log-normal in certain non-technical contexts. That said, we found the log-normal does provide a good fit (in the technical sense) to the data when confined to either the body ($0 \leq k \leq 150$) or tail ($k \geq 20$) of the distribution. We are careful to be precise about which region of the distribution that we mean in technical contexts. In light of these considerations, the reader will do well to keep the informal and technical senses of the log-normal providing a close fit to the APS citation distribution in mind.

The APS Citation Network Attachment Rate. A growing network represents bibliographic data over time in a manner conducive to the quantification of preferential attachment. Before considering how to represent bibliographic data as a growing network it should be understood that what we mean by bibliographic data is a collection of intra-referencing articles complete with timestamps of the form YYYY-MM-DD. While timestamps proved superfluous to the construction of the APS citation network, to a growing network representation of the APS bibliographic data they are essential. This is because a *growing network* is formally defined as a nested sequence of networks, $\mathcal{G} = \{G_t\}_{t=1}^{T}$, that begins with an *initial network*, G_1, with $n_1 > 0$ nodes and $m_1' \geq 0$ edges and ends with a *final network*, $G_T = G$. Nesting means that the network G_t at *time-step t* for $t > 1$ is obtained by augmenting G_{t-1} with $n_t \geq 0$ nodes that form $m_t \geq 0$ connections with the nodes in G_{t-1} and $m_t' \geq 0$ connections among the nodes newly added (see Figure S1 for a graphical depiction of this modelling scheme). A bibliographic dataset is represented as a growing network by specifying a mapping from article timestamps to sequence time-steps that preserves chronological order up to a desired level of time-resolution. Articles are mapped to nodes and references to directed edges within this framework in the obvious fashion.

A few examples of growing network representations will serve to make their workings more comprehensible. Table S1 summarises the examples here described. The APS citation network is a growing network in the trivial sense that all article timestamps from 1893-07-01 to 2003-06-30 are mapped to a single time-step. The network in this case consists of $n_1 = 347{,}083$ nodes and $m_1' = 3{,}063{,}726$ edges. It is sometimes convenient to qualify the APS citation network as being *minimally resolved* to emphasise its growing network nature. What may be called the *maximally resolved* APS citation network falls at the opposite end of the time resolution spectrum. In this case there are as many time-steps as there are articles so that the sequence is grown by $n_t = 1$ node with $m_t \geq 0$ edges at each time-step t. The value of m_t' is equal to 0 for all t since self-citations are prohibited. Identically timestamped articles are discriminated according to the lexicographical ordering of their unique article IDs in a slight abuse of the representation. It is easy to imagine in a similar vein *daily*, *monthly*, and *yearly resolved* APS citation networks lying between these two extremes. For example, yearly resolution means that all articles published in the same calendar year are mapped to nodes in the same time-step. For a given time-step t, n_t is the number of articles published in the corresponding year, m_t the number of citations to articles from previous years, and $m_t'$the number of citations to articles in the same year. Note that journal issue print date timestamps are used to construct the APS citation network at daily resolution. Figure S2 shows a conceptual depiction of the time resolutions here described. The growing networks we have described here will prove key to resolving the preferential attachment paradox in a later section.

But in the present subsection, our focus is squarely on a collection of *bi-epochally resolved* APS citation networks. In principle, bi-epochal resolution describes the scenario when a partition of the article timestamps of a bibliographic dataset into two non-overlapping intervals, labeled T_1 and T_2 hereafter, is used to define a growing network representation comprised of two time-steps. In practice, the time intervals do not always cover the entire data. Table S2 summarises our reconstruction of four bi-epochally resolved growing network representations of the APS citation data that Redner submitted to analysis in his original study[1].

In order to characterise preferential attachment network scientists measure the rates at which articles with k citations are cited by new articles. This they achieve by observing the process of citation formation over time as viewed through the prism of this or that growing network representation. Loosely speaking, the *attachment rate* $\hat{A}(k)$ of a growing network is defined as the likelihood that an edge from among the m_t added at time-step $t > 1$ connects to a node of in-degree k. Jeong *et al.*[7] proposed to measure the attachment rate of a bi-epochally resolved growing network, $\mathcal{G} = \{G_1, G_2\}$, as

$$\hat{A}(k) \overset{\text{def}}{=} \frac{n_1}{m_2} \times \frac{m_2(k)}{n_1(k)}, \tag{2}$$

where $m_2(k)$ is the number of edges from G_2 that connect to an in-degree k node in G_1 of which there are assumed to be $n_1(k)$ in number. The factor n_1/m_2 serves as a mathematically convenient constant of normalisation. The domain is taken to be $\{k|m_2(k)/n_1(k) \neq 0\}$ under the convention that $0/0 = 0$. Mark Newman took steps to generalise Jeong's measure to certain arbitrarily time resolved growing networks[8]. The details of Newman's measure are deferred to a later section. The attachment rate of a growing network is sometimes said to be "preferential" when the trend line of the measured $\hat{A}(k)$ is found to be an increasing function of k. But in this paper, we apply the term "preferential" to measured attachment rates in a more restricted sense. Namely: if $\hat{A}(k)$ increases linearly in k, then attachment rate is said to be *preferential*. In this idealised case the in-degree distribution of the resulting network is bound to follow a power-law under certain regularity conditions[2,3,5].

Redner found the attachment rates for bi-epochally resolved growing network representations of various APS citation data subsets to be nearly linear functions of k with the agreement being especially pronounced for k less than 150[1]. In Fig. 2 we reproduce his experimental measurements in all but a few extraneous details that are discussed in

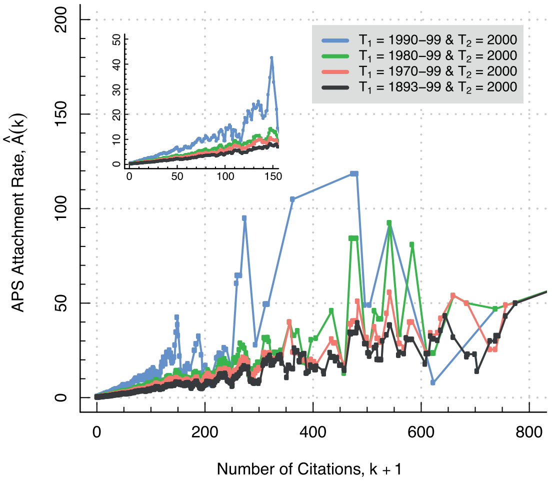

Figure 2. The attachment rate for various bi-epochally resolved APS citation networks appear linear to the naked eye over a wide range of k, especially in the region from $k=0$ to about 150 (inset). Four different measurements of $\hat{A}(k)$ may be distinguished by colour in the plot. Each individual measurement is determined by recording how the publications in time interval $T_2=2000$ cite the publications in time intervals $T_1=1990$–99 (blue), $T_1=1980$–99 (green), $T_1=1970$–99 (red), and $T_1=1893$–99 (black), respectively. The data have been averaged over a range of $k \pm 0.025\,k$. See Table S2 and the surrounding text for details.

Model Name	Attach. Fn.	Eq.	Deg. Dist.	Ref.
Price's model	preferential	(4)	power-law	[2]
Jeong's model	preferential	(4)	unknown	[7]
Callaway's model	uniform	(5)	exponential	[15,16]
Krapivsky's model	log-linear	(6)	diverse	[5,6]
Redner's model	nonlinear	(7)	log-normal	[1]

Table 1. An assortment growing network models relevant to the present work.

Supplementary Note 2. The attachment rate $\hat{A}(k)$, as defined by Jeong's measure, is plotted for bi-epochally resolved growing network representations of four different subsets of the APS citation data. In each case, the measured $\hat{A}(k)$ is observed to approximately follow a straight line. Thus, the presence of preferential attachment is apparently confirmed in each growing network. Redner extrapolates from these bi-epochal outcomes that preferential attachment accounts for the formation of citations in the maximally resolved APS citation network. Let us provisionally accept this conclusion with the understanding that it will be overturned in due course.

A Digression on Growing Network Models.

The bridge between preferential attachment and network in-degree distribution is the growing network model. In this subsection we describe a general growing network modelling scheme that includes a number of important growing network models as special cases. Table 1 summarises the particular growing network models described in detail below.

We define a *growing network model* as a growing network subject to the following constraint: each edge from among the m_t edges added at time-step $t>1$ connects to a given node of in-degree k from G_{t-1} with probability proportional to the *attachment function* $A(k)$, a time-independent function of k that governs the formation of new connections. In particular, the probability that a said edge connects to some in-degree k node from G_{t-1} is given by

$$\pi_t(k) \propto n_{t-1}(k) \times A(k), \tag{3}$$

where $n_{t-1}(k)$ stands for the number of in-degree k nodes in G_{t-1} for $t>1$. It is the form of $A(k)$ together with any structural constraints imposed on the values of T, n_t, m_t, and m_t' that defines a growing network model. The attachment rate $\hat{A}(k)$ of a growing network may be rightly regarded as a realisation of an attachment function, $A(k)$, as defined by a compatible growing network model. Note that we allow for multiple edges to occur between nodes in the above formulation of a growing network model as a matter of mathematical convenience.

In Price's model[2], or rather, a mild generalisation thereof, the attachment function takes the linear form

$$A(k) \propto k + 1, \tag{4}$$

where the unit offset acts as a kind of initial attractiveness, ensuring that zero in-degree nodes stand a fighting chance of acquiring new connections. The form of Eq. (4) makes precise what we mean by a preferential attachment function in this paper. The model definition is completed by taking $n_t = 1$ and assuming the mean value m of the m_t's is constant over time as t becomes large. The average in-degree distribution of networks generated in this manner is known to follow a power-law tail with scaling exponent $\gamma = 2 + 1/m$ in the limit of large T[14]. What we will call Jeong's model is the growing network model analog of the bi-epochally resolved growing network construction from the previous subsection. It consists of two networks G_1 and G_2, i.e., $T = 2$. The preferential attachment function, as defined in Eq. (4), governs the formation of connections between m_2 of the n_2 nodes in G_2 with the n_1 nodes of G_1 at time-step $t = 2$. Callaway's model, formulated here in the language of Price's model without loss of substance, is the random recursive tree defined by substituting the uniform attachment function

$$A(k) \propto 1 \tag{5}$$

into Price's model. Callaway's model has been shown to generate networks with exponentially distributed in-degree distributions[15,16], and its other properties have been examined at length in the classic literature[15,17,18]. An important Price's model generalisation defined by the log-linear attachment function

$$A(k) \propto (k+1)^{\alpha} \tag{6}$$

for *attachment exponent* $\alpha > 0$ was analysed by Krapivsky, Redner, and Leyvraz[5,6]. They fittingly named their model "the growing network model," but we refer to it as Krapivsky's model in this paper. We have slightly redefined the attachment function from the original $k^{\alpha} + 1$ for mathematical convenience. Price's model corresponds to the special case when $\alpha = 1$. For $0 < \alpha < 1$ the resulting in-degree distribution takes the form of a stretched exponential function[6]. For $\alpha > 1$ all nodes connect to a handful of large hubs. Meanwhile the limiting case of $\alpha = 0$ corresponds to Callaway's model. Note that we show the APS citation distribution fitted to the stretched exponential function predicted by Krapivsky's model in Fig. S3.

Finally, Redner writes in passing that the growing network model obtained by substituting the nonlinear attachment function

$$A(k) \propto \frac{k+1}{1+\beta\log(k+1)} \tag{7}$$

with $\beta > 0$ into Price's model generates networks with log-normally distributed in-degree distributions[1]. In Supplementary Note 3, we show that Redner's model, as we will call it, generates networks with in-degree distributions that asymptotically follow the log-normal distribution. The proof is adapted from an outline that was kindly supplied to the authors by Redner via email.

The Preferential Attachment Paradox Illustrated on the APS Citation Data.

It is only in virtue of the preceding digression on growing network models that it has at last become possible to cast the preferential attachment paradox in a reasonably technical light. The paradoxical argument runs as follows:

Premise 1 A preferential attachment rate gives rise to networks with power-law in-degree distributions. Recall that we have defined a preferential rate of attachment to mean that a growing network model attachment function $A(k)$ increases linearly with k.

Premise 2 Measurement suggests a preferential rate of the attachment for the maximally resolved APS citation growing network. In other words, the observed attachment rate $\hat{A}(k)$ is approximately a linear function of k.

Premise 3 The observed APS citation network in-degree distribution is not well-described by a power-law.

Conclusion That the APS citation network has a power-law in-degree distribution follows from a naive application of Premises 1 and 2.

Paradox The stated conclusion is in direct contradiction with Premise 3. In fact, measurement suggests that the APS citation network in-degree distribution is better described by a log-normal distribution, than by a power-law.

The conclusion is trivially seen to follow from the premises. Thus it must be the case that some or another premise is either incoherent or outright false. Redner followed this line of reasoning to its contradictory conclusion for the APS citation network. But it is worth noting that the argument applies to any growing network featuring preferential attachment, which culminates in a network with a log-normally distributed in-degree distribution.

The Preferential Attachment Paradox Resolved.

We have seen that a network cannot enjoy a log-normal in-degree distribution and have grown in accordance with preferential attachment without apparently contradicting network theory. Yet, we are committed to the view that the APS citation network is endowed with exactly these properties. In this section we will see that a critical examination of the premises underlying the argument leads to a ready explanation of the paradox.

First of all, it may be outright denied that the APS citation distribution is log-normally distributed over its full domain; for to maintain otherwise would blindly disregard the statistical testing outcomes presented in the subsection on modelling the APS citation network citation distribution. This line of objection, while technically correct, does not present an interesting challenge to the argument. The APS citation distribution being well-described by the log-normal in the body of k (i.e. $0 \leq k \leq 150$) turns out to be enough to resolve the paradox. In fact, we will see that the extent to which the log-normal falls short of the APS citation distribution in the extreme tail region of k (i.e. $k \geq 150$) is explained by an equal and opposite departure from an ideal in the APS citation network attachment rate. The upshot is that log-normality assumption may be accepted without prejudice to the argument.

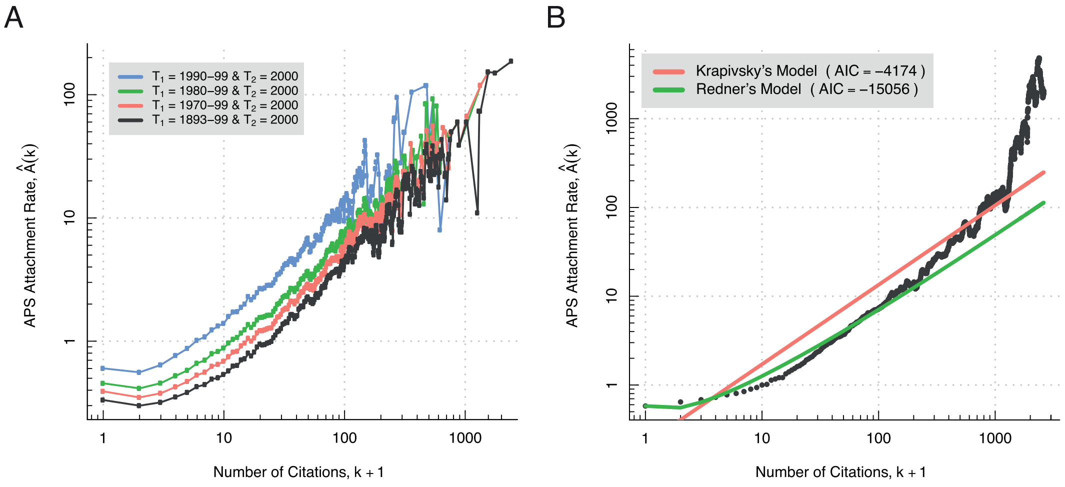

Figure 3. (**A**) Nonlinear tendencies in the attachment rates for various bi-epochally resolved APS citation networks are made apparent on a double logarithmic scale. Four different measurements of $\hat{A}(k)$ may be distinguished by colour in the plot. Each individual measurement is determined by recording how the publications in time interval $T_2 = 2000$ cite the publications in time intervals $T_1 = 1990$–99 (blue), $T_1 = 1980$–99 (green), $T_1 = 1970$–99 (red), and $T_1 = 1893$–99 (black), respectively. The data have been averaged over a range of $k \pm 0.025k$. See Table S2 and the surrounding text for details. (**B**) The attachment rate for the maximally resolved APS citation network is best fit by Redner's model. Shown is the attachment rate for the maximally resolved APS citation network as calculated by Newman's measure (black), the estimated log-linear attachment function of Krapivsky's model (red), and the estimated nonlinear attachment function of Redner's model (green). The best model (the smallest AIC value) is Redner's model. Price's model is not included in the model comparison because it is a special case of Kravipsky's model. The data have been averaged over a range of $k \pm 0.025k$.

The second premise holds that the maximally resolved APS citation network grew in accordance with a preferential rate of attachment. According to Redner's argument, evidence in support of this claim is found in the linear character of the bi-epochally resolved APS citation network attachment rates from Fig. 2. These results, it will be remembered, were extrapolated to the maximally resolved APS citation network as a whole. Redner's conclusion rests on the assumption that the bi-epochally resolved attachment rates are in fact linear. But in Fig. 3(A) the very same attachment rates are shown plotted on a double logarithmic scale. Visual inspection reveals the log-transformed attachment rates to not strictly adhere to straight line relationships. The linear scale plot of Fig. 2 must therefore conceal the nonlinearities made apparent in the log-log plot, since a straight line must again be such on a double logarithmic scale. This shows how the plotting of attachment rates on a linear scale can be misleading. Thus Redner's extrapolation is thrown into jeopardy, and, as a result, his argument for the truth of the second premise collapses.

The question is whether an explanation of the paradox follows from the manner by which the argument for the second premise fails to apply to the maximally resolved APS citation network. In the remainder of this section we venture to answer the question in the affirmative.

This brings us to the connection between the empirical world of growing networks and the theoretical world of growing network models. In particular, the measuring of a preferential rate of attachment is asserted to be a necessary and sufficient condition for concluding that a growing network is well-modelled by Price's model. This test for Price's model, which at first sight might seem unobjectionable, is revealed to be misleading as soon as efforts are made to formulate it carefully. The trouble stems from implicitly assuming that a preferential rate of attachment is intelligible outside the context of a growing network model. However, preferential attachment is always conditional on a growing network model through not only the laws of edge formation, as defined by the preferential attachment function of Eq. (4), but also the specification of model specific structural constraints. Consequently, the most one can hope to say of a given growing network, even in principle, is that it exhibits preferential attachment with respect to this or that particular growing network model. In other words, a preferential rate of attachment is necessary (but not sufficient) for concluding that Price's model describes a growing network.

In fact, a growing network, $\mathcal{G}$, is obliged to satisfy four conditions in order to comply with Price's model. First, $\mathcal{G}$'s initial network G_1 should be small relative to its final network G, i.e., $n_1 \ll N$. Practical experience suggests to us that $N \geq 1000 \times \sqrt{n}_1$ serves as a good rule of thumb, but this by no means rests on a sound theoretical foundation. Second, $\mathcal{G}$ must grow by a single node at each time-step, i.e., $n_t = 1$ for $t > 1$. Third, the number of edges m_t added at time-step t must come from a distribution with a fixed mean and finite variance. Fourth, the formation of connections must be governed by the linear attachment function of Eq. (4), i.e., preferential attachment must prevail in the growth of the network. The test for Price's model that is assumed in the second premise ignores all but the last of these conditions.

Let us reinterpret the Fig. 3(A) attachment rates in the light of these new revelations. The first thing to note is that a casual inspection of Table S2 reveals the corresponding bi-epochally resolved APS citation networks to be in blatant violation of the Price's model structural constraints. They are, however, consistent by definition with the Jeong's model structural constraints. The second thing is that there is a noticeable tendency toward log-linearity

Resolution	Model	Attach. Fn.	Eq.	AIC	BIC
Maximal	Krapivsky	Log-linear	(6)	−4,174	−4,294
	Redner	**Nonlinear**	(7)	−15,056	−12,429
Daily	Krapivsky	Log-linear	(6)	−7,262	−7,252
	Redner	**Nonlinear**	(7)	−12,434	−12,423
Monthly	Krapivsky	Log-linear	(6)	−6,548	−6,538
	Redner	**Nonlinear**	(7)	−7,716	−7,706
Yearly	**Krapivsky**	**Log-linear**	(6)	−4,207	−4,198
	Redner	Nonlinear	(7)	−3,887	−3,878

Table 2. Model comparison results for growing network representations of the APS citation data at various time resolutions. Shown are AIC and BIC values for the fit of the log-linear attachment function of Krapivsky's model and the nonlinear one of Redner's model to the maximally, daily, monthly, and yearly resolved APS citation data attachment rate, respectively. The best model (the smallest AIC/BIC value) for each level of resolution is indicated in bold. Redner's model best describes the data at the three highest levels of resolution (maximal, daily, and monthly). Krapivsky's model best describes the data at the lowest level of resolution (yearly).

in the attachment rates as the $T_2 = 2000$ articles cite the $T_1 = 1893\text{-}99$ (black), $T_1 = 1970–99$ (red), $T_1 = 1980–99$ (green), and $T_1 = 1990–99$ (blue) articles. In the last case, a log-linear fit is especially not out of the question. It is interesting that we came up with a value of $\hat{\alpha} \approx 0.90$ for the corresponding attachment rate exponent. This value is close to $\hat{\alpha} = 1$, which is the mark of a preferential rate of attachment. So there is even a case to be made for the attachment rate plotted in blue being not only log-linear ($\hat{\alpha} = 0.90$), but also approximately linear ($\hat{\alpha} \approx 1.00$). The same, however, cannot be reasonably maintained of the other attachment rates. The point is that Jeong's measure, as applied by Redner to the APS citation data, may be too crude an instrument to permit for the drawing of subtle distinctions in regard to attachment rate functional form.

All this suggests that it is necessary to take seriously the misspecification of the Price's model structural constraints in order to characterise APS attachment rate functional form. Fortunately, Mark Newman devised a way a to measure attachment rates relative to quite a broad class of growing network models[8]. Newman's measure is defined according to

$$\hat{A}(k) \overset{\text{def}}{=} \frac{Z}{W(k)}\sum_{t>1} w_t(k)\frac{m_t(k)}{n_{t-1}(k)}, \tag{8}$$

with weights $w_t(k) = m_t \times [n_{t-1}(k) \neq 0]$ that have sum $W(k) = \sum_{t>1} w_t(k)$ ($[P]$ denotes the Iverson bracket for given proposition P; $[P] = 1$ if P is true and 0 otherwise) and degree independent normalising constant $Z = \sum_{t>1} n_{t-1}/m_{t-1}$; the symbol $n_t(k)$ is used to denote the number of in-degree k nodes in G_t. Newman's measure is consistent with the Price's model structural constraints, because, in contrast with Jeong's measure, it assumes a time resolution consistent with the model. There are several further points regarding the measure that warrant discussion. First, Newman committed a slight error in his original formulation of the measure, the consequence of which was to introduce a waterfall effect in the large k region of measured attachment rates. The measure defined by Eq. (8) incorporates the correction proposed by Pham *et al.*[9] to eliminate this artefact (see Fig. S4 for a dramatic illustration of the said waterfall effect). Second, it is a pleasant exercise to verify that Eq. (8) reduces to Jeong's measure of Eq. (2) in the special case of Jeong's model. Third, Newman's measure assumes that the constant of proportionality implicit to Eq. (3) grows in proportion to the time-step t. This assumption holds true for Price's model which is defined by Eq. (4) with constant m_t on average[19,20]. By contrast, the measure is necessarily approximate in the cases of Krapisky's model (unless $\alpha = 0$ or 1) and Redner's model.

Figure 3(B) shows what happens when Newman's measure is brought to bear on the maximally resolved APS citation network. The results are striking. Visual inspection makes plain that the nonlinear attachment function from Redner's model provides a better fit to the measured attachment rate, than does the log-linear attachment function from Krapivsky's model. The outcome of a model comparison, in which we used the AIC criteria to select the best model, lends numerical support to this conclusion. The AIC score is −4174 for Krapivsky's model and −15056 for Redner's model. It follows that Redner's model compares favorably to that of Price, since the latter forms a special case of Krapivsky's model. The resolution to the paradox is now obvious: The APS citation distribution closely follows a log-normal distribution, because the underlying network's growth is closely described by a growing network model (i.e. Redner's model) that predicts just such an outcome. This explains the paradox.

It is instructive, as an afterthought, to extend our model comparison to the daily, monthly, and yearly resolved APS citation data. Table 2 shows that the AIC and BIC criteria selects Redner's model over Krapivsky's model in all instances with the single exception of the yearly resolution case. This is interesting because it highlights a tendency toward log-linearity in the APS attachment rate as the time resolution decreases. Figure 4 conveys the effect graphically. In Panel A, the measured attachment rates are plotted for the daily, monthly, and yearly resolved data. Panel B shows the same attachment rates overlaid with segmented linear regression lines of best fit we calculated using the R package earth 4.4.7[21]. We defined a log-linearity score heuristic for an attachment rate as the common logarithm of the horizontal component of the longest log-linear segment. Thus the higher the score, the more "log-linear" the attachment rate. As expected, attachment rate log-linearity increases with decreasing time resolution, so that the yearly resolved attachment rate is the most log-linear. Panel C shows the plotted Redner's model

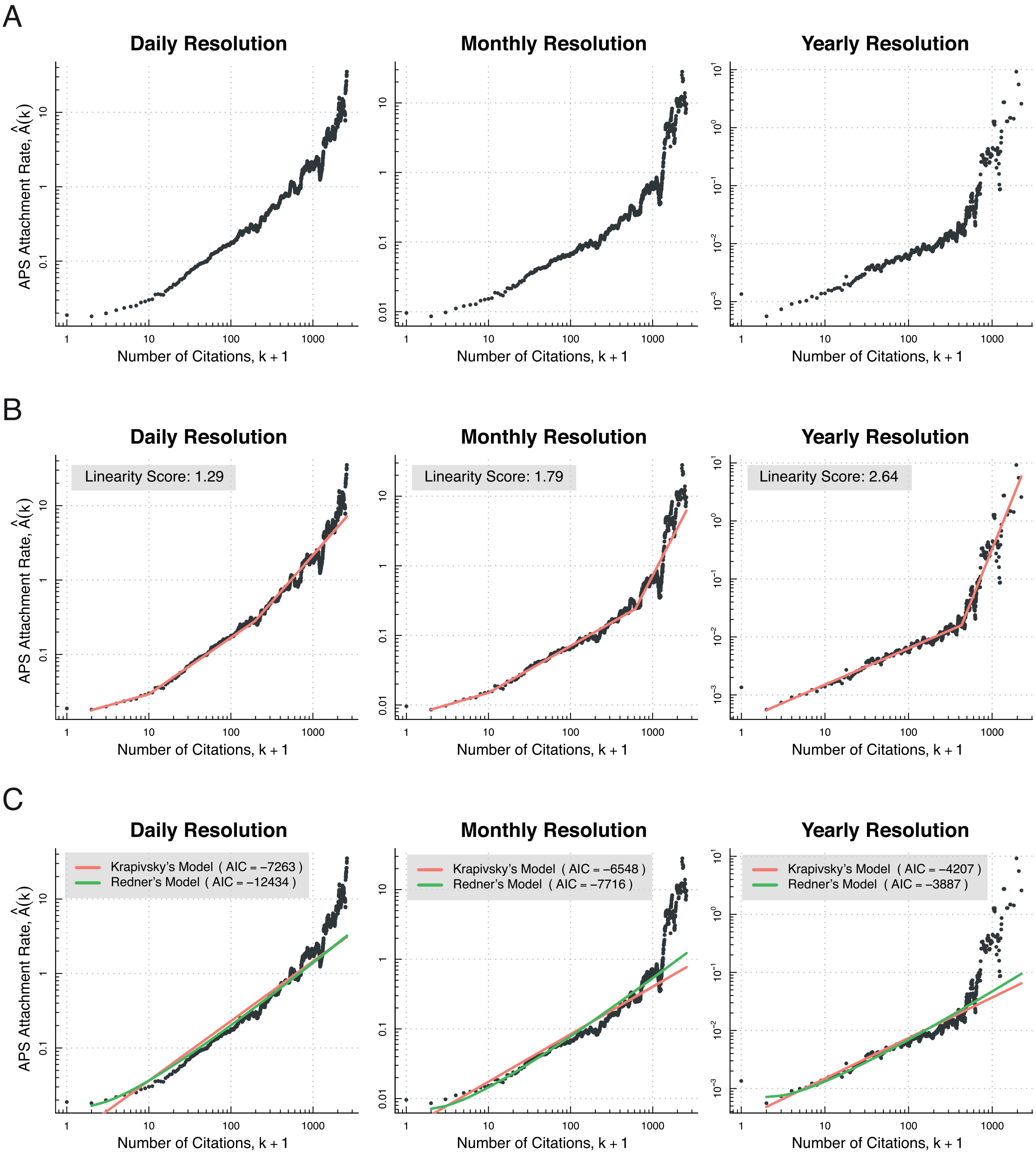

Figure 4. Overview of the measured attachment rates for growing network representations of the APS journal collection data at daily, monthly, and yearly time resolutions. (**A**) The attachment rates plotted in isolation. Each $\hat{A}_k$ was estimated using Newman's measure from APS journal publications dating from July 1893 through 2003 inclusive. The data have been averaged over a range of $k \pm 0.025\,k$. (**B**) The same attachment rates fitted using the segmented linear regression technique discussed in the main text. A larger linearity score reflects a stronger log-linear tendency in a measured attachment rate. The yearly attachment rate is the most log-linear by this score. (**C**) The attachment rates are fitted by the estimated log-linear attachment function of Krapivsky's model (red), and the estimated nonlinear attachment function of Redner's model (green). The best model (the smallest AIC value) is Redner's model in the case of daily and monthly time resolution, but Krapivsky's model in the case of yearly time resolution. Price's model is not included in the model comparison because it is a special case of Kravipsky's model.

and Krapivsky's model attachment functions of best fit. The lesson is that we can expect crudely time resolved data to exhibit a bias toward log-linearity in measured attachment rates.

All that remains is to tie up a few loose ends. First, we have asserted that confining ourselves to the range $0 \leq k \leq 150$ would be sufficient to explain the paradox. For justification, observe how the maximally resolved APS citation network attachment rate plotted in Fig. 3(B) overshoots the Redner's model attachment function after about $k \geq 150$, and the APS citation distribution plotted in Fig. 1 overshoots the log-normal distribution after about $k \geq 150$. These effects are two sides of the same coin: the attachment rate overshooting the predicted attachment function (i.e. $k \geq 150$ nodes acquiring citations at a higher expected rates) automatically leads to the citation distribution overshooting in the log-normal distribution (i.e. $k \geq 150$ nodes are more highly connected than predicted by the log-normal). Thus the lack of agreement between theory and observation can be understood within the modelling framework we have presented, and does not detract from our arguments. Second, the maximally resolved APS citation data, on which we rely to explain the paradox, is consistent with the constant m_t on average assumption on which the models we have considered here depend (i.e. $m_t = 1$ for all t). However, it is important to point out that this assumption is violated for more coarse time resolutions. For example, the number of APS articles have grown exponentially over time with a doubling rate of about 6.5 years. And lastly, the matter of whether the attachment rate remains constant over time in the case of the APS citation data merits some consideration, since this is an assumption of our models. To test the assumption, we partitioned the data from 1901 to 2000 into four non-overlapping time windows (i.e. 1901–74, 1974–88, 1988–95, 1995–2000) and estimated the attachment exponent separately in each case using Newman's method at maximal time resolution. The time windows were selected such that the number of articles are equally distributed. The corresponding estimates for α (i.e. 0.97, 0.94, 1.05, & 1.06, respectively) lend credence to the notion that the constant attachment rate assumption holds at least approximately true.

Discussion

The main purpose of this paper has been to resolve the preferential attachment paradox. Our proposed resolution highlights various pitfalls that the working network scientist would do well to avoid when measuring preferential attachment.

First, we have called attention to the basic fact that an attachment rate is always measured relative to this or that growing model. Granted, this observation is not particularly important as regards the brute assessment of whether or not real-world network attachment rates increase on average with node degree. Recall that this is one way to define preferential attachment. The measurement procedures that Mark Newman[8] and Jeong *et al.*[7] proposed in the early 2000s have proved adequate for confirming this form of preferential attachment for numerous instances as summarised in other sources[9,20]. The same holds true of more recent measurement procedures[9,19,22,23]. But the situation is completely different for the characterisation of attachment rate functional form. In the present work we have seen that the APS citation network attachment rate is better modelled by a nonlinear function under maximal time resolution, and a log-linear function under yearly resolution. This serves as a cautionary tale when it comes to making model-free statements about attachment rate functional form. Second, we have taken pains to state the importance of using the corrected version of Newman's method[9] when assessing attachment rate function form at fine time resolutions. Third, the importance of plotting attachment rates on a double logarithmic scale cannot be overstated in light of the striking contract between the plots of Figs 2 and 3(A).

On a different note, we would be remiss not to comment on the conspicuous lack of statistical formalism employed in the analysis of attachment rate data. The contrast in technical sophistication between the manners in which degree distributions and attachment rates are characterised in the literature is striking. Analysing the APS citation distribution was straightforward thanks to the statistical formalism of Clauset *et al.*[13] as implemented in the poweRlaw R package[12]. More generally, the standardisation of fitting power-laws and other heavy-tailed forms to observed degree distributions was a direct outcome of Clauset *et al.*[13]. No comparable formalism exists for attachment rate analysis to our knowledge. Although important strides in the modelling of citation dynamics are found in the work of Eom and Fortunato[24], and Golosovsky and Solomon[25,26]. This is an intolerable state of affairs seeing that attachment rate and degree distribution are a package deal in so far as growing networks are concerned. An easy-to-use statistical toolkit is needed for fitting and comparing established growing network model attachment functions to observed attachment rates. Fortunately, it should be possible to adapt the maximum likelihood estimation methods and goodness-of-fit tests described in Clauset *et al.*[13] to this purpose. Implementing the proposed methodology in Python and R would go a long way to streamline the analysis of attachment rate data in academic publications.

Lastly, there is a pressing need for a review paper on the measurement of the chief processes describing how complex networks change over time. The measuring of preferential attachment in growing networks, which has so preoccupied our thinking in the present work, is part of a larger enterprise to measure nothing short of all conjectured network evolutionary processes. Preferential attachment is one of many such processes to have been conjectured, including node fitness[27], node duplication coupled with edge rewiring[28], homophily[29], topological distance[8], and node birth/death processes[4]. At least three good reviews have been written on generative network models[30–32], but none on the subject of measuring the processes they embody in real-world networks. It is high time for a survey of the methodological landscape and critical exposition of real-world findings in this area be undertaken.

References

1. Redner, S. Citation Statistics from 110 Years of Physical Review. *Physics Today* **58**, 49–54, https://doi.org/10.1063/1.1996475 (2005).
2. de Solla Price, D. J. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science* **27**, 292–306 (1976).
3. Barabási, A.-L. & Albert, R. Emergence of Scaling in Random Networks. *Science* **286**, 509–512, https://doi.org/10.1126/science.286.5439.509 (1999).
4. Dorogovtsev, S. N., Mendes, J. F. F. & Samukhin, A. N. Structure of Growing Networks with Preferential Linking. *Physical Review Letters* **85**, 4633–4636, https://doi.org/10.1103/physrevlett.85.4633 (2000).
5. Krapivsky, P. L., Redner, S. & Leyvraz, F. Connectivity of Growing Random Networks. *Physical Review Letters* **85**, 4629–4632, https://doi.org/10.1103/physrevlett.85.4629 (2000).
6. Krapivsky, P. L. & Redner, S. Organization of growing random networks. *Physical Review E* **63**, 066123+, https://doi.org/10.1103/physreve.63.066123 (2001).
7. Jeong, H., Néda, Z. & Barabási, A. L. Measuring preferential attachment in evolving networks. *Europhysics Letters* **61**, 567–572, https://doi.org/10.1209/epl/i2003-00166-9 (2003).

8. Newman, M. E. J. Clustering and preferential attachment in growing networks. *Physical Review* E **64** http://arxiv.org/abs/cond-mat/0104209cond-mat/0104209 (2001).
9. Pham, T., Sheridan, P. & Shimodaira, H. Pafit: A statistical method for measuring preferential attachment in temporal complex networks. *PLoS One* **10**, e0137796 https://doi.org/10.1371/journal.pone.0137796 (2015).
10. APS Journals. *APS Data Sets for Research*, http://journals.aps.org/datasets [Online; accessed 1-September-2017] (2017).
11. Radicchi, F., Fortunato, S. &Vespignani, A. Citation networks. In Scharnhorst, A., Börner, K. & van den Besselaar, P. (eds.) *Models of Science Dynamics: Encounters Between Complexity Theory and Information Sciences*, 233–257 https://doi.org/10.1007/978-3-642-23068-4_7 (Springer Berlin Heidelberg, Berlin, Heidelberg, 2012).
12. Gillespie, C. S. Fitting heavy tailed distributions: The poweRlaw package. *Journal of Statistical Software* **64**, 1–16 http://www.jstatsoft.org/v64/i02/ (2015).
13. Clauset, A., Shalizi, C. R. & Newman, M. E. J. Power-Law Distributions in Empirical Data. *SIAM Review* **51**, 661–703 (2009).
14. Newman, M. *Networks: An Introduction* (Oxford University Press, Inc., New York, NY, USA, 2010).
15. Na, H. S. & Rapoport, A. Distribution of nodes of a tree by degree. *Mathematical Biosciences* 6, 313–329 http://www.sciencedirect.com/science/article/pii/0025556470900714 (1970).
16. Callaway, D. S., Hopcroft, J. E., Kleinberg, J. M., Newman, M. E. J. & Strogatz, S. H. Are randomly grown graphs really random? *Physical Review E* **64**, 041902, https://doi.org/10.1103/PhysRevE.64.041902 (2001).
17. Moon, J. W. *The distance between nodes in recursive trees*, 125–132. London Mathematical Society Lecture Note Series (Cambridge University Press, 1974).
18. Meir, A. & Moon, J. On the altitude of nodes in random trees. *Canadian Journal of Mathematics* 997–1015 https://doi.org/10.4153/CJM-1978-085-0 (1978).
19. Massen, C. & Jonathan, P. Preferential attachment during the evolution of a potential energy landscape. *The Journal of Chemical Physics* **127**, 114306 (2007).
20. Sheridan, P., Yagahara, Y. &Shimodaira, H. Measuring preferential attachment in growing networks with missing-timelines using markov chain monte carlo. *Physica A: Statistical Mechanics and its Applications* **391**, 5031–5040 http://EconPapers.repec.org/RePEc:eee:phsmap:v:391:y:2012:i:20:p:5031-5040 (2012).
21. Milborrow, S. Derived from mda:mars by Hastie, T. & Tibshirani, R. *earth: Multivariate Adaptive Regression Splines*, http://CRAN.R-project.org/package=earth (2011).
22. Gómez, V., Kappen, H. J. & Kaltenbrunner, A. Modeling the structure and evolution of discussion cascades. In *Proceedings of the 22nd ACM Conference on Hypertext and Hypermedia*, HT'**11**, 181–190 (ACM, New York, NY, USA, https://doi.org/10.1145/1995966.1995992 (2011).
23. Kunegis, J., Blattner, M. &Moser, C. Preferential attachment in online networks: Measurement and explanations. *In Web Sci'13* (France, (2013).
24. Eom, Y.-H. & Fortunato, S. Characterizing and Modeling Citation Dynamics. *PLoS ONE* **6**, e24926+, https://doi.org/10.1371/journal.pone.0024926 (2011).
25. Golosovsky, M. & Solomon, S. Stochastic dynamical model of a growing citation network based on a self-exciting point process. *Phys. Rev. Lett.* **109**, 098701, https://doi.org/10.1103/PhysRevLett.109.098701. (2012).
26. Golosovsky, M. & Solomon, S. Growing complex network of citations of scientific papers: Modeling and measurements. *Phys. Rev. E* **95**, 012324, https://doi.org/10.1103/PhysRevE.95.012324. (2017).
27. Bianconi, G. & Barabási, A. Competition and multiscaling in evolving networks. *Europhysics Letters* **54**, 436 (2001).
28. Pastor-Satorras, R., Smith, E. & Solé, R. V. Evolving protein interaction networks through gene duplication. *Journal of Theoretical Biology* **222**, 199–210 http://www.sciencedirect.com/science/article/pii/S0022519303000286 (2003).
29. McPherson, M., Lovin, L. S. & Cook, J. M. Birds of a Feather: Homophily in Social Networks. *Annual Review of Sociology* **27**, 415–444, https://doi.org/10.1146/annurev.soc.27.1.415 (2001).
30. Albert, R. & Barabási, A.-L. Statistical mechanics of complex networks. *Reviews of Modern Physics* **74**, 47–97, https://doi.org/10.1103/RevModPhys.74.47. (2002).
31. Boccaletti, S., Latora, V., Moreno, Y., Chavez, M. & Hwang, D. Complex networks: Structure and dynamics. *Physics Reports* **424**, 175–308, https://doi.org/10.1016/j.physrep.2005.10.009 (2006).
32. Holme, P. Modern temporal network theory: a colloquium. *The European Physical Journal B* **88**, 1–30, https://doi.org/10.1140/epjb/e2015-60657-4 (2015).

Acknowledgements

We kindly thank Sidney Redner for supplying us with a sketch of the proof which appears in Supplementary Note 3 and Thong Pham for some helpful discussions.

Author Contributions

P.S. and T.O. conceived the analysis, P.S. conducted the analysis. P.S. and T.O. wrote the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-018-21133-2.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.