Supplementary Materials for

Quantifying Long-Term Scientific Impact

Dashun Wang, Chaoming Song, Albert-László Barabási*

*Corresponding author. E-mail: alb@neu.edu

Published 4 October 2013, *Science* **342**, 127 (2013)
DOI: 10.1126/science.1237825

This PDF file includes:

Supplementary Materials for

Quantifying Long-term Scientific Impact

Dashun Wang, Chaoming Song, and Albert-László Barabási

correspondence to alb@neu.edu

August 30, 2013

Contents

S1 Data Description

S1.1 Physical Review Corpus

The Physical Review (PR) dataset consists of all papers published by journals within the Physical Review corpus from 1893 to 2010 (Table S1). The data is available by request through the APS website. The corpus is comprised of over 450,000 papers with citations. The data is unique in its longitudinal nature, spanning over 100 years. Therefore, it is ideal for understanding the long term aspects of citation histories and impact. Yet, it has two major limitations: (1) It is discipline specific, containing physics papers only. Therefore, the results obtained using this data need to be checked on data pertaining to other disciplines. (2) The data includes only internal citations. Hence, the Web of Science citations of each paper are higher than contained in this data. Such incompleteness introduces a systematic undercount of the impact of interdisciplinary papers. Hence we systematically validate our results on Web of Science data.

S1.2 Web of Science

To correct for the limitations of the Physical Review corpus and to test the generality of our results, we also downloaded papers and citations from Web of Science database from Thomson Reuters. This dataset indexes citations using six major databases with comprehensive coverage. Thus it automatically solves limitation (2). To address limitation (1), we selected 12 journals based on their reach and impact (Table S4). They include general audience journals, like *Nature*, *Science*, and *Proceedings of the National Academy of Sciences* (*PNAS*); leading journals within a certain discipline, like *Cell*, *New England Journal of Medicine* (*NEJM*), *Physical Review Letters* (*PRL*); and review journals, like *Reviews of Modern Physics* (*RMP*). We downloaded all papers published by these journals in three different years (1990, 1995, 2000), and all citations collected by each of these papers until 2011.

To test the mechanism behind the temporal changes in impact factor of *Cell* and *NEJM*, we also downloaded all papers published by these journals from 1995 to 2005 and their citations. Web of Science only provides publication year of papers published prior to 1985, not publication date. This prevents us from accurately (day resolution) estimating the parameters for papers published before 1985. This limitation is corrected by the excellent longitudinality of the PR corpus.

S1.3 Explosive Growth of Publications

The number of scientific publications grows exponentially, a pattern first pointed out by Price in 1963 [31]. Since then, various groups have shown that this pattern not only holds for the overall scientific enterprise, but also within each discipline [32, 33, 34]. We show in Fig. S1 the number of papers published in each year in the PR corpus. The inset of Fig. S1 gives a cumulative view, i.e., total number of papers published up to a certain year, on a log-linear scale. Figure S1 is in good agreement with previous findings [34] that the number of papers published each year increases exponentially, in analogy to Moore's law describing the development of technology, indicating that

$$\mathcal{N}(t) \sim \exp(\beta t), \tag{S1}$$

where $\beta = (17 year)^{-1}$ for PR corpus. Therefore the number of papers increases by 2.73 times after 17 years, or equivalently doubles every $17 \times \ln(2) = 11.8$ years.

S2 Minimal Citation Model

S2.1 Model Definition

In the proposed model, the total number of papers $\mathcal{N}$ grow exponentially as (S1), and every new published paper has m citations to existing papers. The citation probability Π_i of each old paper i

follows

$$\Pi_i(\Delta t_i) \sim \eta_i c_i^t P(\Delta t_i), \tag{S2}$$

where t_i is the publication time of paper i; $\Delta t_i \equiv t - t_i$ indicates the time elapsed since its publication; and $P(\Delta t_i)$ is the temporal relaxation function that measures the probability that a paper gets cited after time Δt elapsed since its publication, capturing the survival rate of a paper's attractiveness to the research community. The c_i^t term in (S2) captures the preferential attachment mechanism. Accordingly a paper i's probability to be cited again is proportional to the total number of citations c_i the paper received previously. In Fig. S3 we document the presence of preferential attachment in our dataset as well. In this context $\Pi_i(t) \sim c_i^t$ corresponds to the scale-free model and $\Pi_i(t) \sim \eta_i c_i^t$ to the fitness model. As we showed before, these models, lacking aging factors, are inconsistent with the jump-decay patterns observed in citation dynamics (Fig. S16).

The temporal relaxation function $P(\Delta t_i)$ can be measured directly from the real data. Given that a paper's citation is driven by three independent forces (S2), that are difficult to separate from each other, we need to control the influence of these factors, isolating the temporal decay. This is similar to measuring preferential attachment from empirical data, where one keeps a constant time window and looks at the growth of degrees as a function of existing degree [3]. To achieve this we should group papers with the same fitness (η) and cumulative citations (c^t), and look at the time when they are cited again. But we do not know η beforehand. Moreover, each paper is likely characterized by different μ and σ parameters in (S3). Therefore, by aggregating different papers, we will measure a superposition of different temporal relaxation functions. We therefore selected papers published between 1950 and 1960 in the PR corpus with fixed cumulative citations c (i.e., controlling for c, publication time and IF), and tracked the moment when their citations changed from c to $c+1$. We then measured Δt, i.e. time between their publication and when $c \to c+1$ took place.

Figure S4 shows both $P(\ln \Delta t \mid c)$ and $P(\Delta t \mid c)$ for fixed $c = 10$ and $c = 20$, finding that the

relaxation function is best approximated by a lognormal function

$$P(\Delta t) = \frac{1}{\sqrt{2\pi}\sigma\Delta t}\exp\left(-\frac{(\ln\Delta t-\mu)^2}{2\sigma^2}\right). \tag{S3}$$

Indeed, a lognormal distribution naturally emerges in multiplicative processes, frequently used to model the temporal relaxation function in diverse settings, from survival times after cancer diagnosis [35] and latency periods of diseases [36] to the duration of marriages [37] and length of both spoken and written conversations [38, 39]. Theoretically, we obtain a lognormal distribution if the relaxation time Δt is a product of independent, identical distributed random variables $\{x_i\}$, $\Delta t = \prod_i^n x_i$ (equivalently, $\ln\Delta t = \sum_i^n \ln x_i$ converges to a normal distribution due to the central limit theorem). Such multiplicative processes often result from independent decision processes [40, 41]. Similar mechanisms are likely at work in the case of citations: a decision to cite a paper involves balancing many different factors, from appropriateness to novelty, relevance and even citation limits, each of which may be approximated as an independent event with random probability, resulting in a random latent waiting time. The final decision to cite requires us to satisfy all these individual conditions, best described by a multiplicative process. This argument offers an intuitive explanation for the origin of the observed lognormal relaxation time. More models that generate lognormal relaxation times are reviewed in Ref. [42].

S2.2 Solving the Model

Starting from Eq. (S2), the time evolution of the expected number of citations c_i^t satisfies

$$\frac{dc_i^t}{dN} = \frac{\Pi_i}{\sum_{i=1}^N \Pi_i}. \tag{S4}$$

Combining (S2-S4) with Eq. (S1) leading to $\Delta t_i = t - t_i = \beta^{-1}\ln(N/i)$, we obtain

$$\frac{dc_i}{dN} = m\frac{c_i\eta_i P_t(\beta^{-1}\ln(N/i))}{\sum_{i=1}^{N} c_i\eta_i P_t(\beta^{-1}\ln(N/i))}. \tag{S5}$$

Assuming $c_i = m(f(\eta_i, \Delta t_i) - 1)$, we have

$$\frac{df(\eta_i, \Delta t_i)}{d\Delta t_i} = \beta\frac{\eta_i f(\eta_i, \Delta t_i)P_t(\Delta t_i)}{A}, \tag{S6}$$

with the initial condition $f(\eta_i, 0) = 1$, where the normalization constant

$$\begin{aligned} A &\equiv \lim_{N\to\infty} N^{-1}\left\langle \sum_{i=1}^{N} \eta_i P_t(\beta^{-1}\ln(N/i)) f(\eta_i, \beta^{-1}\ln(N/i))\right\rangle \\ &= \lim_{N\to\infty}\left\langle \int_1^N \eta_i P_t(\beta^{-1}\ln(N/i)) f(\eta_i, \beta^{-1}\ln(N/i)) d(i/N)\right\rangle \\ &= \beta\int d\eta\rho(\eta)\int_0^\infty \eta P_t(t') f(\eta, t') e^{-\beta t'} dt'. \end{aligned} \tag{S7}$$

The solution of Eq. (S6) is

$$f(\eta_i, \Delta t_i) = e^{\frac{\beta}{A}\eta_i\int_0^{\Delta t_i} P_t(t')dt'}, \tag{S8}$$

thus

$$c_i^{\Delta t_i} = m\left(e^{\frac{\beta}{A}\eta_i\int_0^{\Delta t_i} P_t(t')dt'} - 1\right), \tag{S9}$$

where the constant A can be calculated from

$$\beta\int\rho(\eta)d\eta\int_0^\infty \exp\left(-\beta t + \frac{\beta}{A}\eta\int_0^t P_t(t')dt'\right)dt = 2. \tag{S10}$$

Plugging Eq. (S3) into Eq. (S9), we get

$$c_i^t = m\left(e^{\frac{\beta}{A}\eta_i\Phi\left(\frac{\ln t-\mu_i}{\sigma_i}\right)} - 1\right), \tag{S11}$$

where $\Phi(x)$ is the cumulative normal distribution

$$\Phi(x) \equiv (2\pi)^{-1/2} \int_{-\infty}^{x} e^{-y^2/2} dy. \tag{S12}$$

As β and A are system wide parameters, we use $\lambda_i \equiv \eta_i \beta / A$ the relative fitness for each paper i, arriving at Eq. (3) that describes the citation dynamics of paper i:

$$c_i^t = m\left(e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right). \tag{S13}$$

Equation (S13) suggests two fundamental impact measures. If we take the limit of $t \to \infty$, the formula predicts the ultimate citation of a paper (c_i^∞). When $t \to \infty$, $\Phi \to 1$. Therefore we arrive at a remarkably simple formula,

$$c_i^\infty = m\left(e^{\lambda_i} - 1\right), \tag{S14}$$

indicating that the ultimate impact of a paper is only determined by the relative fitness.

Equation (S13) also allows us to compute a characteristic time scale of a paper's impact (impact time), often chosen as the time required for a paper to reach half of its ultimate citations. Due to the multiplicative nature of citation model (S13), a natural measure is the time necessary for a paper to reach a 'geometric mean' ($\sqrt{mc_i^\infty}$) of its final citations. As m can be viewed as some sort of initial attractiveness, this is equivalent to solving for t in

$$\sqrt{m\left(m + c_i^\infty\right)} = m\left(e^{\lambda_i \Phi\left(\frac{\ln t - \mu_i}{\sigma_i}\right)} - 1\right). \tag{S15}$$

Considering $e^{\lambda_i \Phi(x)} \gg 1$, we can approximate the impact time T_i^* as

$$T_i^* \approx \exp(\mu_i). \tag{S16}$$

S2.3 Maximum Likelihood Estimation of Model Parameters

In order to test how well the model matches empirical data, we need to estimate the best $(\lambda_i, \mu_i, \sigma_i)$ parameters for each individual paper i given its citation history. We show in this section that this can be done by considering a non-homogeneous stochastic process, with the events corresponding to the arrival of individual citations. Imagine a stochastic process $\{x(t)\}$ where $x(t)$ represents the number of events by time t, satisfying

$$\text{Prob}(x(t+h) - x(t) = 1) = \lambda_0(x,t)h + O(h), \tag{S17}$$

where $\lambda_0(x,t)$ is a time dependent rate parameter. Given an empirically observed set of N events $\{t_i\}$ within the time period $[0,T]$, where t_i indicates the moment when the paper gets cited the i^{th} time, the likelihood that a paper's citation dynamics follows the model can be evaluated by the log-likelihood function

$$\begin{aligned} \ln L &= \sum_{i=1}^{N} \ln\left(\lambda_0(i-1,t_i)\right) - \int_0^T \lambda_0(x(t),t)dt \\ &= \sum_{i=1}^{N} \ln\left(\lambda_0(i-1,t_i)\right) - \sum_{i=0}^{N} \int_{t_i}^{t_{i+1}} \lambda_0(i,t)dt. \end{aligned} \tag{S18}$$

From Eq. (S6), we have

$$\lambda_0(x,t) = \frac{\lambda(x+m)}{\sqrt{2\pi}\sigma t} \exp\left(-\frac{(\ln(t)-\mu)^2}{2\sigma^2}\right) \tag{S19}$$

and

$$\int \lambda_0(x,t)dt = \lambda(x+m)\Phi\left(\frac{\ln(t)-\mu}{\sigma}\right). \tag{S20}$$

Combining Eqs. (S18) and (S20), we find

$$\begin{aligned}\ln L &= N\ln\lambda + \sum_{i=1}^{N}\ln(i+m-1) + \sum_{i=1}^{N}\ln\left(P(t_i)\right) - \lambda\sum_{i=0}^{N}(i+m)\left[\Phi\left(\frac{\ln(t_{i+1}-t_0)-\mu}{\sigma}\right) - \Phi\left(\frac{\ln(t_i)-\mu}{\sigma}\right)\right] \\ &= N\ln\lambda + \sum_{i=1}^{N}\ln(i+m-1) + \sum_{i=1}^{N}\ln\left(P(t_i)\right) - \lambda(N+m)\Phi\left(\frac{\ln(T)-\mu}{\sigma}\right) + \lambda\sum_{i=1}^{N}\Phi\left(\frac{\ln(t_i)-\mu}{\sigma}\right).\end{aligned} \tag{S21}$$

After the change of variables $\hat{l} = (\ln L)/N$ and $\hat{m} = m/N$, we have

$$\hat{l} = \ln\lambda + \langle\ln((i-1)/N + \hat{m})\rangle + \left\langle \ln P(t_i) + \lambda\Phi\left(\frac{\ln(t_i)-\mu}{\sigma}\right)\right\rangle - \lambda(1+\hat{m})\Phi\left(\frac{\ln(T)-\mu}{\sigma}\right). \tag{S22}$$

As the goal is to maximize $\ln L$, which is the same as maximizing $\hat{l}$ (S22), we can obtain the set of parameters that best capture a paper's citation records $(\lambda^*, \mu^*, \sigma^*)$,

$$(\lambda^*, \mu^*, \sigma^*) = \arg\max_{\lambda,\mu,\sigma}\hat{l}(\lambda,\mu,\sigma), \tag{S23}$$

or

$$\begin{aligned}\frac{\partial l(\lambda^*,\mu^*,\sigma^*)}{\partial\lambda^*} &= 0 \\ \frac{\partial l(\lambda^*,\mu^*,\sigma^*)}{\partial\mu^*} &= 0 \\ \frac{\partial l(\lambda^*,\mu^*,\sigma^*)}{\partial\sigma^*} &= 0.\end{aligned} \tag{S24}$$

The first equation in (S24) leads to

$$\lambda^* = \left[(1+\hat{m})\Phi\left(\frac{\ln(T)-\mu^*}{\sigma^*}\right) - \left\langle\Phi\left(\frac{\ln(t_i)-\mu^*}{\sigma^*}\right)\right\rangle\right]^{-1}. \tag{S25}$$

and the rest two are

$$\begin{aligned}
&\left\langle \frac{\ln(t_i)-\mu^*}{\sigma^*} - \lambda^* P_G\left(\frac{\ln(t_i)-\mu^*}{\sigma^*}\right)\right\rangle + \lambda^*(1+\hat{m}) P_G\left(\frac{\ln(T)-\mu^*}{\sigma^*}\right) = 0 \\
&\left\langle \frac{\ln(t_i)-\mu^*}{\sigma^*}\left(\frac{\ln(t_i)-\mu^*}{\sigma^*} - \lambda^* P_G\left(\frac{\ln(t_i)-\mu^*}{\sigma^*}\right)\right)\right\rangle + \lambda^*(1+\hat{m})\frac{\ln(T)-\mu^*}{\sigma^*} P_G\left(\frac{\ln(T)-\mu^*}{\sigma^*}\right) = 1,
\end{aligned} \tag{S26}$$

where $P_G(x) \equiv (2\pi)^{-1/2} e^{-x^2/2}$ is the standard normal distribution.

By solving Eqs. (S25–S26) numerically, we obtained the parameter set $(\lambda^*, \mu^*, \sigma^*)$ for each paper based on its historical citation pattern within the time period $[0, T]$.

When estimating parameters in this paper, we have set $m = 30$. Fixing m for all papers makes it easy to compare the parameters for different papers. If we increase m, λ becomes smaller. Yet we find the fitting and prediction results in the paper are not affected by the choice of m when its value is comparable to $< c >$. The intuition behind this is that m measures the typical number of references contained in each new paper, serving as a proxy of initial attractiveness of a paper.

Note that the case of papers that do not receive any citations is still covered in a self-consistent manner by the model. If a paper's fitness is zero, it will not receive any citations. This is also mathematically consistent with our parameter estimation framework. Indeed, although in this case there is no observation of citations in the data, based on the log-likelihood function when estimating model parameters (Eq. S21), for $N \to 0$, we obtain $\lambda \to 0$. Hence for papers without any citations, the model predicts that their fitness is zero.

S2.4 Model Validation

S2.4.1 Statistical test of data collapse

While the collapse of rescaled citation dynamics (Fig. 1J) are visually impressive for papers published by a wide range of journals, we want to measure more precisely how well the model fits the data. For this purpose, we used the Kolmogorov-Smirnov (KS) measure to evaluate the maximum

deviation from the fitted curve to empirical citation data. We computed the KS measure for each paper shown in Fig. 1J, and plotted the distribution (Fig. S5), finding most papers have a KS value close to 0.05, with 85% papers less than 0.1.

To see if the null hypothesis can be rejected at level 0.1, we construct the Kolmogorov-Smirnov test by using the critical values of the Kolmogorov distribution [43]. By measuring $\sqrt{n}D_n$, where D_n is drawn from the distribution in Fig. S5, we find for 97% of the fittings in Fig. 1J the null hypothesis cannot be rejected.

S2.4.2 Data collapse across different journals

To validate the model we rescaled the citation history of individual papers by the parameters estimated using the procedures described in the preceding section. To demonstrate this process we first determined (λ, μ, σ) for four papers selected for their widely different citation histories (Fig. S6A), finding that after rescaling they all collapse into a single curve (Eq. 5) (Fig. S6B). The reason is explained in Fig. S7: by varying λ, μ and σ, Eq. (S13) can account for a wide range of empirically observed citation histories, from jump-decay patterns to delayed impact.

We show in the main text (Fig. 1F) that the model works well for papers published in *PR* and *Science*. Here we test the data collapse for papers published in several other journals (Fig. S8), finding that the model performs consistently well across different disciplines.

S2.4.3 Comparing model simulation with real citation histories

Using the appropriate $(\lambda_i, \mu_i, \sigma_i)$ for each paper, the model is expected to generate a citation history that resemble the real citation of Fig. 1D. We show in Fig. S9 an example of randomly selected papers published between 1960 and 1970, finding excellent agreement between the model and empirical data.

S2.5 Parameter distributions

S2.5.1 Empirical observations

The model indicates that the differences in the citation history of individual papers are encoded in the (λ, μ, σ) parameters, allowing us to separate the factors that influence scientific impact. This is best illustrated by comparing the density functions $P(\lambda)$, $P(\mu)$ and $P(\sigma)$ for papers published in different journals in the same year (1990) (Fig. S10), indicating striking fitness differences among the journals. For example $P(\lambda)$ for *PRB* is peaked at $\lambda \approx 0.5$ and is characterized by a relative paucity of high fitness publications. In contrast most *Cell* papers have high fitness, in the vicinity of 2 and 3.

We also find a modest temporal shift in fitness distributions (Fig. S11a). The observed $P(\mu)$ and $P(\sigma)$ distributions show a remarkable stability in time, across decades (Fig. S11bc). Hence in most cases the immediacy (μ) and the decay (σ) remain unchanged over decades for some journals (but in some periods they can undergo major changes, as we document for *Cell*). We also detect a weak linear correlation between λ_i and σ_i (Fig. S11d), indicating that papers with high fitness are more likely to have a slower decay, enhancing their long-term impact. Fitness λ and immediacy μ_i are independent for most but the high λ papers (Fig. S11e), suggesting that it takes more than 20 years ($\mu > 9$) for truly influential papers (high λ) to reach their citation peak. This explains the lack of correlation between a paper's early (c^2) and long-term (c^{30}) citations for exceptionally high impact papers (Fig. 1C).

S2.5.2 Relation with prior work

Some of the goals in our paper are shared with previous works, like Stringer *et al* [15]. Indeed, it also identifies the need to estimate the long-term impact of recently published papers, and the limitations of Impact Factor. Yet, we take a radically different path in addressing these issues. Indeed, the Stringer model has three components:

1. Let q be a quality parameter assigned to each paper.

2. Let us assume that the quality follows normal distribution, i.e. $P(q)$ is Gaussian.

3. Finally it introduces the hypothesis that the total number of citations obtained by a paper, c, are related to the quality q via the relation $\ln c = q$.

Given (2) and (3), it immediately follows that the citation distribution $P(c)$ has a lognormal shape. Hence, the Stringer model is a static hypothesis based model, designed to predict the citation distribution of a body of publications. In contrast, Model (Eq. 3) is a mechanistic dynamical model that is built to predict the citation history of individual papers. This different mathematical construction deeply affects their predictive power as well.

Stringer *et al* [15] reported that ultimate citations follow lognormal distributions, bearing high-level similarities with our lognormal aging function, raising an important question: could it be that our lognormal temporal decay is anticipated in the lognormal citation distributions proposed by Stringer et al? As we demonstrate next, the lognormal citation distribution of Stringer is neither a necessary nor a sufficient condition for lognormal temporal decay:

1. The lognormal in [15] is rooted in the assumption that the quality parameters follow a Gaussian distribution. As long as the quality parameters are normally distributed, the citations follow lognormal distribution *independent of the form of ageing function*. To show this, we remove our lognormal ageing function and replace it with an exponential form (standard form in Poisson processes). We show in Fig. S12A that citations continue to follow lognormal distributions. This rejects the hypothesis that the aging function is rooted in the citation distribution.

2. Should the quality parameters follow a distribution different from the normal, the resulting citation distribution is predicted to be different from lognormal, *even if the temporal decay follows a lognormal*. As an example consider the case when the fitness distribution $P(\lambda)$

follows an exponential distribution. Assuming a lognormal temporal decay, it predicts that the citations instead follow a power law distribution (Fig. S12B).

S2.6 Predicting Citations

Our ability to accurately capture the mechanisms driving a paper's citation history raises a tantalizing question: can the proposed model predict the future citations of a publication? In principle, we can use a paper's citation history up to year T_t after publication (training period T_t) to estimate the λ_i, μ_i, σ_i parameters associated with the paper and then use Eq. 3 to predict the paper's future citations. This process can be formalized as the following. Using the training period T_t and k_t sampling citations, we try to predict the number of citations at a future time T_p. Equation 3 predicts the expected increment of citations between the period $(T_t, T_p]$

$$\overline{\Delta k} = (k_t + m)\Big(e^{\eta\left(\Phi((\ln T_p-\mu)/\sigma)-\Phi((\ln T_t-\mu)/\sigma)\right)} - 1\Big). \tag{S27}$$

Hence, the expected citation at time T_p is

$$\overline{k}(\eta,\mu,\sigma) = (k_t + m)e^{\eta\left(\Phi((\ln T_p-\mu)/\sigma)-\Phi((\ln T_t-\mu)/\sigma)\right)} - m \tag{S28}$$

where, by assuming uniform prior distributions of (η,μ,σ), the probability of taking parameters (η,μ,σ) follows,

$$P(\eta,\mu,\sigma) \propto L = e^{\ln L(\eta,\mu,\sigma)}, \tag{S29}$$

where the likelihood function L satisfies Eq. (S21).

Therefore, given a citation history, we can use the model to predict the probability for the paper to have k_p citations at the time T_p,

$$P(k_p) = \int \delta\Big(\overline{k}(\eta,\mu,\sigma) - k_p\Big) P(\eta,\mu,\sigma) d\eta d\mu d\sigma. \tag{S30}$$

Here we neglect the fluctuation from the stochastic process itself, and consider only the uncertainties in parameter estimation.

To give an intuition about (S30), we show $P(k_p)$ for a randomly selected paper for different T_p (Fig. S13), illustrating the narrowly peaked nature of $P(k_p)$. Hence, the most probable future citation k_p^* can be obtained from

$$\left.\frac{dP(k_p)}{dk_p}\right|_{k_p=k_p^*} = 0, \tag{S31}$$

and the upper/lower uncertainty can be obtained from the variance of $P(k_p)$:

$$\sigma_p^+ = \sqrt{\int_{k_p^*}^{\infty} (k_p - k_p^*)^2 P(k_p) dk_p} \tag{S32a}$$

$$\sigma_p^- = \sqrt{\int_{k_t}^{k_p^*} (k_p - k_p^*)^2 P(k_p) dk_p} \tag{S32b}$$

Taken together, based on an existing citation history and by combining Equations (S31) and (S32), the model predicts at each future time T_p, the most likely citations at the time (k_p^*) as well as the confidence range $[-\sigma_p^-, \sigma_p^+]$, represented as a citation envelope (Fig. S14). To systematically evaluate the fraction of papers that fall within the predicted citation envelope, we measure $z_T = |c^T - k_p^*|/\sigma_p^+$, that quantifies how many standard deviations away the real citations deviate from predicted most likely citations. $z_T \leq 1$ indicates that the real citation dynamics fall within the citation envelope. If, however, $z_T > 2$, it indicates that the predicted citations exit the envelope far enough that the citations are not predicted correctly by the model. To this end, we compiled a test base of papers, consisting of all papers in the PR corpus published within the same decade (1960s) that have at least 10 citations in 5 years (4492 papers). As shown in Fig. 4C, $P(z_{30})$ decays fast with z_{30}, indicating it is rather rare to have a paper exiting the citation envelope. Indeed, only about 6.5% papers are not captured by the model ($z > 2$). Interestingly, this fraction is hardly affected by the length of training and testing period, demonstrating the remarkable robustness in the model's ability to capture citation dynamics despite its minimalist nature. In comparison, this

fraction ($P(z \leq 2)$) is around 50% for other competing models (See Sec. S3.2 for details).

S3 Potential Models for Citation Dynamics

The observed accuracy of the proposed model prompts us to ask whether the model is unique in its ability to capture future citation histories. We therefore seek models that can account for the observed diversity in citation dynamics, fit citation histories, and predict future citations. While most models are not specifically designed to capture the citation dynamics of individual papers, we examine in this section some of the most relevant models and discuss their strengths and limitations. Two lines of inquiry are relevant in this context: network growth models from statistical physics built to capture citation networks, and models pertaining to diffusion of innovations in social/ecomonic sciences.

S3.1 Network Growth Models

S3.1.1 Scale-Free Model

The scale-free model [20] (also known as Barabási-Albert (BA) model) is designed to reproduce the degree distribution of complex networks. Note that variants of the model were proposed by Price [2] and Simon [44]. At each step an old paper i acquires citations from a new paper with probability proportional to its current citations $\Pi_i \propto c_i^t$, a mechanism known as preferential attachment (PA).

Despite the success of the scale-free model in predicting fat-tailed citation distribution, it has difficulties capturing the citation dynamics of individual papers. Indeed, for paper i the scale-free model predicts its citation growth as [20]

$$c_i^t \sim \mathcal{N}^{1/2}, \tag{S33}$$

where c_i^t is the cumulative number of citations paper i received, given the $\mathcal{N}$ papers in the system. This indicates that (i) all papers follow the same citation dynamics, in contrast with our observation (Fig. 1D) that each paper has a different citation history; (ii) it tells us that the citations should grow indefinitely at $\mathcal{N}^{1/2}$. If we incorporate in the model the fact that we have an exponential growth in the number of papers (see Eq. (S1)), we find that

$$c_i^t \sim \exp(0.5\beta t), \tag{S34}$$

indicating that the paper citations should increase exponentially over time (Fig. S15). Yet, the average number of citations $\overline{c_i(t)}$ of all papers i at time t after publication follows a distinct 'jump-decay' pattern (Fig. S16), indicating that a paper's main impact comes during the first two years after publication and diminishes over time [2]. Therefore, both (S33) and (S34) represent drastic deviations from the empirical observations.

S3.1.2 Fitness Model

In the fitness model [24] (also known as Bianconi-Barabási (BB) model), besides the PA mechanism each paper i has an initial fitness λ_i capturing its unique likelihood to be cited in the future. That is,

$$\Pi_i \propto \lambda_i c_i^t. \tag{S35}$$

The fitness model predicts

$$c_i^t \sim \mathcal{N}^{\alpha_i}, \tag{S36}$$

where the exponent $\alpha_i \propto \lambda_i$, i.e. it is proportional to paper i's fitness. Given exponential growth of papers (Eq. S1), we find that c_i^t again increases exponentially over time, significantly deviating from the observations (Fig. S16).

S3.1.3 Relationship between the proposed model and network growth models

While the proposed model is based on the conceptual basis offered by models developed in the network science literature, like the fitness model [24], the main finding of the model is that the fitness model is not sufficient to capture the real citation histories. The reason is simple: the fitness model is a generic conceptual model, lacking the key ingredients of citation histories.

To be specific, the primary purpose of the fitness model is to reproduce the overall structure of a network under the presence of fitness. In contrast, the proposed model aims at modeling the citation dynamics of the individual papers. In network science, this would require us to correctly predict the growth dynamics of each individual node. This is never pursued, as the time evolution of individual nodes depends on many details that the existing models, like the fitness model simply ignore, as they do not matter in the $t \to \infty$ asymptotic limit. The main difference between the fitness model and the proposed model is driven by the goal to accurately predict future citation patterns. The two models have two common ingredients: preferential attachment (the c^T term in Eq. 2), and the fitness parameter (the η_i term in Eq. 2). The proposed model has two additional ingredients:

1. The aging function $P(t)$.

2. The exponential growth of the arriving papers.

With these two modifications, we show, for the first time, that the citation history of individual papers is characterized by a remarkable degree of regularity.

S3.2 Diffusion of Innovations

The theory of diffusion of innovations aims to explain the adoption of new ideas and technologies. Although its main focus is to determine the success and failure of a product, the models often predict S-curves that are similar to the one presented in the main text. Next we explore the possibility of using diffusion S-curves to describe the citation history of individual papers.

S3.2.1 Logistic Model

The logistic function is widely used to model population growth and product adoptions, with applications in many fields. In the context of citations one could view a paper as a new product, whose adoption leads to an increase in citations. Each paper is characterized by a different increase rate r and a total number of citations c_i^∞ that captures the differences in impact. With time, a paper's attractiveness fades, as the development along the ideas offered by the paper have been adopted by all potential adopters, hence the paper's citations approach c_i^∞. In the rate equation formalism this can be described as

$$\frac{\mathrm{d}c_i^t}{\mathrm{d}t} = r_i c_i^t \left(1 - \frac{c_i^t}{c_i^\infty}\right), \tag{S37}$$

yielding

$$c_i^t = \frac{c_i^\infty}{1 + e^{-r_i(t-\tau_i)}}, \tag{S38}$$

where c_i^∞, r_i and τ_i correspond to ultimate citation, longevity, and immediacy of paper i.

S3.2.2 Bass Model

One of the most famous models in marketing and management sciences is the Bass model [27], that describes the process of new product being adopted by mass populations. The Bass model assumes the adopters of a product are influenced by two aspects: mass media and word of mouth. Hence the buyers comprise two groups. One group, the innovators as coined by Bass, is influenced only by the mass media, while the other group, the imitators, is influenced by others (word of mouth effect). Such assumptions are fairly reasonable in the context of citations. The innovators correspond to people who cite the paper spontaneously, little influenced by how many people have already cited the paper. At the same time, a paper's citations are driven by word-of-mouth diffusion (the imitators). Mathematically, this can be expressed as

$$\frac{dc_i^t}{dt} = (p + qc_i^t/c^\infty)(c^\infty - c_i^t), \tag{S39}$$

where p characterizes "innovators", reflecting an influence that is independent of current citations (c_i^t), and q reflects the imitation part of the model. Solving (S39) yields

$$c_i^t = c^\infty \frac{1 - e^{-(p+q)t}}{1 + \frac{q}{p} e^{-(p+q)t}}. \tag{S40}$$

S3.2.3 Gompertz Model

The Gompertz model [28], named after Benjamin Gompertz, was first proposed to model human mortality. The model generates a skewed diffusion curve with long tails. In this context, early citations pave the way for new citations and drive the citation dynamics, hence the rate of research develop increases at an exponential rate. This can be formulated as

$$\frac{dc_i^t}{dt} = q c_i^t \ln(c^\infty / c_i^t), \tag{S41}$$

yielding,

$$c_i^t = c^\infty e^{-e^{-(a+qt)}}. \tag{S42}$$

In (S42), a sets the displacement in c_i^t, while q characterizes the growth rate of citations.

It is worth noting that, while these three models are perhaps the most famous ones, they are far from complete to cover the list of models in diffusion of innovations. For a more comprehensive review of this body of literatures, refer to [26].

S3.3 Evaluating Different Models

In this section we evaluate the performance of these three models in describing citation dynamics. There are two aspects we need to evaluate. One concerns with fitting. That is, how closely each model matches citation histories. The second one is about the predictive power of these models. As outlined in Table S2, the three models in S3.2, together with the proposed model, all have 3 parameters each, making it a fair comparison of how well each of the models fit the citation dy-

namics. Therefore, given a paper's citation dynamics, we can obtain the three parameters for each paper for a given model. We implemented two methods to estimate the best parameters for fitting for each the models. One is by using non-linear least square fitting, and the other one is by using Maximum Likelihood Estimation. We find these two methods have comparable performance for these three models, corroborating previous finding that in most cases these two methods perform equally well in fitting these models [26].

We fitted these three models to our test base of papers (all papers in the PR corpus published within the same decade (1960s) that have at least 10 citations in 5 years). To quantify how well the model fits the real data, for each paper i with total c_i^T within time period $[0, T]$ we measured the weighted KS measure

$$D_i = \max_{t\in[0,T]} \frac{|c_i^t - \tilde{c}_i^t|}{\sqrt{(1+c_i^t)(c_i^T - c_i^t + 1)}}, \tag{S43}$$

where $\tilde{c}_i^t$ represents the citations computed by the model. A smaller D_i implies a better fit. Figure S17 (same as Fig. 4D) shows the KS distribution $P(D)$ for both the proposed model and the three competing models, finding the best fit is offered by the proposed model.

To illustrate the fitting results and compare the fit by the four model, we use the four papers shown in Fig. 1G and obtain the best fit (Fig. S18). We find, despite radical differences in citation dynamics of these four papers, the proposed model fit all of them consistently well. Logistic model performs the worst, mainly due to the fact that it predicts a symmetric citation curves (same growth and decay). While the Gompertz and the Bass models predict asymmetric citation pattern, they also predict an exponential (Bass) or double-exponential (Gompertz) decay of citations (Table S2), much faster than observed in real data. As a result, they both over-estimate the citation at small time scales (growth region), particularly for papers with low immediacy (red curve in Fig. S18). As Gompertz, Bass, and Logistic models predict citation tails with an exponential or faster decay, it also affects their predictive power. As a consequence, they tend to predict faster saturation of citation growth, under-estimating future citations. This is reflected in two points. First, only about

50% papers have $z < 2$ (Fig. 4C), indicating a large number of papers exit the predicted citation envelope by these models. Second, the error bars in Fig. 4E,F are systematically below $y = x$, indicating these three models underestimate future citations (the predicted citations are always smaller than the real citations).

Next we offer more quantitative evidence regarding the predictive power illustrated in Fig. 4EF:

(i) While it is clear that the proposed model's prediction results in more green bars than the other models, we need to quantify this difference. For this purpose, we measure (Fig. S19A) the precise fraction of points that fall within the green bars, i.e., the ones that are predicted correctly on average in Fig. 4F. If we assume that the green bars correspond to correct predictions, we find that 99% of the points are correctly predicted by the proposed model, while this fraction is about 12% for the other reference models.

(ii) While the scatter plot of Fig. 4EF indicates that the predictions by the proposed model are close to the $y = x$ line, it does not capture the magnitude of the deviations between the predicted and the measured citations. Hence, in addition to the scatter plot, we also computed the error in the predicted citations as a function of real citations (Fig. S19B). That is, for each group of papers with comparable real citations we measured the average predicted citations within the group less the real citations, normalized by the real citations. We find that the proposed model performs consistently for a wide range of citations, with the error hovering in the vicinity of 0. The errors for the other models are mostly negative, indicating that they systematically underestimate future citations. The deviations for those models deteriorate for the highly cited papers.

S3.4 Special case of the proposed model and the role of preferential attachment

Let us now assume that λ is small and let us take the Taylor expansion around $\lambda = 0$ for Eq. 3. As the model predicts $c^t = m(e^{\lambda\Phi} - 1)$, its Taylor expansion can be written as

$$c^t = m(\lambda\Phi + \frac{1}{2}\Phi^2\lambda^2 + ... + \frac{1}{n!}\Phi^n\lambda^n) \tag{S44}$$

The first order term is $c^t = m\lambda\Phi$, the same as the prediction of the lognormal temporal decay modulated by a single parameter. That is, if we assume

$$c_i(t) = q_i \frac{1}{t\sigma_i\sqrt{2\pi}} e^{-\frac{(\ln t - \mu_i)^2}{2\sigma_i^2}}, \tag{S45}$$

which predicts

$$c^t = q_i\Phi(\frac{\ln t - \mu_i}{\sigma_i}). \tag{S46}$$

Hence, the first order approximation of the model is equivalent to a model that takes preferential attachment out of the model but keeps the lognormal temporal decay and one extra parameter for fitness, prompting us to call this the *Lognormal* model. Therefore, for papers with small fitness, the Lognormal model (S46) behaves the same way as the proposed model. Indeed, when papers have very few citations, the role of preferential attachment can be neglected, and the behavior of the model is similar to having simply the lognormal temporal decay modulated by a single parameter.

This raises an important question: could our model offer a theoretical bound below which the effect of preferential attachment can be neglected? Indeed, when the sum of all the neglected higher order terms in (S44) is not sufficient to increase the total citation by 1, the lognormal model and the full proposed model with preferential attachment are equivalent. Mathematically this corresponds to

$$\sum_{n=2}^{\infty} \frac{1}{n!}\Phi^n\lambda^n < 1. \tag{S47}$$

Solving this equation yields $\lambda \approx 0.25$, predicting the citation threshold $c^\infty \approx 8.5$. That is, for papers that collect ultimately less than 8 citations, their citation dynamics are fundamentally indistinguishable from the Lognormal model. Hence our model gives rise to an analytical prediction of the citation threshold for preferential attachment to be in effect, which is also in good agreement

with previous empirical findings [29].

To test the prediction that preferential attachment plays an important role only for high fitness papers, we grouped the papers based on their number of citations after 30 years (c^{30}), and measured the average number of citations within each group as a function of the time, $c^{real}(t|c^{30})$. We obtain the best fit for each paper using our model and Lognormal models, and compute the average number of citations in each year within groups using the best fitted parameters, $c^{model}(t|c^{30})$. In Fig. R2 we show the ratio between real citations and those generated by the two models, $c^{real}(t|c^{30})/c^{model}(t|c^{30})$. If this value equals one, it indicates that the model predicted citations closely match the real citations.

As Fig. S20A shows, for our model the curves hover in the vicinity of 1, indicating that the proposed model performs consistently for papers with rather different impact at any time. In contrast, the Lognormal model shows systematic deviations (Fig. S20B) for papers with intermediate range of citations (from 40 to 500), for which the curves increase dramatically as time increases, with values approaching 2 at year 30. To be specific, the lognormal model significantly underestimates the real citations, thanks to the lack of preferential attachment. For papers with very high citations (around 1000 in 30 years), the differences become even more pronounced.

In summary, the Lognormal model behaves the same way as our model for papers with few citations ($c < 8$), but it fails to characterize the citation patterns of medium to high impact papers. To be consistent, we repeated the same measurements for other competing models documented in the paper, finding that our model consistently outperforms these models according to this new measure as well. It is also worth noting that between Bass and Lognormal, it is hard to say which one is better based on Fig. S20 only.

S4 Quantifying Journal Impact

We used the Web of Science dataset to compare papers from different journals, We only consider the research papers published by each journal, as only their citations are counted in the IF. This can be achieved by looking at *document type* for each paper indexed by Web of Science. Most specifically, we only consider the document types as *Review* and *Article*. We find that for some journals, like for *Physical Review Letters* (*PRL*), the vast majority of papers are in these categories. Yet for other journals, especially the ones for general audience, many are classified as "letter to editor" or "editorial". Therefore, in analogy to the citable items used in measuring a journal's impact factor (IF) by Journal Citation Reports, this distinction is important to understand a journal's impact.

To quantify the impact of a journal, we count the average citations all papers published by the journal acquire over time,

$$C_j^t = \frac{1}{N_j}\sum_i^{N_j} c_i^t, \tag{S48}$$

where N_j is the total number of papers published by journal j. We find that C_j^t is also well approximated by our model (Fig. S21), indicating that

$$C_j^t = m\left(e^{\Lambda_j \Phi\left(\frac{\ln T - M_j}{\Sigma_j}\right)} - 1\right). \tag{S49}$$

Therefore each journal's citations are captured by three parameters (Λ, M, Σ), in analogy with the (λ, μ, σ) parameters derived for individual papers. To check whether (Λ, M, Σ) represent the average of individual papers published by the journal, we computed for each journal shown in Fig. S21 the (λ, μ, σ) parameters for individual papers published in the journal, and the mean of each parameter with its journal average (Fig. S22). We find that $\langle\lambda\rangle$, $\langle\mu\rangle$, $\langle\sigma\rangle$ are in good agreement with Λ, M, and Σ, indicating that (Λ, M, Σ) parameters are representative for an average paper published by the journal.

S4.1 Calculating the Impact Factor (IF)

The IF of a journal is defined as the average number of citations received per paper by that journal during the two preceding years. Let us consider for example calculating a journal's IF in 1992. In the numerator we need to measure the number of times papers published by this journal in 1990 and 1991 are cited during 1992. This includes all papers published by this journal. In the denominator, we need to normalize by the number of papers. But according to the definition from Journal Citation Reports, this normalization is for the number of "citable items" published by that journal in 1990 and 1991. In principle, the exact expression of the IF can be obtained by integrating over all papers published within the two-year time frame using their corresponding parameters. Assuming the publication date of a paper within a year does not affect its citations, we can treat all papers within a year as published on the same date. Imagine we want to calculate a journal's IF in the year T, and this journal published N_1 papers in the year $T_1 = T - 2$ and N_2 papers in $T_2 = T - 1$. Therefore, based on the definition of IF, we have

$$\mathrm{IF}(T) = \frac{\sum_i^{N_1} c_i(T|T_1) + \sum_i^{N_2} c_i(T|T_2)}{N_1 + N_2}, \tag{S50}$$

where $c_i(T|T_1)$ and $c_i(T|T_2)$ are the citations in year T for paper i published in year T_1 and T_2, respectively. In Fig. S23ab, we compared the IFs measured by Eq. (S50) to the reported value for *Cell* and *NEJM* (Fig. 3) between 1998-2005, finding a good agreement except for small deviations for *NEJM* in 1999 and 2000, which are likely caused by the difference in coverage of journals and the specific definition of 'citable items' by Thomson Reuters.

The proposed model allows us to calculate the journal's impact factor analytically. To do it, we

substitute Eqs. (S48) and (S49) into (S50), obtaining

$$\begin{aligned}\mathrm{IF}(T) &= \frac{N_1 C(T|T_1) + N_2 C(T|T_2)}{N_1 + N_2} \\ &= \frac{mN_1}{N_1+N_2}\left(e^{\Lambda(T_1)\Phi\left(\frac{M_1 - M(T_1)}{\Sigma(T_1)}\right)} - e^{\Lambda(T_1)\Phi\left(\frac{M_3 - M(T_1)}{\Sigma(T_1)}\right)}\right) + \frac{mN_2}{N_1+N_2}\left(e^{\Lambda(T_2)\Phi\left(\frac{M_3 - M(T_2)}{\Sigma(T_2)}\right)} - e^{\Lambda(T_2)\Phi\left(\frac{M_2 - M(T_2)}{\Sigma(T_2)}\right)}\right),\end{aligned} \tag{S51}$$

where $(\Lambda(T_1), M(T_1), \Sigma(T_1))$ and $(\Lambda(T_2), M(T_2), \Sigma(T_2))$ are the journal parameters measured at the year T_1 and T_2, respectively, and $M_1 = \ln(3\ years) = \ln(3\times 365) \approx 7.00$, $M_2 = \ln(1\ year) = \ln(365) \approx 5.90$ and $M_3 = \ln(2\ years) = \ln(2\times 365) \approx 6.59$. Figure S23c documents an excellent match between Eq. (S51) and the empirical measurement based on (S50).

To demonstrate how well our formula of Impact factor (IF) agrees with the empirical results, we show in Table S3 the results of three measurements (Table S3): IF reported by ISI, IF measured in our dataset, and IF computed using our formula (S47).

As illustrated in Fig. S24, IF measured in our dataset occasionally differs from the IF reported by ISI. This is mainly due to the difference in journal coverage and the specific definition of citable items by Thomson Reuters. Our dataset is restricted to journals carried by Northeastern University's subscription to ISI, which is a subset of the full range of sources covered by Thomson Reuters. Furthermore, we used only papers labeled as research articles when computing the IF. Yet, the difference is typically small relative to the magnitude of the IF. To be specific, it is always 10% or smaller, except for *Cell* in 2005. Most important, when we compare the IF measured in our dataset with the IF computed using our framework, we find excellent agreement between the two quantities across different journals and years, vividly demonstrating the accuracy of our approach.

To further simplify (S51), we assume that the changes in papers published by a journal are small over the course of two years, in terms of both number of papers published and their citations. Under this assumption, $N_1 = N_2$ and $(\Lambda, M, \Sigma) \equiv (\Lambda(T_1), M(T_1), \Sigma(T_1)) = (\Lambda(T_2), M(T_2), \Sigma(T_2))$, Eq.

(S51) leads to

$$\mathrm{IF} \approx \frac{m}{2}\left(\exp\left[\Lambda\Phi\left(\frac{M_1 - M}{\Sigma}\right)\right] - \exp\left[\Lambda\Phi\left(\frac{M_2 - M}{\Sigma}\right)\right]\right). \tag{S52}$$

This approximation is not able to account for the temporal evolution in IF, but allows us to compute a journal's IF using only one year of data. To see how well (S52) approximates the reported IF, we use the citation data for the journals published in 1990 and approximate their IF in 1992. We then compare the computed IF by using (S52) with the ones reported by ISI (Fig. S25). We find that despite its simplicity, the two quantities largely agree with each other for different journals, indicating that (S52) and thus Eq. (10) serves as a good approximation for a journal's impact factor.

References

[1] E. Garfield. The history and meaning of the journal impact factor. *JAMA: the journal of the American Medical Association*, 295(1):90–93, 2006.

[2] D.J. de Solla Price. Networks of scientific papers. *Science*, 149(3683):510–515, 1965.

[3] S. Redner. Citation statistics from 110 years of physical review. *Physics Today*, 58:49, 2005.

[4] J.E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569, 2005.

[5] S. Lehmann, A.D. Jackson, and B.E. Lautrup. Measures for measures. *Nature*, 444(7122): 1003–1004, 2006.

[6] B.F. Jones, S. Wuchty, and B. Uzzi. Multi-university research teams: shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[7] F. Radicchi, S. Fortunato, and C. Castellano. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272, 2008.

[8] J.A. Evans and J. Reimer. Open access and global participation in science. *Science*, 323 (5917):1025, 2009.

[9] J.A. Evans and J.G. Foster. Metaknowledge. *Science*, 331(6018):721–725, 2011.

[10] Albert-László Barabási, Chaoming Song, and Dashun Wang. Publishing: Handful of papers dominates citation. *Nature*, 491(7422):40, 2012.

[11] L. Egghe. Theory and practise of the g-index. *Scientometrics*, 69(1):131–152, 2006.

[12] A. Fersht. The most influential journals: Impact factor and eigenfactor. *Proceedings of the National Academy of Sciences*, 106(17):6883–6884, 2009.

[13] F. Radicchi, S. Fortunato, B. Markines, and A. Vespignani. Diffusion of scientific credits and the ranking of scientists. *Physical Review E*, 80(5):056103, 2009.

[14] P.O. Seglen. Why the impact factor of journals should not be used for evaluating research. *BMJ: British Medical Journal*, 314(7079):498, 1997.

[15] M.J. Stringer, M. Sales-Pardo, and L.A.N. Amaral. Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, 3(2):e1683, 02 2008.

[16] T.S. Kuhn. *The structure of scientific revolutions*. University of Chicago press, 1996.

[17] D.E. Acuna, S. Allesina, and K.P. Kording. Future impact: Predicting scientific success. *Nature*, 489(7415):201–202, 2012.

[18] G.J. Peterson, S. Pressé, and K.A. Dill. Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences*, 107 (37):16023–16027, 2010.

[19] S. Redner. How popular is your paper? an empirical study of the citation distribution. *The European Physical Journal B*, 4(2):131–134, 1998.

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

[21] G. Caldarelli. *Scale-Free Networks*. Oxford University Press, 2007.

[22] M. Medo, G. Cimini, and S. Gualdi. Temporal effects in the growth of networks. *Physical Review Letters*, 107(23):238701, 2011.

[23] S.N. Dorogovtsev and J.F.F. Mendes. *Evolution of networks: From biological nets to the Internet and WWW*. Oxford, 2003.

[24] G. Bianconi and A.-L. Barabási. Competition and multiscaling in evolving networks. *EPL (Europhysics Letters)*, 54:436, 2001.

[25] G. Caldarelli, A. Capocci, P. De Los Rios, and M.A. Muñoz. Scale-free networks from varying vertex intrinsic fitness. *Physical Review Letters*, 89(25):258702, 2002.

[26] V. Mahajan, E. Muller, and F.M. Bass. New product diffusion models in marketing: A review and directions for research. *The Journal of Marketing*, pages 1–26, 1990.

[27] F.M. Bass. Comments on "a new product growth for model consumer durables the bass mode". *Management science*, 50(12):1833–1840, 2004.

[28] Benjamin Gompertz. On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philosophical transactions of the Royal Society of London*, 115:513–583, 1825.

[29] Y.H. Eom and S. Fortunato. Characterizing and modeling citation dynamics. *PloS one*, 6(9): e24926, 2011.

[30] I. Fuyuno and D. Cyranoski. Cash for papers: Putting a premium on publication. *Nature*, 441(7095):792, 2006.

[31] D.J. de Solla Price. *Little Science, Big Science... and Beyond*. Columbia University, 1963.

[32] D.T. Durack. The weight of medical knowledge. *New England Journal of Medicine*, 298 (14):773–775, 1978.

[33] W.H. Suh, K.S. Suslick, G.D. Stucky, and Y.H. Suh. Nanotechnology, nanotoxicology, and neuroscience. *Progress in neurobiology*, 87(3):133–170, 2009.

[34] N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J.P. Bouchaud, and A.L. Barabási. Dynamics of ranking processes in complex systems. *Physical Review Letters*, 109(12): 128701, 2012.

[35] John W. Boag. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society. Series B (Methodological)*, 11(1):pp. 15–53, 1949. ISSN 00359246.

[36] Philip E. Sartwell. The distribution of incubation periods of infectious disease. *American Journal of Epidemiology*, 51(3):310–318, 1950.

[37] FW Preston. Pseudo-lognormal distributions. *Ecology*, pages 355–364, 1981.

[38] CB Williams. A note on the statistical analysls of sentence-length as a criterion of literary style. *Biometrika*, 31(3-4):356–361, 1940.

[39] G. Herdan. The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika*, 45 (1-2):222–228, 1958.

[40] R. Plamondon. A kinematic theory of rapid human movements. *Biological cybernetics*, 72 (4):295–307, 1995.

[41] G.J.P. VanBreukelen. Parallel information processing models compatible with lognormally distributed response times. *Journal of Mathematical Psychology*, 39(4):396–399, 1995.

[42] R. Ulrich and J. Miller. Information processing models generating lognormally distributed reaction times. *Journal of Mathematical Psychology*, 1993.

[43] William Templeton Eadie and F James. *Statistical methods in experimental physics*. World Scientific Publishing Company, 2006.

[44] H.A. Simon. On a class of skew distribution functions. *Biometrika*, pages 425–440, 1955.

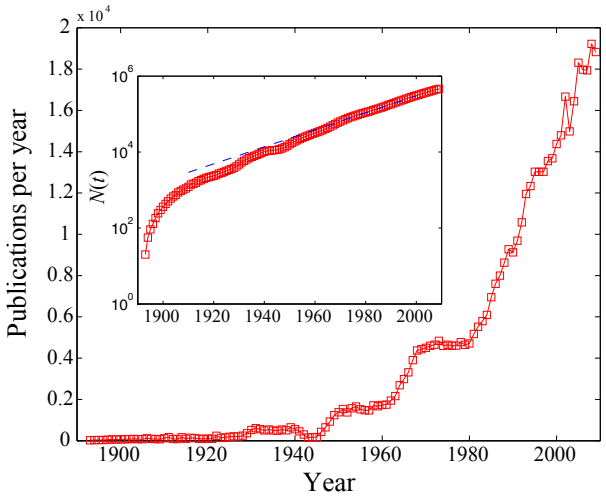

Figure S1: The number of papers published each year in the PR corpus. Inset: cumulative number of papers $N(t)$ published up to year t.

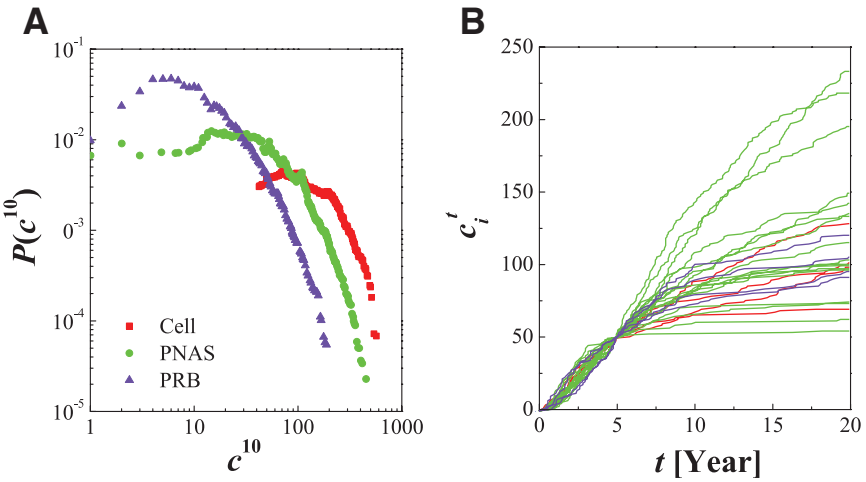

Figure S2: **Impact factor and the number of citations lacks long-term predictability**. (A) Distribution of the cumulative citations ten years after publication (c^{10}) for all papers published in *Cell*, *PNAS*, and *Physical Review B* (*PRB*) in 1990. (B) Citation history of all papers shown in (A) that acquired 50 citations 5 years after publication, illustrating the different long-term impact despite their equal early impact.

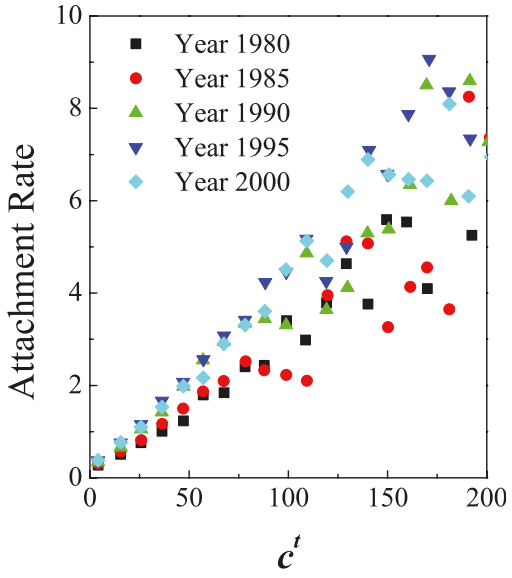

Figure S3: **Empirical validation of preferential attachment**. Attachment rate measures the likelihood for new papers published in different years (color coded) to cite an old paper with c^t citations. That is, for each year, c^t measures the citations of each paper before this year, and attachment rate measures the average number of times each paper with c^t citations was cited in this year. The linearity of the curves offers evidence for preferential attachment. [3]

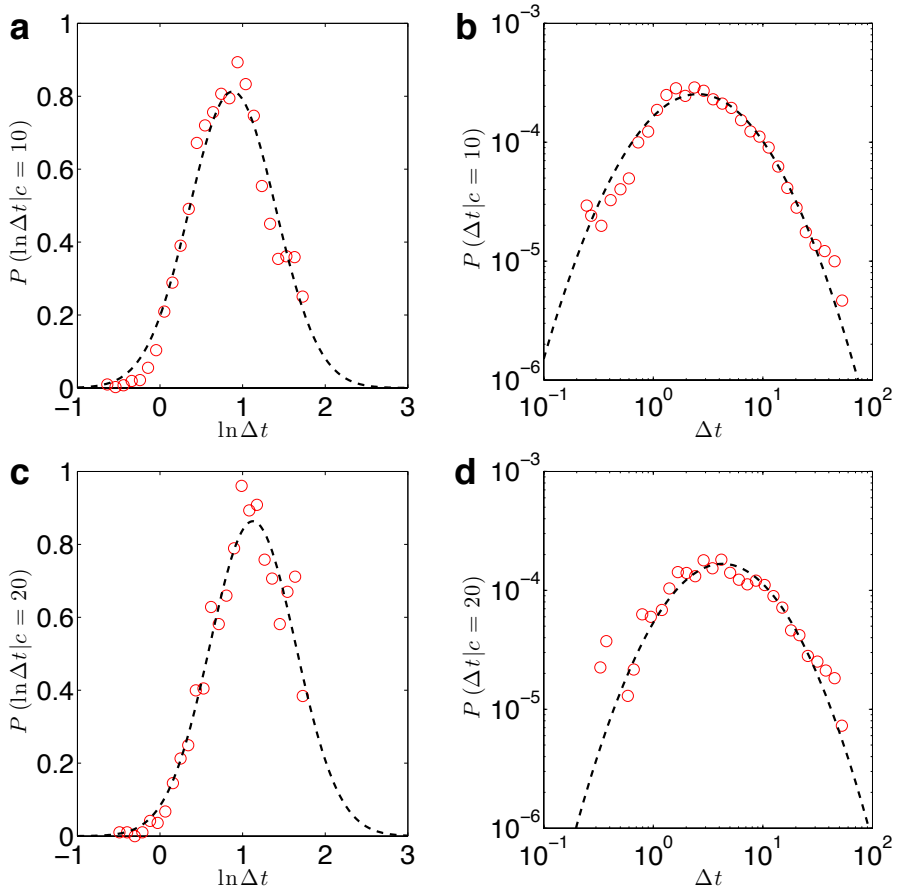

Figure S4: **Empirical validation of the lognormal decay (S3)** (a) $P(\ln\Delta t)$ when papers change from 10 citations to 11 citations. The dashed line corresponds to the best gaussian fitting. (b) Same as (a) but for $P(\Delta t)$. Dashed line corresponds to the best lognormal fitting ($\mu = 7.85$ and $\sigma = 1.01$). Here Δt is measured in unit of years. (c) $P(\ln\Delta t)$ when papers change from 20 citations to 21 citations. The dashed line corresponds to the best gaussian fitting. (d) Same as (c) but for $P(\Delta t)$. The dashed line corresponds to the best lognormal fitting ($\mu = 8.29$ and $\sigma = 0.93$).

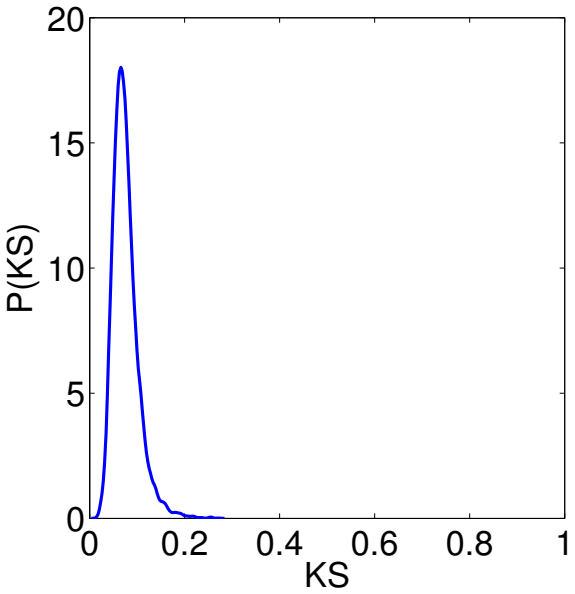

Figure S5: **Statistical test of the data collapse.** Kolmogorov-Smirnov (KS) measure for papers shown in Fig. 1J.

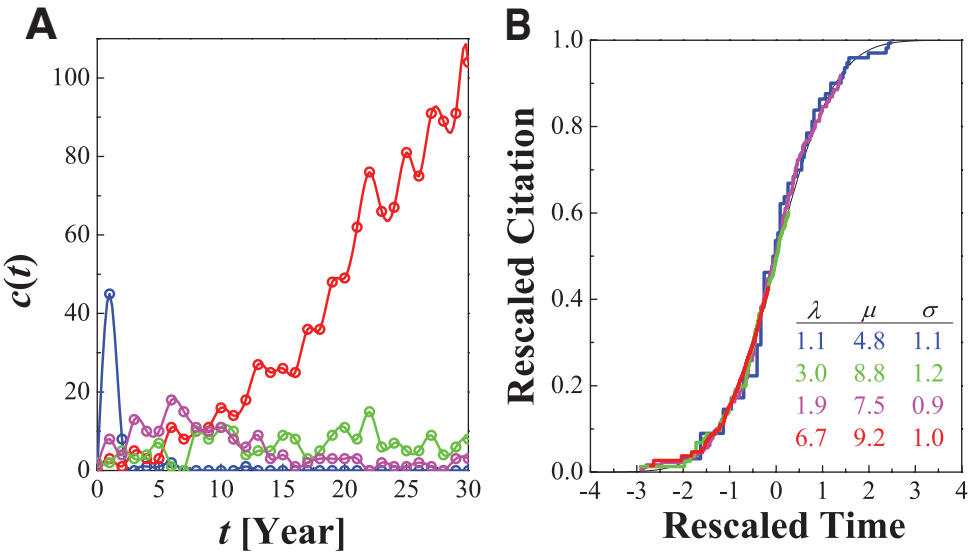

Figure S6: **Demonstration of rescaling for four papers**. (A) Citation history of four papers published in PR in 1964, selected for their distinct dynamics, displaying a 'jump-decay' pattern (blue); delayed peak (magenta); attracting a constant number of citation over time (green), or acquiring an increasing number of citations each year (red). (B) Data collapse for the four papers in (A) using Eq. (5). Legend: the (λ, μ, σ) parameters used to rescale the citation history of each paper.

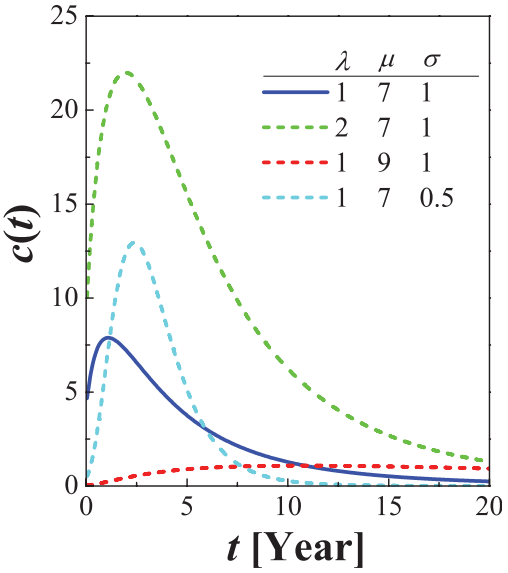

Figure S7: Changes in the citation history $c(t)$ according to (3) after varying the (λ, μ, σ) parameters, indicating that (S13) can account for a wide range of citation patterns.

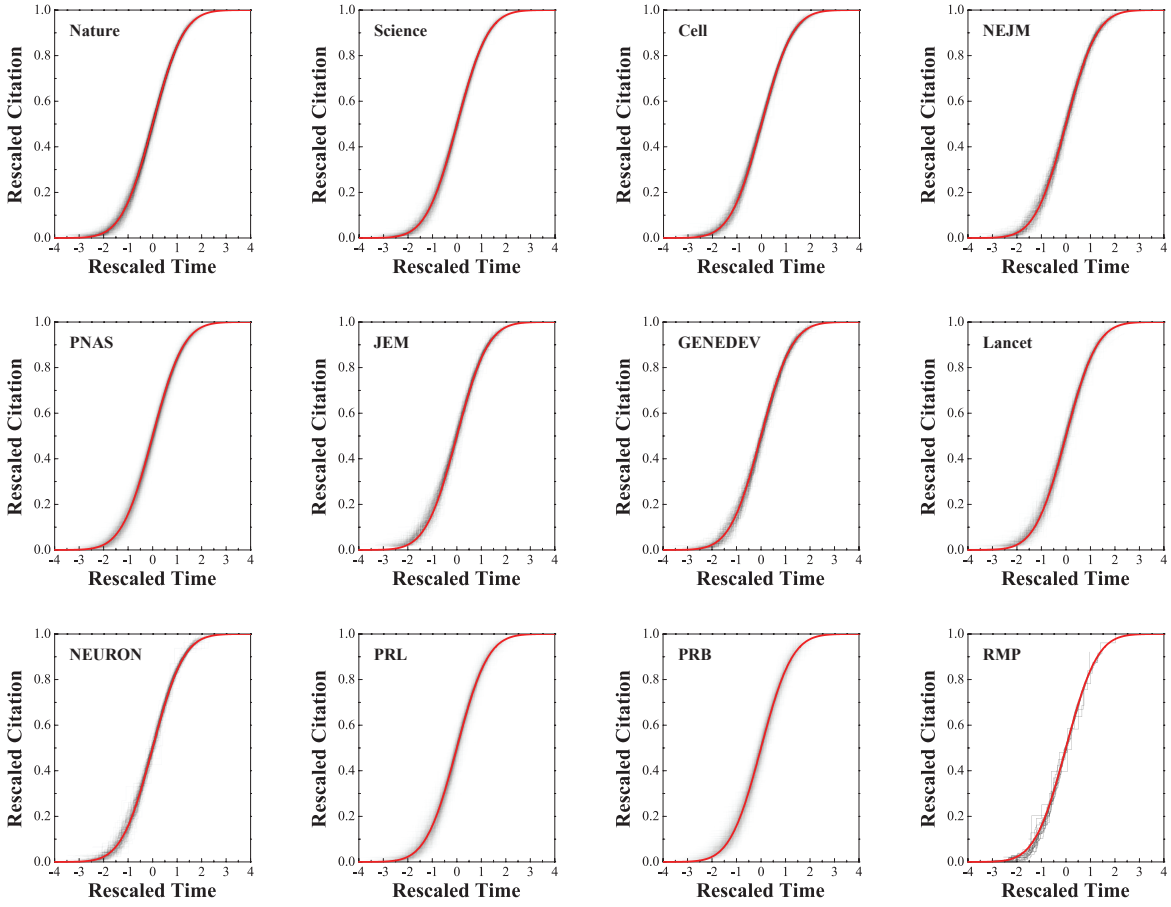

Figure S8: **Data collapse for papers published in several journals**. We applied the model to all papers published by the 12 journals listed in Table S4 in 1990, finding consistently an excellent agreement between model and empirical citation dynamics for journals across different discipline. The number of papers in each dataset is summarized in Table S4.

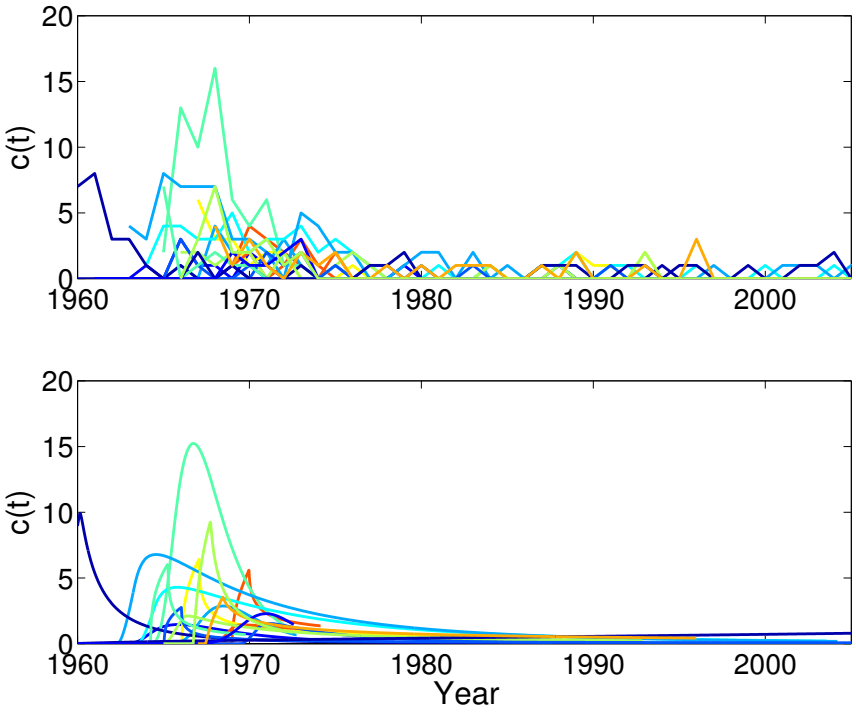

Figure S9: **Simulating individual citation histories.** We randomly selected two papers each year between 1960 to 1970 from the PR corpus. Their citation histories are shown on the top panel. Color code is the same as Fig. 1D, corresponding to the publication year. We estimated the set of (λ, μ, σ) parameters for each paper using the methods described in Section S2.3. The bottom panel shows the citation dynamics predicted by Eq. (S13).

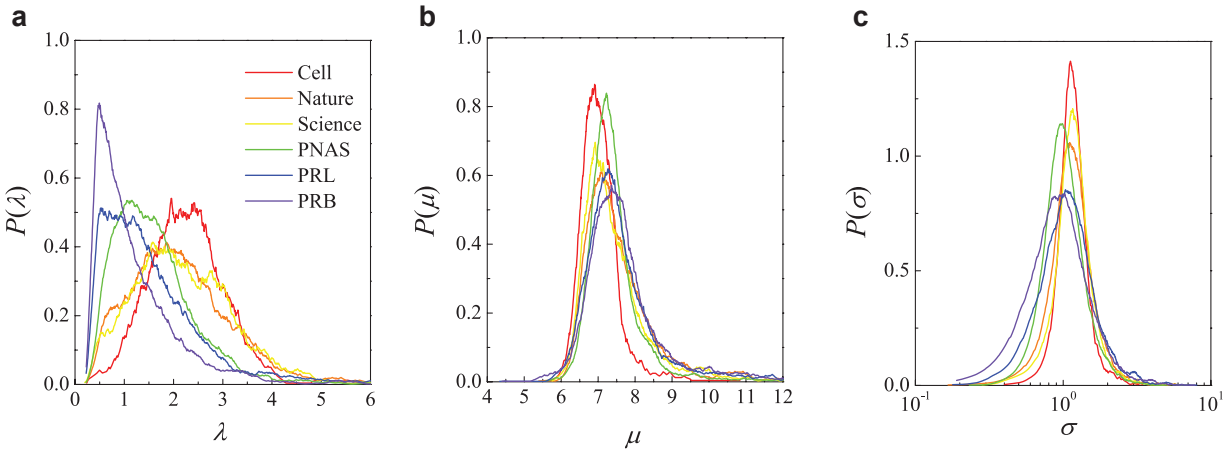

Figure S10: **Parameter distributions for papers published by six journals in 1990**. (a) Fitness distributions are radically different for different journals. (b) Immediacy distributions show modest differences: *Cell* has the smallest average μ among the 6 journals, while the average immediacy of *PRB* is the largest. (c) Longevity distributions of these journals are characterized by similar mean values.

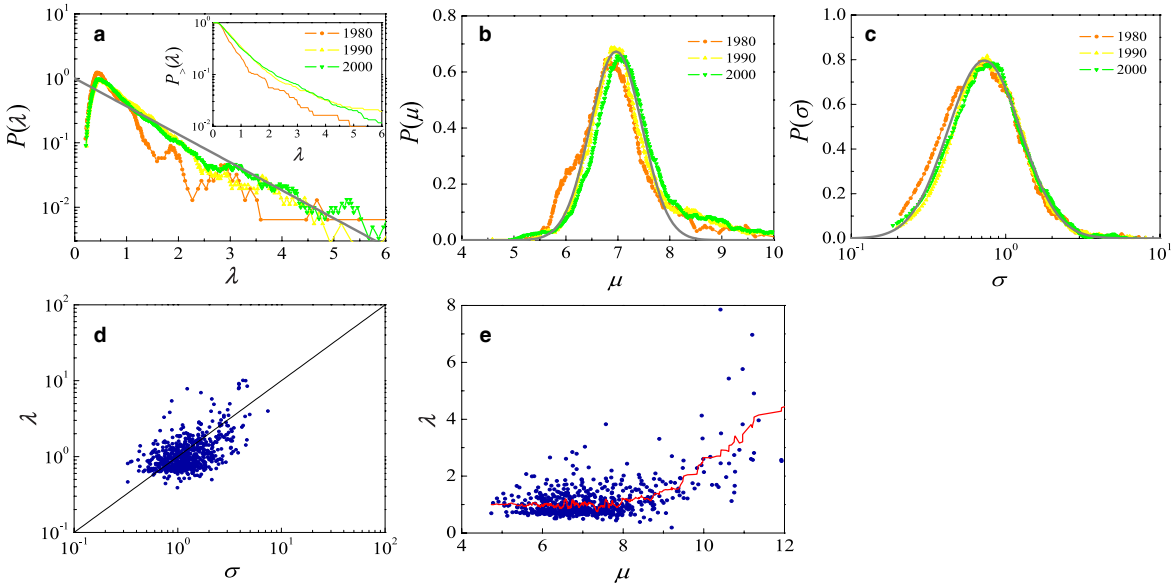

Figure S11: **Temporal evolution of the parameter distributions and correlations between them.** (a) Fitness distribution for papers published in different years (1980, 1990, and 2000) within the PR corpus. (b) Immediacy distributions for papers shown in (a). (c) Same, but for longevity distributions. (d) We observe a weak linear correlation between λ and σ, indicating higher fitness papers tend to have larger longevity. (e) λ and μ are largely uncorrelated, except in the large λ region, indicating that papers with very high fitness are also characterized by a delayed impact.

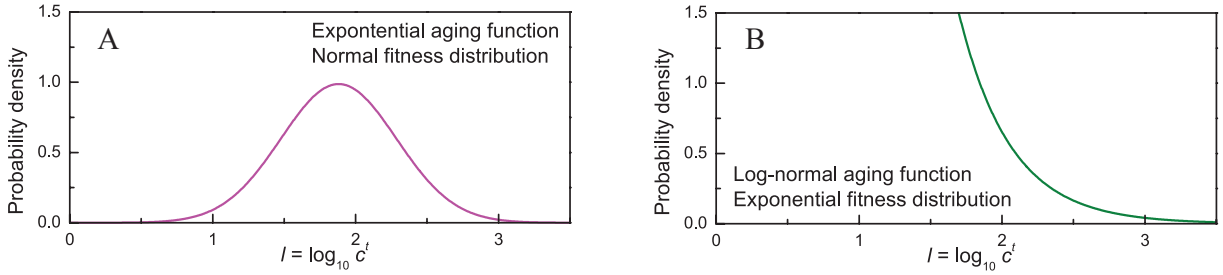

Figure S12: **Untangling lognormal citation distribution and lognormal temporal decay.** (A) To show the ageing function is not rooted in the citation distribution, we remove replace the lognormal ageing function with an exponential form, while keeping the fitness distribution unchanged as normal distribution, the citations still follow lognormal distribution. (B) If we change the fitness distribution to exponential form, and keep the lognormal ageing function, the citations no longer follow lognormal, but power law distribution.

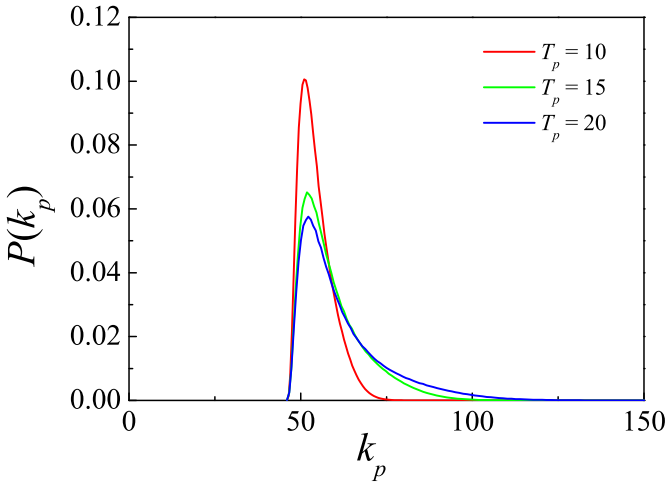

Figure S13: Illustrative example of $P(k_p)$ for a randomly selected paper. Different lines correspond to different testing period (T_p).

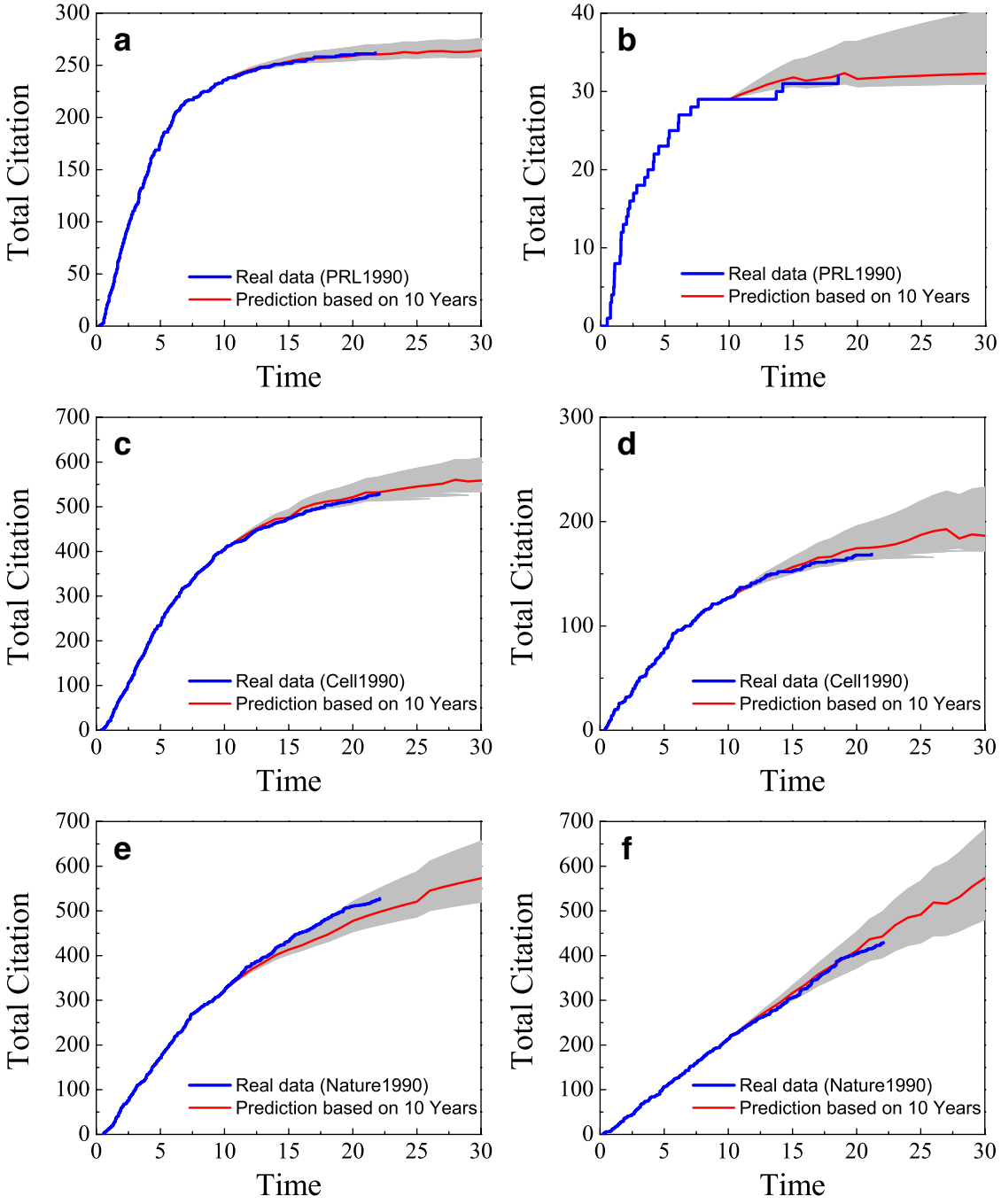

Figure S14: Citation predictions using the proposed model for 6 papers randomly selected form three different journals (*PRL*, *Cell*, and *Nature*).

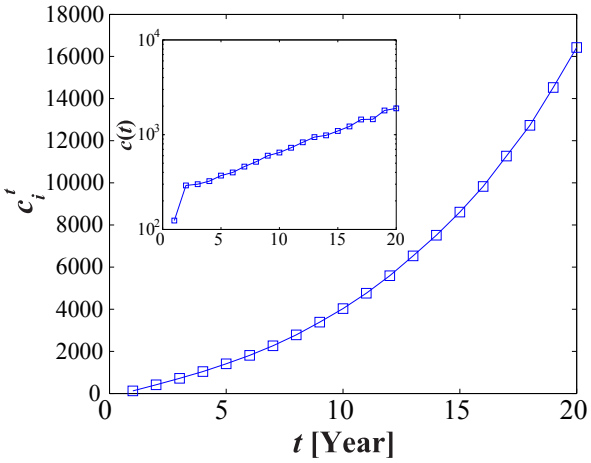

Figure S15: **Simulation results for the scale-free model**. We simulate a scale-free network with 100,000 nodes. Each node is associated with a time stamp such that the number of nodes in each unit time grows exponentially, following Eq. (S1). We group together the nodes with the same time stamps, in analogy to papers published in the same year, and explore how their degrees evolve over time. Inset: the new links acquired by the selected nodes in each time step shown on a log-linear scale, demonstrating the exponential nature of the growth curve, as predicted by Eq. (S34).

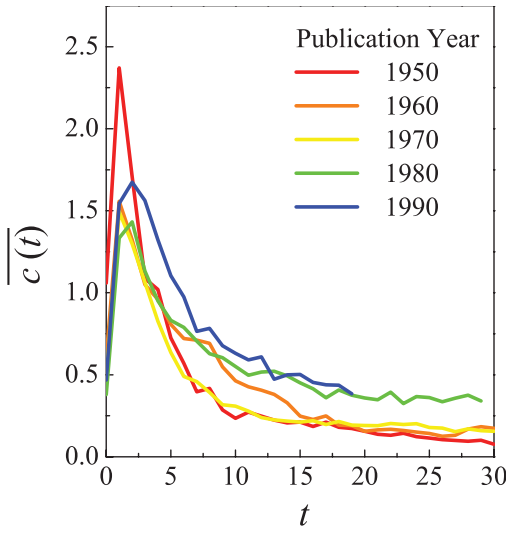

Figure S16: Average citations for papers published in same year in the Physical Review (PR) corpus, demonstrating the typical jump-decay citation pattern of citation histories.

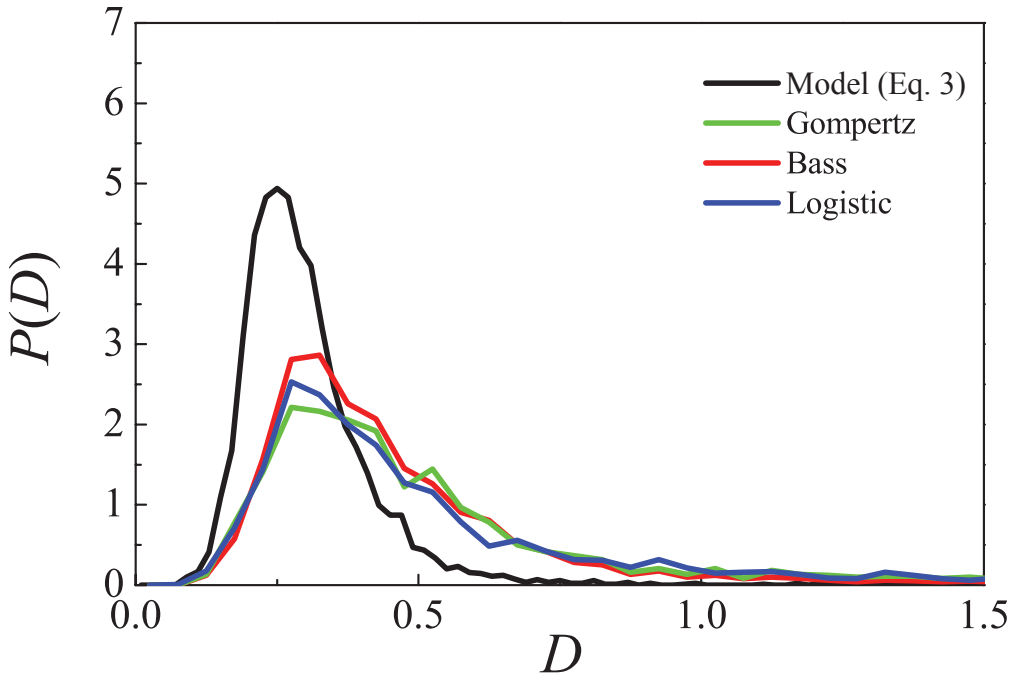

Figure S17: Goodness of fit using weighted Kolmogorov-Smirnov (KS) test.

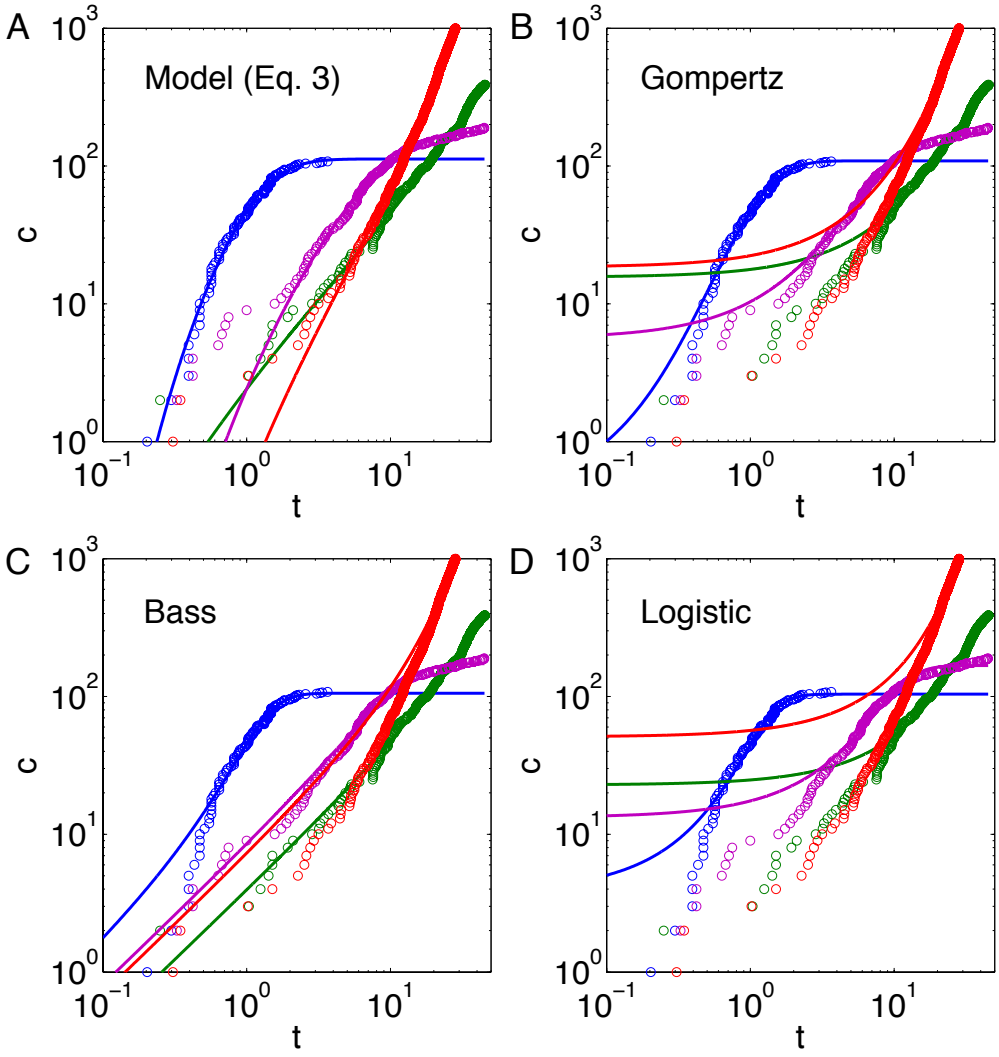

Figure S18: Fitting the four papers in Fig. 1G by using (A) Model (Eq. 3), (B) Gompertz, (C) Bass, and (D) Logistic models.

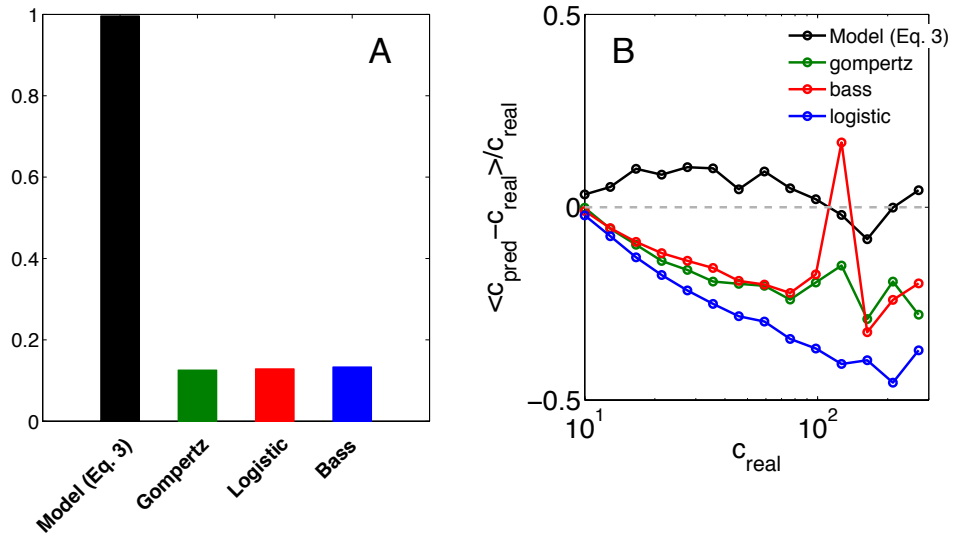

Figure S19: **Comparison between predicted and real citations.** (A) Fraction of points corresponds to the green bars in Fig. 4F. (B) Errors in predicted citations as a function of real citations. We measure the difference between the predicted citations and the real citations, normalized by the real citations, for each group of papers with similar real citations.

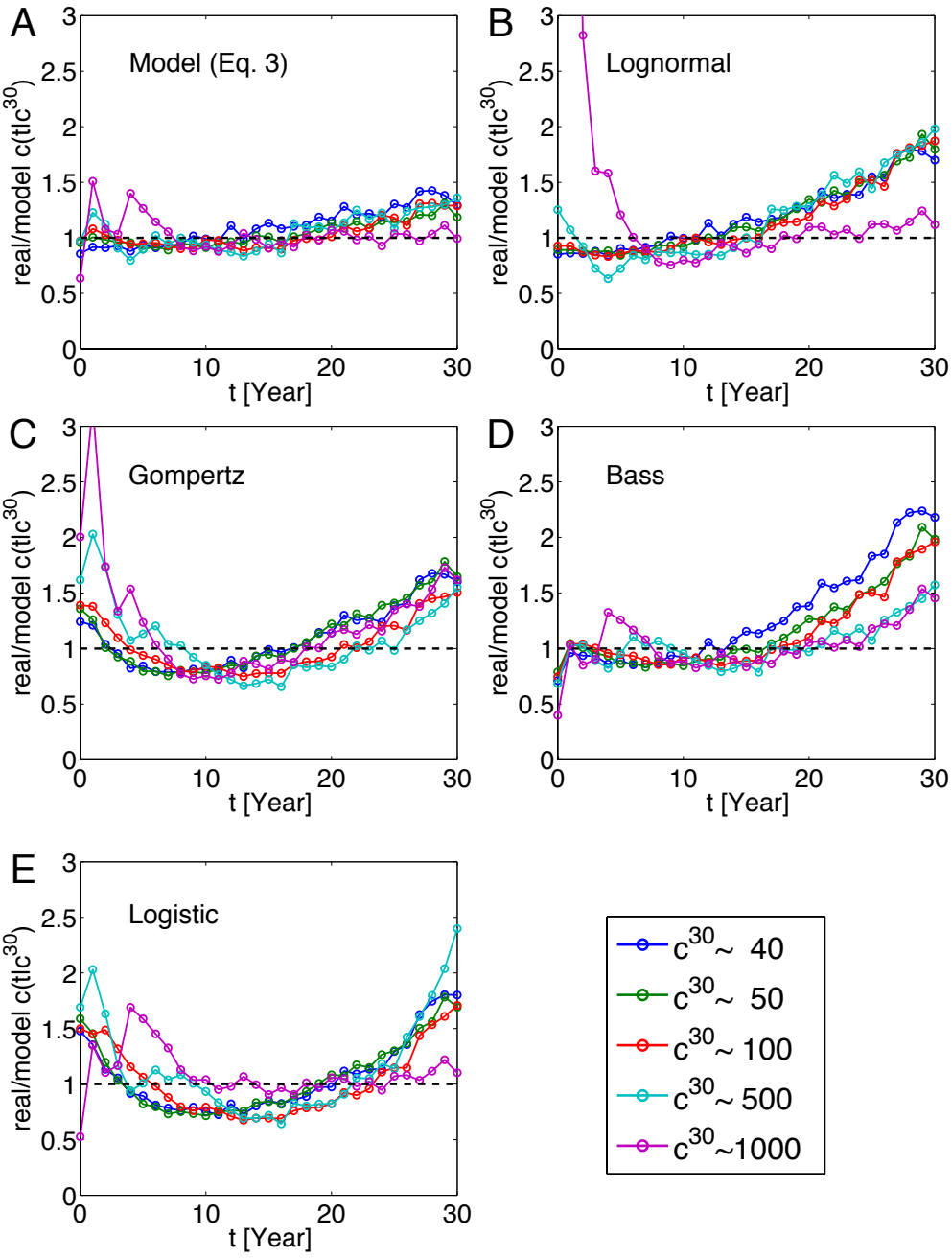

Figure S20: **Comparing model predictions with real citation dynamics for papers with different c^{30}.** We grouped the papers based on their number of citations after 30 years (c^{30}), and measured the average number of citations within each group as a function of the time, $c^{real}(t|c^{30})$. We obtain the best fit for each paper using different models, and compute the average number of citations in each year within groups using the best fitted parameters, $c^{model}(t|c^{30})$. We show the ratio between real citations and those generated by models, $c^{real}(t|c^{30})/c^{model}(t|c^{30})$ in (A—E) for model Eq. 3, Lognormal, Gompertz, Bass, and Logistic model, respectively.

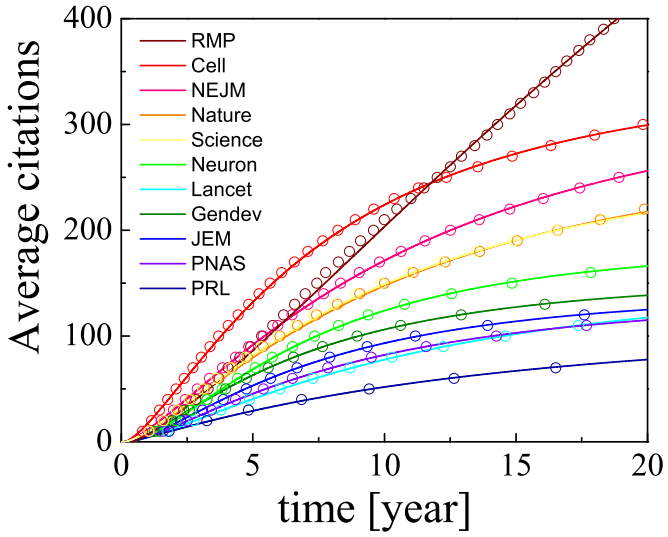

Figure S21: **Average cumulative citations for different journals**. The average number of citations each journal gets to all of its papers published in 1990. Circles correspond to empirically measured citations, and solid lines are based on Eq. (S49)

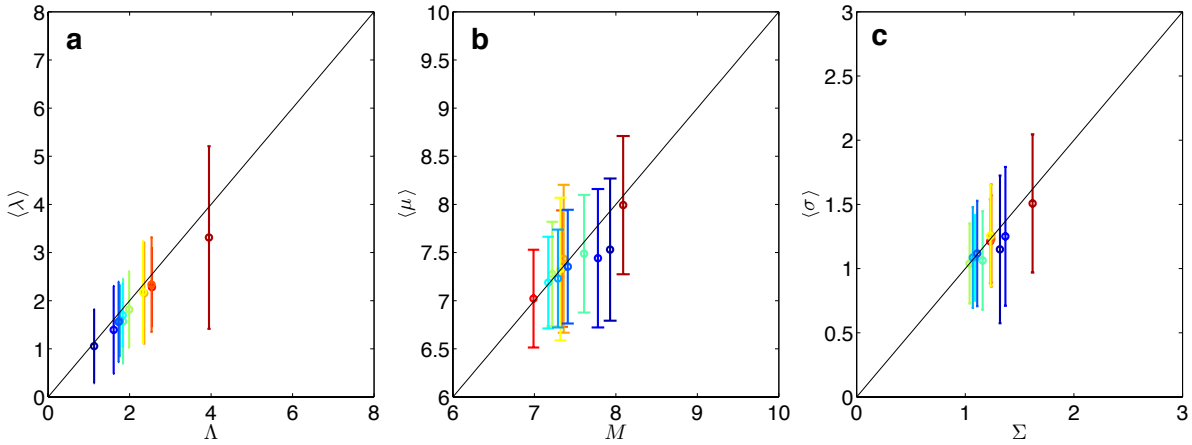

Figure S22: **Comparing journal parameters**. The three parameters characterizing a journal's citation history can be computed in two ways. One is to average over the parameters of individual papers ($\langle\lambda\rangle$, $\langle\mu\rangle$, $\langle\sigma\rangle$). The other is to use the average citation curve (Fig. S21) for a journal (Λ, M, Σ). Here we show an a reasonable agreement between the values offered by these two methods in (a) fitness (b) immediacy (c) longevity.

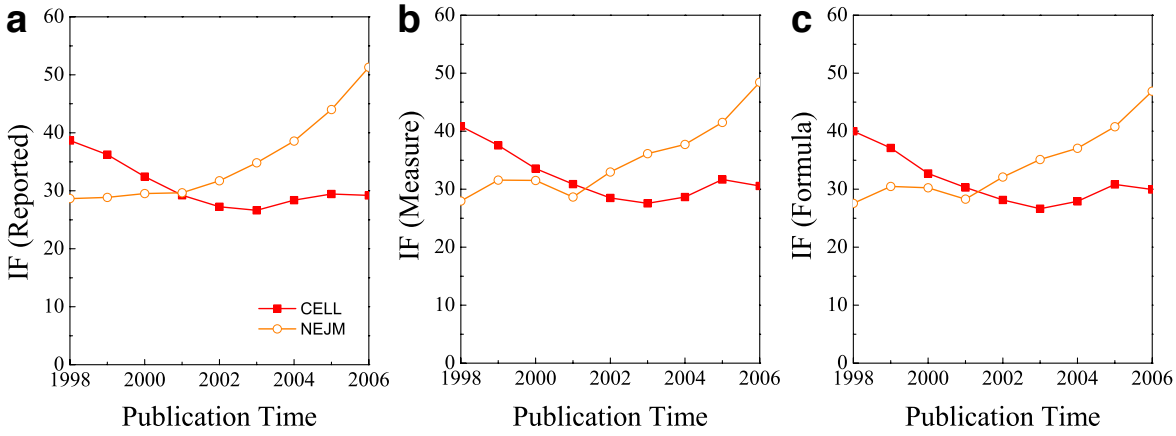

Figure S23: **Comparing** IF **reported by ISI with the (S51) approximation**. (a) IF reported by ISI for *Cell* and *NEJM* from 1998 to 2006. (b) IF measured for these two journals within the time span following the definition of (S50). (c) IF computed by plugging in the corresponding parameters in (S51).

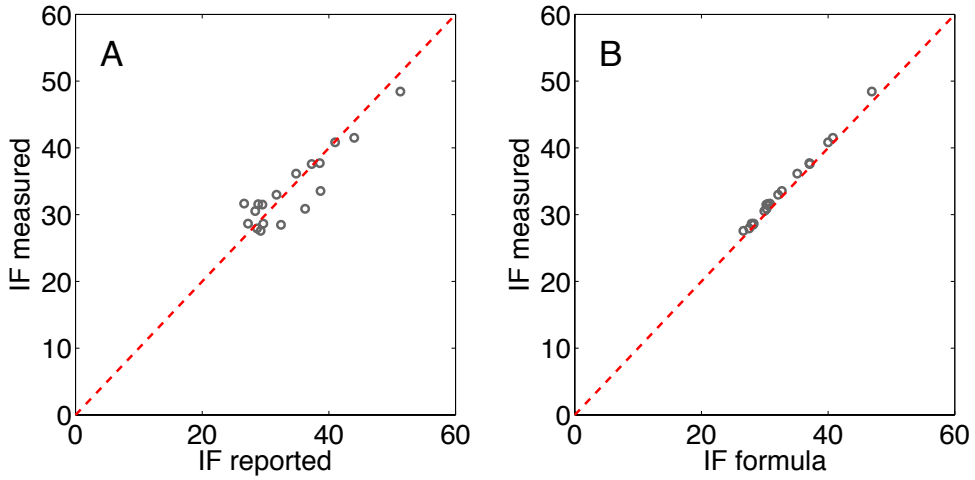

Figure S24: **Comparison between the IF listed in Table S3.** (A) Comparison between the IF reported by ISI and the IF measured by using our dataset. The differences may be due to the subscription coverage difference and the varying definition of 'citable items'. (B) Comparison between the IF computed using our formula (Eq. S47) and the IF measured by using our dataset. We find excellent agreement between these two quantities

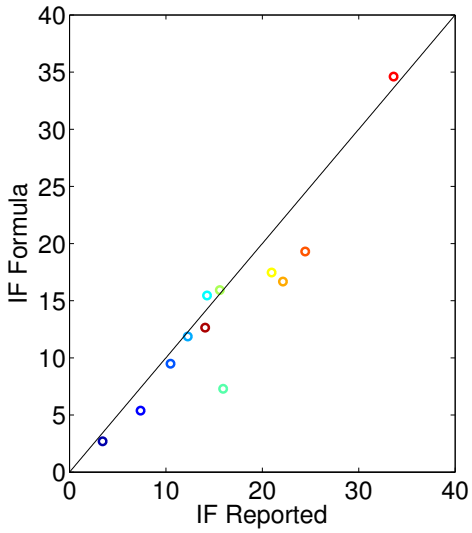

Figure S25: Scatter plot for journals shown in Fig. S9, their reported IF in 1992 and their approximated IF obtained using their parameters in 1990 in (S52).

Table S1: **Statistics of PR Corpus.**

Journal	Start Year	End Year	# Papers
Physical Review (Series I)	1893	1912	1,469
Physical Review	1913	1969	47,941
Reviews of Modern Physics	1929	2009	2,926
Physical Review Letters	1958	2009	95,516
Physical Review A	1970	2009	53,655
Physical Review B	1970	2009	137,999
Physical Review C	1970	2009	29,935
Physical Review D	1970	2009	56,616
Physical Review E	1993	2009	35,944
Physical Review Special Topics - Accelerators and Beams	2002	2009	1,257
Physical Review Special Topics - Physics Education Research	2005	2009	90

Table S2: **Modeling citation dynamics**. We identified three models that can be or have been used to fit citation histories. The table shows the corresponding rate equation and its analytical solution.

Model Name	**Rate Equation**	**Solution**
Minimal Citation Model, Eq. (S13)	$\frac{dc_i^t}{dt} \approx c_i^t \eta_i P(t)$	$c_i^t = m\left(e^{\lambda_i \Phi(\frac{\ln t - \mu_i}{\sigma_i})} - 1\right)$
Logistic [26]	$\frac{dc_i^t}{dt} = r_i c_i^t \left(1 - c_i^t / c^\infty\right)$	$c_i^t = \frac{c^\infty}{1+e^{-r_i(t-\tau_i)}}$
Bass [27]	$\frac{dc_i^t}{dt} = (p + q c_i^t / c^\infty)(c^\infty - c_i^t)$	$c_i^t = c^\infty \frac{1-e^{-(p+q)t}}{1+\frac{q}{p}e^{-(p+q)t}}$
Gompertz [28, 26]	$\frac{dc_i^t}{dt} = q c_i^t \ln(c^\infty / c_i^t)$	$c_i^t = c^\infty e^{-e^{-(a+qt)}}$

Table S3: **Comparison of Impact Factors (IF) for Cell and NEJM from 1998 to 2006.** The last three columns from left to right are: IF reported by ISI, IF measured in our dataset, and IF computed using our formula (S47)

Journal	YEAR	REPORTED	MEASURED	FORMULA
NEJM	1998	28.66	27.9478	27.54758
NEJM	1999	28.857	31.578	30.48742
NEJM	2000	29.512	31.5043	30.23374
NEJM	2001	29.65	28.6434	28.27551
NEJM	2002	31.736	32.971	32.10337
NEJM	2003	34.833	36.1324	35.12369
NEJM	2004	38.57	37.7287	37.04507
NEJM	2005	44.016	41.509	40.75732
NEJM	2006	51.296	48.424	46.90004
CELL	1998	40.997	40.8498	40.00453
CELL	1999	37.297	37.5698	37.10045
CELL	2000	38.686	33.5433	32.70972
CELL	2001	36.242	30.8712	30.30176
CELL	2002	32.44	28.4807	28.1381
CELL	2003	29.219	27.5625	26.62391
CELL	2004	27.254	28.6371	27.91594
CELL	2005	26.626	31.6619	30.81807
CELL	2006	28.389	30.5479	29.94492

Table S4: **Citation statistics for 11 non-review journals and one review journal in 1990.** In line with citation items in definition of IF, we only include here reviews and articles. For 11 non-review journals, highest Λ, M and Σ are in bold faces. C^{∞} is obtained by $C^{\infty} = m\left(e^{\Lambda} - 1\right)$.

Journal	Year	# Papers	Λ	M	Σ	C^{∞}
Cell	1990	485	**2.55**	6.99	1.23	354
NEJM	1990	330	2.54	7.34	1.24	350
Nature	1990	1,099	2.36	7.36	1.24	289
Science	1990	842	2.33	7.32	1.23	280
Neuron	1990	178	1.99	7.22	1.04	189
Lancet	1990	541	1.84	7.61	1.16	159
Gene-Dev	1990	200	1.83	7.17	1.09	157
JEM	1990	313	1.76	7.29	1.07	144
PNAS	1990	2,060	1.73	7.41	1.11	140
PRL	1990	1,633	1.61	7.78	**1.37**	121
PRB	1990	2,189	1.13	**7.93**	1.32	63
RMP	1990	18	3.95	8.09	1.62	1535

References and Notes

1. E. Garfield, The history and meaning of the journal impact factor. *JAMA* **295**, 90–93 (2006). Medline doi:10.1001/jama.295.1.90

2. D. J. Price, Networks of scientific papers. *Science* **149**, 510–515 (1965). Medline doi:10.1126/science.149.3683.510

3. S. Redner, Citation statistics from 110 years of physical review. *Phys. Today* **58**, 49 (2005). doi:10.1063/1.1996475

4. J. E. Hirsch, An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16569–16572 (2005). Medline doi:10.1073/pnas.0507655102

5. S. Lehmann, A. D. Jackson, B. E. Lautrup, Measures for measures. *Nature* **444**, 1003–1004 (2006). Medline doi:10.1038/4441003a

6. B. F. Jones, S. Wuchty, B. Uzzi, Multi-university research teams: Shifting impact, geography, and stratification in science. *Science* **322**, 1259–1262 (2008); 10.1126/science1158357. Medline doi:10.1126/science.1158357

7. F. Radicchi, S. Fortunato, C. Castellano, Universality of citation distributions: Toward an objective measure of scientific impact. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 17268–17272 (2008). Medline doi:10.1073/pnas.0806977105

8. J. A. Evans, J. Reimer, Open access and global participation in science. *Science* **323**, 1025 (2009). Medline doi:10.1126/science.1154562

9. J. A. Evans, J. G. Foster, Metaknowledge. *Science* **331**, 721–725 (2011). Medline doi:10.1126/science.1201765

10. A.-L. Barabási, C. Song, D. Wang, Publishing: Handful of papers dominates citation. *Nature* **491**, 40 (2012). Medline doi:10.1038/491040a

11. L. Egghe, Theory and practise of the g-index. *Scientometrics* **69**, 131–152 (2006). doi:10.1007/s11192-006-0144-7

12. A. Fersht, The most influential journals: Impact Factor and Eigenfactor. *Proc. Natl. Acad. Sci. U.S.A.* **106**, 6883–6884 (2009). Medline doi:10.1073/pnas.0903307106

13. F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, Diffusion of scientific credits and the ranking of scientists. *Phys. Rev. E* **80**, 056103 (2009). Medline doi:10.1103/PhysRevE.80.056103

14. P. O. Seglen, Why the impact factor of journals should not be used for evaluating research. *BMJ* **314**, 498–502 (1997). Medline doi:10.1136/bmj.314.7079.497

15. M. J. Stringer, M. Sales-Pardo, L. A. Nunes Amaral, Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE* **3**, e1683 (2008). Medline doi:10.1371/journal.pone.0001683

16. T. S. Kuhn, *The Structure of Scientific Revolutions* (University of Chicago Press, Chicago, IL, 1996).

17. D. E. Acuna, S. Allesina, K. P. Kording, Future impact: Predicting scientific success. *Nature* **489**, 201–202 (2012). Medline doi:10.1038/489201a

18. G. J. Peterson, S. Pressé, K. A. Dill, Nonuniversal power law scaling in the probability distribution of scientific citations. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 16023–16027 (2010). Medline doi:10.1073/pnas.1010757107

19. S. Redner, How popular is your paper? an empirical study of the citation distribution. *Eur. Phys. J. B* **4**, 131–134 (1998). doi:10.1007/s100510050359

20. A.-L. Barabási, R. Albert, Emergence of scaling in random networks. *Science* **286**, 509–512 (1999). Medline doi:10.1126/science.286.5439.509

21. G. Caldarelli, *Scale-Free Networks* (Oxford Univ. Press, Oxford, 2007).

22. M. Medo, G. Cimini, S. Gualdi, Temporal effects in the growth of networks. *Phys. Rev. Lett.* **107**, 238701 (2011). Medline doi:10.1103/PhysRevLett.107.238701

23. S. N. Dorogovtsev, J. F. F. Mendes, *Evolution of Networks: From Biological Nets to the Internet and WWW* (Oxford Univ. Press, Oxford, 2003).

24. G. Bianconi, A.-L. Barabási, Competition and multiscaling in evolving networks. *Europhys. Lett.* **54**, 436–442 (2001). doi:10.1209/epl/i2001-00260-6

25. G. Caldarelli, A. Capocci, P. De Los Rios, M. A. Muñoz, Scale-free networks from varying vertex intrinsic fitness. *Phys. Rev. Lett.* **89**, 258702 (2002). Medline doi:10.1103/PhysRevLett.89.258702

26. V. Mahajan, E. Muller, F. M. Bass, New product diffusion models in marketing: A review and directions for research. *J. Mark.* **54**, 1–26 (1990). doi:10.2307/1252170

27. F. M. Bass, Comments on "a new product growth for model consumer durables the bass mode". *Manage. Sci.* **50** (suppl.), 1833–1840 (2004). doi:10.1287/mnsc.1040.0300

28. B. Gompertz, On the nature of the function expressive of the law of human mortality, and on a new mode of determining the value of life contingencies. *Philos. Trans. R. Soc. London* **115**, 513–583 (1825). doi:10.1098/rstl.1825.0026

29. Y. H. Eom, S. Fortunato, Characterizing and modeling citation dynamics. *PLoS ONE* **6**, e24926 (2011). Medline doi:10.1371/journal.pone.0024926

30. I. Fuyuno, D. Cyranoski, Cash for papers: Putting a premium on publication. *Nature* **441**, 792 (2006). Medline doi:10.1038/441792b

31. D. J. de Solla Price, *Little Science, Big Science... and Beyond* (Columbia Univ. Press, New York, 1963).

32. D. T. Durack, The weight of medical knowledge. *N. Engl. J. Med.* **298**, 773–775 (1978). Medline doi:10.1056/NEJM197804062981405

33. W. H. Suh, K. S. Suslick, G. D. Stucky, Y. H. Suh, Nanotechnology, nanotoxicology, and neuroscience. *Prog. Neurobiol.* **87**, 133–170 (2009). Medline doi:10.1016/j.pneurobio.2008.09.009

34. N. Blumm, G. Ghoshal, Z. Forró, M. Schich, G. Bianconi, J. P. Bouchaud, A. L. Barabási, Dynamics of ranking processes in complex systems. *Phys. Rev. Lett.* **109**, 128701 (2012). Medline doi:10.1103/PhysRevLett.109.128701

35. J. W. Boag, Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *J. R. Stat. Soc. B* **11**, 15–53 (1949).

36. P. E. Sartwell, The distribution of incubation periods of infectious disease. *Am. J. Epidemiol.* **51**, 310–318 (1950).

37. F. W. Preston, Pseudo-lognormal distributions. *Ecology* **62**, 355–364 (1981). doi:10.2307/1936710

38. C. B. Williams, A note on the statistical analysls of sentence-length as a criterion of literary style. *Biometrika* **31**, 356–361 (1940).

39. G. Herdan, The relation between the dictionary distribution and the occurrence distribution of word length and its importance for the study of quantitative linguistics. *Biometrika* **45**, 222–228 (1958).

40. R. Plamondon, A kinematic theory of rapid human movements. Part I. Movement representation and generation. *Biol. Cybern.* **72**, 295–307 (1995). Medline doi:10.1007/BF00202785

41. G. J. P. VanBreukelen, Parallel information processing models compatible with lognormally distributed response times. *J. Math. Psychol.* **39**, 396–399 (1995). doi:10.1006/jmps.1995.1037

42. R. Ulrich, J. Miller, Information processing models generating lognormally distributed reaction times. *J. Math. Psychol.* **37**, 513–525 (1993). doi:10.1006/jmps.1993.1032

43. W. T. Eadie, F. James, *Statistical Methods in Experimental Physics* (World Scientific Publishing, Singapore, 2006).

44. H. A. Simon, On a class of skew distribution functions. *Biometrika* **42**, 425–440 (1955).