Katchanov and Markova *SpringerPlus* (2015) 4:94
DOI 10.1186/s40064-015-0821-1

SpringerPlus
a SpringerOpen Journal

RESEARCH Open Access

On a heuristic point of view concerning the citation distribution: introducing the Wakeby distribution

Yurij L Katchanov[1]* and Yulia V Markova[2]

Abstract

The paper proposes a heuristic approach to modeling the cumulative distribution of citations of papers in scientific journals by means of the Wakeby distribution. The Markov process of citation leading to the Wakeby distribution is analyzed using the terminal time formalism. The Wakeby distribution is derived in the paper from the simple and general inhomogeneous Choquet – Deny convolution equation for a non-probability measure. We give statistical evidence that the Wakeby distribution is a reasonable approximation of the empirical citation distributions.

AMS Subject Classification: 91D30; 91D99

Keywords: Bibliometrics; Choquet – Deny convolution equation; Citation distributions; Wakeby distribution

Introduction

The number $N(z)$ of scientific papers that has been cited a total of z times is one of the most widely used and strong scientometric indicators. Alternatively, one may consider more sophisticated indicators (see, e.g., (Glänzel and Moed 2013; Leydesdorff et al. 2011, 2013; Radicchi and Castellano 2011; Waltman and van Eck 2013)), but we limit ourselves here to the case in which the underlying variables are defined as the non-negative real numbers z and $N(z)$. This approach has a formal defect that can be easily recognized. As a matter of fact, z and $N(z)$ assume only non-negative integer values. Yet a substantial amount of previous works on the statistical distribution of citations of scientific papers treated z and $N(z)$ as continuous variables in the long-time limit of the observation period, and we pursue the same approach with this paper.

Many problems of Science are describable in terms of a probability distribution. The distribution of citations over papers is important in that it connects more theoretically grounded studies with more practical problems of scientometrics (DeBellis 2009; Moed 2005). Hence, there is a great deal of literature on the distribution of citations to papers in scientific journals. The programmatic article by Lotka 1926 in 1926 was the pioneer paper in scientometric research and continues to be much in demand (see, e.g., (Egghe and Rousseau 2012)). In 1957, Shockley achieved encouraging results (Shockley 1957). Later, in 1965, de Solla Price demonstrated that the citation distribution of scientific papers has strong skewness and heavy tail 1965, and since that time, significant effort has been invested in the study of the citation distribution. De Solla Price explained this "skew distribution" in terms of the cumulative advantage principle (de Solla Price 1976): the probability that a paper will be cited grows with the number of citations it has already received. More precisely, in terms of the probability density function $f(\cdot)$, the cumulative advantage model predicts the following distribution of citations of scientific papers

$$f(z) = \frac{B(z+m,l)}{B(m,l-1)}. \tag{1}$$

Here z indicates the number of citations, $B(\cdot,\cdot)$ is the beta function, and m, l are parameters. It is important to note that the formula (1) is only valid for sufficiently long times. The continuous approximation of (1) can be analytically estimated as a power-law function for some positive number l

$$(z \gg z_{\min}) : f(z) \propto Cz^{-l}, \tag{2}$$

where $z_{\min}$ means a threshold value.

*Correspondence: yurij.katchanov@gmail.com
[1] National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, 101000 Moscow, Russian Federation
Full list of author information is available at the end of the article

One of the classic results of scientometrics is the derivation of a model in which the probability distribution (PD for short) of z, in its asymptotic tail, is equivalent to a power-law PD (2) (Haitun 1982; Yablonsky 1985). A possible mechanism to explain the power-law distribution is a stochastic growth process in which the citation rate of a paper is defined by the total number of received citations and the time after publication (Albert and Barabási 2002; de Solla Price 1976; Dorogovtsev et al. 2000; Golosovsky and Solomon 2013; Krapivsky et al. 2000).

The other main result is the justification of the power-law approximation of the statistical distribution of citations of scientific papers (Albarrán and Ruiz-Castillo 2011; Albarrán et al. 2011; Ausloos 2014; Brzezinski 2014; Egghe 2007, Eom and Fortunato 2011; Peterson et al. 2010; Radicchi and Castellano 2012; Redner 1998; Stringer et al 2010; Waltman et al. 2012; Zhao and Ye 2013). However, the power-law distribution is possessed of a number of characteristics limiting its application (Clauset et al. 2009; Golosovsky and Solomon 2012; Newman 2005).

Power laws (see detail in (Clauset et al. 2009; Newman 2005; Virkar and Clauset 2005; 2014)) are widely used to represent scientometric distributions. In reality, however, certain studies of citation distributions have used various other functional forms to provide best approximations to as wide a variety of bibliometric data as possible (see, e.g., (Burrell 2014; Davies 2002; Golosovsky and Solomon 2012; Gupta et al. 2005; Laherrère and Sornette 1998; Radicchi et al. 2008; Redner 2005; Sangwal 2013; van Raan 2001)). Nevertheless, all the same, power laws had and still have a crucial part to play in scientometrics, not only because they are established but also because they are theoretically well-founded, for reasons arising from the generalized central limit theorem (Uchaikin and Zolotarev 2011), which has very considerable importance in probability theory.

One of the better models for the citation distribution is the Tsallis distribution (Anastasiadis et al. 2010; Bletsas and Sahalos 2009; Tsallis and de Albuquerque 2000; Wallace et al. 2009)

$$(q < 2)(\lambda > 0): f(z) \propto \lambda(2-q)e_q(-\lambda z), \tag{3}$$

where

$$e_q(z) = \begin{cases} \exp(z) & \text{if } q = 1, \\ (1+\rho z)^{\frac{1}{\rho}} & \text{if } ((1+\rho z) > 0) \bigwedge (q \neq 1) \\ 0, & \text{otherwise} \end{cases} \tag{4}$$

is the q-exponential. (Here we use the symbol ρ to denote $(1-q)$). The q-exponential can also be defined by the following equation describing the (temporal) nonlinear relaxation of a system from an unstable point:

$$\frac{d\mathbf{e}}{dt} = -\mathbf{e}^q$$

with $\mathbf{e}$ given by $e_q(-t)$. The meaning of this statement is quite understandable.

In turn, the Tsallis distribution (3) may be regarded as a special case of the generalized Pareto distribution (GPD for short) (Bermudez and Kotz 2010)

$$(z \geq \mu)(\xi \neq 0): f(z) = \frac{1}{\sigma}\left(1 + \frac{\xi(z-\mu)}{\sigma}\right)^{\left(-\frac{1}{\xi}-1\right)},$$

where $\mu = 0$, $\xi = \frac{q-1}{2-q}$, $\sigma = 0$. We also can say that the random variable (or, in abbreviated form, RV) Z has a GPD if (essentially) the RV Z can be expressed as $k + \phi(1-U)^{-\delta}$, where U is a standard uniform RV. We intend to show here a specific but common heuristic model that can be adopted to generalize the GPD.

The practice of citations evolves over time. We can conceive of the process of citation as a way of tracking discrete social acts. Time lends citations direction and meaning (see (Bouabid 2011; Burrell 2002, 2014; Eom and Fortunato 2011; Glänzel 2007; Hsu and Huang 2011; Radicchi et al. 2012; Redner 2005; Simkin and Roychowdhury 2012; Wang et al. 2013) for more details). However, when we analyze bibliometric data sets, we may interpret citations not as a series of discrete acts but rather as a statistical regularity which can be expressed in the language of timeindependent PDs. While the very meaning of the RV Z is difficult to represent in terms of the PD, it acquires a direct intuitive sense in terms of the terminal time formalism, which is developed in a systematic way (a nice general reference book for Markov processes is (Sharpe 1988)). The formal solution may consist in making the terminal time, or the lifetime, the main source of information of the RV Z. For a given process of citation, the terminal time is random. To define a realization (of the Markov process of citation) we must describe the corresponding conditional probability W. There is a natural way to associate with the terminal time problem the conditional probability W that the Markov process of citation does not stop during the fixed time interval, given that all phenomena, connected with this process during the same time interval, are known. It is proved (cf. (Sharpe 1988, [Chap. VII])) that W is connected with (nonnegative and right-continuous with respect to time) additive functionals of the initial Markov process of citation. We recall that an additive functional of a Markov process X is a map which associates with each interval of time $[s,t]$ a RV a_t^s, where a_t^s depends only on the evolution of X during the time $[s,t]$, and also the condition $a_t^s + a_\tau^t = a_\tau^s$ holds for arbitrary $t \in [s, \tau]$.

The approach proposed in this paper consists in letting the probability W play a crucial part by summarizing enough information about social citation system. As a rough guide, we suppose that the RV Z depends on time through the probability W.

The issue addressed in this paper is the development of a citation distribution that can be characterized in terms of the conditional probability W (given the total information concerning the performance of the process of citation for time t) that the Markov process of citation is of a duration longer than the time t. For the moment, we are not concerned with the explicit time dependence of the citations. In this paper, we assume that the RV Z is a function

$$\varphi(w) := \{z \in Z\colon \exists w \ ((w \in W) \wedge (z = \varphi(w)))\}, \qquad (5)$$

which we have yet to treat. We shall adopt an "asymptotic" point of view. We shall only be interested in the relation $\varphi(\cdot)\colon W \to Z$ that holds between W and Z at large times.

The proposed approach is based on the concept of the approximate invariance of the function ($w \in [0, 1]$): $w \mapsto \varphi(w)$ by a translation of w, i.e., we claim that $\varphi(w + \cdot) \approx \varphi(w)\varphi(\cdot)$. The considered heuristic model for the Markov process of citation is formulated as the inhomogeneous Choquet – Deny convolution equation (we shall use the abbreviated notation ICDCE) whose form is apparently determined by the approximate translation invariance. The solution of this equation gives the Wakeby distribution (WD) for citations of scientific papers. Until now, the WD has not been among the distributions employed to model observed bibliometric data.

The rest of this paper is organized as follows. The main result regarding our proposed model and its analytical solution is presented in the 2nd section. The empirical verification is provided in the 3rd section. Finally, concluding remarks are presented in the 4th section. The Appendix 1 introduces certain necessary definitions and reviews results that are needed in the rest of the paper.

Model of citation distribution

The model $w \mapsto \varphi(w)$ can work reasonably well in scientometrics for social citation systems that are either sufficiently "ordered" or sufficiently "disordered". In the limit of a large social citation system, we may at least assume that social citation system can be decomposed into a "structured" subsystem and a "stochastic" subsystem. For sake of concreteness, let us depart from the hypothesis that social citation system includes two types of subsystems whose nature is quite different. One of them could be identified as a social network, the other as a scientific market:

- The social network (sufficiently structured subsystem) is a polycentric complex of interrelated scholars.
- The scientific market (sufficiently stochastic subsystem) contains autonomous scholars who enter into the competition.
- The social network is characterized by structural cohesion, while the scientific market is actually an amorphous medium for sharing information resources.
- The evolution of the scientific market is of a stochastic nature.
- The social network corresponds to the notion of a dynamic system.
- The statistical properties of the citation distribution are partially determined by the nature of interactions between scientific market and social network.

Employing the previous notation, the postulated heuristic propositions, on the basis of which our model of the citation distribution is constructed, are as follows:

(1) In the event horizon where the scientific market "lives", it can be assumed that the function $\varphi(w)$ in the expression (5) is invariant under translation of w

$$\varphi(w + \cdot) = \varphi(w)\varphi(\cdot). \qquad (6)$$

(2) In the event horizon of the social network the function $\varphi(w)$ may be intuitively considered as the positive contraction semigroup $\tau(w)$ on a real one-dimensional Banach space generated by $-\beta$

$$(\beta \in \mathbb{R})\colon \tau(w) = \exp(-\beta w). \qquad (7)$$

(3) The social logic of the citation distribution is such that there is a two-way influence between the scientific market and the social network (Bourdieu 1975) . However, in the limit of long time, social effects of the process of citation bring to screening "long-range" interactions. As a result, the subsystems in social citation system are almost independent and we obtain approximate translation invariance

$$\varphi(w + \cdot) = \varphi(w)\varphi(\cdot) + r(w), \qquad (8)$$

where $r(w)$ indicates a remainder term.

In the framework of previously accepted propositions the following statements are considered:

- The simplest and most intuitive general approach to translate invariance is via convolution. Let T_a be the translation operator defined by $T_a\varphi(w) = \varphi(w + a)$. Translation invariance of the convolution $(\varphi * \chi)$ means that the convolution with a fixed function χ commutes with T_a, i.e.,

$$T_a\left(\varphi * \chi\right) = \left(T_a\varphi\right) * \chi = \varphi * \left(T_a\chi\right).$$

It can involve explicitly the well-known Choquet – Deny convolution equation (CDCE for short, see (20)).

- By virtue of formula (7), whatever the precise form of $r(w)$ may be it will give to (8) a contribution of the form

$$\lim_{w\uparrow 1} r(w) = O\left(\exp(-\beta w)\right).$$

This proposition corresponds to a functional equation that can be rewritten as the ICDCE (see (21)).

The translation invariance is an important concept, so it should be understood in a thorough manner. The probability W, of course, corresponds to terminal time, while the RV Z occurs at random in time. Since the RV Z in the scientific market should be independent of an arbitrary translation a, the constancy of termination rate of the process of citation take place in the scientific market. This is what we mean when we say that $\varphi(w)$ has the translation invariant property (6) in the scientific market.

The motivation of the approximate translation invariance is to take the relation between the scientific market and the social network into consideration. In rough approximation, the scientific market and the social network can be considered as relatively independent. Consequently, their contributions to $\varphi(w)$ are additive. Summing (6), and (7), we obtain (8), i.e., approximate translation invariance. The Eq. 8 therefore expresses some kind of linear superposition of the effect of the scientific market and the social network. This superposition is not valid in the general case.

To find the citation distribution that we seek, we will start off with certain well-known mathematical constructions. Let $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\in I}, \mathbf{P})$ be a filtered probability space that satisfies the usual conditions (for details, see Chap. 1 of (Sharpe 1988)). In constructing a model of the citation distribution, we can imagine the social citation system as a normal Markov process $X = (X_t)_{t\in I}$ in a state space $(S, \mathcal{S})$. Insofar as our interest in the social citation system is confined to a few of its features, the Markov process-based model may be relevant in explaining the citation distribution. Further, we shall suppose that the experimentally observed Markov process $\tilde{X}$ is obtained from X by curtailment of its terminal time up to $\tilde{\zeta}$

$$(I \ni \tilde{\zeta} : \Omega \to \mathbb{R}_+)(t < \tilde{\zeta}) : \tilde{x}(t, \tilde{\omega}) = x(t, \omega).$$

Equivalently, the process $\tilde{X}$ is given by a truncation of the duration of the original process X such that the trajectories of X are terminated in a random manner. One can easily see that, for a proper choice of the filtered probability space $\left(\tilde{\Omega}, \tilde{\mathcal{F}}, (\tilde{\mathcal{F}}_t)_{t\in I}, \tilde{\mathbf{P}}\right)$ and the state space $(S, \mathcal{S})$, the process $\tilde{X}$ is a subprocess of the process X. The duration of the processes X and $\tilde{X}$ are denoted by ζ and $\tilde{\zeta}$, respectively, and

$$(\forall (s, x)) : \mathbf{P}_{s,x}\left(\tilde{\zeta} \leq \zeta\right) = 1.$$

The construction of such a subprocess is minutely described in (Sharpe 1988 [p. 65–74]).

Under appropriate assumptions, we can represent the Markov process $\tilde{X}$ using the concept of the multiplicative functional of the Markov process X. This approach is explained in detail in (Sharpe 1988, [p. 286–301]). Let us now introduce the contracting, multiplicative functional $(s \leq t \leq \infty)$: $m_t^s : I(\omega) \to (S, \mathcal{S})$ continuous from the right on X. It is proved (see Theorem 4 in (Gikhman and Skorokhod 2004, [p. 71–72])) that

$$\begin{aligned}&\left(\Omega_t^s = \{\omega : s, t \in I(\omega)\}\right)\left(\text{a.s. } \Omega_t^s, \mathbf{P}_{s,x}\right) : \\ &m_t^s = \tilde{\mathbf{P}}_{s,x}\left(\tilde{\zeta} > t \,\middle|\, \left(\tilde{\mathcal{F}}^s\right)_{s\in I}\right).\end{aligned} \tag{9}$$

Let $a_t^s : I(\omega) \to (S, \mathcal{S})$ be an additive functional, continuous from the right on X. The formulae $m_t^s = \exp(-a_t^s)$, $a_t^s = -\ln m_t^s$ establish a one-to-one correspondence between a_t^s and m_t^s (Gikhman and Skorokhod 2004 [p. 64]). It follows in the usual way that,

$$(\text{a.s. } \Omega_t, \mathbf{P}_t) : \tilde{\mathbf{P}}_{s,x}\left(\tilde{\zeta} > t \,\middle|\, \left(\tilde{\mathcal{F}}^s\right)_{s\in I}\right) = \exp\left(-a_t^0\right). \tag{10}$$

In the expression (10), of the quantity

$$\tilde{\mathbf{P}}_{s,x}\left(\tilde{\zeta} > t \,\middle|\, \left(\tilde{\mathcal{F}}^s\right)_{s\in I}\right)$$

can be interpreted as the conditional probability W that the trajectory $x(\tau)$ does not terminate during the time interval $[0, t]$. Moreover, to simplify the argument, we set

$$\left(\forall t \in I(\omega)\right) : a_t^0 = \vartheta t.$$

Then we immediately verify that,

$$\tilde{\mathbf{P}}_x\left(\tilde{\zeta} > t \,\middle|\, (\mathcal{F}_t)_{t\in I}\right) = \exp(-\vartheta t), \tag{11}$$

where $\tilde{\mathbf{P}}_x\left(\tilde{\zeta} > t \middle| (\mathcal{F}_t)_{t\in I}\right)$ holds for the conditional probability W that the process of citation is of duration longer than t:

$$W \equiv \tilde{\mathbf{P}}_x\left(\tilde{\zeta} > t \middle| (\mathcal{F}_t)_{t\in I}\right).$$

We assume without essential loss of generality that under a suitable normalization, the RV W has a standard exponential distribution. With the inverse method, we have

$$W = -\ln U. \tag{12}$$

It follows from the above that the properties of the distribution $\mathbf{P}_Z(z)$ depend on w. To be thorough, we must note that the distribution $\mathbf{P}_Z(z)$ is defined on the probability space $(\mathfrak{Z}, \mathcal{B}(\mathfrak{Z}), \mathbf{P}_Z)$. Obviously, the connections between the Markov process X and the distribution $\mathbf{P}_Z(z)$ may be based on the concept of the conditional probability W. A somewhat unrealistic, but simple, schematic idea of these connections is given by the equality

$$\begin{gathered}((\forall z \in \mathbb{R}_+)\,(\{\mathfrak{z}\colon Z(\mathfrak{z}) \le z\} \in \mathcal{B}(\mathfrak{Z})) : Z\colon \mathfrak{Z} \to \mathbb{R}_+)\\ (\varphi(w)\colon \mathbb{R}_+ \to \mathbb{R}_+) : z = \varphi(w),\end{gathered} \tag{13}$$

where, as in Appendix 1, $\varphi(w)$ is locally integrable (with respect to the Lebesgue measure Λ). However, the function $\varphi(w)$ is not yet completely defined. In fact, the general problem of studying the form of $\varphi(w)$ can be reduced to the case in which this function satisfies certain extra conditions. One can attempt to define $\varphi(w)$ implicitly by some functional equation rather than by direct definitions. In particular, the general form of $\varphi(w)$ may be derived uniquely from its invariance.

For the purpose of our study, based upon the denotation introduced in Appendix 1, let μ^n be the n-fold convolution of μ, and let $\varphi(w)$ be a nontrivial positive solution of the ICDCE (21). Observe first that from the paper of (Gu and Lau 1984), we know that for a.a. (mod Λ) $w \in \mathbb{R}_+$, we have the relation

$$\begin{aligned}\varphi(w) = \lim_{n\to\infty} \int_{\mathbb{R}_+} \varphi(w+v)\,\mu^n\,(dv)\\ + \sum_{n=0}^{\infty} \int_{\mathbb{R}_+} \varphi(w+v) r(w+v)\,\mu^n\,(dv).\end{aligned}$$

Suppose μ is a non-probability measure. If we take μ without requiring $\int_{\mathbb{R}_+} \mu(dv) = 1$ and, *mutatis mutandis*, use the arguments employed by (Gu and Lau 1984), we obtain the following expression for $\varphi(w)$:

$$\varphi(w) \propto \kappa_1 \exp(\delta w) + \kappa_2 \exp(-\beta w), \tag{14}$$

where κ_1 and κ_2 are constants. It should be mentioned that the definition (13) allows us to write the function $\varphi(w)$ in an explicit form of the RV Z

$$Z \propto \kappa_1 \exp(\delta w) + \kappa_2 \exp(-\beta w). \tag{15}$$

This expression is the relation we were seeking between the quantities we were interested in, Z and W. As could be expected, the RV Z contains two parts: one corresponds to the incident stream of citations, the other to the scattered stream of citations.

To extract the implications of (15), it is convenient to represent the RV W in terms of the uniform RV U. Now, if we recall the Eq. 12, the expression (15) can be straightforwardly rewritten as

$$Z \propto \kappa_1 U^{-\delta} + \kappa_2 U^{\beta}. \tag{16}$$

The study of the relation (16) makes it possible to obtain the PD of the RV Z. Motivated by the approximate translational invariance of z with respect to the probability w that the process of citation does not terminate, we suggest that this model is appropriate to provide a phenomenologically relevant picture of the citation distribution. Finally, starting from the statistical considerations connected with a common and convenient choice of distribution function (Johnson et al 2010, [Chap. 12]), a natural modification of the relation (16) can be written in the form

$$Z = \upsilon(1-U)^{-\delta} - \theta(1-U)^{\beta} + k. \tag{17}$$

The formula (17) defines the distribution, which is called the WD (Johnson et al. 2010, [p. 44–46]). This distribution was established by H. A. Thomas (Houghton 1978) (who lived on Wakeby pond on Cape Cod, Massachusetts) for hydrological data case studies (Griffiths 1989; Hosking and Wallis 2005). We stress that the explicit formula for the PDF of Z is not generally available.

For the sake of being definite, it would be better to rewrite (17) using the following notation

$$\upsilon = \gamma/\delta,\ \theta = \alpha/\beta,\ k = \xi + \theta - \phi.$$

Suppose all parameters $\alpha,\ \beta,\ \gamma,\ \delta,\ \xi$ are continuous. Then, the WD becomes

$$Z = \xi + \frac{\alpha}{\beta}\left(1-(1-U)^{\beta}\right) - \frac{\gamma}{\delta}\left(1-(1-U)^{-\delta}\right). \tag{18}$$

It is readily seen that the WD has three disposable shape parameters, one location parameter and one scale parameter. Under the following conditions:

$$\begin{gathered}(\alpha \neq 0) \vee (\gamma \neq 0),\\ (\beta + \delta > 0) \vee (\beta = \gamma = \delta = 0),\\ (\alpha = 0) \Rightarrow (\beta = 0),\\ (\gamma = 0) \Rightarrow (\delta = 0),\\ (\gamma \ge 0) \wedge (\alpha + \beta \ge 0)\end{gathered}$$

the Eq. 18 has a unique solution on dom Z; here

$$\operatorname{dom} Z = \begin{cases} [\,\xi, \infty) & \text{if } (\delta \geq 0) \wedge (\gamma > 0), \\ \left[\xi, \xi + \dfrac{\alpha}{\beta} - \dfrac{\gamma}{\delta}\right] & \text{if } (\delta > 0) \vee (\gamma = 0). \end{cases}$$

The WD in (18), when $\alpha = 0$ or $\gamma = 0$ reduces to the GPD. The Eq. 18 is not very tractable for analysis but can yield efficient algorithms for the numerical simulation of the WD.

Nearly all the papers that deal with inference for the WD are based on the theory of L-moments (Hosking 1990, 2006; Hosking and Wallis 2005). The free software statistical environment R contains functions to estimate the parameters of the WD from the data (see, e.g., (Asquith 2011), and packages 'lmom', 'lmomco').

Illustration

To demonstrate the applicability of the proposed heuristic model, we evaluate the goodness-of-fit of the WD to two bibliometric datasets.

Data sets

This study is based on the citation distribution of papers published by the American Physical Society (APS), the American Mathematical Society (AMS), the European Mathematical Society (EMS), and the Institute of Physics (IOP) (see the list of journals in Appendix 2) in the years 1980 — 2008 and indexed in Thomson Reuters Journal Citation Reports, Science Edition 2012. The data on citations was obtained from the Thomson Reuters Web of Science Core Collection. The data on citations of papers of APS, AMS and EMS were obtained in December 2013. The data for IOP were received in April 2014. The number of citations z is counted as the total number of times a paper appears as a reference of a more recently published paper indexed in the Web of Science Core Collection.

Two sets of bibliometric data are tested in the study:

- The first set contains papers published by APS, AMS, and EMS. There are 10,043,731 citations among 356,287 papers.
- The second set consists of 233,570 papers published by IOP. This dataset includes 5,885,458 citations.

Empirical results

Best-fit PDs for both data sets were performed using the Mathwave EasyFit 2014 data analysis software. The 63 PDs were automatically fitted to the empirical distributions of the data sets. The Kolmogorov – Smirnov test and the Anderson – Darling test were performed to assess goodness-of-fit, and the PDs were ranked according to the goodness-of-fit. The values of the test statistics for the top 5 PDs are reported in Tables 1 and 2 (see also Figures 1, 2, 3 and 4).

Comparing the obtained values and goodness-of-fit statistics given in the Tables, it will be seen that the WD offers a greater level of accuracy than the other PDs considered.

Discussion

We conclude that the WD is in some sense the best PD to adequately fit the examined bibliometric data sets.

It should be clear that the proposed heuristic approach is only a phenomenological model of the citation distribution. The Eq. 11 has not been derived yet but has rather been injected into the model. The vector of parameters $(\alpha, \beta, \gamma, \delta, \xi)$, which fixes the WD, is assumed to be given. We can say that the formula (17) does not reproduce the exact citation distribution. We should rather view the expression (17) as an approximate representation, in which the fine details of the citation distribution have been rounded up for clarity. Nevertheless, discrepancies with observation may be caused by errors in data collection or by random influences, which will be explained later. Also, there may be many still unknown secondary effects that could change the shape of the citation distribution. But it does not detract from the consistency or the cognitive value of the mathematical model. The proposed heuristic model of the citation distribution may be considered as a potentially useful amalgamation of mathematical abstraction and scientometric intuition.

Appendixes

Appendix 1. Mathematical preliminaries to model development

In the context of this paper we are interested in mathematical formulations. Therefore, we briefly indicate here how the function $\varphi(\cdot)$ can be treated mathematically.

Let μ and ν be regular Borel measures on a locally compact Abelian group G with a countable basis. The convolution equation $\mu = \mu * \nu$ on G was first thoroughly

Table 1 Goodness of fit — Summary for Dataset 1

Distribution	Kolmogorov – Smirnov ($p = 0.00205, \alpha = 0.1$)		Anderson – Darling ($p = 1.9286, \alpha = 0.1$)	
	Statistic	Rank	Statistic	Rank
WD	0.05036	1	785.96	1
GPD	0.05818	2	878.3	2
Gen. Extreme Value	0.0861	3	2359.3	3
Pareto 2	0.09083	4	47317.0	6
Phased Bi-Exponential	0.09143	5	53274.0	8

Table 2 Goodness of fit — Summary for Dataset 2

Distribution	Kolmogorov – Smirnov ($p = 0.00253, \alpha = 0.1$)		Anderson – Darling ($p = 1.9286, \alpha = 0.1$)	
	Statistic	**Rank**	**Statistic**	**Rank**
WD	0.05845	1	655.6	1
GPD	0.07127	2	796.09	2
Gen. Extreme Value	0.09249	3	1817.5	3
Gen. Logistic	0.09848	4	2022.5	4
Phased Bi-Exponential	0.10584	5	37917.0	10

investigated by (Choquet and Deny 1960). The integral representation of unbounded solutions was generalized by (Deny 1959). For the sake of completeness, we introduce the following notation:

- $\psi : G \to \mathbb{R}_+$: the real-valued non-negative function;
- $C(G, \mathbb{R}_+)$: the space of continuous functions from G to $\mathbb{R}_+$;
- μ: the Radon measure on the Borel σ-field $\mathcal{B}(G)$ that is generated by G;
- Λ: the Lebesgue measure;
- Ψ: the space of all real-valued non-negative functions $\psi(\cdot)\colon G \to \mathbb{R}_+$ such that

$$(\forall x \in G)\ (\psi(\cdot) \in C(G, \mathbb{R}_+)) : \psi(x+y) = \psi(x)\psi(y). \tag{19}$$

The space Ψ is, by construction, a locally compact space with the topology of uniform convergence on compact sets. We define the subset $\Psi_\mu \subset \Psi$ as follows

$$(\forall x \in G)\ (\psi(\cdot) \in C(G, \mathbb{R}_+)) :$$

$$\Psi_\mu := \left\{\psi(\cdot)\colon\ (\psi(\cdot) \in \Psi) \bigwedge \left(\int_G \psi(x)\, \mu\, (dx) = 1 \right) \right\}.$$

From the definition, Ψ_μ is a Borel subset of Ψ. In addition, let G itself be the smallest closed subgroup of G that contains supp(μ).

The generalized version of the Deny's theorem is the following. When the real-valued non-negative continuous function $\phi(\cdot)\colon G \to \mathbb{R}_+$ satisfies the Choquet – Deny convolution equation:

$$(\forall x \in G)\ (\phi(\cdot) \in C(G, \mathbb{R}_+)) : \phi(x) = \int_G \phi(x+y)\, \mu\, (dy), \tag{20}$$

then there exists a unique measure ν_ϕ on Ψ_μ such that

$$(\forall x \in G)\ (\phi(\cdot) \in C(G, \mathbb{R}_+)) : \phi(x) = \int_{\Psi_\mu} \psi(x)\, \nu_\phi\, (d\psi).$$

For an extensive discussion of the whole problem, the reader is referred to (Lukečs et al. 2010, [Chap. 3]). The CDCE (20) and its ramifications occupy a central place in our study.

It should be noted that, according to (Deny 1959), if μ is a probability measure, then every bounded solution of (20) reduces to a constant.

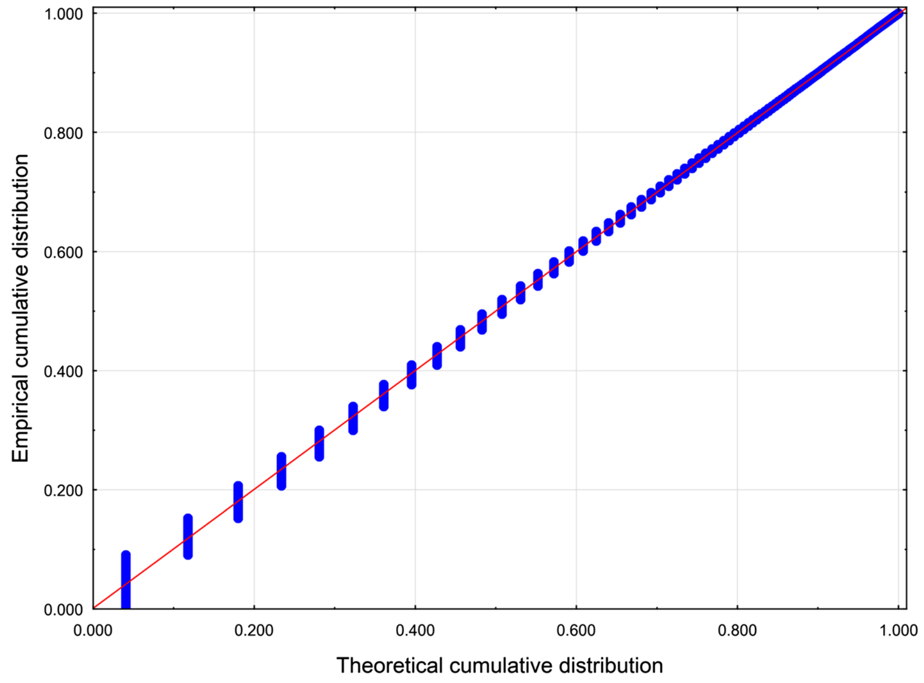

Figure 1 Probability – Probability plot of *Z* for Dataset 1. Distribution: WD.

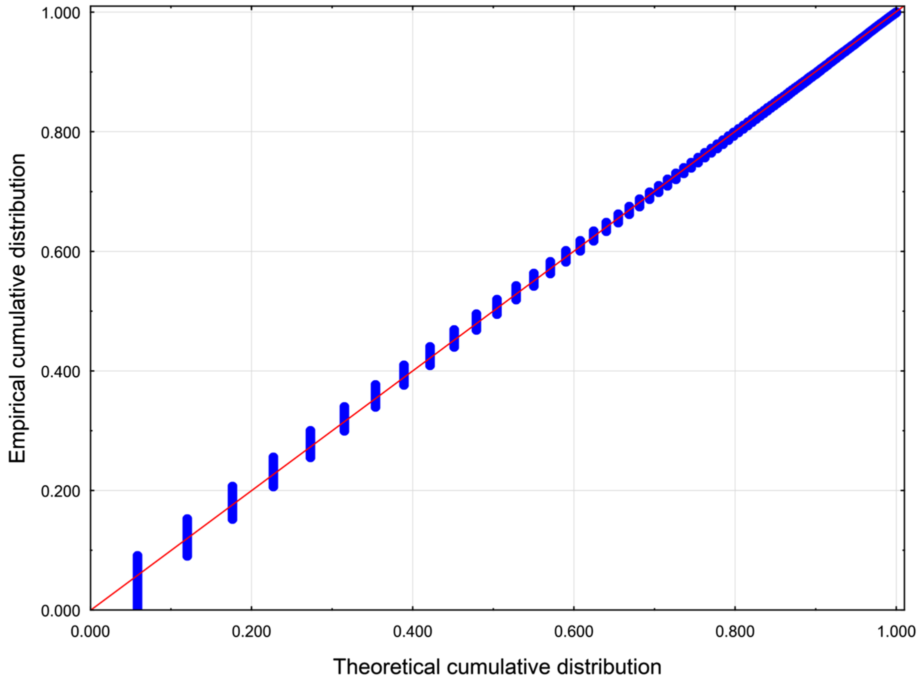

Figure 2 Probability – Probability plot of *Z* for Dataset 1. Distribution: GPD.

In the case $G = \mathbb{R}_+$, μ is assumed to be non-arithmetic such that $\mu(\emptyset) < 1$, and $\phi(\cdot)$ is assumed to be non-negative, real-valued and locally integrable with respect to the Λ function (ignoring the trivial case of $\phi(\cdot) = 0$ a.e. (mod Λ)) such that it satisfies a.a. (mod Λ) to the CDCE (20).

As a corollary of Deny's theorem, Lau and Rao provided the following theorem, specifying the above result: If a nontrivial solution for $\phi(\cdot)$ exists, then it is of the form

$$(\text{a.e. } (\text{mod } \Lambda)\; x \geq x_{\min}) : \phi(x) = p(x)\exp(cx),$$

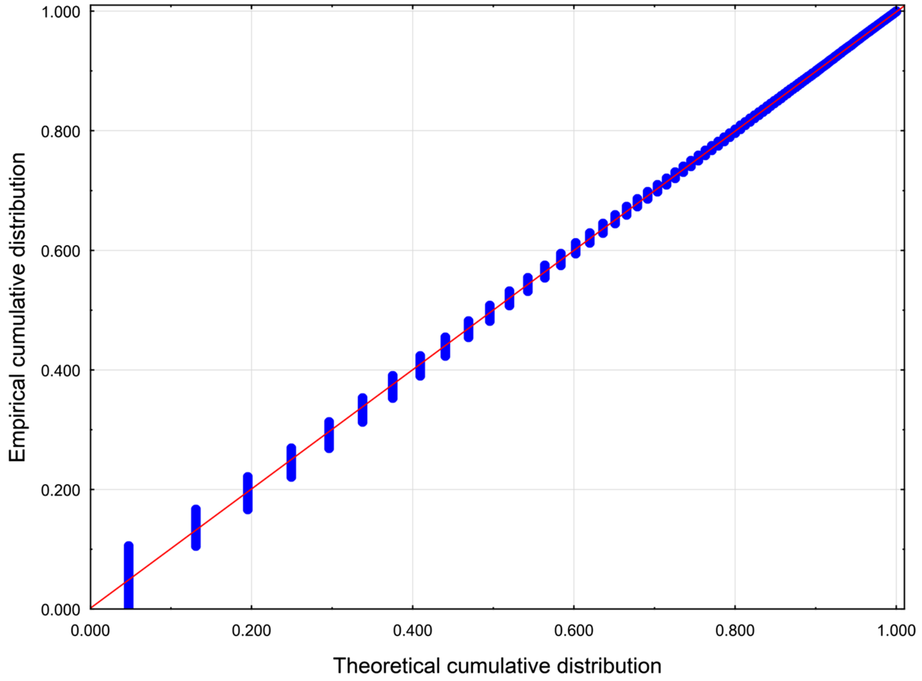

Figure 3 Probability – Probability plot of Z for Dataset 2. Distribution: WD.

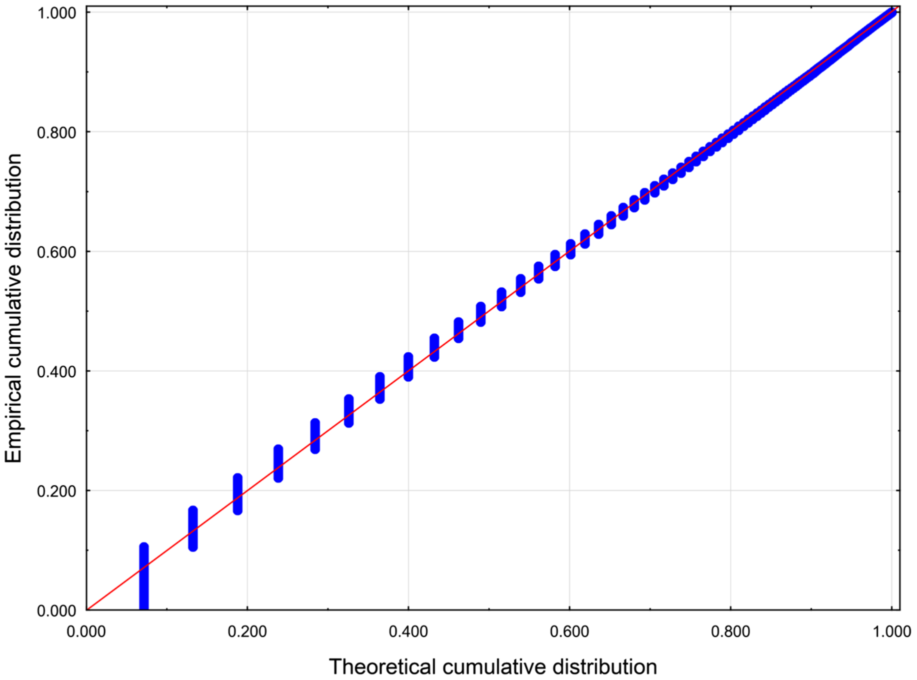

Figure 4 Probability – Probability plot of *Z* for Dataset 2. Distribution: GPD.

where the relation

$$(\forall u \in \operatorname{supp}(\mu)) : p(\cdot + u) = p(\cdot) > 0$$

is fulfilled with c such that

$$(c \in \mathbb{R})\ (\text{a.e. } (\text{mod } \Lambda)\ \forall x \in \mathbb{R}_+) : \int_{\mathbb{R}_+} \exp(cx)\, \mu\,(dx) = 1.$$

The proof of this theorem can be found in (Lau and Rao 1982).

The inhomogeneous Choquet – Deny convolution equation (ICDCE)

$$(\text{a.e. } (\text{mod } \Lambda)\ \forall x \in \mathbb{R}_+) : \phi(x) = \int_{\mathbb{R}_+} \phi(x+y)\, \mu\,(dy) + r(x), \tag{21}$$

where $|r(x)| \leq \kappa \exp(-\beta x)$ is an "error term", is a generalization of the Eq. 20 given by Shimizu. The solutions of the ICDCE on $\mathbb{R}_+$ were considered by (Shimizu 1980)and by (Gu and Lau 1984).

Appendix 2. List of journals

- American Physical Society

1. Physical Review A
2. Physical Review B
3. Physical Review B
4. Physical Review C
5. Physical Review D
6. Physical Review E
7. Physical Review Letters
8. Physical Review Special Topics Accelerators And Beams
9. Physical Review Special Topics Physics Education Research
10. Physical Review X
11. Reviews of Modern Physics

- American Mathematical Society

1. Bulletin of American Mathematical Society
2. Journal of the American Mathematical Society
3. Mathematics of Computation
4. Memoirs of the American Mathematical Society
5. Proceedings of the American Mathematical Society
6. St. Petersburg Mathematical Journal
7. Transactions of the American Mathematical Society

- European Mathematical Society

1. Commentarii Mathematici Helvetici
2. Groups Geometry and Dynamics
3. Interfaces and Free Boundaries
4. Journal of Noncommutative Geometry
5. Journal of the European Mathematical Society

6. Portugaliae Mathematica
7. Rendiconti Lincei — Matematica e Applicazioni
8. Revista Matematica Iberoamericana
9. Zeitschrift für Analysis und Ihre Anwendungen

- Institute of Physics

1. Astronomical Journal
2. Astrophysical Journal
3. Astrophysical Journal Letters
4. Astrophysical Journal Supplement Series
5. Bioinspiration Biomimetics
6. Biomedical Materials
7. Chinese Physics B
8. Chinese Physics Letters
9. Classical and Quantum Gravity
10. Communications in Theoretical Physics
11. Environmental Research Letters
12. European Journal of Physics
13. Fluid Dynamics Research
14. Inverse Problems
15. Journal of Breath Research
16. Journal of Cosmology and Astroparticle Physics
17. Journal of Geophysics and Engineering
18. Journal of Instrumentation
19. Journal of Micromechanics and Microengineering
20. Journal of Neural Engineering
21. Journal of Physics A Mathematical and Theoretical
22. Journal of Physics B Atomic Molecular and Optical Physics
23. Journal of Physics: Condensed Matter
24. Journal of Physics D Applied Physics
25. Journal of Physics G Nuclear and Particle Physics
26. Journal of Radiological Protection
27. Journal of Statistical Mechanics Theory and Experiment
28. Laser Physics
29. Laser Physics Letters
30. Measurement Science Technology
31. Metrologia
32. Modelling and Simulation in Materials Science and Engineering
33. Nanotechnology
34. New Journal of Physics
35. Nonlinearity
36. Physica Scripta
37. Physical Biology
38. Physics in Medicine and Biology
39. Physics World
40. Physiological Measurement
41. Plasma Physics and Controlled Fusion
42. Plasma Science Technology
43. Plasma Sources Science Technology
44. Reports on Progress in Physics
45. Semiconductor Science and Technology
46. Smart Materials and Structures
47. Smart Materials Structures
48. Superconductor Science Technology

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

The authors have made the following declarations about their contributions: YLK conceived and designed the study, and wrote the article; YVM carried out review of literature, collected the data sets and analyzed the data. All authors read and approved the final manuscript.

Acknowledgements

The financial support from the Government of the Russian Federation within the framework of the Basic Research Program at the National Research University Higher School of Economics and within the framework of implementation of the 5-100 Programme Roadmap of the National Research University Higher School of Economics is acknowledged.

Author details

[1]National Research University Higher School of Economics, 20 Myasnitskaya Ulitsa, 101000 Moscow, Russian Federation. [2]Institute of Sociology, Russian Academy of Sciences, 24/35 b.5 Krzhizhanovskogo Ulitsa, 117218 Moscow, Russian Federation.

Received: 12 November 2014 Accepted: 14 January 2015
Published online: 26 February 2015

References

Albarrán P, Ruiz-Castillo J (2011) References made and citations received by scientific articles. J Am Soc Inform Sci Technol 62(1):40–49. doi:10.1002/asi.21448

Albarrán P, Crespo JA, Ortuño I, Ruiz-Castillo J (2011) The skewness of science in 219 sub-fields and a number of aggregates. Scientometrics 88(2):385–397. doi:10.1007/s11192-011-0407-9

Albert R, Barabási A-L (2002) Statistical mechanics of complex networks. Rev Modern Phys 74:47–97. doi:10.1103/RevModPhys.74.47

Anastasiadis AD, de Albuquerque MP, de Albuquerque MP, Mussi DB (2010) Tsallis q-exponential describes the distribution of scientific citations – a new characterization of the impact. Scientometrics 83(1):205–218. doi:10.1007/s11192-009-0023-0

Asquith W (2011) Distributional Analysis with L-moment Statistics Using the R Environment for Statistical Computing. CreateSpace Independent Publishing Platform, US

Ausloos M (2014) Zipf – Mandelbrot– Pareto model for co-authorship popularity. Scientometrics:1–22. doi:10.1007/s11192-014-1302-y

Bermudez PZD, Kotz S (2010) Parameter estimation of the generalized Pareto distribution – Part I. J Stat Plann Inference 140(6):1353–1373. doi:10.1016/j.jspi.2008.11.019

Bletsas A, Sahalos JN (2009) Hirsch index rankings require scaling and higher moment. J Am Soc Inform Sci Technol 60(12):2577–2586. doi:10.1002/asi.21197

Bouabid H (2011) Revisiting citation aging: a model for citation distribution and life-cycle prediction. Scientometrics 88(1):199–211. doi:10.1007/s11192-011-0370-5

Bourdieu P (1975) The specificity of the scientific field and the social conditions of the progress of reason. Soc Sci Inform 14(6):19–47. doi:10.1177/053901847501400602

Brzezinski M (2014) Power laws in citation distributions: Evidence from Scopus. CoRR abs/1402.3890.1402.3890

Burrell QL (2002) The nth-citation distribution and obsolescence. Scientometrics 53:309–323. doi:10.1023/a:1014816911511

Burrell QL (2014) The individual author's publication – citation process: theory and practice. Scientometrics 98(1):725–742. doi:10.1007/s11192-013-1018-4

Choquet G, Deny J (1960) Sur l'équation de convolution $\mu = \mu * \sigma$. Comptes Rendus Hebdomadaires des Séances de l'Académie des Sciences, Paris 250:799–801

Clauset A, Shalizi C, Newman M (2009) Power-law distributions in empirical data. SIAM Rev 51(4):661–703. doi:10.1137/070710111

Davies JA (2002) The individual success of musicians, like that of physicists, follows a stretched exponential distribution. Eur Phys J B — Condens Matter Complex Syst 27(4):445–447. doi:10.1140/epjb/e2002-00176-y

Deny J (1959) Sur l'équation de convolution $\mu = \mu * \sigma$. Séminaire Brelot – Choquet – Deny. Théorie du potentiel 4:1–11

De Bellis N (2009) Bibliometrics and Citation Analysis: From the Science Citation Index to Cybermetrics. Scarecrow Press, Lanham, Md; Toronto; Plymouth, UK

de Solla Price DJ (1965) Networks of scientific papers. Science 149(3683):510–515. doi:10.1126/science.149.3683.510

de Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. J Am Soc Inform Sci 27(5):292–306. doi:10.1002/asi.4630270505

Dorogovtsev SN, Mendes JFF, Samukhin AN (2000) Structure of growing networks with preferential linking. Phys Rev Lett 85:4633–4636. doi:10.1103/PhysRevLett.85.4633

Egghe L (2007) Power Laws in the Information Production Process: Lotkaian Informetrics. 2nd edn. Elsevier/Academic Press, Amsterdam; New York

Egghe L, Rousseau R (2012) Theory and practice of the shifted Lotka function. Scientometrics 91(1):295–301. doi:10.1007/s11192-011-0539-y

Eom Y-H, Fortunato S (2011) Characterizing and modeling citation dynamics. PLoS ONE 6(9):24926. doi:10.1371/journal.pone.0024926

Gikhman II, Skorokhod AV (2004) The Theory of Stochastic Processes: II. Springer Berlin, Heidelberg; New York

Glänzel W (2007) Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. J Informetrics 1(1):92–102. doi:10.1016/j.joi.2006.10.001

Glänzel W, Moed HF (2013) Opinion paper: Thoughts and facts on bibliometric indicators. Scientometrics 96(1):381–394. doi:10.1007/s11192-012-0898-z

Golosovsky M, Solomon S (2012) Runaway events dominate the heavy tail of citation distributions. Eur Phys J Spec Topics 205(1):303–311. doi:10.1140/epjst/e2012-01576-4

Golosovsky M, Solomon S (2013) The transition towards immortality: Non-linear autocatalytic growth of citations to scientific papers. J Stat Phys 151(1-2):340–354. doi:10.1007/s10955-013-0714-z

Griffiths GA (1989) A theoretically based Wakeby distribution for annual flood series. Hydrological Sci J 34(3):231–248. doi:10.1080/02626668909491332

Gu H-M, Lau K-S (1984) Integrated Cauchy functional equation with an error term and the exponential law. Sankhyā, Ind J Stat Ser A (1961–2002) 46(3):339–354

Gupta HM, Campanha JR, Pesce RAG (2005) Power-law distributions for the citation index of scientific publications and scientists. Braz J Phys 35:981–986. doi:10.1590/S0103-97332005000600012

Haitun SD (1982) Stationary scientometric distributions. Scientometrics 4(2):89–104. doi:10.1007/BF02018448

Hosking JR (1990) L-moments: analysis and estimation of distributions using linear combinations of order statistics. J R Stat Soc Ser B (Methodological) 52(1):105–124

Hosking JRM (2006) On the characterization of distributions by their L-moments. J Stat Plann Inference 136(1):193–198. doi:10.1016/j.jspi.2004.06.004

Hosking JRM, Wallis JR (2005) Regional Frequency Analysis: an Approach Based on L-moments. Cambridge University Press, Cambridge; New York

Houghton JC (1978) Birth of a parent: The Wakeby distribution for modeling flood flows. Water Resour Res 14(6):1105–1109. doi:10.1029/WR014i006p01105

Hsu J-W, Huang D-W (2011) Dynamics of citation distribution. Comput Phys Commun 182(1):185–187. doi:10.1016/j.cpc.2010.07.031

Johnson NL, Kotz S, Balakrishnan N (2010) Continuous Univariate Distributions. In: Wiley Series in Probability and Statistics Series, vol. 1, 3rd edn. John Wiley & Sons Incorporated, New York

Krapivsky PL, Redner S, Leyvraz F (2000) Connectivity of growing random networks. Phys Rev Lett 85:4629–4632. doi:10.1103/PhysRevLett.85.4629

Laherrère J, Sornette D (1998) Stretched exponential distributions in nature and economy: "fat tails" with characteristic scales. Eur Phys J B—Condens Matter Complex Syst 2(4):525–539. doi:10.1007/s100510050276

Lau K-S, Rao CR (1982) Integrated Cauchy functional equation and characterizations of the exponential law. Sankhyā: Ind J Stat Ser A (1961–2002) 44(1):72–90

Leydesdorff L, Bornmann L, Mutz R, Opthof T (2011) Turning the tables on citation analysis one more time: Principles for comparing sets of documents. J Am Soc Inform Sci Technol 62(7):1370–1381. doi:10.1002/asi.21534

Leydesdorff L, Zhou P, Bornmann L (2013) How can journal impact factors be normalized across fields of science? An assessment in terms of percentile ranks and fractional counts. J Am Soc Inform Sci Technol 64(1):96–107. doi:10.1002/asi.22765

Lotka AJ (1926) The frequency distribution of scientific productivity. J Wash Acad Sci 16(12):317–323

Lukeš J, Malý J, Netuka I, Spurný J (2010) Integral Representation Theory: Applications to Convexity, Banach Spaces and Potential Theory. Walter de Gruyter, Berlin; New York

MathWave. EasyFit. 5.5 edition (2014). Available from: http://www.mathwave.com/products/easyfit.html

Moed HF (2005) Citation Analysis in Research Evaluation. Springer, Dordrecht. doi:10.1007/1-4020-3714-7

Newman MEJ (2005) Power laws, Pareto distributions and Zipf's law. Contemp Phys 46(5):323–351. doi:10.1080/00107510500052444

Perline R (2005) Strong, weak and false inverse power laws. Stat Sci 20(1):68–88. doi:10.1214/088342304000000215

Peterson GJ, Pressé S, Dill KA (2010) Nonuniversal power law scaling in the probability distribution of scientific citations. Proc Nat Acad Sci 107(37):16023–16027. doi:10.1073/pnas.1010757107

Radicchi F, Castellano C (2011) Rescaling citations of publications in physics. Phys Rev E 83:046116. doi:10.1103/PhysRevE.83.046116

Radicchi F, Castellano C (2012) A reverse engineering approach to the suppression of citation biases reveals universal properties of citation distributions. PLoS ONE 7(3):33833. doi:10.1371/journal.pone.0033833

Radicchi F, Fortunato S, Castellano C (2008) Universality of citation distributions: Toward an objective measure of scientific impact. Proc Nat Acad Sci 105(45):17268–17272. doi:10.1073/pnas.0806977105

Radicchi F, Fortunato S, Vespignani A (2012) Citation networks. In: Scharnhorst A, Börner K, van den Besselaar P (eds). Models of Science Dynamics. Understanding Complex Systems. Springer, Berlin; Heidelberg. pp 233–257. doi:10.1007/978-3-642-23068-4-7

Redner S (1998) How popular is your paper? An empirical study of the citation distribution. Eur Phys J B– Condens Matter Complex Syst 4(2):131–134. doi:10.1007/s100510050359

Redner S (2005) Citation statistics from 110 years of Physical Review. Phys Today 85(6):49–54. doi:10.1063/1.1996475

Sangwal K (2013) Comparison of different mathematical functions for the analysis of citation distribution of papers of individual authors. J Informetrics 7(1):36–49. doi:10.1016/j.joi.2012.09.002

Simkin MV, Roychowdhury VP (2012) Theory of citing. In: Thai MT, Pardalos PM (eds). Handbook of Optimization in Complex Networks. Springer Optimization and Its Applications. Springer, New York, NY. pp 463–505. doi:10.1007/978-1-4614-0754-6-1

Sharpe M (1988) General Theory of Markov Processes. In: Pure and Applied Mathematics, vol. 133. Academic Press, Boston, Mass

Shimizu R (1980) Functional equation with an error term and the stability of some characterizations of the exponential distribution. Ann Inst Stat Math 32(1):1-16. doi:10.1007/BF02480306

Shockley W (1957) On the statistics of individual variations of productivity in research laboratories. Proc Inst Radio Eng 45(3):279–290. doi:10.1109/JRPROC.1957.278364

Stringer MJ, Sales-Pardo M, Amaral LAN (2010) Statistical validation of a global model for the distribution of the ultimate number of citations accrued by papers published in a scientific journal. J Am Soc Inform Sci Technol 61(7):1377–1385. doi:10.1002/asi.21335

Tsallis C, de Albuquerque MP (2000) Are citations of scientific papers a case of nonextensivity? Eur Phys J B – Condens Matter Complex Syst 13(4): 777–780. doi:10.1007/s100510050097

Uchaikin VV, Zolotarev VM (2011) Chance and Stability: Stable Distributions and Their Applications. Walter de Gruyter, Berlin. doi:10.1515/9783110935974

van Raan AFJ (2001) Two-step competition process leads to quasi power-law income distributions: Application to scientific publication and citation distributions. Phys A: Stat Mech Appl 298(3):530–536. doi:10.1016/S0378-4371(01)00254-0

Virkar Y, Clauset A (2014) Power-law distributions in binned empirical data. Ann Appl Stat 8(1):89–119. doi:10.1214/13-AOAS710

Waltman L, van Eck NJ (2013) A systematic empirical comparison of different approaches for normalizing citation impact indicators. J Informetrics 7(4):833–849. doi:10.1016/j.joi.2013.08.002

Wallace ML, Larivière V, Gingras Y (2009) Modeling a century of citation distributions. J Informetrics 3(4):296–303. doi:10.1016/j.joi.2009.03.010

Wang D, Song C, Barabási A-L (2013) Quantifying long-term scientific impact. Science 342(6154):127–132. doi:10.1126/science.1237825

Waltman L, van Eck NJ, van Raan AFJ (2012) Universality of citation distributions revisited. J Am Soc Inform Sci Technol 63(1):72–77. doi:10.1002/asi.21671

Yablonsky AI (1985) Stable non-Gaussian distributions in scientometrics. Scientometrics 7(3):459–470. doi:10.1007/BF02017161

Zhao SX, Ye FY (2013) Power-law link strength distribution in paper cocitation networks. J Am Soc Inform Sci Technol 64(7):1480–1489. doi:10.1002/asi.22846