Physica A 482 (2017) 173–180

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Impact factor distribution revisited

Ding-wei Huang

Department of Physics, Chung Yuan Christian University, Chung-li, Taiwan

HIGHLIGHTS

- Consistent frequency distribution is obtained.
- The goodness-of-fit is evaluated by chi-square value.
- A bell-shaped distribution is restored by a log transformation.
- The tail of distribution can be well described.

ARTICLE INFO

Article history:
Received 9 August 2016
Received in revised form 15 February 2017
Available online 22 April 2017

Keywords:
Journal impact factor
Citation dynamics
Lavalette distribution
Rank distribution
Frequency distribution
Log transformation

ABSTRACT

We explore the consistency of a new type of frequency distribution, where the corresponding rank distribution is Lavalette distribution. Empirical data of journal impact factors can be well described. This distribution is distinct from Poisson distribution and negative binomial distribution, which were suggested by previous study. By a log transformation, we obtain a bell-shaped distribution, which is then compared to Gaussian and catenary curves. Possible mechanisms behind the shape of impact factor distribution are suggested.

1. Introduction

For complex systems, distributions of relevant variables often reflect the underlying basic dynamics. Traditional physic has brought us to understand various kinds of distributions in nature phenomena. In the realm of sociophysics, the physics perspective is extended to social phenomena [1]. Different aspects of empirical data are often demonstrated by different distributions. When the system is complicated and the data are not accurate enough, different distributions might be used without being aware of their inconsistency. This work aims to study the consistency of typical distributions in the citation dynamics.

In the system of academic publishing, a scholarly journal is often ranked by its impact factor. The impact factor is a simple measure of the average number of citations to recent articles published in the journal. With the emphasis on peer opinions, citations reflect the impact, i.e., how often an article being cited by other scholars. The impact factor becomes a convenient measure of the journal's importance in the research field. The rank distribution of impact factors is widely used to compare journals, even with some weaknesses [2]. In many complex systems involving human activities, the rank distribution follows a simple power law with a negative exponent, i.e., the Zipf's law. The rank distribution of impact factors does not conform to the Zipf's law. The power law behavior can only be observed in a limited range in the lower ranks, i.e., the highly cited journals. A cut-off appears in the higher ranks. A recent modification known as the Lavalette distribution gave a satisfactory

E-mail address: dwhuang@cycu.edu.tw.

http://dx.doi.org/10.1016/j.physa.2017.04.038
0378-4371/

description of the empirical data [3,4]. Later, the two-exponent distribution had been proposed as a further modification [5,6]. Data from both science (SCI) and social science (SSCI) journals can be fairly described by the same formulation.

Ever since the impact factor was proposed in 1970s, the distribution has been a focus of study in bibliometrics and scientometrics. Besides the rank distribution, the frequency distribution is also widely used. The frequency distribution of journal impact factors has been utilized for the practical statistical research in library and information science [7]. Related to the Zipf's law, a power law in the frequency distribution is known as the Lotka's law. For the impact factor distribution, significant deviations from the Lotka's law are observed. The distribution is not monotonic but has a maximum. The distribution is positively skewed, SCI journals much more so than SSCI journals. The skewed distribution causes some inconvenience in applying the standard statistical methods, which are often based on the assumption of a normal distribution. Previous study indicated that SCI and SSCI journals follow different formulations: SCI journals follow the negative binomial distribution and SSCI journals follow the Poisson distribution [8].

Rank distribution and frequency distribution are correlated. The relation had been used to investigate the detailed shape of the distribution. An S-shaped rank distribution was related to a bell-shaped frequency distribution [9]. However, the bell-shaped curve cannot be justified in the empirical data of the frequency distribution [10]. Thus, the existence of the S-shaped curve caused a controversy in recent investigations of the rank distribution [11]. Later, a more general relationship between rank distribution and frequency distribution has been investigated [12]. We notice that neither the Poisson distribution nor the negative binomial distribution in the frequency distribution are consistent with the Lavalette distribution or the two-exponent distribution in the rank distribution. In this work, we propose a consistent frequency-distribution. We introduce a variable transformation to restore a non-skewed frequency-distribution. We also evaluate the goodness-of-fit of various formulations. Possible mechanisms behind the shape of impact factor distribution are suggested.

2. Lavalette distribution

The Lavalette distribution is given as

$$g(y) = k\left(\frac{N+1-y}{y}\right)^{a}, \tag{1}$$

where $k > 0$ and $a > 0$ are two parameters to control the shape of the rank distribution. Besides the application in scientometrics, this distribution has also been applied to other rank distributions [13,14]. Various data can be well fitted by the formula, yet the underlying mechanisms are seldom provided. The corresponding frequency distribution becomes

$$f(x) = \frac{k^{\frac{1}{a}}\, x^{\frac{1}{a}-1}}{a\left(x^{\frac{1}{a}} + k^{\frac{1}{a}}\right)^{2}}. \tag{2}$$

The relation between a rank distribution and a frequency distribution is summarized in the Appendix. When $a > 1$, Eq. (2) becomes a monotonically decreasing distribution. When $0 < a < 1$, the frequency distribution has a single peak. The median value locates at $x = k$. The most probable value locates at $x = k\left(\frac{1-a}{1+a}\right)^{a}$, which is less than the median value. The average value locates at $x = k\left(\frac{\pi a}{\sin \pi a}\right)$, which is larger than the median value. The distribution is skewed and distinctly different from the Poisson distribution and the negative binomial distribution. The power law behavior can be observed both in the small x and in the large x. When $x \ll 1$, the distribution increases as $x^{+(1-a)/a}$; when $x \gg 1$, the distribution decreases as $x^{-(1+a)/a}$. The impact factor distribution can be fairly described. Typical results are shown in Fig. 1, where data are taken from 2011 Journal Citation Report (JCR) published by Thomson Reuters. The database includes more than 10^4 scholarly journals, which are divided into 232 subject categories. Fig. 1 shows the typical distributions in 16 subject categories. The first row shows the SCI journals with high impact-factor; the second row shows the SCI journals with low impact-factor; the third row shows the SSCI journals with high impact-factor; the fourth row shows the SSCI journals with low impact-factor.

When the numerator and the denominator of Eq. (1) assume different exponents, the Lavalette distribution is extended to the two-exponent distribution as following,

$$g(y) = k\,\frac{(N+1-y)^{b}}{y^{a}}. \tag{3}$$

The two exponents b and a dictate the power law behavior of the frequency distribution in the small x and in the large x, respectively. When $x \ll 1$, the distribution increases as $x^{+(1-b)/b}$; when $x \gg 1$, the distribution decreases as $x^{-(1+a)/a}$. Compared to the Lavalette distribution, the two-exponent distribution has one more free parameter. The fitting of empirical data improves slightly, as shown by the dotted lines in Fig. 1.

To compare the goodness of fit among different distributions, we use the standard χ^2 value defined as following

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{[f(x_i) - f_i]^2}{f_i}, \tag{4}$$

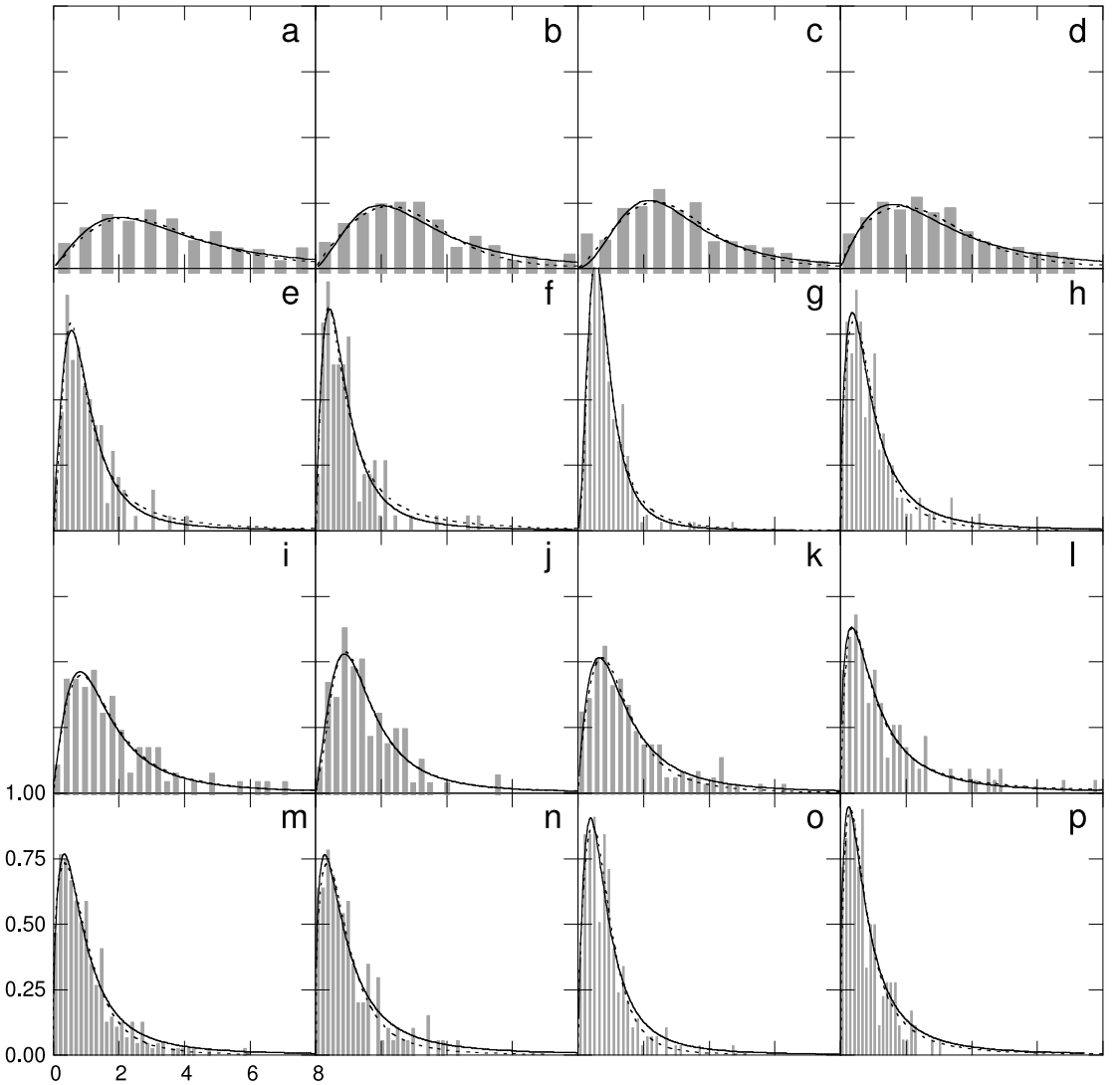

Fig. 1. $f(x)$ distribution of probability density versus impact factor for various subject categories: (a) *Cell Biology* (3.263). (b) *Biochemistry & Molecular Biology* (2.857). (c) *Neurosciences* (2.748). (d) *Oncology* (2.534). (e) *Statistics & Probability* (0.863). (f) *Engineering, Mechanical* (0.743). (g) *Mathematics, Applied* (0.724). (h) *Engineering, Civil* (0.681). (i) *Psychology, Clinical* (1.386). (j) *Public, Environmental & Occupational Health* (1.278). (k) *Management* (1.183). (l) *Psychology, Multidisciplinary* (0.955), (m) *Economics* (0.778). (n) *Law* (0.759), (o) *Education & Educational Research* (0.708), (p) *Political Science* (0.613). The number in parentheses is the median impact factor in each subject category. All figures are in the same scale as shown in the bottom left corner. Solid line shows the results of Eq. (2), which is the corresponding frequency distribution of the Lavalette distribution; dotted line shows that of the two-exponent distribution.

where (x_i, f_i) and $f(x)$ denote data and fitting curve, respectively. The results are listed in Table 1. The χ^2 values are basically the same for Poisson and negative binomial distributions. The Lavalette distribution provides a slightly better description. The two-exponent distribution provides the best description of data. However, the empirical data are not accurate enough to allow an unambiguous distinction.

The frequency distribution is non-symmetrical and positive-skewed. By definition, the impact factor has a lower limit of zero; the maximum is larger than 100 in the data of 2011. It is well known that the citations are not distributed evenly. The cumulative advantage is a generic mechanism in citation dynamics. This principle is also known as the Matthew effect: once an article receives a few citations, it tends to accumulate more citations. As a result, the most probable impact-factor is biased toward the low-impact regime; while the average impact-factor is biased toward the high-impact regime. It is interesting to observe that the bias can be remedied by a variable transformation to restore a non-skewed distribution. By plotting the logarithm of impact factor, the distribution of Eq. (2) becomes

$$h(z) = \frac{1}{a\left(k^{-\frac{1}{2a}}\, e^{\frac{z}{2a}} + k^{\frac{1}{2a}}\, e^{-\frac{z}{2a}}\right)^2} = \frac{1}{2a + 2a\, \cosh\frac{z-\ln k}{a}}, \tag{5}$$

Table 1
χ^2 values for various formulations to fit the data of Fig. 1.

	Poisson	Negative-binomial	Lavalette	Two-exponent
(a)	1.0×10^{-2}	9.3×10^{-3}	1.0×10^{-2}	8.3×10^{-3}
(b)	8.5×10^{-3}	8.5×10^{-3}	1.1×10^{-2}	7.1×10^{-3}
(c)	7.3×10^{-3}	7.3×10^{-3}	1.1×10^{-2}	6.4×10^{-3}
(d)	3.6×10^{-3}	2.6×10^{-3}	4.7×10^{-3}	2.2×10^{-3}
(e)	7.7×10^{-2}	7.7×10^{-2}	2.5×10^{-2}	2.4×10^{-2}
(f)	1.2×10^{-1}	1.2×10^{-1}	5.0×10^{-2}	4.8×10^{-2}
(g)	1.8×10^{-1}	1.8×10^{-1}	1.9×10^{-2}	1.7×10^{-2}
(h)	4.1×10^{-2}	4.1×10^{-2}	2.4×10^{-2}	1.9×10^{-2}
(i)	4.8×10^{-2}	4.8×10^{-2}	2.2×10^{-2}	2.2×10^{-2}
(j)	6.9×10^{-2}	6.9×10^{-2}	3.0×10^{-2}	2.7×10^{-2}
(k)	1.3×10^{-2}	1.3×10^{-2}	1.1×10^{-2}	6.5×10^{-3}
(l)	2.8×10^{-2}	2.6×10^{-2}	2.1×10^{-2}	2.0×10^{-2}
(m)	2.3×10^{-2}	2.3×10^{-2}	1.4×10^{-2}	1.1×10^{-2}
(n)	2.7×10^{-2}	2.7×10^{-2}	2.9×10^{-2}	2.3×10^{-2}
(o)	7.0×10^{-2}	7.0×10^{-2}	3.2×10^{-2}	2.5×10^{-2}
(p)	5.6×10^{-2}	5.6×10^{-2}	4.5×10^{-2}	4.3×10^{-2}

where $z = \ln x$ is the new variable and $h(z)dz$ is the probability to find the data point lying between z and $z + dz$. The distribution is symmetrical to its peak value at $z = \ln k$, where the median value coincides with the most probable value and the average value. The asymmetrical power-law tails in $f(x)$ transform to the symmetrical exponential tails in $h(z)$. Parameter k controls the peak value. The other parameter a controls the shape of the distribution. Typical results from the empirical data are shown in Fig. 2. All the distributions become bell-shaped in the new variable. When the Lavalette distribution is extended to the two-exponent distribution, a slight asymmetry can be observed.

3. Discussions

We obtain analytically the appropriate frequency-distribution for journal impact factors, where the ranks are fairly described by the Lavalette distribution. The distribution is distinct from both Poisson distribution and negative binomial distribution. Based mainly on the variance-to-mean ratio, previous study suggested that SSCI journals follow the Poisson distribution and SCI journals follow the negative binomial distribution. In practice, Poisson distribution is a limiting case of negative binomial distribution. If some data can be described by Poisson distribution, it cannot be worse for negative binomial distribution to describe the same data. Fitting the overall shape, we show that the data conform to the Poisson distribution. As listed in Table 1, only the cases of (a), (d), and (l) deviate from the Poisson distribution. At present, empirical data conform to both Poisson and Lavalette distributions. With Lavalette distribution, both rank distribution and frequency distribution can be described consistently. All journals can be described in the same framework. The obtained distribution provides a fair description to the impact factors across different disciplines.

We demonstrate that a log transformation transfers the skewed distribution to a bell-shaped distribution. As a result, the logarithm of impact factor seems to be a more robust variable to present the frequency distribution. In the conventional practice, citation counts are often considered in a linear scale, i.e., 1, 2, 3, and 4 citations are treated as equal-distanced points. In contrast, the equal-distanced points become 1, 10, 100, and 1000 citations in a logarithmic scale. With the cumulative advantage, it is plausible that the logarithmic scale can be more appropriate than the linear scale to explore the underlying dynamics. As shown in Fig. 2, there is an approximate symmetry between high impact and low impact journals. Some journals are driven to become prestige by the citation mechanism. The same mechanism also causes other journals to be left behind. This symmetry is exact in the Lavalette distribution. However, the data fitting can be further improved by introducing a slight asymmetry as in the two-exponent distribution.

We clarify the relationship among a frequency distribution $f(x)$, a rank distribution $g(y)$, and a log-transformed distribution $h(z)$. The Lavalette distribution looks similar to both a log-normal distribution [15] and a curve of catenary arch [16], which have been widely discussed in physics. For the log-normal distribution, we have

$$f(x) = \frac{1}{xa\sqrt{2\pi}}\, e^{-\frac{(\ln x - \ln k)^2}{2a^2}}, \tag{6}$$

$$g(y) = k\, e^{a\sqrt{2}\cdot \mathrm{erf}^{-1}\left(1-\frac{2y}{N}\right)}, \tag{7}$$

$$h(z) = \frac{1}{a\sqrt{2\pi}}\, e^{-\frac{(z-\ln k)^2}{2a^2}}. \tag{8}$$

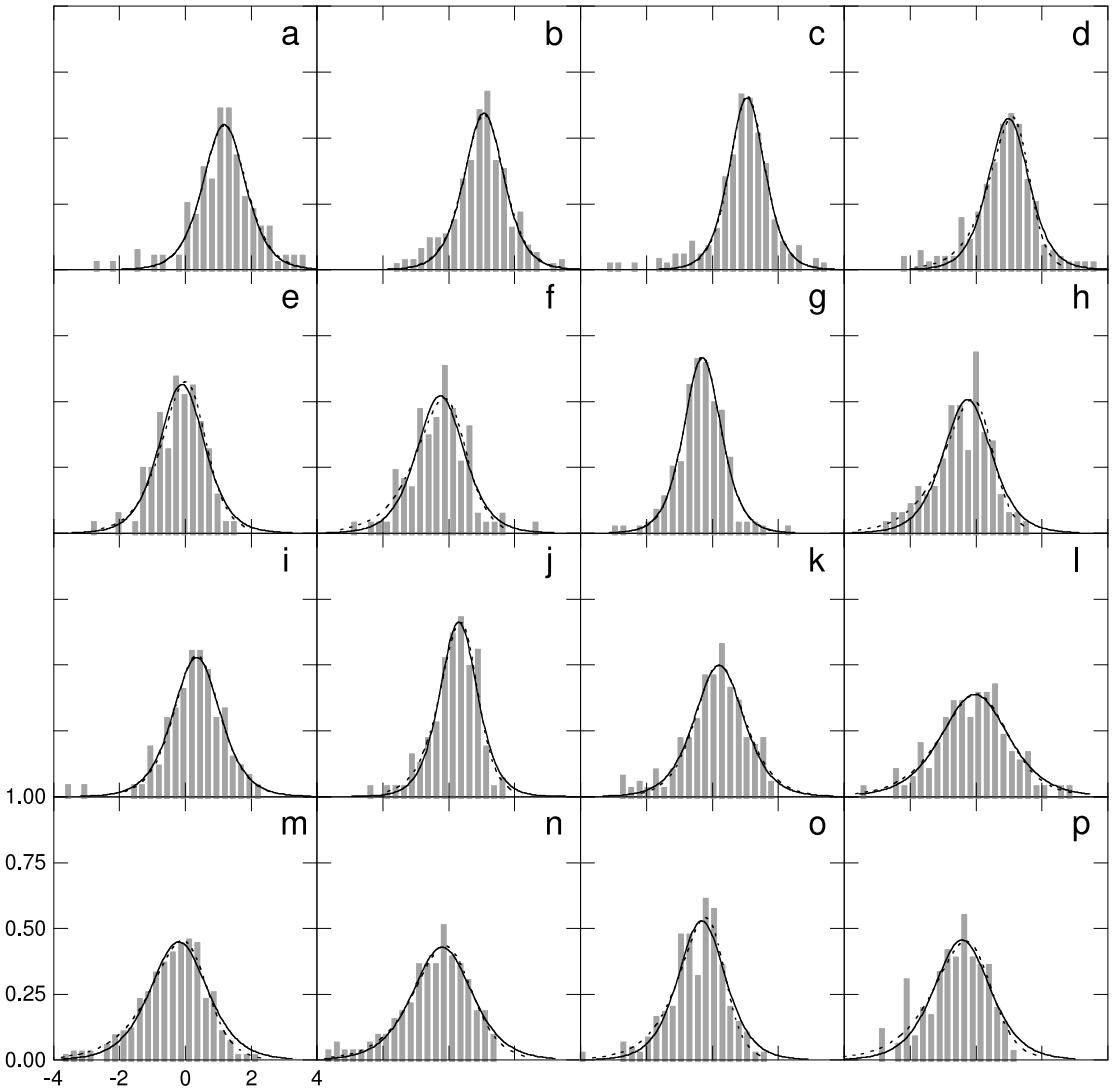

Fig. 2. $h(z)$ distribution of probability density versus the logarithm of impact factor for various subject categories. Data are the same as in Fig. 1. All figures are in the same scale as shown in the bottom left corner. Solid line and dotted line show the corresponding results of the Lavalette distribution and the two-exponent distribution, respectively.

For the catenary arch, we have

$$f(x) = \frac{2\,k^{\frac{1}{a}}\,x^{\frac{1}{a}-1}}{\pi a\left(x^{\frac{2}{a}}+k^{\frac{2}{a}}\right)}, \tag{9}$$

$$g(y) = k\tan^{a}\frac{\pi}{2}\left(1-\frac{y}{N}\right), \tag{10}$$

$$h(z) = \frac{1}{\pi a\cosh\frac{z-\ln k}{a}}. \tag{11}$$

With the uncertainty in data, it is expected that these two distributions can also provide satisfactory description of empirical data. The results of χ^2 values are listed in Table 2. The goodness of the fit is basically the same for different formulations. The empirical data also suggest that the system is in a steady state near equilibrium. Different disciplines have their own practice of publishing and citation, which reflect in the different values of the shape parameters in Fig. 2. It is well known that the impact factors should not be used to compare journals in different disciplines. However, it is also known that the impact factors across disciplines seem to follow the same formulation nicely. In Fig. 3, we plot the frequency distribution of the median impact factors in different subject categories. In Fig. 4, we plot the frequency distribution of the impact

Table 2
χ^2 values for various formulations to fit the data of Fig. 2.

	Log-normal	Catenary	Lavalette	Two-exponent
(a)	1.7×10^{-2}	1.6×10^{-2}	1.6×10^{-2}	1.4×10^{-2}
(b)	1.7×10^{-2}	1.0×10^{-2}	1.2×10^{-2}	1.1×10^{-2}
(c)	1.4×10^{-2}	8.6×10^{-3}	9.8×10^{-3}	8.1×10^{-3}
(d)	1.2×10^{-2}	8.9×10^{-3}	9.3×10^{-3}	6.1×10^{-3}
(e)	2.1×10^{-2}	2.7×10^{-2}	2.4×10^{-2}	1.8×10^{-2}
(f)	2.7×10^{-2}	3.0×10^{-2}	2.8×10^{-2}	2.3×10^{-2}
(g)	1.8×10^{-2}	1.8×10^{-2}	1.7×10^{-2}	1.2×10^{-2}
(h)	2.4×10^{-2}	2.3×10^{-2}	2.3×10^{-2}	1.7×10^{-2}
(i)	9.3×10^{-3}	8.6×10^{-3}	8.4×10^{-3}	7.0×10^{-3}
(j)	2.5×10^{-2}	2.4×10^{-2}	2.3×10^{-2}	1.4×10^{-2}
(k)	1.3×10^{-2}	1.4×10^{-2}	1.3×10^{-2}	1.1×10^{-2}
(l)	1.2×10^{-2}	1.1×10^{-2}	1.1×10^{-2}	1.1×10^{-2}
(m)	6.7×10^{-3}	7.2×10^{-3}	6.7×10^{-3}	3.9×10^{-3}
(n)	4.8×10^{-3}	3.9×10^{-3}	4.1×10^{-3}	3.2×10^{-3}
(o)	1.8×10^{-2}	2.5×10^{-2}	2.1×10^{-2}	1.8×10^{-2}
(p)	2.4×10^{-2}	2.5×10^{-2}	2.4×10^{-2}	1.7×10^{-2}

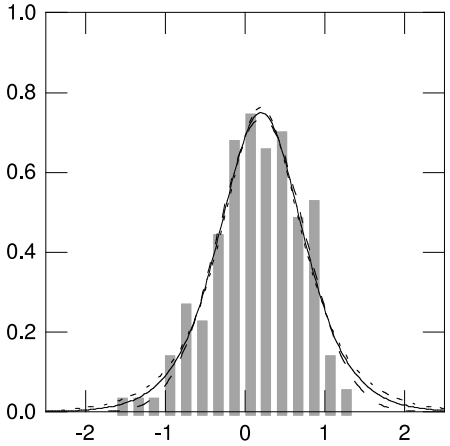

Fig. 3. Distribution $h(z)$ for the median impact factor of 232 subject categories: Lavalette (solid line); catenary (dotted line); Gaussian (dashed line). The log-normal distribution $f(x)$ transforms to a Gaussian distribution $h(z)$.

factors for all the journals from different research fields. It is interesting to note that the same formulation describes both the distribution within a specific discipline and the distribution across different disciplines. It can be the characteristics associated with a system not far away from equilibrium. Compared to Fig. 3, fluctuations in Fig. 4 are significantly suppressed. The data in Fig. 4 are now sufficiently precise to discern the tail of distribution. In Fig. 5, the same data are replotted in the logarithmic scale. The exponential tails become straight lines in this plot. The asymmetry can be well accounted for when the Lavalette distribution is extended to the two-exponent distribution. In contrast, the data on Fig. 3 (and also Fig. 2) are not accurate enough to firmly distinguish among different formulations. Those data conform to log-normal distribution, catenary curve, and Lavalette distribution. Compared to the rather obscure mechanism behind Lavalette distribution, both log-normal distribution and catenary curve are widely used in statistical physics and have concrete interpretations of the underlying processes. The log-normal distribution emphasizes the microscopic fluctuations; while the catenary curve emphasizes the macroscopic equilibrium. Different interpretations might lead to complementary perspectives of the same phenomena. It can be interesting to further explore the possible mechanisms in citation dynamics. A phenomenological model is in progress along this line.

Appendix

The relation between a frequency distribution $f(x)$ and a rank distribution $g(y)$ is summarized as the following. Consider a number of N data points distributed between $x_{\min}$ and $x_{\max}$. The normalization of the frequency distribution $f(x)$ is written as,

$$\int_{x_{\min}}^{x_{\max}} f(x)dx = 1,$$

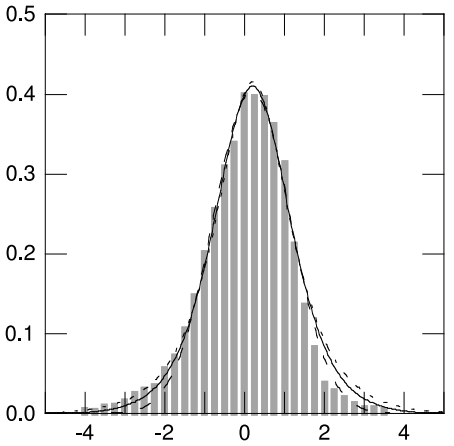

Fig. 4. Distribution $h(z)$ for the impact factor of 11 215 journal in different research fields: Lavalette (solid line); catenary (dotted line); Gaussian (dashed line).

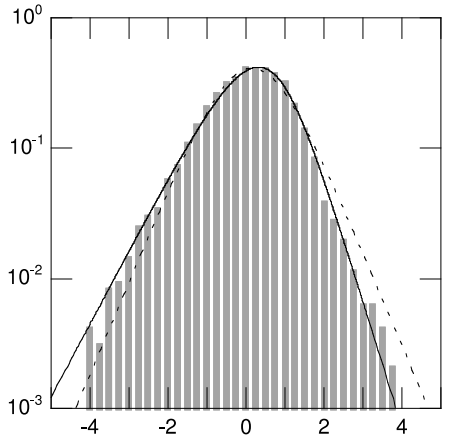

Fig. 5. The same as Fig. 4 in the logarithmic scale: Lavalette (dotted line); two-exponent (solid line).

where $f(x)dx$ is the probability to find the data point lying between x and $x+dx$. The rank distribution $g(y)$ is a monotonically decreasing function as y increases, where $y = 1, 2, 3, \ldots, N$. The boundaries of the distribution are given as $g(1) = x_{\max}$ and $g(N) = x_{\min}$. In the continuum limit, the frequency distribution $f(x)$ and the rank distribution $g(y)$ are related as

$$\int_{g(y)}^{x_{\max}} f(x)dx = \frac{y}{N}.$$

The frequency distribution $f(x)$ can be derived with a given rank distribution $g(y)$, and vice versa. The equation $x = g(y)$ can be inverted to obtain $y = g^{-1}(x)$. With the Leibniz' rule, $f(x)$ can be obtained by the following derivative

$$f(x) \propto -\frac{d}{dx}\, g^{-1}(x).$$

As an example, the Zipf's law and the Lotka's law can thus be related. The Zipf's law is given as

$$g(y) = \frac{k}{y^a},$$

where $a > 0$ is the exponent to prescribe the trend of decreasing. The frequency distribution becomes

$$f(x) = \frac{k^{\frac{1}{a}}}{a\, x^{\frac{1}{a}+1}},$$

which is the Lotka's law. Obviously, the power-law exponents in $f(x)$ and in $g(y)$ are related. When the rank distribution deviates from the Zipf's law, the frequency distribution deviates from the Lotka's law accordingly. A cut off in the tail of large y in the Zipf's law corresponds to a depletion in the distribution of small x in the Lotka's law.

References

[1] For an overview, see S. Galam, Sociophysics: A Physicist's Modeling of Psycho-political Phenomena, Springer, 2012;
B. Philip, Why Society is a Complex Matter: Meeting Twenty-first Century Challenges with a New Kind of Science, Springer, 2012.
[2] J.K. Vanclay, Impact Factor: outdated artefact or stepping-stone to journal certification? Scientometrics 92 (2012) 211–238.
[3] D. Lavalette, Facteur d'impact: impartialité ou impuissance?, Report INSERM U350, Institut Curie-Recherche, Bât. 112, Centre Universitaire, 91405 Orsay, France, 1996.
[4] I. Popescu, On a Zipf's law extension to impact factors, Glottometrics 6 (2003) 83–93.
[5] R. Mansilla, E. Köppen, G. Cocho, P. Miramontes, On the behavior of journal impact factor rank-order distribution, J. Informetr. 1 (2007) 155–160.
[6] J.M. Campanario, Distribution of ranks of articles and citations in journals, J. Am. Soc. Inf. Sci. Technol. 61 (2010) 419–423.
[7] S.J. Bensman, Probability distributions in library and information science: A historical and practitioner viewpoint, J. Am. Soc. Inf. Sci. 51 (2000) 816–833.
[8] S.J. Bensman, Distributional differences of the impact factor in the sciences versus the social sciences: An analysis of the probabilistic structure of the 2005 journal citation reports, J. Am. Soc. Inf. Sci. Technol. 59 (2008) 1366–1382.
[9] L. Egghe, Mathematical derivation of the impact factor distribution, J. Informetr. 3 (2009) 290–295.
[10] L. Waltman, N.J. van Eck, Some comments on Egghes derivation of the impact factor distribution, J. Informetr. 3 (2009) 363–366.
[11] L. Egghe, The impact factor rank-order distribution revisited, Scientometrics 87 (2011) 683–685.
[12] L. Egghe, L. Waltman, Relations between the shape of a size-frequency distribution and the shape of a rank-frequency distribution, Inf. Process. Manage. 47 (2011) 238–245.
[13] I.V. Voloshynovska, Characteristic features of rank-probability word distribution in scientific and belletricstic literature, J. Quant. Linguist. 18 (2011) 274–289.
[14] M. Ausloos, Two-exponent Lavalette function: A generalization for the case of adherents to a religious movement, Phys. Rev. E 89 (2014) 062803.
[15] O. Fontanelli, P. Miramontes, Y. Yang, G. Cocho, W. Li, Beyond Zipf's law: The Lavalette rank function and its properties, PLoS One 11 (2016) e0163241.
[16] For example, see Chapter 6, S.T. Thornton, J.B. Marion, Classical Dynamics of Particles and Systems, fifth ed., Brooks Cole, 2004.