palgrave communications

HUMANITIES | SOCIAL SCIENCES | BUSINESS

ARTICLE

Received 10 Jan 2017 | Accepted 24 May 2017 | Published 4 Jul 2017

DOI: 10.1057/palcomms.2017.64

OPEN

Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent

Giancarlo Ruocco[1,2], Cinzia Daraio[3], Viola Folli[2] and Marco Leonetti[2,4]

ABSTRACT There is now compelling evidence that the statistical distributions of extensive individual bibliometric indicators collected by a scholar, such as the number of publications or the total number of citations, are well represented by a Log-Normal function when homogeneous samples are considered. A Log-Normal distribution function is the normal distribution for the logarithm of the variable. In linear scale it is a highly skewed distribution with a long tail in the high productivity side. We are still lacking a detailed and convincing *ab-initio* model able to explain observed Log-Normal distributions—this is the gap this paper sets out to fill. Here, we propose a general explanation of the observed evidence by developing a straightforward model based on the following simple assumptions: (1) the materialist principle of the natural equality of human intelligence, (2) the *success breeds success* effect, also known as Merton effect, which can be traced back to the Gospel parables about the Talents (Matthew) and Minas (Luke), and, (3) the *recognition* and *reputation* mechanism. Building on these assumptions we propose a distribution function that, although mathematically not identical to a Log-Normal distribution, shares with it all its main features. Our model well reproduces the empirical distributions, so the hypotheses at the basis of the model are *not falsified*. Therefore the distributions of the bibliometric parameters observed *might* be the result of chance and noise (chaos) related to multiplicative phenomena connected to a *publish or perish* inflationary mechanism, led by scholars' recognition and reputations. In short, being a scholar in the right tail or in the left tail of the distribution could have very little connection to her/his merit and achievements. This interpretation *might* cast some doubts on the use of the number of papers and/or citations as a measure of scientific achievements. A tricky issue seems to emerge, that is: *what then do bibliometric indicators really measure*? This issue calls for deeper investigations into the meaning of bibliometric indicators. This is an interesting and intriguing topic for further research to be carried out within a wider interdisciplinary investigation of the science of science, which may include elements and investigation tools from philosophy, psychology and sociology.

[1] Sapienza University of Rome, Rome, Italy [2] Center for Life Nano Science@Sapienza, Istituto Italiano di Tecnologia, Viale Regina Elena, 291 00161, Roma, Italia [3] Department of Computer, Control and Management Engineering Antonio Ruberti (DIAG) University of Rome 'La Sapienza', Rome, Italy [4] CNR NANOTEC-Institute of Nanotechnology c/o Campus Ecotekne, University of Salento, Via Monteroni, 73100 Lecce, Italy Correspondence: (e-mail: giancarlo.ruocco@roma1.infn.it)

Introduction

The *rich get richer* or *success breeds success* effect, also called Matthew's principle from the parable of the Talents in Matthiew 25:14-30), has been invoked many times in the sociology of science to justify highly skewed distributions of bibliometric indicators (often power laws, see Egghe, 2005 and Rousseau, 2010) measuring the scientific production of scholars. The basic underlying idea it is that *if you have more, it's easier to gain more*. This is a consequence of "the process of allocation of rewards to scientists for their contributions" (recognition) "which in turn affects the flow of ideas and findings through the communication networks of science" generating a *reputational effect*, as Merton (1968: 56) put it.

Concerning these mechanisms, Bonaccorsi *et al.* (2017) discussed recognition as a trigger of a cumulative increase in the scientific productivity of scholars and linked the results to the framework proposed by Whitley (2000). According to Whitley (2000) different scientific disciplines, which apply different knowledge production systems, can be investigated in a comparative way, on the base of a common ground, as they are *reputational work organizations*.

A parable that is often considered similar to Matthew's Talents, but which opens toward a different perspective, is the parable of the Ten Mines, in Luke (Luke, 19:11-27). In Matthew, different outcomes are obtained starting by different amounts of Talents given at the initial time to servants with *different* abilities. On the contrary, in Luke, different outcomes are obtained starting from exactly the same amount of *stocks* (one Mina) given at the initial time to each servant, *independently* from their (unspecified) *abilities*. A related view to the latter can be found in Helvetius (1772) which proposes the materialist principle of equality of human intelligence.

The *success breeds success* principle is known, and has been reinvented many times over the last century. In animal and plant taxonomy it is known as the Yule process (Yule, 1924; Raup, 1985; Reed and Hughes, 2007), after Udny Yule (1871–1951) who studied the distribution of the sizes of biological taxa (for instance, how many species are in a genus) in 1925. From a mathematical point of view, the Yule process is a variation of the Polya's urn model (Mahmoud, 2008), attributed to the mathematician George Polya (1887–1985). Subsequently, the Yule process was generalized by the economist Herbert Simon (who won the Turing award in 1975 and Nobel Prize in Economics in 1978) to study the distribution of wealth (1916–2001) (Simon, 1955; Mandelbrot, 1959; Simon, 1960). Simon demonstrated that *the rich get richer* mechanism produces power-law distributions. In Sociology, this principle was introduced by Robert Merton (1910–2003), who named it the "Matthew effect" (Merton, 1968; Wouters and Leydesdorff, 1994), after the quoted passage in the Biblical Gospel of Matthew. In Scientometrics the model was introduced in the 1970 s by the physicist Derek de Solla Price (1922–1983) (de Solla Price, 1965; de Solla Price, 1976). Building on Simon's work, he applied the Yule process to investigate the growth of the citation network, giving the mechanism a different name: "cumulative advantage". In 1984 two Hungarian scholars, Wolfgang Glänzel a mathematician, and Andres Schubert with a background in physical chemistry, propose a model of bibliometric distributions based on the *success breeds success* principle which lead to the less common Waring distribution (Schubert and Glänzel, 1984). Both scholars were later awarded the Scientometrics Derek de Solla Price Medal.

More recently, the physicists Albert-Laszlo Barabasi and Reka Albert once more reinvented Price's network evolution mechanism in a 1999 paper (Barabasi and Albert, 1999; Albert and Barabasi, 2002; Barabasi *et al.*, 2002), renaming it as "preferential attachment". In a recent paper, Glänzel and Schubert (Glänzel and Schubert, 2016) present an overview of their 1984 statistical model. They illustrate the whole family of distributions which can be derived from their original model and show that, in retrospect, it can be considered a precursor of the preferential attachment network model, proposed by Barabasi. Many other examples of applications and many other names of the *success breeds success* mechanism can be found in the current literature. Among others, we quote (1) in system biology, the vertex-copying models recently proposed for the shape of genetic networks, proposed by the physicist Ricard Sole and colleagues (Sole and Montoya, 2001; Sole *et al.*, 2002; Sole and Pastor-Satorras, 2003) and by the mathematician Alexei Vazquez and colleagues (Vazquez, 2003); (2) in the WWW network study, the fitness-based generalization of preferential attachment, proposed by the physicists Ginestra Bianconi and Albert-Laszlo Barabasi in 2001 (Bianconi and Barabasi, 2001); (3) the forest fire model for densification, proposed by the computer scientist Jure Leskovec and colleagues (Leskoves *et al.*, 2005); (4) the local-competition mechanism proposed by the physicist Raissa D'Souza and colleagues (D'Souza *et al.*, 2007); (5) the propagation of scientific memes studied by the physicist Matjaz Perc (Perc, 2013), who also recently reviewed the methodology for measuring the impact of the Matthew effect in social, technical and scientific areas (Perc, 2014).

In Scientometrics, the *Price mechanism* (as it is known) has been mainly focused on the distribution of citations. Price's assumption was that the papers to be cited are chosen at random with a probability that is proportional to the number of citations those same papers already have. Thus, highly cited papers are likely to gain additional citations, giving rise to the *rich get richer* cumulative effect. Several modifications of the basic mechanism have been proposed from time to time, but, aside from small details, Price's original formulation seems to catch the main features of the distribution of citations.

The current literature often focuses on the distribution of citations collected by a given paper. The question of what kind of mathematical function best describes this distribution is crucial. In 1998 Redner (Redner, 1998) considered the articles published in Physical Review D, along with all articles indexed by Thomson Scientific in the period 1981–1997. He found that the right tail of the distribution (corresponding to highly cited papers) follows a power law with exponent -3, in agreement with the conclusions of Price (Wouters and Leydesdorff, 1994). Later, Laherrere and Sornette (Laherrere and Sornette, 1998) studied the top thousand most cited physicists during the same period (1981–1997). The resulting citation distribution is better described by a stretched exponential distribution with $\beta = 0.3$. Tsallis and de Albuquerque (Tsallis and de Albuquerque, 2000) analyzed the same data used by Redner with the addition of all papers published in Physical Review E, and found that the Tsallis distribution[1] with $\xi \approx 10$ and $\beta \approx 1.5$ consistently fits the whole distribution of citations (not just the tail). More recently, Redner performed an analysis over all the papers published in the century-long history of all the journals published in the American Physical Society (Redner, 2005). He reaches the conclusion that the Log-Normal distribution represents the data much better than a power law. In further studies the distributions of citations have been fitted with various functional forms: power laws (Seglen, 1992; Lehmann *et al.*, 2003; Bommarito and Katz, 2010; Perc, 2010; Rodriguez-Navarro, 2011), Log-Normal (Radicchi *et al.*, 2008; Stringer *et al.*, 2008; Bommarito and Katz, 2010), Tsallis distribution (Wallace *et al.*, 2009; Anastasiadis *et al.*, 2010), modified Bessel function (Van Raan, 2001a; Van Raan 2001b) or more complicated distributions (Kryssanov *et al.*, 2007).

It is worth noting that all but the Log-Normal fitting functions used to describe the distribution of citations c are monotonically decreasing functions of c, as the raw data clearly show no

tendency to have a dip around $c=0$. Even in those cases where the Log-Normal shape of the distribution function has been found, the data were fitted to high c tail of the Log-Normal function (see, for example, Fig. 1 of Eom and Fortunato (2011).

In addition to citation distributions, other bibliometric indicators have been shown to be well represented by a Log-Normal function in the whole domain range. When, instead of a single paper, the investigated indicator is referred to a single scholar the distribution, far to be monotonically decreasing, on increasing the variable value, first increases, reaches a maximum, then decreases with a longer right tail. Furthermore, different disciplines and different academic roles share the same Log-Normal distribution when the indicator is scaled by the median (or any other scale parameter) (Ruocco and Daraio, 2013). The same conclusion applies not only to the Hard and Life Science disciplines, but also to Social Sciences and Humanities (Bonaccorsi *et al.*, 2017).

The universality (but for a scaling parameter) of the distribution of bibliometric parameters of scholars is an intriguing finding, and its analysis can provide important information on the Sociology and Science of Science. Also, the ultimate origin of the shape of the distribution, which is highly skewed and well represented by a Log-Normal function, can give some hints on the publishing behavior of scholars and the related scientific production process.

Why must the distribution of, for example, the number of papers published by a full professor in mathematics working on the theory of functions, or the one of an associate professor in astrophysics, or the one of a pathologist, or the one of a Latinist, each be a distribution that closely resembles a Log-Normal function? The origin of the Log-Normal distribution lies in the multiplicative noise (Mitzenmacher, 2004; Limpert *et al.*, 2016), that is, the product of a large number of statistically independent fluctuations (additive noise would give rise to a normal distribution function). This answer is not satisfactory, it is only a reformulation of the original question. Why should the scientific production of a scholar be the result of *multiplicative* random phenomena? Are there other phenomena behind the observed bibliometric distributions?

In this article we propose a very simple model, based on the *rich get richer* rule, which—by the amplification of small initial fluctuations and by the *reputational cumulative advantage* mechanism- gives rise to the observed distribution of bibliometric parameters.

The mathematics of our model is straightforward. It is based on a deterministic differential equation for the individual productivity, being the only statistical variability on the initial conditions. God (Nature) gives an almost equal (number of) talent (small "t"!) to any scholar. Each scholar performs equally well, but the small initial differences, like in an inflationary process, give rise to the huge differences observed in the distributions.

The statement about the near *equality of talents* (note that in the present paper ability, talents and intelligence are considered as synonyms) is counter-intuitive and requires some explanation. Indeed, scholars may be different not only in their abilities (natural talents) but also in their opportunities of doing research. Moreover, scholars are embedded in university departments, universities and countries, all these levels being different in resource allocation, recognition and prestige.

The rationale of our statement is that we would like to test if the model, including this assumption, is still able to replicate the (Log-Normal) distributions observed in many empirical studies. This is important to say something about the *meaning* of bibliometric indicators. The reader is referred to the last section for more discussion on this point.

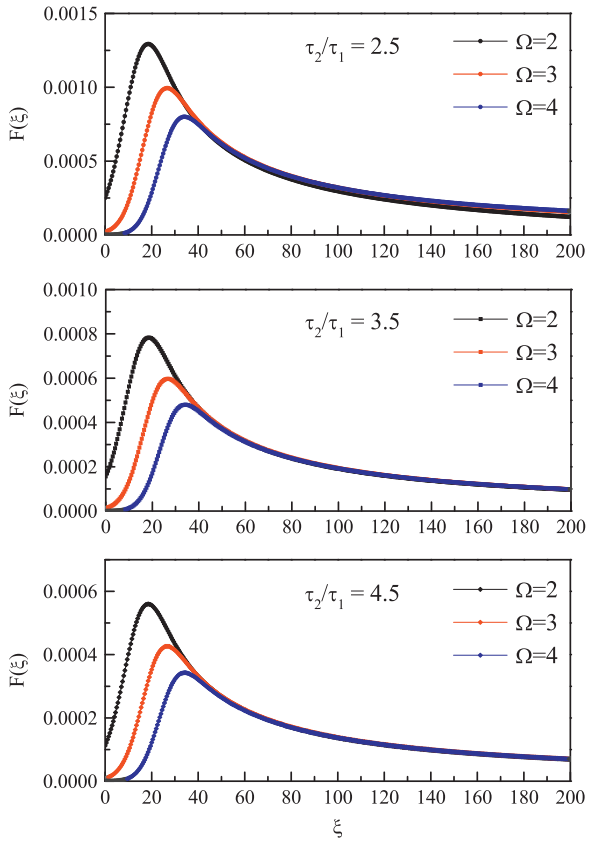

Figure 1 | Examples of distribution functions obtained from eq. (11) for selected values of the parameters. In the upper panel we show the case $\tau_1=2$ and $\tau_2=5$ for three different values of Ω: 2 (black), 3 (red) and 4 (blue). In the middle panel, for the same three Ω values we have $\tau_1=2$ and $\tau_2=7$. Finally, in the lower panel, still at the same Ω's, we report $\tau_1=2$ and $\tau_2=9$.

Model

Our model is inspired to Merton's "Matthew effect", and therefore to Matthew (25: 14–30), which is at the origin of the *success breeds success* effect. However, we also consider Luke's parable of the Ten Mines (Luke, 19:11-27) and the materialist principle of Helvetius (Helvetius, 1772). We assume that there is an equal distribution of talents, abilities and intelligence (all these are considered as synonyms herein) and for that we depart from Matthew which assumes an unequal distribution of abilities. See Table 1 which summarizes the main components of our model.

Note that Luke does not say that individuals have different abilities; he simply does not report anything about the

Table 1 | The main elements of our model

Model	*Initial Productivity* $\alpha+\eta_i$	*Initial Conditions* $x(0)=0$	*Result x(t)*
Gospel	Abilities	Initial Stock	Outcome
Matthew	Unequal	Unequal	Unequal
Luke	Unspecified	Equal	Unequal
Materialism	Equal	—	—
Our model	Almost equal	Equal	Unequal (Log-Normal)

abilities. For this reason we report that an unequal distribution of ability is our interpretation of Luke, on a rational base (Table 1).

Even if a theological interpretation is outside the scope of this paper, a comparative and exegetical analysis of the Gospel of Matthew and Luke shows some differences which are of interest here. Diez Herrera (2003) finds a relevant difference between Matthew and Luke: "pero encontramos también divergencias que no podemos considerar secundarias ya que influyen decisivamente en la interpretación de las parábolas. Así, tenemos primeramente la desigual distribución del dinero entre los siervos que presenta Mateo. Para el lo importante no es que todos reciban la misma cantidad para negociar en igualdad de condiciones (cosa que si aparece en la narración lucana) sino que destaca expresamente que han recibo sumas distintas, y esto, no en virtud de una decisión arbitraria y discriminatoria, sino según su capacidad (Diez Herrera, 2003: 297–298)." That is, the uneven distribution of money among the servants presented by Matthew is purposely related to their ability and not an arbitrary decision. On the other hand, in the Lucan narrative, the important thing is that all receive the same amount to negotiate on equal terms. In particular, this analysis shows some similarity of Luke with Helvetius (1772)'s materialism.

Maggioni (2000) proposes an interpretation of the meaning of the parable of Luke based on the history of the goods left in custody. That is, to take advantage of what God has given you is not simply a matter of preserving it but of *producing fruit*, of being active and productive with enthusiasm and courage. Man is not a simple guardian of God's goods: he/she has the task of trading to multiply them: "Il suo significato [della parabola di Luca] è invece da ricercarsi nella storia dei beni lasciati in custodia. Cioè: sfrutta ciò che Dio ti ha consegnato, perché dovrai renderne conto. E' il tema del giudizio. Che però va ulteriormente precisato: non si tratta semplicemente di conservare, di non perdere, ma di far fruttare. Occorre vivere in attesa di un padrone severo, che vuole raccogliere 'dove non ha seminato', che vuole cioè dall'uomo intraprendenza e coraggio. L'uomo non è un semplice custode dei beni di Dio: ha il compito di commerciare per moltiplicarli (Maggioni, 2000, p. 328–329)".

In our model we adopt Maggioni (2000)'s *entrepreneurial* interpretation of Luke, to be productive, to trade and multiply the goods received in custody to support our hypothesis of the correspondence between productivity and ability/talent/ intelligence. Therefore, in our model, the *operationalization* of scholars' talents (abilities, intelligence) in terms of research productivity is based on Maggioni (2000).

Let's focus on a specific bibliometric indicator, for example, on the total number of papers published by a scholar in her/his whole academic life. None of the concepts introduced in what follows depends on the chosen indicator, and all the considerations and results may apply to any extensive parameter, as for example to the total number of citations received by any author's papers, or to the total IF collected by a scientist.

Let's call $x(t)$ the number of papers published after a time t by a scholar, and define $t=0$ the starting time of their academic career (obviously $x(0)=0$). In order to derive a model for the distribution of x we now need two elements: (1) the time evolution of $x(t)$, and ii) the distribution of the academic ages at the observation time. As we will see, the latter quantity is much less important than the former, at least if no pathological age distributions are chosen.

We first derive a differential equation describing the evolution in time of the variable x(t), which is described in terms of a *productivity* (that is, the number of papers published in a given time), which, in turn, increases with time and is almost the same for all scholars at the beginning of their career. Specifically, the assumptions of the model are the following:

- *Nature gives the same amount of talents to any scholar.* In mathematical terms, productivity at time zero, let's call it α, is the same for all the scholars.
- *A tiny, random, variability of the talents exists.* The previous statement is not strictly true. The initial productivity is $\alpha+\eta_i$, where η_i is a small, addictive, term that depends on the specific scholar *i*. The fluctuation of the initial talent, η_i follows a normal distribution with zero average and standard deviation σ:

$$\begin{aligned} &\langle \eta \rangle = 0 \\ &\langle \eta^2 \rangle = \sigma^2 \\ &\mathcal{P}(\eta) = \frac{1}{\sqrt{2}\sigma} e^{-\frac{\eta^2}{2\sigma^2}} \end{aligned} \tag{1}$$

- *According to a slightly modified version of the* rich get richer *principle, the productivity—not the products—increases proportionally to the amount of products accumulated up to that time.* The *rationale* behind this assumption, which is central to the development of the model, is that the productivity of a scholar is related to her/his *recognition* and *reputation*. It is well known that grant allocation and conference participation, for instance, are based on the international visibility of papers, on their corresponding quality (for example, citations) and on the recognition by the international research community. This is a process which combines quantity and quality. In our model, the *recognition* increases, on average, with the number of papers produced, which in turn allows the scholar to get grants and thus to attract students and Post Docs, who, in turn, will increase her/his productivity. This will increase opportunity to be invited to conferences (with the correlated advertisement of her/his works, publishing additional conference papers, and so on), thereby producing *reputational* cumulative effects. Mathematically, the productivity has a third addendum other than α and η_i, which is $\beta x(t)$, where β has the dimension of an inverse of time. Its inverse ($1/\beta$) represents the *characteristic time* in which the production $x(t)$ increases by a factor e (~2.73). In other words, this parameter specifies how much the *recognition* counts in determining productivity (i.e. the cumulative advantage of reputation generated by recognition). The parameter β indeed determines the value of the productivity (dx/dt) given a collection of output ($x(t)$). The parameter β can also be expressed as the logarithmic increment of production per unit of time: $\beta = dln(x)/dt$. We assume hereafter that β does not depend on the individual characteristics (it does not depend on "*i*"), rather β is the same for all.

Each assumption brings an addendum to the productivity: α, η, and $\beta x(t)$ respectively. The differential equation ruling the time evolution of $x(t)$, thus, is simply the statement that productivity is the sum of the three terms:

$$\frac{dx_i(t)}{dt} = \alpha + \eta_i + \beta x_i(t) \tag{2}$$

where we have retained the pedix i to remember that -due to the presence of the statistical variable η_i—the evolution is different for each individual. This equation is promptly solved, and its solution, with the initial condition $x_i(0)=0$, is:

$$x_i(t) = \frac{\alpha + \eta_i}{\beta}[e^{\beta t} - 1] \tag{3}$$

This equation can be rearranged to be an expression for η_i:

$$\eta_i = \frac{\beta x_i(t)}{[e^{\beta t} - 1]} - \alpha \tag{4}$$

which establishes the identity between a statistical variable η and a quantity which depends on t and x, but that must be equal to η at any time. As we know the distribution function for η (eq. (1)), we can read eq. (4) as change of variable $x \to \eta$, being t a parameter, thus we can work out the distribution function of $x(t)$ via $P(\eta)d\eta = P(x(t))dx(t)$. Therefore:

$$\mathcal{P}(x(t),t) = \frac{d\eta}{dx(t)}\mathcal{P}(\eta) = \frac{\beta}{[e^{\beta t}-1]}\mathcal{P}(\eta) = \frac{\beta}{e^{\beta t}-1}\frac{1}{\sqrt{2\pi}\sigma} exp\left\{-\frac{1}{2\sigma^2}\left(\frac{\beta x(t)}{[e^{\beta t}-1]} - \alpha\right)^2\right\} \tag{5}$$

where we have made explicit that the distribution function $P(x,t)$ not only depends on x, but also explicitly on the time t.

The previous equation represents the statistical distribution of the production $x(t)$, at a given academic time t, for the scholars. Its variability mirrors the small differences in the original productivity associated to the term η. The distribution is a normal distribution, where both mean and standard deviation increase with the time t.

The distribution in eq. (5) depends on the two variables x and t, and has three model dependent parameters: α, β and σ. We can use two of these parameters to scale t and x, and we are therefore left with a single parameter. Defining the scaled time, τ, and the scaled number of papers, ξ, as:

$$\tau = \beta t$$
$$\xi = \frac{\beta t}{\sigma} x \tag{6}$$

and the remaining parameter, Ω, as:

$$\Omega = \frac{\alpha}{\sigma} \tag{7}$$

we get (remembering that $P(\xi) = P(x)dx/d\xi = P(x)\sigma/\beta$):

$$\mathcal{P}(\xi,\tau) = \frac{1}{\sqrt{2\pi}[e^{\tau}-1]} exp\left\{-\frac{1}{2}\left(\frac{\xi}{[e^{\tau}-1]} - \Omega\right)^2\right\} \tag{8}$$

This distribution is normalized, $\int P(\xi,\tau)d\xi = 1$, and its mean and standard deviation are given by

$$\mu_{\mathcal{P}} = \Omega[e^{\tau}-1]$$
$$\sigma_{\mathcal{P}} = [e^{\tau}-1] \tag{9}$$

The second step is to take into account the distribution, let's say $R(\tau)$, of the (scaled) academic ages τ. The distribution of the (scaled) number of papers ξ is therefore:

$$\mathcal{F}(\xi) = \int d\tau \mathcal{R}(\tau)\mathcal{P}(\xi,\tau) \tag{10}$$

In a mature, stationary, world the distribution of the academic ages $R(\tau)$ is stable and, to a good level of approximation, is flat in the time interval between the *average* academic time to reach the specific academic role, and the time to leave this role by promotion (or retirement, if we are considering the full professor role). We are confident that the choice $R(\tau) = \theta([\tau_1 - \tau][\tau - \tau_2])(\tau_2 - \tau_1)^{-1}$, being τ_1 and τ_2 the initial and final (scaled) times for the academic role and $\theta(t)$ the Heavside step function, is a safe approximation at an aggregate level. However, it is well known that this is not exactly the case in centralized academic systems such as the Italian and the French ones (see Lissoni *et al.*, 2011 and Pezzoni *et al.*, 2012). For this reason, we have tested that the results are resilient to modifications of this function, as for example to the smoothing of the harsh discontinuities at τ_1 and τ_2.

In conclusion, we deal with the function:

$$\mathcal{F}(\xi) = \frac{1}{(\tau_1 - \tau_2)}\int_{\tau_1}^{\tau_2} d\tau \mathcal{P}(\xi,\tau) = \frac{1}{\sqrt{2\pi}(\tau_1 - \tau_2)}\frac{1}{}\int_{\tau_1}^{\tau_2} d\tau \frac{1}{[e^{\tau}-1]} exp\left\{-\frac{1}{2}\left(\frac{\xi}{[e^{\tau}-1]} - \Omega\right)^2\right\} \tag{11}$$

Results

In Fig. 1, we show a few examples of the distribution functions obtained in the present paper. These have been obtained by a numerical integration of the expression in eq. (11). Each panel reports three different Ω values (2, black; 3, red; and 4, blue). The different panels refer to different τ_2 values (upper, $\tau_2 = 5$; middle $\tau_2 = 7$; lower $\tau_2 = 9$), while τ_1 is kept fixed to 2. The degree of similarity with the observed Log-Normal distribution depends on Ω, being maximum between $\Omega = 2$ and 3. However, for all the values of the parameters, the present model produces highly skewed distributions.

The present distribution is similar, but not mathematically equivalent to a Log-Normal distribution function:

$$\mathcal{F}(\xi) = \frac{1}{\sqrt{2\pi}\xi\Sigma}\exp\left(-\frac{log^2\left(\frac{\xi}{\mu}\right)}{2\Sigma^2}\right). \tag{12}$$

To better emphasize their similarities, in Fig. 2 we show an example of comparison. We choose a set of parameters for the present model, $\tau_1 = 2$, $\tau_2 = 4$, and $\Omega = 2.5$, and search by a χ^2 minimization, the parameters for the Log-Normal distribution that give the best agreement between the two distributions: $\mu = 41.3$ and $\Sigma = 0.85$.

Having established that the distribution obtained in the present model is undistinguishable from a Log-Normal distribution, it is important to map the set of parameters describing the present model with those describing a Log-Normal. As an example, in Fig. 3 we report the best choice of the Log-Normal' s μ and Σ for each Ω value, for selected τ_1 and τ_2. This mapping has been obtained by a numerical χ^2 minimization.

We illustrate an example of application of the present model to show its ability to represent some real data, although its validity

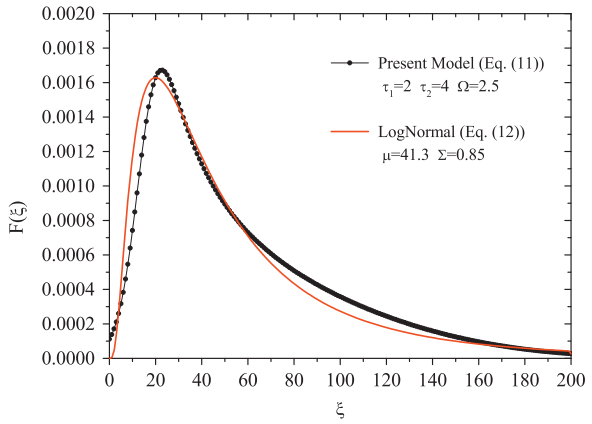

Figure 2 | Comparison of the distribution from eq. (11) and the Log-Normal distribution (eq. (12)). The parameters for the present model are $\tau_1 = 2$, $\tau_2 = 4$, and $\Omega = 2.5$, while the parameters for the Log-Normal distribution, $\mu = 41.3$ and $\Sigma = 0.85$, was chosen to obtain the best agreement between the two distributions.

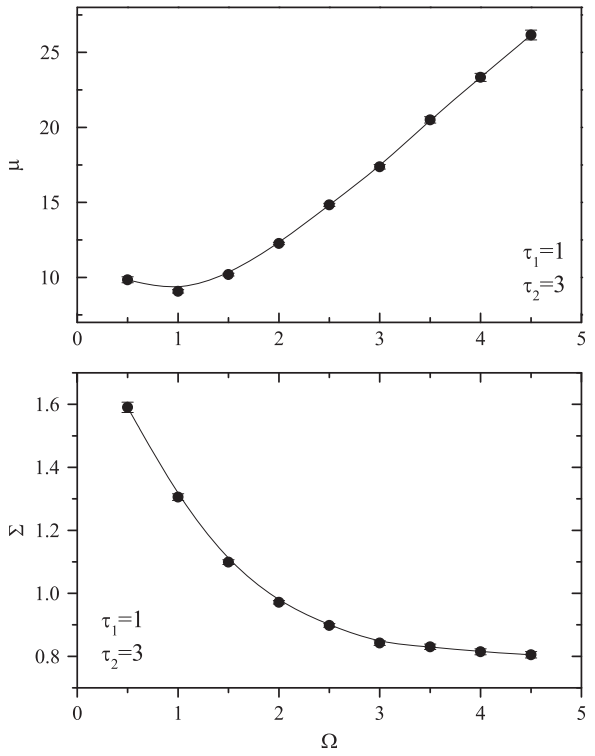

Figure 3 | Example of mapping between the Ω parameter of the present model and the μ and Σ parameters of the Log-Normal distribution that gives the best agreement between the two curves. In the present example, we keep $\tau_1 = 1$ and $\tau_2 = 3$ fixed.

may be broader (see the last section for a discussion on this point). In Fig. 4 we illustrate a comparison between some *experimental* data and the present model. The data, reported as a function of the scaled variable $\xi/\xi o$, represent the distributions of the number of publications, in all of the different disciplines, in a 10 year period, for all the Italian scholars, scaled by their medians, as obtained in Bonaccorsi *et al.* (2017). They studied the scientific production of the universe of Italian academic scholars over a ten-year period across 2002–2012 by using a database built by the Italian National Agency for the evaluation of Universities and Research Institutes. In Italy, each scholar belongs to a disciplinary sector by law. This official classification of scholars separates disciplines according to Life and Hard Sciences (LHS) disciplinary sectors and Social Science and Humanities (SSH) sectors. This classification therefore offers the opportunity to investigate the behavior of scholars without having to create a subjective classification of scholars for the analysis.

For additional information, including descriptive statistics on the data, see Bonaccorsi *et al.* (2017). In Fig. 4, the red points represent the data of scholars belonging to LHS disciplines, the blue points those of the SSH disciplines. The present model has been adjusted to the real data by a numerical χ^2 minimization.

Finally, for practical purposes, we now present an approximation to eq. (11) that leads to a simpler analytic expression for the distribution $F(\xi)$ not involving the numerical integration over τ. In the case the value of $\tau 1$ is large enough, we can exploit the consequences of the approximation $\exp(-\tau) \gg 1$, to rewrite eq. (11) as:

$$\mathcal{F}(\xi) = \frac{1}{\sqrt{2\pi}(\tau_2 - \tau_1)} \int_{\tau_1}^{\tau_2} d\tau e^{-\tau} exp\left\{-\frac{1}{2}(\xi e^{-\tau} - \Omega)^2\right\} \qquad (13)$$

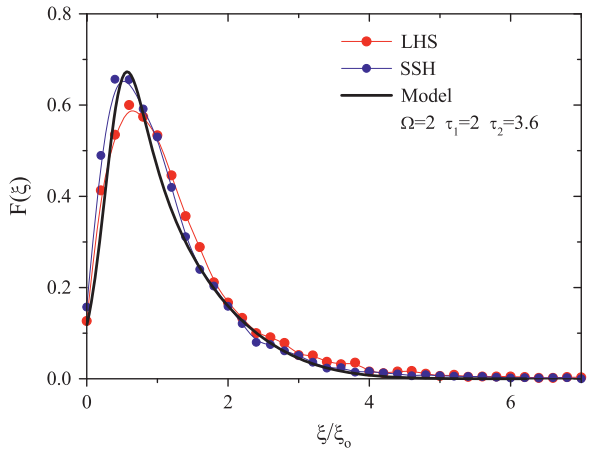

Figure 4 | A comparison between some experimental data and the outcome of the present model. The data, reported as a function of the scaled variable $\xi/\xi o$, represent the distribution of the number of publications in a ten year period for all the Italian scholars, obtained in Bonaccorsi *et al.* (2017) by scaling the distribution of all the different disciplines by their medians. See figure 6 in Bonaccorsi *et al.* (2017). The red points represent the data for scholars belonging to Life and Hard Science disciplines, the blue points those of the Social Science and Humanities disciplines. The present model has been adjusted to the data. The resulting parameters are $\Omega = 2.0$, $\tau_1 = 2.0$, $\tau_2 = 3.6$ and $\xi o = 30$.

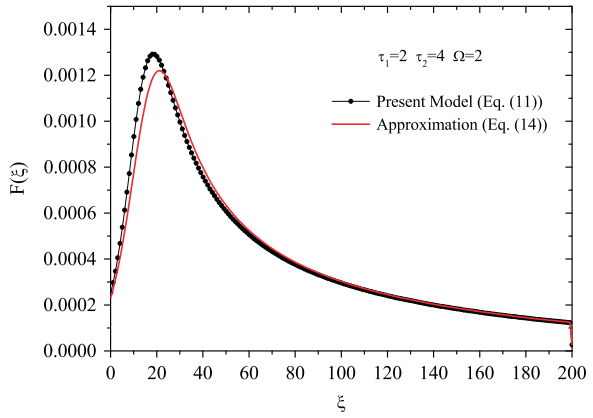

Figure 5 | Comparison of the exact result of the present model from eq. (11) and its approximation, eq. (14), for the selected parameter values: $\Omega = 2$, $\tau_1 = 2$, $\tau_2 = 4$.

Now the integral in this equation can be solved with the substitution $\zeta = \exp(-\tau)$, that is:

$$\mathcal{F}(\xi) = \frac{1}{\sqrt{2\pi}(\tau_2 - \tau_1)} \int_{\exp(\tau_1)}^{\exp(\tau_2)} d\zeta exp\tau\left\{-\frac{1}{2}(\xi\zeta - \Omega)^2\right\} =$$
$$= \frac{1}{(\tau_2 - \tau_1)2\xi}\left[erf\left(\frac{\xi e^{-\tau_1} - \Omega}{\sqrt{2}}\right) - erf\left(\frac{\xi e^{-\tau_2} - \Omega}{\sqrt{2}}\right)\right] \qquad (14)$$

As an example, in Fig. 5 we report the comparison between the exact result in eq. (11) and its approximate counterpart for selected values of the parameters. As expected, for large $\tau_1$the approximation becomes better and better, but already at $\tau_1 = 2$ the two curves appear to be undistinguishable.

Discussion and Conclusion

The distributions reported on the previous section are not coincident with the Log-Normal function, but with this function

they share their main features, to the level that they can be confused. Also, the unavoidable statistical uncertainty of the *experimental* data does not allow us to distinguish the small differences between eq. (11) and a Log-Normal function. In summary, we can talk of eq. (11) as a quasi-Log-Normal distribution. In Fig. 4 we have shown a mapping between the parameters Ω, τ_1 and τ_2 and the genuine Log-Normal parameters and conclude that the distributions observed in the bibliometric parameters in Ruocco and Daraio (2013) and Bonaccorsi *et al.* (2017) can be described by eq. (11) to an high degree of accuracy.

In conclusion, we have presented an (over)simplified model that catches the main features of the observed distributions of different bibliometric indicators. This model is built over the simple assumption that the *natural talent* is (almost) the same for all scholars at the beginning of their career. It is well known that visibility is not just the effect of publication rates. Moreover, it is the effect of only some publications not all of them (Merton, 1968). In our model, only small fluctuations are allowed. These fluctuations inflate with time following the *recognition and reputation* rule à la Merton, mediated by the *entrepreneurial* interpretation of Luke (Maggioni, 2000): the more you publish, the more you are known, the higher your probability of being recognized, the more likely you are to get the right conditions for increasing your publication rate. With these simple ingredients, and with elementary algebra, we derive a functional form that, although not coincident with a Log-Normal function, has all the features of this function, to the extent that they can be confused one with each other. We have called this function *quasi-lognormal*, and we proved that, to any practical purpose, one could use the Log-Normal functional shape to fit the experimental data.

It is worth noting that the assumptions at the basis of the present model, and therefore the implications and the outcome of the model itself, can be extended to other fields outside of the investigation of scientific publishing. The same set of assumptions may apply not only to scientific production, but to numerous other activities as well. Some examples may include the analysis of production and trade, income and wealth distribution, but also more applied political economy matters including public choice or policy advice analyses. In this sense the model may certainly have implications that go beyond the Science of Science.

However, all these considerations leave a tricky issue open: *what do bibliometric indicators really measure?* A discussion on this point follows in the next section.

Policy implications and further research

The investigation on the *ab-initio* causes of the observed empirical distributions of bibliometric indicators is an interesting topic from a philosophical and modelling perspective. On the other hand, policy makers need metrics for, among other things, setting thresholds, establishing criteria of funding allocation or rules for national qualification of scholars. They are not very interested in the philosophical investigation on the origin of the *success breeds success* effect, that is, if all scholars receive the same amount of talents or intelligence, or if they receive different levels of it. Policymakers are mostly interested in understanding what publications and citations really measure; if these metrics are a good proxy of the scientific achievements, ability and efforts of the scholars. For this purpose, our model could provide some hints for further development. According to the hypotheses of our model, the empirical distributions of the bibliometric parameters observed *might be* the result of chance and noise (chaos) related to multiplicative phenomena connected to a *publish or perish* inflationary mechanism, led by the recognition and reputation of scholars. Summing up: being a scholar in the right tail or in left tail of the distribution could have very little connection to her/his merit and achievements. This interpretation might cast some doubts on the use of the number of papers and/or citations as a measure of scientific achievements along the lines of the general critiques against quantitative metrics (see e.g. Wilsdon, 2015, 2016), and may lead to reconsider the method of peer review, despite its well-known limitations.

In the interpretation of our model, however, we follow the deductive induction of Popper (1959). In other words, the assumption of our model about the equality of ability/talents/intelligence, operationalized through an inflationary productivity process, along with the other assumptions, has led to a model that seems to reproduce some observed empirical evidence (Log-Normal distributions). This does not mean that the assumptions of the model (including that of equality of talents) are true, but that simply, according to the *modus tollens*, they are not falsified by our model. A *tricky issue* seems to emerge from this interpretation of our model that is: *what do bibliometric indicators really measure?* The analysis of this issue, calls for deeper investigations on the meaning of the bibliometric indicators. These further analyses are clearly outside the purpose of the present paper. They will require the development of more detailed and accurate models than our (over)simplified model, in which the relationships among intelligence, talents, their historical characterization, ability, merits and their measure (see, for example, Carson, 2007) are more carefully taken into account and modelled. This is an interesting and intriguing topic for further research to be carried out beyond Science of Science and Sociology of Science, including elements and investigation tools from Philosophy, Psychology and Theology. It could also be worthwhile to further investigate from a policy maker's perspective, to understand, model, explain and assess the scholars' behavior and its relation with scientific publication parameters.

Note

1 The Tsallis distribution of a variables x is given by the expression: $P(x) = Po/[1+(\beta-1)(x/\xi)]\beta/(\beta-1)$.

References

Albert R and Barabasi AL (2002) Statistical mechanics of complex networks. *Reviews of Modern Physics*; **74** (1): 47–97.

Anastasiadis AD, de Albuquerque MP, de Albuquerque MP and Mussi DB (2010) Tsallis q-exponential describes the distribution of scientific citations—a new characterization of the impact. *Scientometrics*; **83** (1): 205–218.

Barabasi AL and Albert R (1999) Emergence of scaling in random networks. *Science*; **286** (5439): 509–512.

Barabasi AL, Jeong H, Neda Z *et al.* (2002) Evolution of the social network of scientific collaborations. *Physica A*; **311** (3-4): 590–614.

Bianconi G and Barabasi AL (2001) Competition and multiscaling in evolving networks. *Europhysics Letters*; **54** (4): 436.

Bommarito MJ and Katz DM (2010) A mathematical ap-proach to the study of the united states code. *Physica A*; **389** (19): 4195–4200.

Bonaccorsi *et al.* (2017) Do Social Sciences and Humanities behave like Life and Hard Sciences? *Scientometrics*; **112** (1): 607–653.

Carson J (2007) *The Measure of Merit: Talents, Intel-ligence, and Inequality in the French and American re-publics, 1750-1940*. Princeton University Press: Princeton, NJ.

De Solla Price DJ (1965) Networks of scientific papers. *Science*; **149** (3683): 510–515.

De Solla Price DJ (1976) A general theory of bibliometric and other cumulative advantage processes. *Journal of American Society of Information Science*; **27** (5): 292–306.

Diez Herrera PA (2003) Las Parabolas de los talentos (Mt, 25, 14-30) y de las minas (Lc 19, 11-28): Estudio Comparativo y exegetico. *Isidorianum*; **24** (1): 273–316.

D'Souza RM, Borgs C, Chayes CT, Berger N and Klein-berg R (2007) Emergence of Tempered Preferential At-tachment From Optimization. *Proc. Natl. Acad. of Sci-ences USA*; **104** (15): 6112–6117.

Egghe L (2005) *Power Laws in the Information Production Process: Lotkaian Informetrics*. Elsevier: Oxford.
Eom YH and Fortunato S (2011) Characterizing and model-ing citation dynamics. *PLoS ONE*; **6** (9): e24926.
Glänzel W and Schubert A (2016) From Matthew to Hirsch: A success-breeds-success story. In: Sugimoto CR (ed). *Theories of Informetrics and Scholarly Communication*. de Gruyter: Berlin, Germany, pp 165–179, ISBN 978-3-11-029803-1.
Helvetius. (1772) *De l'homme, de ses facultes intellectuelles et de son Education, London*; Eng. transl., *A Treatise on Man; his Intellectual Faculties and his Education*.
Kryssanov VV, Kuleshov EL, Rinaldo FJ and Ogawa H (2007) We cite as we communicate: A communication model for the citation process. *E-prints arXiv: cs*/0703115.
Laherrere J and Sornette D (1998) Stretched exponential distributions in nature and economy: Fat tails with characteristic scales. *European Physical Journal B*; **2** (4): 525–539.
Lehmann S, Lautrup B and Jackson AD (2003) Citation networks in high energy physics. *Physical Review E*; **68** (2): 026113.
Leskoves J, Kleinberg J and Faloutsos C (2005) Graphs over time: densification laws, shrinking diameters and possible explanations. In: *Proceedings of the International Conference on Knowledge Discovery and Data Mining* (KDD-05). Chicago, IL.
Limpert E, Stahel WA and Abbt M (2016) Log-normal distributions across the sciences: Keys and Clues. *Bio-Science*; **51** (5): 341.
Lissoni F, Mairesse J, Montobbio F and Pezzoni M (2011) Scientific productivity and academic promotion: A study on French and Italian physicists. *Indian and Corporation Changes*; **20** (1): 253–294.
Luke, 19:11-27.
Mahmoud HM (2008) *Polya Urn Models*. Chapman & Hall/CRC: Boca Raton, Florida.
Maggioni B (2000) *Il racconto di Luca*. Cittadella Editrice: Assisi, Italy.
Mandelbrot B (1959) A note on a class of skew distribution functions. *Information and Control*; **2** (1): 90–99.
Matthiew, 25:14-30.
Merton RK (1968) The matthew effect in science. *Science*; **159** (3810): 56.
Mitzenmacher M (2004) A brief history of generative models for power law and lognormal distributions. *Internet Mathematics*; **1** (2): 226–251.
Perc M (2010) Zipfs law and log-normal distributions in measures of scientific output across fields and institutions: 40 years of Slovenia's research as an example. *Journal of Informetrics*; **4** (3): 358–364.
Perc M (2013) Self-Organization of progress across the century of physics. *Scientific Reports*; **3** (1): 1720.
Perc M (2014) The Matthew effect in empirical data. *J. R. Soc. Interface*; **11** (98): 20140378.
Pezzoni M, Sterzi V and Lissoni F (2012) Career progress in centralized academic systems: Social capital and institutions in France and Italy. *Research Policy*; **41** (4): 704–719.
Popper KR (1959) *Logik der Forschung*. Springer: Wien, 1935, English version *The Logic of Scientific Discovery*, Hutchinson, London 1959.
Radicchi F, Fortunato S and Castellano C (2008) Universality of citation distributions: Towards an objective mea-sure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*; **105** (45): 17268–17272.
Raup DM (1985) Mathematical models of cladogenesis. *Paleobiology*; **11** (1): 42–52.
Redner S (1998) How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*; **4** (2): 131–134.
Redner S (2005) Citation statistics from 110 years of physical review. *Physics Today*; **58**, 4954.
Reed WJ and Hughes BD (2007) Theoretical size distribution of fossil taxa: Analysis of a null model. *Theoretical Biology and Medical Modelling*; **4** (1): 12.
Rodriguez-Navarro A (2011) A simple index for the high-citation tail of citation distribution to quantify research performance in countries and institutions. *PLoS ONE*; **6** (5): e20510.
Rousseau R (2010) *Informetric Laws, in Encyclopedia of Library and Information Sciences*, Ronald, Third Edition1:1, 2747–2754.
Ruocco G and Daraio C (2013) An empirical approach to compare the performance of heterogeneous academic fields. *Scientometrics*; **97** (3): 601–625.
Schubert A and Glänzel W (1984) A dynamic look at a class of skew distributions A model with scientometric applications. *Scientometrics*; **6** (3): 149–167.
Seglen PO (1992) The skewness of science. *Journal of the American Society for Information Science*; **43** (9): 628–638.
Simon HA (1955) On a class of skew distribution functions. *Biometrika*; **42** (3/4): 425.
Simon HA (1960) Some further notes on a class of skew distribution functions. *Information and Control*; **3** (1): 80–88.
Sole RV and Montoya JM (2001) Complexity and fragility in ecological networks. *Proceedings of the Royal Society of London Series B*; **268** (1480): 2039–2045.
Sole RV, Pastor-Satorras R, Smith E and Kepler TB (2002) A model of large-scale proteome evolution. *Advances in Complex Systems*; **5** (1): 43–54.
Sole RV and Pastor-Satorras R (2003) Complex networks in genomics and proteomics. In: Bornholdt S and Schuster HG (eds). *Handbook of Graphs and Networks*. Wiley-VCH: Berlin, Germany, pp 145–167.
Stringer MJ, Sales-Pardo M and Amaral LAN (2008) Effec-tiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*; **3** (2): e1683.
Tsallis C and De Albuquerque MP (2000) Are citations of scientific papers a case of nonextensivity? *European Physical Journal B*; **13** (4): 777–780.
Van Raan AFJ (2001a) Two-step competition process leads to quasi powerlaw income distributions - Applica-tion to scientific publication and citation distributions. *Physica A*; **298** (3): 530–536.
Van Raan AFJ (2001b) Competition amongst scientists for publication status: Toward a model of scientific pub-lication and citation distributions. *Scientometrics*; **51** (1): 347–357.
Vazquez A (2003) Growing networks with local rules: Preferential attachment, clustering hierarchy and degree correlations. *Physical Review E*; **67** (5): 056104.
Wallace ML, Lariviere V and Gingras Y (2009) Modeling a century of citation distributions. *Journal of Informetrics*; **3** (4): 296–303.
Whitley R (2000) *The Intellectual and Social Organization of the Sciences*. Oxford University Press: New York.
Wilsdon J (2015) We need a measured approach to metrics. *Nature*; **523** (7559): 129–129.
Wilsdon J (2016) *The metric Tide: Independent Review of the Role of Metrics in Research Assessment and Management*. SAGE: London, UK.
Wouters P and Leydesdorff L (1994) Has pricess dream come true: Is scientometrics a hard science? *Scientometrics*; **31** (2): 193–222.
Yule GU (1924) A mathematical theory of evolution, based on the conclusions of Dr. J. C. Willis. F.R.S. *R Soc Lond. Philosophical Transaction of the Royal Society B*; **213** (1): 21–87.

Data availability

Data sharing is not applicable to this article as no datasets were generated during the current study. Data illustrated in Fig. 4 come from Fig. 6 in Bonaccorsi *et al.* (2017).

Additional information

Competing interests: The authors declare that there are no competing interests.

Reprints and permission information is available at http://www.palgrave-journals.com/pal/authors/rights_and_permissions.html

How to cite this article: Ruocco G, Daraio C, Folli V and Leonetti M (2017) Bibliometric indicators: the origin of their log-normal distribution and why they are not a reliable proxy for an individual scholar's talent. *Palgrave Communications*. 3:17064 doi: 10.1057/palcomms.2017.64.

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