




Ages of cited references and growth of scientific knowledge: an explication of the gamma distribution in business and management disciplines

Anthony G. Stacey¹ 

Received: 3 July 2020 / Published online: 3 November 2020
© Akadémiai Kiadó, Budapest, Hungary 2020

Abstract

The purpose of this study was to assess the gamma distribution as a model of the distribution of ages of cited references in corpora of scientific literature, and to derive inferences from the parameters of the distributions. The ages of cited references in 2867 articles published in 40 distinguished journals in the fields of accounting, economics, finance, management, marketing, operations and information systems, organisation behaviour and human resources were analysed. The distributions of ages of cited references in each subject area were fitted to gamma distributions with the parameters estimated using minimum distance estimation. In contrast to extant literature, it is shown for all subject areas in the study that the goodness-of-fit statistics for gamma distributions were superior to those for lognormal distributions. The rate of growth of knowledge and the temporal profile of the growth of knowledge are derived from the parameters of the gamma distributions and differentiate the subject areas. Longitudinal analysis demonstrates that the gamma distribution models are stable but illustrate the evolution of specific corpora of literature over time. The gamma distribution parameters and derived metrics can be applied diagnostically and descriptively to characterise corpora of literature, or prospectively to set norms, expectations and criteria for new research. The results have implications for future bibliometric studies, authors, editors, reviewers, and knowledge researchers. Opportunities for further research and verification of prior research are created from this novel bibliometric approach.

Keywords Reference distance · Synchronous distribution · Retrospective citation analysis · Lognormal distribution

Introduction

Epistemologically, there is a great deal of diversity across disciplines with regard to the nature of propositional knowledge, and the processes for advancing that knowledge. It is not the intention to delve into the philosophical question of the definition of knowledge,

✉ Anthony G. Stacey
Anthony.Stacey@wits.ac.za

¹ Graduate School of Business Administration, University of the Witwatersrand, Johannesburg, South Africa

but rather to take a pragmatic approach with which present day scientists, academics and researchers can identify.

An approach that has prevailed for many decades and whose legacy continues to influence researchers is the notion that intellectual disciplines can be positioned on a dichotomous scale from *hard* to *soft*. This dimension is supposedly manifest not only in the knowledge itself, but also in the methods used to codify, record, communicate, assimilate, learn, research, analyse, apply, and otherwise utilise the knowledge. Referencing and citation behaviour is perhaps the most accessible means of explaining, characterising, and modelling the cumulativeness of knowledge and the apparent obsolescence of knowledge in the form of published research. Indeed, de Solla Price (1970) explicitly ranked disciplines based on referencing patterns.

It is essential to differentiate between the two complementary modes in which the practice of authors referencing previously published research can be described. One can analyse and evaluate the citing of a specific publication by *subsequent* authors. This is suggestive that the knowledge comprises a portion of the foundation on which each subsequent publication builds. Much of the bibliometric body of knowledge uses this mode of analysis, which is particularly useful in the assessment and evaluation of the influence of particular research outputs. Conversely, one can analyse and evaluate the referencing in a publication to *previously* published work. This mode of analysis alludes to the foundation on which the specific publication itself builds, and is the approach of this study.

Both modes of analysis are found in the literature and typically use descriptive statistics such as the number of citations, the ages of citations, and the decline in the rate of citation with age. In and of themselves, these statistics contribute little to revealing the underlying processes of knowledge accumulation and growth. The distributions of the ages of cited references, being the product of the underlying processes and behaviours, give substantially richer insights and are indicative and diagnostic of these processes and behaviours. This study contributes to bibliometric research which to date has been somewhat equivocal on the appropriateness of various theoretical distributions to describe the distribution of ages of cited references.

Literature review

Hard and soft sciences

Generations of scientists have acknowledged fundamental differences between what have become known as the *hard* and *soft* sciences. Birch (1877, p. 55) for example, contrasts humans' softer attributes with hard science with the words "... while man lives, his imaginative and artistic side no hard science is likely to kill." More recently, the tension between the arts and the sciences is evident when Ransom (1939, p. 197) states that "[in any work of art] its semantic content, the net result in knowledge as hard science would define it, is ... so trifling and commonplace that it is as if the artist were a second-rate parrot."

The perception of hard science as being somehow superior to studies in the arts and humanities is highlighted by Kaplan (1961) who notes the challenge of theory building and confirmation in the field of international politics. The differences between sciences are considered in some detail by Storer (1967) who observes that the extent to which mathematics is applied in the discipline indicates the degree of rigor and therefore the *hardness* of scientific disciplines. They discuss the connotations of the adjectives "hard" and "soft",

and how these may be related to the perceived difficulty of learning in different fields. They allude to a drive in the so-called softer sciences to become more rigorous through the use of mathematics, which is also noted by Lofland (1967).

Burton and Kebler (1960) suggested that published literature has a “half-life” analogous to the decay of radioactive substances, although de Solla Price (1970) considered the rate of decline as less significant than the profile of what he termed the “immediacy effect”. de Solla Price (1970) then proposed the calculation of the proportion of cited references that were less than 5 years old, referred to as Price’s Index, as a more meaningful metric than the half-life. They analysed journals from a wide variety of scientific disciplines and concluded that hard sciences, soft sciences, technology and non-science have decreasing values of Price’s Index.

In the area of psychology, Meehl (1978) identifies and describes 20 characteristics of the discipline that set it apart from the mathematical and statistical aspects of hard (or what they refer to as *developed*) sciences. Cole (1983), acknowledging the long standing hierarchy of sciences, distils six variables that differentiate those at the top of the science hierarchy (i.e. the hard sciences) and those at the bottom. Cole suggests that theory is highly developed, research takes place within a paradigm, and there are high levels of codification of knowledge in the hard sciences. In the soft sciences, ideas are expressed in words whereas mathematical expressions are manifest in the hard sciences, as previously noted by Storer (1967). Cole (1983) further suggests that there are higher levels of consensus (on theory, methods, and significance of problems and individuals’ contributions) and predictability in the hard sciences. Obsolescence and growth of knowledge are of particular relevance to the current study in that the rates of both are deemed to be greater in the hard sciences.

The principle of the cumulateness of science (Nowak 1977) can be separated into *conceptual or theoretical cumulateness* and *empirical cumulateness* (Hedges 1987). Hedges focuses on empirical cumulateness of research in the social and physical sciences. They conclude that there is not a substantial difference in the predictability of experiments in physics (as an exemplar of hard science) and in psychology (as an exemplar of soft science), suggesting that the notion of a hierarchy differentiating hard and soft sciences is not unequivocal.

Growth of knowledge

This study does not adopt a narrow Popperist view of the growth of scientific knowledge through conjectures and refutations (Popper 2014) but rather focuses on the cumulateness of scientific knowledge. This is consistent with what was historically referred to as the third phase of knowledge production: “the digestion of the new knowledge and its absorption into the general mass of information by critical comparison with other experiments on the same or similar subjects” (Mees 1917, p. 1137).

The oft-cited definition of *normal science*, being “research firmly based upon one or more past scientific achievements, achievements that some particular scientific community acknowledges for a time as supplying the foundation for its further practice” (Kuhn 1962, p. 10) is also consistent with the cumulateness of scientific knowledge, albeit within a given paradigm. Small (1999) attributes the notion of structure in science to Kuhn, and noted that paradigm shifts, revolutionary thinking or growth of scientific knowledge are evident in the changes of structure over time (Chen 2003).

Similarly, various authors have referred to the age of cited references as indicative of the rate of growth of scientific knowledge or emerging research field. In ecology and environmental sciences (Jarić et al. 2014) suggested that metrics based on the ages of references can identify emerging scientific fields without the time lag associated with traditional bibliometric indicators based on received citations. Although the unit of analysis for their study was individual researchers not publications nor fields of study, Hamermesh (2018) commented that the age of citations could be an indicator of the rate of growth of knowledge in economics subspecialties.

Distributions of ages of cited references

For clarity and consistency with, for example de Solla Price (1970) and Krauze and Hillinger (1971), a *citation* of a source is a referral to the publication in a subsequent work; a *cited reference* is an older source listed which is referred to in a publication. The age of a cited reference—referred to as the reference distance Δ_r by Pan et al. (2018)—is the difference between the recorded date of a publication and the date of the cited reference. The distributions of ages of cited references is also referred to as the “synchronous distribution” (Nakamoto 1988) or the “retrospective citation approach” (Burrell 2001; Glänzel 2004).

Exponential distribution

The rationale for much early work on the distributions of ages of cited references was to optimise the journal subscriptions and books purchased by academic libraries (e.g. Gross and Gross 1927) and there is an extensive history of applying the exponential distribution in the field of bibliometrics. The suggestion that published literature has a “half-life” (Burton and Kebler 1960) implies an underlying exponential process. In the same vein, de Solla Price (1965) observed that in papers published in 1961, the rate of citation of prior sources decreased by 50% for every 13½ years of age of the cited sources, which also implies an underlying exponential process, noting the relationship with the exponential growth in the number of published papers during the same period.

MacRae Jr (1969, p. 631) commented that the exponential distribution was “good enough to justify the use” to model the distribution of ages of cited references in different disciplines. In addition, references in sociology tend to be older than those in the natural sciences which is in keeping with the findings of de Solla Price (1970).

Various authors (e.g. Krauze and Hillinger 1971) consider the relationship between the exponential growth in the number of publications and the number of cited references; Nakamoto (1988) illustrates the similarity of the distributions of cited references (i.e. synchronous distributions) and the distributions of citations (i.e. diachronous distributions) but evidently overlooks the immediacy phenomenon (de Solla Price 1965) in confirming that “the decrease in citation by age is exponential” (Nakamoto 1988, p. 157). Similarly, Pollman (2000) focuses on the exponential decline in the proportion of cited references after the modal age of cited references, and introduces a correction factor into the simple exponential model to take this into account. Redner (2005) refers somewhat vaguely to the indication that the distribution of ages of cited references in the journal *Physical Review* “roughly decays exponentially with age” (p. 3) but suggestively refers to the goodness-of-fit with the lognormal distribution as “intriguing” (p. 2).

Lognormal distribution

Matricciani (1991) analysed engineering literature from 1988 and concluded that a simple lognormal distribution could be used to model the ages of cited references. They also suggested that the same would apply to literature in the human sciences, and the lognormal distribution should be evident in the contents of libraries and other research archives. This was followed up by Egghe and Ravichandra Rao (1992) who analysed the references cited in three books published in 1979, 1984 and 1990, the results of which indicated that the lognormal distribution is generally applicable to model the ages of cited references.

An analysis of ages of cited references in the field of theoretical population genetics was carried out by Gupta (1997). Having fitted various undisclosed distributions to data at 10 year intervals from 1929 to 1979, they concluded that the best fit on the basis of the Kolmogorov–Smirnov statistic was obtained with a lognormal distribution. Burrell (2002a) came to a similar conclusion comparing the lognormal, Weibull and log-logistic distributions using graphical methods.

It is curious that none of the authors cited above nor Yin and Wang (2017) who provide a reasonably comprehensive analysis of the time dimension of citation as a phenomenon, saw fit to include the gamma distribution in their analyses, particularly as these distributions occurs naturally in processes involving a lapse of time between events.

Alternative distributions

Börner et al. (2004) primarily focussed on modelling the structure and dynamics of the growth of scientific knowledge. In doing so, they note that the Weibull distribution can be used to model the ages of cited references, without referring to Burrell (2002a) nor substantiating why the Weibull distribution may be more appropriate than the lognormal distribution.

In exploring citation distributions and obsolescence, Burrell (2002b) analysed in detail the case where the rate at which citations may be acquired (referred to as the latent rate) has a gamma distribution and obtained results that corresponded with empirical data. Given the likeness of synchronous and diachronous distributions (Nakamoto 1988) it is reasonable to consider the gamma distribution for modelling distributions of ages of cited references.

In a cross-disciplinary study of the ages of cited references in published articles, Stacey (2020) used minimum distance estimation to generate robust estimates of the scale and shape parameters of gamma distributions. These two parameters were shown to characterise research disciplines in a manner similar to but more nuanced than the Price's Index, half-life, and similar metrics. The current study adopts a similar approach to analysing corpora of literature in management disciplines.

Methods

Sample

The unit of analysis for this study is the *subject area* into which published research in business and management fields can be categorised. The sample comprises references

cited in articles published in 2019 volumes and issues of journals that that count towards the Financial Times research rank (so-called FT 50 journals). For the purposes of this study, journals have been grouped into subject areas according to McMaster University (2020). The three subject areas of Entrepreneurship, Ethics and International Business were excluded from the study as there were considered to be too few journals in those subject areas to constitute a representative sample. The subject areas of Organizational Behaviour and Human Resources were aggregated into a single subject area for the same reason. Statistics for the sample data are given in Table 1.

The data themselves (i.e. dates of cited references) were obtained by downloading original research articles or the reference lists from the articles from the journal web sites. These were parsed using rules or algorithms specific to the referencing style used by the journal, to isolate the year of each cited reference. Book reviews, editorials, correspondence, “Point-Counterpoint”, research notes, technical notes, commentaries, and miscellanea were not included in the study. Metadata (e.g. the number of references per article, earliest and latest reference dates) was used to militate against errors that could result from references listed in an inconsistent format or ambiguity in the parsing logic. While the integrity of the data cannot be absolutely guaranteed, due to the number of references, such errors that may exist are randomly distributed through the dataset and do not materially affect the analysis.

Table 1 Statistics for sample data

Subject area	Number of journals	Number of articles	Cited references	Cited references per article
Accounting	6	323	20 090	62.2
Economics ¹	5	497	31 653	63.7
Finance	5	441	23 274	52.8
Management ²	7	607	56 800	93.6
Marketing ³	5	282	19 354	68.6
Operations and Information Systems ⁴	7	781	43 119	55.2
Organisation Behaviour and Human Resources	6	383	33 557	87.6
<i>Scientometrics</i> ⁵	1	259	13 411	51.8
Total ⁶	41	3 126	212 464	68.0

¹Data for *Econometrica* are omitted due to limited access

²The *Harvard Business Review* is excluded as it is a magazine not a journal, and the *MIT Sloan Management Review* is excluded because articles bridge the gap between academia and management practice and references are not cited as comprehensively as in academic articles

³Data for *Journal of Consumer Psychology* are omitted due to limited access

⁴Data for *Journal of Operations Management* are omitted due to limited access

⁵Data for *Scientometrics* for the same period were included for comparative purposes

⁶The totals are not all equal to the sum of the corresponding columns because data for *Research Policy* are included in both the Economics and Management subject areas

Parameter estimation

The three methods of estimating the parameters of a distribution to model the ages of references discussed in Stacey (2020) are the Method of Moments, Maximum Likelihood estimation, and Minimum Distance estimation. It has been argued that Minimum Distance estimation gives more robust estimates of the distribution parameters because both the Method of Moments and Maximum Likelihood estimation are sensitive to extreme values, while the cumulative empirical and theoretical distributions are similarly bounded at 0% and 100% resulting in the Minimum Distance estimates of parameters being relatively insensitive to extreme values.

The basis of Minimum Distance estimation is adjusting the distribution parameters in order to minimise the Kolmogorov–Smirnov statistic, which is given in Eq. (1).

$$D_{\max} = \sup_x \left| \text{CDF}(x) - \widehat{\text{CDF}}(x) \right| \tag{1}$$

where

$$\text{CDF}(x_i) = \frac{\text{Rank}(x_i) - 0.5}{n}$$

with Rank(x_i) corrected for duplicate values, where necessary. This was carried out using the Solver macro in MS Excel (Frontline Systems Inc., 1990–2009).

The age of a reference is the difference between the date of publication of the citing article and the date of publication of the reference. The apparent paradox of fitting a continuous distribution (e.g. lognormal or gamma distributions) to data that are essentially discrete has been highlighted by Clauset et al. (2009). In this study, the date of publication of the citing article is based on the year and month of the issue in which the citing article appears; an earlier date on which the article may have first appeared online is disregarded. The date of publication of the reference is modelled as the year of publication plus a uniformly distributed random number between 0 and 1. The continuous distribution is fitted to the continuous simulated age variable, and Monte Carlo simulation is then used to estimate the distribution parameters.

Confidence interval calculations

Boundaries of the confidence regions resulting from the Monte Carlo simulation for k_i and θ_i for each of the subject areas are formed by all points p satisfying the following identity:

$$(\bar{x} - p)' S^{-1} (\bar{x} - p) = F^* \tag{2}$$

where $\bar{x} = \begin{bmatrix} \bar{k} \\ \bar{\theta} \end{bmatrix}$, $p = \begin{bmatrix} p_k \\ p_\theta \end{bmatrix}$, $S = \begin{bmatrix} s_k^2 & s_{\theta k} \\ s_{k\theta} & s_\theta^2 \end{bmatrix}$, $F^* \cong \frac{2(n-1)}{(n-2)} F_{2,n-2}(1 - \alpha)$ and the $CL = (1 - \alpha)$.

The actual value of F^* is determined such that the proportion of Monte Carlo iterations falling within the boundaries of the confidence region is equal to the CL. By convention, CL = 95% has been used throughout this study.

Table 2 Estimated parameters for corpora of literature based on 1000 Monte Carlo iterations

Subject area	Price's Index PI	Mean age of cited refer- ences, \bar{X}	Gamma distribution			Lognormal distribution $ D _{\max}$
			Estimated shape, \hat{k}	Estimated scale, $\hat{\theta}$	$ D _{\max}$	
Accounting	0.1931	14.34	1.590	8.897	0.0100	0.0286
Economics	0.2574	13.89	1.334	9.921	0.0144	0.0203
Finance	0.2325	13.94	1.457	9.351	0.0146	0.0248
Management	0.1898	15.37	1.567	9.638	0.0072	0.0284
Marketing	0.2258	14.58	1.414	10.161	0.0095	0.0285
Ops & Info Sys	0.2107	13.87	1.572	8.523	0.0163	0.0174
OB and HR	0.1754	15.76	1.598	9.644	0.0094	0.0258

The corresponding statistics for volumes 118 to 121 of *Scientometrics* published in 2019 are $PI=0.3713$, $\bar{X}=10.71$, $\hat{k}=1.184$, $\hat{\theta}=8.448$, $|D|_{\max}=0.0238$ for the gamma distribution and $|D|_{\max}=0.0220$ for the lognormal distribution

Results

The statistics (Price's Index and mean age of cited references) and the estimated gamma distribution shape and scale parameters for the ages of references cited in the corpora of literature in each of the seven identified subject areas are given in Table 2. The table also gives the Kolmogorov–Smirnov goodness-of-fit statistics for the gamma distribution and, for comparative purposes, for the lognormal distribution. The p -values for the Kolmogorov–Smirnov goodness-of-fit statistics are omitted because they would be misleading on account of the well-documented pitfalls of inference from large samples (e.g. Azevedo and de Lima Junior 2019; Baird and Harlow 2016; Chatfield 1995; Lin et al. 2013).

It is noted that although the detailed results are not quoted in Table 2, the results of paired t tests indicate that the goodness-of-fit statistics for the gamma distributions are significantly less (i.e. better) than the goodness-of-fit statistics for the lognormal distributions, for all seven subject areas ($p_i \ll 0.0001 \forall i$).

The empirical distributions and cumulative distributions are illustrated in Fig. 1 for all seven subject areas, together with the corresponding fitted gamma and cumulative gamma distributions; that is, with the parameters that are given in Table 2.

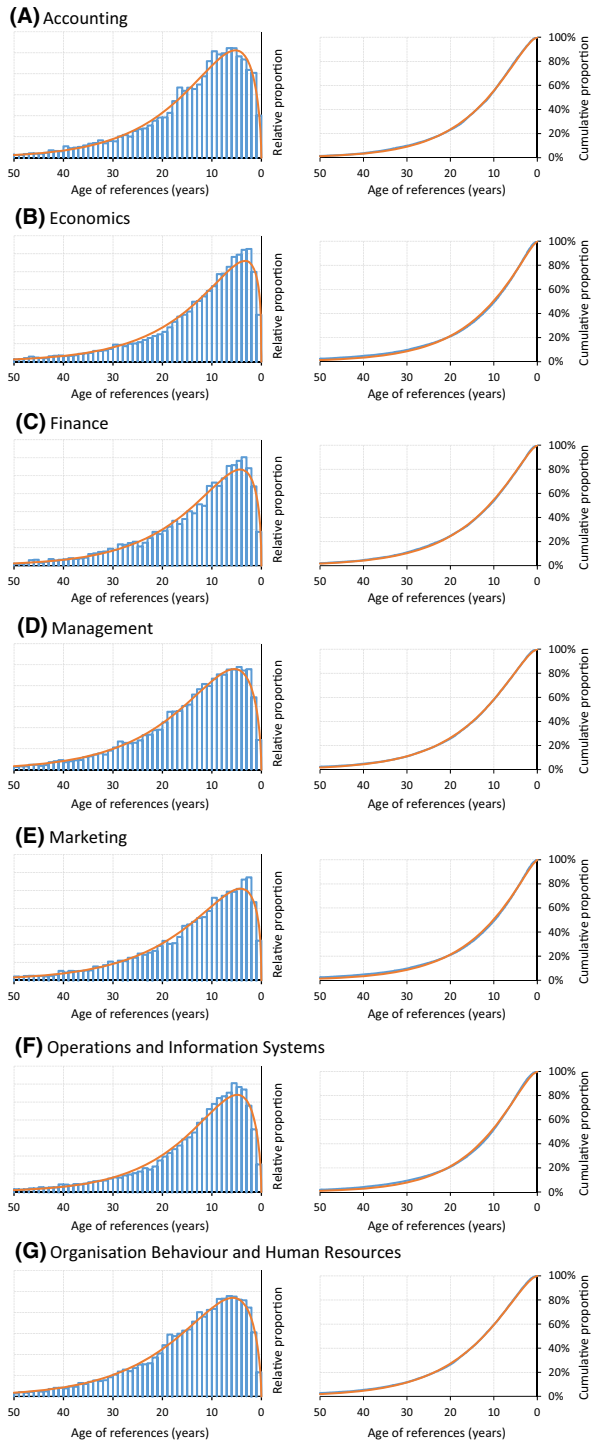
It has been noted that the shape and scale parameters of the gamma distributions fitted to the distributions of ages of cited references in the various corpora of literature have been estimated using a Monte Carlo method. The confidence regions for these parameters based on the Monte Carlo analysis are illustrated in Fig. 2 using $CL=95\%$.

Discussion

Direct interpretation of gamma distribution parameters

The result that the goodness-of-fit statistics for the gamma distributions are significantly better than those for the lognormal distributions supports the finding that the gamma

Fig. 1 Empirical and cumulative distributions of ages of references per subject area with corresponding gamma distributions



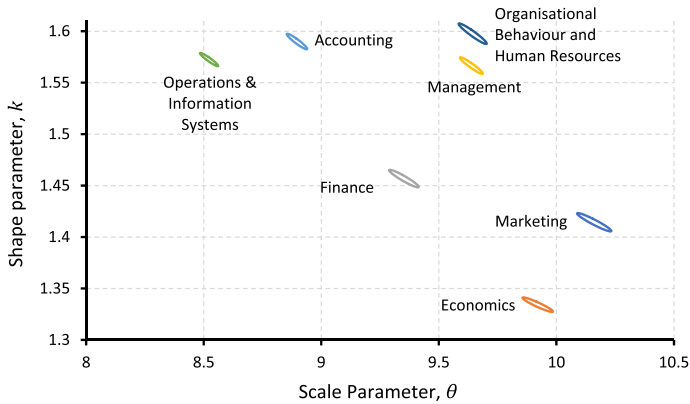


Fig. 2 Confidence regions for the scale and shape parameters by subject area (CL=95%)

distribution is at least as good as a lognormal distribution to model the ages of cited references (Stacey 2020).

The subject areas comprising this study can be ranked from relatively “hard” sciences to more “soft” sciences in terms of the Price’s Index (de Solla Price 1970) as follows:

1. Economics
2. Finance
3. Marketing
4. Operations and Information Systems
5. Accounting
6. Management
7. Organisational Behaviour and Human Resources

This corresponds to a large extent with increasing mean age of references, as expected (Egghé 1997), and with the variables identified by Cole (1983). However, the parameterisation of the distributions of ages of cited references gives more nuanced distinctions between subject areas.

The scale parameter has been tentatively linked to the average time interval between consecutive research publications in the specific thread (Stacey 2020). Considering only the horizontal axis in Fig. 2 it is evident that the subject areas of Management and Organisational Behaviour and Human Resources have similar such time intervals. That their respective Price’s Indices are not substantially different may therefore be expected. However, Operations and Information Systems has the lowest scale parameter of the seven subject areas, while Marketing has the highest, even though their respective Price’s Indices are *not* substantially different.

Similarly, the shape parameter has been linked to the recency of cited references, describing the citing of contemporary or potentially previously uncited references—corresponding to a relatively lesser shape parameter—or a tendency to rely more on recognised, previously cited references corresponding to a relatively greater shape parameter (Stacey 2020). The analysis suggests that articles in the Economics subject area have the greater propensity to cite contemporary or potentially previously uncited references, which is consistent with having the greatest Price’s Index. Articles in Organisational Behaviour and

Human Resources having the greater tendency to rely more on recognised, previously cited references which is also consistent with having the lowest Price's Index.

Supplementary illustrative examples of modelling ages of cited references

Without endeavouring to carry out a comprehensive cross-disciplinary study, additional corpora of literature have been selected and analysed to illustrate the modelling of ages of cited references with the gamma distribution. The empirical and cumulative distributions of ages of references per selected corpus plus corresponding gamma distributions are illustrated in Fig. 3. The gamma distribution parameters have been estimated using minimum distance estimation. Figure 3a is based on the data table given in Brookes (1972); Fig. 3b–d are based on the data tables presented in Egghe and Ravichandra Rao (1992) for books published by Cairns (1979), Egghe (1984) and Egghe and Rousseau (1990) respectively; Fig. 3e is based on Fig. 1 in Förster et al. (2018, p. 1281); and Fig. 3f is based on articles published in volumes 118 to 121 of *Scientometrics* (i.e. during the same period as the corpora of management literature in the study sample).

The differences in the profile or shape of the empirical distributions of ages of cited references result in dissimilar shape and scale parameters, and gamma distributions. The value of the shape parameter for the gamma distribution fitted to the dataset illustrated by Förster et al. (2018) in Fig. 3e is reasonably close to one, and is consistent with their observation that the distribution shows exponential growth up to the most recent few years. However, their claim that this can be observed empirically for all research topics is not supported by the dataset charted in Fig. 3c from Egghe (1984) as cited in Egghe and Ravichandra Rao (1992) and should be regarded with some circumspection.

It is interesting to note that the distribution of ages of cited references in *Scientometrics* in 2019 illustrated in Fig. 3f is quite distinct from those of the various management disciplines for the same time period illustrated in Fig. 1. The values of both the shape and the scale parameters are less than corresponding values for all the subject areas included in this study.

Inferences derived from gamma distribution parameters

Journal articles were the unit of analysis in the study by Stacey (2020); subject areas are the unit of analysis in this study. The finding that gamma distributions can model the distributions of ages of references in individual articles *and* in the corpora of literature in all seven subject areas may suggest an alternative interpretation of the gamma distribution parameters. Analysis of individual journal articles may be interpreted as shedding light on the research practice of the researchers and authors. It is suggested here that analysis of corpora of literature should rather be diagnostic of the respective bodies of knowledge and the process of knowledge creation within those disciplines.

Cited references in a corpus of literature are a sub-set of the body of scientific knowledge in the corresponding discipline. Assuming (i) that all sources within a particular body of scientific knowledge have an equal opportunity to be cited, and (ii) that the distribution of impacts of sources is time-invariant, it follows that the distribution of the ages of cited references in a corpus of literature is equivalent to the age distribution of the corresponding body of knowledge. Both of the foregoing assumptions are reasonable within a practical time domain.

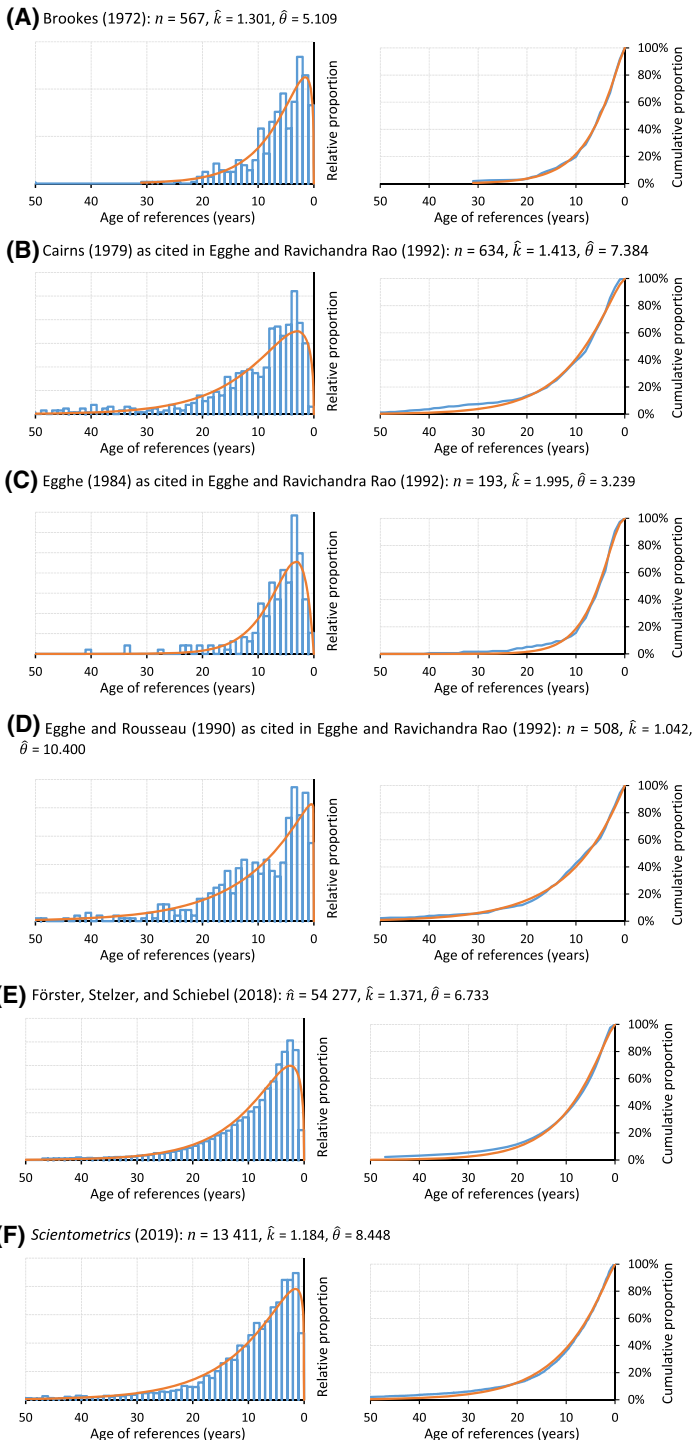


Fig. 3 Empirical and cumulative distributions of ages of references per selected corpus with corresponding gamma distributions

Let X represent the age distribution of the discipline body of knowledge at a given point in time, t_0 . Then, on the basis of the results of the foregoing analysis:

$$X \sim \Gamma(k, \theta) \tag{3}$$

The values of the shape and scale parameters can be estimated by analysing the cited references preceding time t_0 , and the magnitude of the body of knowledge at time t is

$$x(t) = \frac{A}{\Gamma(k)\theta^k} t^{k-1} e^{-\frac{t}{\theta}} \tag{4}$$

where A is an arbitrary constant symbolising the magnitude of the discipline body of knowledge. Now let X' represent the age distribution of the discipline body of knowledge at an earlier point in time, $t_0 - \Delta t$, $\Delta t > 0$. Using the same assumptions as noted previously, it follows that:

$$X' \sim \Gamma(k', \theta'). \tag{5}$$

Because $\lim_{\Delta t \rightarrow 0} k' = k$ and $\lim_{\Delta t \rightarrow 0} \theta' = \theta$, it can be said that for small Δt

$$X' \sim \Gamma(k, \theta)$$

If $\gamma > 0$ represents a proportion of growth of the discipline body of knowledge over time Δt , then

$$\begin{aligned} x'(t) &= \frac{1}{(1 + \gamma)} \frac{A}{\Gamma(k)\theta^k} (t - \Delta t)^{k-1} e^{-\frac{t-\Delta t}{\theta}}, \quad t \geq \Delta t \\ x'(t) &= 0, \forall t < \Delta t \end{aligned} \tag{6}$$

Based on the premises that scientific knowledge is not destroyed and that growth of scientific knowledge over a given time period is finite, the following conditions must apply:

$$\lim_{t \rightarrow \infty} \frac{x(t)}{x'(t)} = 1 \tag{7}$$

and

$$\frac{x(t)}{x'(t)} \geq 1, \quad \forall t \geq 0 \tag{8}$$

From Eqs. (4), (6) and (7)

$$\begin{aligned} \lim_{t \rightarrow \infty} \frac{x(t)}{x'(t)} &= \lim_{t \rightarrow \infty} \frac{\frac{A}{\Gamma(k)\theta^k} t^{k-1} e^{-\frac{t}{\theta}}}{\frac{1}{(1+\gamma)} \frac{A}{\Gamma(k)\theta^k} (t - \Delta t)^{k-1} e^{-\frac{t-\Delta t}{\theta}}} \\ &= (1 + \gamma) \lim_{t \rightarrow \infty} \frac{t^{k-1} e^{-\frac{t}{\theta}}}{(t - \Delta t)^{k-1} e^{-\frac{t-\Delta t}{\theta}}} \\ &= (1 + \gamma) \lim_{t \rightarrow \infty} e^{\frac{(t-\Delta t)-t}{\theta}} \\ &= (1 + \gamma) e^{-\frac{\Delta t}{\theta}} = 1 \end{aligned}$$

$$\therefore \gamma = e^{\frac{\Delta t}{\theta}} - 1 \quad (9)$$

Let us now define the *quasi-mean age of cited references* (μ_r) as the mean of the fitted gamma distribution and is equal to the product of the shape and scale parameters.

$$\mu_r = k \cdot \theta$$

then from (9)

$$\gamma = e^{\frac{\Delta t k}{\mu_r}} - 1 \quad (10)$$

Equation (10) confirms the association between mean age of cited references and growth of knowledge in the previously cited studies by Jarić et al. (2014) and Hamermesh (2018), although not necessarily as these authors envisaged it.

Now from Eqs. (4) and (6):

$$\begin{aligned} \frac{x(t)}{x'(t)} &= \frac{\frac{A}{\Gamma(k)\theta^k} t^{k-1} e^{-\frac{t}{\theta}}}{\frac{1}{(1+\gamma)} \frac{A}{\Gamma(k)\theta^k} (t-\Delta t)^{k-1} e^{-\frac{t-\Delta t}{\theta}}} \\ &= (1+\gamma) \frac{t^{k-1}}{(t-\Delta t)^{k-1}} \frac{e^{-\frac{t}{\theta}}}{e^{-\frac{t-\Delta t}{\theta}}} \\ &= (1+\gamma) \frac{t^{k-1}}{(t-\Delta t)^{k-1}} \frac{e^{-\frac{t}{\theta}}}{e^{-\frac{t-\Delta t}{\theta}}} \\ &= (1+\gamma) \frac{t^{k-1}}{(t-\Delta t)^{k-1}} \frac{1}{(1+\gamma)} \quad \because \gamma = e^{\frac{\Delta t}{\theta}} - 1 \\ &= \left(\frac{t}{t-\Delta t} \right)^{k-1} \geq 1, \quad \forall t \geq 0; \quad k \geq 1 \end{aligned}$$

Therefore the condition in (8) is satisfied for all values of k greater than or equal to 1. This formulation of the growth of knowledge suggests that scientific knowledge will grow exponentially, which is consistent with the exponential growth observed by de Solla Price (1965) and others, and what Popper (2014, p. 216) referred to as “the infinity of our ignorance”.

It is pertinent to note that if the ages of cited references are modelled as lognormal distributions, that is:

$$X \sim \text{Lognormal}(\mu, \sigma^2)$$

$$x(t) = \frac{A}{t\sigma\sqrt{2\pi}} e^{\left[-\frac{(\ln t - \mu)^2}{2\sigma^2}\right]}$$

and

$$X' \sim \text{Lognormal}(\mu, \sigma^2)$$

$$x'(t) = \frac{1}{(1+\gamma)} \frac{A}{(t-\Delta t)\sigma\sqrt{2\pi}} e^{\left[-\frac{(\ln(t-\Delta t) - \mu)^2}{2\sigma^2}\right]}$$

then it can be shown that the condition specified in (7) can only be satisfied if $\gamma=0$ (i.e. there is zero growth of knowledge) and the condition specified in (8) *cannot* be satisfied. It is also relevant to note that simultaneously published cited references cannot be modelled using a lognormal distribution (Hu et al. 2020) because the logarithm of zero is undefined. These analyses suggest that despite many studies illustrating a satisfactory goodness-of-fit with empirical data, the lognormal distribution is *inappropriate* to model the ages of cited references.

Characterisation of scientific bodies of knowledge

Given the derivation above, a body of scientific knowledge can be described in terms of the rate of cumulative growth and the mean age of that body of knowledge, with both these statistics being calculated from the estimated shape and scale parameters of the gamma distribution fitting the ages of cited references in the corresponding corpus of literature. The growth of a body of scientific knowledge over time Δt can be expressed as a function of age, t , as the difference between the body of knowledge at time t_0 and the body of knowledge at time $t_0 - \Delta t$; that is:

$$g_{\Delta t}(t) = x(t) - x'(t) \tag{11}$$

Distributions of the growth of knowledge by age are illustrated in Fig. 4 for relatively low growth (10.5% per annum) and high growth (12.5% per annum), and for quasi-mean ages of cited references (μ_r) of 12 and 15 years. These can be considered to be the profiles of the research front (de Solla Price 1970).

In Fig. 4 it can be seen that the research front is most pronounced for a lower growth rate and lower quasi-mean ages of cited references, in that a large proportion of the growth occurs with recently published work being accumulated into the body of knowledge. Conversely, the novel contributions to a particular body of knowledge are from a

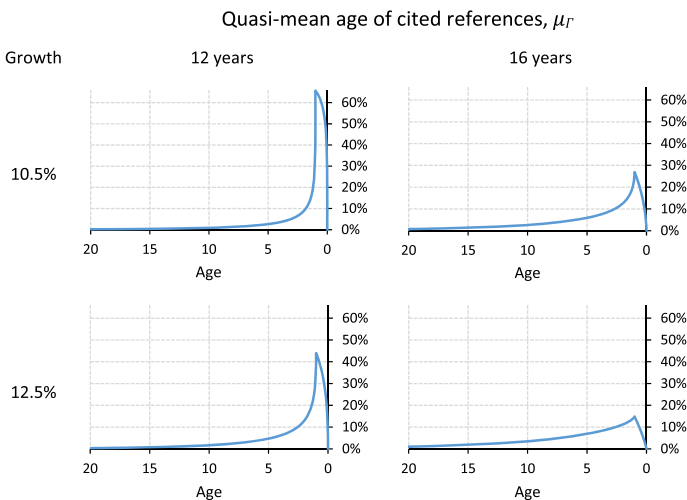


Fig. 4 Distributions of the growth of scientific knowledge as a function of age

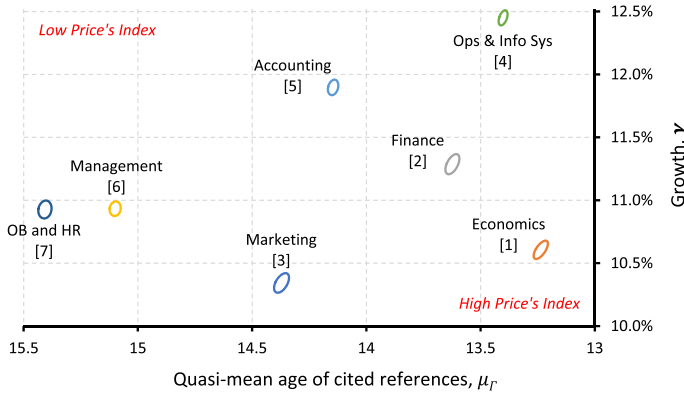


Fig. 5 Confidence region for the growth and quasi-mean age of cited references by subject area (CL=95%)

substantially wider age range for greater growth rates, and for greater quasi-mean ages of cited references.

The confidence regions based on the Monte Carlo analysis for the quasi-mean age of cited references and growth rate by subject areas are illustrated in Fig. 5 using CL=95%. The numerals in parentheses in the subject area labels are the rankings of the subject areas in descending order of Price’s Index.

It is evident that the growth rate of scientific knowledge in Economics and Marketing is relatively low (approximately 10.5% per annum) while that for Operations and Information Systems is relatively high (12.5% per annum). The quasi-mean of age of cited references appears independent of the growth rate with Economics, Operations and Information Systems characterised by relatively younger references and Management, Organisation Behaviour and Human Resources characterised by relatively older references. There is no discernible relationship between the rankings of the subject areas by Price’s Index and either knowledge growth rate or age of references.

Table 3 presents descriptive statistics for bodies of knowledge based on the corpora of literature published in the corresponding subject area in 2019. It is noteworthy that the

Table 3 Descriptive statistics for bodies of knowledge by subject area

Subject area	\hat{k}	$\hat{\theta}$	μ_r	γ (%)	$\widetilde{g}_1(t)$
Accounting	1.590	8.897	14.15	11.90	3.230
Economics	1.334	9.921	13.24	10.61	1.480
Finance	1.457	9.351	13.62	11.29	2.294
Management	1.567	9.638	15.10	10.93	3.256
Marketing	1.414	10.161	14.37	10.34	2.101
Ops & Info Sys	1.572	8.523	13.40	12.45	2.982
OB and HR	1.598	9.644	15.41	10.93	3.525

$\widetilde{g}_1(t)$ = Median age of the growth of the body of knowledge in the subject area, with $\Delta t = 1$ year

The corresponding statistics for volumes 118 to 121 of *Scientometrics* published in 2019 are $\hat{k} = 1.184$, $\hat{\theta} = 8.448$, $\mu_r = 10.00$, $\gamma = 12.57\%$ and $\widetilde{g}_1(t) = 0.831$

ranking of median age of the growth of the body of knowledge in each subject area ($\widetilde{g}_1(t)$) correlates closely with rankings of the subject areas by Price’s Index. This validates the value that has been found in applying Price’s Index in bibliometric analyses, but demonstrates that the two parameter characterisation of subject areas is more nuanced than possible with univariate metrics, including the Price’s Index and half-life of references (Stacey 2020).

Longitudinal analysis of legacy datasets

Two legacy datasets have been selected for longitudinal analysis of ages of cited references by applying the gamma distribution, in order to assess the stability over a period of time. Gupta (1997) tabulated the number of number of references by age at decade

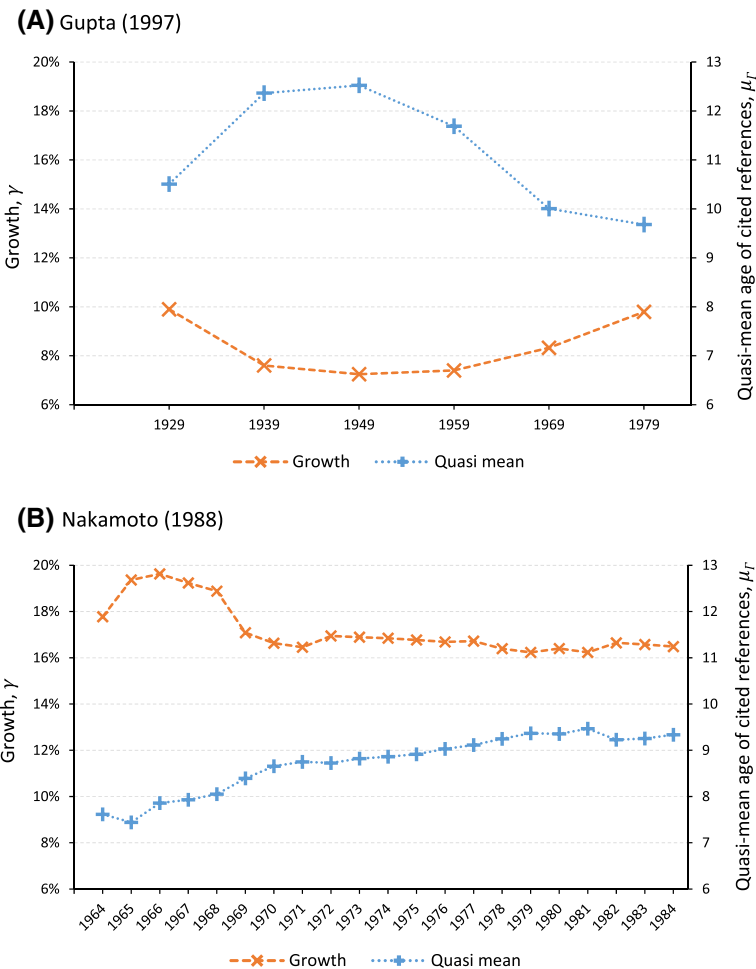


Fig. 6 Trends of growth rate and quasi-mean age of cited references for legacy datasets. **a** Gupta (1997), **b** Nakamoto (1988)

intervals for the period 1929–1979 inclusive and the results of the analysis by fitting the gamma distributions are illustrated in Fig. 6a. A strong inverse relationship is evident between the knowledge growth rate and the quasi-mean age of cited references. Despite the growth rate increasing from 8.33 to 9.79% between 1969 and 1979 and the quasi-mean age of cited references decreasing from 10.0 to 9.7 years over the same time interval, the change in Price's Index from 0.451 to 0.452 is negligible.

Nakamoto (1988, p. 162) tabulated the chronological distribution of citations as percentages from 1961 to 1984 based on the Science Citation Index of 1985. Each column of the table represents with the distribution of ages of cited references for the given year, which have been fitted to gamma distributions and analysed as described in this paper and the results illustrated in Fig. 6b. The inverse relationship between the knowledge growth rate and the quasi-mean age of cited references is less evident than in the previous case. Between 1969 and 1984, the growth rate of knowledge remained in the range of 16–17% while, except for a decrease from 1981 to 1982, the quasi-mean age of cited references increased consistently from 8.4 to 9.3 years over the same period. The Price's Index also decreased consistently except for the same anomaly from 1981 to 1982. The strong autocorrelation at lag 1 year for both growth rate and quasi-mean age of cited references is evidence of the stability of the application of gamma distributions that was not evident in the cross-sectional analyses presented earlier.

Limitations of the study

The scope of this study encompassed original research articles published in selected journals in the fields of accounting, economics, finance, management, marketing, operations and information systems, organisation behaviour and human resources. The following limitations are therefore noted:

- Articles in journals in scientific disciplines other than those listed above were intentionally delimited out of the study.
- Similarly, journals in the above mentioned discipline that are not on the so-called FT-50 list were intentionally delimited out of the study.
- The study was cross-sectional and only analysed in detail cited references from 2019.
- Publication formats such as theses, dissertations, monographs, books, commentaries, book sections, conference papers, etc. were intentionally delimited out of the study.
- Confidence intervals for the distribution parameters for corpora of literature were estimated using 1000 Monte Carlo iterations. This number of iterations was deemed to provide sufficient precision for the purposes of this study but could be increased.
- For pragmatic reasons, the publication date of articles has been taken to be the month and year of the issue in which the articles appear, while the dates of cited references is taken to be the year of publication. A limitation on the accuracy of the ages of cited references arises because these dates do not coincide exactly with the completion and contributions to knowledge of the respective studies.
- The analysis does not attempt to evaluate the pertinence of sources in the citing article and all cited references are equally weighted, although realistically this would not be the case.

Conclusion

The distributions of cited references and citations have been the subject of much research for many decades. Researchers have observed and analysed the exponential growth in the number of publications and cited references, and the apparent exponential fall off in the number of citations received and cited references with increasing age of publications. The time taken to publish and disseminate new knowledge has confounded exponential models and the lognormal distribution has been successfully applied in various contexts.

This study has applied the gamma distribution to the ages of cited references in various corpora of management literature and has demonstrated that the gamma distribution outperforms the lognormal distribution in these instances. The rate of growth and the profile of growth of the body of knowledge can then be derived from the gamma distribution parameters simply by assuming that distribution to the ages of cited references to be a proxy for the age distribution of the corresponding body of knowledge. These insights provide new perspectives on how knowledge is accumulated, documented, and communicated. Obsolescence of publications has been the subject of much research but is not the subject of this article. Nevertheless it is interesting to note that publications become obsolete over time as they comprise an ever decreasing proportion of the total body of knowledge. This manifests because the growth rate of the body of knowledge exceeds the rate at which the number of cited references (i.e. the product of the number of publications and the mean references per publication) increases.

There are several benefits and implications of the finding that ages of cited references fit a gamma distribution better than a lognormal distribution. First, the inference of the growth of knowledge from the gamma distribution scale parameter is a material benefit derived from this result. Then, given the better goodness-of-fit with empirical referencing data, the accuracy of modelling referencing behaviour and growth of knowledge will be improved by applying a gamma distribution rather than a lognormal distribution. Furthermore, the gamma distribution parameters used to characterise corpora of literature are more intuitive and simpler to interpret than those of a lognormal distribution. Specifically, the mean and modal age of cited references are directly proportional to the gamma distribution shape parameter, while the corresponding relationship for the lognormal distribution is considerably more complex. The simpler interpretation of the gamma distribution parameters facilitates inferences, comparisons, and similar purposes. Finally, the use of the gamma shape and scale parameters as metrics of the recency of the references and time lapse between influential publications in a specific field respectively (Stacey 2020) needs further research and validation but constitutes a meaningful contribution to understanding research behaviour.

The metrics derived from the gamma distribution parameters can be applied retrospectively to characterise and differentiate corpora of knowledge, subject areas, disciplines, and subspecialties within disciplines. These metrics may also be used prospectively. For example, researchers may use these metrics to suggest when pertinent sources have been published in their particular discipline, or to benchmark their work against the norms of their discipline. Journal editors may choose to suggest appropriate referencing criteria for authors or to differentiate their publications on the basis of the knowledge growth metrics. Institutions of higher learning and examiners of postgraduate research dissertations and theses may prescribe criteria against which contributions to knowledge are assessed.

There remains much research to be undertaken to realise the full benefit of this perspective on referencing and citing behaviour, and the growth of knowledge. Comparative

cross-sectional and longitudinal studies need to be carried out to document and interpret knowledge accumulation across diverse scientific disciplines and subspecialisations, across academic journal titles, by author, by publication format (i.e. dissertations, theses, monographs, articles, books, etc.), across institutions and geographic regions of origin, and within the “IMRaD” (introduction, methods, results, and discussion) structure of publications (Bertin et al. 2016). Further research is recommended using more accurate dates of the citing article and the cited references, and to explore the practicability of introducing a parameter into the model to account for the lapse of time due to review and publication processes.

Previous research should be revisited from the perspective of the gamma distribution, the parameters, and derived metrics. For example, Stacey (2020) was able to identify relatively few instances of citation of *ephemeral* and *classic* literature (Burton and Kebler 1960) while the decreasing propensity to cite very recent and very old literature (Pan et al. 2018) would be consistent with an increasing value of the gamma shape parameter over time, and could easily be verified. A key challenge that has not yet been unequivocally addressed is the identification of emerging research areas and paradigm shifts (Kuhn 1962). It remains to be seen if publications in emerging research fields or that represent paradigm shifts cite prior sources whose ages can be modelled with uncharacteristically low gamma distribution parameters, or for which the gamma distribution simply cannot provide a satisfactory goodness-of-fit.

Funding No specific funding was received for this research.

Availability of data and material All data used in the study is in the public domain.

Code availability Not applicable.

Compliance with ethical standards

Conflicts of interest The author has no conflicts of interest.

References

- Azevedo, H. V. F., & de Lima Junior, E. T. (2019). Statistical inference techniques applied to large samples. In *Paper presented at the XL Ibero-Latin-American Congress on Computational Methods in Engineering*, Natal, Rio Negro, Brazil.
- Baird, G. L., & Harlow, L. L. (2016). Does one size fit all? A case for context-driven null hypothesis statistical testing. *Journal of Modern Applied Statistical Methods*, 15(1), 100–122.
- Bertin, M., Atanassova, I., Gingras, Y., & Larivière, V. (2016). The invariant distribution of references in scientific articles. *Journal of the Association for Information Science and Technology*, 67(1), 164–177.
- Birch, S. (1877). Our portrait gallery. *Dublin University Magazine, 1833–1877*, 90(535), 54–60.
- Börner, K., Maru, J. T., & Goldstone, R. L. (2004). The simultaneous evolution of author and paper networks. *Proceedings of the National Academy of Sciences*, 101(suppl 1), 5266–5273.
- Brookes, B. C. (1972). The aging of scientific literature. In A. I. Chernyi (Ed.), *Problems of information science: Collection of papers* (Vol. FID-478, pp. 66–90). The Hague: International Federation for Documentation.
- Burrell, Q. L. (2001). Stochastic modelling of the first-citation distribution. *Scientometrics*, 52(1), 3–12.
- Burrell, Q. (2002a). Modelling citation age data: Simple graphical methods from reliability theory. *Scientometrics*, 55(2), 273–285.

- Burrell, Q. (2002b). The n 'th-citation distribution and obsolescence. *Scientometrics*, 53(3), 309–323.
- Burton, R. E., & Kebler, R. (1960). The “half-life” of some scientific and technical literatures. *American Documentation*, 11(1), 18–22.
- Cairns, R. B. (1979). *Social development: The origins and plasticity of interchanges*. San Francisco: WH Freeman.
- Chatfield, C. (1995). *Problem solving: A statistician's guide* (2nd ed.). Boca Raton, Florida: CRC Press.
- Chen, C. (2003). On the shoulders of giants. *Mapping scientific frontiers: The quest for knowledge visualization* (pp. 135–166). London: Springer.
- Clauset, A., Shalizi, C. R., & Newman, M. E. (2009). Power-law distributions in empirical data. *SIAM Review*, 51(4), 661–703.
- Cole, S. (1983). The hierarchy of the sciences? *American Journal of Sociology*, 89(1), 111–139.
- de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149, 510–515.
- de Solla Price, D. J. (1970). Citation measures of hard science, soft science, technology, and nonscience. In C. E. Nelson & D. K. Pollock (Eds.), *Communication among scientists and engineers* (pp. 3–22). Lexington, MA: D. C. Heath and Company.
- Egghe, L. (1984). *Stopping time techniques for analysts and probabilists* (Vol. 100). Cambridge: Cambridge University Press.
- Egghe, L. (1997). Price index and its relation to the mean and median reference age. *Journal of the American Society for Information Science*, 48(6), 564–573.
- Egghe, L., & Ravichandra Rao, I. K. (1992). Citation age data and the obsolescence function: Fits and explanations. *Information Processing and Management*, 28(2), 201–217.
- Egghe, L., & Rousseau, R. (1990). *Introduction to informetrics: Quantitative methods in library, documentation and information science*. Amsterdam: Elsevier Science Publishers.
- Förster, M., Stelzer, B., & Schiebel, E. (2018). Stochastic analysis of citation time series of emergent research topics. In *Paper presented at the 23rd international conference on science and technology indicators*, Leiden University, The Netherlands.
- Frontline Systems Inc. (1990–2009). *Solver add-in*. Incline Village, Nevada. Retrieved from <http://www.solver.com>.
- Glänzel, W. (2004). Towards a model for diachronous and synchronous citation analyses. *Scientometrics*, 60(3), 511–522.
- Gross, P. L., & Gross, E. M. (1927). College libraries and chemical education. *Science*, 66(1713), 385–389.
- Gupta, B. (1997). Analysis of distribution of the age of citations in theoretical population genetics. *Scientometrics*, 40(1), 139–162.
- Hamermesh, D. S. (2018). Citations in economics: Measurement, uses, and impacts. *Journal of Economic Literature*, 56(1), 115–156.
- Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, 42(5), 443.
- Hu, X., Li, X., & Rousseau, R. (2020). Describing citations as a function of time. *Journal of Data and Information Science*, 5(2), 1–12.
- Jarić, I., Knežević-Jarić, J., & Lenhardt, M. (2014). Relative age of references as a tool to identify emerging research fields with an application to the field of ecology and environmental sciences. *Scientometrics*, 100(2), 519–529.
- Kaplan, M. A. (1961). Problems of theory building and theory confirmation in international politics. *World Politics*, 14(1), 6–24.
- Krauze, T. K., & Hillinger, C. (1971). Citations, references and the growth of scientific literature: A model of dynamic interaction. *Journal of the American Society for Information Science*, 22(5), 333–336.
- Kuhn, T. S. (1962). *The structure of scientific revolutions*. Chicago: University of Chicago Press.
- Lin, M., Lucas, H. C., Jr., & Shmueli, G. (2013). Research commentary—too big to fail: Large samples and the p-value problem. *Information Systems Research*, 24(4), 906–917.
- Lofland, J. (1967). Notes on naturalism in sociology. *Kansas Journal of Sociology*, 3(2), 45–61.
- MacRae, D., Jr. (1969). Growth and decay curves in scientific citations. *American Sociological Review*, 34(5), 631–635.
- Matricciani, E. (1991). The probability distribution of the age of references in engineering papers. *IEEE Transactions on Professional Communication*, 34(1), 7–12.
- McMaster University. (2020). *Research guides*. Retrieved from <https://libguides.mcmaster.ca/ft-top50>.
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806–834.
- Mees, C. K. (1917). The production of scientific knowledge. *Journal of Industrial and Engineering Chemistry*, 9(12), 1137–1141. <https://doi.org/10.1021/ie50096a027>.

- Nakamoto, H. (1988). Synchronous and diachronous citation distributions. *Informetrics*, 87(88), 157–163.
- Nowak, S. (1977). The formulation of the research problem and the choice of the right methods. *Methodology of sociological research* (pp. 1–42). Dordrecht: Springer.
- Pan, R. K., Petersen, A. M., Pammolli, F., & Fortunato, S. (2018). The memory of science: Inflation, myopia, and the knowledge network. *Journal of Informetrics*, 12(3), 656–678.
- Pollman, T. (2000). Forgetting and the ageing of scientific publications. *Scientometrics*, 47(1), 43–54.
- Popper, K. (2014). *Conjectures and refutations: The growth of scientific knowledge*. London: Routledge.
- Ransom, J. C. (1939). The arts and the philosophers. *The Kenyon Review*, 1(2), 194–199.
- Redner, S. (2005). *Citation statistics from 110 years of physical review*. (pp. 1–7) arXiv preprint: physics/0506056.
- Small, H. (1999). ASIS award of merit: On the shoulders of giants. *Bulletin of the American Society for Information Science*, 25, 23–25.
- Stacey, A. G. (2020). Robust parameterisation of ages of references in published research. *Journal of Informetrics*, 14(3), 1–19. <https://doi.org/10.1016/j.joi.2020.101048>.
- Storer, N. W. (1967). The hard sciences and the soft: Some sociological observations. *Bulletin of the Medical Library Association*, 55(1), 75–84.
- Yin, Y., & Wang, D. (2017). The time dimension of science: Connecting the past to the future. *Journal of Informetrics*, 11(2), 608–621.