



## Modeling a century of citation distributions

Matthew L. Wallace\*, Vincent Larivière, Yves Gingras

Observatoire des sciences et des technologies (OST), Centre interuniversitaire de recherche sur la science et la technologie (CIRST),  
Université du Québec à Montréal, Case Postale 8888, succ. Centre-Ville, Montréal (Québec) H3C 3P8, Canada

### ARTICLE INFO

#### Article history:

Received 12 January 2009

Received in revised form 27 March 2009

Accepted 31 March 2009

#### Keywords:

Citations

Bibliometrics

Citation distributions

Uncitedness

History of science

### ABSTRACT

The prevalence of uncited papers or of highly cited papers, with respect to the bulk of publications, provides important clues as to the dynamics of scientific research. Using 25 million papers and 600 million references from the Web of Science over the 1900–2006 period, this paper proposes a simple model based on a random selection process to explain the “uncitedness” phenomenon and its decline over the years. We show that the proportion of cited papers is a function of (1) the number of articles available (the competing papers), (2) the number of citing papers and (3) the number of references they contain. Using uncitedness as a departure point, we demonstrate the utility of the stretched-exponential function and a form of the Tsallis  $q$ -exponential function to fit *complete* citation distributions over the 20th century. As opposed to simple power-law fits, for instance, both these approaches are shown to be empirically well-grounded and robust enough to better understand citation dynamics at the aggregate level. On the basis of these models, we provide quantitative evidence and provisional explanations for an important shift in citation practices around 1960. We also propose a revision of the “citation classic” category as a set of articles which is clearly distinguishable from the rest of the field.

© 2009 Elsevier Ltd. All rights reserved.

### 1. Introduction

Since Price's pioneering work on networks of citations (1965) and his subsequent development of the cumulative advantage model to explain observed power-law behavior (1976), much work has been devoted to understanding the formal characteristics of the citation distribution (Redner, 1998, 2004) and to define its underlying mechanisms (Burrell, 2001, 2002; Nadarajah & Kotz, 2007; Simkin & Vwani, 2007). However, these distributions have been studied either from a mainly theoretical perspective (Burrell, 2001, 2002; Glänzel, 1992; Nadarajah & Kotz, 2007; Rousseau, 1994; Simkin & Vwani, 2007) or empirically, using a set of references constructed from a few years of data or based on a single discipline or specialty over a longer period of time (Lehmann, Lautrup, & Jackson, 2003; Redner, 1998, 2004; van Raan, 2006). Redner's examination of complete citation distributions of a century of articles published in *Physical Review* and of articles published in 1981 in journals covered by the Web of Science (WoS) has notably shown that, for a small number of citations, a stretched-exponential provides a better fit, while power-law behavior dominates at a high number of citations (Redner, 1998, 2004). Mathematical models have also succeeded in gaining some insight, *ab initio*, into how these distributions arise, but for the most part, these approaches come at the cost of high levels of parameterization and cannot explain the whole range of the citation distribution, especially the stretched-exponential regime (Nadarajah & Kotz, 2007; Simkin & Vwani, 2007).

\* Corresponding author.

E-mail addresses: [mattyliaam@gmail.com](mailto:mattyliaam@gmail.com) (M.L. Wallace), [lariviere.vincent@uqam.ca](mailto:lariviere.vincent@uqam.ca) (V. Larivière), [gingras.yves@uqam.ca](mailto:gingras.yves@uqam.ca) (Y. Gingras).

Given that most of these studies used age distribution of *cited* material for a given year, by definition, *uncited* articles were excluded from the analysis. Many models based on citation network growth also fail to include the zero-citation case (de Solla Price, 1976; Lehmann et al., 2003). But if one wants to study complete citation distributions, the natural point of departure would be uncited papers which, contrary to what is often believed (Hamilton, 1990, 1991; Macdonald & Kam, 2007; Meho, 2007), do not constitute the majority of the scientific literature. Although several empirical studies (Abt, 1991; Garfield, 1998; Pendlebury, 1991; Stern, 1990; Schwartz, 1997; van Dalen & Henkens, 2004) challenged this erroneous belief in high levels of uncitedness using data from a few fields and small periods of time, no study has yet measured the changes in scientific articles' citedness and uncitedness rates over a long period of time and across fields.

This paper analyzes the changes in articles' citation rates over a 107-year period (1900–2006) for all fields of the natural sciences and engineering (NSE), medicine (MED) and the social sciences (SS), excluding the humanities due to a lack of historical bibliometric data and its very particular citation dynamics compared with the other three disciplines (Larivière, Archambault, & Gingras, 2008). The following section briefly presents the methods; Section 3 presents empirical measures of articles' evolving rates of citedness and uncitedness. It also provides empirical measures of the mean and median citation rates of articles. Based on these data, we focus on the specific case of uncitedness and on *complete* models of citation distributions. Given that data for the full period is only available for NSE and MED, our examination of uncitedness will be centered on these two fields, and the bulk of our subsequent analysis will be limited to NSE data. We demonstrate the importance and utility of the stretched-exponential regime of the citation probability distribution function, and use it to collapse 100 years of citation data onto two master curves. We also demonstrate the robustness of a mathematical model derived from a generalization of Boltzmann–Gibbs statistics and propose an alternative definition of the “citation classic”. We achieve these goals using simple approximations and models with few adjustable parameters, allowing us to elucidate trends and characteristics that have dominated citation practices since 1900.

## 2. Methods

Data for this paper are drawn from Thomson Scientific's Web of Science, including the Century of Science database, along with the Science Citation Index Expanded (SCIE) and the Social Sciences Citation Index (SSCI). For NSE and MED, it covers the period 1900–2006, while the data for the social sciences begin in 1956. Only citations received by research articles, notes and review articles are included in the study; self-citations are also excluded. Citations made to 12 million (M) papers in NSE, 10 M in MED and 2 M in SS are retrieved in a pool of 600 M references recorded in the database using 2-, 5- and 10-year citation windows. The journal classification is based on that used by the National Science Foundation. The matching of citations to source articles was made using Thomson's reference identifier provided with the data, as well as additional matching using the author, publication year, volume number and page numbers.

## 3. Trends in citation distributions

Fig. 1 presents the variations, over the period studied, of the mean number of citations received by papers for 2- and 10-year citation windows. The median values (not shown), less affected by the highly cited papers, follow the same trends. One can readily notice that, since the end of the sixties, there has been a striking increase in the number of citations received by papers. The medical fields (MED), which essentially followed the citation trend of the NSE during the first half of the century, changed their dynamics in the 1960s and diverged from the NSE in terms of citation behavior. An interesting and apparent feature of MED and NSE curves is a decrease in the mean and median number of citations received around World Wars I and II, due to a rapid decline in the number of articles published (Larivière et al., 2008; van Raan, 2000). But perhaps

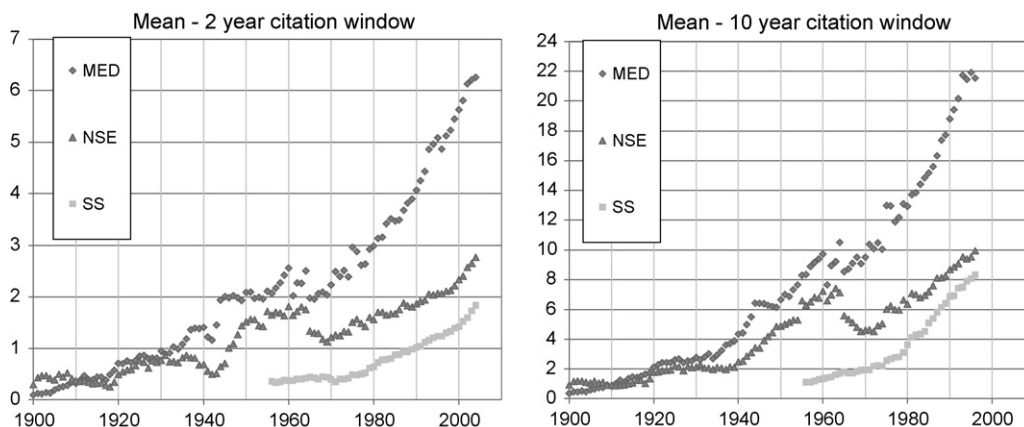


Fig. 1. Mean number of citations per paper, 2 years after publication (1900–2004) and 10 years after publication (1900–1996).

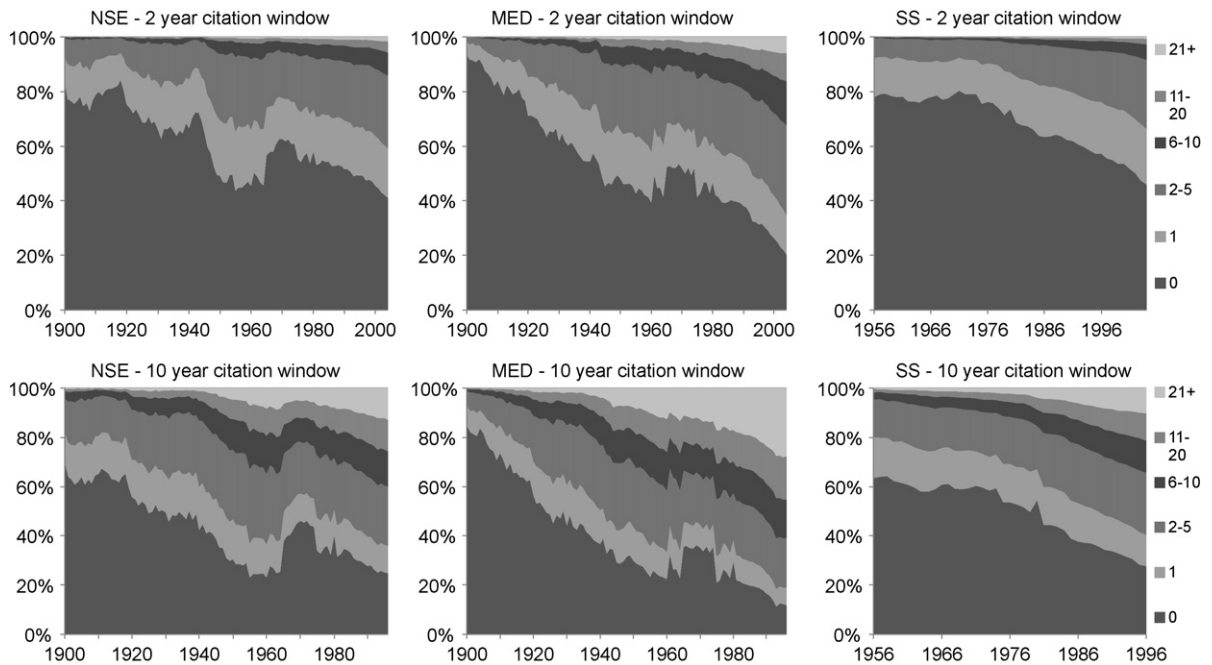


Fig. 2. Share of citations per paper 2 years after publication (1900–2004) and 10 years after publication (1900–1996).

more interesting is that, in these two fields, the post-WWII increase in articles' citation rates was followed by a 5–10-year decrease around 1960. This surprising phenomenon will be discussed in Section 7 of this paper.

Fig. 2 presents the data for 2- and 10-year citation windows, divided into 6 classes that take into account the skewness of citation distributions: 0 citations, 1 citation, between 2 and 5 citations, between 6 and 10 citations, between 11 and 20 citations, and 21 citations or more. These data clearly show that, contrary to a widespread belief, uncitedness has generally *declined* for all disciplines; science is increasingly drawing on the stock of published papers. Though not shown here, data for a 5-year window show the same trends. It is readily apparent that in NSE, for instance, the level of uncitedness is as low today as it was between 1950 and 1965, albeit for entirely different reasons (see Section 7).

#### 4. The decline of uncitedness

The data presented so far in the paper show that for all fields, the number of citations received by papers underwent an overall increase during the 20th century, tempered by a few local fluctuations between 1900 and 1970. More significantly, the proportion of published papers that remain uncited has decreased. In order to explain these changes, the present section provides a simple model that takes into account changes in the number of references per paper and in the growth of the world's scientific production (Larivière et al., 2008).

Previous studies (Barnett, Fink, & Debus, 1989; Burrell, 2001; Egghe, 2000; Rousseau, 1994) have shown that the first citation a paper receives likely occurs shortly after its publication, specifically according to a decay-type differential equation or a robust two or three-parameter function based on the articles' aging distribution. Although more accurate, but more complex formulations are certainly possible, we simply approximate (on the basis of Fig. 2) the first citations as either "immediate" (that is received in the first 2 years) or "latent" (arising at least 3 years after publication).

Our hypothesis is that the first citation depends simply on the number of articles published in a given year ( $N_A$ ) and the number of references ( $N_R$ ) available in the citing papers for the period covered (here a 2-year window). The chance of citing a given paper within the pool of  $N_A$  papers is thus a random selection process, expressed as a simple Poisson distribution (since  $N_R$  and  $N_A$  are large). But only a fraction of references can be considered as randomly "chosen" among recently published uncited papers, so the *effective* number of available references is  $\beta_I N_R$ , where  $\beta_I$  is found empirically, via a least-squares fit, to be around 0.016 in MED and 0.014 in NSE—very similar values to characterize the citation practices at an aggregate level in distinct scientific fields. As expected, the parameter for SS is much smaller (0.009), which means a lower probability of citing as yet uncited papers (given the same  $N_R$  and  $N_A$ ) than in NSE or MED.

The incidence of uncitedness, or the probability of getting zero immediate "successes", is thus expressed as

$$\Phi_I = e^{-\beta_I(N_R/N_A)} \quad (1)$$

In Fig. 3, we plot the empirical data and Eq. (1), which shows very good agreement over the entire period, capturing many of the local trends which appear in the data. The deviations are more important in the case of MED than NSE, but only before

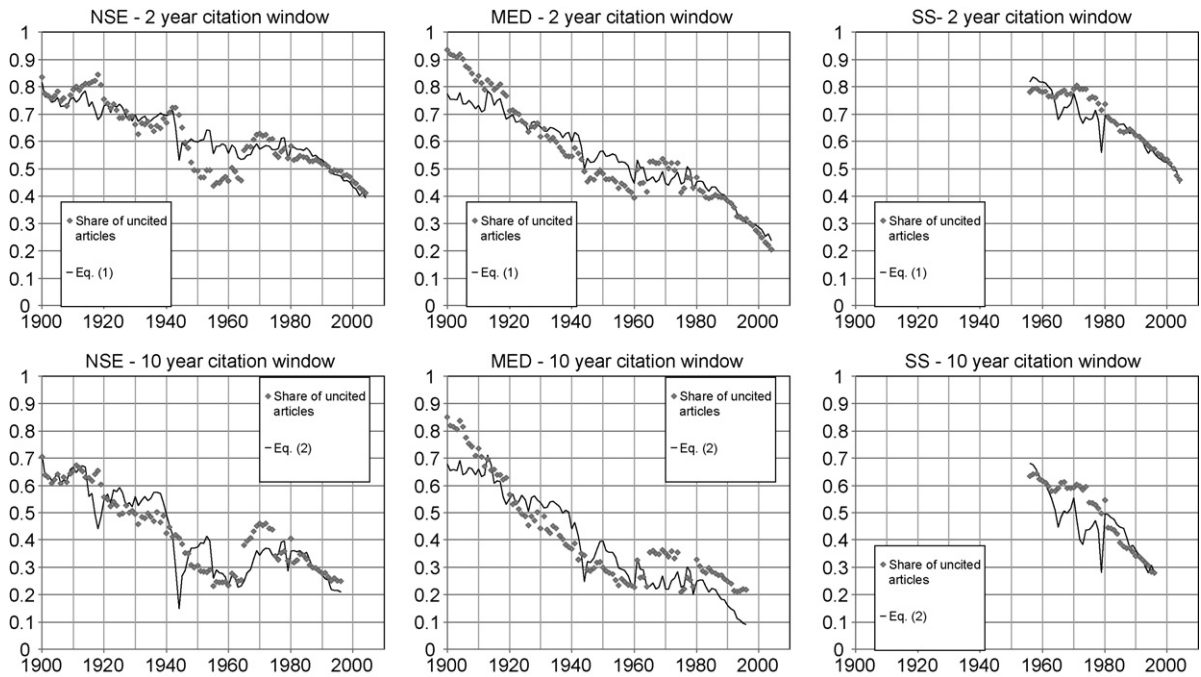


Fig. 3. Uncitedness data, compared with predictions, for 2- and 10-year citation window.

1915. In view of our approximations, this is probably due to the fact that citation practices seem to be less constant in time in MED than in NSE. In the latter, only the period between approximately 1960 and 1965 deviates systematically from the predictions of Eq. (1)—we will return to this particular case in Section 7. The case of SS is less satisfying but it could be related to the fact that primarily citing other articles (instead of books) is a relatively recent practice in many disciplines of the social sciences (Larivière, Archambault, Gingras, & Vignola-Gagné, 2006). We also have to take into account the fact that the WoS database is probably much less complete in these fields and captures less citations than in NSE and MED. The better fit after 1985 seems to be consistent with these hypotheses.

The generally very good agreement between data and Eq. (1) means that not only is uncitedness dominated by the ratio of references available to papers published (similar to the “Relative Publication Growth” developed in Vinkler, 2002), but also that the fraction of references used to determine first citations is approximately invariant over a period of 100 years. At the aggregate level, other mechanisms that could affect whether a given paper will get cited or not seem to have relatively little impact on the distribution.

The number of latent citations (those after the first 2 years) will inevitably depend once again on the ratio of references to remaining uncited articles, with even a smaller number of references from which to randomly choose old uncited articles. In a similar way, the chance of remaining uncited (after 10 years) can be expressed as the fraction of total (“latent” and “immediate”) uncited papers:

$$P(n = 0) = \Phi_T = \Phi_I e^{-\beta_L(N_R/N_A\Phi_I)} \quad (2)$$

Using the simple “latent” citation approximation, Eq. (2) satisfactorily describes first citation practices in MED and NSE (Fig. 3), except during a few years (e.g. the ends of the World Wars and, for MED, in most recent years), when first citations happened relatively quickly due to the fast-growing production. It should also be noted, particularly in the case of SS, that abrupt changes in the size of the database have a significant impact on the accuracy of our assumptions – that uncitedness depends largely on the number of articles per year – and of the application of Eq. (2) to long (e.g. 10 years) citation windows. A second parameter could be added in order to account for apparent changes in scientific production. Unsurprisingly,  $\beta_L$  is much smaller than  $\beta_I$  (by about one order of magnitude), since only a few references will have a chance of being chosen among the old, uncited papers. Furthermore, we have shown that capturing the main mechanisms responsible for uncitedness does not require excessive parameterization; a simple, time-independent approximation is robust enough to explain the bulk of over 100 years of uncitedness, implicitly averaged over many different areas of science.

At first glance, our results show that, in a given year, relatively high uncitedness is a consequence of relatively slow growth (i.e. not a sufficient amount of papers published in subsequent years) and a relatively stable number of references per paper. This is manifest in the case of the two World Wars where, as seen in Fig. 3 for instance, a lack of publication during the wars means higher uncitedness (during a 2-year citation window) beginning around the 2 years preceding the hostilities. A sharp increase in the number of publications immediately after the wars means that papers published up to 2 years before will have a higher chance of being cited. This is not, however, sufficient to explain why uncitedness has continued to decline for

the past 20 years or so, even though the period of rapid exponential growth in publishing has ended since about the mid-seventies. The answer is to be found in the fact that the average number of references per article has almost doubled between 1980 and 2004 in NSE and MED (Larivière et al., 2008). In terms of uncitedness, this phenomenon has thus counterbalanced the slowing of scientific growth in recent years.

### 5. The low and intermediate citation regime modeled with a stretched-exponential function

Naturally, the probability of a paper having  $n > 0$  citations cannot be described by a simple Poisson distribution without incorporating a second probability density function to describe the variable rates of citation (Burrell, 2001, 2002; Glänzel, 1992; Nadarajah & Kotz, 2007), but we can use our results for uncitedness as a starting point to evaluate and elucidate some of the empirical and aggregate citation models. We have found that not only is uncitedness sufficiently well-approximated by simple time-independent distribution, but that the  $P(n=0)$  case also fits the  $P(n > 0)$  data perfectly well, i.e. they can be described by the same function(s) and are therefore derived from the same set of basic mechanisms.

In order to perform a complete analysis of the citation distribution, we have focused our efforts on NSE data from a 10-year citation window, initially divided into 5-year periods (large enough for excellent statistics, but small enough to detect variations over time) from 1900 to 1994. Indeed, we have found that a stretched-exponential function of the form,

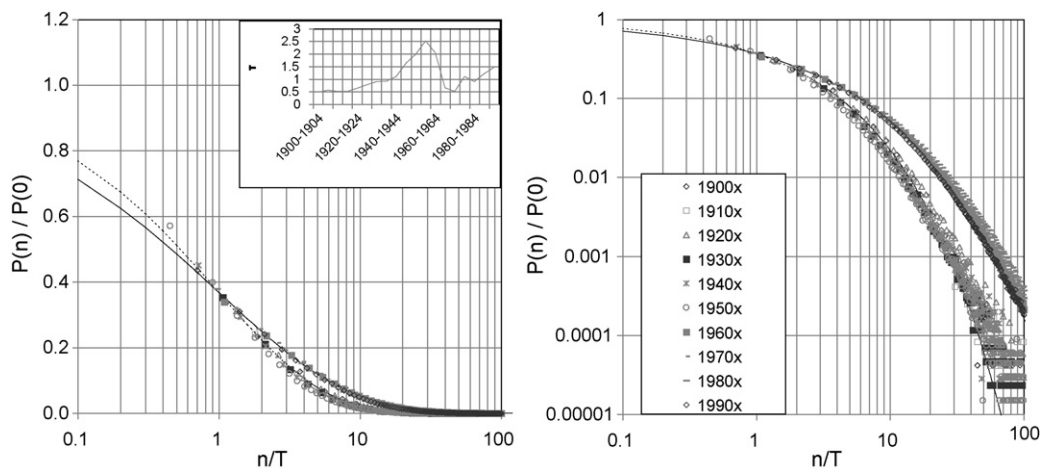
$$P(n) = P(0) \exp[-(n/\tau)^\alpha] \quad (3)$$

fits the data very well ( $R^2 > 0.98$ ) at worst for  $n < 40$  (earlier in the century) and at best for  $n < 200$  (during the 1990s). In the normalized distribution function (Fig. 4),  $P(0)$  can be taken directly from uncitedness data or can be estimated with satisfactory precision using Eqs. (1) and (2). The two parameters  $\tau$  and  $\alpha$  are found empirically by rearranging the equation in terms of logarithms such that,

$$\log \left[ \log \left( \frac{P(0)}{P(n)} \right) \right] = \alpha [\log(n) - \log(\tau)] \quad (4)$$

and performing a simple least-squares linear fit over the relevant regime (low and intermediate values of  $n$ ) to directly extract  $\alpha$  and  $\tau$ . We should emphasize that the purpose here is not to generate a probability distribution, such as the Weibull distribution, which is also based on a stretched-exponential function, but is slightly more restrictive and does not fit the data on uncitedness (the zero class) very well. The same could be said for many Lotkian power-law distributions, which must be modified to accommodate the existence of a zero class. Rather, we analyze the stretched exponential as a means of investigating the “citation chain” that includes uncitedness. This function seems intuitively well-suited to the citation process, by either considering the distribution as the result of a series of stochastic (citation) processes (Laherrère & Sornette, 1998) or, alternately, as an ensemble of papers gaining more citations at different exponential “rates” (analogous to a decay or relaxation time). We believe this stretched-exponential regime to be the most crucial part of the citation “chain”, since it includes virtually all papers and over 95% of citations, and thus captures the bulk of science. Perhaps too much attention has been paid to a small number of papers whose citations could be based on a power-law distribution arising, for instance, from a cumulative advantage process (de Solla Price, 1976) or from the accumulation of “noise” in a set of stochastic processes (Takayasu, Sato, & Takayasu, 1997).

Not only is the quality of the fit excellent during all periods, but there is also very little change in the values of  $\tau$  except, once again, around 1960. When we rescale the number of citations by  $\tau$  and normalize the distribution by  $P(0)$  over the entire



**Fig. 4.** Collapsed citation distribution over all decades, compared with two stretched-exponential fits using values of  $\alpha$ : 0.47 (solid line) and 0.57 (dashed line), shown in semi-log and log–log form (for clarity). *Inset* (left): Evolution of  $\tau$ , found via a linear least-squares fit, after re-arranging Eq. (3).

time period (this time by decade, for clarity), we note that there seem to be two regimes (see inset and master curves in Fig. 4), corresponding to different values of  $\alpha$  with the crossover year around 1960 (see discussion below). The parameter  $\tau$ , on the other hand, indicates the characteristic “citation-scale” of the process, or the number of citations associated with the initial “decay” to higher citedness. By rescaling, we isolate the *manner* in which the papers gain more citations (*how* these citation-scales are distributed). These are not rigorous interpretations of the data, but rather analogies drawn from the physical sciences (Williams & Watts, 1970; Phillips, 1996). The lognormal probability distribution (Radicchi, Fortunato, & Castellano, 2008) fits the overall citation distribution very well, although low numbers of citations are not as accurately predicted and uncitedness is not included at all. We believe that the stretched-exponential approach form is useful – particularly in light of these recent findings regarding universality in such probability distributions – for comparing citation data at the aggregate level, and to shed light on how citation practices operate and how they have evolved over 100 years.

## 6. A robust function to capture citation distribution tails and its application to “citation classics”

There has also been growing evidence suggesting that the large- $n$  tails cannot be fit to the same function that dominates intermediate numbers of citations (Gupta, Campanha, & Pesce, 2005; Lehmann et al., 2003). We would prefer, however, to have a single function to fit the data over the entire range of  $n$ . We find that the most robust fit (over all time periods) of the probability distribution function is found in a  $q$ -exponential function from the general, non-extensive statistical mechanics developed by Tsallis (Tsallis, 1988; Tsallis & de Albuquerque, 2000). As seen in Eq. (4) below, it is empirically equivalent, in our case, to a generalization of an asymptotic power-law form often referred to as the Zipf-Mandelbrot law (cf. Egghe, 2005). The advantage of adopting this particular formalism (the reader is directed to Tsallis & de Albuquerque (2000), and Abe (2002), and references therein for more details) is that it may be connected to the physical dynamics underlying citations (e.g. in terms of a network [cf. Thurner, Kyriakopoulos, & Tsallis, 2007] or other complex system). It is expressed as,

$$P(n) = \frac{P(0)}{[1 + (q-1)\lambda n]^{q/(q-1)}}. \quad (5)$$

In order to make sure we are getting a good fit at long time scales (Newman, 2005), we have used the cumulative probability density function  $P_C(n)$ , integrating Eq. (5) from 0 to a maximum of  $N$  citations, to generate a Zipf plot, and have accordingly transformed Tsallis' equation to,

$$P_C(n) = \int_0^N P(n)dn = \frac{P(0)}{\lambda} \left[ [1 + (q-1)\lambda n]^{-1/(q-1)} - [1 + (q-1)\lambda N]^{-1/(q-1)} \right], \quad (6)$$

where  $\lambda$  and  $q$  are determined empirically from a non-linear fit (using the package provided in Microcal Origin) and  $N$  denotes the maximum rank given to the citations or, equivalently, the maximum number of citations to one paper expected for a given data set. It need not be very precise and only becomes relevant for large  $n$ , since it brings a correction of at most  $O(N^{-2})$ . This approximate size limit on citations (equivalent to a cut-off of the power-law [Gupta et al., 2005; Gupta, Campanha & Schinaider, 2008; Takayasu et al., 1997]) essentially introduces a finite-size effect to the system, generating the tails at large  $n$  seen for many distributions in Fig. 5. The exponent  $q$  varies between around 1.2 and 1.5, depending on the period studied, so there can be no strictly universal power-law citation distribution; this seems more consistent with the known facts of varying citation practices over time and between disciplines.

We have thus shown that the Tsallis distribution is robust enough to fit the intermediate and high range of citation distributions over all time periods (without an excessive use of parameters or “extraneous” functions), and suggest further research into the details of specific citation mechanisms which can give rise to it. Contrary to Tsallis' claim, however, the function performs poorly at very small  $n$  (see also Fig. 1 of Tsallis & de Albuquerque, 2000), where the bulk of citation occurs, but it is difficult to see on a log-log-plot (Fig. 5). When analytically calculating the average number of citations from Eq. (5), there appears to be a systematic overshoot, on the order of 5–20%, compared with the empirical value, which is nonetheless sufficiently accurate to observe all the main trends over 100 years (Fig. 1).

An alternative method was recently proposed (Naumis & Cocho, 2007), consisting in explicitly ranking papers in terms of their citations and applying a beta-like function, based on the “hierarchy” of a stochastic multiplicative process, over the entire range of ranks. We have found that this also fits citation data for most years extremely well, but a detailed discussion of the advantages of a rank-based approach (as opposed to standard probability distribution functions) remains beyond the scope of the present paper.

We propose to interpret the very large  $n$ -tails of the distribution as a useful and systematic way of determining what constitutes a citation “classic” (Garfield, 1984, 1989). From Eq. (6), this is simply dependent on the distance from the maximum rank  $N$  and, once again, is equivalent to assigning some importance to the cutoffs of a power-law (Gupta et al., 2005; Takayasu et al., 1997). Clearly, the number of citations which can intuitively (and mathematically) define a “classic” is not the same in 1920 as in 1980 (in MED, for instance, 20 citations has now become “average”!). The finite size of the pool of references means that this category of articles must drop off quickly (hence the tails). Our approach would thus empirically distinguish the “cream” from the rest of the crop, without requiring that the same number, or fraction, of articles every year achieve this status. As is shown in Fig. 5, certain periods, namely 1905–1909, display little evidence of these “classics” as distinct from

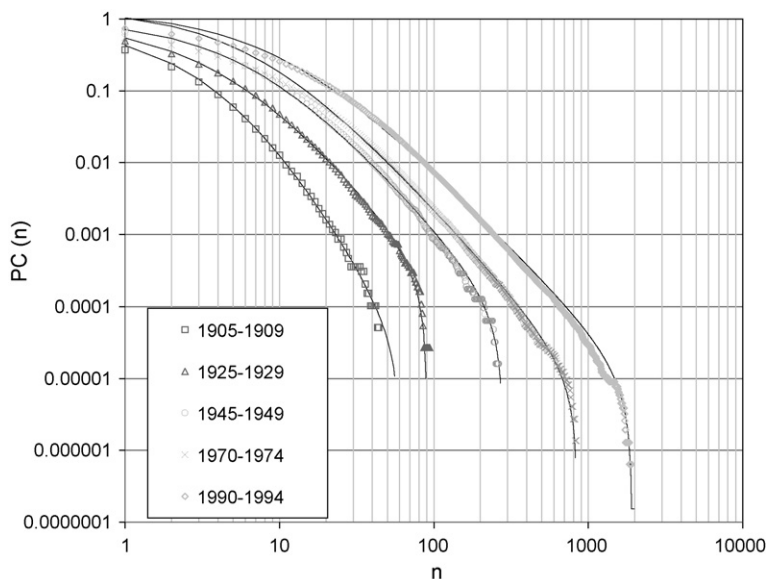


Fig. 5. Cumulative citation distributions and Tsallis fits from Eq. (5) at various times.

the rest of the citation distribution. Overall, if one wishes to unambiguously define *any* given class of citations, applicable to different time periods or scientific fields, it is imperative to have a precise idea of the form of the *entire* citation distribution.

## 7. Bibliometric data and the peak of postwar science

The particularities in the citation data between around 1958 and 1965 in NSE (and, to a lesser extent, MED, although we have not examined this case in the same detail) are worthy of some special attention. First, certain characteristics, unmatched in the 20th century, can be ascribed to the evolution of science (but, seemingly, not of the social sciences—see Fig. 2) during this time. Furthermore, this exceptional case study serves to elucidate important elements of the trends we have observed over 100 years and to understand the significance (and limitations) of the models that can describe them.

Our model for the proportion of uncited papers – which performs surprisingly well during most of the 20th century – predicts that, 20 years into the postwar boom of science, there would be an increase in uncitedness due to a *relative* and inevitable slowing-down of production. However, the model fails because the scientific community has proven itself to be particularly “efficient” during this period, meaning that even “marginal” authors were producing “citable” papers. Another way of looking at this is to consider certain periods of scientific activity to be more “compact” or cohesive than others, perhaps indicating that very few actors are working in the periphery of science. Take, for instance, the dominance of a few main scientific problems or extraordinarily fruitful avenues of research during this time: in physics, the development of the standard model and semiconductors; in biochemistry and biology, research on enzymes and proteins (thanks in part to rapid developments in organic chemistry methods), the discovery of the structure of DNA, and important progress in molecular biology (including molecular evolution and immunology).

In addition, Fig. 4 suggests that, around 1960, a sharp change in the distribution of citations occurred, separating two “steady-states” of citation practices. In other words, the impetus caused by the productivity boom forced a jump in the “citation-system”, which then reconfigured to a state with a slightly lower stretching exponent  $\alpha$ . This generally indicates that the distribution of “characteristic citation-scales” (analogous to relaxation or decay times) changes to include larger citation-scales. Evidence for some type of transition appears in all our citation data, including simple indicators, such as the average number of citations per paper during 2- and 10-year citation windows (Fig. 1).

## 8. Conclusion

We have constructed a complete dataset of all citations to papers published between 1900 and 2006, highlighting the similarities and differences in citation practices in MED, NSE and SS. An overall decrease in uncitedness (for 2, 5 and 10 years after publication) since 1900, as well as several local trends observed over 100 years can be satisfactorily explained by assuming that, each year, *the same fraction of references* contributes to randomly giving papers their first citation. Specifically, the trends observed during the wars, for instance, were due to changes in the number of *papers* published, while the recent decrease in uncitedness (contrary to what is often believed) is due to a higher number of *references per paper*. The relatively low level of uncitedness around the 1960s (compared with the mathematical predictions) might be explained by the efficiency or cohesiveness of the scientific community around this time.

Taking a global and macroscopic perspective on citation distributions (including uncitedness), we have insisted on the importance of the stretched-exponential function which dominates at small and intermediate  $n$  (the bulk of citations). We have used simple fits to extract data from this function and rescale the 100 years of citation data onto two master curves, finding a crossover point and exceptionally large values of  $\tau$  once again around 1960. We have shown that, by using the cumulative distribution function, applying Tsallis' non-extensive statistical mechanics and considering a finite size limit for the number of citations, we can accurately model most of the citation distribution function (including the tails), at the expense of the very small  $n$  regime, and that such a fit is robust enough to be applied to citation data from all periods of the 20th century. The tails of the Tsallis function, as deviations from a power-law, are furthermore proposed as a way to systematically and objectively identify highly cited, "exceptional" papers.

## References

- Abe, S. (2002). Stability of Tsallis entropy and instabilities of Rényi and normalized Tsallis entropies: A basis for  $q$ -exponential distributions. *Physical Review E*, 66, 046134.
- Abt, H. A. (1991). Science, citation, and funding. *Science*, 251, 1408–1409.
- Barnett, G. A., Fink, E. L., & Debus, M. B. (1989). A mathematical model of academic citation age. *Communication Research*, 16, 510–531.
- Burrell, Q. L. (2001). Stochastic modeling of the first-citation distribution. *Scientometrics*, 52, 3–12.
- Burrell, Q. L. (2002). The  $n$ th-citation distribution and obsolescence. *Scientometrics*, 53, 309–323.
- de Solla Price, D. J. (1965). Networks of scientific papers. *Science*, 149, 510–515.
- de Solla Price, D. J. (1976). A general theory of bibliometric and other cumulative advantage processes. *Journal of the American Society for Information Science*, 27, 292–306.
- Egghe, L. (2000). A heuristic study of the first-citation distribution. *Scientometrics*, 48, 345–359.
- Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. New York: Elsevier Academic Press.
- Garfield, E. (1984). The 100 most-cited papers ever and how we select citation classics. *Current Contents*, 23, 3–9.
- Garfield, E. (1989). Citation classics and citation behavior revisited. *Current Contents*, 12, 3–8.
- Garfield, E. (1998). I had a dream. About uncitedness. *The Scientist*, 12, 10.
- Glänzel, W. (1992). On some stopping times of citation processes. *Information Processing and Management*, 28, 53–60.
- Gupta, H. M., Campanha, J. R., & Pesce, R. A. G. (2005). Power-law distributions for the citation index of scientific publications and scientists. *Brazilian Journal of Physics*, 35, 981–986.
- Gupta, H. M., Campanha, J. R., & Schinaider, S. J. (2008). Size limiting in Tsallis statistics. *Physica A*, 387, 6745–6751.
- Hamilton, D. P. (1990). Publishing by – and for? – the numbers. *Science*, 250, 1331–1332.
- Hamilton, D. P. (1991). Research papers, who's uncited now? *Science*, 251, 25.
- Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy "fat tails" with characteristic scales. *European Physical Journal B*, 2, 525–539.
- Larivière, V., Archambault, É., & Gingras, Y. (2008). Long-term variations in the aging of scientific literature, from exponential growth to steady-state science (1900–2004). *Journal of the American Society for Information Science and Technology*, 59, 288–296.
- Larivière, V., Archambault, É., Gingras, Y., & Vignola-Gagné, É. (2006). The place of serials in referencing practices, comparing natural sciences and engineering with social sciences and humanities. *Journal of the American Society for Information Science and Technology*, 57, 997–1004.
- Lehmann, S., Lautrup, B., & Jackson, A. D. (2003). Citation networks in high energy physics. *Physical Review E*, 68, 026113.
- Macdonald, S., & Kam, J. (2007). Aardvark et al., quality journals and gamesmanship in management studies. *Journal of Information Science*, 33, 702–717.
- Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World*, 20, 32–36.
- Nadarajah, S., & Kotz, S. (2007). Models for citation behavior. *Scientometrics*, 72, 291–305.
- Naumis, G. G., & Cocho, G. (2007). The tails of rank-size distributions due to multiplicative processes, from power laws to stretched exponentials and beta-like functions. *New Journal of Physics*, 9, 286–293.
- Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, 46, 323–351.
- Pendlebury, D. (1991). Science, citation, and funding. *Science*, 251, 1410–1411.
- Phillips, J. C. (1996). Stretched exponential relaxation in molecular and electronic glasses. *Reports on Progress in Physics*, 59, 1133–1207.
- Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions, toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of United States of America*, 105, 17268–17272.
- Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, 2, 131–134.
- Redner, S. (2004). Citation statistics from more than a century of Physical Review. arXiv:physics/0407137.
- Rousseau, R. (1994). Double exponential models for first-citation processes. *Scientometrics*, 30, 213–227.
- Schwartz, C. A. (1997). The rise and fall of uncitedness. *College & Research Libraries*, 58, 19–29.
- Simkin, M. V., & Vwani, P. R. (2007). A mathematical theory of citing. *Journal of the American Society for Information Science and Technology*, 58, 1661–1673.
- Stern, R. E. (1990). Uncitedness in the biomedical literature. *Journal of the American Society for Information Science and Technology*, 4, 193–196.
- Takayasu, H., Sato, A.-H., & Takayasu, M. (1997). Stable infinite variance fluctuations in randomly amplified Langevin systems. *Physical Review Letters*, 79, 966–969.
- Thurner, S., Kyriakopoulos, F., & Tsallis, C. (2007). Unified model for network dynamics exhibiting nonextensive statistics. *Physical Review E*, 76, 036111.
- Tsallis, C. (1988). Possible generalization of Boltzmann–Gibbs statistics. *Journal of Statistical Physics*, 52, 479.
- Tsallis, C., & de Albuquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? *European Journal of Physics B*, 13, 777–780.
- van Dalen, H. P., & Henkens, K. E. (2004). Demographers and their journals, who remains uncited after ten years? *Population and Development Review*, 30, 489–506.
- van Raan, A. F. J. (2000). On growth, ageing, and fractal differentiation of science. *Scientometrics*, 47, 347–362.
- van Raan, A. F. J. (2006). Statistical properties of bibliometric indicators, research group indicator distributions and correlations. *Journal of the American Society for Information Science and Technology*, 57, 408–430.
- Vinkler, P. (2002). Dynamic changes in the chance for citedness. *Scientometrics*, 54, 421–434.
- Williams, G., & Watts, D. C. (1970). Non-symmetrical dielectric relaxation behavior arising from a simple empirical decay function. *Transactions of the Faraday Society*, 66, 80–85.