Scientometrics (2011) 88:385–397
DOI 10.1007/s11192-011-0407-9

The skewness of science in 219 sub-fields and a number of aggregates

Pedro Albarrán · Juan A. Crespo · Ignacio Ortuño · Javier Ruiz-Castillo

Received: 10 December 2010 / Published online: 18 May 2011

Abstract This paper studies evidence from Thomson Scientific (TS) about the citation process of 3.7 million articles published in the period 1998–2002 in 219 Web of Science (WoS) categories, or sub-fields. Reference and citation distributions have very different characteristics across sub-fields. However, when analyzed with the Characteristic Scores and Scales (CSS) technique, which is replication and scale invariant, the shape of these distributions over three broad categories of articles appears strikingly similar. Reference distributions are mildly skewed, but citation distributions with a 5-year citation window are highly skewed: the mean is 20 points above the median, while 9–10% of all articles in the upper tail account for about 44% of all citations. The aggregation of sub-fields into disciplines and fields according to several aggregation schemes preserve this feature of citation distributions. It should be noted that when we look into subsets of articles within the lower and upper tails of citation distributions the universality partially breaks down. On the other hand, for 140 of the 219 sub-fields the existence of a power law cannot be rejected. However, contrary to what is generally believed, at the sub-field level the scaling parameter is above 3.5 most of the time, and power laws are relatively small: on average, they represent 2% of all articles and account for 13.5% of all citations. The results of the aggregation into disciplines and fields reveal that power law algebra is a subtle phenomenon.

P. Albarrán
Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, Alicante, Spain

J. A. Crespo
Departamento de Economía Cuantitativa, Universidad Autónoma de Madrid, Madrid, Spain

I. Ortuño
Departamento de Economía, Universidad Carlos III, Madrid, Spain

J. Ruiz-Castillo (✉)
Departamento de Economía, Universidad Carlos III & Research Associate of the CEPR Project SCIFI-GLOW, Madrid, Spain
e-mail: jrc@eco.uc3m.es

Keywords Research performance · Citation analysis · Power laws · Characteristic Scores

Introduction

It is well known that, among other factors, differences in publication practices across research areas regarding the length of the average article in the periodical literature and the number of articles per person, are responsible for large differences in area sizes measured by the number of articles per area. It is also well known that, due to vastly different citation practices, reference distributions have very different mean rates and other characteristics. In turn, it is equally well known that citation distributions have very different characteristics across scientific fields.

This diversity seems to be compatible with the belief among Scientometrics' practitioners that citation distributions share some fundamental characteristics. As originally suggested in Price (1965) and afterwards analyzed in Seglen's (1992) seminal contribution, it is generally believed that citation distributions are highly skewed. Moreover, it is widely held that citation distributions can be represented by power laws (see Egghe 2005, for a treatise on the importance of power laws for information production processes of which citation distributions are only one type). More recently, in two important contributions Radicchi et al. (2008) and Glänzel (2010) suggest that since citation distributions only differ by a scale factor, after appropriate normalization we can speak of a (highly skewed) universal citation distribution. The problem is that the empirical evidence sustaining these beliefs is, although valuable, not conclusive. This paper contributes to setting the record straight at different aggregation levels for a large sample of 3.7 million scientific articles published in the period 1998–2002, acquired from Thomson Scientific (TS hereafter).

The shortcomings of the present situation are of two types. Firstly, claims about citation distributions' regularities appearing in the literature are based on an accumulation of case studies. Most of the evidence is not systematic. As far as the skewness of citation distributions is concerned, together with the illustrations in Seglen's paper from a random sample of articles drawn from the 1985–1989 Science Citation Index, and Magyar's (1973) data on the small sub-field of dye laser research, a first set of papers only includes the contributions of Irvine and Martin (1984) and Lehmann et al. (2003) on high energy physics, and Burke and Butler (1996) on the entire fields of the natural sciences and the social sciences and the humanities in Australian universities. On the other hand, beyond the graphical illustrations included in Narayan (1971) and Seglen (1992), the only directly estimated results that we have found in the fitting of power laws to citation distributions are for papers on all fields published in 1981 and listed in the Science Citation Index (Redner 1998; Clauset et al. 2009), papers published in *Physical Review* during long time periods (Redner 1998, 2005), 18,000 publications in Chemistry in the Netherlands during 1991–2000 with a 3-year citation window (Van Raan 2006), and papers in high energy physics (Lehmann et al. 2003, 2008). Laherrère and Sornette (1998) study the citation record of the most cited physicists, while Clauset et al. (2009) include the publication record of mathematicians.

Secondly, the available evidence does not confront what we call the aggregation issue. On one hand, the smaller the set of closely linked journals used to define a given research field, the greater the homogeneity of citation patterns among the articles included must be. This homogeneity guarantees that the relative merit of articles in a given field can be measured by their number of citations. Moreover, when questioned, most scientists would

answer that they belong to one, or at most a few, well-defined research areas. Consequently, one should always work at the lowest aggregation level that the data allows for. In this paper, research areas at that level are referred to as *sub-fields*. On the other hand, given the plethora of scientific sub-fields that easily reach between two and three hundred, for many practical problems the interest of investigating larger aggregates is undeniable. Above sub-fields, this paper distinguishes between an intermediate category—referred to as *disciplines*, such as Internal Medicine or Dentistry; Particle and Nuclear Physics or Physics of Solids; and Organic or Inorganic Chemistry—and traditional, broad fields of study such as Clinical Medicine, Physics and Chemistry, referred to simply as *fields*. In this context, the aggregation issue is whether the existence of common characteristics for all—or most—citation distributions, is a phenomenon dependent on the aggregate level of analysis. In other words, one should investigate whether these characteristics are only present at a high aggregation level and disappear at the sub-field level or, on the contrary, whether they are present at the lowest aggregation level and, in this case, whether they are preserved or not at higher aggregate levels.

It must be recognized that the task of deciding what a sub-field should be at the lowest level of aggregation, as well as the drawing of the lines that connect each sub-field to a single discipline and a single field, constitute formidable practical problems that must be solved prior to the study of citation distributions at different aggregation levels. In this scenario, the limitations of the only three groups of systematic studies available in the literature should be readily apparent. (i) Schubert et al. (1987) analyze the papers published in the period 1981–1985 in the journals covered by the Science Citation Index, and the citations received during this same period. These authors work at a low aggregation level, consisting of the 114 sub-fields distinguished at the time in the Journal Citation Reports. They study the shape of citation distributions by applying the Characteristic Scores and Scales (CSS hereafter) technique that permits the partition of any distribution of articles into a number of classes as a function of their members' citation characteristics. However, no statistical test of the presence of common characteristics at this level is provided, and no aggregation into scientific fields or disciplines is attempted. Glänzel (2007a) studies 450,000 citable papers published in 1980, cited in the 1980–2000 period, and classified into 60 disciplines and 12 major fields. However, this study only reports results for the application of CSS to 12 of the 60 disciplines. None of these papers directly estimate a power law. Instead, under the hypothesis that a citation distribution consisting of those articles receiving at least one citation follows a power law, Glänzel (2007a) obtains an equation relating the scaling parameter of this distribution and the parameters of the CSS technique. With direct estimates of the latter, the former are computed.[1] (ii) Radicchi et al. (2008) focus on the evaluation of citation performance of single publications in different research areas. They conclude that, "…in rescaling the distribution of citations for publications in a scientific discipline by their average number, a universal curve is found, independent of the specific discipline" (p. 17269). They indicate that the universal normalized citation distribution resembles a lognormal distribution, whose single parameter is estimated. However, the empirical evidence they provide only refers to 14 sub-fields or World of Science categories that, nevertheless, span broad areas of science. Based on an observation in an earlier paper by Schubert et al. (1989), Glänzel (2010) studies a normalization similar to the one suggested by Radicchi et al. (2008). The problem, again, is that evidence is only presented for the 12 disciplines analyzed in Glänzel (2007a). (iii)

[1] Under the same restrictive hypothesis, Schubert and Glänzel (2007) and Glänzel (2007b, 2008) deduce the scaling parameter from an equation relating the *h*-index and the parameters of the assumed power law.

Using the same dataset as this paper, Albarrán and Ruiz-Castillo (2011) study the shape of reference and citation distributions using the CSS technique, and estimate power laws using state-of-the-art, maximum likelihood techniques. However, the work is only conducted at a high aggregation level, namely, the 22 broad fields distinguished by TS.

This paper takes an important step towards settling these issues by investigating a large dataset organized in 219 Web of Science (WoS hereafter) categories or sub-fields at the lowest aggregation level. To deal with the problem of multiple assignments of articles to WoS categories, we adopt a multiplicative strategy in which items classified into several sub-fields are wholly counted in all of them. On the other hand, given the inexistence of a hierarchical Map of Science organizing sub-fields, disciplines, and fields in a way agreed upon by the international scientific community, we consider two alternative routes inspired in Tijssen and van Leeuwen (2003) and Glänzel and Schubert (2003) to ascend from the sub-field to the discipline and the field levels.

The paper investigates the following two questions. Firstly, whether there exists a typical shape of reference and citation distributions with the same stylized features for sub-fields, disciplines, and fields.[2] Using the CSS approach the answer is that, indeed, as long as we focus on the distribution of articles over three broad categories reference and citation distributions at all aggregation levels present the same shape. In particular, citation distributions are characterized by a highly skewed shape. However, as long as we focus on smaller segments inside the lower and upper tails of citation distributions higher measures of dispersion indicate a lack of universality across sub-fields. Secondly, we establish that in 140 out of 219 sub-fields the existence of a power law cannot be rejected. However, their main features are rather different from what is generally believed. We also study the intriguing question of whether power laws at the sub-field level are preserved or not at upper aggregation levels, and whether sub-fields that cannot be represented by a power law give rise to a discipline or a field that exhibits this interesting behavior.

The rest of the paper is organized into three sections. "Data, methods, and descriptive statistics" section presents the data, discusses the assignment of articles to sub-fields, disciplines, and fields using two convenient aggregation schemes, and summarizes some basic descriptive statistics that illustrate how different sub-field reference distributions and citation distributions at every aggregation level really are. "Empirical results" section contains the main results of the paper about (i) the striking similarities of the shapes of citation distributions at all aggregation levels, and (ii) the results of the estimation of power laws at all aggregation levels. "Conclusions and extensions" section offers some concluding comments, and discusses a number of possible extensions.

Data, methods, and descriptive statistics

The dataset

Since we wished to address a homogeneous population, in this paper only research articles, or simply articles, are studied. A relatively large sample was needed to ensure a minimum size for all areas of study at all aggregation levels. We choose the 3,767,378 articles published between 1998 and 2002, a sample that was shown in Albarrán and Ruiz-Castillo

[2] The vast majority of articles written in citation analysis deal exclusively with citations received. For an exception, apart from Price's (1965) seminal contribution and Albarrán and Ruiz-Castillo (2011), see Liang and Rousseau (2010) and the references quoted there.

(2011) to be representative of the 1998–2007 dataset. A fixed, common 5-year citation window is chosen for all articles.

The classification of articles into sub-fields

Table 1 in the Working Paper version of this paper (see Albarrán et al. 2011) informs in detail about the multi-WoS category structure of the 22 fields distinguished by TS. The main fact is that only about 58% of all articles are assigned to a single WoS category. Therefore, we must confront the question of how to classify articles with multi-WoS categories into sub-fields. A crucial requirement is that all articles within a sub-field should count the same. Otherwise, if an article assigned to several WoS categories were fractionally assigned to them, then its place in the various citation distributions would be dramatically affected. In particular, fractionally assigned articles would have a much smaller chance of occupying the upper tail of citation distributions than articles assigned to a single WoS category. Therefore, we opt for a multiplicative strategy, where each article is classified into as many sub-fields as WoS categories in the original dataset. In this way, the space of articles is expanded as much as necessary beyond the initial size. As a matter of fact, the total number of articles in what we call the *extended count* for the 219 TS sub-fields is 5,509,510, or 57% larger than the original dataset. This artificially large number is not that worrisome in the sense that, since the multiplicative strategy does not create any interdependencies among the sub-fields involved, it is still possible to separately investigate every sub-field in isolation, independently of what takes place in any other sub-field.

The classification of articles into disciplines and fields

Assume for a moment that we are given a reasonable classification of sub-fields into disciplines and fields, that is, assume that we have a Map of Science to work with. The next question is how to classify articles into disciplines and fields. Consider first the assignment of articles to disciplines. Articles originally assigned to a single sub-field are directly assigned to the discipline indicated by the Map of Science. For articles assigned to multiple sub-fields we adopt again a multiplicative strategy. For example, consider the case of an article assigned to two subfields. If both belong to the same discipline, then the assignment of the article to a discipline poses no problem. Otherwise, that is, if the two sub-fields belong to two different disciplines, then the article is wholly assigned to both of them. In the case of other multiple sub-field assignments, we would proceed likewise. In this way, the space of articles is again expanded as much as necessary beyond the initial size. However, because whenever two or more sub-fields belong to the same discipline no multiplication of the article is necessary, the total number of articles in the disciplines case will be closer to the initial one than in the sub-fields case. A similar process for the assignment of articles to fields should lead to a still lower number of expanded articles.

The question that remains to be answered is how to construct a Map of Science—a task that is known to have no easy answer (see Albarrán et al. 2011 for a brief discussion of the literature on this issue). As indicated in the "Introduction" section, in this paper we use two alternatives: a scheme inspired in Tijssen and van Leeuwen (2003)—TvL hereafter—and a rather different one due to Glänzel and Schubert (2003)—GS hereafter—that exclusively refers to the natural sciences. In the first case, we end up with 38 disciplines and 12 fields, while in the second case there are 61 disciplines and 12 fields (see the details in the Appendix and Table 2 in Albarrán et al. 2011). In our version of the TvL scheme, disciplines and fields lead to extended counts about 37 and 24% larger than the original

dataset. The extended counts in the GS scheme for disciplines and fields are about 44 and 20% larger than the original number of articles in the natural sciences.

Descriptive statistics

Before we proceed to search for similarities across reference and citation distributions, it is important to document the differences they present at each aggregation level. It should be noted that our dataset does not indicate how references made by articles published in year t actually become citations received by other articles in years t, $t + 1$, up to $t + 4$ during a 5 year citation window. Our information is about the references made by articles published in each of the years from 1998 to 2002, and the citations they receive afterwards during the period 1998–2002, 1999–2003, up to 2002–2006, respectively. Nevertheless, we believe that the study of both types of distributions is worthwhile. The information can be summarized in the following three points (see Tables A and B in the Appendix in Albarrán et al. 2011, about individual characteristics, and Tables 3, 4, and 5 in that same paper about average values).

1. Publication practices are very different indeed. In some research areas authors publishing one article per year would be among the most productive, while in other instances authors—either alone or as members of a research team—are expected to publish several papers per year. On the other hand, since the different aggregation categories are not designed at all to equalize the number of articles published in a given period of time, distribution sizes are expected to differ within all aggregation levels. In particular, at the sub-field level the mean is equal to 26,984 articles and the standard deviation is 29,669, while the range of variation goes from a minimum of 423 articles (Biology, Miscellaneous), or 893 (Ethnic Studies), and seven sub-fields with fewer than 3,000 articles, to a maximum of 213,448 articles in Biochemistry and Molecular Biology and seven sub-fields with more than 100,000 articles.
2. Due to vastly different citation practices, reference distributions are very different across sub-fields. On average, the mean reference rate is equal to 26.3 and the standard deviation is 8. In turn, the ratio of references made over citations received is equal to 6.1 with a standard deviation of 4.1.[3] This should be an important factor explaining the dramatic changes experienced by the percentage of uncited articles and the mean citation rate when we turn from reference to citation distributions: the first variable increases (up to 24.7%) and the second decreases (down to 5.7) by a factor of five.
3. The main point that should be emphasized is the high values of absolute and relative dispersion measures associated with characteristics of citation distributions at every aggregation level. This clearly indicates that *within* all aggregation levels these distributions are very different indeed.

Empirical results

The skewness and universality of science

In the previous section it was observed that, as a consequence of vastly different publication and citation practices, reference and citation distributions at different aggregation

[3] Recall that references are made to many different items: articles in TS-indexed journals, as well as articles in conference volumes, books, and other documents, none of them covered by TS. Moreover, some references are to articles published in TS journals before 1998 and, hence, outside our dataset.

levels are very different in two crucial dimensions: distribution size, and mean citation rate. However, as soon as replication and scale invariant measurement instruments are used and, consequently, as soon as we focus on the shape of reference and citation distributions, we discover that reference and citation distributions are strikingly similar.

Denote by s_1 the mean citation rate; by s_2 the mean citation rate of articles with more than s_1 citations, and by s_3 the mean citation rate of articles with more than s_2 citations. In the CSS approach, five categories are distinguished: articles without citations; *poorly cited* articles if their citations are below s_1; *fairly cited* if they are between s_1 and s_2; *remarkably cited* if they are between s_2 and s_3, and *outstandingly cited* if they are above s_3. The average and the standard deviation of the percentage represented by articles in the two lowest and the two highest categories, as well as the percentage of references or citations accounted for by the key categories are reported in Table 1.

At the sub-field level (Panel A in Table 1), the conclusions are the following:

- Reference distributions are moderately skewed: on average, the mean is only 7.5% percentage points above the median, while articles with a remarkable or outstanding number of references that in a uniform distribution would constitute 25% of the total, actually represent 16%; these articles account for 35% of all references.
- As expected, citation distributions are highly skewed: approximately 69% of all articles receive citations below the mean and on average account for 21% of all citations, while articles with a remarkable or outstanding number of citations represent about 9 or 10% of the total, and account for approximately 44% of all citations.
- Small standard deviations for the partition of articles into three broad categories in Table 1 indicate that most differences across sub-fields dramatically diminish.

Since sub-field shapes are so similar, any reasonable aggregation scheme can be expected to preserve its main characteristics. This is exactly what is found when sub-fields are aggregated into what we call disciplines and fields according to the two schemes

Table 1 Characteristic scores and scales: means (and standard deviations)

	Percentage of articles in categories		Percentage of references in categories	
	1 + 2	4 + 5	2	4 + 5
A. TS sub-fields				
Reference distributions	57.5 (3.1)	16.0 (1.8)	31.3 (5.3)	35.0 (4.6)
			Percentage of citations in categories	
Citation distributions	68.6 (3.7)	10.0 (1.7)	21.1 (5.0)	44.9 (4.6)
Citation distributions	Percentage of articles in categories		Percentage of citations in categories	
B. TvL scheme				
Disciplines	69.5 (2.8)	9.3 (1.5)	20.4 (5.2)	45.6 (4.3)
Fields	70.3 (2.7)	9.0 (1.5)	20.6 (4.0)	46.1 (3.5)
C. GS scheme				
Disciplines	68.8 (3.0)	9.8 (1.4)	22.6 (3.1)	43.8 (2.9)
Fields	69.7 (1.8)	8.9 (0.7)	22.2 (2.4)	43.8 (1.8)

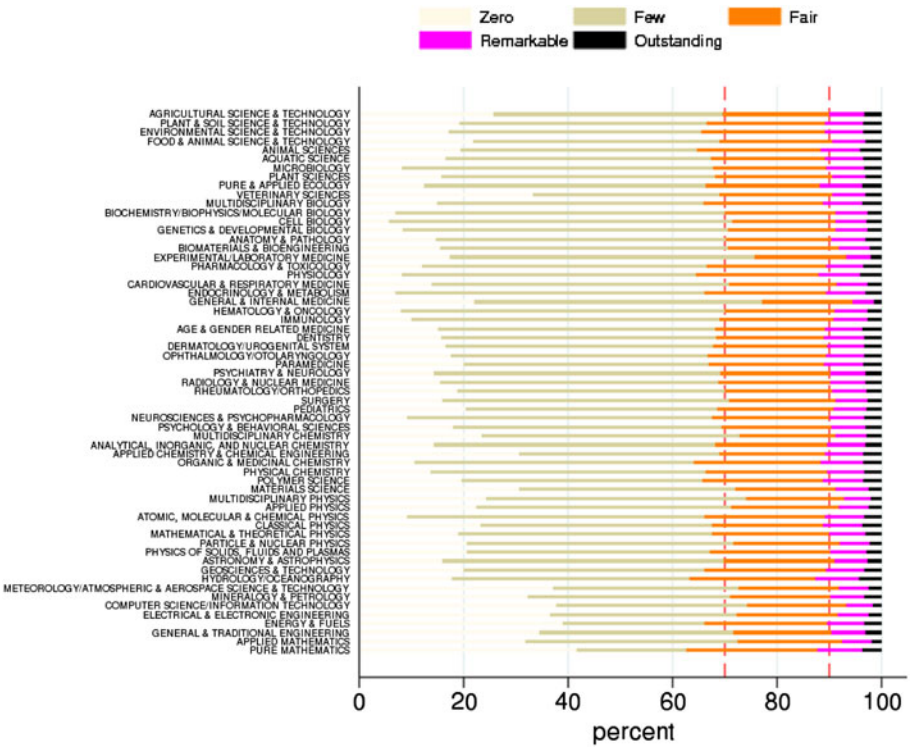

Fig. 1 Citations received by articles in the natural sciences published in 1998–2002 with a five-year citation window, classified into the 61 disciplines distinguished by Glänzel and Schubert (2003)

mentioned in the Introduction (Panels B and C in Table 1).[4] Figure 1 illustrates the classification of citation distributions into the five citation categories for the 61 disciplines in the GS scheme (for the same information about the 38 disciplines in the TvL scheme, see Fig. 2 in Albarrán et al. 2011).

In a nutshell: when citation distributions are partitioned into three classes—including uncited and poorly cited articles below the mean citation rate, fairly cited articles, and remarkably and outstandingly cited articles—they exhibit a strikingly similar shape, and this is the shape of a highly skewed distribution in which a relatively small percentage of highly cited articles account for a large percentage of all citations. However, the fact that the coefficient of variation—namely, the ratio of the standard deviation to the mean—tends to increase as we move inside the union of categories 1 and 2 and categories 4 and 5 reveal the lack of universality of citation distributions' shapes across sub-fields. This conclusion qualifies the analysis in Glänzel (2010) and contrasts with the more optimistic view offered by Radicchi et al. (2008) (see Tables 7 and 8, as well the discussion in Albarrán et al. 2011).

[4] It is important to emphasize that these results coincide with those obtained in Albarrán and Ruiz-Castillo (2011) when analyzing reference and citation distributions using the original dataset in which each article is assigned by TS to only one of 22 broad fields.

Power law characteristics

Consider articles in a given field, and let x be the number of citations received by an article. This quantity is said to obey a power law if it is drawn from a probability density function $p(x)$ such that

$$p(x)\mathrm{d}x = \Pr(x \leq X \leq x + \mathrm{d}x) = Cx^{-\alpha}$$

where X is the observed value, C is a normalization constant, and α is known as the exponent or scaling parameter. This density diverges as $x \to 0$, so that there must be some lower bound to the power law behavior, denoted by $\rho > 0$. Then, provided $\alpha > 1$, it is easy to recover the normalization constant that guarantees that the conditional distribution (given that $x \geq \rho$) integrates up to one. Assuming that our data are drawn from a distribution that follows a power law exactly for $x \geq \rho$, and assuming for the moment that ρ is given, the maximum likelihood estimator of the scaling parameter can be derived.

Using the methods discussed in Clauset et al. (2009) to estimate parameters α and ρ, it can be concluded that the existence of a power law representing citation distributions is a prevalent but not a universal phenomenon: in 140 out of 219 sub-fields, covering about 62% of the total number of articles in the sample, the existence of a power law cannot be

Table 2 Power law estimation results

Panel A: Cases where the existence of a power law cannot be rejected

	Number of items/total (percentage of the total) (1)	Percentage of articles in % (2)
TS sub-fields	140/219 (63.9)	61.9
TvL disciplines	26/38 (68.4)	71.5
TvL fields	9/12 (75)	72.8
GS disciplines	39/61 (63.9)	55.1
GS fields	9/11 (81.8)	71.3
New disciplines	59/80 (73.8)	71.8
New fields	16/19 (84.2)	75.5

Panel B: Power law characteristics

	Number of power laws with:			Power law as % total number of articles (Std. deviation) (4)	Percentage citations accounted for power laws (Std. deviation) (5)
	$\alpha \leq 3$ (1)	$\alpha \in (3, 4)$ (2)	$\alpha > 4$ (3)		
TS sub-fields	4	80	56	2.09 (2.74)	13.49 (13.53)
TvL disciplines	0	13	13	0.78 (0.90)	8.38 (7.67)
TvL fields	0	6	3	0.75 (0.92)	9.60 (9.32)
GS disciplines	1	20	18	1.11 (1.43)	8.77 (8.29)
GS fields	0	5	4	0.52 (0.58)	6.92 (6.81)
New disciplines	1	33	25	1.20 (1.33)	10.34 (8.22)
New fields	0	8	8	0.63 (0.54)	8.33 (6.61)

rejected. Table 2 includes some summary results at different aggregation levels. It should be emphasized that, when they exist, power laws at the sub-field level (i) have a scaling parameter α larger than usually believed, which implies that the citation inequality among the articles in the power law is smaller than what was previously thought, (ii) have a relatively large ρ so that they only represent a small proportion of the upper tail of citation distributions, and (iii) account for a considerable percentage of all citations. Although subject to a large dispersion, on average power laws represent 2% of all articles in a sub-field, and account for about 13.5% of all citations (individual results can be found in Table E in Albarrán et al. 2011, while Fig. 3 in that paper graphically illustrates the distribution followed for some key characteristics).

When moving up from the sub-field level to other aggregate categories, we find that the power law algebra operates in a very subtle way: sub-fields for which a power law does not exist may be aggregated into a category for which the existence of a power law cannot be rejected; on the other hand, power law behavior at the sub-field level is not always preserved in aggregation; in particular, a single sub-field may be responsible for the power law behavior of a large number of sub-fields disappearing. Heterogeneous broad fields, such as Engineering, Physics, or Chemistry, can be fruitfully partitioned into a number of disciplines, many of which present power law behavior. On the contrary, disciplines in the Biomedical Sciences and Clinical Medicine often fail to be represented by a power law. At any rate, higher aggregates for which the existence of a power law cannot be rejected tend to cover between 70 and 80% of all articles in the sample and, when they exist, power laws at these aggregate levels tend to be flatter, smaller and accountable for smaller percentages of citations than those at the sub-field level.

Finally, it is possible to use the experience obtained with the TvL and GS schemes to devise an aggregation scheme into disciplines and fields that maximizes power law behavior. For example, we can adopt the TvL breakdown of Engineering into ten disciplines, seven of which exhibit power law behavior, as well as the GS breakdown of Physics and Chemistry into thirteen disciplines, ten of which exhibit power law behavior. Using this strategy, this paper suggests a third aggregation scheme (Table F in the Appendix in Albarrán et al. 2011 describes how the 219 sub-fields are classified into 80 disciplines and 19 fields). The existence of a power law cannot be rejected in 59 of 80 disciplines and 16 of 19 fields, accounting for 71.8 and 75.5% of all articles in the respective extended samples.

Conclusions and extensions

Summary of results

In brief, using a large dataset we have presented convincing systematic evidence about the existence of fundamental regularities in the shape of reference and citation distributions at different aggregation levels. This is important because, as anticipated in Albarrán and Ruiz-Castillo (2011), this massive evidence points towards a single theoretical explanation of the decentralized process whereby scientists make references that a few years later will translate into a highly skewed citation distribution crowned in many cases by a power law. Recent contributions using a social network approach by, for example, Dorogovstev and Mendes (2001), Jackson and Rogers (2007), and Peterson et al. (2010) constitute a formidable first attempt in this direction.

Nevertheless, this paper has also established that when we look into subsets of articles in the lower and upper tails of citation distributions the appearance of relatively large

coefficients of variation reveal that the existence of common features partially breaks down. This conclusion contrasts with the more optimistic and universalistic view offered by Radicchi et al. (2008), and presents an added challenge, for example, to any attempt at explaining the formation of a power law at the very end of citation distributions.

Extensions

Together with Glänzel (2007a, 2010), Redner (1998, 2005), Lehmann et al. (2003, 2008), Van Raan (2006), Radicchi et al. (2008), and Albarrán and Ruiz-Castillo (2011), these results provide the most complete evidence available in the Scientometrics literature about the skewness of science and the prevalence of power laws in the citation distributions arising from the academic periodicals indexed by TS (or other comparable periodicals collections). The following issues are left for further research.

(a) As indicated in Albarrán and Ruiz-Castillo (2011), from a statistical point of view there are two directions in which this work can be extended. Firstly, the fact that a power law cannot be rejected does not guarantee that a power law is the best distribution to fit the data at the upper tail of citation distributions. New tests must be applied confronting power laws with alternative distributions, such as the lognormal distribution for which Radicchi et al. (2008) present some evidence, the escort distribution suggested in Tsallis and de Alburquerque (2000), or other extreme distributions. Secondly, the ML estimation approach used so far might be quite vulnerable to the existence of a few, but potentially influential extreme observations consisting of a small set of highly cited articles. Consequently, robust estimation methods are worthwhile exploring. In addition, a dynamic model of the citation process may allow us to select variable citation windows to ensure that a common percentage of the process is completed in all distributions in the dataset. This may strengthen the similarities at the lower tail of citation distributions.

(b) It has been observed that, when a power law is present, it only covers a relatively small percentage of articles. Therefore, the rest of the citation distribution needs to be systematically studied. For the results obtained taking a global and macroscopic perspective, see Wallace et al. (2009), as well as the references quoted there. Also, given the parallelism between citation distributions and income distributions (articles are interchangeable with individuals, and citations with incomes), a reasonable suggestion is to apply in Scientometrics the same statistical methods that have been proved useful in Economics (see *inter alia*, Kleiber and Kotz 2003).

(c) The abundance of sub-fields motivates the search for schemes that allow us to work with a smaller number of aggregate categories. This paper has studied the consequences of different aggregation procedures for the distribution of articles into citation categories, and for the existence of a power law representing the very upper tail of citation distributions. However, it remains to be investigated whether the upper tail of aggregate categories is a fair mix of the upper tail of the constituent sub-fields, or whether it is dominated by a single sub-field or a small subset of them.

(d) At present, the assignment of articles to sub-fields is often done through the assignment of the journals where the articles are published. In the TS case, this leads to 42% of all articles being assigned to two or more sub-fields. In the multiplicative strategy followed in this paper, where each article is wholly counted as many times as sub-fields it is assigned to, the resulting extended count is 57% larger than the number of articles in the original dataset. Naturally, since the extent of the multi-assignment

problem decreases as we proceed upwards in the aggregation scheme, there exists a different extended count at every aggregation level. This breaks down the natural connection between aggregation levels, a fact that may not affect the results much on the skewness of science using the CSS approach, but it may have unknown consequences for power law behavior across aggregation levels, and it may also affect other research in which aggregation issues are critical. To solve this problem, it is crucial to construct schemes in which each article is directly assigned to a single sub-field (see *inter alia*, Glänzel and Schubert 2003; Waltman et al. 2010) on the basis of its references, its key words, and other techniques that may include the testing for the existence of a power law. The obvious difficulty of truly interdisciplinary research belonging to several very closely related sub-fields might be solved by creating new mixed sub-fields containing them. In turn, for research in aggregation issues it would be extremely convenient if each sub-field were assigned to a single discipline, and each discipline to a single field, on the basis of experts' opinions, as well as bibliometric techniques that may include the preservation, or generation as the case may be, of power law behavior.

Acknowledgments The authors acknowledge financial support from the Spanish MEC through grants SEJ2007-63098, SEJ2007-67436, ECO2009-11165, and ECO2010-19596. The database of Thomson Scientific (formerly Thomson-ISI; Institute for Scientific Information) has been acquired with funds from Santander Universities Global Division of Banco Santander. This paper is part of the SCIFI-GLOW Collaborative Project supported by the European Commission's Seventh Research Framework Programme, CoSSH7-CT-2008-217436, and was presented in a Poster Session of the STI Conference held in Leiden, 9-11 September, 2010. References and suggestions by a referee led to an improved version of the paper.

References

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2011). The skewness of science in 219 sub-fields and a number of aggregates. Working Paper 11-09, Universidad Carlos III.

Albarrán, P., & Ruiz-Castillo, J. (2011). References made and citations received by scientific articles. *Journal of the American Society for Information Science and Technology, 62*, 40–49.

Burke, P., & Butler, L. (1996). Publication types, citation rates, and evaluation. *Scientometrics, 37*, 473–494.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review, 51*, 661–703.

Dorogovstev, S., & Mendes, J. (2001). Scaling properties of scale-free evolving networks: Continuous approach. *Physical Review E, 85*, 4633–4636.

Egghe, L. (2005). *Power laws in the information production process: Lotkaian informetrics*. Kidlington: Elsevier Academic Press.

Glänzel, W. (2007a). Characteristic scores and scales: A bibliometric analysis of subject characteristics based on long-term citation observation. *Journal of Informetrics, 1*, 92–102.

Glänzel, W. (2007b). Some new applications of the *h*-index. *ISSI Newsletter, 3*, 28–31.

Glänzel, W. (2008). On some new bibliometric applications of statistics related to the *h*-index. *Scientometrics, 77*, 187–196.

Glänzel, W. (2010). The application of characteristics scores and scales to the evaluation and ranking of scientific journals. In *Proceedings of INFO 2010* (pp. 1–13). Havan, Cuba.

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56*, 357–367.

Irvine, J., & Martin, B. R. (1984). CERN: Past performance and future prospects. II. The scientific performance of the CERN accelerators. *Research Policy, 13*, 247–284.

Jackson, M., & Rogers, B. (2007). Meeting strangers and friends of friends: How random are social networks? *American Economic Review, 97*, 890–915.

Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. Hoboken: Wiley.

Laherrère, J., & Sornette, D. (1998). Stretched exponential distributions in nature and economy: Fat tails with characteristic scales. *European Physical Journal B, 2*, 525–539.
Lehmann, S., Lautrup, B., & Jackson, A. D. (2003). Citation networks in high energy physics. *Physical Review, E68*, 026113–026118.
Lehmann, S., Lautrup, B., & Jackson, A. D. (2008). A quantitative analysis of indicators of scientific performance. *Scientometrics, 76*, 369–390.
Liang, L., & Rousseau, R. (2010). Reference analysis: A view in the mirror of citation analysis. *Geomatics and Information Science of Wuhan University, 35*, 6–9.
Magyar, G. (1973). Bibliometric analysis of a new research sub-field. *Journal of Documentation, 30*, 32–40.
Narayan, S. (1971). Power law relations in science bibliography—a self consistent interpretation. *Journal of Documentation, 27*, 83–97.
Peterson, G., Presse, S., & Dill, K. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *Proceedings of the National Academy of Sciences of the United States of America, 107*, 16023–16027.
Price, D. J. D. S. (1965). Networks of scientific papers. *Science, 149*, 510–515.
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 the 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, 4*, 131–134.
Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today* (pp. 49–54).
Schubert, A., & Glänzel, W. (2007). A systematic analysis of Hirsh-type indices for journals. *Journal of Informetrics, 1*, 2179–2184.
Schubert, A., Glänzel, W., & Braun, T. (1987). A new methodology for ranking scientific institutions. *Scientometrics, 12*, 267–292.
Schubert, A., Glänzel, W., & Braun, T. (1989). Scientometric datafiles: A comprehensive set of indicators on 2,649 journals, 96 countries in all major fields, sub-fields 1981–1985. *Scientometrics, 16*, 3–478.
Seglen, P. (1992). The skewness of science. *Journal of the American Society for Information Science, 43*, 628–638.
Tijssen, J. W., van Leeuwen, T. (2003). *"Bibliometric analysis of world science", extended technical annex to chapter 5 of the third European report on science and technology indicators*. Directorate-General for Research. Luxembourg: Office for Official Publications of the European Community.
Tsallis, C., & de Alburquerque, M. P. (2000). Are citations of scientific papers a case of nonextensivity? *European Physical Journal B, 13*, 777–780.
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.
Wallace, M., Larivière, V., & Gingras, Y. (2009). Modeling a century of citation distributions. *Journal of Informetrics, 3*, 296–303.
Waltman, L., van Eck, N. J., & Noyons, E. (2010). A unified approach to mapping and clustering of bibliometric networks. *Journal of Informetrics, 4*, 629–635.