References Made and Citations Received by Scientific Articles

Pedro Albarrán
Departamento de Fundamentos del Análisis Económico, Universidad de Alicante, Alicante, Spain. E-mail: albarran@ua.es

Javier Ruiz-Castillo
Departamento de Economía, Universidad Carlos III, Madrid, Spain. E-mail: jrc@eco.uc3m.es

This article studies massive evidence about references made and citations received after a 5-year citation window by 3.7 million articles published in 1998 to 2002 in 22 scientific fields. We find that the distributions of references made and citations received share a number of basic features across sciences. Reference distributions are rather skewed to the right while citation distributions are even more highly skewed: The mean is about 20 percentage points to the right of the median, and articles with a remarkable or an outstanding number of citations represent about 9% of the total. Moreover, the existence of a power law representing the upper tail of citation distributions cannot be rejected in 17 fields whose articles represent 74.7% of the total. Contrary to the evidence in other contexts, the value of the scale parameter is above 3.5 in 13 of the 17 cases. Finally, power laws are typically small, but capture a considerable proportion of the total citations received.

Introduction

This article studies the following problem: Are the citation distributions of different sciences very different among themselves or do they share a number of essential characteristics despite differences in publication and citation practices across scientific fields? The answer is important for any attempt at explaining how these distributions get formed. Whether citation distributions are very different or can be described in terms of a few stylized features would determine whether we must search for as many explanations as distribution types or for a single explanation capable of accounting for the fundamental characteristics shared by all the distributions in question.

The article searches for regularities across sciences in two dimensions. First, we investigate how the distribution of references made by articles in a given field becomes a highly skewed distribution of citations received in which a large proportion of articles gets none or few citations while a small percentage of them account for a disproportionate amount of all citations.[1] We are able to provide a much more complete view of this process than the picture drawn in Price's (1965) pioneer contribution with the newly available (but limited) data during the early 1960s, or in Seglen's (1992) seminal contribution where the skewness of citation distributions is only illustrated for a random sample of articles drawn from the 1985 to 1989 *Science Citation Index*, and for Magyar's (1973) data on the small subfield of dye laser research.[2] The case of Vinkler (2009) is paradigmatic. He stated that "As is well known, the distribution of citations by paper . . . may be rather skewed" (p. 602), but his only references are to Seglen and to papers by Burke and Butler (1996) on the entire fields of the natural sciences and the social sciences and humanities in Australian universities, and to Irvine and Martin (1984) and Lehmann et al. (2003), both on high-energy physics. Two clear exceptions are the important contributions by Schubert, Glänzel, and Braun (1987), which described the skewness of articles published and cited in 1981 to 1985 in 114 subfields, and by Glänzel (2007), who studied 450,000 citable papers published in 1980, cited in the 1980 to 2000 period, and classified into 60 subfields and 12 major fields (However, Glänzel, 2007, only reported results for 12 subfields while Glänzel, in press, studied papers published in 2006 with a 3-year citation window, but only reported results for three subfields.)

[1] In the eloquent summary by Lehmann, Lautrup, and Jackson (2003), "The picture which emerges is thus a small number of interesting and significant papers swimming in a sea of dead papers" (p. 7).

[2] Seglen (1992) also illustrated the skewness of citations to articles from single journals and from single authors, a type of citation distribution beyond this article's scope.

Second, it is generally believed that the citation process in the periodical literature is one of the aspects of scientific activity in which power laws (or other extreme distributions) are prevalent (An extensive discussion of the properties of power laws can be found in reviews by Mitzenmacher, 2004, and Newman, 2005, and in the references therein.) However, the available evidence is very scant indeed. To our knowledge, there are only results for a few samples of articles belonging to certain scientific fields such as physics or high-energy physics, or all fields combined.[3] We investigate the existence of power laws for a broad array of scientific disciplines, including how they are inserted in the rest of the citation distribution.

In other words, this article searches for a compact and systematic description of the distribution of references made and that of citations received by articles in different scientific fields, with special attention to the existence of power laws. A key feature of this empirical investigation is that it provides massive evidence about these issues using a large sample acquired from Thomson Scientific (TS), consisting of about 3.9 million articles published in 1998 to 2002, the almost 10-million references they made, and the more than 28-million citations they received using a 5-year citation window. After excluding the arts and humanities for its intrinsic peculiarities, we are left with the 20 natural sciences and the two social sciences distinguished by TS.

The skewness of the distribution of references made or citations received in any field is described using the characteristic scores and scales (CSS) technique, which permits the partition of any distribution of articles into a number of classes as a function of its members' citation characteristics. Schubert et al. (1987), Glänzel and Schubert (1988), and Glänzel (2007, in press) applied this technique to classify articles into five categories according to whether they received no citations, or were poorly cited, fairly cited, or remarkably or outstandingly cited (discussed later). This classification method has two important invariance properties: The results do not change if the citations received by all articles are multiplied by a common scalar greater than zero (scale or unit invariance), or if the original distribution of articles and the citations they received is replicated any discrete number of times (replication or size invariance).[4] These properties permit us to focus on skewness as a property of the shape of the distributions involved.

The estimation of a power law presents more subtle technical problems. From a statistical point of view, the estimation of a power law and the evaluation of the goodness-of-fit are known to be much more complex than is the direct linear fit of the log–log plot of the full raw histogram of the data, let alone the mere inspection of the histogram plotted on logarithmic scales to check whether it looks like a straight line.[5] In this respect, there seems to be unanimity that a maximum likelihood (ML) approach provides the best solution to the estimation problem.

The rest of the article is organized in three sections. The first section presents the 1998 to 2002 sample as well as the classification of reference and citation distributions in all fields into five characteristic classes following the CSS approach. The next section presents the results of the power law estimation in 22 fields (excluding arts and humanities) and all sciences as a whole. Finally, the last section discusses the main findings and a number of possible extensions.

The Data and a Characterization of the Reference and Citation Distributions

The Data

TS-indexed journal articles include research articles, reviews, proceedings papers, and research notes. In this article, only research articles, or simply articles, are studied, so 390,097 review articles and three notes are disregarded. The 52,789 articles without information about some variables (number of authors, Web of Science category, or TS field) also are eliminated from the analysis. Thus, the initial sample size consists of 8,470,666 articles published in 1998 to 2007, or 95% of the number of items in the original database. For this article, we have restricted ourselves to the sample of articles published in 1998 to 2002. How representative is this sample, consisting of 3,912,097 articles? And how large is the number of articles in the smallest sciences? The information on these issues is in Table 1, where the 1998 to 2007 and 1998 to 2002 samples are compared. The 20 fields in the natural sciences are organized in three large groups: Life Sciences, Physical Sciences, and Other Natural Sciences. Physical Sciences and Other Natural Sciences in the larger sample represent approximately 28 and 26%, respectively, of the total while Life Sciences represent about 37%. The remaining 9% corresponds to the two Social Sciences and Arts and Humanities. The distribution of the 1998 to 2002 sample by fields is very similar: It contains more Life and Social Sciences articles (1.1 and 0.4%, respectively), and somewhat less from the Physical and the Other Natural Sciences. Therefore, the 1998 to 2002 sample can be taken to be representative of the larger sample. On the other hand, for most fields, the 1998 to 2002 sample size is rather large: Twelve fields have more than 100,000 articles; 10 fields have between this number and 49,000 articles; and only the Multidisciplinary field has about 21,000 articles.

[3] Beyond the graphical illustrations included in Seglen (1992), the only directly estimated results we have found are those of Redner (1998, 2005), Lehmann et al. (2003; Lehmann, Lautrup, & Jackson, 2008), and Clauset, Shalizi, and Newman (2009); Laherrère and Sornette (1998) studied the citation record of the most cited physicists. Under the hypothesis that citation distributions follow a power law, Glänzel (2007) obtained an equation relating the scale parameter of a power law and the parameters of the characteristic scores and scales technique (discussed later); with direct estimates of the latter, estimates of the former are computed.

[4] Of course, these properties also are satisfied for the partition of articles into classes according to the references they make.

[5] See inter alia Nicholls (1987), Pickering, Bull, and Sanderson (1995), Clark, Cox, and Laslett (1999), Goldstein, Morris, and Yen (2004), Bauke (2007), Clauset et al. (2009), and White, Enquist, and Green (2008).

TABLE 1. Articles by Thomson Scientific field in the entire 1998 to 2007 dataset and in the 1998 to 2002 sample.

	1998–2007		1998–2002	
	Dataset	%	Sample	%
Life Sciences	3,165,734	37.4	1,507,634	38.5
1. Clinical Medicine	1,667,362	19.7	791,723	20.2
2. Biology & Biochemistry	470,483	5.6	228,908	5.9
3. Neuroscience & Behavioral Science	244,508	2.9	116,100	3.0
4. Molecular Biology & Genetics	216,835	2.6	102,800	2.6
5. Psychiatry & Psychology	198,225	2.3	91,905	2.3
6. Pharmacology & Toxicology	135,116	1.6	64,271	1.6
7. Microbiology	130,458	1.5	60,754	1.6
8. Immunology	102,747	1.2	51,173	1.3
Physical Sciences	2,365,084	27.9	1,056,552	27.0
9. Chemistry	1,004,835	11.9	458,373	11.7
10. Physics	809,301	9.6	375,075	9.6
11. Computer Science	233,757	2.8	76,460	2.0
12. Mathematics	212,496	2.5	97,309	2.5
13. Space Science	104,695	1.2	49,335	1.3
Other Natural Sciences	2,186,875	25.8	987,794	25.2
14. Engineering	701,423	8.3	318,504	8.1
15. Plant & Animal Sciences	466,587	5.5	218,385	5.6
16. Materials Science	388,218	4.6	168,724	4.3
17. Geoscience	228,221	2.7	101,783	2.6
18. Environment & Ecology	207,795	2.5	90,520	2.3
19. Agricultural Sciences	155,466	1.8	69,051	1.8
20. Multidisciplinary	39,165	0.5	20,827	0.5
Social Sciences	469,799	5.5	220,014	5.6
21. Social Sciences, General	337,041	4.0	156,523	4.0
22. Economics & Business	132,758	1.6	63,491	1.6
Arts & Humanities	283,174	3.3	140,103	3.6
23. Arts & Humanities	283,174	3.3	140,103	3.6
All Fields	8,470,666	100.0	3,912,097	100.0
Reviews and Notes	390,100			
Articles without information about some variables	52,789			
No. of “Items” in the original database	8,913,555			

The original dataset consists of articles published in a certain year and the citations they receive from that year until 2007; that is, articles published in 1998 and their citations during the 10-year period 1998 to 2007, articles published in 1999 and their citations in the 9-year period 1999 to 2007, and so on until articles published in 2007 and their citations during that same year. The time pattern of citations varies a lot among the different disciplines. In this situation, ideally, the citation window in each field should be estimated along other features of the stationary distribution in a dynamic model; however, the estimation problem is beyond the scope of this article. Therefore, it was decided to take all fields equally by taking a fixed, common window for all of them. The standard length of citation windows in the literature is 3 years, possibly because it is large enough for the citation process to be settled in the quickest disciplines that include most natural sciences (see inter alia Moed, De Bruin, & van Leeuwen, 1995). However, we wanted to make sure that the slowest sciences were relatively well covered. But the greater the citation window, the smaller the sample size had to be. We settled on a common 5-year citation window for all articles published in 1998 to 2002.

Note that the simplification of taking a common citation window implies that certain idiosyncratic features that differentiate some fields from each other will be preserved in our data: Five years is a long-enough period for the completion of a sizable part of the citation process for some disciplines but rather short for others, notably the Social Sciences and other “slower” fields such as Psychiatry and Psychology, Geosciences, and Environmental and Ecology. However, Glänzel (2007) established that except for a short initial period of 4 years—under our 5-year choice—the particular length of a citation window was not important for the class sizes determined in the CCS approach applied (discussed later). Having selected a rather large citation window, together with a large sample size, we conjecture that we also are on the “safe side” for the estimation of a power law.

Differences Across Fields in the Citation Process

For each field, Table 2 presents descriptive statistics about the two sides of the citation process: 99,767,108 references made as well as 28,426,632 citations received in the 1998 to 2002 sample. Naturally, the citations received by articles in a

TABLE 2. The distribution of references made and citations received.

	References			Citations			
	M	CV	Ratio References/ Citations	%zeros	*M*	CV	h-index
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Life Sciences							
1. Clinical Medicine	25.5	0.67	2.7	16.4	9.4	2.27	323
2. Biology & Biochemistry	33.7	0.52	2.7	9.9	12.3	1.62	187
3. Neuroscience & Behavioral Science	37.1	0.56	2.7	7.5	13.5	1.35	161
4. Molecular Biology & Genetics	38.2	0.50	1.9	7.4	20.2	1.63	253
5. Psychiatry & Psychology	34.8	0.62	5.2	18.7	6.7	1.63	107
6. Pharmacology & Toxicology	28.6	0.60	3.7	13.1	7.7	1.41	94
7. Microbiology	32.4	0.52	2.9	8.1	11.3	1.23	108
8. Immunology	35.5	0.48	2.2	4.6	16.0	1.41	161
Physical Sciences							
9. Chemistry	24.6	0.69	3.4	18.2	7.3	1.75	156
10. Physics	20.7	0.71	3.0	22.0	6.8	2.23	198
11. Computer Science	18.1	0.76	6.4	43.2	2.8	4.75	85
12. Mathematics	16.8	0.70	7.1	37.6	2.4	1.90	50
13. Space Science	31.1	0.66	2.9	18.1	10.8	1.76	138
Other Natural Sciences							
14. Engineering	15.9	0.85	5.6	39.8	2.8	1.90	85
15. Plant & Animal Sciences	28.4	0.66	5.7	22.3	4.9	1.59	97
16. Materials Science	17.3	0.75	4.2	31.3	4.1	1.93	97
17. Geoscience	31.7	0.71	5.1	22.0	6.3	1.57	92
18. Environment & Ecology	31.2	0.65	4.6	15.6	6.7	1.42	88
19. Agricultural Sciences	23.6	0.69	5.2	26.0	3.5	1.54	69
20. Multidisciplinary	15.5	1.06	4.5	46.3	3.4	3.06	69
Social Sciences							
21. Social Sciences, General	30.9	0.80	10.5	36.1	3.0	1.81	71
22. Economics & Business	24.0	0.90	7.6	40.7	3.2	2.00	63
Arts & Humanities							
23. Arts & Humanities	19.4	1.12	38.2	21.8	0.5	6.63	67
All Sciences	25.7	0.72	3.4	82.9	7.5	2.13	170

certain field would depend on the reference distribution in that field. In particular, the higher the mean (or the median, not shown in Table 2 but available on request), the higher the total citations received will be—and, presumably, the smaller the percentage of articles with zero citations. But 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 will be to articles published in TS journals before 1998 and, hence, outside of our dataset. The larger the number of references made to recently published articles, the larger the number of citations received tend to be, and the smaller the ratio references made/citations received in Column 3 in Table 2.

Fields can be classified in three groups according to the value of the references/citations ratio: (Group A) six of the eight Life Sciences and Space Science, characterized by a relatively low value (between 1.9–3) of the ratio; (Group B) the two remaining Life Sciences and another seven natural sciences with a ratio between 3 and 5.2; and (Group C) a group of seven fields with a ratio greater than 5.2 (including Engineering, Plant and Animal Sciences, Computer Sciences, Mathematics, the two Social Sciences, plus Arts and Humanities with a value equal to 38.2). With few exceptions, the means of the reference distributions in Group C are relatively small, ranging from 15.8 to 30.9, and relative high in Group A, ranging from 25.5 to 38.2, with intermediate values in Group B. On the other hand, reference and citation inequality are measured by the coefficient of variation (CV); that is, the *SD* normalized by the mean. A negative association across fields can be seen between the mean in the reference distribution and the CV (The correlation coefficient between Columns 1 and 2 in Table 1 is −0.73.) Correspondingly, the dispersion of the former is greater than the dispersion of the latter. Mean differences across fields are important: They range from fewer than 17 references per article for Engineering and Mathematics to more than 37 for Neuroscience and Behavioral Science, and Molecular Biology and Genetics. The CV ranges from 0.48 for Immunology to more than 1 for Multidisciplinary and Arts and Humanities; however, it is between 0.5 and 0.7 for 13 disciplines and between 0.71 and 0.80 for the remaining seven.

Thus, fields in Group C make fewer references on average and receive fewer citations. Correspondingly, they are characterized by a relatively high percentage of articles with no citations at all, a relatively low mean, and a relatively low h-index (Columns 4, 5, and 7 in Table 2). Indeed, for six of these seven fields, the percentage of articles without citations ranges from 22.3 to 43.2% while for the remaining field in Group C, Arts and Humanities, the percentage is an astronomical 82.9%. With few exceptions, the opposite is the case for Life Science fields in Group A: The percentage of articles with zero citations ranges from 4.6 to 16.4% while Group B is characterized by intermediate values. Since greater mean references are associated with smaller reference/citations ratios, the dispersion of mean citations increases: Apart from an uncommon low mean of 0.5 citations per article for Arts and Humanities, mean citation ranges from a low 2.4 per article in Computer Science to a value greater than 9 in most fields in Group A, with Molecular Biology and Genetics the highest with 20.2 citations per article. Similarly, the h-index in Column 7 ranges from 50 in Mathematics (or 63 in Economics and Business, and 67 in Arts and Humanities) to 253 in Molecular Biology and Genetics, and 323 in Clinical Medicine. On the other hand, when we go from the reference to the citation distribution, the CV dramatically increases by a factor greater than 3 or 4 generally, and greater than 6 in Arts and Humanities and Computer Science (Column 6). Citation inequality now ranges from 1.2 in Microbiology to 4.7 in Computer Science and 6.6 in Arts and Humanities. But, as before, once the extreme values are taken away, the range is very limited: There are 17 fields with a CV between 1.35 and 1.99, and 3 more with this measure between 2 and 3.1.

The overall conclusion is that as expected, the reference and citation processes present large differences across fields. The reference distribution of fields in Group A are characterized by low reference/citation ratios, a high mean, and a relatively low CV; correspondingly, these fields tend to have lower percentages of articles without citations, higher citation means, and higher h-indices. Fields in Group C present the opposite pattern while fields in Group B constitute an intermediate case. Citation inequality is always much greater than is reference inequality; however, as soon as we normalize by the mean in the CV, both distributions become considerably more similar across fields.

The 1998 to 2002 and the 1998 to 2007 reference distributions are very similar indeed (Results for the original 1998–2007 dataset are available on request.) Likewise, a 5-year citation window for the articles published in 1998 to 2002 appears to be large enough for the sample's citation distribution to closely resemble that of the entire dataset. Taking into account that the sample's distribution by field also is very similar to that of the dataset (see Table 1), we are confident that the 1998 to 2002 sample constitutes a good testing bank to explore the empirical issues that motivate this article.

A special case should be singled out: It is clear that Arts and Humanities constitutes an entirely different, or an extreme, case of a scholarly field that makes relatively few references, a very small part of which appear as citations received by articles published only a few years later in TS-indexed journals. This leads us to eliminate this field from further analysis and to define the all-sciences category as the sum of the remaining 22 TS scientific fields; namely, 3,771,994 articles that made 97,043,743 references and received 28,355,343 citations.

Similarities Across Fields: References Made

In this subsection, the CSS methodology is applied to the ordered distribution of references made by the articles published in 1998 to 2002, $r = (r_1, \ldots, r_n)$ with $r_1 \leq r_2 \leq \ldots \leq r_n$, where r_i is the number of references made by the ith article, $i = 1, \ldots, n$. The following *characteristic scores* are determined:

$s_0 = 0$.
$s_1 =$ mean references per article.
$s_2 =$ mean references of articles with references above average.
$s_3 =$ mean references of articles with references above s_2.

These scores are used to partition the set of articles into five categories:

Category 0 = articles that make *no* references;
$r = s_0$

Category 1 = articles that make *few* references; namely,
$r \in (s_0, s_1]$ references lower than average;

Category 2 = articles that make a *fair* number of references;
$r \in [s_1, s_2)$ namely, at least average references but below s_2;

Category 3 = articles that make a *remarkable* number of
$r \in [s_2, s_3)$ references; namely, no lower than s_2 but below s_3;

Category 4 = articles that make an *outstanding* number of
$r \geq s_3$ references; namely, no lower than s_3.

As indicated in the Introduction, the classification of any distribution into these five categories satisfies two important properties, also satisfied by the CV: The classification is invariant when the references each article makes are multiplied by any positive scalar, and when the initial distribution is replicated any discrete number of times. The first property implies that the classification method is independent of the units in which references are measured. Consequently, it allows for a comparison of two distributions with different means. The second property implies that the classification method only responds to references per article. Consequently, it allows for a comparison of distributions of different sizes.[6]

[6] Suppose there are two distributions x and y with sizes n and m, respectively. Distributions x and y can be replicated m and n times, respectively, so that each will be of size $n \times m$ after the operation is performed. However, the replication will leave unchanged the classification into five categories of either x or y. Thus, the two distributions could be compared using their corresponding $n \times m$ replicas.

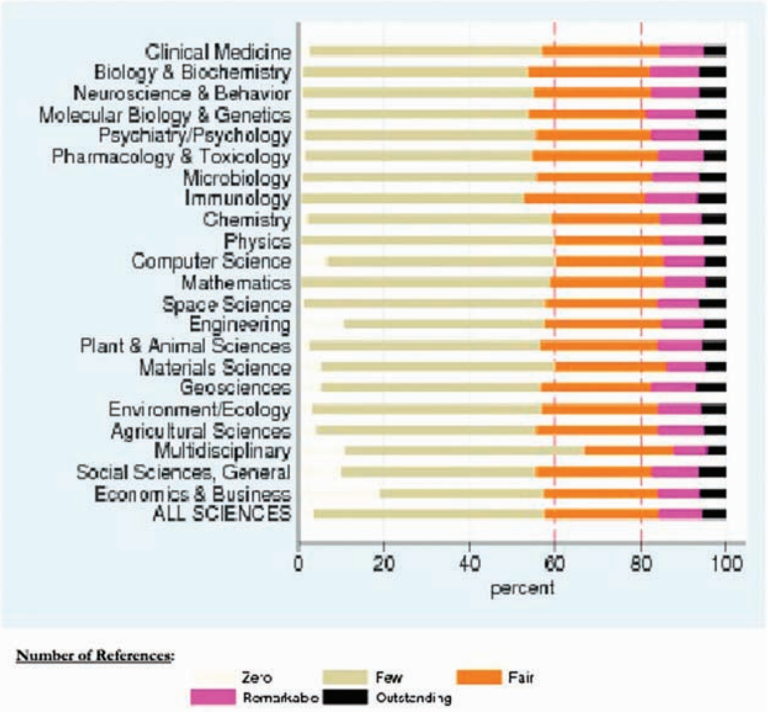

FIG. 1. References made by articles published in 1998 to 2002.

The classification of the reference distributions into five categories for TS fields is in Figure 1. Two comments are in order. First, taking as reference the distribution for ALL SCIENCES combined, it is observed that it is a rather skewed distribution: The mean is well to the right of the median while the last two categories represent about 15% of all articles. Second, after the normalization involved in the classification method, most differences across fields essentially vanish. On average, the first two categories represent 57.4% in the 22 fields, with a minimum value of 53% for Immunology and a maximum value of 67.1% for Multidisciplinary.

Similarities Across Fields: Citations Received

The classification into five categories of articles without citations or poorly cited, fairly cited, remarkably cited, and outstandingly cited articles for the 22 TS fields is in Figure 2. Again, two comments are in order. First, the essential change from Figure 1 is that now all distributions are even more skewed to the right than they were before. Taking ALL SCIENCES as a representative example, a large percentage of articles without citations is observed, the mean is shifted about 10 percentage points to the right, and the last two categories constituting the upper tail of the distribution represent only about 9% of all articles. Second, the only difference across scientific fields is the percentage of articles without citations; however, these differences essentially disappear when the sum of the first two categories is compared. This long lower tail represents on average 70.3% of all articles, with a minimum of 66.3% for Plant and Animal Sciences, and a maximum of 78.2% for Multidisciplinary.

Taking into account the considerable changes in scientific communication during the lat 2 decades (see Persson, Glänzel, & Danell, 2004, who documented the intensification of research collaboration and coauthorship), this 70–21–9 rule for 3.7 million articles published in 1998 to 2002 with a 5-year citation window and classified into 22 TS fields is not that different from the 75–18–7 rule reported in Glänzel (2007) for 450,000 papers published in 1980 with a 20-year citation window and classified into 60 subfields and 12 major fields.

To complete this discussion, one also could ask about the percentage of references made and citations received by each category (beyond the first that by definition, accounts for no references or citations at all). First, on average, Categories 1 and 2 of the reference distributions account for 32 and 33.7% of all references, respectively, while the upper tail formed by 15.9% of all articles in Categories 3 and 4 accounts for the remaining 34.3% of all references. Second, as noted earlier, citation distributions show an even greater skewness to the right than do the reference distributions. Thus, on average, Categories 1 and 2 account only for 22.7 and 33.3%, respectively, of all citations while the upper tail formed by 9.2% of all articles in Categories 3 and 4 accounts for the remaining 44% of all citations.[7]

[7] The skewness of the citation distribution is even more pronounced in the high-energy physics subfield, where Lehmann et al. (2003) reported that 4% of the papers account for half of the citations.

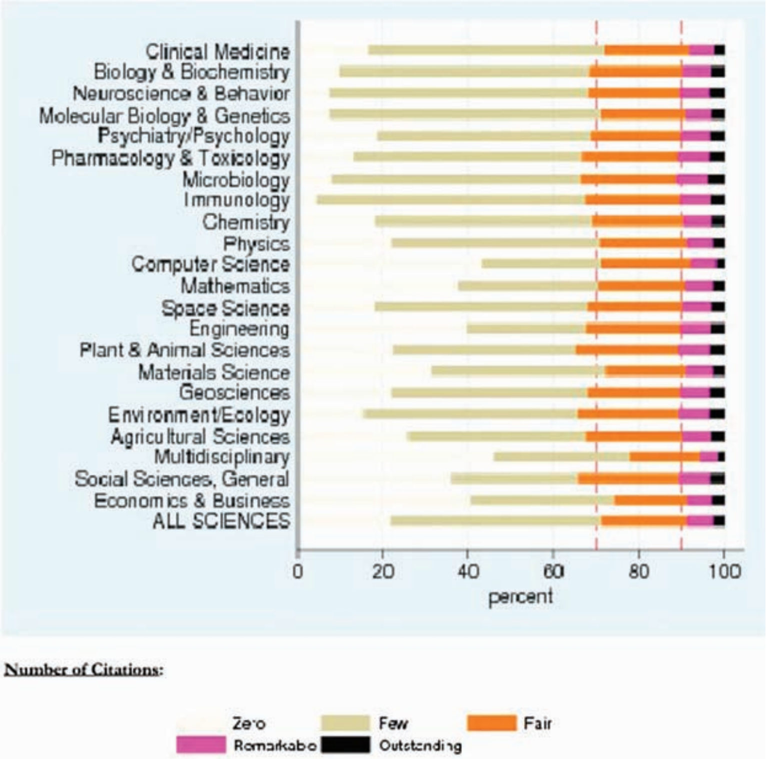

FIG. 2. Citations received by articles published in 1998 to 2002 with a 5-year citation window.

Estimation of the Power Law

The ML Approach

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

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

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 ρ. Then, provided $\alpha > 1$, it is easy to recover the normalization constant, which in the continuous case is shown to be

$$C = (\alpha - 1)\rho^{\alpha-1}.$$

Assuming that in each field 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 ML estimator (MLE) of the scaling parameter can be derived. For instance, the MLE in the continuous case can be shown to be (see Appendix B in Clauset et al., 2009):

$$\hat{\alpha}_{MLE} = 1 + T\left[\sum_{i=1}^{T} \ln \frac{x_i}{\rho}\right]^{-1} \tag{1}$$

where T is the sample size for values $x \geq \rho$. These authors tested the ability of the MLEs to extract the known scaling parameters of synthetic power law data, finding that the MLEs give the best results when compared with several competing methods based on linear regression. Nevertheless, for very small datasets, the MLEs can be significantly biased. Clauset et al. (2009) suggested that $T \geq 50$ is a reasonable rule of thumb for extracting reliable parameter estimates.

The large percentage of articles with no citations at all as well as the low value of the mean in most fields (see Column 5 in Table 2) indicate that we are in the typical case where there is some non-power law behavior at the lower end of the citation distributions. In such cases, it is essential to have a reliable method for estimating the parameter ρ; that is, the power law's starting point. In this article, as in Clauset et al. (2009), we choose the value of ρ that makes the probability distributions of the measured data and the best-fit power law as similar as possible above ρ. To quantify the distance to be minimized between the two probability distributions, the Kolmogorov–Smirnov (KS) statistic is used. Again, Clauset et al. (2009) generated synthetic data and examined their method's ability to recover the known values of ρ. They obtain good results, provided the power law is followed by at least 1,000 observations.

The method described allows us to fit a power law distribution to a given dataset and provides good estimates of the parameters involved.[8] An entirely different question is to

[8] As a matter of fact, to estimate the parameters α and ρ, we use the program that Clauset et al. (2009) made available in http://www.santafe.edu/∼aaronc/powerlaws/

TABLE 3. Power law estimation results: Articles published in 1998 to 2002 with a 5-year citation window.

				No. of power	% of Total	
	α	ρ	*p*-value	law articles	articles	% of Citations
Life Sciences						
1. Clinical Medicine	3.28	136	0.879	2,408	0.30	7.78
2. Biology & Biochemistry	3.82	71	0.233	3,219	1.41	12.64
3. Neuroscience & Behavioral Science	5.05	137	0.304	305	0.26	3.51
4. Molecular Biology & Genetics	3.86	152	0.089	1,073	1.04	11.81
5. Psychiatry & Psychology	3.77	42	0.097	1,495	1.63	15.50
6. Pharmacology & Toxicology	3.73	33	0.000	2,037	3.17	20.61
7. Microbiology	4.56	66	0.457	626	1.03	8.27
8. Immunology	3.57	73	0.367	1,223	2.39	17.37
Physical Sciences						
9. Chemistry	4.02	72	0.099	1,777	0.39	5.79
10. Physics	3.35	55	0.028	4,253	1.13	15.74
11. Computer Science	2.92	18	0.672	1,701	2.22	29.55
12. Mathematics	3.83	20	0.614	841	0.86	11.18
13. Space Science	3.37	62	0.552	909	1.84	17.72
Other Natural Sciences						
14. Engineering	3.59	20	0.015	4,953	1.56	17.21
15. Plant & Animal Sciences	4.16	50	0.157	900	0.41	6.01
16. Material Science	3.62	37	0.245	1,460	0.87	12.20
17. Geosciences	4.02	39	0.254	1,253	1.23	11.38
18. Environment & Ecology	4.14	48	0.633	645	0.71	7.42
19. Agricultural Sciences	3.85	27	0.008	1,111	1.61	14.25
20. Multidisciplinary	3.23	48	0.918	166	0.80	19.51
Social Sciences						
21. Social Sciences, General	3.63	19	0.001	2,928	1.87	18.53
22. Economics & Business	4.63	46	0.667	207	0.33	6.46
All Sciences	3.58	136	0.850	6,119	0.16	4.80

decide whether the power law distribution is even a reasonable hypothesis to begin with; that is, whether the data we observe could have been drawn from a power law distribution. The standard way to answer this question is to compute a *p*-value, defined as the probability that a dataset of the same size that is truly drawn from the hypothesized distribution would have a KS statistic larger than or equal to the one that has been observed. Thus, the *p*-value summarizes the sample evidence that the data were drawn from the hypothesized distribution, based on the observed goodness of fit. Therefore, if the *p*-value is very small, then it is unlikely that the data are drawn from a power law.

To implement this procedure, we again follow Clauset et al. (2009). First, take the value of the KS statistic minimized in the estimation procedure as a measure of its goodness of fit. Second, generate a large number of synthetic datasets that follow a perfect power law with scaling parameter equal to the estimated α above the estimated ρ, but which have the same non-power law behavior as the observed data below it. Third, fit each synthetic dataset according to the estimation method already described, and calculate the KS statistic for each fit. Fourth, calculate the *p*-value as the fraction of the KS statistics for the synthetic datasets whose value exceeds the KS statistic for the real data. If the *p*-value is sufficiently small, say below 0.1, then the power law distribution can be ruled out.

Estimation Results

For the 1998 to 2002 sample with a 5-year citation window, the results of the ML approach are presented in Table 3. Judging by the *p*-value, the results are very satisfactory: In 17 fields—as well as ALL SCIENCES—with a *p*-value close to 0.1 or greater, the existence of a power law cannot be rejected. These fields represent 74.7% of all articles in the natural and the social sciences. In the remaining five fields (Pharmacology and Toxicology, Physics, and Agricultural Sciences from Group B, as well as Engineering, and Social Sciences, General from Group C), the *p*-value is clearly below the critical value 0.1.[9]

With regard to the 17 fields for which the existence of a power law cannot be ruled out, the following three comments are in order:

- Only for Computer Science is the estimated scale parameter between 2 and 3. For three fields, $\hat{\alpha}$ is below 3.5, for seven fields is between 3.5 and 4, for five fields is between 4 and 5, and for the remaining field (Neuroscience and Behavioral Sciences), $\hat{\alpha}$ is greater than 5. This is rather at variance with

[9] This is important when for seven of the 24 datasets rigorously investigated in Clauset et al. (2009)—HTTP connections, earthquakes, Web links, fires, wealth, Web hits, and the metabolic network—the *p*-value is sufficiently small that the power law model can be firmly ruled out.

previous research in bibliometrics: Redner (1998) reported that $\hat{\alpha}$ is approximately 3 for papers published in a single year in a variety of scientific fields while Lehmann et al. (2003) found that for papers with 50 or more citations in high-energy physics, $\hat{\alpha}$ is equal to 2.31. Through indirect methods, Glänzel (2007) concluded that the most relevant range for $\hat{\alpha}$ is [1.5, 3.5].[10]

- As expected, the estimated value of ρ that determines the beginning of the power law is rather low in Group C—ranging from 18 citations in Computer Science to 50 in Plant and Animal Sciences—and very high in Group A—ranging from 66 in Microbiology to 152 in Molecular Biology and Genetics. The estimated value of ρ in Group B ranges from 37 in Materials Science to 72 in Chemistry.
- Perhaps more interestingly, all power laws are of a relatively small size, but account for a considerable percentage of all citations in their field. The power laws in eight fields represent between 0.2 and 0.9% of all articles and account for 3.5 to 12.2% of all citations. In six fields, the power laws represent between 1 and 1.9% and capture between 8.3 and 17.7% of all citations. Computer Science and Immunology represent 2.2 and 2.4%, respectively, of all articles, and account for 29.5 and 17.4%, respectively, of all citations. Finally, the power law in the Multidisciplinary field accounts for 19.5% of all citations.[11]

Discussion

Summary and Results

This article has been concerned with the question of whether the distributions of references made and citations received by scientific articles have many things in common. Publication and citation practices are very different across disciplines. As a result, certain key statistics—such as the mean reference or the mean citation ratio, the percentage of articles without citations, or indicators of scientific excellence such as the h-index—exhibit a large degree of variation across scientific fields. However, this article has demonstrated that from another perspective, the shape of the reference and citation distributions of different sciences share many basic features.

The article has analyzed the largest dataset ever investigated in search of basic differences or similarities across 22 broad fields, consisting of about 3.7-million articles published in 1998 to 2002 with a 5-year citation window. We have used state-of-the-art techniques; namely, we have ranked references made and citations received into five classes using the CSS approach, and we have searched for the existence of a power law in the upper tail of citation distributions using ML methods. The main results can be summarized by the following three observations:

- Reference distributions are rather skewed to the right: The mean is almost 10 percentage points to the right of the median, and articles with a remarkable or an outstanding number of references represent less than 18% of the total.
- Part of the references made during a certain period (the so-called *citation window*) becomes the citations received by earlier published articles. These citation distributions are highly skewed: About 70% of all articles receive citations below the mean, and articles with a remarkable or outstanding number of citations represent about 9% of the total. The corresponding figures reported in Glänzel (2007) for papers published 20 years ago are 75 and 7%, respectively—a very small difference indeed, which speaks about the stability of these features of the citation process in a large number of fields and during a long period of time. At any rate, in our sample, this 9% of highly cited articles accounts for 44% of all citations received.
- The existence of a power law cannot be rejected in ALL SCIENCES taken together as well as in 17 of 22 fields whose articles represent 74.7% of the total. Contrary to the evidence in other contexts, the value of the scale parameter is above 3.5 in 13 of the 17 cases. Due to the prevalence of articles with no or few citations, power laws are typically small (representing 0.2–2.4% of all articles), but receive between 3.5 and 19.5% of all citations, with a maximum of 29.5% in Computer Science.

It can be concluded that what is needed is a single explanation of the decentralized process whereby scientists made references that a few years later translate into a highly skewed citation distribution crowned in most cases by a power law.

Future Research

It is natural to work at the aggregate level of the 22 scientific fields distinguished by TS. Quite apart from other alternatives at this level (see inter alia Adams et al., 1998; Glänzel & Schubert, 2003; Tijssen & van Leeuwen, 2003), it would be interesting to investigate these issues at the subfield level—a topic addressed in Schubert et al. (1987), where 114 subfields were analyzed, Glänzel (2007, in press), who studied 60 subfields, and Albarrán, Crespo, Ortuño, and Ruiz-Castillo (2010), who studied the 219 Web of Science categories within the 22 fields analyzed here.

The preliminary results obtained in this study constitute the most complete evidence available in the scientometrics literature about the prevalence of power laws among citation distributions arising from the academic periodicals indexed by TS (or other comparable journal collections). Three points are left for further research. First, as noted by Clauset et al. (2009), the fact that a power law cannot be rejected does not guarantee that a power law is the best distribution that fits the data. New tests must be applied confronting power laws with alternative distributions, such as the log-normal or the exponential distributions. Moreover, confidence intervals around the parameter estimates must be obtained.

[10] For the very different 17 phenomena for which a power law cannot be rejected in Clauset et al. (2009), in four cases the scale parameter is below 2, in eight cases between 2 and 3, and in five cases above 3.

[11] There are seven phenomena in Clauset et al. (2009) where the sample size is larger than 10,000 observations and a power law cannot be rejected. Ordered by sample size, these are solar flair intensity, count of word use, population of cities, Internet degree, papers authored, citations to papers from all sciences, and telephone calls received. In the last three phenomena, the size of the power law is less than 1% of the sample size; in two cases, this percentage is between 1 and 3%, and in the remaining three cases, this percentage is between 8 and 16%.

Second, the ML approach might be quite vulnerable to the existence of a few, but potentially influential extreme observations consisting of a small set of highly cited articles at the very end of the citation distribution. A possibility currently being investigated is that for a citation distribution following a power law, an estimation method that uses the relationship that has been shown to exist between the h-index for that sample, the sample size, and the scale parameter of the power law (Egghe & Rousseau, 2006; Glänzel, 2006). The rationale for this strategy lies in the fact that the h-index, of course, is robust to the presence of extreme observations.

Third, in the case of high-energy physics, Lehmann et al. (2003) estimated a second power law for the lower impact articles not included in the first one—a possibility that needs to be further explored.

Acknowledgments

The authors acknowledge financial support from Spanish MEC Grants SEJ2007–67436, SEJ2007–63098, and SEJ2006–05710. The database of Thomson Scientific (formerly Thomson-ISI; Institute for Scientific Information) was acquired with funds from Santander Universities Global Division of *Banco Santander*. This article is part of the SCIFI-GLOW Collaborative Project supported by the European Commission's 7th Research Framework Programme, Contract No. SSH7-CT-2008–217436. Comments by an anonymous reviewer helped to improve the original version of the article.

References

Adams, J., Bailey, T., Jackson, L., Scott, P., Pendlebury, D., & Small, H. (1997). Benchmarking of the international standing of research in England. Report of a consultancy study on bibliometric analysis [mimeo], University of Leeds, Leeds, United Kingdom.

Albarrán, P., Crespo, J., Ortuño, I., & Ruiz-Castillo, J. (2010). The skewness of science in 219 sub-fields, and a number of aggregates (Working Paper 10-16). Madrid, Spain: Universidad Carlos III.

Bauke, H. (2007). Parameter estimation for power-law distributions by maximum likelihood methods. The European Physical Journal B, 58, 167–173.

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

Clark, R.M., Cox, S.J.D., & Laslett, G.M. (1999). Generalizations of power-law distributions applicable to sampled fault-trace lengths: Model choice, parameter estimates, and caveats. Geophysical Journal International, 136, 357–372.

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

Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. Scientometrics, 69, 121–129.

Glänzel, W. (2006). On the h-index—A mathematical approach to a new measure of publication activity and citation impact. Scientometrics, 67, 315–321.

Glänzel, W. (2007). Characteristic scores and scales. Journal of Informetrics, 1, 92–102.

Glänzel, W. (in press). The application of characteristic scores and scales to the evaluation and ranking of scientific journals. Proceedings of INFO, Havana, Cuba.

Glänzel, W., & Schubert, A. (1988). Characteristic scores and scales in assessing citation impact. Journal of Information Science, 14, 123–127.

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

Goldstein, M.L., Morris, S.A., & Yen, G.G. (2004). Problems with fitting to the power-law distribution. European Physical Journal B, 41, 255.

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.

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, 1–8.

Lehmann, S., Lautrup, B., & Jackson, A.D. (2008). A quantitative analysis of indicators of scientific performance. Scientometrics, 76, 369–390.

Magyar, G. (1973). Bibliometric analysis of a new research sub-field. Journal of Documentation, 30, 32–40.

Mitzenmacher, M. (2004). A brief history of generative models for power law and lognormal distributions. Internet Mathematics, 1, 226–251.

Moed, H.F., De Bruin, R.E., & van Leeuwen, Th.N. (1995). New bibliometrics tools for the assessment of national research performance: Database description, overview of indicators, and first applications. Scientometrics, 33, 381–422.

Newman, M.E.J. (2005). Power laws, pareto distributions, and Zipf's law. Contemporary Physics, 46, 323–351.

Nicholls, P.T. (1987). Estimation of Zipf parameters. Journal of the American Society for Information Science, 38, 443–445 [also see Nicholls, P.T. (1988). Errata. Journal of the American Society of Information Science, 39, 287].

Persson, O., Glänzel, W., & Danell, R. (2004). Inflationary bibliometric values: The role of scientific collaboration and the need for relative indicators in evaluation studies. Scientometrics, 60, 421–432.

Pickering, G., Bull, J.M., & Sanderson, D.J. (1995). Sampling power-law distributions. Tectonophysics, 248, 1–20.

Price, D.J. de S. (1965). Networks of scientific papers. Science, 149, 510–515.

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, 58, 49–54.

Schubert, A., Glänzel, W., & Braun, T. (1987). A new methodology for ranking scientific institutions. Scientometrics, 12, 267–292.

Seglen, P. (1992). The skewness of science. Journal of the American Society for Information Science, 43, 628–638.

Tijssen, R.J.W., & van Leeuwen, T.N. (2003). Bibliometric analyses of world science. Extended Technical Annex to chapter 5 of the 3rd European Report on Science & Technology Indicators [mimeo], Leiden University, Leiden, The Netherlands.

Vinkler, P. (2009). The π-index: A new indicator for assessing scientific impact. Journal of Information Science, 35, 602–612.

White, E., Enquist, B., & Green, J. (2008). On estimating the exponent of power-law frequency distributions. Ecology, 89, 905–912.