Eur. Phys. J. B **4**, 131–134 (1998)

THE EUROPEAN PHYSICAL JOURNAL B

Rapid Note

How popular is your paper? An empirical study of the citation distribution

S. Redner[a]

Center for BioDynamics, Center for Polymer Studies, Boston University, Boston, MA, 02215, USA

Received: 12 May 1998 / Accepted: 12 May 1998

Abstract. Numerical data for the distribution of citations are examined for: (i) papers published in 1981 in journals which are catalogued by the Institute for Scientific Information (783,339 papers) and (ii) 20 years of publications in Physical Review D, vols. 11-50 (24,296 papers). A Zipf plot of the number of citations to a given paper *versus* its citation rank appears to be consistent with a power-law dependence for leading rank papers, with exponent close to $-1/2$. This, in turn, suggests that the number of papers with x citations, $N(x)$, has a large-x power law decay $N(x) \sim x^{-\alpha}$, with $\alpha \approx 3$.

PACS. 02.50.-r Probability theory, stochastic processes, and statistics – 01.75.+m Science and society – 89.90.+n Other areas of general interest to physicists

In this article, I consider a question which is of relevance to those for whom scientific publication is a primary means of scholarly communication. Namely, how often is a paper cited? While the average or total number of citations are often quoted anecdotally and tabulations of highly-cited papers exist [1,2], the focus of this work is on the more fundamental *distribution of citations*, namely, the number of papers which have been cited a total of x times, $N(x)$. In spite of the fact that many academics are obliged to document their citations for merit-based considerations, there have been only a few scientific investigations on quantifying citations or related measures of scientific productivity. In a 1957 study based on the publication record of the scientific research staff at Brookhaven National Laboratory, Shockley [3] claimed that the scientific publication rate is described by a log-normal distribution. Much more recently, Laherrere and Sornette [4] have presented numerical evidence, based on data of the 1120 most-cited physicists from 1981 through June 1997, that the citation distribution of individual authors has a stretched exponential form, $N(x) \propto \exp[-(x/x_0)^\beta]$ with $\beta \approx 0.3$. Both papers give qualitative justifications for their assertions which are based on plausible general principles; however, these arguments do not provide specific numerical predictions.

Here, the citation distribution of scientific publications based on two relatively large data sets is investigated [5]. One (ISI) is the citation distribution of 783,339 papers (with 6,716,198 citations) published in 1981 and cited between 1981 – June 1997 that have been cataloged by the Institute for Scientific Information. The second (PRD) is the citation distribution, as of June 1997, of the 24,296 papers cited at least once (with 351,872 citations) which were published in volumes 11 through 50 of Physical Review D, 1975–1994. Unlike reference [4], the focus here is on citations of publications rather than citations of specific authors. A primary reason for this emphasis is that the publication citation count reflects on the publication itself, while the author citation count reflects ancillary features, such as the total number of author publications, the quality of each of these publications, and co-author attributes. Additionally, only most-cited author data is currently available; this permits reconstruction of just the large-citation tail of the citation distribution.

The main result of this study is that the asymptotic tail of the citation distribution appears to be described by a power law, $N(x) \sim x^{-\alpha}$, with $\alpha \approx 3$. This conclusion is reached indirectly by means of a Zipf plot (to be defined and discussed below), however, because Figure 1 indicates that the citation distribution is not described by a single function over the whole range of x.

Since the distribution curves downward on a double logarithmic scale and upward on a semi-logarithmic scale (Figs. 1a and b respectively), a natural first hypothesis is that this distribution is a stretched exponential, $N(x) \propto \exp[-(x/x_0)^\beta]$. Visually, the numerical data fit this form fairly well for $x \leq 200$ (PRD) and $x \leq 500$ (ISI) as indicated in Figure 1b, with best fit values $\beta \approx 0.39$ (PRD) and $\beta \approx 0.44$ (ISI). However, the stretched exponential is unsuitable to describe the large-x data. Here,

[a] redner@sid.bu.edu

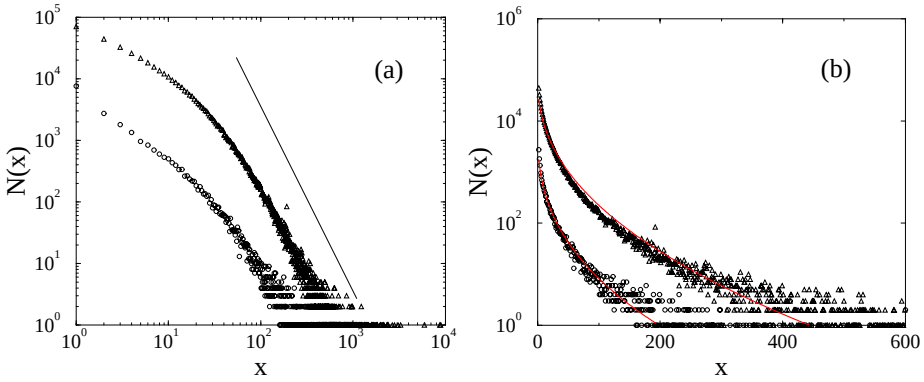

Fig. 1. (a) Citation distribution from the 783,339 papers in the ISI data set (△) and the 24,296 papers in the PRD data set (○) on a double logarithmic scale. For visual reference, a straight line of slope −3 is also shown. (b) Same as (a), except on a semi-logarithmic scale. The solid curves are the best fits to the data for $x \leq 200$ (PRD) and $x \leq 500$ (ISI).

data points are widely scattered, reflecting the paucity of well-cited papers. For example, in the ISI data, only 64 out of 783,339 papers are cited more than 1000 times, 282 papers are cited more than 500 times, and 2103 papers are cited more than 200 times, with the most-cited paper having 8907 citations. Such a sparsely populated tail is not amenable to being directly fit by a smooth function. (Amusingly (or soberingly) 633,391 articles in the ISI set are cited 10 times or less and 368,110 are uncited.)

Another test to determine the functional form of $N(x)$ is to compare numerical values for the moments of the citation distribution

$$\langle x^k \rangle = \frac{\int x^k N(x)\, dx}{\int N(x)\, dx}, \tag{1}$$

with those obtained by assuming a given form for $N(x)$. For example, if the citation distribution is a stretched exponential, then the dimensionless ratios $\mathcal{M}_k \equiv \langle x^k \rangle / \langle x \rangle^k = \Gamma\left(\frac{k+1}{\beta}\right) \Gamma\left(\frac{1}{\beta}\right)^{k-1} / \Gamma\left(\frac{2}{\beta}\right)^k$, where $\Gamma(x)$ is the gamma function. Notice that the scale factor x_0 in the exponential cancels. For each k, an estimate for β can be inferred by matching the value of $\mathcal{M}_k$ obtained from the above gamma function formula with the corresponding numerical data. For both the ISI and PRD data, the corresponding estimates for β for $k = 2, 3, \ldots, 6$ depend weakly but non-systematically on k, and further do not match the values for β obtained from a least-squares fit to a stretched exponential (Fig. 1b). Similarly, the numerical data for $\langle x^k \rangle$ also do not match a power-law form for the citation distribution, $N(x) \sim x^{-\alpha}$. These results provide evidence that the citation distribution is not described by a single function over the entire range of citation count.

More fundamentally, it is natural to expect different underlying mechanisms and different statistical features between minimally-cited and heavily-cited papers. The former are typically referenced by the author and close associates, and such papers are typically forgotten a short time after publication. Evidence for such a short lifetime of minimally-cited papers can be found, *e.g.*, by comparing the small-citation tail of $N(x)$ for the first 4 years (1975-79) and the last 4 years (1990-1994) of the PRD data set. For $x \lesssim 200$, these data (appropriately normalised) and the complete PRD data are virtually identical. On the other hand, well-cited papers become known through collective effects and their impact also extends over long time periods. This is reflected in the significant differences among the large-citation tails of $N(x)$ for papers of different eras.

To help expose these differences in the citation distribution, it is useful to construct a Zipf plot [6], in which the number of citations of the kth most-ranked paper out of an ensemble of M papers is plotted *versus* rank k (Fig. 2). By its very definition (see Eq. (2)), the Zipf plot is closely related to the cumulative large-x tail of the citation distribution. This plot is therefore well-suited for determining the large-x tail of the citation distribution. The integral nature of the Zipf plot also smooths the fluctuations in the high-citation tail and thus facilitates quantitative analysis.

Given an ensemble of M publications and the corresponding number of citations for each of these papers in rank order, $Y_1 \geq Y_2 \geq \ldots \geq Y_M$, then the number of citations of the kth most-cited paper, Y_k, may be estimated by the criterion [7]

$$\int_{Y_k}^{\infty} N(x)\, dx = k. \tag{2}$$

This specifies that there are k publications out of the ensemble of M which are cited at least Y_k times. Equation (2) also represents a one-to-one correspondence between the Zipf plot and the citation distribution. From the dependence of Y_k on k in a Zipf plot, one can test whether it accords with a hypothesised form for $N(x)$.

In Figure 2a, a Zipf plot of the rank-ordered citation data is presented on a double logarithmic scale for 4 data sets: (a) ISI data (top 200,000 papers only), (b) complete PRD data (24,296 papers), (c) first 4 years of PRD data, vols. 11-18 (5044 papers), and (d) last 4 years of PRD data, vols. 43-50 (5467 papers). As alluded to previously, there is a considerable difference between the first and last 4 years of the PRD data. As might be anticipated, the more recent highly-cited papers (up to approximately rank 700) are cited less than papers in the earlier subdata. (There are two exceptions, however. These are the two top papers in the first 4 years which are cited 1741 and 1294 times, while in the last 4 years of data the two leading papers are cited 2026 and 1420 times.) The larger citation count of heavily-cited older papers reflects the obvious fact that popular but recent PRD papers are still relatively