ELSEVIER

Physica A 298 (2001) 530–536

PHYSICA A

www.elsevier.com/locate/physa

Two-step competition process leads to quasi power-law income distributions Application to scientific publication and citation distributions

Anthony F.J. van Raan

Centre for Science and Technology Studies, Leiden University, Wassenaarseweg 52, P.O. Box 9555, 2300 RB Leiden, The Netherlands

Received 16 March 2001

Abstract

We apply a two-step competition process as a model to explain the distribution of citations ('income') over publications ('work'). The first step is the competition amongst scientists to get their work published in better journals, and the second to get this work cited in these journals. Generally, citation distributions are supposed to follow a power law, like most other 'income' distributions. So far, no satisfactory theoretical model of citation distribution has been developed. On the basis of two Boltzmann type distribution functions of source publications, we derive a distribution function of citing publications over source publications. This distribution function corresponds very well to the empirical data. It is not a power law, but a modified Bessel-function. In our view, the model presented in this article has a more generic value, particularly in economics to explain observed income distributions.

We developed a new model to explain the distribution of citations over publications. Bibliometric measurements of the distribution of citations over publications suggest a power-law function (see, for instance, Refs. [1,2]). But so far, no satisfactory theoretical model of citation distribution has been developed.

Our model consists of two steps. First, the competition amongst scientists for 'publication status'. This status is determined by the way the journal is cited by other journals. We argue that the underlying distribution originates from the journal in which a publication appears and it is operationalized in the form of an equilibrium distribution

E-mail address: vanraan@cwts.leidenuniv.nl (A.F.J. van Raan).

0378-4371/01/$ - see front matter
PII: S0378-4371(01)00254-0

of publications according to their 'status'. Second, within their status level, scientists again have to compete with their publications (i.e., with their 'work'), in terms of getting cited ('income'). On the basis of these two distribution laws, a third one results, the distribution of citations (i.e., citing publications) over source publications. A more detailed discussion and comparison of the model with further empirical findings based on bibliometric measurements is given elsewhere [3].

The basic concept of our model is the idea that scientific communication is characterized by a large number of publications that has to be divided according to attributed *status*. This concept is based on the following assumptions:

(1) The total system of scientific communication contains a limited amount of attributable status.
(2) The status of a publication is represented in a significant way by the status of the journal in which it is published.
(3) The status of a journal is operationalized significantly by the way it is cited by other journals ('bibliometric' operationalization).

Given these assumptions, it is possible to calculate the most probable distribution of publications over status levels.

The probability of any specific distribution is proportional to the number of ways this distribution can be realized. We now calculate this distribution following the lines of statistical mechanics, which will lead us to a Boltzmann distribution of publication numbers over journal status.

Say we have n levels $\boldsymbol{L}_1,\ldots,\boldsymbol{L}_n$ with an amount of status $\boldsymbol{W}_1,\ldots,\boldsymbol{W}_n$, and $\boldsymbol{N}$ publications. As indicated in our second assumption, status levels correspond to journals. If we start with the first level $\boldsymbol{L}_1$, there are $\boldsymbol{N}$ possibilities to choose the first publication. The second publication can be chosen in $\boldsymbol{N}-\boldsymbol{1}$ ways, the third in $\boldsymbol{N}-\boldsymbol{2}$ ways, and so on, up to $\boldsymbol{N}_1$ publications. The total number of possibilities is $\boldsymbol{N}!/\boldsymbol{N}_1\boldsymbol{!}(\boldsymbol{N}-\boldsymbol{N}_1)\boldsymbol{!}$. For the next status-level $\boldsymbol{L}_2$, we have $\boldsymbol{N}-\boldsymbol{N}_1$ publications available. We may continue this procedure until we have considered all status levels. The total probability for all status levels together is found by multiplication of the partial probabilities per level, which yields

$$\boldsymbol{P}=\boldsymbol{N}!/\boldsymbol{N}_1\boldsymbol{!}\boldsymbol{N}_2\boldsymbol{!}\boldsymbol{N}_3\boldsymbol{!}\ldots\ . \tag{1}$$

A more general model is given by the inclusion of an *a priori* probability for a status level. However, we expect differences in a priori probabilities only in exceptional cases, particularly for journals with very strict restrictions in their acceptance policy, such as *Nature and Science*. Thus, in good approximation we neglect the a priori probabilities and continue to use Eq. (1).

In order to find the *most probable* distribution, we have to identify the *maximum* value of $\boldsymbol{P}$. The most effective way to solve this problem is to calculate the maximum of $\ln \boldsymbol{P}$, instead of $\boldsymbol{P}$. From Eq. (1) it follows that

$$\ln \boldsymbol{P}=\ln \boldsymbol{N}!-\ln \boldsymbol{N}_1\boldsymbol{!}-\ln \boldsymbol{N}_2\boldsymbol{!}-\cdots\ .$$

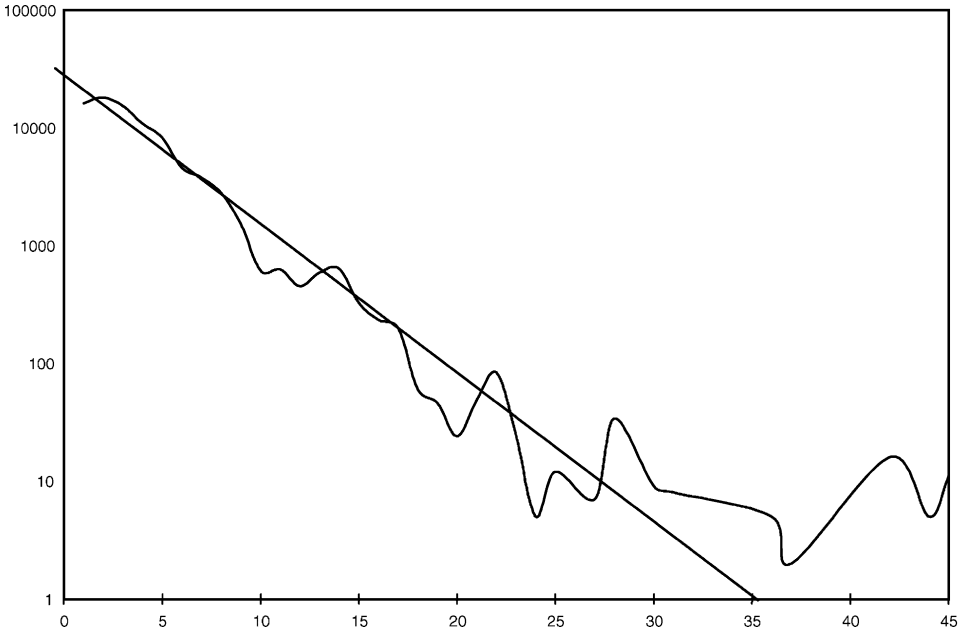

Fig. 1. Number of chemistry publications as a function of the ***JCSm*** values of the journals in which these publications have appeared. Publication years: 1985–1993. The ordinate indicates the number of publications, the abscissa indicates the ***JCSm*** values.

By using the Stirling formula $\ln \boldsymbol{x}! \sim \boldsymbol{x} \ln \boldsymbol{x} - \boldsymbol{x}$, and given that $\boldsymbol{N}_1 + \boldsymbol{N}_2 + \boldsymbol{N}_3 + \cdots = \boldsymbol{N}$, we find

$$\mathrm{d}(\ln \boldsymbol{P}) = -\sum_i \ln \boldsymbol{N}_i \, \mathrm{d}\boldsymbol{N}_i = 0, \tag{2}$$

taking into account that $\mathrm{d}\boldsymbol{N} = 0$, from which it follows that $\sum_i \mathrm{d}\boldsymbol{N}_i = 0$.

We solve Eq. (2) under conditions (1) $\sum_i \mathrm{d}\boldsymbol{N}_i = 0$ (as discussed above), and (2) the limited amount of status available in the total system: $\boldsymbol{W}_{\mathrm{tot}} = \boldsymbol{N}_1 \boldsymbol{W}_1 + \boldsymbol{N}_2 \boldsymbol{W}_2 + \cdots = \sum_i \boldsymbol{N}_i \boldsymbol{W}_i$, so that $\sum_i \boldsymbol{W}_i \, \mathrm{d}\boldsymbol{N}_i = 0$. Applying Lagrange's method, we multiply both condition-equations by an arbitrary constant, β and α, respectively, and add them to Eq. (2):

$$\sum_i (\ln \boldsymbol{N}_i + \beta + \alpha \boldsymbol{W}_i) \, \mathrm{d}\boldsymbol{N}_i = 0 \tag{3}$$

which yields

$$\boldsymbol{N}_i = \mathrm{e}^{-\beta - \alpha \boldsymbol{W}_i} = \boldsymbol{A} \mathrm{e}^{-\alpha \boldsymbol{W}_i}, \tag{4}$$

where $\boldsymbol{A}$ is a constant, which follows from Eq. (4).

In Fig. 1, we present the number of publications $\boldsymbol{N}$ of chemistry research in the Netherlands in a period of 8 years (in total about 15,000 publications) as a function of the ***JCSm*** values of the journals in which the publications have appeared. The ***JCSm*** value is the number of citations per publication of the journal in a specific period of

time (e.g., four years after publication). It is a 'bibliometric' operationalization of the journal status W_i as meant in our third assumption of the statistical model. For an ample discussion of ***JCSm*** we refer to Van Raan [4].

We clearly observe the exponential character of the function, which empirically supports the above model to explain the distribution of publication numbers over journal status, as given by Eq. (4). Only for very high ***JCSm***-values we observe a deviation, due to a very low number of publications in this region.

For the further mathematical development of our model, we rewrite the *distribution function* given in Eq. (4) as a *density function*:

$$\rho(W) = N\alpha \exp(-\alpha W) \quad \text{with} \int_0^\infty \rho(W)\,\mathrm{d}W = N. \tag{5}$$

We now suppose that the probability for publications to be cited *within a journal*, i.e., within a specific status-level L_i, is in fact the probability to occupy internal status levels with the same rules as discussed in Part 1. As for the first part of our statistical model, this assumption is clearly supported by empirical findings, see, for instance, Ref. [2].

Thus we find for this probability

$$p_i(c) = b_i \exp(-b_i c) \quad \text{with} \int_0^\infty p_i(c)\,\mathrm{d}c = 1\ . \tag{6}$$

With the help of Eq. (6), the average number of citations per publication $\langle c\rangle_i$ can be written as

$$W_i = \langle c\rangle_i = 1/b_i\ . \tag{7}$$

Given the empirical fact (Fig. 1) that our status parameter W can be considered in very good approximation as a continuous variable, we rewrite the probability function for the distribution of publications within a specific journal over the received citations c with the help of Eqs. (6) and (7):

$$p(c) = (1/W)\exp(-c/W)\ . \tag{8}$$

The probability that a publication in a given journal will receive a specific number of citations is given by

$$\rho(W,c) = \rho(W)p(c) = Na\exp(-\alpha W)\ (1/W)\exp(-c/W)\ . \tag{9}$$

Finally, we arrive at the distribution of all publications over citations

$$N(c) = \rho(c) = \int_0^\infty \rho(W,c)\,\mathrm{d}W = N\alpha \int_0^\infty \exp(-\alpha W - c/W)\ (1/W)\,\mathrm{d}W\ . \tag{10a}$$

The integral in Eq. (10a) is a *modified Bessel function of the zeroth order*, and thus we find

$$N(c) = 2N\alpha \mathbf{K}_0(2\sqrt{\alpha c})\ . \tag{10b}$$

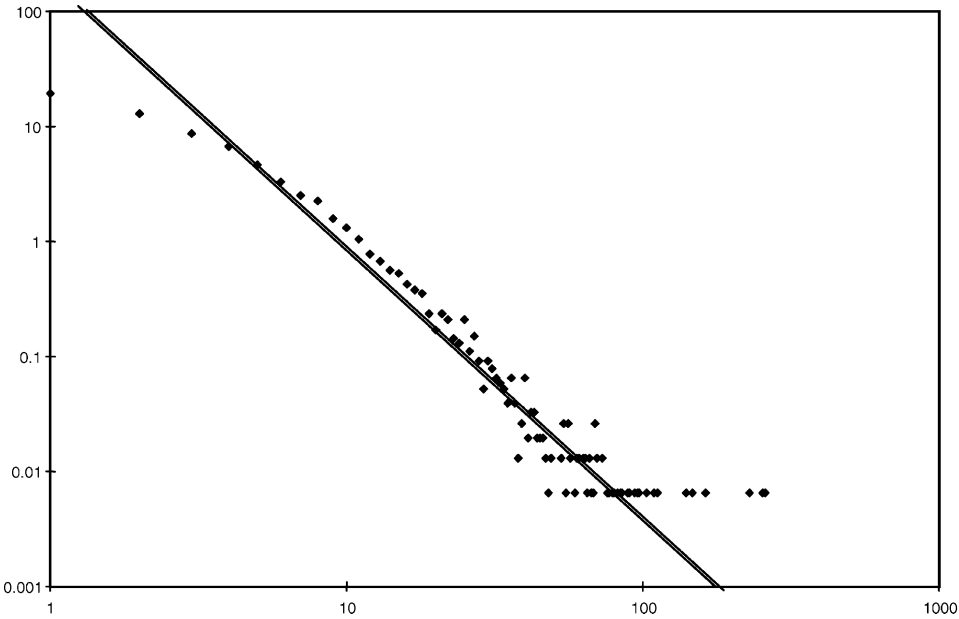

Fig. 2. Number of chemistry publications as a function of number of citations. Publication years: 1985–1993, citations are counted with 3-years 'window' after publication year, self-citations excluded. The ordinate indicates the relative number of publications, the abscissa indicates the absolute number of citations.

For the same set of publications and citations as presented in Fig. 1, we show in Fig. 2 the 'final' distribution: the number of publications as a function of the number of citations. The distribution suggests a power law relation, indicated by the straight line in the figure, particularly for the higher number of citations.

However, the empirical distribution function does *not* follow a power law for the lower number of citations. This is a serious problem, as most of the publications receive just a few citations. Our model solves this problem. The same distribution as in Fig. 2 is now shown in Fig. 3 (up to $\boldsymbol{c} = 30$), compared with the fitted modified Bessel function as given in Eq. (10b). We conclude that our model very well explains the empirical data. We find for the parameter α (Eq. (10b)) the value 0.32.

Even the value for zero citations is predicted excellently. This value cannot be represented in the log–log scales of Figs. 2 and 3. We find this value by the following argument. The number of citations is by definition an integer. Thus we deal with a discrete distribution, whereas the Bessel function holds for a continuous distribution. So we approximate the $\boldsymbol{c}$-values with the nearest integer, which means an integration of the Bessel function. For instance, the probability for zero citations is given by the integration of $\boldsymbol{N}(\boldsymbol{c})\,\mathrm{d}\boldsymbol{c}$ from $\boldsymbol{c} = 0$ to 0.5, i.e., the 'cumulative chance'

$$\int_0^{0.5} \boldsymbol{N}(\boldsymbol{c})\,\mathrm{d}\boldsymbol{c}\ .$$

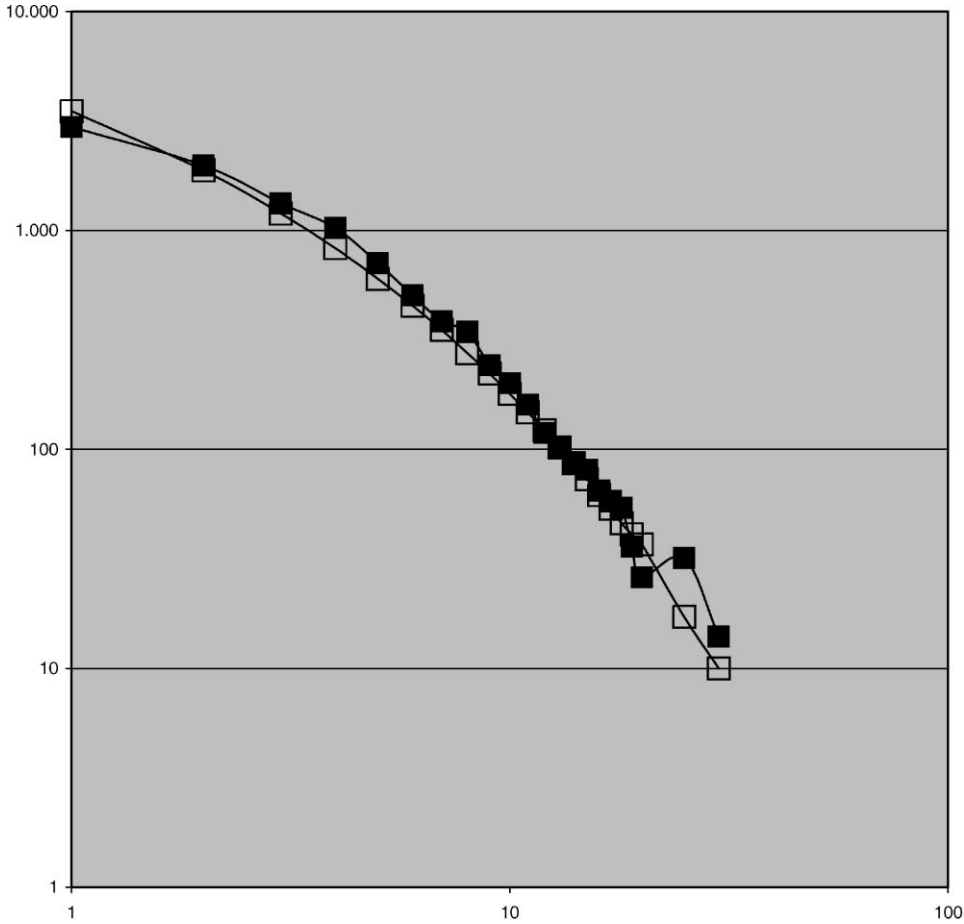

Fig. 3. Number of chemistry publications as a function of number of citations, empirical data compared with the fitted modified Bessel-function as given in Eq. (10b). Publication years: 1985–1993, citations are counted with 3-years 'window' after publication year, self-citations excluded. The ordinate indicates the absolute number of publications, the abscissa indicates the absolute number of citations. Black squares: empirical data, open squares: theoretical calculations.

With parameter $\alpha = 0.32$ as discussed above, this integration of the Bessel function yields 0.310, and the measured (relative) number is 0.292.

With the help of our two-step competition model, we have found that the distribution of citations over publications does not follow a power law, but is represented by a modified Bessel function. We find a very good agreement between the outcomes of our model and empirical data.

The citation distribution process can be seen as a specific representation of a more generic process of income distribution. Thus, our two-step competition model may be of interest for the understanding of complex social and economic phenomena. For instance, the income distribution may result from a process in which people first have

to compete (with education, talent, etc.) for occupations of different 'status' in society, and, secondly, within these occupations for their own position in terms of salary, revenues, etc. Thus, we wonder if the famous Pareto distributions are indeed power law distributions, or, according to a more generic form of our two-step competition model, a modified Bessel function.

Acknowledgements

I thank Prof. Carlo Beenakker for his crucial suggestions concerning the mathematics of this model and Thed van Leeuwen for his careful data-analytical work. This research was supported by the Netherlands Organization of Scientific Research (NWO) and Elsevier Science.

References

[1] S. Naranan, Power law relations in science bibliography: a self-consistent interpretation, J. Doc. 27 (1971) 83–97.
[2] P.O. Seglen, The skewness of science, J. Am. Soc. Inf. Sci. (JASIS) 43 (1992) 628–638.
[3] A.F.J. van Raan, Competition amongst scientists for publication status. Toward a model of scientific publication and citation distributions, Scientometrics 51 (2001) 347–357.
[4] A.F.J. van Raan, Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises, Scientometrics 36 (1996) 397–420.