Jointly published by Kluwer Academic Publishers, Dordrecht and Akadémiai Kiadó, Budapest

Scientometrics, Vol. 51, No. 1 (2001) 347–357

Competition amongst scientists for publication status: Toward a model of scientific publication and citation distributions

ANTHONY F. J. VAN RAAN

Centre for Science and Technology Studies, University of Leiden, Leiden (The Netherlands)

We present a model in which scientists compete with each other in order to acquire status for their publications in a two-step-process: first, to get their work published in better journals, and second, to get this work cited in these journals. On the basis of two Maxwell-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. In contrast to all observations so far, we conclude that this distribution of citations over publications, which is a crucial phenomenon in scientometrics, is not a power law, but a modified Bessel-function.

Introduction

Bibliometric measurements of the scientific communication system reveal many statistical characteristics at a higher aggregation level. One of the most striking phenomena is the distribution of citations over publications. The empirical findings suggest a power-law function (see for instance *Naranan* 1971, *Seglen* 1992).

We developed a model to explain the distribution of citations over publications. The model consists of two steps. In this model we introduce a new concept, the 'status' of publications. We argue that the underlying and most basic distribution law originates from an equilibrium distribution of publications according to their 'status'. This 'status' is determined by the journal in which a publication appears and it is operationalised by the way the journal is cited by other journals.

Second, within their status level, publications again have to compete for status, in terms of getting cited. On the basis of these two distribution laws, a third one results, the distribution of citations (i.e., citing publications) over source publications.

0138–9130/2001/US $ 15.00

The statistical model, Part 1: Publication distribution over status levels

We here discuss the concept 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 operationalised significantly by the way it is cited by other journals ('bibliometric' operationalisation).

Given these assumptions, it is possible to calculate the most probable distribution of publications over the above defined status levels. A most probable distribution means an equilibrium of the total system, and given the fact that journal status has a high stability, it is reasonable to conjecture that this most probable distribution is a major characteristic of the scientific communication system. Not only the number of publications is large but limited, the same is true for journals. So the total amount of status which has to be distributed among publications is also large but still limited.

The above concepts suggests a strong analogy with the statistical mechanics approach in thermodynamics (we follow *Alonzo* and *Finn*, 1974). 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 Maxwell-Boltzmann distribution of publication numbers over journal status.

Say we have n levels $\boldsymbol{L}_1,\ldots\ldots\boldsymbol{L}_n$ with an amount of status $\boldsymbol{W}_1,\ldots\ldots,\boldsymbol{W}_n$, 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 chose the first publication. The second publication can be chosen in $\boldsymbol{N-1}$ ways, and the third in $\boldsymbol{N-2}$ ways. The total number of possibilities is $\boldsymbol{N!/(N-3)!}$. As there are 3! = 6 different permutations of the three chosen publications, all resulting in the same distribution, the real total number of possibilities for the same distribution is $\boldsymbol{N!/3!(N-3)!}$.
Extending the number of publications from 3 to $\boldsymbol{N}_1$ yields

$$\boldsymbol{N!/\ N_1!(N-N_1)!} \qquad (1)$$

For the next status-level $\boldsymbol{L}_2$ we have $\boldsymbol{N-N}_1$ publications available. So if $\boldsymbol{N}_2$ publications have to populate $\boldsymbol{L}_2$, we can calculate this probability simply by replacing $\boldsymbol{N}$ by $\boldsymbol{N-N}_1$,

and N_1 by N_2 in Eq. 1. This yields $(N\text{-}N_1!)/N_2!(N\text{-}N_1\text{-}N_2)!$, and for the third status level in a similar way $(N\text{-}N_1\text{-}N_2!)/N_3!(N\text{-}N_1\text{-}N_2\text{-}N_3)!$. We may continue this procedure until we have considered all status levels. Now the total probability for all status-levels together is found by multiplication of the partial probabilities per level, which yields:

$$P = N!/N_1!N_2!N_3! \qquad (2)$$

In the above, we supposed that all status-levels have the same probability for publications to be 'occupied'. But this might be too simple: a more general model is given by the inclusion of an *a-priori* probability p_i for a status-level L_i. An example is the case of journals with very strict restrictions in their acceptance-policy, such as *Nature* and *Science*. So the probability to find one publication in L_i is p_i, two publications $p_i \times p_i = p_i^2$, and so on. Therefore we have to replace Eq. 2 by

$$P = N!\, p_1^{N_1} p_2^{N_2} p_3^{N_3} \ldots / N_1!N_2!N_3!\ldots \qquad (3)$$

with $\sum_i p_i = 1$

for the total probability of the distribution.

In order to find the most probable distribution, we have to identify the maximum-value of P. This means a situation whereby a small change of the number of publications in the different status-levels (i.e., dN_1, dN_2, dN_3,…) does not change the total probability P, i.e., $dP = 0$. The most effective way to solve this problem is to calculate the maximum of lnP, instead of P.

From Eq. 3 it follows that

$$lnP = N_1.lnp_1 + N_2.lnp_2 + \ldots - lnN_1! - lnN_2! - \ldots$$

By using the Stirling formula $lnx! \sim xlnx - x$, and given that $N_1 + N_2 + N_3 + \ldots = N$, we find

$$\ln P = N_1 \ln p_1 + N_2 \ln p_2 + .. - (N_1 \ln N_1 - N_1) - (N_2 \ln N_2 - N_2) .. =$$
$$- N_1 \ln(N_1/p_1) - N_2 \ln(N_2/p_2) - .. + (N_1 + N_2 + ..) = N \ln N - \sum_i N_i \ln(N_i/p_i).$$

Upon differentiating this expression we find $d(\ln P) = d\ln N - \sum_i \ln(N_i/p_i).dN_i - \sum_i dN_i$, and since $dN = 0$, it follows that $\sum_i dN_i = 0$. Thus, the maximum is determined by

$$d(\ln P) = - \sum_i \ln(N_i/p_i)\, dN_i = 0 \qquad (4)$$

We have to solve this equation under two conditions. The first is given above, $\sum_i dN_i = 0$. The second is the limited amount of status available in the total system, namely

$$W_{tot} = N_1W_1 + N_2W_2 + .. = \sum_i N_iW_i \text{, so that } \sum_i W_i dN_i = 0.$$

This conditional solution means that we have to apply Lagrange's method of undetermined multipliers: we multiply both condition-equations by an arbitrary constant, *b* and *a* respectively, and add them to Eq. 4. Thus we obtain:

$$\sum_i (\ln \boldsymbol{N}_i/\boldsymbol{p}_i + b + a\boldsymbol{W}_i).d\boldsymbol{N}_i = 0 \qquad (5)$$

and now we can solve this equation by taking all coefficients equal to zero, i.e.,

$$\ln \boldsymbol{N}_i/\boldsymbol{p}_i + b + a\boldsymbol{W}_i = 0$$

or

$$\boldsymbol{N}_i = \boldsymbol{p}_i \, e^{-b - a\boldsymbol{W}i}, \qquad (6)$$

which means that the number of publications is distributed over the status-levels in an exponential manner: there are many more publications in the lower status than in the higher status levels. We expect differences in a priori probabilities only in exceptional cases, such as discussed above. Thus, in good approximation, we may write

$$\boldsymbol{N}_i = \boldsymbol{A} \, e^{-a\boldsymbol{W}i}, \qquad (7)$$

where $\boldsymbol{A}$ is a constant, which follows from Eq. 6.

The attractiveness of Eq. 7 is that it immediately allows empirical verification. In Figure 1 we present as a striking example 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 $\boldsymbol{W}_i$ as meant in our third assumption of the statistical model. The ***JCSm*** is related to the 'impact factor' of a journal but it is defined differently, in order to cover a larger time-period for citations. For an ample discussion of ***JCSm*** we refer to work of our group, e.g., *Van Raan* (1996).

We clearly observe the exponential character of the function, which empirically supports the above Maxwell-Boltzmann model to explain the distribution of publication numbers over journal status, as given by Eq. 7. Only for very high ***JCSm***-values we observe a deviation, due to very low numbers of publications in this region.

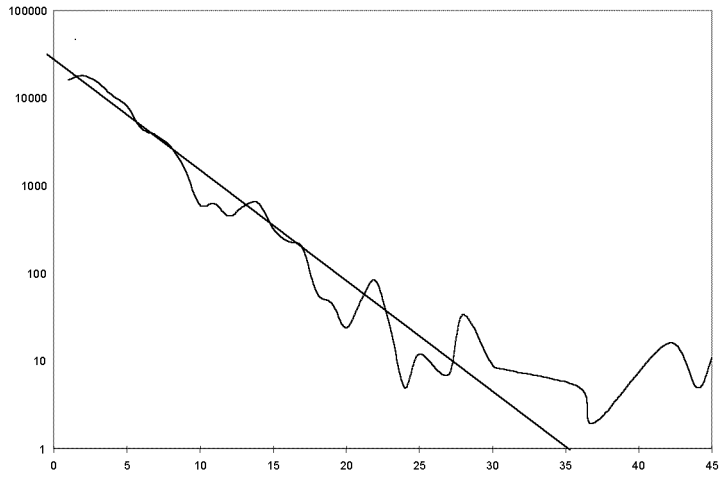

Figure 1. Number of chemistry publications from the Netherlands as a function of the ***JCSm*** values of the journals in which these publications have appeared. (Publication years: 1985-1993. Ordinate indicates the number of publications, the abscissa indicates the ***JCSm*** values.)

For the further mathematical development of our model, we rewrite the *distribution function* given in Eq. 7 as a *density function*:

$$\rho(\boldsymbol{W}) = \boldsymbol{N}\, a \exp(-a\boldsymbol{W}), \text{ with } \int_0^\infty \rho(\boldsymbol{W})\, \mathrm{d}\boldsymbol{W} = \boldsymbol{N}\,. \tag{8}$$

In the next section we will use this density function in order to arrive at a model for the distribution of citations over publications.

The statistical model, Part 2: Citation distribution over publications

We now suppose that the probability for publications to be cited *within a journal*, i.e., within a specific status-level $\boldsymbol{L}_i$, 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 *Seglen* (1992).

Thus we find for this probability

$$\boldsymbol{p}_i(\boldsymbol{c}) = \boldsymbol{b}_i.\exp(-\boldsymbol{b}_i\,\boldsymbol{c}), \quad \text{with} \int_0^\infty \boldsymbol{p}_i(\boldsymbol{c})\,d\boldsymbol{c} = 1\ . \tag{9}$$

Now we have a link between the distribution functions in Part 1 and in this part: the average number of citations per publication $< \boldsymbol{c} >_i$ can be written as

$$\boldsymbol{W}_i = < \boldsymbol{c} >_i = 1/\boldsymbol{b}_i\ , \tag{10}$$

where the relation with the parameter $\boldsymbol{b}_i$ follows from the integral equation given in Eq. 9.

Given the empirical fact (Figure1) that our status parameter $\boldsymbol{W}$ 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 $\boldsymbol{c}$ with help of Eqs 9 and 10:

$$\boldsymbol{p}(\boldsymbol{c}) = (1/\boldsymbol{W}).\exp(-\boldsymbol{c}/\boldsymbol{W})\ . \tag{11}$$

The results of both parts of our statistical model so far are represented by Eqs 8 and 11. This means that we created an abstract publication space with two dimensions: status of journals $\boldsymbol{W}$ (represented by the ordinate of the space) and $\boldsymbol{c}$ (represented by the abscissa of the space). This approach allows us to find the probability that a publication in a given journal will receive a specific number of citations:

$$\rho(\boldsymbol{W}, \boldsymbol{c}) = \rho(\boldsymbol{W}).\,\boldsymbol{p}(\boldsymbol{c}) = \boldsymbol{N}\,a\exp(-a\boldsymbol{W})\,(1/\boldsymbol{W})\exp(-\boldsymbol{c}/\boldsymbol{W})\ . \tag{12}$$

With help of this function we derive for the distribution of citations over journals

$$\boldsymbol{c}(\boldsymbol{W}) = \int_0^\infty \rho(\boldsymbol{W}, \boldsymbol{c})\,.\,\boldsymbol{c}\,d\boldsymbol{c} = \boldsymbol{N}\,a \int_0^\infty \exp(-a\boldsymbol{W} - \boldsymbol{c}/\boldsymbol{W})\,(1/\boldsymbol{W})\,\boldsymbol{c}\,d\boldsymbol{c} = \tag{13}$$

$$\boldsymbol{N}\,a\,\boldsymbol{W}\exp(-a\boldsymbol{W})\ .$$

Again, we have clear empirical support for this particular outcome of our model.

Figure 2 shows the distribution of citations received by the chemistry publications mentioned earlier as a function of the ***JCSm***-values of the journals in which the cited publications have appeared. As in the case of Figure1, we see a deviation for the very high ***JCSm***-values due to the relatively low numbers of cited publications in this region. We notice the difference with Figure 1 due to the linear part of the distribution.

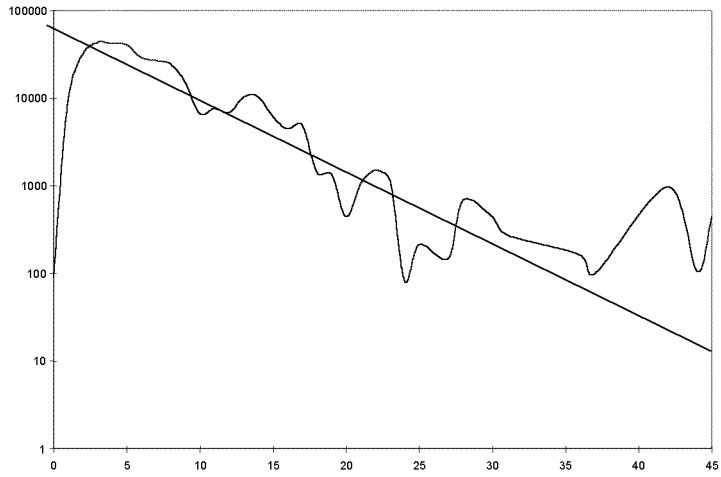

Figure 2. Citations received by chemistry publications from the Netherlands as a function of the ***JCSm*** values of the journals in which the cited publications have appeared. (Publication years: 1985-1993, citations are counted with 3-years 'window' after publication year, self-citations excluded. Ordinate indicates the number of citations, the abscissa indicates the ***JCSm*** values.)

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

$$N(c) = \rho(c) = \int_0^\infty \rho(W, c)\, dW = N a \int_0^\infty \exp(-aW - c/W)\,(1/W)\, dW \quad . \tag{14}$$

The integral in Eq. 14 is a *modified Bessel-function of the 0-th order*, and thus we find

$$N(c) = 2 N a \mathbf{K_0}(2\sqrt{ac}) \, . \tag{15}$$

Let us look at the empirical findings. For the same set of publications and citations as presented in Figures1 and 2, we show in Figure 3 the 'final' distribution: the number of publications as a functions of the number of citations. The distribution suggests a power-law relation, indicated by the straight line in the figure:

$N(c) = c^{-\beta}$ (where β is a parameter to be determined empirically)

particularly for the higher numbers of citations. Often it is more or less taken for granted that scientometric distributions 'must' have a power-law character as scientometric phenomena are considered to be comparable to economic phenomena, such as the Pareto income distributions, which are also considered to be power laws.

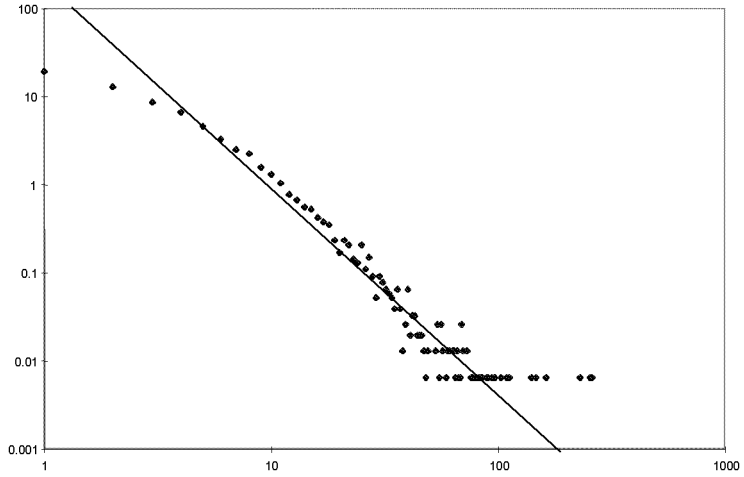

Figure 3. Citations received by chemistry publications from the Netherlands (Publication years: 1985-1993, citations are counted with 3-years 'window' after publication year, self-citations excluded. Ordinate indicates the relative number of publications, the abscissa indicates the absolute number of citations.)

However, the empirical fact that the distribution function does *not* follow a power law for the lower number of citations, is a serious problem, as most of the publications received just a few citations. Our model solves this problem. The same distribution as in Figure 3 is now shown in Figure 4 (up to $\boldsymbol{c} = 30$), compared with the fitted modified Bessel-function as given in Eq. 15. We conclude that our model explains very well the empirical data. We find for the parameter a (Eq.15) the value 0.32.

Even the value for zero citations is predicted excellently. This value cannot be represented in the log-log scale of Figure 4. 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 in the Bessel function with the nearest integer, which means 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'

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

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

We are currently investigating a larger group of data-sets to identify field- and organization-specific characteristics of this parameter *a*, and how it varies as a function of time.

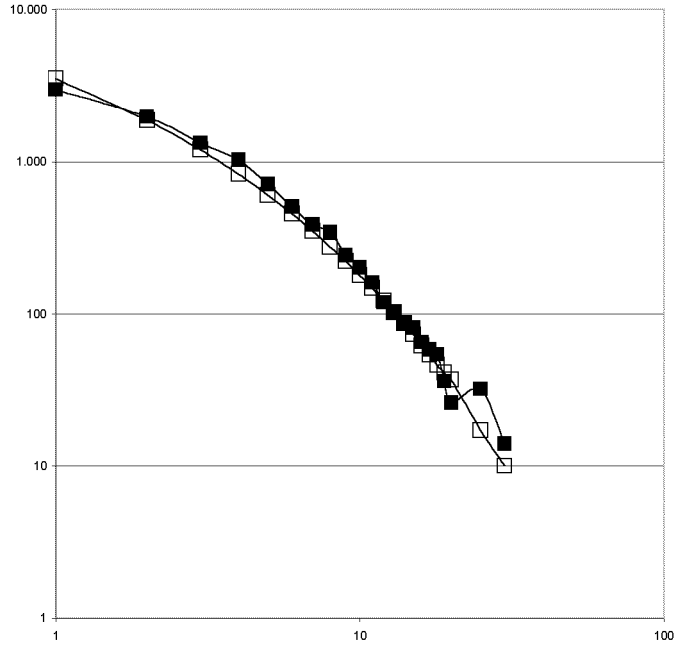

Figure 4. Absolute number of publications as a function of number of citations, empirical data compared with the fitted modified Bessel-function as given in Eq. 15. (The empirical data are the same as in Figure 3, but now we have absolute numbers of publications on the ordinate; the abscissa indicates the absolute number of citations. Black squares: empirical data , open squares: theoretical calculations.

Conclusions

In this paper we have studied a model in which scientists are competing for publication status by trying to get their work published in the best journals. From this model, we have derived a statistical approach for the distribution of citations over publications. Our work is inspired by methods in statistical physics. 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 good agreement between the outcomes of our model and empirical data. Current research is now being focused on the identification of field-, organization- and time-specific characteristics of the Bessel function parameter *a*.

What could be the meaning of the discovered distribution function? Are there any other 'natural' processes governed by the same distribution functions as found in the application of our status competition model? An interesting association are diffusion processes in physics (neutron diffusion in absorbing material) and, particularly, biology. *Pielou* (1969) describes the outward diffusion of larvae from a compact cluster of eggs laid by an insect. The spatial patterns resulting from this diffusion process (relationship between number of individual organisms and distance from the point when organisms started to diffuse outward) is described by the same modified Bessel function as found in our model. This analogy would imply the picture of publications diffusing into an 'impact-space' with number of citations as a simple metric.

We hope, given the high current level of interest in 'power-law-like' distribution functions in complex social and economic phenomena, that this work will shed new light on other types of competition-based 'income' distributions and will stimulate follow-up research also in other fields of science.

*

I would particularly like to thank my Leiden colleague Professor Carlo *Beenakker* for his crucial suggestions concerning the mathematics of this model. I am also thankful to Thed *van Leeuwen* at our institute for his careful data-analytical work.

The research reported in this paper was made possible in part by grants of the Netherlands Organization of Scientific Research (NWO) and also by the freedom we have in our contract research with Elsevier Science to devote a nice amount of the resources to strange, exotic and at first sight completely useless research.

References

ALONSO, M., E. J. FINN (1974), *Fundamental University Physics. Vol. I. Mechanics and Thermodynamics*, Addison-Wesley.

NARANAN, S. (1971), Power law relations in science bibliography: A self-consistent interpretation, *Journal of Documentation*, 27:83–97.

PIELOU, E. C. (1969), An Introduction to Mathematical Ecology, New York, John Wiley & Sons.

RAAN, A. F. J. VAN (1996), Advanced bibliometric methods as quantitative core of peer review based evaluation and foresight exercises, *Scientometrics*, 36:397–420.

SEGLEN, P. O. (1992), The skewness of science, *Journal of the American Society for Information Science*, 43:628-638.

Received February 16, 2001.

Address for correspondence:
ANTHONY F. J. VAN RAAN
Centre for Science and Technology Studies,
University of Leiden, Wassenaarseweg 52, P.O. Box 9555
2300 RB Leiden, The Netherlands
E-mail:vanraan@cwts.leidenuniv.nl