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. © 2001 Elsevier Science B.V. All rights reserved.

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).

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 $\boldsymbol{L}_1,\dots,\boldsymbol{L}_n$ with an amount of status $\boldsymbol{W}_1,\dots,\boldsymbol{W}_n$, 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 $\boldsymbol{N}-\mathbf{1}$ ways, the third in $\boldsymbol{N}-\mathbf{2}$ ways, and so on, up to $\boldsymbol{N}_1$ publications. The total number of possibilities is $\boldsymbol{N}!/\boldsymbol{N}_1!(\boldsymbol{N}-\boldsymbol{N}_1)!$. 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

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

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

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