Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics
DOI: 10.1007/s11192-008-2175-8

Paretian publication patterns imply Paretian Hirsch index

KRISZTINA BARCZA, ANDRÁS TELCS

Department of Computer Science, Faculty of Electric Engineering,
Budapest University of Technology and Economics, Goldman tér 3., 1111 Budapest, Hungary

The paper pursues the rigorous mathematical study of the Hirsch index and shows that it has power law upper tail distribution and determines the exponent provided that the underlying publication and citation distributions have fat tails as well. The result is demonstrated on the distribution of the Hirsch index of journals. The paper is concluded with some further remarks on the Hirsch index.

Introduction

The Hirsch index [HIRSCH, 2005], *h*, has occupied its place as an indicator of scientific excellence among other standard scientific indicators like publication and citation score, impact factor or relative impact factor. Rigorous investigation of the Hirsch index revealed some fundamental statistical properties. GLÄNZEL [2006] showed that the expected Hirsch index can be expressed by the number of underlying publication score, *n*, governed by a Paretian distribution and the Paretian exponent of the citation distribution β of the given field:

$$h \cong cn^{\frac{1}{\beta+1}}.$$ [1]

BEIRLANT & EINMAHL [2007] deduced a nonstandard asymptotic of the index. The nice result shows that the population size dependent observed index, $\hat{H}$, is normally distributed about the theoretical value, *h*:

$$\frac{1+nf(h)}{\sqrt{h}}(\hat{H}-h)\xrightarrow{d}N(0,1), \tag{1}$$

where *f*(*k*) is the citation distribution. Several papers deal with the practical aspects of the Hirsch index. Rescaling with respect to scientific fields discussed by IGLESIAS & PECHARROMÁN [2007], Hirsch index of journals introduced by BRAUN & AL. [2006] and SCHUBERT & GLÄNZEL [2007] just to mention some of the studies.

[1] Let $a_n \approx b_n$mean that $a_n/b_n \to 1$ as $n\to\infty$ and $a_n \cong b_n$ that there is a $C>1$ such that $1/C \le a_n/b_n \le C$ for all n

Received October 22, 2008

Address for correspondence:
ANDRÁS TELCS
E-mail: telcs@szit.bme.hu

0138–9130/US $ 20.00

The present paper pursues the rigorous mathematical study of the Hirsch index and shows that it has a power law upper tail distribution and determines the exponent provided that the underlying publication and citation distributions have fat tails as well. Based on a wide dataset it is shown that the journal h-index asymptotically fat tailed and the exponent of the tail close to the value suggested by the theory. The paper is concluded with some further investigation of the information which can be extracted from the statistics (*n*,*h*); the conditional distribution of the most cited papers is derived and the implied Hirsch exponent is introduced.

The tail behavior

Let x be a randomly chosen author of the scientific community under scrutiny, $n=n(x)$ is the number of his/her papers (in the whole life span or in a defined period). Denote y_i for $i=1,..,n$ the individual papers and $cit(y_i)$ their citation score (ordered decreasing in i), i.e., $cit(y_1)\geq cit(y_2)\geq cit(y_n)$.

$h(x)$ is the Hirsch index of x:

$$h(x)=\max\{k\ :\ cit(y_k)\geq k\}.$$

Assume that

$$G_l^P=\mathbf{P}(n(x)\geq l)\approx cl^{-\alpha}$$

is the tail distribution of the productivity for some $\alpha>1$, $F_l=1-G_l$ and

$$G_l^Q=\mathbf{P}(cit(y)\geq l)\approx cl^{-\beta}$$
$$h_k=\mathbf{P}(h(x)=k)$$

the corresponding notions for the citation score and h-index.[2] In addition we assume that all the publication and citation scores are independent. The main result of the paper is

$$\boxed{\mathbf{P}(h\geq k)\cong k^{-\alpha(\beta+1)}.} \qquad (2)$$

The detailed proof is deferred to the appendix.

[2] Here and in the sequel $c>0$ is an arbitrary constant which may change from place to place.

The journal case study

Since there is no free access to entire databases of author's publication and citation, the available alternative, the Scimago database [SCIMAGO GROUP] of journal scores has been used. Scimago provides a quick and easily accessible performance evaluation of scientific journals including the Hirsch index. On the downloaded data simple statistics have been performed; the tail exponent of the distribution of published papers, citation and Hirsch index extracted. The log-log plot of the distribution of the *h* index is presented in Figure 1.

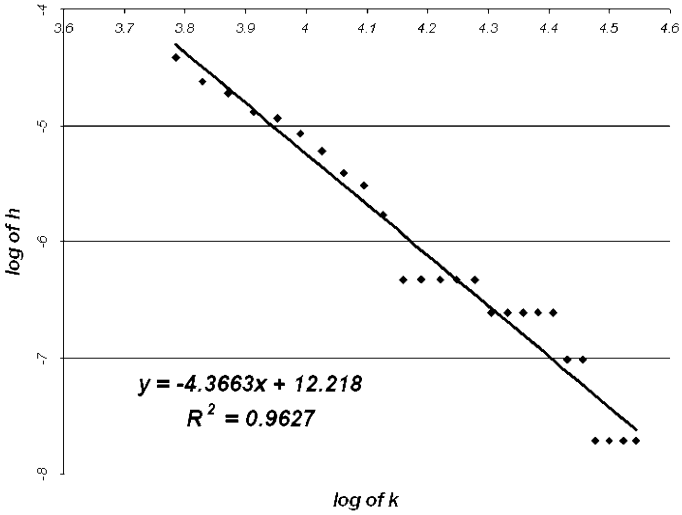

Figure 1. The log-log plot of the distribution of the h-index

For the tail exponent estimate data series truncated from below by 23 and the zero entries omitted from above. The slope of the fitted line provides the estimate of the exponent $\hat{\eta}$ while the expected value based on α=1.82 and β=1.45 is $\eta=\alpha(\beta+1)=4.46$..

Similar estimates are summarized in Table 1. η stands for $\alpha(\beta+1)$ and $\hat{\eta}$ for the exponent provided by the linear fit in the log-log plot of the distribution of *h* .

Table 1. Exponents of all combined and some particular fields

Field	α	β	η	$\hat{\eta}$
All	1.82	1.45	4.46	4.36
Mathematics	1.01	1.40	2.42	2.40
Medicine	2.33	1.48	4.53	4.41
Physics	1.53	1.07	2.70	2.68

Discussion

The citation distribution within the h-core

Given that the citation score has a fat tailed distribution, the h-index provides an other aspect to evaluate the performance. Let us study the citation distribution of the papers belonging to the h-core, i.e., the papers cited more than or equal to the h-index. If the paper y is in the h-core (assuming $\hat{H} = k$) then the truncated distribution (by paper) of the citation score follows a fat tailed distribution again: for $l>k>0$

$$\begin{aligned} &\mathbf{P}(cit(y) \geq l \mid y \text{ is in the h-score and } \hat{H} = k) \\ &= \mathbf{P}(cit(y) \geq l \mid cit(y) \geq k) \\ &= \frac{\mathbf{P}(cit(y) \geq l \text{ and } cit(y) \geq k)}{\mathbf{P}(cit(y) \geq k)} \\ &= \frac{\mathbf{P}(cit(y) \geq l)}{\mathbf{P}(cit(y) \geq k)} \approx c\left(\frac{l}{k}\right)^{-\beta}. \end{aligned} \tag{3}$$

This truncated distribution is a good base of evaluation of the h-core and partially eliminates the valid objection that the Hirsch-index can not differentiate between different citation patterns. As an example consider that $\hat{H} = 4$, and the citation profiles of the sores are: Case 1: (6,6,4,4), Case 2: (41,27,19,14). If we plot $c(l/k)^{-\beta}$ and the actual observed citations for the h-core it is easy to recognize the differences.

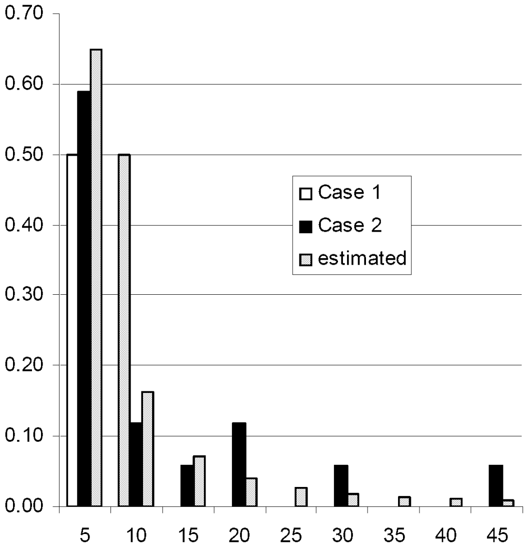

Figure 2. Comparison of h-cores and the theoretical distribution

The implied exponent, $\tilde{\eta}$

Finally the implied citation distribution exponent provides a concise information on the performance of individuals. We can start again with the observation

$$h \approx cn^{\frac{1}{\eta+1}}.$$

A tail exponent estimate can be built on this relation:

$$\tilde{\eta} = \frac{\ln n}{\ln h} - 1$$

It is clear that the higher the h-index, the lower the implied exponent $\tilde{\eta}$ and as a consequence one can expect that the h-core has fatter tail.

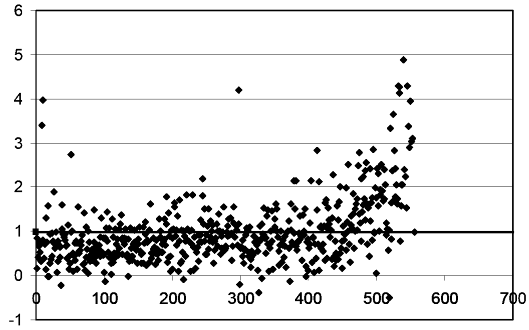

Figure 3. The implied exponents

In Figure 3 the individual implied exponent of journals is depicted. The solid line represents the average η=0.98 value of mathematical journals. In extreme cases the implied exponent might be negative indicating very high h-index values.

The iterative Hirsch index

PRATHAP [2006] and SCHUBERT [2007] proposed the notion of iterative Hirsch index. In the same spirit let us denote h_k by $h_k^{(1)}$ and define $h_k^{(2)}$, the Hirsch index of second order of pools of authors (like departments, institutions or countries). Each pool p is a set of scientists $\{x_1, x_2, ..., x_m\}$ and they are arranged in decreasing order of $h^{(1)}(x_i)$

$$h^{(2)}(p) = \max\{i \; : \; h^{(1)}(x_i) \geq i\}$$

If the pool size distribution has a heavy tail with exponent γ one can deduce along the lines of the proof of the main result that $h^{(2)}$ has a fat tailed distribution again and the exponent is $\gamma(\alpha(\beta+1)+1)$. It is clear that one can continue the iteration until $h^{(j)}$ has any meaning.

*

This work is motivated by a number of delightful discussions with A. Schubert. These discussions started a long-long time ago and covered a very broad area including scientometrics, network theory and mathematics. None of us can tell which was the initial point of the present work, but I am very grateful for all of them. The author thanks for the valuable comments and suggestions of the anonymous referees.

References

BEIRLANT, J., EINMAHL, J. J. H. (2007), Asymptotics for the Hirsch index, *CentER Discussion Paper Series*, No. 2007-86.

BRAUN, T., GLÄNZEL, W., SCHUBERT, A. (2006), A Hirsch-type index for journals, *Scientometrics*, 69 (1) : 169–173.

HIRSCH, J. E. (2005), An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the United States of America*, 102 (46) : 16569–16572.

GLÄNZEL, W. (2006), On the h-index – A mathematical approach to a new measure of publication activity and citation impact. *Scientometrics*, 67 (2) : 315–321.

IGLESIAS, J.E, PECHARROMÁN, C. (2007), Scaling the h-index for different scientific ISI fields, *Scientometrics*, 73(3) : 303–320.

KRAPIVSKY, P. L., REDNER, S., LEYVRAZ, F. (2000), *Phys Rev Lett.*, 85 (21) : 4629–4632.

PRATHAP, G. (2006), Hirsch-type indices for ranking institutions scientific research output, *Current Science*, 91(11) : 10.

SCHUBERT, A. (2007), Successive h-indices, *Scientometrics*, 70 (1) : 183–200.

SCHUBERT, A. GLÄNZEL, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3) : 179–184.

SCIMAGO GROUP, http://www.scimagojr.com/journalrank.php

Appendix

In what follows we provide rigorous derivation of (2). Assume that k≥1.

$$\begin{aligned} h_k &= \sum_{l=0}^{\infty} P\big(h(x)=k, n(x)=k+l\big) \\ &\approx c\left(G_k^Q\right)^k \sum_{l=0}^{\infty} ck+l^{-(\alpha+1)} \binom{k+l}{l} \left(F_{k+1}^Q\right)^l. \end{aligned}$$

First we develop a lower estimate for h_k.

$$\begin{aligned} h_k &\approx c\left(G_k^Q\right)^k \sum_{l=0}^{\infty} (k+l)^{-(\alpha+1)} \binom{k+l}{l} \left(F_{k+1}^Q\right)^l \\ &= cd^k k^{-\beta k} \sum_{l=0}^{\infty} (k+l)^{-(\alpha+1)} \binom{k+l}{l} \left(1-d(k+1)^{-\beta}\right)^l \end{aligned}$$

Let us treat the last term first.

$$e^{-d(k+1)^{-\beta}} \approx 1 - d(k+1)^{-\beta}.$$

We estimate the binomial coefficient as follows:

$$\binom{k+l}{l} = \frac{(k+l)\ldots(l+1)}{k!} \geq \frac{l^k}{k!}.$$

$$\binom{k+l}{l} = \frac{(k+l)\ldots(l+1)}{k!} \leq \frac{(k+l)^k}{k!}.$$

$$\begin{aligned}
h_k &\approx c\left(G_k^Q\right)^k \sum_{l=0}^{\infty} (k+l)^{-(\alpha+1)} \binom{k+l}{l} \left(F_{k+1}^Q\right)^l \geq c\frac{d^k k^{-\beta k}}{k!} \sum_{l=0}^{\infty} l^{k-(\alpha+1)} e^{-\frac{dl}{k^\beta}} \\
&\geq c\frac{d^k k^{-\beta k}}{k!} \int_0^\infty x^{k-(\alpha+1)} e^{-\frac{dx}{k^\beta}} dx \\
&\geq cd^k k^{-\beta k} \frac{\Gamma(k-(\alpha+1)+1)}{k!} \left(\frac{k^\beta}{d}\right)^{k-(\alpha+1)+1} \\
&= ck^{-\alpha\beta} \frac{\Gamma(k-\alpha)}{k!} \geq ck^{-\alpha\beta} \frac{(k-\alpha)^{k-\alpha-1/2} k^{-\alpha} e^{k-\alpha+\Theta/12(k-\alpha)}}{k^{k+1/2} e^{k+\Theta/(12k+1)}} \\
&\approx ck^{-\alpha(\beta+1)} k^{-1} = ck^{-\alpha(\beta+1)-1},
\end{aligned}$$

where 0<Θ<1. The upper estimate works similarly:

$$\begin{aligned}
h_k &= \left(G_k^Q\right)^k \sum_{l=0}^{\infty} c(k+l)^{-(\alpha+1)} \binom{k+l}{l} \left(F_{k+1}^Q\right)^l \\
&\leq c\frac{k^{-\beta k} d^k}{k!} \sum_{l=0}^{\infty} (k+l)^{-(\alpha+1)} (k+l)^k e^{-\frac{dl}{(k+1)^\beta}} \\
&\leq c\frac{k^{-\beta k} d^k}{k!} \sum_{i=1}^{\infty} i^{k-(\alpha+1)} e^{-\frac{di}{(k+1)^\beta}} \leq c\frac{k^{-\beta k} d^k}{k!} \int_0^\infty x^{k-(\alpha+1)} e^{-\frac{dx}{(k+1)^\beta}} dx \\
&= ck^{-\beta k} d^k \left(\frac{(k+1)^\beta}{d}\right)^{k-\alpha} \frac{\Gamma(k-\alpha)}{k!} \\
&\approx ck^{-\alpha\beta} \left(1+\frac{1}{k}\right)^{-k\beta} \frac{(k-\alpha)^{k-\alpha-1/2} e^{-k+\alpha}}{k^{k+1/2} e^{-k}} \\
&\approx ck^{-\alpha\beta} \left(\frac{k+1}{k}\right)^{-\alpha\beta} \frac{(k-\alpha)^{k-\alpha-1/2} e^{-k+\alpha}}{k^{k+1/2} e^{-k}} \\
&\approx ck^{-\alpha(\beta+1)-1}
\end{aligned}$$