CITATION ENTROPY AND RESEARCH IMPACT ESTIMATION

Z.K. Silagadze

Budker Institute of Nuclear Physics
and
Novosibirsk State University, 630 090, Novosibirsk, Russia

(Received November 10, 2009; revised version received July 20, 2010)

A new indicator, a real valued s-index, is suggested to characterize a quality and impact of the scientific research output. It is expected to be at least as useful as the notorious h-index, at the same time avoiding some of its obvious drawbacks. However, surprisingly, the h-index is found to be quite a good indicator for majority of real-life citation data with their alleged Zipfian behaviour for which these drawbacks do not show up. The style of the paper was chosen deliberately somewhat frivolous to indicate that any attempt to characterize the scientific output of a researcher by just one number always has an element of a grotesque game in it and should not be taken too seriously. I hope this frivolous style will be perceived as a funny decoration only.

PACS numbers: 01.30.–y, 01.85.+f

Sound, sound your trumpets and beat your drums! Here it is, an impossible thing performed: a single integer number characterizes both productivity and quality of a scientific research output. Suggested by Hirsch [1], this simple and intuitively appealing h-index has shaken academia like a storm, generating a huge public interest and a number of discussions and generalizations [2–12].

A Russian physicist with whom I was acquainted long ago used to say that the academia is not a Christian environment. It is a pagan one, with its hero-worship tradition. But hero-worshiping requires ranking. And a simple indicator, as simple as to be understandable even by dummies, is an ideal instrument for such a ranking.

h-index is defined as given by the highest number of papers, h, which has received h or more citations. Empirically

$$h \approx \sqrt{\frac{C_{\rm tot}}{a}} \qquad (1)$$

with a ranging between three and five [1]. Here $C_{\rm tot}$ stands for the total number of citations.

And now, with this simple and adorable instrument of ranking on the pedestal, I am going into a risky business to suggest an alternative to it. Am I reckless? Not quite. I know a magic word which should impress pagans with an irresistible witchery.

Shannon introduced the quantity

$$S = -\sum_{i=1}^{N} p_i \ln p_i \qquad (2)$$

which is a measure of information uncertainty and plays a central role in information theory [13]. On the advice of John Von Neumann, Shannon called it entropy. According to Feynman [14], Von Neumann declared to Shannon that this magic word would give him "a great edge in debates because nobody really knows what entropy is anyway".

Armed with this magic word, entropy, we have some chance to overthrow the present idol. So, let us try it! Citation entropy is naturally defined by (2), with

$$p_i = \frac{C_i}{C_{\rm tot}}\,,$$

where C_i is the number of citations on the ith paper of the citation record. Now, in analogy with (1), we can define the citation record strength index, or s-index, as follows

$$s = \tfrac{1}{4}\sqrt{C_{\rm tot}}\; e^{S/S_0}\,, \qquad (3)$$

where

$$S_0 = \ln N$$

is the maximum possible entropy for a citation record with N papers in total, corresponding to the uniform citation record with $p_i = 1/N$.

Note that (3) can be rewritten as follows

$$s = \tfrac{1}{4}\sqrt{C_{\rm tot}}\; e^{(1-S_{\rm KL}/S_0)} \approx \tfrac{2}{3}\sqrt{C_{\rm tot}}\; e^{-S_{\rm KL}/S_0}, \qquad (4)$$

where

$$S_{\rm KL} = \sum_{i=1}^{N} p_i \ln\frac{p_i}{q_i}\,, \qquad q_i = \frac{1}{N} \qquad (5)$$

is the so-called Kullback–Leibler relative entropy [15], widely used concept in information theory [16–18]. For our case, it measures the difference between the probability distribution p_i and the uniform distribution $q_i = 1/N$. The Kullback–Leibler relative entropy is always a non-negative number and vanishes only if p_i and q_i probability distributions coincide.

That is all. Here it is, a new index s afore of you. Concept is clear and the definition simple. But can it compete with the h-index which already has gained impetus? I do not know. In fact, it does not matter much whether the new index will be embraced with delight or will be coldly rejected with eyes wide shut. I sound my lonely trumpet in the dark trying to relax at the edge of precipice which once again faces me. Nevertheless, I feel s-index gives more fair ranking than h-index, at least in the situations considered below.

Some obvious drawbacks of the h-index which are absent in the suggested index are the following:

- h-index does not depend on the extra citation numbers of papers which already have h or more citations. Increasing the citation numbers of most cited h papers by an order of magnitude does not change h-index. Compare, for example, the citation records $10, 10, 10, 10, 10, 10, 10, 10, 10, 10$ and $100, 100, 100, 100, 100, 100, 100, 100, 100, 100$ which have $h = 10, s = 6.8$ and $h = 10, s = 21.5$ respectively.

- h-index will not change if the scientist losses impact (ceases to be a member of highly cited collaboration). For example, citation records $10, 10, 10, 10, 10$ and $10, 10, 10, 10, 10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0$ both have $h = 5$, while s-index drops from 4.8 to 3.0.

- h-index will not change if the citation numbers of not most cited papers increase considerably. For example, $10, 10, 10, 10, 10, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0$ and $10, 10, 10, 10, 10, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4, 4$ both have $h = 5$, while s-index increases from 3.0 to 6.9.

Of course, s-index itself also has its obvious drawbacks. For example, it is a common case that an author publishes a new article which gains at the beginning no citations. In this case the entropy will typically decrease and so will the s-index. I admit such a feature is somewhat counter-intuitive for a quantity assumed to measure the impact of scientific research output. However, the effect is only sizable for very short citation records and in this case we can say that there really exists some amount of objective uncertainty in the estimation of the impact.

Anyway, you can hardly expect that a simple number can substitute for complex judgments implied by traditional peer review. Of course, the latter is subjective. Nothing is perfect under the Moon. It is tempting, therefore, to try to overcome this possible subjectivity of peer review by using "simple and objective" numerical measures. Especially because some believe academic managers "don't have many skills, but they can count" [19]. In reality, however, it is overwhelmingly evident [20] that simple numerical indicators, like s-index proposed here, neither can eliminate subjectivity in management science nor prevent a dull academic management. But they can add some fun to the process, if carefully used: "citation statistics, impact factors, the whole paraphernalia of bibliometrics may in some circumstances be a useful servant to us in our research. But they are a very poor master indeed" [21].

It will be useful to compare two "servants" on some representative set of real-life citation records and Fig. 1 gives a possibility. One hundred citation records were selected more or less randomly from the *Citebase* [22] citation search engine. Fig. 1 shows h-index plotted against s-index for these records.

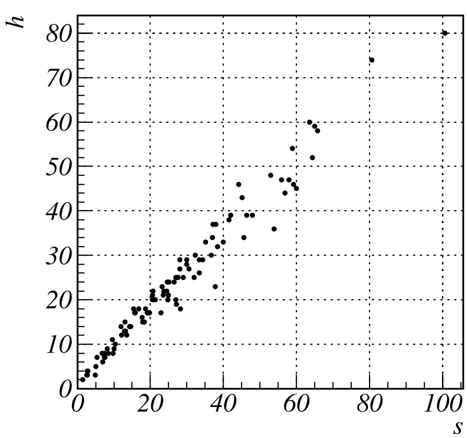

Fig. 1. h-index *versus* s-index for one hundred *Citebase* records.

What a surprise! h- and s-indexes are strongly correlated and almost equal for a wide range of their values. Of course, the coefficient, 1/4, in (3) was chosen to make these two indexes relatively equal for *some* citation records, but I have not expected them to remain close for *all* citation records.

There is some mystery here. Let us try to dig up what it is. Recent studies indicate that on the one hand the normalized entropy does not change much from record to record with $S/S_0 \approx 0.8$ [23], and on the other hand, the scaling law (1) for the h-index is well satisfied with $a = 4$ [24]. These two facts and a simple calculation imply that s- and h-indexes are expected to be approximately equal. Therefore, the real question is why these regularities are observed in the citation records?

Common sense and some experience on the citation habits tell us that these habits are the subject of preferential attachment — the papers that already are popular tend to attract more new citations than less popular papers. It is well known that the preferential attachment can lead to power laws [25]. Namely, with regard to citations, if the citations are ranked in the decreasing order $C_1 \geq C_2 \geq \ldots C_N$ then the Zipf's law [26] says that

$$C_i = \frac{C}{i^\alpha}\,. \tag{6}$$

Empirical studies reveal [27, 29] that the Zipf's law is a ubiquitous and embarrassingly general phenomenon and in many cases $\alpha \approx 1$. The citation statistics also reveals it [30–32]. Therefore, let us assume the simplest case, distribution (6) with $\alpha = 1$. Then $C_h = h$ condition determines the Hirsch index

$$h = \sqrt{C}\,. \tag{7}$$

And we see that, if the individual citation records really follow the Zipf distribution $C_i = C/i$, the Hirsch index is a really good indicator as it determines the only relevant parameter C of the distribution. In fact, the number of papers are finite and we have a second parameter, the total number of papers N. For sufficiently large N

$$C_{\text{tot}} = \sum_{i=1}^{N} \frac{C}{i} \approx C \int_1^N \frac{dx}{x} = C \ln N\,.$$

Therefore, from (7) we get the following scaling law

$$h \approx h_N = \sqrt{\frac{C_{\text{tot}}}{\ln N}} \tag{8}$$

instead of (1). However, $\sqrt{\ln N}$ varies from 1.84 to 2.21 then N varies from 30 to 130 and this explains the observations of [24].

Note that the relation (7) was already suggested by Egghe and Rousseau, on the basis of Zipf's law, in [33]. For other references where the connections between Zipf's law and h-index are discussed see, for example, [5, 34, 35].

As for s-index, Zipfian distribution $C_i = C/i$ implies the probabilities (assuming N is large)

$$p_i = \frac{1}{i \ln N}$$

and, hence, the following entropy

$$S = \frac{1}{\ln N} \sum_{i=1}^{N} \frac{1}{i} \ln\left(i \ln N\right) \approx \ln\left(\ln N\right) + \frac{1}{\ln N} \sum_{i=1}^{N} \frac{1}{i} \ln i\,.$$

But

$$\sum_{i=1}^{N} \frac{1}{i} \ln i \approx \int_{1}^{N} \frac{\ln x}{x}\, dx = \frac{1}{2} \ln^2 N\,.$$

Therefore,

$$\frac{S}{S_0} \approx \frac{1}{2} + \frac{\ln(\ln N)}{\ln N} = \frac{\ln(\sqrt{N}\ln N)}{\ln N}\,. \tag{9}$$

This expression gives 0.86 for $N=30$ and 0.82 for $N=130$ which are quite close to what was found in [23] (although for a small data set).

Because

$$\frac{S_{\rm KL}}{S_0} = 1 - \frac{S}{S_0} \approx \frac{1}{2} - \frac{\ln(\ln N)}{\ln N}$$

is small, we have the following scaling behaviour for the s-index from (4)

$$s \approx s_N = \frac{2}{3}\sqrt{C_{\rm tot}}\,\frac{\ln(\sqrt{N}\ln N)}{\ln N}\,. \tag{10}$$

Fig. 2 and Fig. 3 demonstrate an empirical evidence for these scaling rules from the *Citebase* citation records mentioned above.

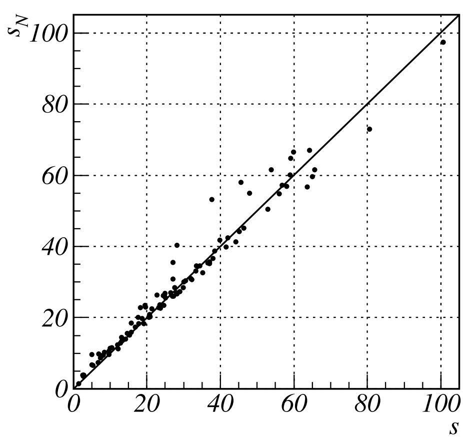

Fig. 2. s-index scaling: s_N *versus* s-index. The solid line corresponds to the ideal scaling.

As we see, the scalings (10) and (8) are quite pronounced in the data. However, there are small number of exceptions. Therefore, citation patterns are not always Zipfian. Inspection shows that in such cases the Zipfian

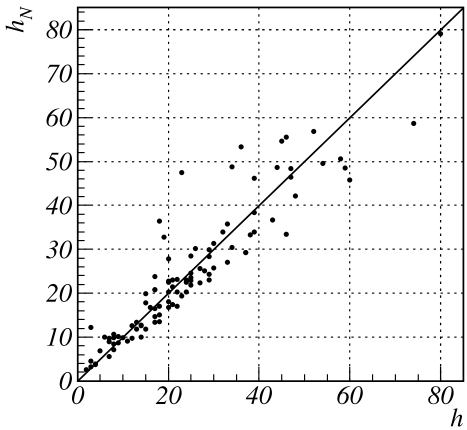

Fig. 3. h-index scaling: h_N *versus* h-index. The solid line corresponds to the ideal scaling.

behaviour is spoiled by the presence of several renowned papers with very high number of citations. If they are removed from the citation records, the Zipfian scalings for s- and h-indexes are restored for these records too.

To conclude, we wanted to overthrow the King but the King turned out to be quite healthy. The secret magic that makes h-index healthy is Zipfian behaviour of citation records. Under such behaviour, a citation record really has only one relevant parameter and the Hirsch index just gives it. However, not all citation records exhibit Zipfian behaviour and for such exceptions the new index related to the entropy of the citation record probably makes better justice. But, I am afraid, this is not sufficient to sound our trumpets and cry *Le Roi est mort. Vive le Roi!*

The magic has an another side however. The Zipfian character of citation records probably indicate the prevalence of preferential attachment and some randomness in the citation process. We can use a propagation of misprints in scientific citations to estimate how many citers really read the original papers and the striking answer is that probably only 20% do, the others simply copying the citations [36].

Of course, it will go too far to assume that "copied citations create renowned papers" [37], but these observations clearly give a caveat against taking all the indexes based on the number of citations too seriously. "Not everything that can be counted counts, and not everything that counts can be counted" [38]. Such local measures of the research impact estimation should be taken with a grain of salt and at least supplemented by different instruments for analyzing the whole citation network, like the one given in [39].

The author thanks to anonymous referee for helpful and constructive comments.

REFERENCES

[1] J.E. Hirsch, *Proc. Nat. Acad. Sci.* **46**, 16569 (2005) [arXiv:physics/0508025].

[2] P. Ball, *Nature* **436**, 900 (2005).

[3] D.F. Taber, *Science* **309**, 4 (2005).

[4] L. Bornmann, H.-D. Daniel, *EMBO reports* **10(1)**, 2 (2009).

[5] W. Glänzel, *Scientometrics* **67**, 315 (2006).

[6] L. Bornmann, H.-D. Daniel, *Scientometrics* **65**, 391 (2005).

[7] S. Lehmann, A.D. Jackson, B.E. Lautrup, arXiv:physics/0512238; *Nature* **444**, 1003 (2006).

[8] J.E. Hirsch, *Proc. Nat. Acad. Sci.* **104**, 19193 (2007) [arXiv:0708.0646v [physics.soc-ph]].

[9] L. Egghe, *ISSI Newsletter* **2(1)**, 8 (2006).

[10] S.B. Popov, arXiv:physics/0508113.

[11] D. Katsaros, L. Akritidis, P. Bozanis, *J. Am. Soc. Inf. Sci. Tech.* **60N5**, 1051 (2009).

[12] M. Schreiber, *Europhys. Lett.* **78**, 30002 (2007) [arXiv:physics/0701231].

[13] C.E. Shannon, *Bell Syst. Tech. J.* **27**, 379 (1948); **27**, 623 (1948).

[14] R.P. Feynman, *Feynman Lectures on Computation*, eds. A.J.G. Hey, R.W. Allen, Addison-Weslay, Reading 1996, p. 123.

[15] S. Kullback, R.A. Leibler, *Ann. Math. Statist.* **22**, 79 (1951).

[16] T.M. Cover, J.A. Thomas, *Elements of Information Theory*, Wiley-Interscience, New York 1991.

[17] R. Gray, *Entropy and Information Theory*, Springer-Verlag, New York 1990.

[18] S. Guiasu, *Information Theory with Applications*, McGraw-Hill, New York 1977.

[19] P.G. Hall, *Statist. Sci.* **24**, 25 (2009) [arXiv:0910.3546 [stat.ME]].

[20] R. Adler, J. Ewing, P. Taylor, *Statist. Sci.* **24**, 1 (2009) [arXiv:0910.3529 [stat.ME]].

[21] B.W. Silverman, *Statist. Sci.* **24**, 15 (2009) [arXiv:0910.35321 [stat.ME]].

[22] http://www.citebase.org/

[23] M. Schreiber, *Ann. Phys.* **522**, 536 (2010) [arXiv:1005.5227 [physics.soc-ph]].

[24] S. Redner, *J. Stat. Mech.* L03005 (2010) [arXiv:1002.0878 [physics.data-an]].

[25] M. Mitzenmacher, *Internet Mathematics* **1N2**, 226 (2004).

[26] G.K. Zipf, *Human Behavior and the Principle of Least Effort*, Addison-Wesley, Cambridge 1949.

[27] W. Li, *Glottometrics* **5**, 14 (2003).

[28] M.V. Simkin, V.P. Roychowdhury, `arXiv:physics/0601192v2`.

[29] M.E.J. Newman, *Contemp. Phys.* **46**, 323 (2005) [`arXiv:cond-mat/0412004`].

[30] D.J.de S. Price, *Science* **149**, 510 (1965).

[31] S. Redner, *Eur. Phys. J.* **B4**, 131 (1998) [`arXiv:cond-mat/9804163`].

[32] Z.K. Silagadze, *Complex Syst.* **11**, 487 (1997) [`arXiv:physics/9901035`]. This paper demonstrates a violation of causality: written in 1999, it was published in 1997!

[33] L. Egghe, R. Rousseau, *Scientometrics* **69**, 121 (2006).

[34] K. Barcza, A. Telcs, *Scientometrics* **81**, 513 (2009).

[35] J.E. Iglesias, C. Pecharromán, *Scientometrics* **73**, 303 (2007) [`arXiv:physics/0607224`].

[36] M.V. Simkin, V.P. Roychowdhury, *Complex Syst.* **14**, 269 (2003) [`arXiv:cond-mat/0212043`].

[37] M.V. Simkin, V.P. Roychowdhury, *Ann. Improb. Res.* **11**, 24 (2005) [`arXiv:cond-mat/0305150`].

[38] Albert Einstein, quoted in [11].

[39] F. Radicchi, S. Fortunato, B. Markines, A. Vespignani, *Phys. Rev.* **E80**, 056103 (2009) [`arXiv:0907.1050v2 [physics.soc-ph]`].