Citation entropy and research impact estimation

Z. K. Silagadze
Budker Institute of Nuclear Physics and Novosibirsk State University, 630 090, Novosibirsk, Russia

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.

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 Jorge 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, 4, 5, 6, 7, 8, 9, 10, 11].

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 which has received \( h \) or more citations. Empirically,

\[
h = \sqrt{\frac{C_{tot}}{a}} ,
\]

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

And now, with this simple and adorable instrument of ranking on the pedestal, I’m 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.

Claude Shannon introduced the quantity

\[
S = - \sum_{i=1}^{N} p_i \log p_i ,
\]

which is a measure of information uncertainty and plays a central role in information theory [12]. On the advice of John Von Neumann, Shannon called it entropy. According to Feynman [13], 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

\[
p_i = \frac{C_i}{C_{tot}} ,
\]

where \( C_i \) is the number of citations on the \( i \)-th paper of the citation series. Now, in analogy with [12], we can define the citation record strength index, or \( s \)-index, as follows

\[
s = \frac{1}{2} \sqrt{\frac{SC_{tot}}{S_0}} ,
\]

where

\[
S_0 = \log N
\]

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

That’s 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 situation considered below.

Let us consider the following citation records from the SPIRES database (see Table I). What we can say about these unpersonalized records? Just looking at the numbers? I think the citation record \( A \) is good, \( B \) - average, \( C \) - quite remarkable, \( D \) - splendid, \( E \) - very good, and \( F \) - excellent.
Not surprising, because \( D \) is for Paul Dirac and \( F \) for Richard Feynman (as represented in the SPIRES database). You would be surprised to know who \( A \) is, but I postpone revealing his personality. I know physicist \( B \) very well and he is, alas, not-prominent (at least up to now). \( C \) is for Petr Hořava, who is well known for his articles with Edward Witten in string theory. At last \( E \) is for William Unruh, a well established researcher with many creative ideas.

How all this is captured by \( h \) and \( s \) indexes? See Table I to be convinced that \( s \)-index conveys our intuitive ranking (emerging solely from the unpersonalized citation records, before revealing personalities of the physicists behind the records) better than \( h \)-index.

| Citation Record | \( C_{tot} \) | \( S/S_0 \) | \( s \)-index | \( h \)-index |
|-----------------|--------------|-------------|-------------|-------------|
| A               | 2375         | 0.72        | 20.7        | 19          |
| B               | 1647         | 0.82        | 18.4        | 22          |
| C               | 5070         | 0.61        | 27.8        | 21          |
| D               | 7550         | 0.76        | 37.8        | 27          |
| E               | 4044         | 0.72        | 26.9        | 27          |
| F               | 9984         | 0.75        | 43.2        | 28          |

Table II: Total citation number, relative entropy, \( s \) and \( h \) indexes for citation records from the Table I.

One thing which catches the eye is that citation series seem to be completely random. That is knowing even all citation numbers \( C_i \) for \( i = 1, \ldots, n \), we can not predict the next citation number \( C_{n+1} \). In fact there exist an objective criterion to decide whether the series is random or deterministic. This criterion is also related to the magic word, entropy, and is called permutation entropy [14].

The permutation entropy is defined as follows. Let us split the citation series in overlapping groups of \( n \) elements in each group. For example, for \( n = 3 \) and citation record \( C \) we have the following groups

\[
\{258, 202, 0\}, \quad \{202, 0, 0\}, \quad \{0, 0, 13\}, \quad \{0, 13, 58\}, \quad \{13, 58, 12\}, \quad \{58, 12, 18\}, \quad \{12, 18, 4\}, \\
\{18, 4, 21\}, \quad \{4, 21, 25\}, \quad \{21, 25, 26\}, \quad \{25, 26, 30\}, \quad \{26, 30, 1778\}, \quad \{30, 1778, 1316\}, \quad \{1778, 1316, 219\}, \\
\{1316, 219, 65\}, \quad \{219, 65, 107\}, \quad \{65, 107, 17\}, \quad \{107, 17, 235\}, \quad \{17, 235, 41\}, \quad \{235, 41, 45\}, \quad \{41, 45, 97\}, \\
\{45, 97, 116\}, \quad \{97, 116, 108\}, \quad \{116, 108, 61\}, \quad \{108, 61, 21\}, \quad \{61, 21, 0\}, \quad \{21, 0, 25\}, \quad \{0, 25, 16\}, \\
\{25, 16, 20\}, \quad \{16, 20, 7\}, \quad \{20, 7, 6\}, \quad \{7, 6, 6\}, \quad \{6, 6, 9\}, \quad \{6, 9, 10\}, \quad \{9, 10, 17\}, \\
\{10, 17, 21\}, \quad \{17, 21, 26\}, \quad \{21, 26, 14\}.
\]
For each group, there exists a unique permutation \( \pi \) which orders the elements of this group in an ascending manner. For example, for the first group, \( \{258, 202, 0\} \), the permutation which orders it is \( \{3, 2, 1\} \), because \( 0 < 202 < 258 \). In the second group, \( \{202, 0, 0\} \), we have two equal elements. In such a case we assume that the first of them is in fact a bit smaller, according to greater time span and, therefore, low accumulation rate which it is assumed to correspond. Under such agreement, the order permutation of the \( \{202, 0, 0\} \) group is \( \{2, 3, 1\} \). Analogously, we can find order permutations for other groups also and get the following table:

\[
\begin{align*}
\pi_6, \pi_3, \pi_1, \pi_2, \pi_5, \pi_4, \pi_6, \pi_5, \pi_2, \pi_6, \pi_4, \pi_2, \pi_3, \pi_6, \pi_4, \pi_6, \pi_6, \\
\pi_5, \pi_4, \pi_3, \pi_2, \pi_6, \pi_3, \pi_1, \pi_1, \pi_1, \pi_2,
\end{align*}
\]

where

\[
\pi_1 = \{1, 2, 3\}, \quad \pi_2 = \{3, 1, 2\}, \quad \pi_3 = \{1, 2, 3\}, \quad \pi_4 = \{1, 3, 2\}, \quad \pi_5 = \{2, 1, 3\}, \quad \pi_6 = \{3, 2, 1\}.
\]

Now we can calculate relative frequencies by which particular order permutations occur: \( p_1 = 13/38 = 0.342 \), \( p_2 = 5/38 = 0.132 \), \( p_3 = 6/38 = 0.158 \), \( p_4 = 4/38 = 0.105 \), \( p_5 = 3/38 = 0.079 \) and \( p_6 = 7/38 = 0.184 \). Then the permutation entropy of order \( n \) is given by the Shannon formula

\[
S_p^{(n)} = - \sum_{i=1}^{n!} p_i \log p_i. \tag{4}
\]

In our above example of the citation record \( C \), we get \( S_p^{(3)}/\log 6 = 0.93 \), where \( \log 6 \) is the maximum permutation entropy for order \( n = 3 \), corresponding to the completely random series where each order permutation \( \pi_i \) will appear with equal frequency \( p_i = 1/6 \). For other citation records, the normalized permutation entropies are even higher, ranging from 0.96 to 0.99. Taking in mind that citation series considered are relatively short, \( N \sim 100 \), I don’t think that the deviations from unity are statistically significant.

I even tried to overcome the principle of least effort [15], which leads to the fancy Zipf’s law [16], and analyze longer citation record of Edward Witten (\( N \sim 300 \)). Having \( h = 141 \) and \( s = 136.7 \), this citation record has permutation entropy \( S_p^{(3)}/\log 6 > 0.99 \). Even if we increase the order of the permutation entropy and take \( n = 4 \), we get almost the same result for the normalized permutation entropy: \( S_p^{(4)}/\log 24 = 0.99 \). Therefore, we can safely conclude that citation series are expected to be random. An interesting question is whether self-citations introduce some deterministic element in the citation record and can be, therefore, to some extent detected by measuring the permutation entropy.

It remains to surprise you by revealing that the physicist \( A \) is in reality Albert Einstein! Do you need a caveat more profound against taking all these indexes and the corresponding ranking of scientists too seriously? “Not everything that can be counted counts, and not everything that counts can be counted” [17]. Sound your trumpets!

---

[1] J. E. Hirsch, “An index to quantify an individual’s scientific output,” Proc. Nat. Acad. Sci. 102, 16569 (2005) [arXiv:physics/0508025].
[2] P. Ball, “Index aims for fair ranking of scientists,” Nature 436, 900 (2005).
[3] D. F. Taber, “Quantifying publication impact,” Science 309, 4 (2005).
[4] L. Bornmann and H.-D. Daniel, “The state of h index research. Is the h index the ideal way to measure research performance?” EMBO reports 10(1), 2 (2009).
[5] W. Glänzel, “On the h-index – A mathematical approach to a new measure of publication activity and citation impact,” Scientometrics 67, 315 (2006).
[6] L. Bornmann and H.-D. Daniel, “Does the h-index for ranking of scientists really work?” Scientometrics 65, 391 (2005).
[7] S. Lehmann, A. D. Jackson and B. E. Lautrup, “Measures and Mismeasures of Scientific Quality,” [arXiv:physics/0512238v2 [physics.soc-ph]]; “Measures for measures,” Nature 444, 1003 (2006).
[8] J. E. Hirsch, “Does the h index have predictive power?” Proc. Nat. Acad. Sci. 104, 19193 (2007). [arXiv:0708.0646v2 [physics.soc-ph]].
[9] L. Egghe, “An improvement of the h-index: the g-index,” ISSI Newsletter 2, 8 (2006). 8.
[10] S. B. Popov, “A parameter to quantify dynamics of a researcher’s scientific activity,” [arXiv:physics/0508113v1 [physics.soc-ph]].
[11] D. Katsaros, L. Akritidis and P. Bozanis, “The f Index: Quantifying the Impact of Coterminal Citations on Scientists’ Ranking,” J. Am. Soc. Inf. Sci. Tech. 60N5, 1051 (2009).
[12] C. E. Shannon, “A mathematical theory of communication,” Bell Syst. Tech. J. 27, 379 (1948); 27, 623 (1948).
[13] R. P. Feynman, *Feynman Lectures on Computation*, edited by Anthony J.G. Hey and Robin W. Allen (Addison-Wesley, Reading, 1996), p. 123.

[14] C. Bandt and B. Pompe, “Permutation Entropy: A Natural Complexity Measure for Time Series,” Phys. Rev. Lett. **88**, 174102 (2002).

[15] G. K. Zipf, *Human behavior and the principle of least effort*, (Addison-Wesley, Cambridge, 1949).

[16] Z. K. Silagadze, “Citations and the Zipf-Mandelbrot’s law,” Complex Syst. **11**, 487 (1997) [arXiv:physics/9901035](https://arxiv.org/abs/physics/9901035).

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