On the meaning of the *h*-index

J. Stat. Mech. (2010) L03005

(http://iopscience.iop.org/1742-5468/2010/03/L03005)

Journal of Statistical Mechanics: Theory and Experiment
An IOP and SISSA journal

LETTER

On the meaning of the h-index

S Redner

Center for Polymer Studies and Department of Physics, Boston University, Boston, MA 02215, USA
E-mail: redner@buphy.bu.edu

Received 6 February 2010
Accepted 19 February 2010
Published 16 March 2010

Online at stacks.iop.org/JSTAT/2010/L03005
doi:10.1088/1742-5468/2010/03/L03005

Abstract. The h-index—defined as the value such that an individual has published at least h papers with at least h citations—has become a popular metric for assessing the citation impact of scientists. As already noted in the original work of Hirsch and as evidenced from data for a representative sample of 255 physicists, $\sqrt{c}$ scales as h, where c is the total number of citations for an individual. Thus $\sqrt{c}$ appears to be equivalent to the h-index. As a further check of this equivalence, the distribution of the ratio $s \equiv \sqrt{c}/2h$ for this sample is sharply peaked about 1. The outliers in this distribution reveal fundamentally different kinds of individual publication records.

Keywords: scaling in socio-economic systems

ArXiv ePrint: 1002.0878

1742-5468/10/L03005+6$30.00

What is the best way to assess the influence of scientific publications of individual scientists? Traditionally, this assessment has been based on the number of publications of a scientist or the total number of citations received. However, in any creative endeavor, such as physics research, the total amount of output is not necessarily the right metric for productivity. In fact, Landau himself [2] kept a list of physicists who were ranked on a logarithmic scale of achievement.

Recently, Hirsch [1] introduced the h-index that attempts to capture the overall impact of an individual's publication record as a researcher in a single number. The total number of publications can be misleading because an individual could simply publish a large number of worthless articles. Conversely, the total number of citations could also be misleading because an individual might publish a single highly cited article in a hot but transient subfield but then nothing else of scientific value. Such a citation record may not be as valuable as that of someone who steadily authors good publications that are reasonably well cited.

The idea underlying the h-index is that an equitable integral measure of citation impact is provided by the value h, defined as the value such that an individual has published at least h papers with at least h citations. It is obvious that the h-index of a prolific author of trivial publications and that of a researcher with a single great publication will be much less than that of someone who publishes good papers at a steady rate. Because of its obvious appeal, the h-index has become a universally used metric for overall citation impact. As one example of the prominence of the h-index, it is immediately quoted in the Web of Science citation reports [3]. Moreover, the original idea of the h-index has spawned various of efforts to make the h-index more 'fair' [4] by correcting for some of the obvious biases that are part of the citation record, such as there being many co-authors, self-citations, the role of a thesis advisor, etc.

However, as noted by Hirsch in his original publication [1], the h-index of an individual should scale as the square root of the total number citations for this individual. This square-root scaling arises in the simplest model of citations in which an individual publishes papers at a constant rate and each publication is cited at a constant rate. As a result, the total number of citations grows quadratically with time while the h-index grows linearly with time, i.e. $\sqrt{c}$ scales linearly with h. Here, we test this observation for a representative sample of 255 condensed-matter and statistical physics theorists in North America and Europe.

The data were obtained by starting with the names of well-known condensed-matter and statistical physics theorists and looking up their citation records in the ISI Web of Science. By scanning the author lists of the top-cited publications of these initial authors, the initial list of authors was extended to include their main collaborators, and then to include collaborators of collaborators, etc. After about 250 people, it became difficult to find new people or people who could be unambiguously resolved in the ISI database with the limited knowledge of the author. Primarily because of limited personal knowledge, the data set also under-represents junior people. Moreover, because the Boston University institutional subscription for ISI extends only to citations after 1973, individuals who began publishing before this year were excluded to avoid the use of incomplete citation data for their publications. The data were gathered during a two-day period, 30–31 January 2010, between updates of the science citation index database.

J. Stat. Mech. (2010) L03005

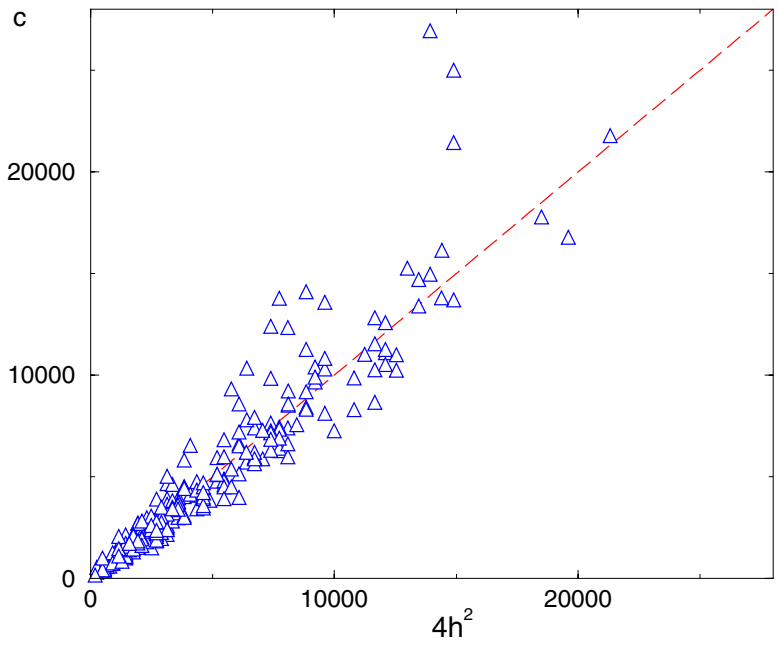

Figure 1. Plot of c versus $4h^2$ for the 255 individuals in the data set. The line $c = 4h^2$ is shown dashed.

If $\sqrt{c}$ scales linearly with h, then a plot of these two quantities should yield a straight line. Figure 1 illustrates this behavior for all the individuals in the data set. To highlight the outliers for the linear behavior that will be discussed below, figure 1 actually shows c versus $4h^2$. A linear least-squares fit to all the $\sqrt{c}$ versus $2h$ data gives a best fit value of the slope $s \equiv \sqrt{c}/2h$ of $s \approx 1.045$. The data therefore suggest that $\sqrt{c}$ is essentially equivalent to the h-index, up to an overall factor that is close to two.

As a further test of the linearity of the dependence of h versus $\sqrt{c}$, the quantity $s = \sqrt{c}/2h$ is computed for each individual in the data set of 255 physicists and the resulting distribution, $P(s)$, is shown in figure 2. This distribution is fairly symmetric and most of the data lie within the range $|s - 1| < 0.2$. The tightness of the range of s again suggests that the relation $\sqrt{c} = 2h$ accounts for most of the citation data.

The outliers in the distribution $P(s)$ with $s < 1$ and with $s > 1$ are particularly interesting. In the scatter plot of c versus $4h^2$ in figure 1, consider first the outliers with $s < 1$; these are data points that lie below the diagonal. As illustrated in table 1, the citation patterns of best-cited publications for the individuals with the smallest ten values of s are remarkably similar even though the h-indices of this group of researchers range over a factor of more than two. In particular, the difference in number of citations of successive top-cited papers is relatively small in all cases. For example, the ratio of the number of citations to the top-cited and third-cited paper for each individual is in the range 1.025–2.072.

For the 20 individuals with the largest value of s (table 2), the citation patterns are also quite similar within this subpopulation. Almost all have one (or a few) papers whose citations are a substantial factor larger than their second-ranked paper. For example, the largest ratio between the number of citations of the top-cited and

J. Stat. Mech. (2010) L03005

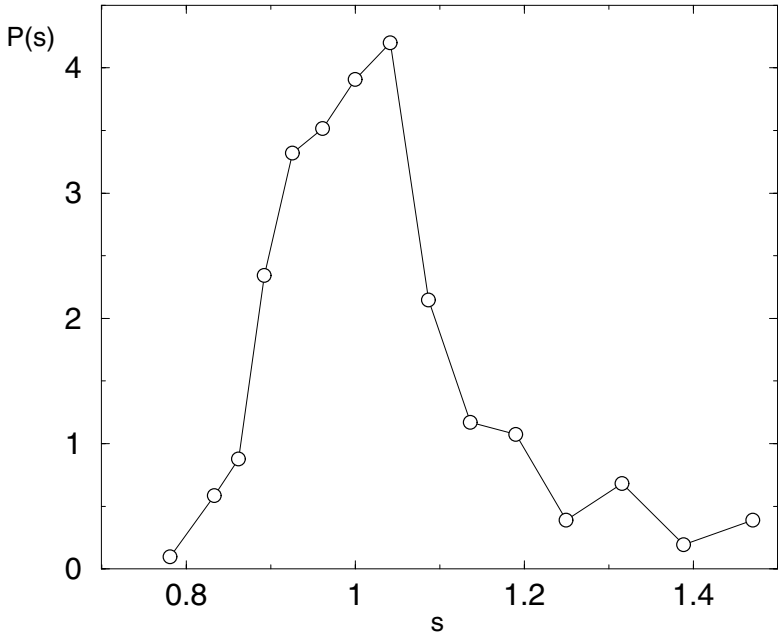

Figure 2. Plot of the probability density $P(s)$ that an individual is characterized by a value $s = \sqrt{c}/2h$.

Table 1. List of the ten top-cited publications of the individuals with the ten smallest values of $s = \sqrt{c}/2h$. The first three columns give the h-index, the total number of citations c, and $s = \sqrt{c}/2h$. The columns labeled c_i for $i = 1, 2, \ldots, 10$ are the respective numbers of citations of the ten best-cited papers for each individual.

h	c	$\sqrt{c}/2h$	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
25	1510	0.777	84	81	62	48	46	43	42	39	37	36
39	3983	0.809	260	177	144	127	126	92	91	90	89	85
18	853	0.811	172	153	83	72	49	39	36	35	33	23
27	1966	0.821	197	191	139	110	66	66	52	51	48	44
26	1854	0.828	83	81	81	72	70	68	63	56	55	52
28	2169	0.832	100	95	92	89	83	75	73	67	64	64
19	1002	0.833	68	66	64	56	51	51	50	43	42	39
26	1879	0.833	148	141	84	76	75	65	64	62	56	54
23	1480	0.836	94	64	64	62	58	51	49	47	46	42
54	8209	0.839	316	297	285	199	198	198	181	177	162	153

third-cited paper is now 10.03. This wide disparity arises because each individual in this subpopulation (co-)authored one (or a few) famous publications whose citation frequency outstrips the remaining publications. Among the individuals that (co-)author these famous publications, there are three clearly defined situations: (i) individuals that wrote a ground-breaking publication on their own or were the driver of a publication with a junior co-author, (ii) those that collaborated with a more senior author in a famous publication, and (iii) those whose famous publication was a particularly timely or authoritative review article.

doi:10.1088/1742-5468/2010/03/L03005

J. Stat. Mech. (2010) L03005

Table 2. List of the citation records of the individuals with the 20 largest values of $s = \sqrt{c}/2h$; the data format is the same as for table 1. Italicized entries denote review articles.

h	c	$\sqrt{c}/2h$	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}
8	544	1.458	141	135	50	34	31	17	13	13	8	8
11	1 011	1.445	329	220	105	75	73	37	28	24	24	17
20	3 163	1.406	480	303	276	264	257	212	198	191	165	157
59	26 937	1.391	2259	1830	1310	1220	784	777	606	355	54	312
44	13 789	1.334	1824	1469	1393	1042	570	560	504	480	327	316
17	2 058	1.334	550	255	197	194	123	97	81	73	70	70
27	4 903	1.297	2004	371	316	243	157	133	114	100	98	97
61	25 003	1.296	4461	*3778*	1444	1333	1176	1104	1101	835	651	400
43	12 403	1.295	4148	1561	551	495	452	405	399	339	217	214
40	10 347	1.271	2118	2004	857	433	292	281	274	238	223	221
38	9 331	1.271	2721	828	530	472	466	451	324	271	205	178
32	6 537	1.263	1105	735	650	525	516	320	174	154	151	138
47	14 090	1.263	*3232*	815	699	620	477	466	420	353	329	274
45	12 347	1.235	*2357*	765	641	563	495	462	405	377	350	322
28	4 660	1.219	2260	274	206	140	116	86	84	83	81	79
19	2 137	1.271	766	301	182	77	74	71	61	58	43	41
61	21 446	1.200	7014	1102	699	626	502	427	331	325	304	296
15	1 274	1.190	242	232	140	96	66	57	48	41	34	33
49	13 582	1.189	3051	985	883	864	698	374	349	349	302	241
22	2 732	1.188	569	343	271	192	165	98	96	90	72	63
39	8 584	1.188	2260	980	658	451	296	289	269	149	147	144
22	2 699	1.181	507	340	192	184	145	130	121	93	92	90

One basic conclusion from this study is that the square root of the total number of citations that an individual receives very nearly coincides with twice his or her h-index. A still open question is that of why $\sqrt{c}$ should provide the same integrated measure of the breadth and depth of an individual's citation record as the h-index itself.

A second conclusion is that it is possible to identify outstanding researchers as the outliers above the diagonal in the scatter plot of figure 1. While there are roughly the same number of points below the diagonal as above the diagonal, the above-diagonal points with roughly 9000 citations or more are visually prominent and correspond to individuals with seminal publications. This simple characteristic appears to provide a useful predictor of research excellence.

A final caveat: while the outliers discussed here correspond to researchers with excellent publications to their credit, there are many examples of excellent researchers that do not fit this outlier criterion. It is important to be aware of the limitations of using citations alone, or some function of the number of citations, as a measure of research excellence.

I gratefully acknowledge financial support from the US National Science Foundation grant DMR0906504. I also thank S Dorogovtsev for initial correspondence that kindled my old interest in this subject, J E Hirsch for friendly correspondence and advice and S Olejnik for helpful comments and for informing me about ref [5].

References

[1] Hirsch J E, 2005 *Proc. Nat. Acad. Sci.* **46** 16569
[2] See http://en.wikipedia.org/wiki/Lev_Landau
[3] http://apps.isiknowledge.com/
[4] See e.g., Egghe L, 2006 *Scientometrics* **69** 131
Van Raan A F J, 2006 *Scientometrics* **67** 491
Hirsch J E, arXiv:0911.3144 [physics.soc-ph]
[5] See e.g., Nielsen M, http://michaelnielsen.org/blog/why-the-h-index-is-virtually-no-use/

J. Stat. Mech. (2010) L03005