

Using the H-index to Rank Influential Information Scientists

Blaise Cronin and Lokman I. Meho

School of Library and Information Science, Indiana University Bloomington, IN 47401, USA
{bcronin; meho}@indiana.edu

Abstract: We apply a new bibliometric measure, the *h*-index (Hirsch, 2005), to the literature of information science. Faculty rankings based on raw citation counts are compared with those based on *h*-counts. There is a strong positive correlation between the two sets of rankings. We show how the *h*-index can be used to express the broad impact of a scholar's research output over time in more nuanced fashion than straight citation counts.

Introduction

A new measure, the *h*-index, has been proposed to quantify the impact of individual scientists' research output (Hirsch, 2005). It has generated considerable interest (e.g., Dumé, 2005) and attracted largely favorable comment (e.g., Ball, 2005). The measure is elegantly simple: "*A scientist has index h if h of his/her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have fewer than h citations each*" where N_p is the number of papers published over n years (Hirsch, 2005, p. 1) (italics in original). That is to say, a scientist with an *h*-index of 30 has published 30 papers that have each attracted at least 30 citations. Of course, some of those 30 papers may have attracted considerably more than 30 citations, just as others of the author's publications may have attracted considerably fewer than 30 citations. The value of *h* is likely to increase linearly with time. To state the obvious, a scientist's *h*-index will never decrease, but may well increase as new papers are published, as 'sleepers' come to life and as the individual's corpus incrementally attracts citations.

Hirsch provides a brief overview of the limitations of commonly used publication and citation measures before demonstrating the potential utility (and validity) of his *h*-index on a population of physicists and also a group of life scientists. The distinguishing feature of the *h*-index is that it captures the *broad* impact of a scientist's work: "... Nobel prizes do not originate in one stroke of luck but in a body of scientific work" (Hirsch, 2005, p. 4). In other words, an

author with very few high-impact papers or, alternatively, many low-impact papers will have a weak h -index. Crudely put, the h -index helps us distinguish between a one-hit wonder and an enduring performer. Hirsch estimates that a “successful scientist” will have an h -index of 20 after 20 years, an “outstanding scientist” one of 40 after 20 years, and a “truly unique” individual one of 60+ after 20 years (p. 3). However, he is quick to point out that values of h will be field dependent. As with all such reductionist measures of presumptive research quality, caution is required in both application and interpretation.

Approach and Methods

Here, for the first time, we apply the h -index to information science (IS) and compare rankings based on raw citation counts with those based on h -counts. In IS, as in many other fields, there is robust, ongoing debate on the pros and cons of evaluative bibliometrics and the associated techniques (e.g., Budd, 2000; Cronin & Overfelt, 1996; Meho & Spurgin, 2005). We identified 31 influential information science faculty from the U.S. (see Tables 1 and 2) based on our personal domain knowledge, augmented by a list of the 100+ most cited IS scholars in the *Social Sciences Citation Index (SSCI)*, 1999-2005. We excluded highly-cited disciplinary outsiders (e.g., Jakob Nielsen, Wanda Orlikowski) and deceased grandees (e.g., Rob Kling, Gerald Salton). A majority of our names featured in the Budd (2000) and Meho and Spurgin (2005) studies of faculty productivity. Our sample, which is *not* strictly speaking a list of the most highly cited faculty in the field, includes a number of recently retired but active scholars (e.g., Marcia Bates, Howard White). Because a relatively high h -count depends to some extent on time-in-field, our list comprises mainly mid- to late-career researchers. Junior faculty will typically not have achieved the volume of output and level of citation impact necessary to generate a competitive h -index. For convenience, time-in-field is determined by the date of each faculty member’s earliest citation found in the ISI databases.

Two sets of data were gathered: one for the h -index and another for the total citation counts. For both measures, citation data were generated from the *Arts & Humanities Citation Index* (1980-2005), *Science Citation Index* (1974-2005) and the *Social Sciences Citation Index* (1972-2005) using Dialog. All the searches were done twice, first by a research assistant and subsequently by one of the authors (LM). No significant differences were found between the two sets of results. To generate h -counts, a list of all cited papers was created for each faculty

member using the “EXPAND CR=...” function or command (CR stands for Cited Reference). Because cited items often had multiple entries (e.g., sometimes with and sometimes without the author’s middle initial), an effort was made to collate all entries that referred to the same cited item. In cases where a cited item did not provide enough information to determine where it belonged (due to citing or indexing errors), we relied on the author’s list of publications or on online database searching to verify the information. [It is interesting to note in passing that approximately two-thirds of the study sample had their bibliographies/curricula vitae available on the Web.]

For practical reasons, searches were limited to first authors. After searching for citations for each individual item, duplicate records were removed and the citation count was reported first with and then without self-citations. We included citations from all documents types: journal articles, conference papers, review articles, letters to editors, and so on. On average, it took roughly three hours of searching to generate the *h*-index for each individual author. Although we used the same searching method to generate the total citation counts, it took less than an hour on average to generate these data for each author. Hirsch and others have mentioned that generating *h*-indices using ISI’s *Web of Science* is relatively straightforward. While this may be true, the scores generated from it are inaccurate. This is because *Web of Science* does not collate the different entries for the same cited item. Table 3 illustrates the different results achieved when comparing *h*-counts derived from *Web of Science* with those generated manually using the method employed in this study.

Results

Our 31 names are ranked according to four measures: (i) total number of citations, including self-citations; (ii) total number of citations, excluding self-citations; (iii) *h*-index, with self-citations included; (iv) *h*-index, with self-citations excluded (see Tables 1 and 2). The citation range is from 84 to 1,048. When self-citations are removed, the range contracts slightly, from 79 to 1,025. In both cases, Carbo and Belkin occupy the respective pole positions. Generally, the elimination of self-citations does not much influence the rank ordering, though both Cronin and Kantor drop two notches in the rankings, from 3 to 5 and 20 to 22, respectively. The *h*-index range is from 5 to 20 (from Carbo to Belkin, respectively), with citations included, and from 5 to 19 with self-citations excluded. In the latter case, Belkin and Saracevic are tied on

19 while Carbo, Griffiths, and Schamber are tied at 5. The Spearman rank order correlation between citation ranks and *h*-index (with self-citations excluded) was 0.9, significant at the 0.01 level.

There are no tied ranks when citation counts are used, but bunching does occur with the *h*-index. For instance, eight individuals have an *h*-value of 11. The citation scores (self-citations excluded) for this octet range from 250 to 599. The year of the first item cited in the ISI databases for each author ranges from 1963 to 1991. Four are date stamped in the sixties, 13 each in the seventies and eighties. In only one case (Spink) does the earliest cited item date from the nineties. Irrespective of which of the four measures is used to generate rankings, five names are omni-present: Belkin, Saracevic, Bates, Borgman, and Cronin. These five authors had their first paper cited in either the 1960s or 1970s.

Discussion and Conclusions

The strong positive correlation between the *h*-index and citation counts suggests that the total number of citations (with or without self-citations) is, indeed, a reliable indicator of scholarly impact and influence. Nonetheless, the *h*-index provides some additional discriminatory power. By way of example, Schamber and Eisenberg have citation scores of 241 and 142, yet their *h*-indices are 5 and 8, respectively. In other words, although Schamber has accrued considerably more citations over approximately the same timeframe, more of Eisenberg's publications have had a broader impact. To take another example, Smith and McCain have roughly comparable citation counts of 281 and 317, yet their *h*-indices are 7 and 12, respectively, despite the fact that McCain has had shorter time-in-field. A different example is provided by Spink, whose citation score (389)—though appreciably lower than either Cronin's (782) or Marchionini's (671)—has an *h*-index of 12, only one less than theirs. Additionally, her first cited paper (1991) is much more recent than either Cronin's (1978) or Marchionini's (1985). Spink may be a comparative newcomer, one who is moderately rather than heavily cited, but her publications are having a relatively high impact. A final illustration: despite considerable differences in citation yields and time-in-field, Fidel, McCain, Spink and White all have the same *h*-count.

Information science is a small field (certainly by comparison with either physics or the life sciences) and it is not surprising to find relatively modest citation yields and correspondingly

modest *h*-indices. The median *h*-index (self-citations excluded) for our sample was 11, the highest 19, and the lowest five. These numbers are considerably lower than those generated by Hirsch (2005, p. 4) for either physics (median = 46) or the life sciences (median = 157), which should make it abundantly clear that cross-field comparisons are inappropriate. It would be inappropriate to stratify our sample the way Hirsch did, as successful, outstanding, or truly unique. As Hirsch (2005, p. 4) has already noted, “more research in understanding similarities and differences of *h*-index distributions in different fields of science” is needed. Others may wish to extend and refine the approach we have piloted here by, for example, enlarging the sample to accommodate both more junior and senior IS scholars or by including more detailed analysis of authors’ publication data.

Acknowledgment

We are most grateful to Sara Franks for assistance with online database searching and to both Yvonne Rogers and Debora Shaw for helpful comments.

References

- Ball, P. (2005). Index aims for fair ranking of scientists. Available online at:
<http://www.nature.com/nature/journal/v436/n7053/full/436900a.html>
- Budd, J. M. (2000). Scholarly productivity of U.S. LIS faculty: An update. *Library Quarterly*, 70(2), 230-245.
- Cronin, B. & Overfelt, K. (1996). Postscript on program rankings. *Journal of the American Society for Information Science*, 47(2), 173-176.
- Dumé, B. (2005). Number theory. Available online at:
<http://physicsweb.org/articles/news/9/8/9/1>
- Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Available online at: <http://xxx.arxiv.org/abs/physics/0508025>
- Meho, L. I. & Spurgin, K. M. (2005). Ranking the research productivity of LIS faculty and schools: An evaluation of data sources and research methods. *Journal of the American Society for Information Science and Technology*, 56(12), 1314-1331.

TABLE 1
Faculty Members Ranked by Number of Citations

Name	Total number of citations				Year first item was cited
	Including self-citations		Excluding self-citations		
	Total	Rank	Total	Rank	
Belkin, Nicholas J.	1,048	1	1,025	1	1974
Saracevic, Tefko	1,013	2	972	2	1963
Bates, Marcia J.	830	5	795	3	1971
Borgman, Christine L.	839	4	786	4	1975
Cronin, Blaise	875	3	782	5	1978
Marchionini, Gary	706	6	671	6	1985
Tenopir, Carol	655	7	599	7	1979
Buckland, Michael K.	618	9	571	8	1967
Hernon, Peter	627	8	527	9	1968
McClure, Charles R.	583	10	508	10	1974
Kuhlthau, Carol C.	480	11	463	11	1983
White, Howard D.	475	12	441	12	1974
Fidel, Raya	450	14	427	13	1981
Spink, Amanda	458	13	389	14	1991
Dillon, Andrew	373	15	352	15	1987
McCain, Katherine W.	349	16	317	16	1981
Budd, John M.	324	17	302	17	1981
Smith, Linda C.	287	19	281	18	1974
Koenig, Michael E.D.	294	18	254	19	1971
Larson, Ray R.	263	21	250	20	1981
Van House, Nancy	255	22	243	21	1983
Kantor, Paul B.	284	20	242	22	1972
Schamber, Linda	250	23	241	23	1987
Soergel, Dagobert	198	25	194	24	1967
Case, Donald O.	210	24	192	25	1981
Griffiths, José-Marie	157	27	152	26	1975
Eisenberg, Mike	148	28	142	27	1986
Losee, Robert M.	166	26	136	28	1986
Liddy, Elizabeth D.	128	29	116	29	1987
Wildemuth, Barbara M.	121	30	108	30	1977
Carbo, Toni	84	31	79	31	1975

TABLE 2
Faculty Members Ranked by H-Index

Name	H-Index				Year first item was cited
	Including self-citations		Excluding self-citations		
	H-Index	Rank	H-Index	Rank	
Belkin, Nicholas J.	20	1	19	1T	1974
Saracevic, Tefko	19	2	19	1T	1963
Bates, Marcia J.	18	3T	17	3	1971
Borgman, Christine L.	18	3T	15	4	1975
Cronin, Blaise	14	5T	13	5T	1978
Marchionini, Gary	14	5T	13	5T	1985
Fidel, Raya	13	11T	12	7T	1981
McCain, Katherine W.	14	5T	12	7T	1981
Spink, Amanda	14	5T	12	7T	1991
White, Howard D.	14	5T	12	7T	1974
Buckland, Michael K.	12	13T	11	11T	1967
Budd, John M.	11	17T	11	11T	1981
Dillon, Andrew	12	13T	11	11T	1987
Hernon, Peter	14	5T	11	11T	1968
Kuhlthau, Carol C.	12	13T	11	11T	1983
Larson, Ray R.	11	17T	11	11T	1981
McClure, Charles R.	13	11T	11	11T	1974
Tenopir, Carol	12	13T	11	11T	1979
Kantor, Paul B.	10	19T	9	19T	1972
Koenig, Michael E.D.	9	21T	9	19T	1971
Van House, Nancy	9	21T	9	19T	1983
Case, Donald O.	9	21T	8	22T	1981
Eisenberg, Mike	8	24	8	22T	1986
Losee, Robert M.	10	19T	7	24T	1986
Smith, Linda C.	7	25T	7	24T	1974
Soergel, Dagobert	7	25T	7	24T	1967
Wildemuth, Barbara M.	7	25T	7	24T	1977
Liddy, Elizabeth D.	6	28T	6	28	1987
Carbo, Toni	5	29T	5	29T	1975
Griffiths, José-Marie	5	29T	5	29T	1975
Schamber, Linda	5	29T	5	29T	1987

TABLE 3
Comparison of H-indices* Derived from *Web of Science* and Dialog Searches

Name	<i>Web of Science</i>	Dialog Searching
Belkin	17	20
Bates	16	18
Borgman	14	18
Saracevic	13	19
Cronin	12	14
Budd	11	11
Kantor	8	10
Losee	8	10
Soergel	5	7

*Includes self-citations.