Study of different h-indices for groups of authors

Peer-reviewed author version

EGGHE, Leo & RAO, Ravichandra (2008) Study of different h-indices for groups of authors. In: JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 59(8). p. 1276-1281.

DOI: 10.1002/asi.20809
Handle: http://hdl.handle.net/1942/7918

Study of different h-indices for groups of authors

by

L. Egghe [1,2,4]

and

I.K. Ravichandra Rao[1,3]

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium [1]

Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk, Belgium [2]

Indian Statistical Institute (ISI), 8th Mile, Mysore Road, R.V. College P.O., Bangalore-560059, India [3]

ABSTRACT

In this paper, for any group of authors, we define three different h-indices. First there is the successive h-index h_2 based on the ranked list of authors and their h-indices h_1 as defined by Schubert. Next there is the h-index h_P based on the ranked list of authors and their number of publications. Finally there is the h-index h_C based on the ranked list of authors and their number of citations.

We present formulae for these three indices in Lotkaian informetrics from which it also follows that $h_2 < h_P < h_C$. We give a concrete example of a group of 167 authors on the topic

---

[4] Corresponding author.

Key words and phrases: h-index, Hirsch, group of authors.

Acknowledgement: The second named author is grateful to the Universiteit Hasselt for financial support during the time he was a visiting professor.

"optical flow estimation". Besides these three h-indices we also calculate the two-by-two Spearman rank correlation coefficient and prove that these rankings are significantly related.

I. Introduction

The h-index (or Hirsch index) of an author has been introduced in Hirsch (2005) as follows. If we rank the publications of an author in decreasing order of the number of citations that these articles received then this author's h-index h is the largest rank such that all the articles on rank 1,2,…,h have h or more citations.

The h-index, introduced only 2 years ago, has become a real hype in and even outside informetrics: Ball (2005, 2007), Bornmann and Daniel (2005, 2007a), Braun, Glänzel and Schubert (2005, 2006) (introducing the h-index for journals, yielding a new journal indicator to be preferred above the impact factor – see Miller (2006)), Egghe and Rousseau (2006), Glänzel (2006a,b), Popov (2005), van Raan (2006), Bar-Ilan (2006), Rousseau (2007), Burrell (2007a,b), Glänzel and Persson (2005), Egghe (2007c), Saad (2006), Oppenheim (2007), Hirsch (2007), Barendse (2007), Wan, Hua and Rousseau (2007), Rao and Rousseau (2007), Vinkler (2007), Vanclay (2007) and see also the papers in the special issue on the Hirsch index in Journal of Informetrics 1(3), 2007: Schubert and Glänzel (2007), Beirlant, Glänzel, Carbonez and Leemans (2007), Costas and Bordons (2007) and Bornmann and Daniel (2007b).

Banks (2006) introduces the interesting notion of the h-index for topics and compounds – see also Egghe and Rao (2007) and the STIMULATE6 Group (2007). Let us, finally note that both the Web of Science and Scopus offer the h-index in their databases (remarkably quick after its introduction in 2005!).

Schubert and Prathap (independently and almost simultaneously) introduced the interesting idea of "successive h-indices" (see Schubert (2007), Prathap (2006) – see also http://11011110.livejournal.com/10507.html and www.cs.utah.edu/~shirley/hindex/ and Egghe (2007b). Denote the h-index of an author in a group of authors by h_1. If we rank these authors in decreasing order of their h-index h_1 we can then apply the definition of the h-index to this ranked list and obtain the h-index h_2 of this group of authors (e.g. an institute, hereby giving a visibility score of this institute). This is the first h-index for a group of authors that we will use in this paper. Now we introduce two other ones for the same group of authors.

h-indices can be calculated on any system of sources and items (a so-called information production process (IPP)) where the sources are ranked in decreasing order of their number of items. As sources we will always take the group of authors under consideration (as we did also above). For our next h-index of this group we rank these authors in decreasing order of their number of publications. The h-index of this ranked group will be denoted by h_P.

For our third h-index we rank the same group of authors in decreasing order of their total number of citations that where received by the publications of these authors. The h-index of this ranked group will be denoted by h_C.

Apart from the paper Liu and Rousseau (2007), where the h-index was applied to library circulation data, the present paper – we think – is the first to also apply the h-index to non-citation data.

In the next section we will present models for these three indices based on results in Lotkaian informetrics. We will derive from these models that $h_2 < h_P < h_C$.

In the third section a concrete example will be given: 167 authors in the field "optical flow estimation". We find in this case: $h_2 = 10$, $h_P = 20$ and $h_C = 67$. We also calculate the Spearman rank correlation coefficient for each couple of author rankings and show that the ranks correlate significantly.

II. Models for the h-indices h_2, h_P and h_C

Let us start with the simplest model: the one for h_P.

II.1 Model for h_P

Let there be T authors and suppose that their author-publication IPP is ruled by a Lotkaian law of the form

$$f(j) = \frac{C}{j^{\alpha_1}} \qquad (1)$$

where $C > 0$, $\alpha_1 > 1$, $j \in [1, +\infty[$: $f(j)$ denotes the density of the authors with publication density j – see Egghe (2005), Chapter II. The requirement $\alpha_1 > 1$ makes sure that

$$T = \int_1^{\infty} f(j)dj = \frac{C}{\alpha_1 - 1} \qquad (2)$$

is finite. It is generally true that the Lotka exponent α_1 is larger than 1 in practice.

In Egghe and Rousseau (2006) we showed that in such Lotkaian systems, the h-index is given by

$$h_P = T^{\frac{1}{\alpha_1}} \qquad (3)$$

So this result follows from Egghe and Rousseau (2006) in a straightforward way.

II.2 Model for h_C

Of course, as for h_P, we can here make the same argument, supposing a Lotka law as in (1) but with another value for α_1, say δ and we then have, by Egghe and Rousseau (2006):

$$h_C = T^{\frac{1}{\delta}} \qquad (4)$$

But we can do better than that: we can present a formula for δ in function of α_1 and a "positive reinforcement" parameter β, to be explained further on. Indeed, the situation described in this subsection is a typical case of a positive reinforcement of the situation described in the previous subsection – cf. Egghe (2005), Chapter III: the author-publications relation is composed with the publications-citations relation yielding the author-citations relation. It is proved in Theorem III.1.3.1.3 (page 164-167) that this yields a positive reinforcement of the item densities: j in (1) is transformed into $\varphi(j)$ where $\varphi(j) \geq j$ for all $j \geq 1$ since we only consider publications with at least 1 citation for the calculation of the h-index h_C.

As we did in Section III.2 of Egghe (2005) we propose a power law for the function φ:

$$\varphi(j) = Bj^{\beta} \tag{5}$$

where $B \geq 1$ and $\beta > 1$ since we explained above that $\varphi(j) \geq j$ is a requirement for all $j \geq 1$ (with $0 < \beta < 1$ we have that $\varphi(j) < j$ from some j on (j large enough) and $\beta = 1$ is not considered since the publication-citation relation is a strictly positive reinforcement!).

We now invoke Corollary III.2.1.1 in Egghe (2005) proving that our authors-citations relation is also Lotkaian but with f replaced by $\left(j_1 =: \varphi(j)\right)$

$$f^*(j_1) = \frac{D}{j_1^{\delta}} \tag{6}$$

where $D > 0$, $j_1 \geq B \geq 1$ and with

$$\delta = \frac{\alpha_1 + \beta - 1}{\beta} \tag{7}$$

Since now we have that $j_1 \geq B$ we have to use the extension of the result proved in Egghe and Rousseau (2006) on the h-index. This extension was proved in Egghe (2007a): we now have

$$h_C = B^{\frac{\delta - 1}{\delta}} T^{\frac{1}{\delta}} \tag{8}$$

(since the authors = sources remain the same we have taken $\psi = 1$ in Egghe (2007a)).

Readers who are not familiar with this extension can simply take $B = 1$ and apply the well-known result in Egghe and Rousseau (2006):

$$h_C = T^{\frac{1}{\delta}} \tag{9}$$

with δ as in (7). The reader will agree that (8) and (9) shed more light on h_C than (4) because we have the knowledge of (7). So we have the following theorem:

Theorem 1: In the notation above, we have (take $B = 1$):

$$h_C = T^{\frac{\beta}{\alpha_1 + \beta - 1}} \tag{10}$$

where $\alpha_1 > 1$ is the same as in (3) and where $\beta > 1$ is as in (5).

This has the following corollary:

Corollary 1:

$$h_C > h_P \tag{11}$$

Proof: This follows from the fact that

$$\frac{\beta}{\alpha_1 + \beta - 1} > \frac{1}{\alpha_1}$$

(since $\beta > 1$ and $\alpha_1 > 1$) and by (3) and (10). $\square$

Of course, in practise the above corollary is trivial since $\varphi(j) \geq j$ enforces the item numbers in positively transformed table to be larger than the ones in the original table and hence, the h-index of the transformed table must be strictly larger than the one of the original table!

Finally, we study h_2.

II.3 Model for h_2

In Egghe (2007b) we have modelled successive h-indices which we described in the introductory section. This means that, for a group of authors we assume that, per author, the publication-citation relation is ruled by a law of Lotka of the form

$$F(k) = \frac{E}{k^{\alpha_2}} \tag{12}$$

where $E > 0$ and $\alpha_2 > 1$. This expresses – per author – the density of the papers with a density of k citations received. Since we also assume that the group of authors is homogeneous (e.g. all authors are from the same field) we can, as a simplification, assume that α_2 is the same for all authors.

Let there be j publications:

$$j = \int_1^{\infty} F(k)dk$$

$$j = \frac{E}{\alpha_2 - 1} \tag{13}$$

Hence, from Egghe and Rousseau (2006) we derive that the h-index h_1 is given by

$$h_1 = j^{\frac{1}{\alpha_2}} \tag{14}$$

Hence, by (1), taking j as a publication density of an author, the author-h_1 relation is ruled by a distribution which is equal, up to a constant, to (combine (1) and (14))

$$\frac{1}{h^{\alpha_1 \alpha_2}} \tag{15}$$

The complete theory is presented in Egghe (2007b). Applying again Egghe and Rousseau (2006), assuming as in (1) that there are T authors, we have, by definition of h_2 that

$$h_2 = T^{\frac{1}{\alpha_1 \alpha_2}} \tag{16}$$

, based on (15). We have indicated the proof of the following theorem (see Egghe (2007b) for a complete proof):

Theorem 2: If we have (12) for the publication-citation relation and (1) for the author-publication relation, then the successive h-index h_2 of the group of authors is given by

$$h_2 = T^{\frac{1}{\alpha_1 \alpha_2}}$$

where T is as in (3) or (10).

Corollary 2: In all cases

$$h_2 < h_P < h_C \tag{17}$$

Proof: That $h_P < h_C$ has already been proved in Corollary 1. But, by (16)

$$h_2 = T^{\frac{1}{\alpha_1 \alpha_2}}$$

$$< T^{\frac{1}{\alpha_1}} = h_P$$

since $\alpha_2 > 1$ and by (3). □

This closes the theoretical part of this paper.

III. Application

We apply these indices to a bibliography of 167 authors on the topic "optical flow estimation". The bibliography was compiled from a list of 252 references selected from the website http://itbl.biologie.huerlin.de/~wiskott/Bibliographics/flowestimation.htma=ml in 2006. For each of the 167 authors, the number of papers was collected. The number of citations received by each of the papers of every author was determined using the Web of Science and so, the total number of citations received by every author was determined.

Six authors received no citations (hence $h = 0$). One author (J.J. Koenderink) had a maximum of $h = 25$. The data on the authorranks are shown in the appendix. We have obtained the following results.

$$h_2 = 10 < h_p = 20 < h_C = 67,$$

in accordance with Corollary 2.

It is also interesting to check the three different author rankings: in all three cases the set of authors is the same. Let us denote by $r(h_1, P)$ the Spearman rank order correlation coefficient of the comparison of the author ranks according to the author h-index h_1 and according to the number of publications per author. We have $r(h_1, P) = 0.854611$ with a corresponding t-value of 21.1406463, hence far beyond the classical critical values. We can say that, with a probability of nearly 1, that there is a positive correlation between the two ranks.

For $r(h_1, C)$, the Spearman rank order correlation coefficient of the comparison of the author ranks according to h_1 and according to the number of citations (per author), we have: $r(h_1, C) = 0.818131$ with a corresponding t-value of 18.2756527, showing almost 100% a positive correlation.

Finally, for $r(P, C)$, the Spearman rank order correlation coefficient of the comparison of the author ranks according to the number of publications and according to the number of citations, we have $r(P,C) = 0.643224$ which is, due to the large value $N = 167$ of number of authors, almost 100% a positive correlation.

It is clear that a positive correlation between h_1 and P, h_1 and C and P and C is logical, but the above results indicate that the correlation is very strong.

Remark

A h-index for a group of authors, different in nature from the ones studied here, is the so-called global h-index, denoted by h_G. The index h_G is defined as the h-index of the group of authors considered as one meta-author. This index was used in van Raan (2006) in the connection of university research groups in chemistry in the Netherlands.

If we assume that (12) also applies in this global publication-citation setting, then we have that

$$h_G = S^{\frac{1}{\alpha_2}} \tag{18}$$

where S denotes the total number of papers (of all authors). Since clearly $S \geq T$ (total number of authors), we have, by (16), that

$$h_G > h_2 \tag{19}$$

An inequality between h_G and the other indices h_P and h_C is not provable in general due to the appearance of α_1 in the formulae for h_P and h_C (and since α_1 and α_2 have no a priori relation between each other).

IV. Conclusions

In this paper we showed that there are, at least, three different ways to calculate an h-index of a group of authors: the successive h-index h_2, based on the h-indices h_1 of the authors, the h-index h_P based on the author-publication ranking and the h-index h_C based on the author-citation ranking.

Concrete formulas for these indices are presented in the framework of Lotkaian informetrics, thereby also proving the inequalities $h_2 < h_P < h_C$.

A case study of 167 authors in optometry is presented confirming these results and we also show that the three different author rankings (according to h-index h_1, according to number of publications and according to number of citations) have a highly positive correlation value.

It is hard to constitute author data sets as described above. Yet it would be very interesting to have more data of this kind in order to better understand these three indices and to have the results confirmed on the positive correlation between the different rankings.

The first time – we think – that the h-index has been applied to non-citation data was in Liu and Rousseau (2007), where the h-index was applied to library circulation data. To the best of our knowledge, this is the second time that the h-index has been applied to non-citation data (such as for h_P) in a practical case (that the h-index could be defined in a general theoretical IPP setting was already noticed in Egghe and Rousseau (2006)). It is intriguing to explore these simple applications of the h-index to other IPP types. Examples could include calculating the h-index based on papers vs. downloads (replacing citations) or calculating the h-index of websites based on their (in-) links. One could even think of applying the h-index to econometric topics (e.g. calculating the h-index to social groups (e.g. countries, companies, …) based on their rankings according to wealth or income.

Although it is not clear at the moment which of the h-indices (h_2, h_P or h_C) should be preferred, it is clear that h_P, the only h-index not based on citation data, is the easiest to calculate since it only uses author publication data.

References

P. Ball (2005). Index aims for fair ranking of scientists. Nature, 436, 900.

P. Ball (2007). Achievement index climbs the ranks. Nature 448, 737, 2007.

M.G. Banks (2006). An extension of the Hirsch index: indexing scientific topics and compounds. Scientometrics 69(1), 161-168, 2006.

J. Bar-Ilan (2006). H-index for Price medallists revisited. ISSI Newsletter 2(1), 3-5, 2006.

W. Barendse (2007). The strike rate index: a new index for journal quality based on journal size and the h-index of citations. Biomedical Digital Libraries 4(3), 2007. http://www.bio-diglib.com/content/4/1/3

J. Beirlant, W. Glänzel, A. Carbonez and H. Leemans (2007). Scoring research output using statistical quantile plotting. Journal of Informetrics 1(3), 185-192, 2007.

L. Bornmann and H.-D. Daniel (2005). Does the h-index for ranking of scientists really work? Scientometrics, 65(3), 391-392.

L. Bornmann and H.-D. Daniel (2007a). What do we know about the h-index ? Journal of the American Society for Information Science and Technology 58(9), 1381-1385, 2007.

L. Bornmann and H.-D. Daniel (2007b). Convergent validation of peer review decisions using the h-index. Extent of and reasons for type I and type II errors. Journal of Informetrics 1(3), 204-213, 2007.

T. Braun, W. Glänzel and A. Schubert (2005). A Hirsch-type index for journals. The Scientist, 19(22), 8.

T. Braun, W. Glänzel and A. Schubert (2006). A Hirsch-type index for journals. Scientometrics 69(1), 169-173.

Q.L. Burrell (2007a). Hirsch's h-index: a stochastic model. Journal of Informetrics 1(1), 16-25, 2007.

Q.L. Burrell (2007b). On the h-index, the size of the Hirsch core and Jin's A-index. Journal of Informetrics 1(2), 170-177, 2007.

R. Costas and M. Bordons (2007). The h-index: advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics 1(3), 193-203, 2007.

L. Egghe (2005). Power Laws in the Information Process: Lotkaian Informetrics. Elsevier, Oxford (UK).

L. Egghe (2007a). The influence of transformations on the h-index and the g-index. Preprint.

L. Egghe (2007b). Modelling successive h-indices. Scientometrics, to appear.

L. Egghe (2007c). Dynamic h-index: the Hirsch index in function of time. Journal of the American Society for Information Science and Technology 58(3), 452-454, 2007.

L. Egghe and I.K. Ravichandra Rao (2007). The influence of the query formulation of a topic on its h-index: models and examples of the h-index of N-grams. Preprint.

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch-index. Scientometrics 69(1), 121-129.

W. Glänzel (2006a). On the opportunities and limitations of the h-index. Science Focus 1(1), 10-11 (in Chinese).

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

W. Glänzel and O. Persson (2005). H-index for Price medallists. ISSI Newsletter 1(4), 15-18, 2005.

J.E. Hirsch (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, 16569-16572.

J.E. Hirsch (2007). Does the h-index have predictive power ? arXiv:0708.0646v1 [physics.soc-ph] 6 Aug 2007.

Y. Liu and R. Rousseau (2007). Hirsch-type indices and library management: the case of Tongji university library. In: Proceedings of ISSI 2007. CSIC, Madrid, Spain, 514-522.

C.W. Miller (2006). Superiority of the h-index over the impact factor for physics. arXiv:physics/0608183v1 17 Aug 2006.

C. Oppenheim (2007). Using the h-index to rank influential British researchers in information science and librarianship. Journal of the American Society for Information Science and Technology 58(2), 297-301, 2007.

S.B. Popov (2005). A parameter to quantify dynamics of a researcher's scientific activity. arXiv:physics/0508113.

G. Prathap (2006). Hirsch-type indices for ranking institutions' scientific research output. Current Science 91(11), 1439, 2006.

I.K. Ravichandra Rao and R. Rousseau (2007). Empirical series of journal h-indices. Preprint.

R. Rousseau (2007). The influence of missing publications on the Hirsch index. Journal of Informetrics 1(1), 2-7, 2007.

G. Saad (2006). Exploring the h-index at the author and journal levels using bibliometric data of productive consumer scholars and business-related journals respectively. Scientometrics 69(1), 117-120, 2006.

A. Schubert (2007). Successive h-indices. Scientometrics 70(1), 201-205.

A. Schubert and W. Glänzel (2007). A systematic analysis of Hirsch-type indices for journals. Journal of Informetrics 1(3), 179-184, 2007.

The STIMULATE6 Group (2007). The Hirsch index applied to topics of interest to developing countries. First Monday 12(2), 2007. http://www.firstmonday.org/issues/issues/issues12 2/stimulate/

A.F.J. van Raan (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. Scientometrics 67(3), 491-502.

J.K. Vanclay (2007). On the robustness of the h-index. Journal of the American Society for Information Science and Technology 58(10), 1547-1550, 2007.

P. Vinkler (2007). Eminence of scientists in the light of the h-index and other scientometric indicators. Journal of Information Science 33(4), 481-491, 2007.

J. Wan, P. Hua and R. Rousseau (2007). The pure h-index: calculating an author's h-index by taking co-authors into account. Preprint.

Appendix

Table 1. Ranks according to h-index values, number of papers and number of citations

author code	rank-h	rank-p	rank-c	author code	rank-h	rank-p	rank-c
				a43	91	91	118
a1	123	49	129	a44	47	67	73
a2	5	17	1	a45	92	129	110
a3	2	3	4	a46	137	130	138
a4	86	74	102	a47	93	106	68
a5	14	5	36	a48	94	78	125
a6	124	146	95	a49	164	154	164
a7	15	26	9	a50	138	107	146
a8	125	123	141	a51	95	108	120
a9	87	124	58	a52	67	21	94
a10	126	102	142	a53	96	79	82
a11	34	43	12	a54	139	131	147
a12	28	27	49	a55	68	52	119
a13	88	125	124	a56	97	80	132
a14	127	147	96	a57	98	92	100
a15	41	75	14	a58	140	155	158
a16	42	50	3	a59	69	93	27
a17	128	103	130	a60	37	45	63
a18	129	104	99	a61	21	39	6
a19	130	148	88	a62	99	109	69
a20	20	36	34	a63	70	94	89
a21	64	88	57	a64	71	11	92
a22	131	149	54	a65	72	110	107
a23	162	126	162	a66	48	46	42
a24	8	9	17	a67	100	82	111
a25	43	89	65	a68	141	156	133
a26	163	150	163	a69	101	132	80
a27	35	51	52	a70	142	111	159
a28	89	105	143	a71	102	133	109
a29	44	90	30	a72	29	59	19
a30	45	44	5	a73	30	68	32
a31	26	28	51	a74	73	112	47
a32	65	65	114	a75	143	134	116
a33	132	127	144	a76	38	82	15
a34	46	76	71	a77	103	135	74
a35	33	151	131	a78	144	157	77
a36	66	66	97	a79	49	22	79
a37	134	152	145	a80	145	136	148
a38	135	128	152	a81	104	53	121
a39	16	20	35	a82	146	137	87
a40	4	5	8	a83	22	37	16
a41	136	153	155	a84	105	113	134
a42	90	77	115	a85	106	114	117

a86	74	95	93	a127	156	161	98
a87	23	40	21	a128	115	61	140
a88	75	60	72	a129	116	85	126
a89	147	158	157	a130	13	24	18
a90	107	138	112	a131	152	116	160
a91	12	18	33	a132	153	162	150
a92	50	41	82	a133	117	117	127
a93	51	69	64	a134	25	34	26
a94	52	30	75	a135	7	19	10
a95	17	7	31	a136	81	54	103
a96	24	13	50	a137	154	163	78
a97	18	6	28	a138	57	62	22
a98	76	83	83	a139	19	14	44
a99	1	2	2	a140	26	35	55
a100	148	139	149	a141	155	143	161
a101	108	96	101	a142	118	100	106
a102	9	8	23	a143	82	118	48
a103	109	140	86	a144	83	119	108
a104	77	70	139	a145	58	63	56
a105	149	159	153	a146	156	120	151
a106	53	15	84	a147	119	101	128
a107	110	47	135	a148	59	55	85
a108	3	1	7	a149	39	56	45
a109	78	32	105	a150	40	57	29
a110	6	4	24	a151	84	86	41
a111	111	84	61	a152	60	87	46
a112	165	97	165	a153	157	164	156
a113	54	48	76	a154	61	72	70
a114	112	141	39	a155	167	165	167
a115	55	38	67	a156	158	144	104
a116	56	31	91	a157	11	29	11
a117	113	98	122	a158	159	166	113
a118	114	142	123	a159	120	73	136
a119	31	71	40	a160	27	33	20
a120	150	99	154	a161	160	167	38
a121	166	160	166	a162	62	25	66
a122	32	42	60	a163	85	88	37
a123	79	115	53	a164	121	145	59
a124	10	16	13	a165	63	64	90
a125	80	23	43	a166	161	121	137
a126	33	12	25	a167	122	122	62