Fractionalization of *h*-index for multiple authorship – an impact-based interpretation conserving counts

Gangan Prathap*

A. P. J. Abdul Kalam Technological University,
Thiruvananthapuram 695 016, India

The *h*-index can be fractionalized to take into account multiple authorship. We discuss the problems associated with fractionalization and point out that only one method satisfies the count conservation rule. We illustrate with examples taking care to use a subtle interpretation based on specific impact and not citations.

Keywords: Bibliometrics, count conservation rule, fractional counting, *h*-index.

THE Hirsch index (*h*-index) combines impact (quality) with productivity (size or quantity) into a single number as a bibliometric indicator of scholarly performance when a citation distribution is given for a publication set[1]. It has now become overwhelming popular as a performance indicator[2]. The *h*-index is found from a particular heuristic construction that accounts for productivity (quantity or size), namely the number of papers *P*, and quality. Although initially quality was equated to impact as measured in an overall sense as the total number of citations, we emphasize here that this should be measured by the specific impact *i* considering the values of the citations of the individual papers c_k in the sequence $k = 1$ to *P* arranged in a monotonically decreasing order of citations.

An early effort to provide a mathematical framework for the Hirsch index assumed a standard Lotka model for citation distribution[3]. Egghe and Rousseau[4,5] later modified this framework by introducing the shifted Lotka model to make allowance for uncited papers. Burrell[6] showed that a simple Lotka/Pareto-like model could give misleading results as the formulae actually gave similar results whether or not the uncited papers are included, and severely underestimated the empirically estimated *h*-index. Note that all the indices, *h*, *i*, c_k and *P*, have the units and dimensions of *P* as proposed by Prathap[7]. Also, all the original indices are based on whole counting and do not recognize that most papers have multiple authorship. Here, we shall examine a consistent protocol for setting up the fractionalized *h*-index that adheres to the count conservation rule[8]. After fractionalization, the individual papers are still arranged in the same original sequence of a monotonically decreasing order of impact, taking into account the fact that impact is an intensive property of each paper that does not change with fractionalization. We use three case studies from the published literature to illustrate this interpretation.

The *h*-index, as originally introduced, used whole counting, i.e. publications and citations were assigned fully to each author contributing to the paper. This is because the procedure to compute *h*, which was performed by arranging citations in descending order according to rank, does not take into account the fact of multiple authorship[9] and this shortcoming was already anticipated in Hirsch's original proposal[1].

An early protocol for fractionalizing or individualizing the *h*-index[9] was intended to correct for disciplinary differences[9]. Batista *et al.*[9] used the mean number of authors of the papers in the *h*-core as the factor with which to fractionalize the *h*-index, and obtained a fractional value that accounts for multiple authorship, i.e. the h_I-index was obtained by dividing the *h*-index by the average number of authors in the *h*-core set. The argument for this was that co-authorship allows academics to write more papers and at the same time increase citations to these papers[10], and that the publication practices of different disciplines promote different patterns of multiple authorship. There is no conservation rule adhered to here.

Harzing[11] introduced the h_I, norm by first normalizing citations for each paper by dividing the number of citations by the number of authors for that paper, and only then calculate the *h*-index of the normalized citation counts. This is a fractionalized version of the *h*-index, where only citations are normalized according to the number of authors. Here, while there is conservation of citation counts, the count of papers is not conserved.

Recently, Hirsch[12] reopened the discussion on multi-authorship by proposing a h_α-index, where the α person is the dominant author among all the co-authors. A high *h*-index in conjunction with a high h_α/h ratio is a hallmark of scientific leadership. The discussion was on establishing an index to measure leadership and not on ensuring count conservation. This prompted Tietze *et al.*[13] to revisit the Galam conservation rule[8] to credit papers fractionally to a single author in order to test early career achievement or scientific leadership.

Schreiber[14,15], Egghe[16] and Galam[8] variously defined indices h_m, fractional *h* and *gh* based on fractional credit allocation in multi-authored papers. Many of the methods in the literature on this topic relate to different ways to allocate credit to co-authors of a multi-authored paper, rather than to ensure that in the process multiple counting does not inflate the count. Galam[8] was the first to insist that any quantitative modification must keep the number of published papers and the total count of citations invariant under multiple authorship, i.e. when fractional allocations are attributed to each co-author, the summation must equal one. This is analogous to the various conservation principles on which physics is founded.

For the purpose of this study, we shall focus attention on the Schreiber[14,15] and Galam[8] schemes. Schreiber[14,15]

*e-mail: gangan_prathap@hotmail.com

Table 1. Illustration of computation of fractional value of h-index for dataset V from table 1 of Schreiber[15]

Authors	Whole counting			Fractional counting			
a_k	k	c_k	i_k	n_{Fk}	N_{Fk}	c_{Fk}	i_{Fk}
3	1	79	79	0.333	0.33	26.33	79
4	2	34	34	0.250	0.58	8.50	34
4	3	32	32	0.250	0.83	8.00	32
2	4	25	25	0.500	1.33	12.50	25
4	5	16	16	0.250	1.58	4.00	16
4	6	13	13	0.250	1.83	3.25	13
10	7	12	12	0.100	1.93	1.20	12
2	8	11	11	0.500	2.43	5.50	11
3	9	11	11	0.333	2.77	3.67	11
3	**10**	11	11	0.333	3.10	3.67	11
3	11	8	8	0.333	3.43	2.67	8
1	12	8	8	1.000	4.43	8.00	8
2	13	8	8	0.500	4.93	4.00	8
4	14	8	8	0.250	5.18	2.00	8
2	15	7	7	0.500	5.68	3.50	7
1	16	7	7	1.000	**6.68**	7.00	**7**
2	17	6	6	0.500	7.18	3.00	6
2	18	6	6	0.500	7.68	3.00	6
2	19	5	5	0.500	8.18	2.50	5
3	20	5	5	0.333	8.52	1.67	5

a_k, Number of authors of paper at kth rank; c_k as well as i_k, Number of citations of the paper at the kth rank; n_{Fk}, Effective fractional count of the kth paper; N_{Fk}, Cumulative count up to k papers; c_{Fk}, Fractional count of citations from the kth paper; i_{Fk} Which is the specific fractional impact will be the same as c_k and i_k.

proposed an approach whereby each paper is counted fractionally according to the inverse of the number of co-authors. Thus, papers are fractionalized and citations are then proportionately accounted for, i.e. fractionalized. The ranking scheme needed to compute the h-index now depends on the original unfractionalized citations, in other words, on the original impact value of each paper in the h-core. Egghe[16] pointed out that either citations or papers can be counted in a fractional manner to take into account the number of co-authors, and this would lead to two ranking schemes and thus to two values of fractional h-indices. Chai *et al.*[17] also devised a scheme to allocate partial credit to each co-author of a paper. We see from the above that there is some confusion about the protocol for fractionalization – should papers be fractionalized, or citations, or both? The confusion arises from the original definition of the h-index, as the highest number h of papers of a scientist that have been cited h or more times. By implication, the construction for h is performed by arranging citations in descending order according to rank and displayed graphically with citations on the y-axis and rank of papers on the x-axis. That is, a paper at rank k that has c_k citations is displayed by a bar of unit width and a height c_k. The h-index is then read-off this sequence as

$$c_h \geq h \geq c_{h+1}.$$

As long as whole counting is used, there is no problem – each contributing author to the paper placed at rank k is given full credit for authorship and assigned all the citations c_k. It is important to note here that in whole counting, the impact of the kth paper and the citations it receives are identical, i.e. $i_k = c_k$. Assume now that this paper at rank k has a_k authors. Then the author is given a fractional credit to $1/a_k$ papers and also to c_k/a_k citations. In this manner, the count of papers and the count of citations is conserved. Further, the fractionalized impact remains $i_k = c_k$. That is, impact is an intensive property that cannot be fractionalized, while papers and citations are extensive properties that are fractionalized. This is fully consistent with Schreiber's protocol[14,15]. It will therefore be more meaningful, in the context of fractionalization, to read Schreiber's h_m-index using the logic that it is the largest number of effective papers h_m for which h_m is larger than the impact at that rank, using the definition of effective number of papers.

We illustrate the count conserving protocol by applying it directly to the data in Table 1 based on data set V from table 1 of Schreiber[15]. Let c_k, $k = 1$ to P, represent the citation sequence of all P papers from a publication set belonging to an author V. Let a_k be the number of authors for a paper at the kth rank. At the kth rank, the author has an effective share $n_{Fk} = 1/a_k$ of the paper and $c_{Fk} = c_k/a_k$ share of the citations for that paper. The fractionalized impact, i_{Fk} is the same as the original impact i_k, confirming that impact is an intensive property that cannot be fractionalized. Up to the kth rank, the effective number of papers is $N_{Fk} = \Sigma n_k$.

Table 2. Illustration of computation of fractional value of h-index for a dataset from table 5 of Galam[8]

Authors	Whole counting			Fractional counting			
a_k	k	c_k	i_k	n_{Fk}	N_{Fk}	c_{Fk}	i_{Fk}
2	1	187	187	0.500	0.50	93.50	187
1	2	181	181	1.000	1.50	181.00	181
3	3	179	179	0.333	1.83	59.67	179
1	4	145	145	1.000	2.83	145.00	145
1	5	145	145	1.000	3.83	145.00	145
3	6	132	132	0.333	4.17	44.00	132
1	7	132	132	1.000	5.17	132.00	132
3	8	120	120	0.333	5.50	40.00	120
2	9	104	104	0.500	6.00	52.00	104
3	10	98	98	0.333	6.33	32.67	98
2	11	94	94	0.500	6.83	47.00	94
3	12	90	90	0.333	7.17	30.00	90
3	13	81	81	0.333	7.50	27.00	81
1	14	75	75	1.000	8.50	75.00	75
2	15	72	72	0.500	9.00	36.00	72
3	16	71	71	0.333	9.33	23.67	71
3	17	68	68	0.333	9.67	22.67	68
3	18	66	66	0.333	10.00	22.00	66
2	19	63	63	0.500	10.50	31.50	63
3	20	55	55	0.333	10.83	18.33	55
1	21	51	51	1.000	11.83	51.00	51
2	22	50	50	0.500	12.33	25.00	50
2	23	48	48	0.500	12.83	24.00	48
1	24	45	45	1.000	13.83	45.00	45
1	25	43	43	1.000	14.83	43.00	43
2	26	42	42	0.500	15.33	21.00	42
1	27	39	39	1.000	16.33	39.00	39
3	28	38	38	0.333	16.67	12.67	38
2	29	38	38	0.500	17.17	19.00	38
2	30	35	35	0.500	17.67	17.50	35
2	31	35	35	0.500	18.17	17.50	35
2	32	34	34	0.500	18.67	17.00	34
6	33	33	33	0.167	18.83	5.50	33
2	34	31	31	0.500	19.33	15.50	31
3	35	30	30	0.333	19.67	10.00	30
2	36	30	30	0.500	20.17	15.00	30
3	37	30	30	0.333	20.50	10.00	30
2	38	30	30	0.500	21.00	15.00	30
2	39	29	29	0.500	21.50	14.50	29
2	40	29	29	0.500	22.00	14.50	29

Table 3. Publication and citation details of five authors: V, W, X, Y and Z from example 2 and table 1 of Wan *et al.*[18]

Papers s	Authors a_s	Citations c_s	Authors
1	1	10	V
2	2	2	V, W
3	2	1	W, X
4	1	5	V
5	1	2	W
6	3	1	X, Y, Z
7	3	2	X, Y, Z
8	2	2	V, Y
9	3	30	W, X, Z

As a first case study, we use dataset V from table 1 of Schreiber[15]. Note that citation records are available for only 20 most cited papers and hence the fractionalized index is calculated based on this restriction. In the case shown here (dataset for V of table 1 of Schreiber[15]), we see that the fractional values are smaller than the whole counting values. Also, the citations after fractionalization need not be rearranged in a descending fashion as it is the impact, which is an intensive property, which is used to ensure decreasing monotonicity. The h-indices are computed in the fashion recommended by Schreiber[14,15], as this is the only protocol which is consistent with the fractionalization methodology used in this study. We also see from Table 1 that it is more meaningful, in the context of fractionalization, to read the h-index off the impact sequence rather than the citation sequence. The fractional h-index is obtained as the value h_m, which is the effective number of papers which has an impact equal to or greater than h_m (ref. 15). The fractional h-index is 6.68 instead of

Table 4. Illustration of computation of fractional value of *h*-index for a dataset from example 2 and Table 1 of Wan *et al.*[18]

Rank	Paper	*V*-whole					Rank	Paper	*V*-fractional				
k	s	n_k	c_k	i_k	N_k	C_k	k	s	n_{Fk}	c_{Fk}	i_{Fk}	N_{Fk}	C_{Fk}
1	1	1	10	10	1	10	1	1	1	10	10	1	10
2	4	1	5	5	2	15	2	4	1	5	5	2	15
3	2	1	2	2	3	17	3	2	0.5	1	2	2.5	16
4	8	1	2	2	4	19	4	8	0.5	1	2	3	17
$h = 2$							$h_f = 2$						

Rank	Paper	*W*-whole					Rank	Paper	*W*-fractional				
k	s	n_k	c_k	i_k	N_k	C_k	k	s	n_{Fk}	c_{Fk}	i_{Fk}	N_{Fk}	C_{Fk}
1	9	1	30	30	1	30	1	9	0.333	10	30	0.333	10
2	2	1	2	2	2	32	2	2	0.5	1	2	0.833	11
3	3	1	1	1	3	33	3	3	0.5	0.5	1	1.333	11.5
$h = 2$						$h_f = 0.833$							

Rank	Paper	*X*-whole					Rank	Paper	*X*-fractional				
k	s	n_k	c_k	i_k	N_k	C_k	k	s	n_{Fk}	c_{Fk}	i_{Fk}	N_{Fk}	C_{Fk}
1	9	1	30	30	1	30	1	9	0.333	10	30	0.333	10
2	7	1	2	2	2	32	2	7	0.333	0.667	2	0.667	10.67
3	3	1	1	1	3	33	3	3	0.5	0.5	1	1.167	11.17
4	6	1	1	1	4	34	4	6	0.333	0.333	1	1.5	11.5
$h = 2$							$h_f = 0.667$						

Rank	Paper	*Y*-whole					Rank	Paper	*Y*-fractional				
k	s	n_k	c_k	i_k	N_k	C_k	k	s	n_{Fk}	c_{Fk}	i_{Fk}	N_{Fk}	C_{Fk}
1	7	1	2	2	1	2	1	7	0.333	0.667	2	0.333	0.667
2	8	1	2	2	2	4	2	8	0.5	1	2	0.833	1.667
3	6	1	1	1	3	5	3	6	0.333	0.333	1	1.167	2
$h = 2$							$h_f = 0.833$						

Rank	Paper	*Z*-whole					Rank	Paper	*Z*-fractional				
k	s	n_k	c_k	i_k	N_k	C_k	k	s	n_{Fk}	c_{Fk}	i_{Fk}	N_{Fk}	C_{Fk}
1	9	1	30	30	1	30	1	9	0.333	10	30	0.333	10
2	5	1	2	2	2	32	2	5	1	2	2	1.333	12
3	7	1	2	2	3	34	3	7	0.333	0.667	2	1.667	12.67
4	6	1	1	1	4	35	4	6	0.333	0.333	1	2	13
$h = 2$							$h_f = 1.667$						
Check for conservation of counts			18	126			Check for conservation of counts					9	55

a whole counted value of 10. Figure 1 shows graphically how the construction heuristic works in this case.

As a second case study, we use a dataset from table 5 of Galam (Table 2)[8]. Now, citation records are available for only 40 most cited papers and the fractionalized index is calculated based on this restriction of the dataset. In this case we see again that the fractional values are smaller than the whole counting values. Again, after fractionalization, the fractionalized citations need not be rearranged in a descending fashion as the monotonicity is determined by the impact and this does not change as it is an intensive property. The *h*-indices are computed in the fashion recommended by Schreiber[14,15]. We read the *h*-index directly off the impact sequence. Because of the unavailability of data beyond 40 records, the fractional *h*-index based on an egalitarian sharing is definitely greater than 22.00, which is the fractionalized total count of articles, instead of a whole counted value of 33. Galam[8] used various non-egalitarian schemes and instead of an *h*-index value of 33, found $gh(2/3) = 21$, $gh(3/4) = 19$, $gh(1/2) = 23$, $gh(0) = 20$. However, it is to be noted that Galam rearranged the fractionalized citations in descending order as the *gh*-indices were read off against the fractionalized citations and not the impact at that rank.

The total number of articles for the author was 19.91, 18.94, 22.31, and 19.13 respectively, instead of the inflated value of 40. In our egalitarian scheme, the total count of articles was 22.00. Figure 2 shows graphically how the construction heuristic works in this case.

As a final case study, we take the full publication and citation details of five authors: *V*, *W*, *X*, *Y* and *Z* from example 2 and table 1 of Wan *et al.*[18]. These five authors have published nine unique papers (numbered using the index $s = 1$ to 9) for a total of 55 citations and Table 3 collects the summary statistics. For each paper, a_s and c_s are the number of authors and citations respectively. Table 4 illustrates computation of fractional value of the *h*-index for the five authors. If fractionalization had not been adopted, the *h*-indices for all five authors are identically equal to 2, and in the process the count of the number of papers and citations has been inflated to 18 and 126 respectively. Instead, if fractional counting is used, there is complete conservation of the counts of papers and citations and the fractional *h*-indices are 2, 0.833, 0.667, 0.833 and 1.667 respectively.

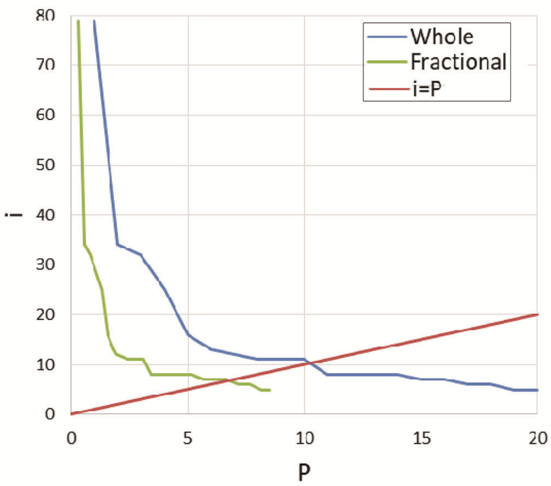

Figure 1. Heuristic construction of the original *h*-index and fractional value of *h*-index for dataset V from table 1 of Schreiber[15].

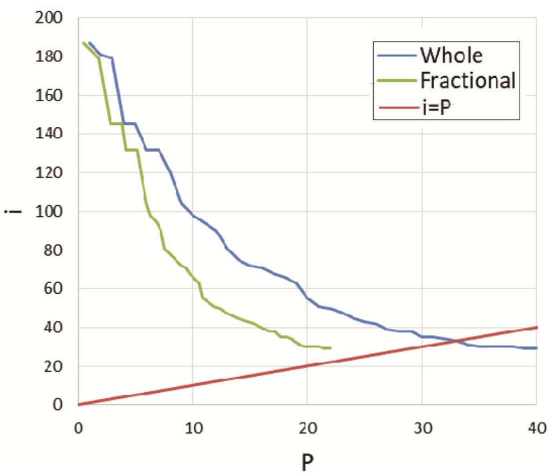

Figure 2. Heuristic construction of the original *h*-index and the fractional value of *h*-index for the dataset from table 5 of Galam[8].

Many approaches for fractionalizing the *h*-index taking into account multiple authorship have been proposed. We see that it is more meaningful, in the context of fractionalization, to read the *h*-index off the impact sequence, rather than the citation sequence as the former is based on an intensive property that does not change with fractionalization. The fractional *h*-index is obtained as the value h_m, which is that largest value of the effective number of papers which has an impact equal to or greater than h_m (ref. 15). We have demonstrated the procedure with three examples taken from the published literature.

1. Hirsch, J. E., An index to quantify an individual's scientific research output. *Proc. Natl. Acad. Sci. of USA*, 2005, **102**(46), 16569–16572.
2. Bornmann, L. and Marx, W., The *h*-index as a research performance indicator. *Eur. Sci. Edit.*, 2011, **37**(3), 77–80.
3. Egghe, L. and Rousseau, R., An informetric model for the Hirsch index. *Scientometrics*, 2006, **69**(1), 121–129.
4. Egghe, L. and Rousseau, R., Theory and practice of the shifted Lotka function. *Scientometrics*, 2012, **91**(1), 295–301.
5. Egghe, L. and Rousseau, R., The Hirsch index of a shifted Lotka function and its relation with the impact factor. *J. Am. Soc. Infor. Sci. Technol.*, 2012, **63**(5), 1048–1053.
6. Burrell, Q. P., Formulae for the *h*-Index: a lack of robustness in Lotkaian informetrics? *J. Am. Soc. Inf. Sci. Technol.*, 2013, **64**(7), 1504–1514; doi:10.1002/asi.22845.
7. Prathap, G., Eugene Garfield: from the metrics of science to the science of metrics. *Scientometrics*, 2018, **114**(2), 637–650.
8. Galam, S., Tailor based allocations for multiple authorship: a fractional g*h*-index. *Scientometrics*, 2011, **89**(1), 365–379.
9. Batista, P. D., Campiteli, M. G., Kinouchi, O. and Martinez, A. S., Is it possible to compare researchers with different scientific interests? *Scientometrics*, 2006, **68**(1), 179–189.
10. Glänzel, W. and Thijs, B., Does co-authorship inflate the share of self-citations? *Scientometrics*, 2004, **61**(3), 395–404.
11. Harzing, A. W., Publish or Perish, 2007; http://www.harzing.com/pop.htm
12. Hirsch, J. E., h_α: an index to quantify an individual's scientific leadership. *Scientometrics*, 2019, **118**, 10.1007/s11192-018-2994-1.
13. Tietze, A., Galam, S. and Hofmann, P., Crediting multi-authored papers to single authors, 2019, arXiv:1905.01943v1.
14. Schreiber, M., A modification of the *h*-index: The h_m-index accounts for multi-authored manuscripts. *J. Informetri.*, 2008, **2**(3), 211–216.
15. Schreiber, M., A case study of the modified Hirsch index h_m accounting for multiple coauthors. *J. Am. Soc. Infor. Sci. Technol.*, 2009, **60**, 1274–1282.
16. Egghe, L., Mathematical theory of the *h*- and *g*-index in case of fractional counting of authorship. *J. Am. Soc. Infor. Sci. Technol.*, 2008, **59**, 608–1616.
17. Chai, J. C., Hua, P. H., Rousseau, R. and Wan, J. K., Real and rational variants of the *h*-index and the *g*-index. In Proceedings of the WIS (eds Kretschmer, H. and Havemann, F.), 2008, vol. 64, p. 71.
18. Wan, J. K., Hua, P. H. and Rousseau, R., The pure *h*-index: calculating an author's *h*-index by taking co-authors into account. 2007; http://eprints.rclis.org/10376/1/pure_h.pdf

Received 22 June 2019; accepted 26 November 2019

doi: 10.18520/cs/v118/i6/961-965