The pure *h*-index: calculating an author's *h*- index by taking co-authors into account

Jin-kun WAN [1,2], Ping-huan HUA[1,2] and Ronald ROUSSEAU [3]

[1] Library of Tsinghua University, Beijing,100084,China
E-mail: wanjk@lib.tsinghua.edu.cn
[2] China Scientometrics and Bibliometrics Research Center
P O BOX 84-48, Tsinghua University ,Beijing,100084,China
E-mail: gfhs@cnki.net
[3] KHBO (Association K.U.Leuven), Industrial sciences and Technology, B-8400 Oostende, Belgium
E-mail: ronald.rousseau@khbo.be

Abstract
We introduce a new Hirsch-type index for a scientist. This so-called pure *h*-index, denoted by h_P, takes the actual number of co-authors, and the scientist's relative position in the byline into account. The transformation from *h* to h_P can also be applied to the *R*-index, leading to the pure *R*-index, denoted as R_P. This index takes the number of collaborators, possibly the rank in the byline and the actual number of citations into account.

Introduction

The *h* index proposed by J. E. Hirsch (Hirsch, 2005) combines productivity with impact. In this article we will not discuss advantages and disadvantages, see e.g. (Glänzel, Jin *et al*., 2007) for this aspect, but will propose an adaptation of the original proposal. This adaptation takes the number of co-authors into account.

Recall that, when a researcher's articles are ranked according to the number of citations received, his or her Hirsch index is *h* if *h* is the highest rank (largest natural number) such that the first *h* publications received each at least *h* citations. The Hirsch core is the set consisting of the first *h* publications, where, in case of ties, a choice has to be made. In this article preference is given to articles with the least number of authors. In other situations preference has been given to the most recent articles (Jin, 2007; Jin *et al*., 2007). The Hirsch core of author A will be denoted by H(A).

Papers belonging to a scientist's Hirsch core may be written by this author as a single author or in collaboration with colleagues. The question we want to study in this note is: how can the *h*-index be adapted in order to take account of the number of collaborators? Indeed, it goes without saying that it is much easier to get a high *h*-index when one has written many papers with many collaborators. We will moreover take an author's rank in the byline into account and propose a new index, denoted as h_P, for evaluating the so-called pure contribution of a given author.

The idea of taken the number of co-authors into account has already been considered by Batista *et al*. (2006). They simply divide *h* by the average number of researchers in the publications of the Hirsch core. Quentin Burrell (2007) proposes to discount the h-index for career length, multi-authorship and self-citations. He notes that if discounting is performed before the determination of the Hirsch core this core itself can be reduced. This is one possible approach. We will take another approach by first determining the *h*-index and Hirsch core in the usual way, and then determining a complementary index. Egghe (2007) presents a mathematical theory of the *h*-index (and also of the *g*-index) in case of fractional counting (see next section for a definition). He considers fractional counting of citations as well as fractional counting of publications.

Methods for accrediting publications to authors

In this section we present a short overview of some scoring methods (Egghe *et al*., 2000). The number of co-authors of an article is denoted by *N*. The term 'normalized score' is used to indicate that the sum of the scores of all co-authors is equal to one.

(1) First-author counting (Cole & Cole, 1973)

Only the first of the *N* authors of a paper receives a credit equal to one. The other authors do not receive any credit. This method is also known as straight counting. It has been argued, again and again, that this is not an acceptable method for assigning credits to authors (Lindsey, 1980).

(2) Total counting

Here, each of the *N* authors receives one credit. This counting method is also called normal, or standard counting.

(3) Fractional counting (Price, 1981; Oppenheim, 1998)

Now, each of the *N* authors receives a score equal to 1/*N*. This counting method is sometimes called adjusted counting. Fractional counting has been studied e.g. in (Burrell and Rousseau, 1995; Van Hooydonk, 1997).

(4) Proportional or arithmetic counting (Van Hooydonk, 1997)

If an author has rank *R* in the author list of an article with *N* collaborators (*R* = 1, …, *N*), then she/he receives a score of *N*+1-*R*. This score can be normalized in such a way that the total score of all authors is equal to 1. In this normalized version the score is: $\frac{2}{N}\left(1-\frac{R}{N+1}\right)$.

(5) Geometric counting (Egghe *et al*., 2000)

If an author has rank *R* in an article with *N* co-authors (*R* = 1,…,*N*) then she/he receives a credit of 2^{N-R}. In its normalized version this score becomes $\frac{2^{N-R}}{2^N-1}$.

(6) Noblesse oblige, cf. (Zuckerman, 1968)

In this approach it is assumed that the most important author closes the list. She/he receives a credit of 0.5, while the other *N*-1 authors receive a credit of 1/(2(*N*-1)) each (this is but one suggestion, among many more that are possible here). Clearly, this concept makes only sense if an article has more than one author. In the case of one author this counting method assigns a score of one to the single author.

We note that methods (4), (5), (6) assume that the rank of the authors in the byline accurately reflects their contribution. If authors adapt alphabetical ordering, or take turns in being first and second author, these counting schemes may not be applied.

The co-author adapted h-index, based on the concept of the equivalent number of co-authors

In the previous list of scoring methods, only total counting is not normalized. This method will not be used further as our approach is based on normalized scores. Also first-author counting will not be considered further. We will now introduce the concepts leading to the definition of an *h*-index representing the so-called pure contribution of an author.

Definition: the equivalent number of co-authors of author A in document D.

This concept, denoted by N_E (A,D) is defined as $\frac{1}{S(A_D)}$, where $S(A_D)$ denotes the normalized score of author A in document D.

Clearly, N_E(A,D) is at least equal to 1. It has no theoretical upper limit. For a single-authored article N_E(A,D) is always equal to 1. When using fractional counting N_E(A,D) is always equal to N, the actual number of co-authors of the article. For proportional counting N_E(A,D) = $\frac{N(N+1)}{2(N+1-R)}$. This value lies between (N+1)/2 (for rank 1) and N(N+1)/2 (for rank N). In the case of geometric counting N_E(A,D) = $\frac{2^N-1}{2^{N-R}}$. This values lies between $\frac{2^N-1}{2^{N-1}}$ (rank 1, which is about 2 for N large) and 2^N - 1 (rank N). Finally in the case of noblesse oblige the most important author (closing the list; and assuming we are not dealing with a single-authored article) always has an N_E(A,D) equal to 2, while the other authors' N_E(A,D) is 2(N-1). This number is at least equal to 2 (the case of two authors).

Definition: The equivalent Hirsch core average number of authors

The equivalent Hirsch core average number of authors for author A, denoted as E(A) is defined as:

$$E(\mathrm{A}) = \frac{\sum_{\mathrm{D} \in \mathrm{H(A)}} N_{\mathrm{E}}(\mathrm{A,D})}{h} \qquad (1)$$

Definition: The pure or co-author adapted h-index

We define the pure h-index of author A, denoted by h_P(A) as:

$$h_{\mathrm{P}}(A) = \frac{h}{\sqrt{E(\mathrm{A})}} = h\sqrt{\frac{h}{\sum_{\mathrm{D} \in \mathrm{H(A)}} N_{\mathrm{E}}(\mathrm{A,D})}} \qquad (2)$$

Clearly, when author A has written all his/her articles in the Hirsch core as sole author, h(A) = h_P(A). In all other cases h_P(A) < h(A).

Some examples

Example 1

Assume that three authors, A, B and C always publish together and always in the same order, namely B – C – A. Assume further that their *h*-index is equal to *h*. Observe that, because of our assumptions, this *h*-index must be the same for these three authors.

What is their pure *h*-index? If fractional counting is used, their h_P-value is still equal, but it is now reduced to $\frac{h}{\sqrt{3}}$. If arithmetic counting is applied $E(B) = 2$, hence $h_P(B)=\frac{h}{\sqrt{2}}$, $E(C) = 3$, hence $h_P(C)=\frac{h}{\sqrt{3}}$, and $E(A) = 6$, leading to $h_P(A)=\frac{h}{\sqrt{6}}$.

Example 2

Assume that the following Table 1 gives the full publication and citation details of five authors: V, W, X, Y and Z; authors are given in the order they are mentioned in the byline. Table 2 gives the details for the calculation of the pure *h*-index.

Table 1

Authors	V	W-V	W-X	V	Z	X-Y-Z	X-Y-Z	V-Y	X-Z-W
Citations	10	2	1	5	2	1	2	2	30

Besides the data necessary for calculating h and h_P Table 2 also shows the values of these authors' *R*-index, introduced in (Jin *et al*., 2007). The R-index is equal to the square root of the sum of the actual number of citations of articles in the Hirsch core. For author A it is given as shown in formula (3):

$$R(A) = \sqrt{\sum_{D \in H(A)} \mathrm{cit}(A,D)} \qquad (3)$$

Also this index can be divided by the square root of $E(A)$, leading to an index denoted as R_P (last two rows of Table 2). This new indicator is called a pure *R*-index, see formula (4):

$$R_P(A) = \sqrt{\frac{\sum_{D \in H(A)} \text{cit(A,D)}}{E(A)}} \tag{4}$$

Table 2. Calculation of h_P and R_P using fractional and arithmetic counting

Authors	V	W	X	Y	Z
Citations	10 5 2 2	30 2 1	30 2 1 1	2 2 1	30 2 2 1
h-index	2	2	2	2	2
N_E (fract.)	1 1	3 2	3 3	3 2	3 1
N_E (prop.)	1 1	6 1.5	2 2	3 3	3 1
E (fract)	1	2.5	3	2.5	2
E (prop.)	1	3.75	2	3	2
h_P (fract.)	2	1.26	1.15	1.26	1.41
h_P (prop.)	2	1.03	1.41	1.15	1.41
R	3.87	5.66	5.66	2	5.66
R_P (fract)	3.87	3.58	3.27	1.26	4.00
R_P (prop)	3.87	2.92	4.00	1.15	4.00

Note also that, for author Z, we have given preference to the article with the least number of authors (here one).

According to the standard *h*-index, these five authors attain the same score. Table 3 shows the rankings of these five authors, based on the five other methods. These different rankings again illustrate that different counting methods lead to different rankings.

Table 3. Rankings of the five authors of Table 1, according to different *h*-type indices.

Authors	V	W	X	Y	Z
h_P (fract.)	1	3	5	3	2
h_P (prop.)	1	5	2	4	2
R	4	1	1	5	1
R_P (fract)	2	3	4	5	1
R_P (prop)	3	4	1	5	1

The h_P-index, based on fractional counting, ranks these authors as V, followed by Z, then W and Y (tied) and finally X; h_P-index, based on arithmetic counting, ranks these authors as V, followed by X and Z (tied), then Y and finally W. According to the *R*-index authors W, X and Z score equal (5.66 $\approx \sqrt{32}$), followed by authors V and Y, in that order. This result illustrates the (obvious) fact that taking actual citations into account gives a different (in our opinion, better) view on the achievements of these authors. Using the pure *R*-index, an indicator that incorporates also the number of collaborators, leads to an even more refined appreciation.

Additional observations

When fractional counting is used the exact rank occupied by an author does not play any role. Yet, even then our proposal does not coincide with that by Batista *et al.* (2006). We reduce the effect of a large number of authors by taken the square root. In this way, authors are less 'punished' for having collaborated in a mega-authored, highly-cited article.

It is sometimes possible for an author to obtain a higher h_P-value by replacing an article in the Hirsch core by one outside the core but with less collaborators. We propose not to allow this, as we only seek to complement the *h*-index. Moreover, it would make the procedure considerably more difficult, as many combinations would have to be tried in order to find the optimal one. The next example shows that it is indeed possible to increase the h_P-value in this way.

Assume that author T has the following author list

Authors	A-T	A-B-T	T	T
Citations	3	3	2	1

Then $h(T) = 2$, $E(T) = 2.5$, $h_P(T) = 1.265$ and $R_P(T) = 1.55$; using fractional counting. Using arithmetic counting $E(T) = 4.5$, $h_P(T) = 0.94$ and $R_P(T) = 1.15$.

Considering T's publications in the order:

Authors	A-T	T	A-B-T	T
Citations	3	2	3	1

one could say that $h(T)$ is still equal to 2 (this is, of course not the correct way of calculating h), $E(T) = 1.5$ and $h_P(T)$ would be 1.633 > 1.265 (fractional counting); or $E(T) = 2$ and $h_P(T) = 1.41 > 0.94$ (arithmetic counting). This line of approach is usually counterproductive for the calculation of the R_P-index, as the total number of citations is lowered, yet in this example $R_P(T)$ would be 1.83 > 1.55 (fractional counting); and $R_P(T) = 1.58 > 1.15$ (arithmetic counting). As stated before, we do not encourage this calculating method.

Conclusion

We have introduced an adaptation of the *h*-index, which takes the actual number of co-authors and the relative position of an author into account. It is a practical way of discounting the *h*-index as suggested by Burrell (2007). In real applications many authors may have the same *h*-index. Applying a complementary index such as the pure *h*-index introduces a method of discriminating among such authors. The pure *R*-index, denoted as R_P, takes moreover the number of collaborators, possibly the rank in the byline and the actual number of citations into account.

It is well-known (Egghe *et al.*, 1999; Burrell, 2007) that different counting methods lead to different rankings. This is also true in the context of *h*-type indices. Hence, the concrete counting method should be determined (and preferably validated) in advance. When the order of authors in the byline does not reflect the actual contribution then only fractional counting can be applied.

Acknowledgements

The work presented by Ronald Rousseau in this paper was supported by the National Natural Science Foundation of China by grant no. 70673019.

References

Pablo D. Batista, Mônica G. Campiteli, Osame Kinouchi, and Alexandre S. Martinez (2006). “Is it possible to compare researchers with different scientific interests?,” *Scientometrics*, volume 68, number 1, pp.179-189.

Quentin Burrell (2007). “Should the *h*-index be discounted?,” In: W. Glänzel and A. Schubert (editors). *The multidimensional world of Tibor Braun*. Leuven: ISSI, pp. 65-68.

Quentin Burrell, and Ronald Rousseau (1995). “Fractional counts for authorship attribution: a numerical study,” *Journal of the American Society for Information Science*, volume 46, number 2, pp. 97-102.

Jonathan R. Cole, and Stephen Cole (1973). “*Social stratification in science”*. Chicago: The University of Chicago Press.

Leo Egghe (2007). “Mathematical theory of the *h*- and *g*-index in case of fractional counting of authorship”. Preprint

Leo Egghe, Ronald Rousseau, and Guido Van Hooydonk (2000). “Methods for accrediting publications to authors or countries: consequences for evaluation studies," *Journal of the American Society for Information Science*, volume 51, number 2, pp. 145-157.

Wolfgang Glänzel (2006)."On the opportunities and limitations of the *h*-index," *Science Focus*, volume 1, number 1, pp. 10-11 (in Chinese). English version available at E-LIS: ID-code 9535.

Jorge 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,* vol. 102, number 46, pp. 16569-16572.

Bihui Jin (2007). "The *AR*-index: complementing the *h*-index," *ISSI Newsletter*, volume 3, number 1, p.6.

Bihui Jin, Liming Liang, Ronald Rousseau, and Leo Egghe (2007). "The *R*- and *AR*-indices: complementing the *h*-index, " *Chinese Science Bulletin*, volume 52, number 6, pp. 855-863.

Duncan Lindsey (1980). "Production and citation measures in the sociology of science: the problem of multiple authorship," *Social Studies of Science*, volume 10, number 2, pp. 145-162.

Charles Oppenheim (1998). "Fractional counting of multiauthored publications,“ *Journal of the American Society for Information Science,* volume 49, number 5, p. 482.

Derek de Solla Price (1981). "Multiple authorship," *Science*, volume 212, issue 4498, p. 987.

Guido Van Hooydonk (1997). "Fractional counting of multi-authored publications: consequences for the impact of authors, " *Journal of the American Society for Information Science,* volume 48, number 10, pp. 944-945.

Harriet Zuckerman (1968). "Patterns of name-ordering among authors of scientific papers: a study of social symbolism and its ambiguity, " *American Journal of Sociology*, volume 74, number 3, pp. 276-291.