Mathematical results on the H-index and H-sequence of Randić

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium [1]
and
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
leo.egghe@uhasselt.be

ABSTRACT

We present mathematical formulae for the H-sequence and H-index of Randić in a Lotkaian framework. We also present a variant of the H-index of Randić.
We prove that the following assertions are equivalent, given two persons with the same h-index of Hirsch: the Randić H-index H_1 of the first person is larger than H_2, the one of the second person, if and only if the H-sequence of the first person dominates the one of the second person. These properties are equivalent with $\alpha_1 < \alpha_2$ (where $\alpha_i \left(i=1,2\right)$ are the Lotka exponents of the two persons) and also equivalent with $\mu_1 > \mu_2$ (where $\mu_i \left(i=1,2\right)$ is the average number of citations per article of the two persons).

[1] Permanent address

Key words and phrases: H-index, H-sequence, Lotka.

I. Introduction

The Hirsch-index (or h-index), introduced in Hirsch (2005) by Jorge Hirsch (a physicist) is one of the most remarkable measures of impact in informetrics. Hundreds of articles are already devoted to the h-index (see e.g. Egghe (2010c) for an extensive review up to and including 2008). The h-index can be applied to any source-item production system (i.e. an information production process – see Egghe (2005)) but in this paper (as in Hirsch (2005)) we will consider the case of a researcher and his/her articles and citations to these articles. We refer to van Raan (2006) for an early application of the h-index to the evaluation of research groups.

A researcher has h-index h if, in the list of articles arranged in decreasing order of the number of citations to these articles, $r = h$ is the highest rank such that the papers on rank $1, 2, ..., h$ (the h-core) each have at least h citations. We refer again to Egghe (2010c) for an overview of advantages and disadvantages of the h-index.

One clear disadvantage of the h-index is that, once a paper is in the h-core it does not matter how many citations it receives (as long as it is at least h). It is the view of many informetricians that the number of citations to the papers in the h-core should count more in an impact measure than it does in the h-index – see Bar-Ilan (2006), Bornmann and Daniel (2005), Braun, Glänzel and Schubert (2005, 2006), Egghe (2006, 2009), Glänzel (2006) and many more references in Egghe (2010c). Variants of the h-index (such as the g-index (Egghe (2006)) or the R-index (Jin, Liang, Rousseau and Egghe (2007)) have been introduced to (partially) overcome this disadvantage of the h-index.

In addition to the above mentioned disadvantage of the h-index it is clear that – as any index – the h-index is a single number. Instead of one number one can consider a sequence of numbers. I do not mean the h-index (or variant) in function of time (e.g. the h-index of a person, calculated each year of the career – for this see e.g. Egghe (2009) and references therein) but h-type index sequences in the sense of characteristic scores and scales (as introduced in Glänzel and Schubert (1988) and further studied in Glänzel (2008, 2009) and Egghe (2010a,b).

The idea is, e.g. on the ranked list of articles of a researcher, arranged in decreasing order of the number of received citations, to give some remarkable ranks in this list. We repeat two examples. Glänzel and Schubert (1988) consider the average number of citations per paper, disregard papers with less citations than this average yielding a cut-off point in the list and this procedure is repeated several times yielding the sequence (here: increasing in rank). In Egghe (2010b) a decreasing sequence, based on the h-index, is obtained as follows. First calculate the h-index and then delete the h-core. Of the remaining list we again calculate the h-index. This is continued several times yielding the sequence.

Randić (2009) has another (related) idea where one better uses the actual numbers of citations as well as where a list is produced. The method goes as follows. The first number is the h-index. Then, on the same list one doubles all the ranks (so obtaining the "ranks" 2,4,6,…). On this list one calculates the h-index (which, obviously, is smaller than or equal to the first h-index), yielding the second number in the sequence. Then the ranks are doubled again (yielding the "ranks" 4,8,12,…). On this list one calculates the h-index, yielding the third number in the sequence, and so on. This is a clever idea, where one better uses the actual number of citations of the papers in the h-core (see Randić (2009), Table 1, p. 813).

Then Randić takes the partial sums of these numbers (i.e. if ($h = {}^{0}H, {}^{1}H, {}^{2}H, ...$) is the above sequence (notation of Randić), Randić takes ($h = {}^{0}H, {}^{0}H + {}^{1}H, {}^{0}H + {}^{1}H + {}^{2}H, ...$) and this is called the H-sequence. Finally, the last number in the latter sequence (i.e. the sum of the ${}^{0}H, {}^{1}H, {}^{2}H, ...$) is called H.

In the next section, we present a mathematical model for these indices in the Lotkaian framework. That means that we assume that Lotka's law is valid:

$$f(j) = \frac{C}{j^{\alpha}} \tag{1}$$

$C > 0, \ \ \alpha > 1, \ \ j \geq 1$ where $f(j)$ is the density of sources (here articles) with item-density j (items are here citations). In this connection, Egghe and Rousseau (2006) proved that, in systems where (1) is valid, the h-index is given by

$$h = T^{\frac{1}{\alpha}} \tag{2}$$

where T is the total number of sources (here articles). We will calculate the variants of (2) for the h-indices of Randić as well as for his H-sequence and the index H.

In Randić (2009), the author further defines the notion of "dominance" of one author over another one. That occurs when all terms in the H-sequence of one author are larger than the corresponding terms (on the same coordinate) of the other author. In the third section we give a necessary and sufficient condition for this to happen: Lotka's α of the dominating author must be smaller than the one of the dominated author. The same characterization is given for $H_1 > H_2$, i.e. for the Randić H-index of the first author to be larger than the one of the second author. Further these conditions are also equivalent with $\mu_1 > \mu_2$, i.e. the average number of citations per article of the first author must be larger than the average number of citations per article of the second author.

The paper closes with conclusions and suggestions for further research.

II. Lotkaian models for the Randić indices

We first give a model for the different Randić h-indices and then we use this to give formulae for the H-index and H-sequence of Randić.

Theorem II.1

Let us denote the kth-h-index of Randić by ${}^{k}H$ (same notation as in Randić (2009)), $k = 0,1,2,...$. Then

$${}^{k}H = \frac{T^{\frac{1}{\alpha}}}{2^{k\frac{\alpha-1}{\alpha}}} = \frac{h}{2^{k\frac{\alpha-1}{\alpha}}} \tag{3}$$

Proof: see the Appendix.

Note that, for $\alpha = 2$ (a turning point in informetrics – see Egghe (2005)), (3) reduces to

$$^{k}H = \frac{T^{\frac{1}{\alpha}}}{2^{\frac{k}{2}}} \tag{4}$$

Note that, for $k = 0$, (3) reduces to $^{0}H = T^{\frac{1}{\alpha}} = h$, by (2), as it should.

We can now present a formula for the H-index of Randić.

Theorem II.2:

$$H = T^{\frac{1}{\alpha}} \frac{1}{1-2^{-\frac{\alpha-1}{\alpha}}} = \frac{h}{1-2^{-\frac{\alpha-1}{\alpha}}} \tag{5}$$

Proof: see the Appendix.

Note that for $\alpha = 2$ we have that $H = \frac{\sqrt{2}}{\sqrt{2}-1} h$.

Remark:

Instead of adding all the h-indices in (A4) one could give a higher weight to the first h-indices (such as $h = {}^{0}H$) by defining a new measure H^{*} as the sum of the squares of the numbers $^{0}H, {}^{1}H, {}^{2}H...$:

$$H^{*} = \sum_{k=0}^{\infty} \left(\frac{T^{\frac{1}{\alpha}}}{2^{k\frac{\alpha-1}{\alpha}}} \right)^{2}$$

$$H^{*} = T^{\frac{2}{\alpha}} \frac{1}{1-4^{-\frac{\alpha-1}{\alpha}}} = \frac{h^{2}}{1-4^{-\frac{\alpha-1}{\alpha}}} \tag{6}$$

Now we turn our attention to the calculation of the H-sequence of Randić.

Theorem II.3:

The H-sequence of Randić is

$$\mathscr{H} = \left({}^{0}H = h, {}^{0}H + {}^{1}H, \ldots, \sum_{i=0}^{k} {}^{i}H, \ldots \right) \tag{7}$$

where the $\left(k+1\right)^{th}$ term is given by

$$\sum_{i=0}^{k} {}^{i}H = h \frac{1-2^{-(k+1)\frac{\alpha-1}{\alpha}}}{1-2^{-\frac{\alpha-1}{\alpha}}} \tag{8}$$

Proof: By (3) we have

$$\sum_{i=0}^{k} {}^{i}H = h \sum_{i=0}^{k} \frac{1}{2^{k\frac{\alpha-1}{\alpha}}}$$

which is (8) based on the formula for a finite geometric series. □

III. Author dominance in the sense of Randić

Let us have two authors with H-sequence respectively

$$\mathscr{H}_1 = \left({}^{0}H_1 = h_1, {}^{0}H_1 + {}^{1}H_1, \ldots, \sum_{i=0}^{k} {}^{i}H_1, \ldots \right) \tag{9}$$

and

$$\mathscr{H}_2 = \left({}^{0}H_2 = h_2, {}^{0}H_2 + {}^{1}H_2, \ldots, \sum_{i=0}^{k} {}^{i}H_2, \ldots \right) \tag{10}$$

Randić's H-index and H-sequence have been defined, as explained in the introduction, to refine the classical h-index so as to better take into account the actual number of citations to the papers in the h-core. Therefore, the following problem is interesting: given two authors as above with equal h-indices $h_1 = h_2$, characterise the situations

$$\mathscr{H}_1 > \mathscr{H}_2$$

i.e. for which

$$\sum_{i=0}^{k} {}^{i}H_1 > \sum_{i=0}^{k} {}^{i}H_2$$

for all $k = 1, 2, ...$ (for $k = 0$ we have ${}^{0}H_1 = h_1 = h_2 = {}^{0}H_2$ as supposed).

We suppose that the article-citation system is Lotkaian with Lotka-exponent α_1 for the first author respectively α_2 for the second author. Denote by μ_1, respectively μ_2 the average number of citations per article for author 1, respectively author 2. We have the following characterization

Theorem III.1:
The following assertions are equivalent, given that $h_1 = h_2$.

(i) $\mathscr{H}_1 > \mathscr{H}_2$

(ii) $\alpha_1 < \alpha_2$

If $\alpha_1, \alpha_2 > 2$ then these assertions are equivalent with

(iii) $\mu_1 > \mu_2$

Proof: see the Appendix.

The same characterization can be given for author dominance via the measure H: let

$$H_1 = \frac{h_1}{1-2^{-\frac{\alpha_1 - 1}{\alpha_1}}} \tag{11}$$

and

$$H_2 = \frac{h_2}{1-2^{-\frac{\alpha_2 - 1}{\alpha_2}}} \tag{12}$$

be the Randić H-indices of the two authors. Then we have

Theorem III.3:

The following assertions are equivalent, given that $h_1 = h_2$.

(i) $H_1 > H_2$

(ii) $\alpha_1 < \alpha_2$

If $\alpha_1, \alpha_2 > 2$ then these assertions are equivalent with

(iii) $\mu_1 > \mu_2$.

Proof: see the Appendix.

Note: One can also show that $H_1^* > H_2^*$ is equivalent with the assertions in Theorem III.3 (based on (6)).

IV. Conclusions and suggestions for further research

The Randić H-sequence and H-index have been studied in a Lotkaian framework. We have given formulae for the H-sequence and H-index and we have given necessary and sufficient conditions for dominance via the H-sequence or H-index, given that the Hirsch indices are equal.

It is hence clear that the H-sequence and H-index of Randić is capable to describe the citation scores of articles in the Hirsch h-core and to distinguish authors with the same Hirsch h-index.

The advantage of Randić's method is that a sequence of impact measures is defined using the actual number of citations to papers in the h-core. Hence it refines the h-index and similar h-type indices. As such, Randić's method is similar to the method of characteristic scores and scales.

It is too early to say which methodology best describes the citation impact of an author. Further research is needed to understand the practical useability of the H-index and H-sequence in different fields and to see how H-sequence dominance can be used in practical cases.

An additional point of research is to make a statistical study on the significance of the difference between two H-indices, given that their Hirsch indices are the same. By extension this also goes for the other variants of the h-index such as the g-index or the R-index.

References

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

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

T. Braun, W. Glänzel and A. Schubert (2005). A Hirsch-type index for journals. The Scientist 19(22), 8-10.

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

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

L. Egghe (2006). Theory and practise of the g-index. Scientometrics, 69(1), 131-152.

L. Egghe (2009). Mathematical study of h-index sequences. Information Processing and Management 45(2), 288-297.

L. Egghe (2010a). Characteristic scores and scales in a Lotkaian framework. Scientometrics, to appear.

L. Egghe (2010b). Characteristic scores and scales based on h-type indices. Journal of Informetrics, to appear.

L. Egghe (2010c). The Hirsch-index and related impact measures. Annual Review of Information Science and Technology, 44 (B. Cronin, ed.), 65-114. Information Today, Inc., Medford, New Jersey, USA

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

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

W. Glänzel (2008). What are your best papers ? ISSI Newsletter 4(4), 64-67.

W. Glänzel (2009). The role of the h-index and characteristic scores and scales in testing the tail properties of scientometric distributions. Proceedings of the 12th International Conference of Scientometrics and Informetrics (ISSI 2009) (B. Larsen and J. Leta, eds.), 120-130, BIREME/PAHO/WHO and Federal University of Rio de Janeiro, Brasil.

W. Glänzel and A. Schubert (1988). Characteristic scores and scales in assessing citation impact. Journal of Information Science 14, 123-127.

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(46), 16569-16572.

B. Jin, L. Liang, R. Rousseau and L. Egghe (2007). The R- and AR-indices: Complementing the h-index. Chinese Science Bulletin 52(6), 855-863.

M. Randić (2009). Citations versus limitations of citations: beyond Hirsch index. Scientometrics 80(3), 809-818.

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.

Appendix

Proof of Theorem II.1

By (1), the total number of articles with n or more citations is given by

$$\int_n^\infty f\left(j\right)dj = \int_n^\infty \frac{C}{j^\alpha}dj$$

$$= \frac{C}{\alpha - 1} n^{1-\alpha} \tag{A1}$$

But the total number T of articles is given by

$$T = \int_1^\infty f\left(j\right)dj = \frac{C}{\alpha - 1} \tag{A2}$$

(A2) in (A1) yields

$$\int_n^\infty f\left(j\right)dj = Tn^{1-\alpha} \tag{A3}$$

By definition of ${}^kH$ we have, by (A3)

$$T\left(2^k \,.\, {}^kH\right)^{1-\alpha} = {}^kH$$

from which (3) follows, using also (2). □

Proof of Theorem II.2

By definition of H, we have, using Theorem II.1

$$H=\sum_{k=0}^{\infty}\frac{T^{\frac{1}{\alpha}}}{2^{k\frac{\alpha-1}{\alpha}}} \tag{A4}$$

$$H=T^{\frac{1}{\alpha}}\sum_{k=0}^{\infty}\left(2^{-\frac{\alpha-1}{\alpha}}\right)^{k}$$

which is a geometric series with $0<2^{-\frac{\alpha-1}{\alpha}}<1$ since $\alpha>1$. By the well-known formula of a converging geometric series we have (5), also using (2). □

Proof of Theorem III.1.

Denote

$$x_1=2^{-\frac{\alpha_1-1}{\alpha_1}} \tag{A5}$$

$$x_2=2^{-\frac{\alpha_2-1}{\alpha_2}} \tag{A6}$$

Then (i) is equivalent with (by definition) (since $h_1=h_2$), by Theorem II.3:

$$\frac{1-x_1^{k+1}}{1-x_1}>\frac{1-x_2^{k+1}}{1-x_2} \tag{A7}$$

for all $k=1,2,...$. The Lemma below shows that (A7) is true for all cases where $x_1>x_2$ (since $k\geq 1$). If $x_1=x_2$ then (17) is an equality and if $x_1<x_2$ then the opposite strict inequality is true in (17). This means that (17) is valid if and only if $x_1>x_2$.

Now $x_1>x_2$ is equivalent with

$$\left(\frac{1}{2}\right)^{\frac{\alpha_1-1}{\alpha_1}} > \left(\frac{1}{2}\right)^{\frac{\alpha_2-1}{\alpha_2}}$$

which is equivalent with (since $\alpha_1,\alpha_2 > 1$)

$$\frac{\alpha_1-1}{\alpha_1} < \frac{\alpha_2-1}{\alpha_2}$$

which is equivalent with $\alpha_1 < \alpha_2$.

Now (1) implies that

$$\mu = \frac{\int_1^\infty jf\left(j\right)dj}{\int_1^\infty f\left(j\right)dj} = \frac{\alpha-1}{\alpha-2} \qquad \text{(A8)}$$

, if we suppose $\alpha > 2$. Now (19) applied to α_1 and α_2 (supposed to be > 2) yields for (iii):

$$\frac{\alpha_1-1}{\alpha_1-2} > \frac{\alpha_2-1}{\alpha_2-2}$$

but this is equivalent with (ii) as is readily seen. □

<u>Lemma III.2</u>: The function

$$f\left(x\right) = \frac{1-x^{k+1}}{1-x}$$

is strictly increasing in $x \in \left]0,1\right[$ for all $k \geq 1$, $k \in \square$.

<u>Proof</u>:

$$f'(x) = \frac{(1-x)\left(-(k+1)x^k\right) - \left(1-x^{k+1}\right)(-1)}{(1-x)^2}$$

which has the same sign as

$$-(k+1)x^k + (k+1)x^{k+1} + 1 - x^{k+1}$$

which is strictly positive if and only if

$$kx^k(x-1) > x^k - 1$$

or

$$kx^k < \frac{x^k - 1}{x-1} \tag{A9}$$

since $0 < x < 1$. But

$$\frac{x^k - 1}{x-1} = \sum_{i=0}^{k-1} x^i \tag{A10}$$

The sum in (21) contains k terms, each of which are strictly larger than x^k since $0 < x < 1$. This proves (20) and hence the Lemma. □

<u>Proof of Theorem III.3</u>

Since $h_1 = h_2$, (i) is valid if and only if

$$\frac{1}{1-2^{-\frac{\alpha_1 - 1}{\alpha_1}}} > \frac{1}{1-2^{-\frac{\alpha_2 - 1}{\alpha_2}}}$$

This is true if and only if

$$2^{-\frac{\alpha_1-1}{\alpha_1}} > 2^{-\frac{\alpha_2-1}{\alpha_2}}$$

But the function $f\left(x\right)=2^{-\frac{x-1}{x}}$ is a strictly decreasing function since

$$f\,'\left(x\right)=\left(\frac{1}{2}\right)^{\frac{x-1}{x}}\ln\left(\frac{1}{2}\right)\frac{1}{x^2}<0$$

This proves (i) $\Leftrightarrow$ (ii) and the equivalence with (iii) is proved as in Theorem III.1. □