Made available by Hasselt University Library in https://documentserver.uhasselt.be

Influence of adding or deleting items and sources on the h-index

Peer-reviewed author version

EGGHE, Leo (2010) Influence of adding or deleting items and sources on the h-index. In: JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY, 61(2). p. 370-373.

DOI: 10.1002/asi.21239
Handle: http://hdl.handle.net/1942/10236

The influence of adding or deleting items and sources on the h-index

Leo Egghe

Universiteit Hasselt (Uhasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium

leo.egghe@uhasselt.be

ABSTRACT

Adding or deleting items, such as self-citations has an influence on the h-index of an author. This influence will be proved mathematically in this paper. We hereby prove the experimental finding in [Gianoli and Molina-Montenegro, Journal of the American Society for Information Science and Technology 60(6), 1283-1285, 2009] that the influence of adding or deleting self-citations on the h-index is greater for low values of the h-index. Why this is logic is also shown by a simple theoretical example.

Adding or deleting sources, such as adding or deleting minor contributions of an author, also has an influence on the h-index of this author. Also this influence is modelled in this paper. This model explains some practical examples found in [Hu, Rousseau and Chen, Journal of Information Science, to appear].

I. Introduction

The h-index (or Hirsch-index), introduced in Hirsch (2005) is the most popular indicator of impact of an author. Soon after its introduction it was noted that the h-index can also be defined for journals (Braun, Glänzel and Schubert (2005,2006)), institutes or groups of authors (van Raan (2006), Egghe and Rao (2008)), topics and compounds (Banks (2006)), countries (Csajbók, Berhidi, Vasas and Schubert (2007)) and even libraries (Liu and Rousseau (2009)) - see also the extensive review Egghe (2009).

In general, the h-index can be defined for general "information production processes" (IPPs) where we have sources producing items (example, as defined by Hirsch: an author is an IPP, his/her articles are the sources and the citations to these articles are the items) - see Egghe and Rousseau (1990) or Egghe (2005). In this sense we rank the sources in decreasing order of their number of items and the h-index is the largest rank $r = h$ such that all the sources on ranks $r = 1,...,h$ have at least h items.

It is clear that the h-index depends on the used databases. Scopus, Web of Science (WoS) and Google Scholar (via www.harzing.com/pop.htm) publish the h-index based on their citation and publication data and this can lead to different h-values (see Bar-Ilan (2008), Jacsó (2008a,b)) since these databases have different publication data and citation data (e.g. of an author).

As always when working with citation data there is the question how to deal with self-citations of an author. We do not go into the debate whether or not one should include self-citations in the calculation of the h-index of an author (Schreiber (2007), Zhivotovsky and Krutovski (2008)). But it is clear that the h-index based on all citation data (including self-citations) will be higher ($\geq$) than the h-index based on citation data excluding self-citations. In Gianoli and Molina-Montenegro (2009) one finds, experimentally, that the influence of self-citations on the h-index is greater for low h-index values than for high h-index values.

In the next section we give a mathematical explanation of this property based on Egghe, Liang and Rousseau (2009) where a mathematical relation between the h-index and the impact factor (IF) of a journal (or more general, where IF is interpreted as the average number of items per source) is proved. We further present a theoretical example showing that the Gianoli and Molina-Montenegro finding is logical.

Adding or deleting self-citations is an example of adding or deleting items. But also sources can be added or deleted. An example is given in Hu, Rousseau and Chen (2009). Here one compares the h-index of an author, based on all publications (and citations), with the so-called major contribution h-index (denoted $h\text{-}maj$) which is only based on the publications (and citations) in which the author contributes in a major way (e.g. as first author or corresponding author - see Hu, Rousseau and Chen (2009)). Of course, $h\text{-}maj \leq h$. In the third section we study the mathematical difference between $h\text{-}maj$ and h based on the mathematical model for the h-index, proved in Egghe and Rousseau (2006). The obtained model yields values of $h\text{-}maj$ in function of h which are confirmed by the practical data in Hu, Rousseau and Chen (2009).

The paper closes with conclusions and suggestions for further research.

II. The influence of adding or deleting items on the h-index

Here and in the next section we will work in Lotkaian IPPs - Egghe (2005). This means that the source-item size-frequency function is given by a decreasing power law as in (1)

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

where $C > 0$ and $\alpha > 1$. In Egghe and Rousseau (2006) we proved that in this case we have formula (2) for the h-index.

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

where T denotes the total number of sources.

Based on this result and some basic Lotkaian properties (Egghe (2005)) the following relation between the h-index and the impact factor IF (in the case of the h-index for journals) was proved, if $\alpha > 2$ (see Egghe, Liang and Rousseau (2009))

$$h = C\left(1 - \frac{1}{2IF}\right)^{\frac{IF-1}{IF-1}} \tag{3}$$

In the general case of IPPs the IF is nothing else than the average number $\mu = \frac{A}{T}$ of items per source (A = total number of items, T = total number of sources). So (3) reads, in the general setting, as in (4)

$$h = C\left(1 - \frac{1}{2\mu}\right)^{\frac{\mu-1}{2\mu-1}} \tag{4}$$

In Egghe, Liang and Rousseau (2009) it is proved that (4) is a concavely increasing function of μ, for every fixed $C > 0$. We will use formula (4) for the explanation of the Gianioli and Molina-Montenegro finding since (4) is an exact mathematical deduction from (2) (Egghe, Liang and Rousseau (2009)) which in turn is an exact mathematical deduction from (1) (Egghe and Rousseau (2006)), the established law of Lotka. If (4) can explain the Gianoli and Molina-Montenegro finding we hence will have a mathematically correct and simple explanation (which does not exclude that other arguments are also possible to explain the same finding).

Let us now return to the case where we add or delete items but no sources. Let us consider the case where we delete items (as is the case where items are self-citations where we assume that there are no sources with only self-citations!). The case of adding items is treated in the same way. Then T, the total number of sources remains the same while A, the total number of items decreases. This is true in any case where sources are kept the same but where items are deleted (as in the example of deleting self-citations).

Consequently, the average number of items per source, μ, decreases. Because of the concave increasing relationship (4) between h and μ we hence have that h decreases more for small values than for large values of h. We note that, in this change of number of items, also C can change (since $C = f(1)$ it e.g. decreases due to deletion of items in sources with 1 item but it possibly increases due to deletion of items in sources with more than 1 item) but then we move to another function represented in the sheave (4) and each function has the concavely increasing relationship between h and μ and hence the above reasoning still applies.

This explains the finding in Gianoli and Molina-Montenegro (2009). Why the above mathematical argument is also intuitively logic is shown in the next (theoretical) example. Let us have a first IPP with h_1 sources each with h_1 items (hence with h-index h_1). Let us have a second IPP with h_2 sources each with h_2 items (hence with h-index h_2) and let $h_1 < h_2$. Let us now delete in each source in the first IPP n items (where $n < h_1$, $n \hat{I}$ ¥) and the same for the second IPP. This means that in the first IPP we delete nh_1 items and in the second IPP we delete nh_2 items, where we have that $nh_2 > nh_1$. Hence in the second IPP more items are deleted when compared with the first IPP.

Now it is clear that the h-index of the first IPP (where the items are deleted) is $h'_1 = h_1 - n$ while the h-index of the second IPP (where the items are deleted) is $h'_2 = h_2 - n$. Although more items are deleted in the second IPP, we have $h'_2 - h_2 = h'_1 - h_1$ but, more importantly,

$$\frac{h'_2}{h_2} = 1 - \frac{n}{h_2} > 1 - \frac{n}{h_1} = \frac{h'_1}{h_1} \qquad (5)$$

since $h_2 > h_1$ so that the change in h_2 is relatively smaller than in h_1.

In the next section we turn our attention to the addition or deletion of sources.

III. The influence of adding or deleting sources on the h-index

In Hu, Rousseau and Chen (2009) we have an example where one studies the "classical" h-index h, where all publications (and citations to these publictions) are used, in comparison with the "major contribution h-index" (denoted $h - maj$) where only the articles are considered in which the author has a major contribution (this needs more clarification but as an example one gives articles in which the author is first author (in

case of non-alphabetical order of the authors of course!) or corresponding author. This is an example where sources are added or deleted from the IPP.

We do not go further into the interpretation of "major contribution" but here we can suffice in assuming that $h\text{-}\ maj$ is based on a subset of the articles of an author. So we have two situations. First we have the classical h-index h, based on all publications where formula (2) is valid in the Lotkaian framework where there are T articles in total and where we have Lotka's exponent $\alpha > 1$. Then we have the major contribution h-index $h\text{-}\ maj$ based on, say T_m major contributions. If we assume that the same Lotka exponent α applies to this situation, we hence have, using again the result in Egghe and Rousseau (2006)

$$h\text{-}\ maj = T_m^{\frac{1}{\alpha}} \tag{6}$$

We underline that it is not clear whether or not the same Lotka-exponent α applies but this is left as an open problem and is supposed as a first approximation (we will check the agreement of our model's result with the practical results in Hu, Rousseau and Chen (2009)). Hence, we have, by (2) and (6) that

$$\frac{h\text{-}\ maj}{h} = \left(\frac{T_m}{T}\right)^{\frac{1}{\alpha}} \tag{7}$$

being a formula for the comparison of $h\text{-}\ maj$ and h. Note that formulae (6) and (7) are related to h-index concatenation as studied in Glänzel (2008). In our terminology, if one takes the union of disjoint sets of sources (belonging to two IPPs) then the new IPP is a "concatenation" of the two IPPs and the h-index of the new IPP is defined as the concatenated h-index. Denote by T_1 the total number of sources in the first IPP and by T_2 the total number of sources in the second IPP, then the concatenated IPP has $T = T_1 + T_2$ sources. Assuming that the same Lotka exponent α applies, we have, by (2)

$$h^\alpha = h_1^\alpha + h_2^\alpha \tag{8}$$

(h_i, i = 1, 2 are the h-indices of the two IPPS), hence

$$h = \left(h_1^{\alpha} + h_2^{\alpha}\right)^{\frac{1}{\alpha}} \tag{9}$$

which also appears in Glänzel (2008). An application of this is given by the above: if we denote by T_r the total number of sources, which are not major contributions, we have $T = T_m + T_r$ and, by (9)

$$h = \left((h - maj)^{\alpha} + (h - rest)^{\alpha}\right)^{\frac{1}{\alpha}} \tag{10}$$

where h-rest is the h-index of this set of non-major contributions. Also

$$\frac{h - rest}{h} = \left(\frac{T_r}{T}\right)^{\frac{1}{\alpha}} \tag{11}$$

is valid.

There are two examples in Hu, Rousseau and Chen (2009). In each case on considers Chinese projects and scientists in the framework of the 11th national short plan in health sciences. Using the WoS one finds 781 papers out of 2,439 papers as major contributions, i.e. 32 %. Using a Chinese database (CNKI-Citation Index Subset (the citation index of the China National Knowledge Infrastructure), denoted CNKI-C) one finds 3,873 papers out of 10,476 papers as major contributions, i.e. 37 %. So we can conclude that, in this example, about one third of the papers are major contributions leading to an estimate of $T_m = \frac{T}{3}$. If we apply (7) with an estimated classical value of $\alpha = 2$ (see Egghe (2005)) we reach the estimate of (based on (7))

$$h - maj = \sqrt{\frac{1}{3}}h = 0.58h \tag{12}$$

We note the robustness of the h-index: a deletion of about 2/3 of the papers leads to a decrease of the h-index of less than 50 %. This robustness was already remarked in Rousseau (2007):

$$\frac{dh}{dT} = 1 \tag{13}$$

and

$$\lim_{T \to \infty} \frac{dh}{dT} = 0 \tag{14}$$

which readily follows form (2) and, based on this,

$$\frac{dh}{dT} = \frac{1}{\alpha} T^{\frac{1}{\alpha} - 1} \tag{15}$$

since $\alpha > 1$.

We checked result (12) in the Tables 1 and 2 in Hu, Rousseau and Chen (2009). In Table 1 in the case of WoS, for the weighted average over the disciplines, we found a weighted average of the h-index $\bar{h} = 3.41$ and a weighted average major contribution h-index $\overline{h\text{-}maj} = 1.71$, hence $\overline{h\text{-}maj} \approx 0.5\bar{h}$, reasonably close to (12). In case of CNKI-C we are even more close: $\bar{h} = 9.92$ and $\overline{h\text{-}maj} = 5.70$, hence $\overline{h\text{-}maj} \approx 0.57\bar{h}$, very close to (12). In Table 2 one considers individual Chinese scientists in one field (oncology) (15 scientists). In the case of WoS we have $\bar{h} = 7.40$ and $\overline{h\text{-}maj} = 2.87$, hence $\overline{h\text{-}maj} \approx 0.39\bar{h}$ which is different from (12) but in the case of CNKI-C we have $\bar{h} = 11.27$ and $\overline{h\text{-}maj} = 6.27$ yielding $\overline{h\text{-}maj} \approx 0.56\bar{h}$, very close to (12). Why there is a difference between the WoS and CNKI-C case for individual authors is left as an open problem (perhaps the small sample of 15 scientists is one of the reasons).

This section shows that also the influence of adding or deleting sources on the h-index can be modelled adequately.

For more general models on merging and its influence on the h-index we refer to Egghe (2008).

IV. Conclusions and suggestions for further research

We have modelled (in a general source-item relation in IPPs) the influence of adding or deleting self-citations on the h-index of an author and we could prove the experimental

finding of Gianoli and Molina-Montenegro (2009) that this influence is larger for the lower h-indices. A simple example also shows the logic of this.

We have also modelled (again in a general source-item relation in IPPs) the influence of adding or deleting minor contributions on the h-index of an author and where able to predict the experimental results as shown in Hu, Rousseau and Chen (2009).

It would be interesting in both cases (adding or deleting sources or items) how a Lotkaian IPP changes. We mean by this: given a Lotkaian system and given a mathematically defined way of adding or deleting of sources and/or items, how is the exponent α, as appearing in Lotka's power law (1) changing. This would give even more insight in the two problems that were discussed here. The same study could also be executed for other indices such as (e.g.) the g-index (Egghe (2006)) or the R-index (Jin, Liang, Rousseau and Egghe (2007)).

References

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

J. Bar-Ilan (2008). Which h-index? - A comparison of WoS, Scopus and Google Scholar. Scientometrics 74(2), 257-271.

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.

E. Csajbók, A. Berhidi, L. Vasas and A. Schubert (2007). Hirsch-index for countries based on essential science indicators data. Scientometrics 73(1), 91-117.

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 (2008). The influence of merging on h-type indices. Journal of Informetrics 2(3), 252-262.

L. Egghe (2009). The Hirsch-index and related impact measures. Annual Review of Information Science and Technology, Chapter 2, 65-115 (to appear).

L. Egghe, L. Liang and R. Rousseau (2009). A relation between h-index and impact factor in the power law model. Journal of the American Society for Information Science and Technology, to appear.

L. Egghe and I.K.R. Rao (2008). Study of different h-indices for groups of authors. Journal of the American Society for Information Science and Technology 59(8), 1276-1281.

L. Egghe and R. Rousseau (1990). Introduction to Informetrics. Quantitative Methods in Library, Documentation and Information Science. Elsevier, Amsterdam, the Netherlands.

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

E. Gianoli and M.A. Molina-Montenegro (2009). Insights into the relationship between the h-index and self-citations. Journal of the American Society for Information Science and Technology 60(6), 1283-1285.

W. Glänzel (2008). H-index concatenation. Scientometrics 77(2), 369-372.

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.

X. Hu, R. Rousseau and J. Chen (2009). In those fields where multiple authorship is the rule, the h-index should be supplemented by a major contribution h-index. Journal of Information Science, to appear.

P. Jacsó (2008a). The plausibility of computing the h-index of scholarly productivity and impact using reference-enhanced databases. Online Information Review 32(2), 266-283.

P. Jacsó (2008b). Testing the calculation of a realistic h-index in Google Scholar, Scopus and Web of Science for F.W. Lancaster. Library Trends 56(4), 784-815.

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.

Y. Liu and R. Rousseau (2007). Hirsch-type indices and library management: The case of Tongji University library. Proceedings of the 11th International Conference of the International Society for Scientometrics and Informetrics (D. Torres-Salinas and H.F. Moed, eds.), 514-522.

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

M. Schreiber (2007). Self-citation corrections for the Hirsch index. Europhysics Letters 78(3). Retrieved February 2, 2009 from dx.doi.org/10.1209/0295-5075/78/30002

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

L.A. Zhivotovsky and K.V. Krutovsky (2008). Self-citation can inflate h-index. Scientometrics 77(2), 373-375.