Author Evaluation Based on H-index and Citation Response

Miloš Kudělka, Jan Platoš, and Pavel Krömer

Department of Computer Science
VŠB Technical University of Ostrava
Ostrava, Czech Republic
{milos.kudelka, jan.platos, pavel.kromer}@vsb.cz

Abstract. An accurate and fair assessment of the efficiency and impact of scientific work is, despite a lot of recent research effort, still an open problem. The measurement of quality and success of individual scientists and research groups can be approached from many different directions, none of which is universal. A reason for this is inherently different behavior of different scientists within the global research community. A complex evaluation of ones publication activities requires a careful consideration of a wide variety of factors. The well-known H-index is one of the most used bibliometric indices. Despite its many imperfections, its simplicity and ease of interpretation make it a popular scientometric method. This short paper uses the ideas behind the H-index to analyze communities of authors who cite publishing scientists. A new author evaluation measure named aH-indexis proposed, and intuitive interpretations of its properties and semantics are presented. Preliminary experiments with authors with high H-index active in the area of computer science are presented to demonstrate the properties of the proposed measure.

Keywords: H-index,citation,bibliometric,scientometric.

1 Introduction

Publication activities and citation response of active scientists are affected by a number of different factors. Natural preconditions for successful publication of research results include a deep understanding of a research topic, creativity, diligence, and last but not least an ability to comprehend current trends and latest advances in a field of study. Many new factors influencing the nature of a scientist's publication behavior can be attributed to modern technologies. Internet databases and search engines can be used to quickly explore information sources such as conference papers and journal articles. Electronic communication simplifies networking between scientists and their research groups and supports preservation of long and short-term contacts. Natural consequences include an increase in the average number of research paper co-authors, the emergence of strongly connected but sometimes opportunistic research groups (communities), and the rise of multidisciplinary research topics and hybrid methods.

In response to this development, many new conferences and journals are launched every year. They are created in reaction to the wide need for the massive and rapid dissemination of research results. This need, however, is sometimes motivated by the academic and peer pressure exercised on individual scientists (publish or perish) rather than by the quality of their work.

A number of scientometric measures have been developed to evaluate ones research performance and impact of his or her work [1,2]. The major issue of many simple author performance metrics, based on citation response, is that they often do not reflect the quality of research papers that triggered the response, and the publication venues (conferences and journals) that published them. Moreover, they do not reflect how the author reached his or her level of citation response. Important information such as with how many co-authors the scientist usually collaborates, how often and how much do different authors respond to his or her research in their work, and in how many research areas is he or she active, is not considered by most traditional scientometric measures at all.

Our approach, outlined in this short paper, is different. Even though we are very well aware of the known imperfections of the original *H-index*, we consider its underlying principles excellent and especially value its simplicity and ease of interpretation. In this work, we extend the H-index by an analysis of the citation response received by scientists. The proposed evaluation measure, termed *aH-index*, complements the H-index with a new type of information reflecting the quality and quantity of a scientist's citation response. Due to its design, heavily inspired by the principles of the H-index, it suffers from the same problems. However, it also retains the simplicity and interpretability of the original Hirsch index.

2   Related Work

Hirsch proposed in [3] a single number, H-index, as a particularly simple and useful way to characterize the scientific output of a researcher. A purpose of the H-index was to describe both the productivity and impact of the published work of a scientist. However, there are some well-known drawbacks of using the H-index to evaluate and compare individual scientists. They are e.g. neglecting the quality of publications, a number of co-authors of a citing paper, comparing scientists working in different research fields, the number of citations of most cited papers, etc. That is why H-index characteristics were extensively investigated. Costas et al. analyzed in [4] the relationship of the H-index to other well-known bibliometric indicators.

After the introduction of the H-index, many improvements that addressed its fundamental drawbacks were proposed. In [5], Zhang proposed a new index that is suitable for evaluating highly cited scientists and comparing groups of scientists with an identical H-index. Alonso et al. present in [6] a comprehensive review on the H-index and related indicators. They studied their main advantages, drawbacks, and main applications. In [7], Bornmann et al. present a study

of 37 different variants of H-index. They show a high correlation between the H-index and most of its variants.

The complicated relationship between the scientist and his or her co-authors is one of the problems of the H-index. Hirsch proposes in [8] a new version of the H-index that takes into account the effect of multiple co-authors and solves a well-known problem with the so-called Hirsch core. A similar problem is solved and a new variant of the H-index is introduced by Wan et al. in [9]. In [10], McCarty et al. apply social network analysis on ego co-authorship network. They show that the highest H-index can be achieved by working with many co-authors.

Analysis of the relation between the H-index and the behavior of citing authors (citers) is also suitable for a better understanding how a community of citing authors influences H-index. Brooks study complex citer motivations in [11]. Seven citer motives are analyzed, and more than 70% of references surveyed are the result of a complex interplay of multiple citer motives. Amancio et al. investigate in [12] the dependency of a quantity of citations on author reputation (visibility). They show that the reputation can affect a temporal evolution of H-index. In [13], Bras-Amorós et al. present a new index in which the evaluated objects are the citations received by an author and the quality function is based on a collaboration distance between the authors of the cited and the citing papers. The new index takes into account only significant citations; significance is proportional to collaboration distance.

3 Author Evaluation

In this section, we first recall the H-index and then propose a new citation measure evaluating certain *properties* of the citation response received by publishing scientists.

Definition 1 (H-index). *A scientist has (Hirsch) index* h *if* h *of his or her* N_p *papers have at least* h *citations each and the other* $(N_p - h)$ *papers have* $\leq$ h *citations each [3].*

All papers by a scientist that have at least h *citations form his or her* Hirsch core *[14].*

The proposed citation measure, called aH-index, is designed to numerically evaluate the following intuition: the impact of a scientist on a research community is proportional to the number of his or her publications that are cited by the members of the community. We are seeking for a measure reflecting how much and how often do different researchers cite the publications of the evaluated scientist.

Definition 2 (aH-index). *A scientist has aH-index* a *if* a *of the* N_c *researchers, that cite his or her work, cite at least* a *his or her publications each and the other* $(N_c - a)$ *researchers cite* $\leq$ a *publications.*

The aH-index comprehends certain qualitative aspects of the citation response. A high aH-index is awarded to scientists who exhibit prolific publication activities (i.e. they produce a high number of papers) and at the same time achieve a significant impact on the research community (i.e. a high number of other researchers cite a high number of their publications). Such a high aH-index indicates that there is a large group of other researchers that follow the work of the evaluated scientist and cite many of his or her publications.

Nevertheless, a high value of the aH-index can be also obtained by an extremal behavior, contradicting the intuition mentioned above. Let us assume that a group of 20 authors writes a single research paper in which they cite 20 different publications of a single scientist. Such a scientist is immediately awarded 20 points of aH-index due to this research paper alone. The impact of such publication on the scientific community as a whole is, however, questionable (it has been referred to in a single work only). The likelihood of such extremal behavior in current bibliographic datasets is a subject of our ongoing investigation.
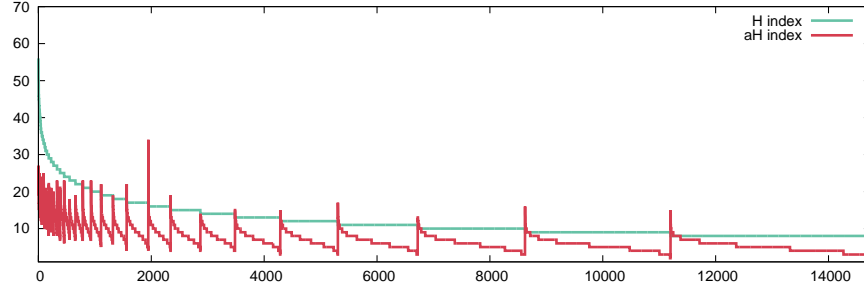
4 Experiments

The properties of the aH-index is initially investigated using a real-world bibliographical dataset. The dataset was aggregated under the Arnet Miner project[15] and is publicly available on its website[1]. It consists of 2,092,209 papers, 8,024,869 citations between papers, and 1,712,433 authors with related information. This dataset is large enough to allow deep analysis but suffers from several problems, e.g. missing author from a publication, author duplication and misspelling, etc. Nevertheless, it is a unique real-world bibliographic dataset that can be used to validate properties of scientometric indicators. After pre-processing, we used a subset of dataset containing $14,744$ scientist with an H-index of at least 8 for our initial computational experiments. The results of this dataset are used to formulate several hypotheses about the proposed aH-index. The confirmation of these hypotheses requires a thorough statistical analysis of detailed author data that is beyond the scope of this short paper. A thorough validation of the outlined concepts and propositions is the subject of our current research.

Proposition 1. *If the aH-index is taken as a complement to the original H-index, it can uncover and quantify certain publication behavior that a scientist used to attain his or her H-index. It can be also used to evaluate how much and how strongly (persistently) he or she influenced other researchers in the community.*

Fig. 1 compares the distributions of the H-index and the proposed aH-index, respectively. It clearly shows that the same value of the H-index can be achieved with both very high and very low values of the aH-index. In other words, the aH-index provides a more detailed and more sensitive assessment of publication activities with respect to received citation response.

---

[1] http://aminer.org/big-scholar-challenge/

Fig. 1. The distribution of H-index and aH-index in test data.

The implications of the aH-index for the evaluation of scientists are, however, subject to further investigation. One possible interpretation of the effect of a high aH-index is that achieving a high H-index is easier when one has a high aH-index. Our motivation for this hypothesis is the following. If there are not many citers that would cite a large number of a scientist's publications (i.e., he or she has low aH-index), but if the scientist has a high H-index, there had to be a wide group of citers that respond to only a few of his publications each. If there is a group of citers that cite a high number of his or her publications, they all contribute to the H-index and attaining a high h is easier. In our work, we expect that very high values of aH-index, close to the values of the H-index, indicate an abnormal behavior of citers. The spikes in the value of aH-index on fig. 1 correspond to such abnormal situations. We also consider a combination of low aH-index and high H-index a sign of abnormal behavior of citers.

As shown in fig. 1, the majority of authors in the test dataset has a higher H-index than aH-index. The arithmetic mean of the ratio of the H-index, h, and aH-index, a, in this dataset for authors with $h \geq 8$ is approximately 1.848. We use this mean as a correction coefficient r, and evaluate its relation to h.

Let $n = \frac{h}{r}$ be a normal value of the aH-index. The value of n for each scientist with H-index h tells, what should his or her aH-index a be if he or she received a citation response typical (average) for a given community. Next, for each scientist consider an *xA-ratio*, x, as a proximity (similarity) of aH-index, a, to the normal value n:

$$x = \begin{cases} \frac{a}{n}, & \text{if } a \leq n \\ \frac{n}{a}, & \text{if } a > n \end{cases}.$$ (1)

In the eq. (1), we expect that each evaluated scientist has at least one publication and at least one citation. We can now use the *xA-ratio* to modify the H-index. Their product, $x \cdot h$, can be interpreted as a correction of the H-index with respect to the behavior of citers. For authors with typical publication activities and receiving citation response typical for given community, this product should not introduce large changes to h. Rather, it provides a subtle correction to the H-index that applies a penalty for the following abnormal types of citation response:

1. citers respond to an above average number of scientist's publications,

2. citers respond to a bellow average number of scientist's publications.

As described earlier, an above average citation response can also be obtained by a single publication in which a large number of authors cite a large number of a scientist's papers. Such publications can be understood as *citation bombs* and it is a question whether they should be omitted when computing a (and eventually h). Table 1 shows the top 20 scientists in the test dataset ordered by the H-index and $x \cdot h$, respectively. In this experiment, we expect that the value of h should be 1.848 times higher than a. As previously mentioned, this was the average ratio of h and a in the test dataset.

We can use the discussed measures to identify scientists with a high h and at the same time an exceptional, i.e. above or bellow average, a. The former case can indicate that such scientist is favored by a community of citers with an outstanding behavior characterized by a tendency to cite a large number of ones publications. An example of in the test dataset is Alessandro F. Garcia with $h = 16$ and $a = 34$. The latter case indicates a very wide impact of the publications of the evaluated scientist, e.g. David R. Karger with $h = 36$ and $a = 11$.

In the next experiments, we set the correction coefficient, r, to the minimum and maximum value in our dataset. The minimum value corresponds to $h = 16, a = 34$ and the maximum value to $h = 18, a = 4$ (Richard A. Caruana). For the minimum value we expect that a is higher than h, the maximum value assumes that h is more than four times higher than a. Table 2 shows that even after correction, most of the top 20 authors feature a higher H-index.

The correlation between $x \cdot h$ and h in the test dataset for 14,744 scientists with $h \geq 8$ and correction coefficient $r = 1.848$ is equal to 0.937. If we focus on the top scientists in the test dataset, we can see that the correlation between $x \cdot h$ and h is 0.839 for 1111 scientists with $h \geq 20$ and 0.720 for 184 scientists with $h \geq 30$. It means that despite being based on the H-index, the corrected evaluations produce different rankings of authors and provide different assessments of a scientist's publication activities.

5   Conclusions

This short paper presents a novel author evaluation measure based on the principles of the well-known H-index and citation response. It outlines the principles of a new citation measure evaluating the behavior of groups of citers responding to scientist's publications. The new measure, called the aH-index, was conceived as a complement to the original H-index and allows the extraction of a new type of information about the evaluated scientists. The properties of this measure were studied using a comprehensive real-world bibliometric dataset. Based on the results of these initial experiments, the role and interpretation of the aH-index was formulated and discussed. A notion of the H-index modified by an aH-index-based coefficient evaluating the citation response received by individual scientists, with respect to typical citation patterns in their community, was investigated.

Table 1. The first 20 authors according the H-index and with reduction for $r = 1.848$

Name	h	aH	Name	h	aH	r=1.848
Hector Garcia-Molina	60	19	Jiawei Han	53	27	49.90
Scott Shenker	56	21	David E. Culler	51	25	46.20
Jiawei Han	53	27	Chris Faloutsos	50	24	44.35
David E. Culler	51	25	Ian Foster	46	24	44.35
Chris Faloutsos	50	24	Anil K. Jain	50	22	40.66
Anil K. Jain	50	22	P S Yu	46	22	40.66
Jeffrey D. Ullman	49	19	Mihir Bellare	42	24	39.77
Ian Foster	46	24	Scott Shenker	56	21	38.81
P S Yu	46	22	W Bruce Croft	44	21	38.81
D Estrin	45	20	Moshe Y. Vardi	42	21	38.81
Jennifer Widom	45	19	M. Naor	42	21	38.81
Ch. H. Papadimitriou	45	17	David J. DeWitt	41	21	38.81
Hari Balakrishnan	45	17	Tom Henzinger	42	25	38.18
Jon M. Kleinberg	45	15	E. M. Clarke	40	23	37.64
W Bruce Croft	44	21	HongJiang Zhang	39	20	36.96
Rakesh Agrawal	43	19	D Estrin	45	20	36.96
Ben Shneiderman	43	18	I. Stoica	41	20	36.96
T. Anderson	43	17	Dan Suciu	40	20	36.96
Mihir Bellare	42	24	Serge Abiteboul	37	20	36.96
Moshe Y. Vardi	42	21	Dan Boneh	37	20	36.96

Table 2. The first 20 authors according the reduced H-index for $r = \frac{16}{34}$ and $r = \frac{18}{4}$

Name	h	aH	$r = \frac{16}{34}$	Name	h	aH	$r = \frac{18}{4}$
Alessandro F. Garcia	16	34	16.00	Hector Garcia-Molina	60	19	42.11
Jiawei Han	53	27	12.71	Scott Shenker	56	21	33.19
David E. Culler	51	25	11.76	Jon M. Kleinberg	45	15	30.00
Tom Henzinger	42	25	11.76	Jeffrey D. Ullman	49	19	28.08
W. van der Aalst	33	25	11.76	Randy Katz	40	13	27.35
Chris Faloutsos	50	24	11.29	Ch. H. Papadimitriou	45	17	26.47
Ian Foster	46	24	11.29	Hari Balakrishnan	45	17	26.47
Mihir Bellare	42	24	11.29	David R. Karger	36	11	26.18
Oded Goldreich	40	24	11.29	L Zhang	37	12	25.35
N. R. Jennings	36	24	11.29	Anil K. Jain	50	22	25.25
Jack J. Dongarra	33	23	10.82	Richard M. Karp	30	8	25.00
Milind Tambe	26	23	10.82	David A. Patterson	41	15	24.90
Micha Sharir	24	23	10.82	David Wagner	33	10	24.20
E. M. Clarke	40	23	10.82	T. Anderson	43	17	24.17
Brad A. Myers	36	23	10.82	Jennifer Widom	45	19	23.68
M. Harman	21	23	10.82	Mihalis Yannakakis	37	13	23.40
Francky Catthoor	20	23	10.82	Ronald Fagin	41	16	23.35
Anil K. Jain	50	22	10.35	Chris Faloutsos	50	24	23.15
Luca Benini	33	22	10.35	Richard Lipton	25	6	23.15
Silvio Micali	33	22	10.35	David E. Culler	51	25	23.12

The work, summarized in this short paper, is indeed preliminary. However, it clearly outlines a number of promising options for a fair and accurate assessment of publication activities combining the best-of-breed scientometric methods with a novel idea coming in part from the area of network science.

References

1. Martin, B.R.: The use of multiple indicators in the assessment of basic research. Scientometrics **36**(3) (1996) 343–362
2. Rinia, E.J., Van Leeuwen, T.N., Van Vuren, H.G., Van Raan, A.F.: Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the netherlands. Research policy **27**(1) (1998) 95–107
3. Hirsch, J.E.: 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) (2005) 16569–16572
4. Costas, R., Bordons, M.: The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level. Journal of Informetrics **1**(3) (2007) 193–203
5. Zhang, C.T.: The e-index, complementing the h-index for excess citations. PLoS One **4**(5) (2009) e5429
6. Alonso, S., Cabrerizo, F.J., Herrera-Viedma, E., Herrera, F.: h-index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics **3**(4) (2009) 273–289
7. Bornmann, L., Mutz, R., Hug, S.E., Daniel, H.D.: A multilevel meta-analysis of studies reporting correlations between the h index and 37 different h index variants. Journal of Informetrics **5**(3) (2011) 346–359
8. Hirsch, J.: An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. Scientometrics **85**(3) (2010) 741–754
9. Wan, J.K., Hua, P.H., Rousseau, R.: The pure h-index: calculating an author's h-index by taking co-authors into account. COLLNET Journal of Scientometrics and Information Management **1**(2) (2007) 1–5
10. McCarty, C., Jawitz, J.W., Hopkins, A., Goldman, A.: Predicting author h-index using characteristics of the co-author network. Scientometrics **96**(2) (2013) 467–483
11. Brooks, T.A.: Evidence of complex citer motivations. Journal of the American Society for Information Science **37**(1) (1986) 34–36
12. Amancio, D.R., Oliveira, O.N., da Fontoura Costa, L.: Three-feature model to reproduce the topology of citation networks and the effects from authors' visibility on their h-index. Journal of informetrics **6**(3) (2012) 427–434
13. Bras-Amorós, M., Domingo-Ferrer, J., Torra, V.: A bibliometric index based on the collaboration distance between cited and citing authors. Journal of Informetrics **5**(2) (2011) 248–264
14. Rousseau, R.: New developments related to the hirsch index. Science Focus **1** (2006) 23–25
15. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: Extraction and mining of academic social networks. In: KDD'08. (2008) 990–998