Novel scientometric indicators to characterize an individual's scientific research output

Antonis Sidiropoulos[⋆], Dimitrios Katsaros[†], and Yannis Manolopoulos[‡]

[⋆]Alexander Technological Education Institute of Thessaloniki, Greece
[†]University of Thessaly, Greece,
[‡]Aristotle University of Thessaloniki, Greece
asidirop@gmail.com,dkatsar@inf.uth.gr,manolopo@csd.auth.gr

Abstract. Despite the rich ecosystem of scientometric indicators that quantify an individual's scientific research output, these indicators either loose singificant information that resides in the citation curve of a scientist's work or they can not capture the multi-dimensional aspects of the scientist's performance. Based on these observations, we outline novel concepts that lead to the introduction of a new scientometric indicator and to the presentation of a new concept that can be used to discover star scientists.

Key words: h-index, perfectionism index, skyline, scientometrics

1 Introduction

The evaluation of the scientific work though scientometric indicators has long attracted significant scientific interest, but recently has become of ground practical and scientific importance. An increasing number of academic institutions are using such indicators to decide faculty promotions, and automated methodologies have been developed to calculate such indicators. Also, funding agencies use them to allocate funds.

Traditionally, the impact of a scholar is measured by the number of authored papers and/or the number of citations. The early metrics are based on some form of (arithmetics upon) the total number of authored papers, the total number of citations, the average number of citations per paper, and so on. Due to the power-law distribution followed by these metrics, they present one or more of the following drawbacks (see also [1]): a) they do not measure the impact of papers, b) they are affected by a small number of "big hits" articles, and c) they have difficulty to set administrative parameters. J. E. Hirsch attempted to collectively overcome all these disadvantages and proposed the *h-index* [1].

Definition 1. *A researcher has h-index h if h of his/her N_p articles have received at least h citations each, and the rest $(N_p - h)$ articles have received no more than h citations.*

This metric calculates how broad the research work of a scientist is. The *h-index* accounts for both productivity and impact. The *h-index* was a really pathbreaking idea, and inspired several research efforts to cure various deficiencies of it, e.g., its aging-ignorant behaviour [2].

The original *h-index* does not take into account the "age" of an article. It may be the case that some scientist contributed a number of significant articles that produced a large *h-index*, but now s/he is rather inactive or retired. Therefore, senior scientists, who keep contributing nowadays, or brilliant young scientists, who are expected to contribute a large number of significant works in the near future but now they have only a small number of important articles due to the time constraint, are not distinguished by the original *h-index*. Thus, arises the need to define a generalization of the *h-index*, in order to account for these facts.

We have defined IN [2] a score $S_c(i)$ for an article i based on citation counting, as follows:

$$S_c(i) = \gamma * (Y(now) - Y(i) + 1)^{-\delta} * |C(i)| \qquad (1)$$

where $Y(i)$ is the publication year of article i and $C(i)$ are the articles citing the article i. If we set δ=1, then $S_c(i)$ is the number of citations that the article i has received, divided by the "age" of the article. Since, we divide the number of citations with the time interval, the quantities $S_c(i)$ will be too small to create a meaningful *h-index*; thus, we use the coefficient γ.

This way, an old article gradually loses its "value", even if it still gets citations. In other words, in the calculations we mainly take into account the newer articles[1]. Therefore, we define a novel citation index for scientist rankings, the *contemporary h-index*, expressed as follows:

Definition 2. *A researcher has contemporary h-index h_c, if h_c of its N_p articles get a score of $S_c(i) \geq h_c$ each, and the rest $(N_p - h_c)$ articles get a score of $S_c(i) \leq h_c$.*

The original *h-index* does not take into account the year when an article acquired a particular citation, i.e., the "age" of each citation. For instance, consider a researcher who contributed to the research community a number of really brilliant articles during the decade of 1960, which, say, got a lot of citations. This researcher will have a large *h-index* due to the works done in the past. If these articles are not cited anymore, it is an indication of an outdated topic or an outdated solution to the problem. On the other hand, if these articles continue to be cited, then we have the case of an *influential mind*, whose contributions continue to shape newer scientists' minds. There is also a second very important aspect in aging the citations. There is the potential of disclosing *trendsetters*, i.e., scientists whose work is considered pioneering and sets out a new line of research that currently is hot ("trendy"), thus this scientists' works are cited very frequently.

To handle this, we take the opposite approach than *contemporary h-index*'s; instead of assigning to each scientist's article a decaying weight depending on its

[1] Apparently, if δ is close to zero, then the impact of the time penalty is reduced, and, for $\delta = 0$, this variant coincides with the original *h-index* for $\gamma = 1$.

age, we assign to each citation of an article an exponentially decaying weight, which is as a function of the "age" of the citation. This way, we aim at estimating the impact of a researcher's work in a particular time instance. We are not interested in how old the articles of a researcher are, but whether they still get citations. We define an equation similar to Equation 1, which is expressed as follows:

$$S_t(i) = \gamma * \sum_{\forall x \in C(i)} (Y(now) - Y(x) + 1)^{-\delta} \tag{2}$$

where γ, δ, $Y(i)$ and $S(i)$ for an article i are as defined earlier. We define a novel citation index for scientist ranking, the *trend h-index*, expressed as follows:

Definition 3. *A researcher has trend h-index h_t if h_t of its N_p articles get a score of $S_t(i) \geq h_t$ each, and the rest $(N_p - h_t)$ articles get a score of $S_t(i) \leq h_t$ each.*

Apparently, for $\gamma = 1$ and $\delta = 0$, the *trend h-index* coincides with the original *h-index*.

The concept of *h-index* has been proposed to easily assess a researcher's performance with a single number. However, by using only this number, we lose significant information about the distribution of citations per article in an author's publication list. In the next section, we study an author's citation curve and we define two new areas related to this curve. We call these "penalty areas", since the greater they are, the more an author's performance is penalized. We exploit these areas to establish a new indice, namely PI, aiming at categorizing researchers in two distinct categories: "influentials" and "mass producers"; the former category produces articles which are (almost all) with high impact, and the latter category produces a lot of articles with moderate or no impact at all.

2 Assessment based on the Perfectionism index

The *h-index* has been a well honored concept since it was proposed by Jorge Hirsch, and a lot of variations have been proposed in the literature. Many efforts enhanced the original *h-index* by taking into account age-related issues [2], multi-authorship [3], fractional citation counting [4]. Other works explored its predictive capabilities [5], its robustness to self-citations [6], etc.

Even though there are several hundreds of articles developing variations to the original *h-index*, there is notably little research on making a better and deeper exploitation of the "primitive" information that is carried by the citation curve itself and by its intersection with the 45^o line defining the *h-index*. The projection of the intersection point on the axes creates three areas that were termed in [7], [8], and [9] as the *h-core* area[2], the *tail* area and the *excess* area

[2] We slightly depart from the original terminology, and we use the term core to refer only to the h^2 square area.

(see Figure 1). The core area is a square of size h (depicted by grey color in the figure), includes h^2 citations; the area that lies to the right of the core area is the tail or *lower area*, whereas the area above the core area is the excess or *upper* or e^2 area [9]. Both the absolute and the relative sizes of these areas carry significant information.

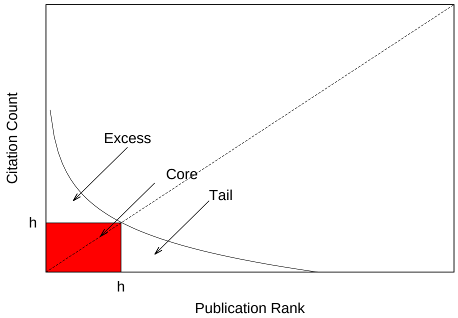


Fig. 1. Citation curve depicting the excess, core and tail areas.

In the next paragraphs, we will define the *penalty area* which forms the basis for the development of the respective scientometric index. Before proceeding further, we summarize in Table 1 some symbols, their interpretations, and the relationships among them, which will be used here.

Symbol	Description	Relations
h	*h-index* of an author	
p	number of articles of an author	
P	set of articles of an author	$\lvert P\rvert = p$
P_H	set of articles that belong in the core area	$\lvert P_H\rvert = h$
P_T	set of articles that belong in the tail area	$\lvert P_T\rvert = p_T = p - h$
p_T	number of articles that belong in P_T	
C	number of citations of an author	
C_i	number of citations for publication i	
C_H	number of citations for publications in P_H	$C_H = \sum_{\forall i \in P_H} C_i$
C_T	number of citations for publications in P_T	$C_T = \sum_{\forall i \in P_T} C_i$
C_E	number of citations in the upper (excess) area	$C_E = C_H - h^2$

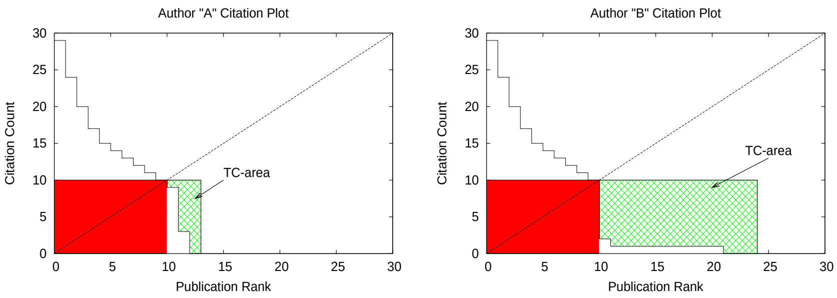
Table 1. Basic symbols and their interpretation.

It is intuitive that long tails and light-weight tails reduce an author's articles' collective influence. Therefore, we argue such kind of a tail area should be considered as a “negative” characteristic when assessing a scientists's performance. The closer the citations of the tail's articles get to the line $y = h$, the more probable it is for the scientist to increase his *h-index*, and at the same time to be able to claim that practically each and every article he publishes does not get unnoticed by the community.

For this purpose, we define a new area, the *tail complement penalty area*, denoted as *TC-area* with size C_{TC}. The size of the tail complement penalty area is computed as follows:

$$C_{TC} \;=\; \sum_{\forall i \in P_T} (h - C_i) \;=\; h \times (p - h) - C_T. \tag{3}$$

This area is depicted with the green crossing-lines pattern in Figure 2, and fulfilling the motivation behind its definition, it is much bigger for author B than for author A.



(a) (b)

Fig. 2. Citation curves for two sample authors A and B.

The definition of the penalty area allows us to design a new measure which will act as the filter to separate influentials from mass producers. Firstly, let us introduce the concept of *Parameterized Count*, PC, as follows:

$$PC \;=\; \kappa * h^2 + \lambda * C_E + \mu * C_T \tag{4}$$

where κ, λ, μ are integer values. Apparently:

- when $\kappa = \lambda = \mu = 1$, then it holds that $PC = C$,
- when $\kappa = 1 \;\wedge\; \lambda = \mu = 0$, then $PC = h^2$,
- when $\lambda = 1 \;\wedge\; \kappa = \mu = 0$, then $PC = C_E = e^2 = C_h - h^2$,
- when $\mu = 1 \;\wedge\; \kappa = \lambda = 0$, then $PC = C_T$.

By assigning positive values to κ and λ, but negative values to μ, we can favor authors with short and thick tails in the citation curve. However, even this way, we cannot differentiate between the authors A and B of our example.

For this reason, instead of using the tail of the citation curve, we use the tail complement penalty area. Thus, similarly to Equation 4, we define the concept of *Perfectionism Index* as follows:

$$PI \;=\; \kappa * h^2 + \lambda * C_E - \nu * C_{TC} \tag{5}$$

We use the values of $\kappa = \lambda = \nu = 1$. These default values give a straightforward geometrical notion of the newly defined metric. Noticeably, it will appear that PI can get negative values. Thus:

– if an author has $PI < 0$, then we characterize him as a *mass producer*,

– if an author has $PI > 0$, then we characterize him as an *influential*.

Using publication and citation data from the Microsoft Academic Search (MAS) database we present in Table 2 the rank of the top-20 authors by *h-index*. They are truly remarkable scientists with significant contributions to their field. The table also shows their corresponding PI values; it is remarkable that about half of them are characterized as "Mass Producers" (i.e., they have negative PI values).

Author	h		PI		p	C	C/p	change
	val	pos	val	pos				h-PI
Shenker Scott	97	1	5754	52	508	45621	89.81	-51
Foster Ian	93	2	-15510	1287	768	47265	61.54	-1285
Garcia-Molina Hector	92	3	-17423	1299	605	29773	49.21	-1296
Estrin Deborah	90	4	5348	62	479	40358	84.25	-58
Ullman Jeffrey	86	5	11267	18	460	43431	94.42	-13
Culler David	84	6	7552	38	386	32920	85.28	-32
Tarjan Robert	83	7	2888	117	405	29614	73.12	-110
Towsley Don	82	8	-31929	1318	793	26373	33.26	-1310
Kanade T.	81	9	-20753	1309	742	32788	44.19	-1300
Haussler David	81	10	10952	19	335	31526	94.11	-9
Jain Anil	81	11	-11474	1236	590	29755	50.43	-1225
Papadimitriou Christos	80	12	-5897	968	506	28183	55.70	-956
Katz Randy	78	13	-27820	1317	757	25142	33.21	-1304
Pentland Alex	77	14	-1242	724	509	32022	62.91	-710
Han Jiawei	77	15	-15410	1285	653	28942	44.32	-1270
Jordan Michael	75	16	-1062	717	499	30738	61.60	-701
Karp Richard	75	17	7231	41	377	29881	79.26	-24
Zisserman A.	75	18	210	263	421	26160	62.14	-245
Jennings Nick	74	19	-15718	1289	626	25130	40.14	-1270
Thrun S.	74	20	-5789	958	445	21665	48.69	-938

Table 2. Ranking by *h-index* (top-20 scientists).

Therefore, the development of scientometric indicators can be used for discovering the scientist's "publishing habits".

3 Assessment based on skylines

Admittedly, despite the plethora of scientometric indices proposed to rank scientists, none of them can fully capture the performance and impact of a scientist, since each index quantifies only one or a few aspects of his/her multifarious performance. Therefore, it would be better if we could identify those individuals that are not worse than any other individual with respect to all considered indicators. This is the concept described by the *skyline*, and calculated by the respective operator [10]. We apply this concept to scientometric ranking and using the following indicators

(h) *h-index*: as a proxy for both impact and productivity [1],

(h^t) contemporary *h-index*: as an indicator for identifying trend-setters [2],

(h^c) trend *h-index*: as an indicator of rising stars [2],

(PI) perfectionism index: as an indicator of highly-cited, laconic scientists,

(C) number of citations: as an indicator of high-impact scientists,

($\frac{C}{p}$) C divided by the number of articles: as an indicator of high-impact normalized to productivity.

we examine whether there is any correlation among a scientist winning the "Edgar F. Codd Innovations", the "SIGMOD Test-of-Time", the "VLDB 10 Years", and "PODS Test-of-Time" awards. Table 3 presents the results.

Name	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	00	01	02	03	04	05	06	07	08	09	10	11	12	13
Abiteboul Serge												s	s	s	sC	s					S				P					
Agrawal Rakesh																	C			S	V	s	s	s	s	s	s	s	s	s
Babcock Brian																									s	s	s	s	s	s
Bancilhon Francois								s	s	s	s	sV		s	s	s	s	s	s											
Bayer Rudolf																		C												
Beeri Catriel				s	s	s	s	s	s																					
Bernstein Philip	s	s	s	s	s	s	s	s	s	s	sC	s	s	s	s	s	s	s										V		
Carey Michael													V								S	C								
Ceri Stefano																	V													C
Chamberlin Don	s	s	s	s	s	s	s	s	s											C										
Chaudhuri Surajit																								V				C		
Codd E.F.	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s
Datar Mayur																											s	s	s	s
Dayal Umeshwar																		V									C			
DeWitt David							s	s	s	s	s	sC	sV	s	s	s					S						S			
Fagin Ronald	s	s	s	s	s	s	s	s					s	s	s	s	s				C							P		
Faloutsos Christos														V		s												s		
Garcia-Molina H.																C			s	s	s	sS	s	s	s	s	s	s	s	s
Goodman Nathan	s	s				s		s						s	s	s	s	s	s	s										
Gray Jim										sC	s				V															
Halevy Alon																							sV	s	s	s	s	s	s	s
Hammer Michael															s	s	s	s	s	s										
Imielinski Tomasz																s	s	s	sV	sS										
Kim Won								s	s	s	s	sV	s	s					S											
Kitsuregawa Masaru																										C				
Lindsay Bruce																													C	
Lorie Raymond		s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s		s	s									
Maier David								s	s			s	s	C	s	s														
Mohan C.													C			V														
Motwani Rajeev																							s	s	s	s	s	s	sV	s
Papadimitriou C.																							s	s	s	s	s	s	s	
Papakonstantinou																	s													
Quass Dallan																		s	s	s										
Selinger Patricia																			C											
Stonebraker M.				s	s	s	s	s	sC	s																				
Swami Arun																		s	s	sS	s	s	s	s	s	s	s	s	s	s
Ullman Jeffrey	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	s	sCS	s	s	s	s	s	s	s
Vardi Moshe																									CP					
Widom Jennifer																s	SV					S	s	sC	s	s	s	s	s	s

Table 3. 6-D skylines (h-$index$, PI, h^t, h^c, C, C/p) members vs. "E.F. Codd Innovations", "SIGMOD Test-of-Time", "VLDB 10 Years", and "PODS Test-of-Time" award winners.

We use the symbols 's' and 'C' to denote presence in the skyline and a "Codd" award winner, respectively, and the symbols 'S', 'V' and 'P' to denote a "SIGMOD Test-of-Time", "VLDB 10 Years", and "PODS Test-of-Time" award winner at the specific year where the symbol appears. We are able to observe a really very good match between a "skyline person" and an award winner. Nevertheless, there are winners of (multiple) awards such as M. Carey, S. Chaudhuri, U. Dayal, C. Mohan which do not appear in the skyline; this fact reinforces the value of peer review in academic performance assessment, even though we can not preclude the case that there is some combination of skyline dimensions where these persons might appear in.

4 Conclusion

In this article, we briefly presented some scientometric indicators for assessing the work of individual scientists. We believe that scientometric indicators are not a panacea, and we should work a lot before applying a set of them to characterize the achievements of a scholar. However, as Lord Kelvin said "When you can measure what you are speaking about, and express it in numbers, you know something about it. But when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind."

Acknowledgments

D.Katsaros acknowledges the support of the Research Committee of the University of Thessaly through the project "Web observatory for research activities in the University of Thessaly".

References

1. J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National Academy of Sciences*, vol. 102, no. 46, pp. 16 569–16 572, 2005.
2. A. Sidiropoulos, D. Katsaros, and D. Manolopoulos, "Generalized Hirsch h-index for disclosing latent facts in citation networks," *Scientometrics*, vol. 72, no. 2, pp. 253–280, 2007.
3. J. E. Hirsch, "An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship," *Scientometrics*, vol. 85, no. 3, pp. 741–754, 2010.
4. D. Katsaros, L. Akritidis, and P. Bozanis, "The f index: Quantifying the impact of coterminal citations on scientists' ranking," *Journal of the American Society for Information Science and Technology*, vol. 60, no. 5, pp. 1051–1056, 2009.
5. J. E. Hirsch, "Does the h index have predictive power?" *Proceedings of the National Academy of Sciences*, vol. 104, no. 49, pp. 19 193–19 198, 2007.
6. M. Schreiber, "Self-citation corrections for the Hirsch index," *Europhysics Letters*, vol. 78, no. 3, 2007.
7. R. Rousseau, "New developments related to the Hirsch index," *Science Focus*, vol. 1, no. 4, pp. 23–25, 2006.
8. F. Y. Ye and R. Rousseau, "Probing the h-core: An investigation of the tail-core ratio for rank distributions," *Scientometrics*, vol. 84, no. 2, pp. 431–439, 2010.
9. C.-T. Zhang, "The e-index, complementing the h-index for excess citations," *PLoS One*, vol. 4, no. 5, 2009.
10. S. Börzsönyi, D. Kossmann, and K. Stocker, "The Skyline operator," in *Proceedings of the IEEE International Conference on Data Engineering (ICDE)*, 2001, pp. 421–430.