Identification of Influential Scientists vs. Mass Producers by the Perfectionism Index

Antonis Sidiropoulos[1], Dimitrios Katsaros*[2], and Yannis Manolopoulos[3]

[1]Dept. of Information Technology, Alexander Technological Educational Institute of Thessaloniki, Greece
[2]Dept. of Electrical & Computer Engineering, University of Thessaly, Volos, Greece
[3]Dept. of Informatics, Aristotle University, Thessaloniki, Greece

Abstract

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 this article, 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 new indices, namely PI and XPI, 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. We evaluate the merits mainly of PI as a useful tool for scientometric studies.

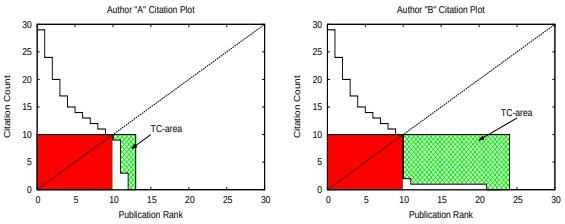
1 Introduction

The *h-index* has been a well honored concept since it was proposed by Jorge Hirsch [1]. 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 [2], [3], and [4] as the *h-core-square* area[1], the *tail* area and the *excess* area (see Figure 2). 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 [4]. Both the absolute and the relative sizes of these areas carry significant information.

1.1 Motivation and contributions

During the latest years an abundance of scientometric indices have been published to evaluate the academic merit of a scientist. Despite the debate around the usefulness of any index in general, they remain an indispensable part of the evaluation process of a scientist's academic merit. The ideas behind the *h-index* philosophy was so influential, that the vast majority of the proposed indices are about some variant or extension of the *h-index* itself. Despite the wealth and sophistication of the proposed indices, we argue that the relevant literature did not strive for an *holistic* consideration of the information carried by the citation curve and by the 45^o line. In the next paragraph we will present the motivating idea with a simple example.

Let us consider author *A* who has published 13 articles, and author *B* who has published 24 articles with citation distributions {29, 24, 20, 17, 15, 14, 13, 12, 11, 10, 9, 3, 0} and { 29, 24, 20, 17, 15, 14, 13, 12, 11, 10, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0} respectively. Both authors have the same "macroscopic" characteristics in terms of the number of citations, i.e., they both have the same total number of citations, identical core areas and h-indices equal to 10, identical excess areas with 65 citations there, and the same number of citations in the tail area, namely 12. However, author *A* has only 3 articles in his tail area, whereas author *B* has 14 articles.



(a) (b)

Figure 1: Citation curves for two sample authors *A* and *B*.

2 Penalty areas and the Perfectionism Indices

We now get back to the motivating example presented in the previous section, and we illustrate graphically their citation distributions (see Figure 1). We depict with red color the h-core area of each author. 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}. This area is depicted with the green crossing-lines pattern in Figure 1, and fulfilling the motivation behind its definition, it is much bigger for author *B* than for author *A*.

If we push further the idea of the tail complement penalty area, we can think that "ideally" an author could publish p papers with p citations each and get an *h-index* equal to p. Thus, a square $p \times p$ could represent the minimum number of citations to achieve an *h-index* value equal to p. Along the spirit of penalizing long and thin citation curves, we can define another area in the citation curve: the *ideal complement penalty area* (*IC-area*), which is the complement of the citation curve with respect to the square $p \times p$. Figure 2 illustrates

*Corresponding author: Dimitrios Katsaros (dkatsar@inf.uth.gr)

[1]In the sequel of the article for the sake of simplicity, we use the term h-core and h-core-square interchangeably.

graphically the IC-area with the green crossing-lines pattern.

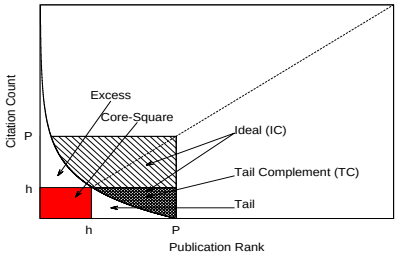


Figure 2: Graph illustrating all (existing and proposed) areas.

The definition of the penalty areas in the previous subsection, allows us to design two new metrics which will act as the filter to separate influential from mass producers.

We define the concept of *Perfectionism Index based on TC-area* as follows:

$$PI = \kappa * h^2 + \lambda * C_E - \nu * C_{TC} \quad (1)$$

Where h^2 is the h-square area, C_E is the excess area and T_C is the tail complement area. In the experiments that will be reported in the next sections, we will 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:

- a negative PI characterizes a *mass producer*,
- a positive PI characterizes an *influential*.

In the same way as the PI's definition, we define an extremely perfectionism metric, the *Extreme Perfectionism Index*, taking into account the ideal complement penalty area, as follows:

$$XPI = \kappa * h^2 + \lambda * C_E + \mu * C_T - \nu * C_{IC}. \quad (2)$$

Where C_T is the tail area and C_{IC} is the Ideal complement area. As in the previous case, we will assume that $\kappa = \lambda = \mu = \nu = 1$.

3 Experiments

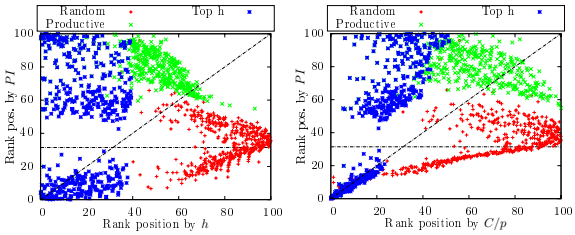
We have performed various experiments based on the Microsoft Academic Search (MAS) database acquired during the period December 2012 to April 2013. We compiled 3 datasets. The first one consists of randomly selected authors (named “Random" henceforth). The second one includes highly productive authors (named “Productive"). The last consists of authors in the top *h-index* list (named “Top h").

The first question that needs to be answered is whether a new index offers something new and different compared to the existing (hundreds of) indices. The answer is positive; our metric separates the rank tables into two parts independently from the rank positions.

In Figure 3(a) the x-axis denotes the rank position (normalized percentagewise) of an author by *h-index*, whereas the y-axis denotes the rank position by PI. Each point denotes the position of an author ranked by the two metrics.

All of our experiments show that PI and XPI are not correlated with any existing index. Also, our new metrics split the authors set into two sets based on the threshold of the value of zero for PI (or XPI). Thus, our new metrics split the author set into two groups: The Perfectionisms and the Mass producers.

We have also performed an experiment to study the behavior of *h-index* and PI with respect to self-citations. We have shown that PI and XPI are not affected by self-citations. This is another advantage of the proposed metric.



(a) h vs. PI (unioned) (b) C/p vs. PI (unioned)

Figure 3: Correlation of PI to standard bibliometric indices. (Q-Q plots: X- and Y-axis denote normalized rank positions (%).)

4 Conclusions

The development of indices to characterize the output of a scientist is a significant task not only for funding and promotion purposes, but also for discovering the scientist’s “publishing habits". Motivated by the question of discovering the steadily influential scientists as opposed to mass producers, we have defined two new areas on an scientist’s citation curve:

- The *tail complement penalty area* (TC-area), i.e., the complement of the tail with respect to the line $y = h$.
- the *ideal complement penalty area* (IC-area), i.e., the complement with respect to the square $p \times p$.

Using the aforementioned areas we defined two new metrics:

- The *perfectionism index based on the TC-area*, called the PI index.
- The *extreme perfectionism index based on the IC-area*, called the XPI index.

We have performed an experimental evaluation of the behavior of the PI and XPI indices. For this purpose, we have generated three datasets (with random authors, prolific authors and authors with high *h-index*) by extracting data from the Microsoft Academic Search database. Our contribution is threefold:

- We have shown that the proposed indices are uncorrelated to previous ones, such as the *h-index*.
- We have used these new indices, in particular PI, to rank authors in general and, in particular, to split the population of authors into two distinct groups: the “influential" ones with $PI > 0$ vs. the “mass producers" with $PI < 0$.
- Also, we have shown that ranking authors with the PI index is more robust than *h-index* with respect to self-citations, and we applied it to rank individual scientists offering some explanations for the reasons behind their publishing habits.

References

[1] J. E. Hirsch. An index to quantify an individual’s scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572, 2005.

[2] R. Rousseau. New developments related to the Hirsch index. *Science Focus*, 1(4):23–25, 2006.

[3] F. Y. Ye and R. Rousseau. Probing the *h*-core: An investigation of the tail-core ratio for rank distributions. *Scientometrics*, 84(2):431–439, 2010.

[4] C.-T. Zhang. The *e*-index, complementing the *h*-index for excess citations. *PLoS One*, 4(5), 2009.