Scientometrics
DOI 10.1007/s11192-014-1515-0

Ranking and identifying influential scientists versus mass producers by the Perfectionism Index

Antonis Sidiropoulos · Dimitrios Katsaros · Yannis Manolopoulos

Received: 26 May 2014

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 Perfectionism Index and eXtreme Perfectionism Index (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. Using data from Microsoft Academic Service, we evaluate the merits mainly of PI as a useful tool for scientometric studies. We establish its effectiveness into separating the scientists into influentials and mass producers; we demonstrate its robustness against self-citations, and its uncorrelation to traditional indices. Finally, we apply PI to rank prominent scientists in the areas of databases, networks and multimedia, exhibiting the strength of the index in fulfilling its design goal.

Keywords Ranking · h-Index · Citation analysis · Bibliometrics

PI, note here that PI is not related and should not be confused with the term Perfect Index (Woeginger in Math Soc Sci 56: 224–232, 2008).

A. Sidiropoulos (✉)
Department of Information Technology, Alexander Technological Educational Institute of Thessaloniki, Thessaloniki, Greece
e-mail: asidirop@it.teithe.gr

D. Katsaros
Department of Electrical and Computer Engineering, University of Thessaly, Thessaly, Greece
e-mail: dkatsar@uth.gr

Y. Manolopoulos
Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
e-mail: manolopo@csd.auth.gr

Published online: 23 December 2014

Introduction

The *h-index* has been a well honored concept since it was proposed by Hirsch (2005). A lot of variations have been proposed in the literature, see for instance the references within (Alonso et al. 2009). Many efforts enhanced the original *h-index* by taking into account age-related issues (Sidiropoulos et al. 2007), multi-authorship (Hirsch 2010), fractional citation counting (Katsaros et al. 2009), the highly cited articles (Egghe 2006). Other works explored its predictive capabilities (Hirsch 2007), its robustness to self-citations (Schreiber 2007), etc. Some of the proposals have been implemented in commercial and free software, such as Matlab[1] and the Publish or Perish software.[2]

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° line defining the *h-index* . The projection of the intersection point on the axes creates three areas that were termed in (Rousseau 2006; Ye and Rousseau 2010; Zhang 2009) as the *h-core-square* area,[3] the *tail* area and the *excess* area (see Fig. 1). The core area is a square of size h (depicted by grey color in the figure), includes h^2 citations and it is also called Durfee square area (Anderson et al. 2008); 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 (Zhang 2009). Both the absolute and the relative sizes of these areas carry significant information. The absolute size of the excess and core areas were directly used for the definition of e-index and *h-index* ; part of the absolute size of the tail area was used by García-Pérez (2012) to create a vector of h-indices; the relative size of the core to the tail area (without taking into account the tail length) was used by Ye and Rousseau (2010) for similar purposes, etc. (For a complete review of the relevant bibliography cf. section "Relevant work".)

The common characteristic of all these works is that they develop indices to "break ties", i.e., to differentiate between scientists with equal h-indices.

We believe that the latent information carried by these areas is not adequately explored, and most significantly, it can be used in a different way, not as a plain tie-breaker, but as a "first-class" citizen in the scientometric indices ecosystem.

Rosenberg (2011) took the first step towards this goal; he provided a *qualitative characterization* for the scientists with many citations in the upper area and a few citations in the tail area by referring to them as *perfectionists*. He referred to the authors with few citations in the upper area and many citations in the tail area as *mass producers*, since they have a lot of publications but mostly of low impact. Finally, he referred to the rest of the scientists as the *prolific* ones. The origin of this terminology is quite old; it was proposed by Cole and Cole (1967), and subsequently studied further by Feist (1997).

Motivated by Rosenberg's classification scheme, we pose the following question: "Can we develop a quantitative methodology for identifying those scientists who are truly laconic and constantly influential compared to those who produce a mass of papers but relatively few of them contribute to their *h-index* ?"

In the present paper, we will present a methodology and an easy calculated criterion to categorize a scientist in one of two distinct categories: either an author is a "mass producer", i.e., he has authored many papers with relatively few citations or he is

[1] http://www.mathworks.de/matlabcentral/fileexchange/28161-bibliometrics-the-art-of-citations-indices.

[2] http://www.harzing.com/pop.htm.

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

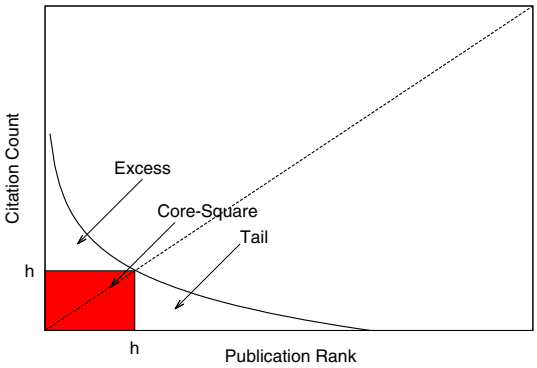

Fig. 1 Citation curve depicting the excess, core and tail areas. (Color figure online)

"influential", i.e., most of his papers have an impact because they have received a significant number of citations. This methodology will indirectly highlight the "attitude" towards publishing. Some scientists are acting in a laconic way, in the sense that they are not fond of having published "half-baked" articles that are soon superseded by mature and extended versions of their work. Others develop their work in a slow and incremental way, publishing their ideas in a step-by-step fashion producing a lot of moderate impact articles until they hit the big contribution. This attitude may be due to other reasons as well, e.g., the pressure to present published articles as deliverables to a project. In any case, it is not the purpose of the present article to discover and explain those reasons. The sole purpose of our work is to develop metrics that can be used complementary to the traditional ones such as the *h-index* , in order to separate the steadily influential authors from the mass producers. At this point we need to emphasize that the concept of "influential" scientists we develop here is not related to the notion of influential nodes in a social network of actors as considered by Basaras et al. (2013).

The area of scientometric performance indicators is very rich, and it is continuously flourishing. Vinkler (2011) provides a brief classification of the traditional and modern scientometric indicators explaining their virtues and shortcomings; it is shown there that Hirsch index is not the only indicator that combines impact and quantity, but π-index (Vinkler 2009) which introduced the concept of "elite set" is another competitor of it. Nevertheless, in this article we use Hirsch index as a basis to expose our ideas claiming that neither Hirsch index is the 'best' one nor that our core methodology applies exclusively to it. We are strongly confident that our ideas can be applied also to the family of indices based on the Impact Factor by penalizing those journals that publish articles which accumulate far less citations than the Impact Factor of the journal they appear in.

The rest of the article is organized as follows. In the next section we will present the relevant literature and then define two new areas in the citation curve. Based on these two new areas, we will establish two new metrics for evaluating the performance of authors in terms of impact. In section "Penalty areas and the Perfectionism Indices" we will present our datasets, which were built by extracting data from the Microsoft Academic Search database, and analyze these data to view the dataset characteristics. Subsequently, we will present the distributions of our new metrics for the above datasets and compare them with other metrics proposed in the literature. Finally, in section "PI in action: Ranking scientists" we will present some of the resulting ranking tables based on the new metrics and *h-index* . Section "Conclusions" will conclude the article.

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° 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 $\{29, 24, 20, 17, 15, 14, 13, 12, 11, 10, 2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0, 0\}$ 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.

At a first glance, we can simply use the number of articles in the tail as a tie-breaker to differentiate between the two authors, and characterize the first one as constantly "influential", and the second one as a "mass producer". But, how can we capture the fact that, in the short term, the first author's *h-index* is more likely to increase. At the same time, we need a way to describe—actually, to penalize—the second author for this long and lightweight tail. Starting from these questions we will define the *penalty areas* and then develop the respective indices. In this article, we do not consider temporal issues, e.g., the time of publication of the articles in the tail area; such issues are part of our on-going work. Specifically, the article makes the following contributions:

- It defines two areas to quantify the fact that some authors publish articles which eventually do not have analogous impact with those that contribute to their *h-index* .
- It develops two new perfectionism indices taking into account the size of the penalty areas. There are the PI and XPI indices, which are statistically uncorrelated to the *h-index* , thus proving that they measure something that is different from what the *h-index* measures.
- Using these indices, it proposes a filter to separate the authors into influential ones and mass producers. This filter partitions the authors irrespectively of their *h-index* , i.e., it can classify two authors as influentials, even if the values of their h-indices are significantly different. This two-way segmentation of the scientists is a significant departure from the earlier, rich classification schemes (Cole and Cole 1967; Feist 1997), since with a *single* integer number and its sign (plus or minus) it can provide rankings, contrary to intuitive mapping schemes such as that by Zhang (2013b).
- It provides a thorough investigation of the indices against the *h-index* for three datasets retrieved by the Microsoft's Academic Search.

At this point, we need to emphasize again that the proposed indices are neither a substitute for any of the already existing metrics nor a tool for identifying "bad" publishing behaviour. They are one more tool in the indices toolbox of someone who wishes to capture the multi-dimensional facet of a scientist's performance.

Relevant work

The original article by Hirsch (2005) created a huge wave of proposals for indices attempting to capture the academic performance of a scientist. It is characteristic that at the time of writing the present article, the *h-index* 's article had more than 3850 citations in Google Scholar. Since the focus of the present manuscript is not about the *h-index* in general, but about the exploitation of the information in the *tail area*, we will survey only the articles relevant to the usage and mining of that part of the citation.

Ye and Rousseau (2010) studied the evolution of tail-core ratio as a function of time, and later extended their study by Liu et al. (2013) including a few more ratios among the three areas. Similar in spirit is the work reported by Chen et al. (2013), which examines variations of the ratios across scientific disciplines. Baum (2012) introduced the ratio (the relative citedness) of the few, highly-cited articles in a journal's h-core and the many, infrequently-cited articles in its h-tail as a way to improve journals' Impact Factors.

Having as motive to consider each and every citation under the *whole* citation curve (and therefore under the tail area as well), Anderson et al. (2008) proposed a fractional citation counting scheme based on Ferrers graphs. Later, Franceschini and Maisano (2010) recognized the weaknesses of that scheme and proposed the Citation Triad method; both indices are striving to exploit the information under the whole citation curve in a way that creates a strictly monotonic (increasing) index for every new citation added to the curve, which is completely different to what we propose.

A kind of "quantization" scheme for the citation curve and the creation of multiple Durfee squares under that curve was proposed by García-Pérez (2012). The output of that method was a vector (i.e., multiple h-indexes) as a measure of the scientific performance. However, the method simply transformed the task of comparing different citation curves into the problem of comparing vectors, without setting clear rules. A study of the contribution of the excess, core and tail areas to the entire citation curve was performed by Bornmann et al. (2010) proving that this contribution varies across scientists. The study provided also a regression model for determining the most visible article of a scientist. The position of the centroids of the core and tail area was used by Kuan et al. (2011) as an index for comparing different scientists providing only straightforward characterizations for high-low impact and productivity. Along these lines of research, Zhang (2013b) proposed a triangle mapping technique to map the three percentages (of the excess, core and tail area) of these citations onto a point within a regular triangle; by viewing the distribution of the mapping points, different shapes of citation curves can be studied in a perceivable form. The work described by Dorta-González and Dorta-González (2011) sought selective and large producers considering only a part of the excess and a part of the tail area, thus again neglecting a part of the tail area which carries significant information.

The most closely relevant articles to our work are these from Rosenberg (2011) and Zhang (2013a). Rosenberg (2011) described a three-class scheme for scientists' classification based on the length and thickness of the tail of the citation curve. Zhang (2013a) proposed the h' index as a quantitative measure to discover which scientist belongs to which one of those three categories.

Collectively, the present work differentiates itself by the previous studies in a number of factors: (a) it exploits the full spectrum of information under the citation curve, (b) it is based on the definition of new areas (not below, but above the citation curve), (c) it penalizes those scientists with long and thin tails, (d) it proposes an index that can be used as a filter to separate the constantly influential scientists from the mass producers.

Penalty areas and the Perfectionism Indices

In this section, we will define the *penalty areas* which form the basis for the development of the respective scientometric indices. Before proceeding further, we summarize in Table 1 some basic symbols related to the productivity and impact of a scientist, along with their description and the relations among them.

The tail complement penalty area

We now get back to the motivating example presented in the previous section, and we illustrate graphically their citation distributions (see Fig. 2). We depict with red solid 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}. The size of the tail complement penalty area is computed as follows:

$$C_{TC} = \sum_{\forall i \in P_T} (h - C_i) = h \times (p - h) - C_T. \tag{1}$$

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

The ideal complement penalty area

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,

Table 1 Basic symbols and their interpretation

Symbol	Description	Relations
h	*h-index* of an author	
p	Number of articles of an author	
P	Set of articles of an author	$\|P\| = p$
P_H	Set of articles that belong in the core area	$\|P_H\| = h$
P_T	Set of articles that belong in the tail area	$\|P_T\| = p_T = p - h$
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	$C_H = \sum_{\forall i \in P_H} C_i = R^2$ (Jin et al. 2007)
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	$C_E = C_H - h^2$

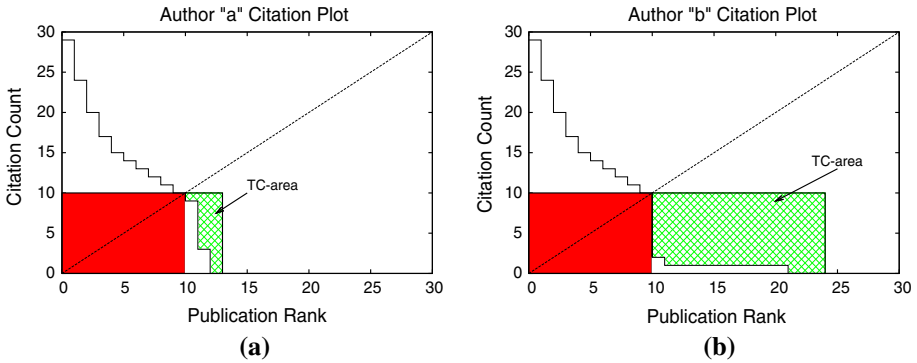

Fig. 2 Citation curves for two sample authors (**a**, **b**). (Color figure online)

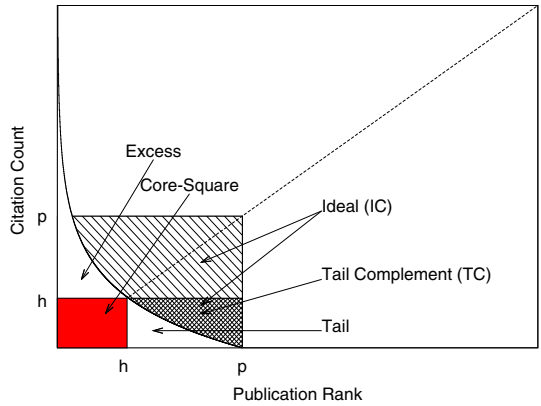

Fig. 3 Graph illustrating all (existing and proposed) areas. (Color figure online)

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 tails, 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 3 illustrates graphically the IC-area with the green crossing-lines pattern. The size of the IC-area (C_{IC}) can be computed as follows:

$$C_{IC} = \sum_{\forall i \in P \wedge C_i < p} (p - C_i). \tag{2}$$

Apparently, this area does not depend on the *h-index* value, as it holds for the case of the TC-area. Notice that the IC-area includes the TC-area defined in the previous paragraph. We realize of course that it is hard (if possible at all) to find scientists—with sufficiently large *h-index* , and—with zero-sized TC-area. Therefore, the index derived by this area is not expected to provide significant insights into scientists' performance.

The new scientometric perfectionism indices: PI and XPI

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. Firstly, let us introduce the concept of *Parameterized Count*, *PC*, as follows:

$$PC = \kappa * h^2 + \lambda * C_E + \mu * C_T \tag{3}$$

where κ, λ, μ are real numbers. Therefore, we define PC as the parameterized addition of the three areas we defined earlier. 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. Even in 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 Eq. 3, we define the concept of *Perfectionism Index based on TC-area* as follows:

$$\mathrm{PI} = \kappa * h^2 + \lambda * C_E - \nu * C_{TC} \tag{4}$$

In the experiments that will be reported in the next sections, we will use the values of $\kappa = \lambda = \nu = 1$. We experimented with various combination of values for the parameters, but we used integer values equal to one, because 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 $\mathrm{PI} < 0$, then we characterize him as a *mass producer*,
- if an author has $\mathrm{PI} > 0$, then we characterize him as 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:

$$\mathrm{XPI} = \kappa * h^2 + \lambda * C_E + \mu * C_T - \nu * C_{IC}. \tag{5}$$

As in the previous case, we will assume that $\kappa = \lambda = \mu = \nu = 1$. We will show in the experiments, that very few authors have positive values for this metric. Using the previously defined perfectionism indices, the resulting values for authors *a* and *b* are shown in Table 2. Author *a* has greater values than author *b* for both XPI and PI perfectionism indices. This is a desired result.

Before proceeding to the next section, which describes the detailed experiments that demonstrate the merits of the new indices, we provide an additional example of five authors with different publishing patterns, as an extension to our artificial motivating example which was presented in the beginning of the article. We use only initials but they refer to real persons and their data. In Table 3 we present the raw data (i.e., *h-index* , number of publications *p* and number of citations *C*) of 5 authors[4] selected from Microsoft

[4] We selected authors with relatively small number of publications and citations for better readability of the figures.

Table 2 Traditional and proposed indices for authors a and b

Author	p	C	h	C_T	C_E	C_H	C_{TC}	PI	C_{IC}	XPI
a	13	177	10	12	65	165	18	147	33	144
b	24	177	10	12	65	165	128	37	404	−227

Academic Search.[5] The last column shows the calculated PI values, which can be positive as well as negative numbers. In Fig. 4 we present citation plots for these five authors.

In Fig. 4a we compare three authors: AuthorA, AuthorB and AuthorC. They correspond to real names but we have preferred to present them anonymously. They all have an *h-index* equal to 10. Note that AuthorA has a comparatively large number of publications but the citation curve is cropped to focus on the lower values. As he has the bigger and longest tail (red line with crosses), he could be characterized as a "mass producer". This is reflected in a PI value of −2,505 as shown in Table 3. AuthorB (green line with diagonal crosses) has shorter tail than AuthorA and higher excess area. From the same table we remark that his PI value is 11 (i.e. close to zero). Finally, the last author of the example, AuthorC (blue line with asterisks), has similar tail with AuthorB but a bigger excess area (e^2). Definitely, he demonstrates the "best" citation curve out of the three authors of the example. In fact, his PI score is 717, higher than the respective figure of the other two authors.

In Fig. 4b, again we compare three authors: AuthorD, AuthorE and AuthorC. The first two have *h-index* value equal to 15. Comparing those, it seems that AuthorD (red line with crosses) has a better citation curve than AuthorE (green line with diagonal crosses) because he has a shorter tail and a bigger excess area. Indeed, the first one has PI = 717, whereas the second one has PI = −2,523. AuthorC (blue line with asterisks) has a smaller tail and a big excess area but since there is a difference in *h-index* we cannot say for sure if he must be ranked higher or not than the others.

In Fig. 4c we have scaled the citation plots so that all lines cut the line $y = x$ at the same point. From this plot, it is shown that AuthorC has a better curve than AuthorE because he has a shorter tail and a bigger excess area. When comparing AuthorC to AuthorD, we see that the latter has a longer tail but also a bigger excess area. Both curves show almost the same symmetry around the line $y = x$. That is why they both have similar PI values. This is a further positive outcome as authors with different quantitative characteristics (say, a senior and a junior one) may have similar qualitative characteristics, and thus be classified together.

In general, there is no upper (lower) limit for the size of the PI index. This depends on the productivity (number of articles) and impact (citation) distribution of each scientist. However, a positive PI means that the scientist is a perfectionist one, and a negative PI indicates a mass-producer (Table 4).

Experiments

In this section, we will present the results of the evaluation of the proposed indexes. The primary goal of our experimentation is to study the merits of the PI index, since our results confirmed that the severe penalty that XPI imposes makes it a less useful scientometric tool. Firstly, we will explain the procedures for dataset acquisition, then we will present their characteristics, and finally we will give the results that concern the evaluation of PI.

[5] http://academic.research.microsoft.com/.

Table 3 Computed *h-index* and PI values for 5 sample authors

Author	*h*	*p*	*C*	PI
AuthorA	10	319	585	−2,505
AuthorB	10	49	391	11
AuthorC	10	48	1097	717
AuthorD	15	105	2,040	690
AuthorE	15	259	1,137	−2,523

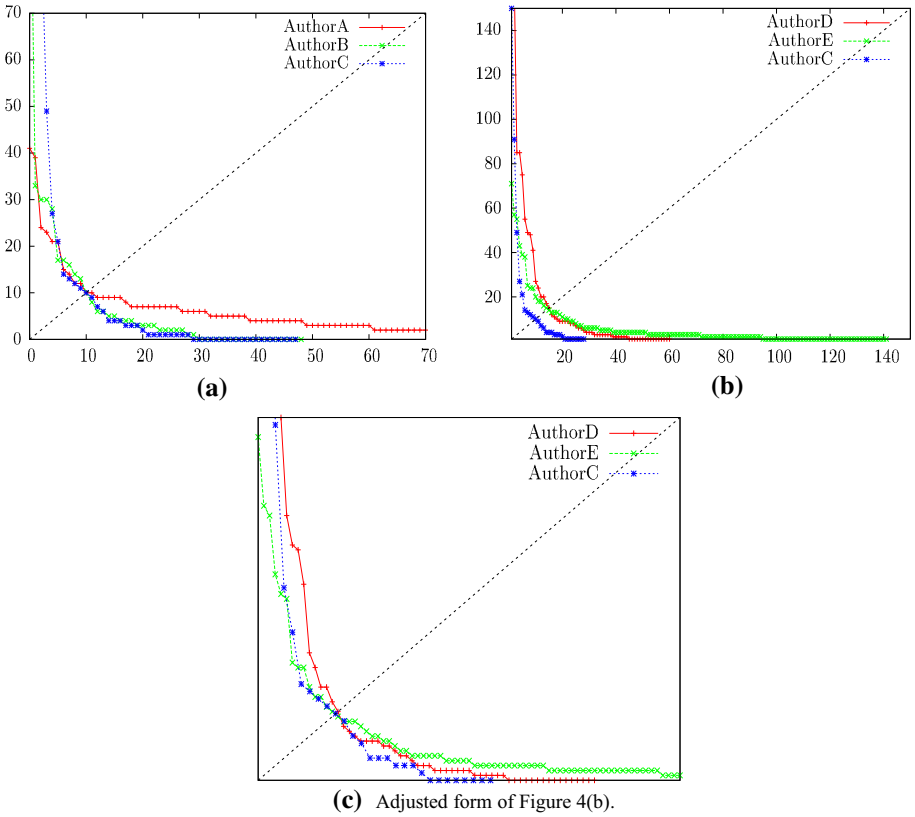

Fig. 4 Real examples. (Color figure online)

Datasets acquisition and characterization

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 one consists of authors in the top *h-index* list (named "Top h"). The publication and the citation data were extracted from the Microsoft Academic Search (MAS) database using the MAS API.

Table 4 Symbols used and their interpretation

Symbol	Description	Relations
C_{TC}	Tail complement area	$C_{TC} = \sum_{\forall i \in P_T} (h - C_i)$
C_{IC}	Ideal complement area	$C_{IC} = \sum_{\forall i \in P \wedge C_i < p} (p - C_i)$
PC	Parameterized Count	$PC = \kappa * h^2 + \lambda * C_E + \mu * C_T$
PI	Perfectionism Index	$\text{PI} = \kappa * h^2 + \lambda * C_E - \nu * C_{TC}$
XPI	eXtreme Perfectionism Index	$\text{XPI} = \kappa * h^2 + \lambda * C_E + \mu * C_T - \nu * C_{IC}$
κ	h-core area factor	
λ	Excess area factor	
μ	Tail area factor	
ν	Penalty area (Tail complement or Ideal complement) factor	

Table 5 Statistics of the datasets used in our study

	Random	Productive	Top h
No. of authors	500	500	500
No. of publications	25,679	223,232	149,462
No. of P/Author	51	446	298
Min No. of P/Author	10	354	92
Max No. of P/Author	768	1,172	1,172
No. of Citations	410,280	3,197,880	5,015,971
No. Cit/Author	820	6,395	10,031
Min No. of Cit/Author	1	25	4,405
Max No. of Cit/Author	47,263	47,263	47,263

The dataset "Random" was generated as follows: We fetched a list of about 100,000 authors belonging to the "Computer Science" domain as tagged by MAS. MAS assigns at least three sub-domains to every author. These three sub-domains may not all belong to the same domain (e.g., Computer Science). For example, an author may have two sub-domains from Computer Science and one from Medicine. We kept only the authors who have their first three sub-domains belonging to the domain of Computer Science. From this set, we randomly selected 500 authors with at least 10 publications and at least 1 citation.

The dataset "Productive" was generated as follows: from the set of 100,000 Computer Science authors we selected the top-500 most productive. The least productive author from this sample has 354 publications.

The third dataset named "Top h" was generated by querying the MAS Database for the top-500 authors in the "Computer Science" domain ordered by *h-index* .

Table 5 summarizes the information about our datasets with respect to the number of authors (line: # of authors), number of publications (line: # of publications), number of citations (line: # of Citations) and average/min/max numbers of citations and publications per author.

Figure 5 shows the distributions for the values of *h-index* , m, C and p. The m index was defined in Hirsch's original article (Hirsch 2005) and is explained (quoting Hirsch's text) in

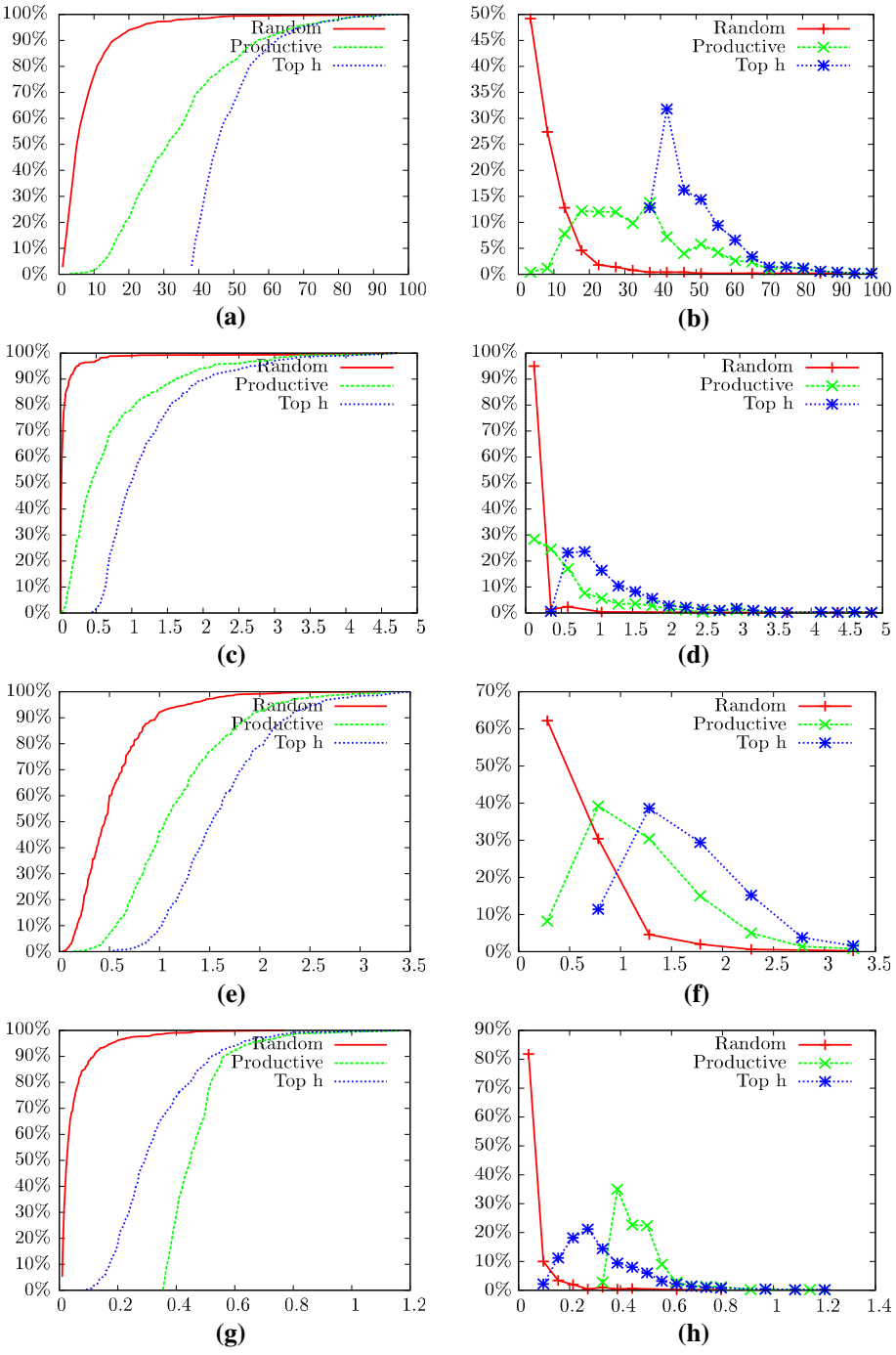

Fig. 5 Distributions of *h-index* , *m*, *p*, *C* indices (*Left plots* CDFs. *Right plots* PDFs) **a** *h*-index, **b** *h*-index, **c** $C\,(*10,000)$, **d** $C\,(*10,000)$, **e** *m*, **f** *m*, **g** $p\,(*1,000)$ **h** $p\,(*1,000)$. (Color figure online)

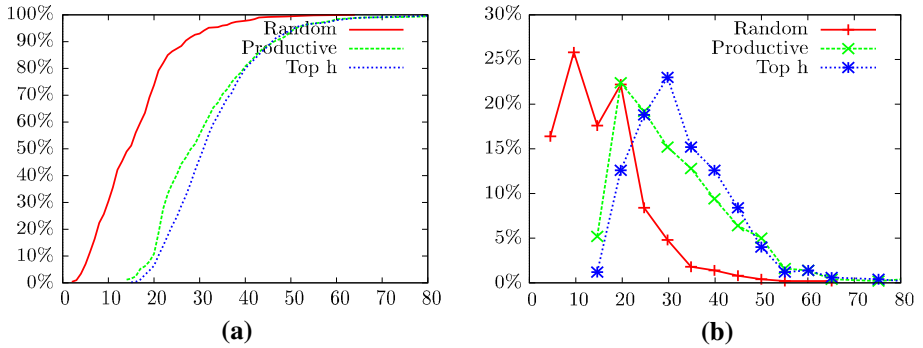

Fig. 6 Distributions of career length (*Left plot* CDFs. *Right plots* PDF) **a** Career Length **b**Career Length. (Color figure online)

the next paragraph. Plots are illustrated in pairs. The ones on the left show cumulative distributions. For example, in Fig. 5a we see that 80 % of the authors in the sample "Random" (red solid line) have *h-index* less than 10. It is obvious that the sample "Top h" (blue dotted line) has higher values for the *h-index* . Figures 5c, d show the distributions for the total number of citations. As expected the sample "Top h" has the highest values.

Figures 5e, f show the distributions for the m value. Recall its definition from Hirsch (2005): A value of $m \approx 1$ (i.e., an *h-index* of 20 after 20 years of scientific activity), characterizes a successful scientist. A value of $m \approx 2$ (i.e., an *h-index* of 40 after 20 years of scientific activity), characterizes outstanding scientists, likely to be found only at the top universities or major research laboratories. A value of $m \approx 3$ or higher (i.e., an *h-index* of 60 after 20 years, or 90 after 30 years), characterizes truly unique individuals. Indeed, only a few authors have $m > 3$.

Figure 5 g, h illustrate the distributions for the total number of publications. It is obvious that in the "Random" sample (red solid line) there are relatively low values for the total number of publications. Also, as expected, the distribution for the "Productive" sample has the highest values for the total number of publications.

For the completeness of the dataset description we present Fig. 6. In this plots is shown the distribution of the Career length of researchers in our samples. This information was necessary in order to compute the $m - index$ presented in Fig. 5. The figures 6a, b depicting carreer length, show that since the distributions of carreers of top and productive are similar, this means that the PI index mainly characterizes the scientist's behaviour towards publising and not the length of its scientific life

We have conducted further experiments to study the behavior of other indices such as α (Hirsch 2005) and e^2 (Zhang 2009). However, the results did not carry any significant information, and therefore, the figures for these factors are not presented.

Does PI offer new insights about the impact and publication habits of scientists?

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 Fig. 7a the x-axis denotes the rank position (normalized percentagewise) of an author by *h-index* , whereas the y-axis denotes the rank position by the total number of citations (C). Each point denotes the position of an author ranked by the two metrics. Note that all

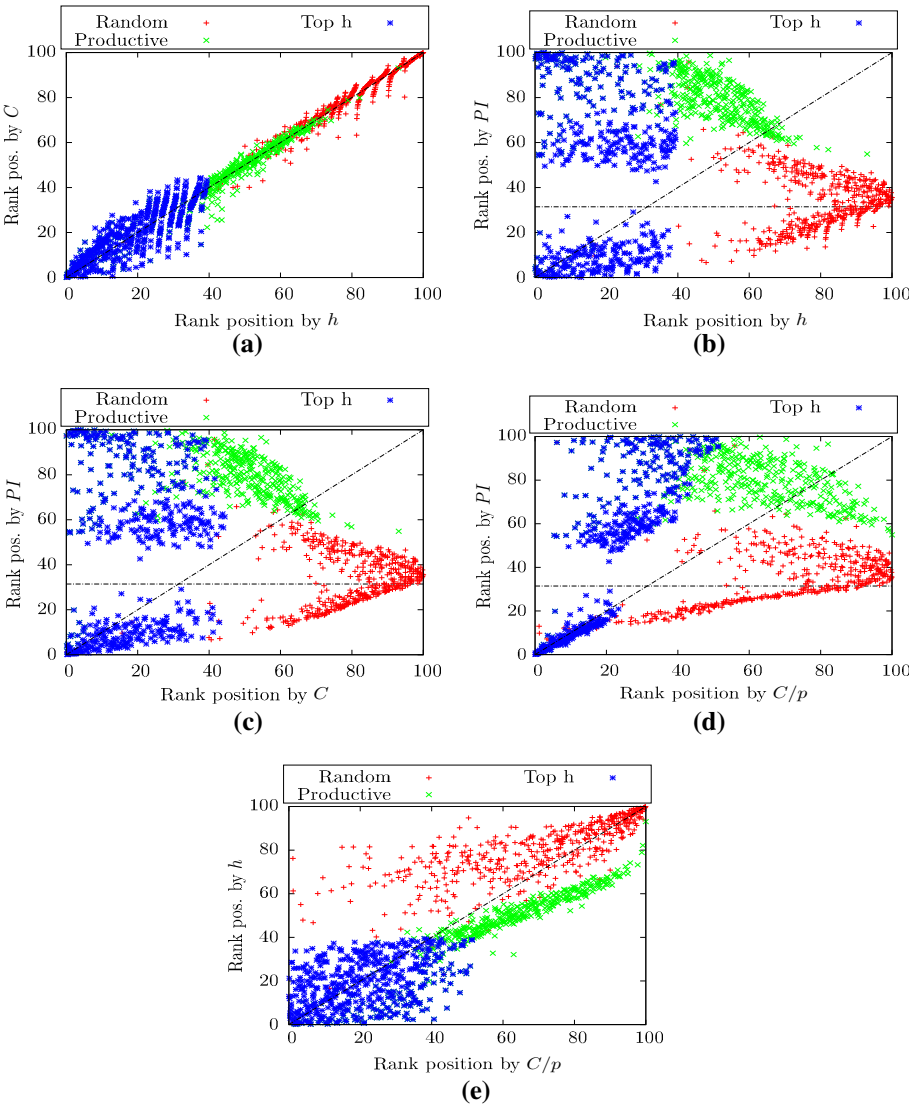

Fig. 7 Correlation of PI to standard bibliometric indices. (Q-Q plots: X- and Y-axis denote normalized rank positions (%).) **a** h versus C (unioned), **b** h versus PI (unioned), **c** C versus PI (unioned), **d** C/p versus PI (unioned), **e** C/p versus h (unioned). (Color figure online)

three samples are merged but if the point is blue asterisk, then the author belongs to the "Top h" sample, if the point is green diagonal cross then he belongs to the "Productive" sample etc. If an author belongs to more than one sample, then only one color is visible since the bullet overwrites the previous one. From Fig. 7a the outcomes are:

- "Top h" authors are ranked in the first 40 % of the rank table by *h-index* , as well as in the top 40 % by the total number of citations (C).

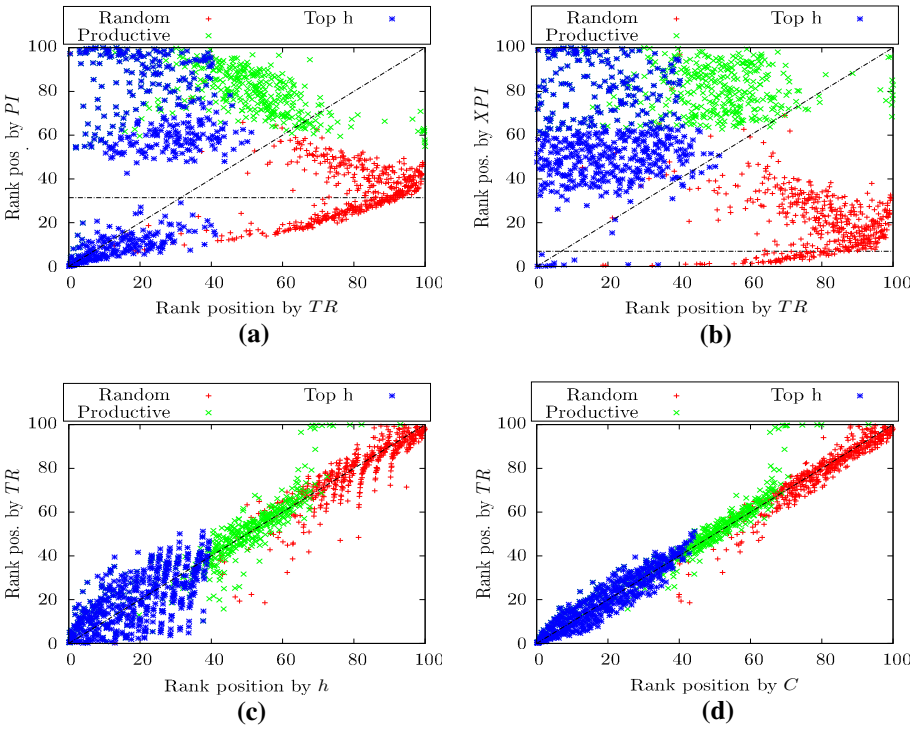

Fig. 8 Correlation of Academic Trace to PI, XPI, C, hindex. (Q-Q plots: X- and Y-axis denote normalized rank positions (%).) **a** T versus PI, **b** T versus XPI **c** h versus T **d** C versus T. (Color figure online)

- "Productive" authors are mainly ranked by *h-index* between 30 and 70 %. The rank positions by C are between 20 and 70 %.
- "Random" authors are mainly ranked below 60 % for both metrics with some outliers in the range 0–60 %, mostly by C.

The aforementioned conclusions are as expected; it also occurs that *h-index* ranking does not differ significantly from the C ranking; i.e., they are correlated which is consistent[6] with earlier findings (Spruit 2012).

In Fig. 7b the *h-index* ranking is compared to PI ranking. It can be seen that there is no correlation between PI and *h-index* . Note that the horizontal line at about 32 % (also shown later in Table 6) shows the cut point of PI for the zero value. Authors that reside below this line have PI > 0 and authors above this line have PI < 0. This observation strengthens the motivation of the article; only one out of three authors is a perfectionist one. Even among those with high h-index (Table 6) only half of them are truly laconic.

- "Top h" authors are split to two groups. The first group is ranked in the top 20 % of the PI rank table. The second group is ranked in the last 50 %. These two groups are also separated by the zero line of PI.

[6] http://michaelnielsen.org/blog/why-the-h-index-is-virtually-no-use/.

Table 6 PI and XPI statistics

Sample	PI		$PI_{\kappa=2}$		$PI_{\kappa=4}$		XPI		$XPI_{\kappa=2}$		$XPI_{\kappa=4}$	
	<0	≥ 0	<0	≥ 0	<0	≥ 0	<0	≥ 0	<0	≥ 0	<0	≥ 0
Random	284	216	213	287	122	378	418	82	408	92	383	117
	57 %	43 %	43 %	57 %	24 %	76 %	84 %	16 %	82 %	18 %	77 %	23 %
Productive	485	15	474	26	439	61	500	0	500	0	500	0
	97 %	3 %	95 %	5 %	88 %	12 %	100 %	0 %	100 %	0 %	100 %	0 %
Top h	292	208	226	274	114	386	488	12	484	16	473	27
	58 %	42 %	45 %	55 %	23 %	77 %	98 %	2 %	97 %	3 %	95 %	5 %
Unioned	904	419	767	556	563	760	1230	93	1216	107	1180	143
	68 %	32 %	58 %	42 %	43 %	57 %	93 %	7 %	92 %	8 %	89 %	11 %

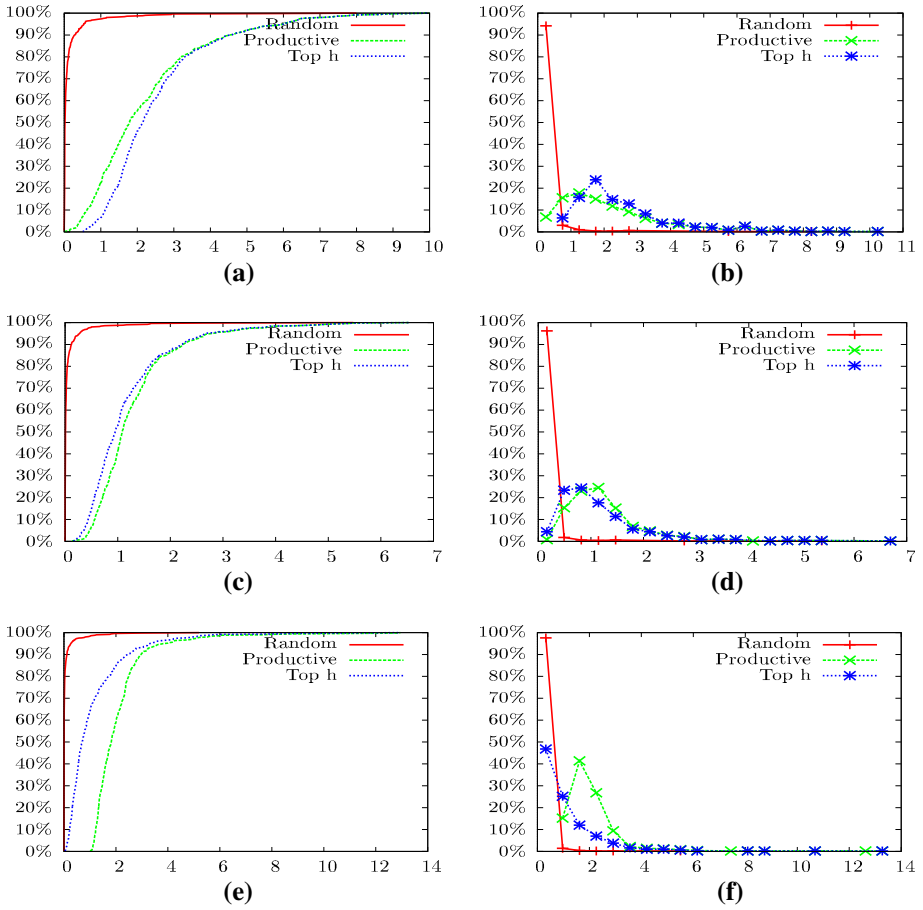

Fig. 9 Distributions of C_T, C_{TC} (tail complement), C_{IC} (ideal complement). (Left plots: CDFs. Right plots: PDFs) **a** C_T (*1,000) **b** C_T (*1,000) **c** C_{TC} (*10,000) **d** C_{TC} (*10,000), **e** C_{IC} (*100,000), **f** C_{IC} (* 100,000). (Color figure online)

- "Productive" authors are almost all ranked at lower positions by PI than by *h-index* . Almost all points reside above the PI zero line and also above the line $y = x$ (with some exceptions at about 65–70 % of the rank list).
- "Random" authors are also generally higher ranked by PI than by *h-index* . They are also split into two groups by the line PI $= 0$.

From the above, it seems that PI is not correlated to *h-index* , whereas the line PI $= 0$ plays the role of a symmetric axis. Thus, it emerges as the key value that separates the "influential" authors from the "mass producers".

In Fig. 7c we compare PI ranking against C (total number of citations) ranking. It is expected that the plot would be similar to Fig. 7b based on the similarity of *h-index* with C.

In Figs. 7d, e we compare *h-index* and PI with the average number of citations per publication (C/p) ranking. It is apparent that PI is not correlated to C/p. *h-index* is also uncorrelated to C/p, however the points of the qq-plot in Fig. 7e are closer to the line $x = y$ than the points of Fig. 7d.

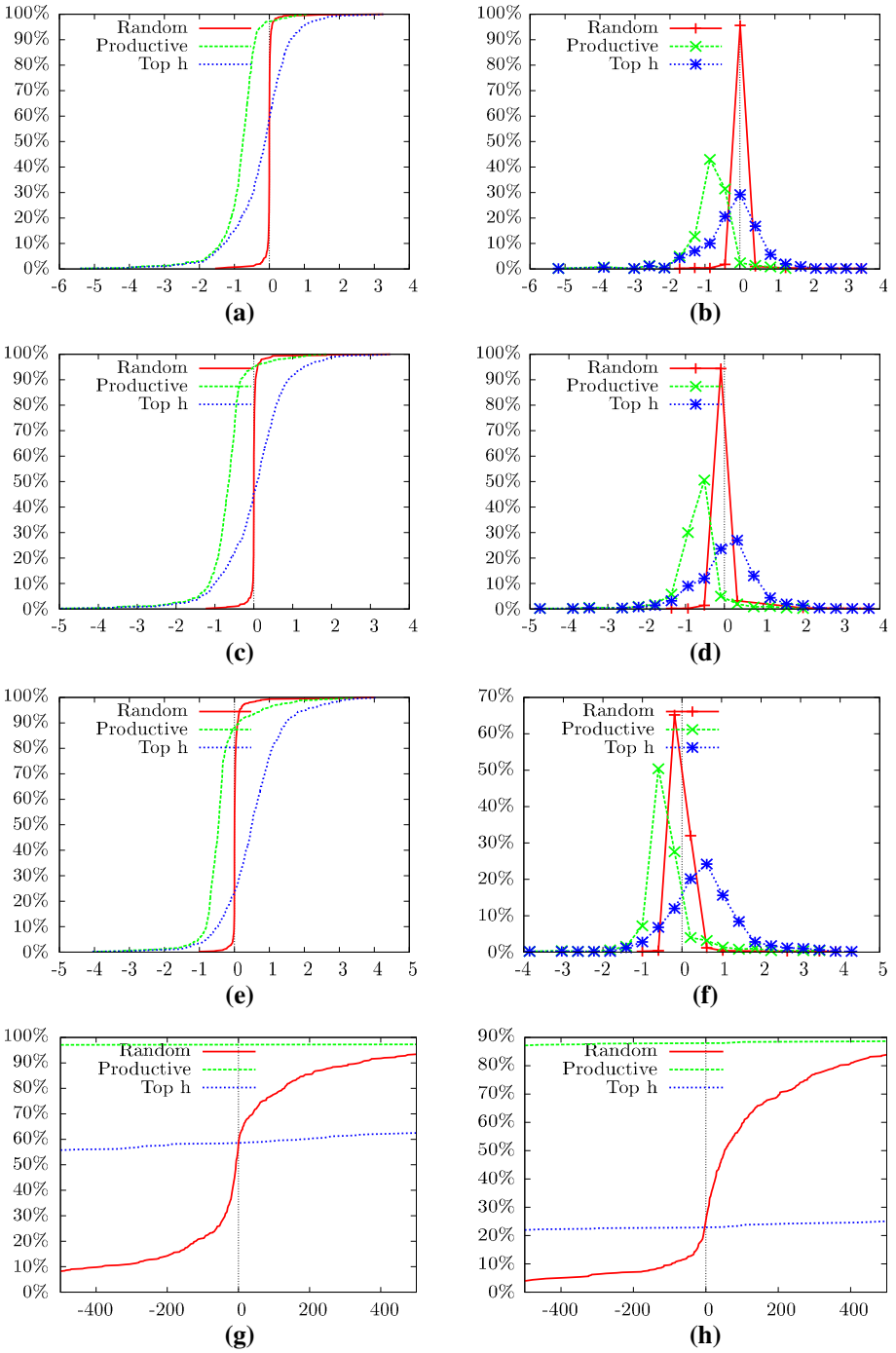

Fig. 10 Distributions of PI, $\mathrm{PI}_{(\kappa=2)}$ and $\mathrm{PI}_{(\kappa=4)}$ **a** $\mathrm{PI}(*10,000)$ **b** $\mathrm{PI}(*10,000)$ **c** $\mathrm{PI}_{(\kappa=2)}(*10,000)$ **d** $\mathrm{PI}_{(\kappa=2)}(*10,000)$ **e** $\mathrm{PI}_{\kappa=4}(*10,000)$ **f** $\mathrm{PI}_{\kappa=4}(*10,000)$ **g** $\mathrm{PI}(\textit{limited}\,\mathrm{x}\textit{range})$ **h** $\mathrm{PI}_{\kappa=4}(\textit{limited}\,\mathrm{x}\,\textit{range})$. (Color figure online)

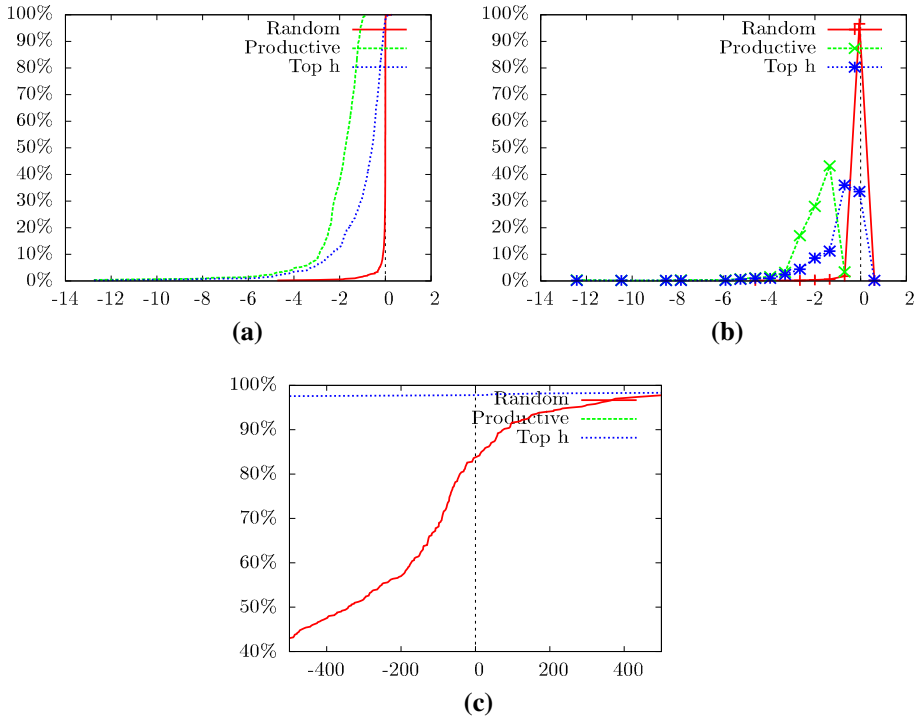

Fig. 11 Distributions of XPI, **a** XPI($*100,000$), **b** XPI($*100,000$), **c** XPI(*limited* x *range*). (Color figure online)

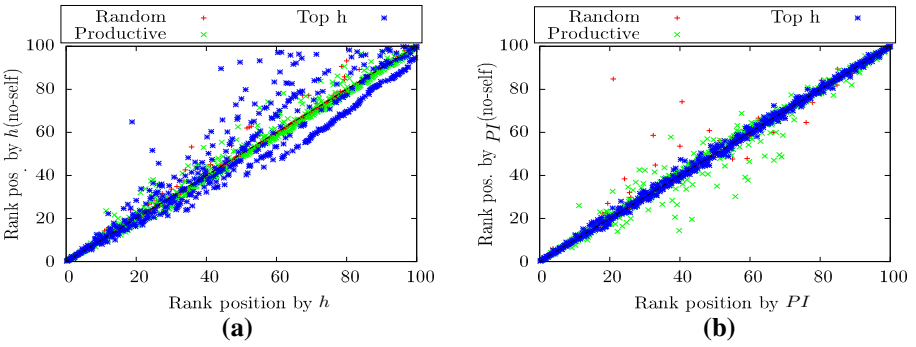

Fig. 12 Robustness of *h-index* and PI to self-citations (Q-Q plots: X and Y axis denote the rank position normalized in percent) **a** h versus h(no-self) **b**PI versus PI(no-self). (Color figure online)

Conclusively, the PI ranking is not correlated to *h-index* , C and C/p.

We have also implemented several comparisons of PI and XPI with various type of ranking indices. We do not include all of them for brevity. All them show that our new metric is not tied correlated with none of them. One interesting metric is the "Academic Trace" (Ye and Leydesdorff 2013). This metric looks similar to PI in the way that it has a

Table 7 Ranking by *h-index* (top-20 scientists)

Author	h		PI		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Shenker Scott	97	1	5754	52	508	45,621	89.81	−51
Foster Ian	93	2	−15510	1,287	768	47,265	61.54	−1,285
Garcia-Molina Hector	92	3	−17423	1,299	605	29,773	49.21	−1,296
Estrin Deborah	90	4	5348	62	479	40,358	84.25	−58
Ullman Jeffrey	86	5	11267	18	460	43,431	94.42	−13
Culler David	84	6	7552	38	386	32,920	85.28	−32
Tarjan Robert	83	7	2888	117	405	29,614	73.12	−110
Towsley Don	82	8	−31929	1,318	793	26,373	33.26	−1,310
Kanade T.	81	9	−20753	1,309	742	32,788	44.19	−1,300
Haussler David	81	10	10952	19	335	31,526	94.11	−9
Jain Anil	81	11	−11474	1,236	590	29,755	50.43	−1,225
Papadimitriou Christos	80	12	−5897	968	506	28,183	55.70	−956
Katz Randy	78	13	−27820	1,317	757	25,142	33.21	−1,304
Pentland Alex	77	14	−1242	724	509	32,022	62.91	−710
Han Jiawei	77	15	−15410	1,285	653	28,942	44.32	−1,270
Jordan Michael	75	16	−1062	717	499	30,738	61.60	−701
Karp Richard	75	17	7231	41	377	29,881	79.26	−24
Zisserman A.	75	18	210	263	421	26,160	62.14	−245
Jennings Nick	74	19	−15718	1,289	626	25,130	40.14	−1,270
Thrun S.	74	20	−5789	958	445	21,665	48.69	−938

negative factor for the zero cited publications. For this reason we prove that neither PI nor XPI is correlated to "Academic Trace" metric.

Ye and Leydesdorff (2013) defined the metric named "Academic Trace" as:

$$T = tr(V) = \frac{h^2}{p} + \frac{C_T^2}{C} + \frac{C_E^2}{C} - \frac{p_z^2}{p}$$

where p_z is the number of the zero cited publications. Figure 8 shows the comparison of Academic Trace (T) with PI, XPI, C and *hindex*. As it is shown in Fig. 8a, b, PI and XPI are not correlated to T. That is because the negative factor of T (the p_z^2/p) is usually too small so it does not affect significantly the resulting score. Academic Trace can only be used with the notion of a time window (like Impact Factor). When used with a time window the average number of zero cited papers increases and this negative factor affects the ranking. So, Academic Trace can mainly be used with a temporal notion. In addition, Fig. 8c, d show that it is correlated with *hindex* and especially with the total number of citations (C), except for some outliers of the "Productive" sample which are ranked last with T but they are about at 70th (of 100) position by *hindex* and C. These researchers probably have a very big set of zero cited publications. As a conclusion we can say that Academic Trace is strongly correlated with the total number of citations, and when used

Table 8 Ranking by PI index (top-20 influential scientists)

Author	PI		*h*		*p*	*C*	*C*/*p*	Change
	val	pos	val	pos				*h*-PI
Vapnik Vladimir	32542	1	50	171	126	36,342	288.43	+170
Rivest Ronald	29340	2	62	53	320	45,336	141.68	+51
Zadeh L.	25613	3	59	70	320	41,012	128.16	+67
Kohonen Teuvo	19880	4	51	157	160	25,439	158.99	+153
Floyd Sally	18059	5	66	38	222	28,355	127.73	+33
Kesselman Carl	17054	6	60	64	272	29,774	109.46	+58
Schapire Robert	16169	7	56	90	186	23,449	126.07	+83
Milner Robin	16019	8	54	108	202	24,011	118.87	+100
Shamir A	15926	9	53	125	213	24,406	114.58	+116
Tuecke Steven	14747	10	44	281	96	17,035	177.45	+271
Balakrishnan Hari	14444	11	72	21	272	28,844	106.04	+10
Agrawal Rakesh	14375	12	67	30	353	33,537	95.01	+18
Hinton G.	13415	13	63	45	314	29,228	93.08	+32
Aho Alfred	13048	14	50	173	193	20,198	104.65	+159
Lamport Leslie	12254	15	59	71	273	24,880	91.14	+56
Hopcroft John	12088	16	45	258	198	18,973	95.82	+242
Morris Robert	11685	17	57	81	305	25,821	84.66	+64
Ullman Jeffrey	11267	18	86	5	460	43,431	94.42	−13
Haussler David	10952	19	81	10	335	31,526	94.11	−9
Joachims T.	10767	20	41	377	134	14,580	108.81	+357

with a time window is an improvement to Impact Factor as it penalizes the big sets of zero cited publications (but not the once or twice cited ones[7]).

Aggregate analysis of the datasets

Figure 9 shows the distributions for the areas defined in the previous section. In particular, Fig. 9a, b illustrate the distributions for the C_T (tail) area. It seems that the "Top h" cumulative distribution is very similar to the "Productive" one, however, the "Top h" distribution has slightly higher values.

Figure 9c , d illustrate the distributions for the C_{TC} (tail complement) area. It seems that C_{TC} has the same distribution as C_T for all samples except for the sample "Productive", for which C_{TC} has slightly higher values than C_T does. Note, also, that the "Productive" distribution has lower values for *h-index* than "Top h". This means that the height of the C_{TC} areas is smaller for the "Productive" authors than for "Top H' ones. The previous two remarks lead to the (rather expected) conclusion that the "Productive" authors have long and thin tails.

The C_{IC} distribution is shown in Fig. 9e, f. In these plots, it is clear that the "Productive" authors have clearly higher values than any other sample, since C_{IC} is strongly related with the total number of publications.

[7] Usually when a publication is cited once or twice during its total "life", these citations are self-citations.

Table 9 Ranking by PI (bottom-20 by PI, i.e., top-20 mass producers)

Author	PI		h		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Ikeuchi Katsushi	−18173	1,303	43	327	638	7,412	11.62	−976
Thalmann D.	−18356	1,304	46	249	632	8,600	13.61	−1,055
Reddy Sudhakar	−18369	1,305	43	322	659	8,119	12.32	−983
Gao Wen	−18494	1,306	26	649	907	4,412	4.86	−657
Prade Henri	−18692	1,307	65	42	633	18,228	28.80	−1,265
Liu K.	−19063	1,308	42	355	672	7,397	11.01	−953
Kanade T.	−20753	1,309	81	9	742	32,788	44.19	−1,300
Rosenfeld Azriel	−21023	1,310	59	73	707	17,209	24.34	−1,237
Gupta Anoop	−23959	1,311	64	43	739	19,241	26.04	−1,268
Miller J.	−24112	1,312	40	433	807	6,568	8.14	−879
Shin Kang	−24125	1,313	57	85	731	14,293	19.55	−1,228
Schmidt Douglas	−24153	1,314	56	94	729	13,535	18.57	−1,220
Bertino Elisa	−27058	1,315	49	194	805	9,986	12.40	−1,121
Yu Philip	−27727	1,316	63	48	789	18,011	22.83	−1,268
Katz Randy	−27820	1,317	78	13	757	25,142	33.21	−1,304
Towsley Don	−31929	1,318	82	8	793	26,373	33.26	−1,310
Kuo C.	−36848	1,319	40	425	1148	7,472	6.51	−894
Gerla Mario	−37464	1,320	67	32	945	21,362	22.61	−1,288
Dongarra Jack	−39901	1,321	67	31	982	21,404	21.80	−1,290
Poor H.	−40492	1,322	55	100	1069	15,278	14.29	−1,222
Huang Thomas	−54047	1,323	67	33	1172	19,988	17.05	−1,290

In Fig. 10 we see the distributions for the previously defined PI index. For all plots, the zero y-axis is the center of the figure. As seen in Fig. 10a, b most of the authors are located around zero. Note that in the right plots, a point at $x = 0, y = 95\,\%$ with a previous value of $x = -3{,}000$ means that the 95 % of the authors have values in the range $-1{,}500, \ldots, 1{,}500$. The first two plots show that the "Top h" authors have the highest values for PI (about 10 % of them have values greater than 8,000). Interestingly, it seems that the value 0 is a key value. This practically means that the majority of the authors follows a conservative approach toward publishing; they publish significant articles hoping to attract also a significant number of citations. However, the fact that we can encounter authors far away (to the left and to the right) from zero, strengthens the value of the present research, because it shows two things: (a) there are scientists which publich very aggressively and they end up being "mass producers"; (b) there are a few scientists who only publish when they have produced ground-breaking results that really advance their field. We strongly believe that this is not a random concidence. The bell-shaped curves of "Top-h" and "Productive" authors confirm that there is a publishing pattern; otherwise, the curves would be flat ones. These outliers (scientists at the far left and at the far right) reveal the information that we are seeking, i.e., the mass-producers and the perfectionists. Another generic conclusion drawn from the figures is that it is very likely to find a perfectionist who is also a "Top-h" rather than a "Productive" one. This means that the

Table 10 Rank table by PI of sample "Networks"

Author	PI		h		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Jacobson Van	19982	1	44	44	161	25,130	156.09	+43
Floyd Sally	18059	2	66	9	222	28,355	127.73	+7
Balakrishnan Hari	14444	3	72	7	272	28,844	106.04	+4
Johnson David	12180	4	54	21	263	23,466	89.22	+17
Morris Robert	11685	5	57	16	305	25,821	84.66	+11
Handley M.	10763	6	47	35	201	18,001	89.56	+29
Perkins C.	9609	7	52	25	373	26,301	70.51	+18
Paxson Vern	8871	8	60	12	233	19,251	82.62	+4
Stoica Ion	8558	9	63	11	266	21,347	80.25	+2
Heidemann John	8059	10	47	36	237	16,989	71.68	+26
Culler David	7552	11	84	3	386	32,920	85.28	−8
Shenker Scott	5754	12	97	1	508	45,621	89.81	−11
Govindan Ramesh	5356	13	55	19	287	18,116	63.12	+6
Estrin Deborah	5348	14	90	2	479	40,358	84.25	−12
Crovella Mark	4886	15	46	38	172	10,682	62.10	+23
Perrig Adrian	4304	16	58	14	247	15,266	61.81	−2
Lu Songwu	3430	17	44	45	129	7,170	55.58	+28
Akyildiz Ian	3089	18	53	23	401	21,533	53.70	+5
Kleinrock Leonard	1986	19	51	31	282	13,767	48.82	+12
Knightly Edward	263	20	41	50	172	5,634	32.76	+30
Peterson L.	−652	21	54	22	292	12,200	41.78	+1
Hubaux Jean-Pierre	−653	22	45	43	247	8,437	34.16	+21
Vaidya Nitin	−1242	23	50	32	337	13,108	38.90	+9
Zhang Lixia	−1609	24	55	20	374	15,936	42.61	−4
Low Steven	−1796	25	45	42	291	9,274	31.87	+17
Boudec Jean-Yves	−2338	26	44	46	258	7,078	27.43	+20
Win Moe	−2619	27	46	37	341	10,951	32.11	+10
Rexford Jennifer	−2632	28	49	33	269	8,148	30.29	+5
Zhang Hui	−3344	29	52	27	352	12,256	34.82	−2
Srikant R.	−3827	30	46	39	328	9,145	27.88	+9
Diot Christophe	−4054	31	52	30	290	8,322	28.70	−1
Simon Marvin	−4450	32	42	49	370	9,326	25.21	+17
Ammar Mostafa	−4547	33	43	48	308	6,848	22.23	+15
Kurose Jim	−5114	34	59	13	391	14,474	37.02	−21
Campbell Andrew	−6036	35	46	41	348	7,856	22.57	+6
Chlamtac I.	−6274	36	43	47	357	7,228	20.25	+11
Crowcroft Jon	−6863	37	48	34	404	10,225	25.31	−3
Whitt W.	−7759	38	52	29	394	10,025	25.44	−9
Goldsmith A.	−7819	39	57	17	479	16,235	33.89	−22
Srivastava Mani	−8139	40	57	18	423	12,723	30.08	−22
Paulraj A.	−8421	41	64	10	442	15,771	35.68	−31
Schulzrinne Henning	−11050	42	53	24	555	15,556	28.03	−18

Table 10 continued

Author	PI		h		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Garcia-Luna-Aceves J.	−11169	43	46	40	460	7,875	17.12	−3
Nahrstedt Klara	−12286	44	52	28	492	10,594	21.53	−16
Cioffi J.	−14685	45	52	26	575	12,511	21.76	−19
Mukherjee B.	−15702	46	58	15	535	11,964	22.36	−31
Katz Randy	−27820	47	78	5	757	25,142	33.21	−42
Towsley Don	−31929	48	82	4	793	26,373	33.26	−44
Gerla Mario	−37464	49	67	8	945	21,362	22.61	−41
Giannakis Georgios	−44707	50	77	6	932	21,128	22.67	−44

"Top-h" scientists are more "selective" towards publishing, but this not the rule. Its exceptions also strenghten the value of the PI index.

Figure 10g is a zoomed-in version of Fig. 10a. It is clear that about 96 % of the "Productive" authors have $PI < 0$. This means that in this sample there are a lot of "mass producers" (people with high number of publications but relatively low *h-index*- or at least not in "Top h-indexers"). The other samples cut the zero y-axis at about 50 to 60 %, which means that 40 to 50 % are positive. It is also noticeable that about 70 % (15–85 %) of the "Random sample have values very close to zero within the range −200, …, 200.

In Fig. 10c–f) we present the distributions for $PI_{\kappa=2}$ and $PI_{\kappa=4}$. We remind that factor κ is the core area multiplier. In these plots, it is shown that these distributions behave like the basic PI distribution except that they are slightly shifted to the right. The "Productive" sample is affected less than the others. This outcome is understandable since they are the authors with small *h-index* core areas compared to their tail and excess areas.

Comparing subfigure 10h to g we can better see the differences. The number of authors in the negative side of samples "Random" and "Top h" has decreased from 57 and 58 to 24 and 23 % respectively, meaning that about 33–35 % of the sample members moved from the negative to the positive side. The number of "Productive" authors in the negative side has been decreased from 97 to 88 %, i.e. an additional 11 % of the sample members moved to the positive side.

In addition to the distribution plots, Table 6 presents the number of authors that have the mentioned metrics below or above zero for each sample. As mentioned before, 97 % of the "Productive" authors have $PI < 0$, whereas only 3 % reside in the positive side of the plot. This amount increases as we increase the core factor κ. For $\kappa = 4$ the increment is 9 % (12 % from 3 %). In all other samples the increment is greater, i.e. for "Top h" the increment is 35 %, for "Random" is 33 %.

In Fig. 11 the same kinds of plots are presented for the metric XPI. As expected, the difference is that most of the authors lie in the negative side of the graph. The cut points of y-axis are also presented in Table 6. About 2 % of the "Top *h-index* " authors have $XPI > 0$ but **none** of the "Productive" authors do. The cut point for "Random" authors is at 6 %. Also, at this point we repeat the experiment of varying the κ value. The results do not match with those of the PI case. Incrementing κ does not increase the number of positive authors in the same way as the PI case. The increment is negligible for the "Productive" and "Top h" and very small for the sample "Random". This leads to the conclusion that varying the κ factor does not affect XPI significantly. Probably different

Table 11 Rank table by PI of sample "DataBases"

Author	PI		h		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Agrawal Rakesh	14375	1	67	8	353	33,537	95.01	+7
Ullman Jeffrey	11267	2	86	2	460	43,431	94.42	0
Motwani Rajeev	9349	3	69	6	271	23,287	85.93	+3
Fagin Ronald	4400	4	59	16	215	13,604	63.27	+12
Widom Jennifer	4031	5	71	4	280	18,870	67.39	−1
Florescu Daniela	3058	6	40	43	132	6,738	51.05	+37
Bernstein Philip	2917	7	52	22	279	14,721	52.76	+15
Buneman Peter	2001	8	43	39	158	6,946	43.96	+31
Hellerstein Joseph	1941	9	51	25	272	13,212	48.57	+16
Naughton J.	640	10	48	29	221	8,944	40.47	+19
Dewitt David	308	11	63	10	308	15,743	51.11	−1
Koudas Nick	58	12	35	50	168	4,713	28.05	+38
Sagiv Yehoshua	−196	13	42	40	209	6,818	32.62	+27
Chaudhuri Surajit	−278	14	41	41	239	7,840	32.80	+27
Egenhofer Max	−314	15	47	30	223	7,958	35.69	+15
Livny Miron	−597	16	61	12	310	14,592	47.07	−4
Suciu Dan	−659	17	54	19	285	11,815	41.46	+2
Papadias Dimitris	−809	18	38	47	200	5,347	26.73	+29
Lakshmanan Laks	−914	19	37	48	196	4,969	25.35	+29
Lenzerini M.	−1074	20	50	26	269	9,876	36.71	+6
Abiteboul Serge	−1111	21	59	15	321	14,347	44.69	−6
Ioannidis Yannis	−1647	22	39	46	209	4,983	23.84	+24
Sellis Timos	−2747	23	36	49	264	5,461	20.69	+26
Jagadish H.	−2924	24	52	24	303	10,128	33.43	0
Dayal Umeshwar	−2975	25	44	34	306	8,553	27.95	+9
Maier David	−3096	26	45	32	331	9,774	29.53	+6
Wiederhold Gio	−3320	27	43	38	315	8,376	26.59	+11
Ramakrishnan Raghu	−4249	28	52	23	348	11,143	32.02	−5
Snodgrass Rick	−4293	29	41	42	297	6,203	20.89	+13
Srivastava Divesh	−4333	30	44	35	317	7,679	24.22	+5
Ceri Stefano	−4355	31	45	33	345	9,145	26.51	+2
Kriegel Hans-Peter	−5034	32	46	31	451	13,596	30.15	−1
Stonebraker M.	−5643	33	62	11	380	14,073	37.03	−22
Halevy Alon	−5858	34	71	5	392	16,933	43.20	−29
Abbadi Amr	-6906	35	39	45	361	5,652	15.66	+10
Gray Jim	−7953	36	54	18	508	16,563	32.60	−18
Faloutsos Christos	−8509	37	68	7	484	19,779	40.87	−30
Jensen Christian	−8566	38	44	37	389	6,614	17.00	−1
Agrawal Divyakant	−9199	39	40	44	433	6,521	15.06	+5
Aalst W.	−9811	40	48	28	468	10,349	22.11	−12
Weikum Gerhard	−11700	41	44	36	467	6,912	14.80	−5
Sheth Amit	−12193	42	58	17	488	12,747	26.12	−25

Table 11 continued

Author	PI		*h*		*p*	*C*	*C*/*p*	Change
	val	pos	val	pos				*h* − *PI*
Carey Michael	−14606	43	60	14	488	11,074	22.69	−29
Franklin Michael	−14765	44	60	13	559	15,175	27.15	−31
Han Jiawei	−15410	45	77	3	653	28,942	44.32	−42
Jajodia Sushil	−15483	46	53	21	554	11,070	19.98	−25
Mylopoulos John	−15513	47	53	20	569	11,835	20.80	−27
Garcia−Molina Hector	−17423	48	92	1	605	29,773	49.21	−47
Bertino Elisa	−27058	49	49	27	805	9,986	12.40	−22
Yu Philip	−27727	50	63	9	789	18,011	22.83	−41

default values for the factors of Eq. 4 (especially for κ and/or ν) may be needed for tuning the XPI metric. However, this task remains out of the scope of the present article.

PI robustness to self-citations

Self-citations (a citation from an article to another article when there is at least one common author between the citing and the cited paper) is a common way for authors to increase the visibility of their works. It has been documented that self-citations can have significant impact upon *h-index* (Schreiber 2007). However, self-citations do not necessarily represent a bad practise, that the scientometric indices should punish. In many cases (Katsaros et al. 2009), "they can effectively describe the authoritativeness of an article." Therefore, the aim is to design robust metrics (Katsaros et al. 2009) that will be unaffected by self-citations (Katsaros et al. 2009).

We performed an experiment to study the behavior of *h-index* and PI with respect to self-citations. In Fig. 12(a) a qq-plot is shown, which compares the ranking produced by *h-index* . The x-axis represents the rank produced by the computed *h-index* including self-citations, whereas the y-axis represents the rank of *h-index* after excluding self-citations. We have performed several experiments with different types of ranking and they all show similar behavior with respect to *h-index* . In Fig. 12(b) the same kind of qq-plot for the PI as a rank criterion is displayed. It is apparent that PI is much less affected by self-citations than the *h-index* . This is another advantage of the proposed metric; it is not affected by self-citations.

PI in action: Ranking scientists

In the previous two subsections, we performed an analysis of the datasets at a coarse level. In this section, we will provide an analysis at a finer level, that of individual scientists. We have emphasized from the beginning of the article that it is not this article's purpose to explain the roots of the publishing behaviour of individual scientists. However, we will attempt to record those characteristics of the scientists (if there are such characteristics) that make them exhibit particular behaviours.

Table 7 shows the rank table for the top-20 authors by *h-index* from all our samples. They are truly remarkable scientists with significant contributions to their field. The table

Table 12 Rank table by PI of sample "Multimedia"

Author	PI		h		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Donoho David	7508	1	72	2	350	27,524	78.64	+1
Cox Ingemar	3464	2	41	15	210	10,393	49.49	+13
Simoncelli Eero	2619	3	47	12	227	11,079	48.81	+9
Yeo Boon-lock	2131	4	27	44	77	3,481	45.21	+40
Rui Yong	1745	5	33	32	168	6,200	36.90	+27
Jain Ramesh	1637	6	36	25	243	9,089	37.40	+19
Yeung Minerva	1490	7	24	48	66	2,498	37.85	+41
Goljan Miroslav	1401	8	28	41	64	2,409	37.64	+33
Wiegand Thomas	602	9	32	33	262	7,962	30.39	+24
Fridrich Jessica	472	10	27	45	118	2,929	24.82	+35
Elad Michael	156	11	36	26	216	6,636	30.72	+15
Naphade Milind	136	12	24	49	106	2,104	19.85	+37
Manjunath B.	−46	13	39	20	279	9,314	33.38	+7
Orchard M.	−784	14	34	30	187	4,418	23.63	+16
Wu Min	−1079	15	27	46	169	2,755	16.30	+31
Li Mingjing	−1180	16	28	42	150	2,236	14.91	+26
Zhang Ya-Qin	−2205	17	36	28	236	4,995	21.17	+11
Hauptmann Alexander	−3183	18	34	31	243	3,923	16.14	+13
Smith John	−3277	19	40	17	282	6,403	22.71	−2
Zakhor Avideh	−3468	20	38	23	268	5,272	19.67	+3
Ebrahimi Touradj	−3520	21	31	37	272	3,951	14.53	+16
Memon Nasir	−3736	22	32	34	286	4,392	15.36	+12
Li Shipeng	−3925	23	26	47	271	2,445	9.02	+24
Hua Xian−sheng	−4252	24	24	50	285	2,012	7.06	+26
Ma Wei−ying	−4292	25	46	13	335	9,002	26.87	−12
Ortega Antonio	−4894	26	31	36	330	4,375	13.26	+10
Xiong Zixiang	−4950	27	35	29	308	4,605	14.95	+2
Bouman C	−5592	28	27	43	380	3,939	10.37	+15
Wu Xiaolin	−6332	29	31	39	337	3,154	9.36	+10
Bovik Alan	−8008	30	39	19	507	10,244	20.21	−11
Ramchandran Kannan	−8111	31	49	11	421	10,117	24.03	−20
Liu Bede	−8784	32	38	22	436	6,340	14.54	−10
Strintzis M.	−8871	33	29	40	454	3,454	7.61	+7
Chang Edward	−9099	34	31	38	448	3,828	8.54	+4
Delp Edward	−10001	35	37	24	438	4,836	11.04	−11
Chen Liang-Gee	−10311	36	32	35	478	3,961	8.29	−1
Tekalp A.	−10552	37	40	18	448	5,768	12.88	−19
Unser Michael	−10801	38	54	6	465	11,393	24.50	−32
Vetterli M.	−11139	39	63	4	547	19,353	35.38	−35
Jain Anil	−11474	40	81	1	590	29,755	50.43	−39
Katsaggelos Aggelos	−11662	41	36	27	504	5,186	10.29	−14
Wang Yao	−11705	42	39	21	484	5,650	11.67	−21

Table 12 continued

Author	PI		h		p	C	C/p	Change
	val	pos	val	pos				$h - PI$
Chang Shih−Fu	−11941	43	52	7	507	11,719	23.11	−36
Nahrstedt Klara	−12286	44	52	9	492	10,594	21.53	−35
Girod Bernd	−13613	45	52	8	529	11,191	21.16	−37
Pitas Ioannis	−13849	46	44	14	515	6,875	13.35	−32
Chellappa Rama	−14092	47	50	10	604	13,608	22.53	−37
Zhang Hongjiang	−15112	48	63	5	556	15,947	28.68	−43
Kuo C.	−36848	49	40	16	1148	7,472	6.51	−33
Huang Thomas	−54047	50	67	3	1172	19,988	17.05	−47

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). We will seek an explanation for that by contrasting these results in Table 8.

Table 8 shows the rank list ordered by PI; all authors have high ranking positions by *h-index* as well. If we try to find what is common in all these persons, we could say that (most of) these scientists spend significant time of their careers in industrial environments making groundbreaking contributions, and being recognized as inventors whose ideas have been incorporated into many products that penetrated our lives. Examples include Tuecke, Rivest, Shamir, Agrawal, and Lamport. The personnel in these environments are highly trained, working on "real" problems whose solutions are part of business products. Thus, these groups are not publishing-prone, and (most of the time) whenever they publish their results, these are path-breaking and influential. Others, such as Vapnik, Zadeh, Kohonen, Aho and Schapire are pioneers, inventing brand new knowledge and developing it in a long series of articles. It might also be the case that these scientists work only with experienced researchers, thus being *elitists* (Cormode et al. 2013), because for instance their topics are very advanced. On the contrary, people coming solely from academic environments have the role of a mentor (Cormode et al. 2013) and are charged with the task of training young PhD students whose initial works (usually) do not have high impact. Moreover, sometimes they are involved in projects of exploratory nature, which eventually do not open new avenues. Finally, we should not forget the publish-or-perish pressure upon their students and themselves.

Table 9 shows the top-20 "Mass Producers" from our samples. In this table we also present the average number of citations per paper (C/p column). It can be seen that there is a big range of average values from 4 to 45 citations per publication in the top "Mass Producers". In this table we will recognize—consistent with what we said in the previous paragraph – some excellent academics who have trained many PhD students.

In Tables 10, 11 and 12 we present the "toppers" with respect to the fields of "Networks" "Databases" and "Multimedia", respectively. Starting from the "Networks' table, we see Van Jacobson and Sally Floyd ranked first and second respectively; they are well-known inventors who contributed fundamental algorithms to the design and operation of the Internet. They both had careers in industry: in Cisco, Xerox, AT&T Center for Internet Research at ICSI, and worked extensively on developing standards (RFC) in the areas of TCP/IP congestion control. Similarly, looking at Table 11 for the database field, we will find in the top positions persons such as Rakesh Agrawal, Ronald Fagin, Jeffrey Ullman

and Rajeen Motwani who also have spent their careers in companies such as Google, IBM and Microsoft, or have contributed fundamental algorithms in fields such as compilers, databases and algorithms. We can make similar observations from Table 12 where we find some entrepreneurs such as Ramesh Jain who founded or co-founded multiple startup companies including Imageware, Virage and Praja. The type of career is certainly a factor that helps categorize a scientist as an influential, since we can see that Van Jacobson ("Networks") and Nick Koudas (from "Databases" who spend part of his career in AT&T) are the ones who gained the greatest rise in PI ranking compared to the h-ranking: 43 and 38 positions, respectively.

If we turn our attention to the bottom rows of these tables we will recognize some excellent mentors, but mass producers: Elisa Bertino and Jiawei Han from databases, Georgios Giannakis and Jack Dongarra from the networking community, Thomas S. Huang, Rama Chellappa and Ioannis Pitas from multimedia.

But, is it really the case that only inventors and industry persons are influentials, whereas academia persons are mass producers? In that case, the PI index would be of little usefulness since the separation of influentials and mass producers would be quite straightforward. The answer to this question is definitely negative. From the beginning of our article we emphasized that this is a generic attitude of the scientists towards publishing, rather than an outcome of their type of careers. Thus, we can see in the "Networks" field some academia persons such as Hari Balakrishnan, David Johnsonand Ion Stoica, or Peter Buneman from "Databases" who are quite high in the PI ranking, even though they did not develop their careers in companies working with highly trained colleagues. On the other hand, P. S. Yu ("Databases") who spend many years in IBM is found at the end of Table 11, remaining in the top-50 of "Databases".

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.

We are already involved in the consideration of temporal issues into PI by integrating the concepts of *contemporary h-index* (Sidiropoulos et al. 2007) into the PI index.

Acknowledgments The authors wish to thank Professor Sofia Kouidou, Vice-rector of the Aristotle University of Thessaloniki, for stating the basic question that led to the present research. The authors would also wish to thank Professor Vana Doufexi for reviewing and editing the final release of this article. The offer of Microsoft to provide gratis their database API is appreciated. Finally, 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

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E., & Herrera, F. (2009). h-index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289.

Anderson, T. R., Hankin, R. K. S., & Killworth, P. D. (2008). Beyond the Durfee square: Enhancing the h-index to score total publication output. *Scientometrics*, *76*, 577–588.

Basaras, P., Katsaros, D., & Tassiulas, L. (2013). Detecting influential spreaders in complex, dynamic networks. *IEEE Computer magazine*, *46*(4), 26–31.

Baum, J. A. C. (2012). The excess-tail ratio: Correcting Journal Impact Factors for Citation Quality. SSRN: http://ssrn.com/abstract=2038102.

Bornmann, L., Mutz, R., & Daniel, H. D. (2010). The h index research output measurement: Two approaches to enhance its accuracy. *Journal of Informetrics*, *4*(3), 407–414.

Chen, D. Z., Huang, M. H., & Ye, F. Y. (2013). A probe into dynamic measures for h-core and h-tail. *Journal of Informetrics*, *7*(1), 129–137.

Cole, S., & Cole, J. R. (1967). Scientific output and recognition—Study in operation of reward system in science. *American Sociological Review*, *32*(3), 377–390.

Cormode, G., Ma, Q., Muthukrishnan, S., & Thompson, B. (2013). Socializing the *h*-index. *Journal of Informetrics*, *7*(3), 718–721.

Dorta-González, P., & Dorta-González, M. I. (2011). Central indexes to the citation distribution: A complement to the h-index. *Scientometrics*, *88*(3), 729–745.

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

Feist, G. J. (1997). Quantity, quality, and depth of research as influences on scientific eminence: Is quantity most important? *Creativity Research Journal*, *10*, 325–335.

Franceschini, F., & Maisano, D. (2010). The citation triad: An overview of a scientist's publication output based on Ferrers diagrams. *Journal of Informetrics*, *4*(4), 503–511.

García-Pérez, M. A. (2012). An extension of the *h*-index that covers the tail and the top of the citation curve and allows ranking researchers with similar *h*. *Journal of Informetrics*, *6*(4), 689–699.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, *102*(46), 16,569–16,572.

Hirsch, J. E. (2007). Does the h index have predictive power? *Proceedings of the National Academy of Sciences*, *104*(49), 9,193–19,198.

Hirsch, J. E. (2010). An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship. *Scientometrics*, *85*(3), 741–754.

Jin, B., Liang, L., Rousseau, R., & Egghe, L. (2007). The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, *52*(6), 855–863. doi:10.1007/s11434-007-0145-9.

Katsaros, D., Akritidis, L., & Bozanis, P. (2009). The *f* index: Quantifying the impact of coterminal citations on scientists' ranking. *Journal of the American Society for Information Science and Technology*, *60*(5), 1051–1056.

Kuan, C. H., Huang, H. H., & Chen, D. Z. (2011). Positioning research and innovation performance using shape centroids of h-core and h-tail. *Journal of Informetrics*, *5*(4), 515–528.
Liu, J. Q., Rousseau, R., Wang, M. S., & Ye, F. Y. (2013). Ratios of h-cores, h-tails and uncited sources in sets of scientific papers and technical patents. *Journal of Informetrics*, *7*(1), 190–197.
Rosenberg, M. S. (2011). *A biologist's guide to impact factors*. Arizona: Tech. rep., Arizona State University.
Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus*, *1*(4), 23–25.
Schreiber, M. (2007). Self-citation corrections for the Hirsch index. *Europhysics Letters*, *78*(3), 30002.
Sidiropoulos, A., Katsaros, D., & Manolopoulos, D. (2007). Generalized Hirsch *h*-index for disclosing latent facts in citation networks. *Scientometrics*, *72*(2), 253–280.
Spruit, H. C. (2012). The relative significance of the *h*-index. Tech. rep., http://arxiv.org/abs/1201.5476v1.
Vinkler, P. (2009). The π-index: A new indicator for assessing scientific impact. *Journal of Information Science*, *35*(5), 602–612.
Vinkler, P. (2011). Application of the distribution of citations among publications in scientometric evaluations. *Journal of the American Society for Information Science and Technology*, *62*(10), 1963–1978.
Woeginger, G. J. (2008). An axiomatic characterization of the Hirsch-index. *Mathematical Social Sciences*, *56*, 224–232.
Ye, F. Y., Leydesdorff, L. (2013). The "Academic Trace" of the Performance Matrix: A Mathematical Synthesis of the h-Index and the Integrated Impact Indicator (I3) http://arxiv.org/abs/1307.3616.
Ye, F. Y., & Rousseau, R. (2010). Probing the *h*-core: An investigation of the tail-core ratio for rank distributions. *Scientometrics*, *84*(2), 431–439.
Zhang, C. T. (2009). The *e*-index, complementing the *h*-index for excess citations. *PLoS One*, *4*(5), e5429.
Zhang, C. T. (2013a). The h′-index: Effectively improving the h-index based on the citation distribution. *PLOS One*, *8*(4), e59,912.
Zhang, C. T. (2013b). A novel triangle mapping technique to study the h-index based citation distribution. *Nature Scientific Reports*, *3*(1023).