Metrics and Rankings: Myths and Fallacies

Yannis Manolopoulos[1(✉)] and Dimitrios Katsaros[2]

[1] Department of Informatics, Aristotle University, 54124 Thessaloniki, Greece
manolopo@csd.auth.gr
[2] Department of Electrical and Computer Engineering, University of Thessaly, 38221 Volos, Greece
dkatsar@inf.uth.gr

Abstract. In this paper we provide an introduction to the field of Bibliometrics. In particular, first we briefly describe its beginning and its evolution; we mention the main research fora as well. Further we categorize metrics according to their entity scope: metrics for journals, conferences and authors. Several rankings have appeared based on such metrics. It is argued that these metrics and rankings should be treated with caution, in a light relative way and not in an absolute manner. Primarily, it is the human expertise that can rigorously evaluate the above entities.

Keywords: Scientometric indicators · Academic performance · Impact factor · CiteScore index · *h*-index · DORA · Rankings · Informetrics · Scientometrics

1 Introduction

The term "*Bibliometrics*" has been proposed by Alan Pritchard in 1969 [50]. According to Wikipedia, "Bibliometrics is statistical analysis of written publications, such as books or articles" [7]. A relevant field is "*Scientometrics*", a term coined by Vasily Nalimov in 1969 [45]. According to Wikipedia, "Scientometrics is the study of measuring and analysing science, technology and innovation" [53]. Finally, "*Citation analysis*" is a fundamental tool for Bibliometrics and deals with the "examination of the frequency, patterns, and graphs of citations in documents" [14].

A milestone in the development of the field of Bibliometrics was the introduction of the "*Journal Impact Factor*" (IF) by Eugene Garfield in 1955 [25]. Garfield founded the Institute for Scientific Information (ISI) in 1964, which published the Science Citation Index and the Social Science Citation Index. ISI was acquired by Thomson Reuters in 1992.

For about four decades IF was the standard tool for academic evaluations. Despite the fact that it was proposed as a metric to evaluate journals' impact, it was used as a criterion to evaluate the quality of scholarly work by academicians and researchers as well. It was only in 2005 that Jorge Hirsch, a physicist, proposed the *h-index* as a simple and single number to evaluate the production and the impact of a researcher's work [32]. During the last 10–15 years the field has flourished and significant research has appeared in competitive journals and conferences.

L. Kalinichenko et al. (Eds.): DAMDID/RCDL 2016, CCIS 706, pp. 1–16, 2017.
DOI: 10.1007/978-3-319-57135-5_19

Based on these metrics, several rankings have appeared in the web, e.g., for journals, conferences and authors. On the other hand, university rankings appear in popular newspapers; actually, they are the beloved topic of journalists and politicians. University rankings are commercial artifacts of little scientific merit as they are based on arbitrary metrics (like the ones previously mentioned) and on other unjustified subjective criteria.

The purpose of this position paper is to explain that these metrics and rankings should be used with great skepticism. To a great extent, they shed light only to some particular facets of the entity in question (be it a journal, an author etc.); moreover, they are often contradictory to each other. The suggestion is to use this information with caution and pass it through an expert's filtration to come up with a scientific and objective evaluation.

The rest of the paper has the following structure. In the next section we provide more information about the Bibliometric/Scientometric community. Then, we introduce and annotate several metric notions for the evaluation of journals, conferences and authors. In Sect. 4 we mention assorted rankings and pinpoint their contradictions, limitations and fallacies. We devote Sect. 5 to discussing university rankings. Section 6 focuses on a recent negative phenomenon, that of selling publications or citations. In the last section we introduce the Leiden manifesto, an article that tries to put academic evaluations in an academic (not commercial, not mechanistic) framework. In addition, we introduce the main parts of the DORA report, and finally we state the morals of our analysis.

This paper is based on an earlier look on these topics [43]. In particular, the main augmentation parts in the current paper are the following ones: metrics by the Nature publishing group (Subsect. 3.3), moving ahead with citation-based metrics (Subsect. 3.6), altmetrics (Subsect. 3.7), science for sale (Sect. 6), and a large part of the final discussion (Sect. 7).

2 The Bibliometrics Community

During the last two decades a new research community was formed focusing on bibliometric and scientometric issues.

The research output of this community appears in specialized journals and conferences. In particular we note the following journals: (i) Journal of the Association for Information Science and Technology published by Wiley, (ii) Scientometrics by Springer, (iii) Journal of Informetrics by Elsevier, (iv) COLLNET Journal of Scientometrics and Information Management by Taylor and Francis, (v) Research Evaluation by Oxford Journals. Other less established outlets include: (i) Journal of Data and Information Science – formerly, Chinese Journal of Library and Information Science, (ii) Frontiers in Research Metrics and Analytics by Frontiers, (iii) Journal of Scientometric Research by Phcog.Net, (iv) Cybermetrics published by CINDOC and CSIC organizations in Spain, and (v) ISSI Newsletter by the ISSI society.

Two major annual conferences are running for more than a decade. For example, the International Conference of the International Society for Scientometrics and Informetrics (ISSI) is organizing its 16th event at Yuhan/China in October 2017, whereas the 12th International Conference on Webometrics, Informetrics, and Scientometrics (WIS) of the COLLNET community has been organized in December 2016 at Nancy/France.

In addition, assorted papers appear in other major outlets related to Artificial Intelligence, Data Mining, Information Retrieval, Software and Systems, Web Information Systems, etc.

Notably, there are several major databases with bibliographic data. Among others we mention: Google Scholar [28] and the tool Publish or Perish [51], which runs on top of Google Scholar, MAS (Microsoft Academic Search) [44], Scopus [55] by Elsevier and Web of Science [68] (previous known as ISI Web of Knowledge) by Thomson Reuters and DBLP (Data Bases and Logic Programming) [19] by the University of Trier.

3 The Spectrum of Metrics

3.1 Impact Factor

As mentioned earlier, IF is the first proposed metric aiming at evaluating the impact of journals. For a particular year and a particular journal, its IF is the average number of citations calculated for all the papers that appeared in this journal during the previous 2 years. More specifically, this is the 2-years IF as opposed to the 5-years IF, which has been proposed relatively recently as a more stable variant.

IF is a very simple and easily understood notion; it created a business, motivated researchers and publishers and was useful for academicians and librarians. However, IF has been criticized for several deficiencies. For instance,

- it is based mostly on journals in English,
- it considers only a fraction of the huge set of peer-reviewed journals,
- it did not take into account (until recently) conference proceedings, which play an important role in scholar communication in computer science for example,
- it fails to compare journals across disciplines,
- it is controlled by a private institution and not by a democratically formed scientific committee, etc.

On top of these remarks, studies suggest that citations are not *clean* and therefore the whole structure is weak [42]. Also, a recent book illustrates a huge number of flaws encountered in IF measurements [62].

IF can be easily manipulated by the journal's editors-in-chief, who are under pressure in the competitive journal market. For example, the editors-in-chief:

- may ask from authors to add extra references of the same journal,
- may invite/accept surveys as these articles attract more citations that regular papers, or
- may prefer to publish articles that seem to be well cited in the future.

Proposals to fight against IF's manipulation, include the *Reliability IF (RIF)* [39], which considers citation and the length of impact.

There is yet another very strong voice against the over-estimation of IF. Philip Campbell, Editor-in-Chief of the prestigious *Nature* journal, discovered that few papers make the difference and increase the IF of a specific journal. For example, the IF value of Nature for the year 2004 was 32.2. When Campbell analyzed Nature papers over the

relevant period (i.e., 2002–2003), he found that 89% of the impact factor was generated by just 25% of the papers [12]. This is yet another argument against the use of IF to evaluate authors. That is, an author should not be proud just because he published an article in a journal with high IF; on the contrary he should be proud if his paper indeed contributed in this high IF value. However, it is well known that the distribution of the number of citations per paper is exponential [29]; therefore, most probably the number of citations of a paper per year will be less than half of the IF value.

In this category another score can be assigned: the *Eigenfactor* developed by Jevin West and Carl Bergstrom of the University of Washington [6]. As mentioned in Wikipedia: "The Eigenfactor score is influenced by the size of the journal, so that the score doubles when the journal doubles in size" [24]. Also, Eigenfactor score has been extended to evaluate the impact at the author's level.

More sophisticated metrics have been proposed, not only for reasons of elegance but also in the course of commercial competition as well.

3.2 Metrics by Elsevier Scopus

It is well-known that IF values range significantly from one field to another. There are differences in citation practices, in the lag time between publication and its future citation and in the particular focus of digital libraries. It has been reported that "the field of Mathematics has a weighted impact factor of IF = 0.56, whereas Molecular and Cell Biology has a weighted impact factor of 4.76 - an eight-fold difference" [4].

Scopus, the bibliometric research branch of Elsevier uses two important new metrics: *SNIP* (Source Normalized Impact per Paper) [59] and *SJR* (Scimago Journal Rank) [54]. Both take into account two important parameters.

SNIP has been proposed by the Leiden University Centre for Science and Technology Studies (CWTS) based on the Scopus database. According to SNIP citations are normalized by field to eliminate variations; IFs are high in certain fields and low in others. SNIP is a much more reliable indicator than the IF for comparing journals among disciplines. It is also less open to manipulation. Therefore, normalization is applied to put things in a relative framework and facilitate the comparison of journals of different fields.

On the other hand, SJR has been proposed by the SCImago research group from Consejo SCImago research group from the Consejo Superior de Investigaciones Científicas (CSIC), University of Granada, Extremadura, Carlos III (Madrid) and Alcalá de Henares. SJR indicates which journals are more likely to have articles cited by prestigious journals, not simply which journals are cited the most, adopting a reasoning similar to that of Pagerank's algorithm [11].

On December 8th 2016, Elsevier launched the *CiteScore index* to assess the quality of academic journals. It is very similar to IF, i.e., to score any journal in any given year, both tot up the citations received to documents that were published in previous years, and divide that by the total number of documents. However, CiteScore incorporates two major differences that results in quite diverse rankings with respect to those produced by the traditional IF; firstly it considers the "items" published during the last three years (instead of the last two years considered by IF), and secondly, CiteScore counts all

documents as potentially citable, including editorials, letters to the editor, corrections and news items [65]. Despite the simplicity of the differences, they may make a huge difference in rankings. For instance, it has been observed that mathematics articles are slow in acquiring citations (it takes around five years to build up their counters). Also, including more items, especially those that are less cited, may have an adverse effect on the average. For instance, The Lancet, gets a 44 in IF, but a 7.7 in CiteScore, thus it is outside the top-200 in CiteScore.

These metrics have been put into practice as the reader can verify by visiting websites of journals published by Elsevier.

3.3 Metrics by the Nature Publishing Group

The Nature Publishing Group, publisher of some of the top-quality journals in various fields, made its entrance in the world of metrics in 2014 by introducing the *Nature index* [47]. Nature Index maintains a collection of author affiliations collected from articles published in a selected group of 68 high-quality scientific journals. This collection is aggregated and maintained by Nature Research. The Nature Index provides a proxy for high-quality research at the institutional, national and regional level. The Nature Index is updated monthly, and a 12-month rolling window of data is openly available.

There are three measures provided by the Nature Index. The *Article Count* (AC) is the simplest of them; we say that a country or institution has AC equal to 1, if the country or institution has one scientist that (co-)authored one article. Thus, with AC the same article can contribute to multiple countries or institutions. To remove this effect, the Nature Index provides the second measure, namely the Fractional *Count* (FC). The total FC per article is 1, and it is split equally among the co-authors; if some authors have multiple affiliations, then the share is split equally among them. Finally, the *Weighted Fractional Count* (WFC) applies a weighting scheme to properly adjust the FC to the overrepresentation of articles from astronomy and astrophysics, because the four journals (out of the 68) in these disciplines account for about 50% of all the articles in the journals of these fields.

Various objections concern the validity of the Nature Index, such the identity and the way the 68 journals were selected, the citation counting per author, and the weighting scheme, and so on. Nevertheless, this index carries the authoritativeness of its promoter.

3.4 Metrics for Conferences

Conferences are not treated in a uniform way from one discipline to another and even from subfield to subfield. For example, there are conferences where only abstracts are submitted, accepted and published in a booklet, whereas there are other conferences where full papers are submitted and reviewed by ~3 referees in the same manner as journals. Apparently, the latter articles can be treated as first class publications, in particular if the acceptance ratio is as high as 1 out 5, or 1 out of 10 as it happens in several prestigious conferences (such as WWW, SIGMOD, SIGKDD, VLDB, IJCAI, etc.).

Thus, an easy metric to evaluate the quality of a conference is the acceptance ratio (i.e. number of accepted vs. number of submitted papers). Several publishing houses

(e.g. Springer) specify that the acceptance ratio should be <33%. It is a common practice to report such numbers in the foreword of conference proceedings.

Several websites collect data about the acceptance ratios of conferences of several fields such as Theoretical CS, Computer Networks, Software Engineering etc. [1–3]. In general, there is a trend towards events with stricter acceptance policies. The humoristic study of [18] seriously deconstructs such approaches.

Apart from the acceptance ratio, an effort to quantify the impact of conferences has been first initiated by Citeseer, a website and service co-created by Lee Giles, Steve Lawrence and Kurt Bollacker at NEC Research Institute [26]. In particular, using its own datasets Citeseer calculates the IF of a rich set journals and conferences in a unique list. Nowadays, Citeseer is partially maintained by Lee Giles at the Pennsylvania State University [15]; practically, it has been surpassed by Google Scholar.

3.5 Metrics for Authors

In 1985, Jorge Hirsch, a physicist at UCSD, invented the notion of the *h-index* [32]. According to Wikipedia: "a scholar with an index of *h* has published *h* papers each of which has been cited in other papers at least *h* times" [34]. Thus, the *h*-index illustrates both the production and the impact of a researcher's work. It is not just a single number but a 2-dimensional number. *h*-index was a breakthrough; it was a brand new notion that broke the monopoly of IF in academic evaluations.

A propos, it came up that the *h*-index was just a re-invention of a similar metric. Arthur Eddington, an English astronomer, physicist, and mathematician of the early 20th century, was an enthusiastic bicycler. In the context of cycling, Eddington's number is the maximum number E such that the cyclist has cycled E miles on E days. Eddington's own E-number was 84 [22].

Although a breakthrough notion, the *h*-index received some criticism when it was put in practice. In particular, the following issues were brought up:

- it does not consider peculiarities of each specific field,
- it does not consider the order of an author in the list of authors,
- it has a reduced discriminative power as it is an integer number,
- it can be manipulated with self-citations, which cannot be revealed in Google Scholar,
- it has a correlation with the number of the author's publications,
- it constantly increases with time and cannot show the progress or stagnation of an author.

Soon after the invention of the h-index, the field of Bibliometrics flourished and a lot of variants were proposed. The following is only a partial list: g-index, a-index, h(2)-index, hg-index, q^2-index, r-index, ar-index, m-quotient, k-index, f-index, m-index, h_w-index, hm-index, h_{rat}-index, v-index, e-index, π-index, RC-index, CC-index, ch-index, n-index, p-index, w-index, and so on and so forth [35]. The present author's team proposed the following 3 variants: contemporary *h*-index, trend *h*-index, normalized *h*-index [57]. After this flood of variants, several such studies were reported aiming at analyzing, comparing and categorizing the multiplicity of indicators [8, 10, 69, 70]. Two interesting studies are [48,

64], where *h*-index is criticized as easily manipulated by authors in an effort to improve their metrics.

In passing, there have been efforts in studying and devising metrics for the whole citation curve of the works by an author, as a metric supplementary to the *h*-index. In this direction, we proposed two new metrics: the perfectionism index [58] and the fractal dimension [27] to penalize long tails and to dissuade authors from writing papers of low value.

The books by Nikolay Vitanov [67] and Roberto Todeschini, Alberto Baccini [61] give extensive insight into these metrics for authors. In addition, in Publish or Perish [51], a website (maintained by Harzing) and book [30], the most common of these variants have been implemented. Finally, Matlab includes several such implementations as well.

3.6 Moving Ahead with Citation-Based Metrics

One of the major obstacles, for any citation-based metric is the radically different citation patterns from discipline to disciple, which causes problems in comparing journals, persons from different disciplines, and calls for an effective normalization method. To alleviate this problem, the *Relative Citation Ratio* (RCR) [37] was proposed recently; it is a field-normalized metric that shows the citation impact of an article relative to the average NIH-funded paper. It is accompanied by a free software for its calculation, namely iCite, The normalization procedure is done by using each article's co-citation network to field- and time-normalize the number of citations it has received; this topically linked group of articles is used to derive an expected citation rate, which serves as the ratio's denominator. The article citation rate (ACR) is used as the numerator. The basic idea behind RCR is clever, however the study reported in [9], establishes that the RCR correlates highly with established field-normalized indicators, but the correlation between RCR and peer assessments is only low to medium.

The skewness in citation distributions is a universal law, and thus makes metrics stemming from averages to misrepresent the citation curve. So, journals, such as those published by the Royal Society and EMBO Press, already publicize citation distribution [13]. Citation distributions are more relevant than impact factors for high-stakes decisions, such as hiring and promotion, but they can be useful for researchers who are trying to decide which among a pile of papers to read. Protocols for the publication of whole citation distributions appeared recently in literature [41], which can reveal the full extent of the skew of distributions and variation in citations received by published papers.

3.7 Altmetrics: The Alternative to Citation-Based Metrics

The term *altmetrics* was proposed in 2010 [49] as an alternative to article level metrics. Usually, altmetrics are thought of as metrics about articles, but they can be gracefully applied to persons, publication fora, presentations, videos, source code repositories, Web pages, etc. A classification of altmetrics was proposed by ImpactStory in September 2012 [38]:

- Viewed: HTML views and downloads.
- Discussed: journal comments, science blogs, Wikipedia, Twitter, Facebook and other social media.
- Saved: Mendeley, CiteULike and other bookmarking services.
- Cited: citations in the scholarly literature, tracked by Web of Science, Scopus, CrossRef and others.
- Recommended: for example used by F1000Prime.

Altmetrics are in principle more difficult to standardize compared to standard impact measures such as citations. One example is the number of tweets linking to a paper where the number can vary widely depending on how the tweets are collected. Like other metrics, altmetrics are prone to manipulation, by self-citation, gaming, and other mechanisms to boost one's apparent impact. For instance, altmetrics can be gamed in the following ways: likes and mentions can be bought.

4 The Spectrum of Rankings

Ranking is a popular game in academic environments. One can easily find rankings about authors, journals, conferences, and universities as well. Here, we comment on some interesting rankings drawn from several websites. In particular, we will comment on university rankings in the next section.

DBLP website [19] is maintained by Michael Ley at the University of Trier. As of May 2016, its dataset contains more than 1.7 million authors and 3.5 million articles. Based on this dataset, DBLP posts a list of prolific authors in terms of publications of all sorts, e.g. journal and conference papers, books, chapters in books etc. It is interesting to note that Vincent Poor, a researcher at Princeton University, is the most productive person in this ranking with an outcome of 1348 publication (as of 21/10/2016) [20]. Another ranking with the same dataset ranks authors according to the average production per year. Vincent Poor can be found in the 19th position in this ranking (as of 21/10/2016) [21].

Jens Palsberg of UCLA maintains a website where, by using the DBLP datasets, a list of authors ordered according to decreasing *h*-index is produced. First name in this list is Herbert Simon, a Professor at CMU, Nobel Laureate, Turing Award recipient, ACM Fellow; his *h*-index is 164. In this list, Vincent Poor appears with *h*-index equal to 70 [36].

MAS provides a variety of rankings using a dataset of 80 million of articles [44]. For example, it provides two ranked lists of authors according to productivity and according to impact. When the whole dataset is taken into account, then in terms of the number of publications Scott Shenker is 1st, Ian Foster is 2nd and Hector Garcia-Molina is 3rd. According to the number of citations Ian Foster is 1st, Ronald Rivest is 2nd and Scott Shenker is 3rd. Other ranking can be produced by limiting the time window during the last 5 or 10 years. For instance, Vincent Poor is ranked 2nd in terms of productivity for the period of the last 10 years.

Similarly, MAS provides rankings of conferences according to the number of publications or citations for certain periods (i.e., 5 years, 10 years, or the whole dataset). Steadily, INFOCOM, SIGGRAPH, CVPR, ICRA, ICASSP appear at the top.

In an analogous manner, MAS provides rankings for journals. When considering the whole data set, the top journals are CACM, PAMI and TIT. During the last 5 years, new fields came up and, thus, new journals gained acceptance: see for example Expert Systems with Applications and Applied Soft Computing. It is important to notice that these rankings use raw numbers, i.e. without any normalization. However, they show trends in science with time.

The above paragraphs show that there are several kinds of ranking, each with a different emphasis and as such they should be treated with caution. Another example of misuse of rankings concerns the classification of journals and conferences. CORE is an Australian website/service, where journals and conferences are divided in 5 categories as illustrated in the following table [17]. Numbers show the percentages of journals or conferences at their *corresponding* category. Similar categorizations exist in other websites. Even though it is not transparent how the percentages were calculated and the rankings are based on somewhat arbitrary listings and categorizations, such rankings have great acceptance and in several instances state funding may be based on them.

	A*	A	B	C	Other
Journals	7%	17%	27%	46%	3%
Conferences	4%	14%	26%	51%	5%

We give another example where caution is needed. We present two tables. The first table contains data from Aminer [5], which runs on top of DBLP. This table shows the top-10 outlets for "database and data mining" sorted by the *h5*-index, a variation of *h*-index for journals. *h5* is the largest number h such that h articles published in 2011–2015 have at least h citations each. In an analogous manner, the following table gives the top-10 outlets for "Database and Data Mining" sorted by *h5* according to Google Scholar for "Database and Information System".

1	WWW conference	66
2	Information Sciences	62
3	ACM KDD	56
4	IEEE TKDE	53
5	ACM WSDM conference	50
6	JASIST	47
7	ACM SIGIR	42
8	IEEE ICDE conference	40
9	ACM CIKM conference	38
10	IEEE ICDM conference	33

1	WWW conference	74
2	VLDB conference	67
3	IEEE TKDE	66
4	arXiv Social & Infor. Networks (cs. SI)	66
5	ACM SIGMOD conference	65
6	arXiv Databases (cs DB)	61
7	ICWSM (weblog) conference	60
8	ICWSM (web) conference	59
9	ACM WSDM conference	58
10	IEEE ICDE conference	52

We note that the two lists have only 5 items in common, and in different order. At first, one might think that the two lists were not comparable since they were produced by querying different key-words. However, since the first table contains outlets related to Information Systems, whereas the second one contains outlets related to Data Mining the two lists are indeed comparable. This example illustrated that the adoption of a ranking versus another is a subjective matter.

5 University Rankings

Nowadays education is considered as a product/service and, thus, there is a growing financial interest in this global market. Universities try to improve their position in the world arena. Thus, university rankings try to satisfy the need of universities for visibility. These ranking are a popular topic for journalists and, therefore, for politicians as well. However, beforehand we claim that there is little scientific merit in these rankings.

Some rankings are widely-known from mass media. We mention alphabetically the most commercial ones:

- Academic Ranking of World Universities (Shanghai or ARWU),
- QS World University Rankings (QS),
- Times Higher Education World University Rankings (THE).

Other rankings originate from academic research teams, such as:

- Leiden Ranking,
- Wikipedia Ranking of World Universities,
- Professional Ranking of World Universities (École Nationale Supérieure des Mines de Paris),
- SCImago Institutions Ranking,
- University Ranking by Academic Performance (Middle East Technical University),
- Webometrics (Spanish National Research Council),
- Wuhan University.

A full list of such rankings exists at Wikipedia [63].

University rankings are intensively criticized for a number of reasons. For example:

- All rankings are based on a number of subjective criteria.
- In all cases, the choice of each particular criterion and its weight are arbitrary.
- To a great extent, these criteria are correlated.
- Evaluation for some criteria is based on surveys, e.g. "academic reputation" or "employer reputation" by QS, which count for 50% of the total weight. The same holds for the "reputation survey" by THE, which counts for 17.9% or 19.5% or 25.3%, if the examined institution is a medical, an engineering or an arts/humanities school, respectively. Such surveys are totally not-transparent.
- THE devotes a 7.5% of the total weight for the international outlook, subcategorized into "ratio of international to domestic staff", "international co-authorship" and "ratio of international to domestic students". In the same way, QS considers "international student ratio" and "international staff ratio" with a special weight of 5% + 5%. Clearly, such criteria favor Anglo-Saxon universities.
- The number of publications and the number of citations (without normalization) favor big universities; this is probably a reason for a general trend in merging universities in Europe.
- No ranking considers whether a university is an old or a new institution, big or small, a technical university or a liberal arts one, etc. Thus, different entities are compared.
- In general, ranking results are not reproducible, an absolutely necessary condition to accept an evaluation as methodologically reliable.
- QS adopts the *h*-index at a higher level, i.e. not at the author's level but for a group of academicians. This is beyond the fair use of the original idea by Hirsch since it does not consider the size of the examined institution, neither has it performed any normalization.
- The rankings exert influence on researchers to submit papers to "prestigious" journals (e.g. Nature, Science). Since such journals follow particular policies as to what is in fashion researchers may not work on what they truly think is worthwhile but according to external/political criteria acting as sirens [56].
- Finally, and probably the most important point of this criticism is that university rankings are misleading proportionally to the degree that they are based on (a) collections of citations from English-language digital libraries, (b) erroneous collections of citations, (c) IF calculations, which ignore whole statistical distributions of a single number, (d) higher level *h*-index calculations, which are conceptually wrong. In other words, "garbage in, garbage out".

All rankings are not equally unacceptable. Several independent studies agree that ARWU is probably the most reliable in comparison to other commercial rankings [40], whereas QS is the most criticized ranking. On the other hand, between the rankings originated from academic institutions, Leiden is considered as the most reliable as it stems from a strong research team with significant academic reputation and tradition in the field of Bibliometrics/Scientometrics. On the other hand, the ranking of Webometrics is criticized for the adoption of non-academic criteria, such as the number of web pages and files and their visibility and impact according to the number of received inlinks.

Based on the above discussion, one can understand why the question "science or quackery" arises [52]. In a recent note by Moshe Vardi, Editor-in-Chief of CACM and professor with Rice University, same skepticism was reported [66]. Moreover, some

state authorities are critical against these methodologies [46]. However, it is sad that rankings are "*here to stay*" because strong financial interests worldwide support such approaches.

At this point, we mention a very useful website which, based on the DBLP dataset, ranks American CS departments in terms of the number of faculty and the number of publications in selected fora, by picking certain CS subfields [16].

6 Science for Sale

There is yet another problem with the sole existence of any research impact/productivity indicator. Scholars in their struggle to increase their personal ranking according to such indicators, e.g., to publish articles in journals with high IF, to increase citation numbers, resort to the worst possible practice, for instance, to buy authorship! It has been reported in [33] that this highly unethical act is far beyond an isolated event. It is a developing market involving "shady agencies, corrupt scientists, and compromised editors", where the prices vary depending on whether the buyer wishes to be the first/primary author, or merely a coauthor. Similar unethical acts have been documented in the context of citation buying [60].

7 Discussion and Morals

The intention of this position paper is the following. Bibliometrics is a scientific field supported by a strong research community. Although the term is not new, during the last years there is an intense research in the area due to the web and open/linked data.

The outcome of Bibliometrics is most often misused by mass media and journalists, state authorities and politicians, and even in the academic world. Criticism has been expressed for several metrics and rankings, not without a reason.

In 2015, a paper was published in Nature entitled: "The Leiden Manifesto for research metrics". More specifically, the paper states 10 principles to guide research evaluation [31]. We repeat them here in a condensed style:

1. Quantitative evaluation should support qualitative, expert assessment.
2. Measure performance against the research missions of the institution, group or researcher.
3. Protect excellence in locally relevant research.
4. Keep data collection and analytical processes open, transparent and simple.
5. Allow those evaluated to verify data and analysis.
6. Account for variation by field in publication and citation practices.
7. Base assessment of individual researchers on a qualitative judgment of their portfolio.
8. Avoid misplaced concreteness and false precision.
9. Recognize the systemic effects of assessment and indicators.
10. Scrutinize indicators regularly and update them.

Probably, the last principle is the most important. Since it is easy for humans to cleverly adapt to external rules and to try to get the most benefit out of them, the Bibliometrics community has to devise and promote new metrics for adoption by academia and others.

At this point, it is worth mentioning the San Francisco Declaration of Research Assessment (DORA) [22], which was initiated by the American Society for Cell Biology (ASCB) together with editors and publishers of scholarly journals. They altogether recognize the need to improve the ways in which the outputs of research are evaluated. The group developed a set of recommendations, known as the San Francisco Declaration on Research Assessment. In summary, they published recommendations for four "bodies":

- For funding agencies:
  - To clarify the criteria used in evaluating the scientific output of grant applicants and state firmly that an article's content is much more important than metrics or the identity of the journal in which the article was published.
  - To account also for the significance and impact of collected datasets and developed software, apart from the published articles, and consider a wide range of impact metrics, including also qualitative indicators such as influence on decision making and practice.
- For institutions:
  - To be explicit about the criteria used to make decisions about hiring, tenure, and promotion, and state firmly that an article's content is much more important than metrics or the identity of the journal in which the article was published.
  - For the purposes of research assessment, to account for the significance and impact of collected datasets and developed software, apart from the published articles, and consider a wide range of impact metrics, including also qualitative indicators such as influence decision making and practice.
- For publishers:
  - To avoid giving too much emphasis on the IF, or at least to present IF as a member of a set of journal-related metrics (e.g., EigenFactor, SCImago, *h*-index, etc.) that provide a multidimensional perspective on journal performance.
  - To make available several measures concerning article impact so as to push toward assessment based on the content of an article rather than metrics of the journal in which it was published.
  - To enforce at the extent possible responsible authorship practices, and the clear statement about the specific contributions of each author.
  - To remove all reuse limitations on reference lists and make them available under the Creative Commons Public Domain Dedication.
  - To remove the constraints on the number of references in research articles.
- For organizations that supply metrics:
  - To be open by providing data and methods used to calculate all measures.
  - To provide the data under a license that allows for its unrestricted reuse.
  - To be clear that manipulation of metrics is unacceptable, and that any manipulation of them will have severe consequences.
  - To account for the variation in article types.
- For researchers:

- To make assessments involving funding, hiring, promoting based on true scientific contributions rather than publication metrics.
- To cite the articles where the research originated instead of reviews articles.
- To use several article measures to evaluate the impact of articles.
- To challenge research assessment practices that relies only on IFs.

Finally, we close this paper with a proposed personal list of do's and don'ts.

1. Do not evaluate researchers based on the number of publications or the IF of the journals they appeared.
2. Evaluate researchers with their *h*-index and variants (resolution according to competition).
3. To further evaluate researchers, focus on the whole citation curve and its tail in particular (relevant metrics: perfectionism index and fractal dimension).
4. Do not evaluate journals based on their IF.
5. Evaluate journals with the SCIMAGO and EIGENFACTOR scores as they are robust and normalized.
6. Further, ignore journal metrics and choose to work on the topics that inspire you.
7. Metrics are not panaceas; metrics should change periodically.
8. Do not get obsessed with contradictory rankings for authors, journals and conferences.
9. Ignore university rankings; they are non-scientific, non-repeatable, commercial, and unreliable.
10. Follow your heart and research what attracts and stimulates you.

Acknowledgments. Thanks are due to our ex and present students and colleagues. Many of the ideas expressed in this article are the outcome of research performed during the last 15 years. In particular, we would like to thank Eleftherios Angelis, Nick Bassiliades, Antonia Gogoglou, Vassilios Matsoukas, Antonios Sidiropoulos and Theodora Tsikrika.

References

1. Acceptance Ratio of Networking Conferences: https://www.cs.ucsb.edu/~almeroth/conf/stats
2. Acceptance Ration of SW Engineering Conferences: http://taoxie.cs.illinois.edu/seconferences.htm
3. Acceptance Ratio of TCS Conferences: http://www.lamsade.dauphine.fr/~sikora/ratio/confs.php
4. Althouse, B., West, J., Bergstrom, T., Bergstrom, C.: Differences in impact factor across fields and over time. J. Am. Soc. Inf. Sci. Technol. **60**(1), 27–34 (2009)
5. Aminer: https://aminer.org/ranks/conf
6. Bergstrom, C.T., West, J.D., Wiseman, M.A.: The Eigenfactor metrics. J. Neurosci. **28**(45), 11433–11434 (2008)
7. Bibliometrics: https://en.wikipedia.org/wiki/Bibliometrics
8. Bollen, J., van de Sompel, H., Hagberg, A., Chute, R.: A principal component analysis of 39 scientific impact measures. PLoS ONE **4**, e6022 (2009)
9. Bornmann, L., Haunschild, R.: Relative Citation Ratio (RCR): an empirical attempt to study a new field-normalized bibliometric indicator, J. Assoc. Inf. Sci. Technol. (2017, to appear)

10. Bornmann, L., Mutz, R., Hug, S., Daniel, H.D.: A multilevel meta-analysis of studies reporting correlations between the *h*-index and 37 different *h*-index variants. J. Inform. **5**(3), 346–359 (2011)
11. Brin, S., Page, L.: The anatomy of a large-scale hypertextual web search engine. Comput. Netw. ISDN Syst. **30**(1–7), 107–117 (1998)
12. Campbell, P.: Escape from the impact factor. Ethics Sci. Environ. Politics **8**, 5–7 (2008)
13. Callaway, E.: Publishing elite turns against impact factor. Nature **535**, 210–211 (2016)
14. Citation Analysis: https://en.wikipedia.org/wiki/Citation_analysis
15. CiteSeer Digital Library: http://citeseerx.ist.psu.edu/index
16. Computer Science Ranking: http://csrankings.org
17. Computing Research and Evaluation (CORE): http://www.core.edu.au
18. Cormode, G., Czumaj, A., Muthukrishnan S.: How to increase the acceptance ratios of top conferences? http://www.cs.rutgers.edu/~muthu/ccmfun.pdf
19. DBLP: http://dblp.uni-trier.de
20. DBLP: Prolific authors. http://dblp.uni-trier.de/statistics/prolific1
21. DBLP: Prolific authors per year. http://dblp.l3s.de/browse.php?browse=mostProlificAuthorsPerYear
22. DORA: http://www.ascb.org/dora/
23. Eddington Arthur: https://en.wikipedia.org/wiki/Arthur_Eddington
24. Eigenfactor Metric: https://en.wikipedia.org/wiki/Eigenfactor
25. Garfield, E.: Citation indexes for science: a new dimension in documentation through association of ideas. Science **122**, 108–111 (1955)
26. Giles, C.L., Bollacker, K., Lawrence, S.: CiteSeer: an automatic citation indexing system. In: Proceedings 3rd ACM Conference on Digital Libraries, pp. 89–98 (1998)
27. Gogoglou, A., Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Quantifying an individual's scientific output using the fractal dimension of the whole citation curve, In: Proceedings 12th International Conference on Webometrics, Informetrics & Scientometrics (WIS), Nancy (2016)
28. Google Scholar: http://www.scholar.google.com
29. Gupta, H., Campanha, J., Pesce, R.: Power-law distributions for the citation index of scientific publications and scientists. Braz. J. Phys. **35**(4a), 981–986 (2005)
30. Harzing, A.W.: Publish or Perish, Tarma Software Research (2010)
31. Hicks, D., Wouters, P., Waltman, L., de Rijke, S., Rafols, I.: The Leiden Manifesto for research metrics. Nature **520**(7548), 429 (2015)
32. Hirsch, J.E.: An index to quantify an individual's scientific research output. Proc. Natl. Acad. Sci. **102**(46), 16569–16572 (2005)
33. Hvistendahl, M.: China's publication bazaar. Science **342**(6162), 1035–1039 (2013)
34. *h*-index: https://en.wikipedia.org/wiki/H-index
35. *h*-index Variants: http://sci2s.ugr.es/hindex
36. *h*-index for CS Scientists: http://web.cs.ucla.edu/~palsberg/h-number.html
37. Hutchins, B.I., Yuan, X., Anderson, J.M., Santangelo, G.M.: Relative Citation Ratio (RCR): a new metric that uses citation rates to measure influence at the article level. PLoS Biol. **14**(9), e1002541 (2016)
38. ImpactStory Blog. A new framework for altmetrics (2012)
39. Kuo, W., Rupe, J.: R-impact factor: reliability-based citation impact factor. IEEE Trans. Reliab. **56**(3), 366–367 (2007)
40. Lages, J., Patt, A., Shepelyansky, D.: Wikipedia Ranking of World Universities (2016). Arxiv https://arxiv.org/abs/1511.09021

41. Lariviere, V., Kiermer, V., MacCallum, C.J., McNutt, M., Patterson, M., Pulverer, B., Swaminathan, S., Taylor, S., Curry, S.: A simple proposal for the publication of journal citation distributions, Technical report. http://dx.doi.org/10.1101/062109
42. Lee, D., Kang, J., Mitra, P., Giles, L., On, B.W.: Are your citations clean? Commun. ACM **50**(12), 33–38 (2007)
43. Manolopoulos, Y.: On the value and use of metrics and rankings: a position paper. In: Selected Papers of the 18th International Conference on Data Analytics & Management in Data Intensive Domains (DAMDID 2016), vol. 1752, pp. 133–139. CEUR Workshop Proceedings (2016)
44. Microsoft Academic Search: http://academic.research.microsoft.com
45. Nalimov, V., Mul'chenko, Z.M.: Naukometriya, the study of the development of science as an information process in Russian, p. 191. Nauka, Moscow (1969)
46. Norwegian Universities: http://www.universityworldnews.com/article.php?story=20140918170926438
47. Nature Publishing Group: A guide to the nature index. Nature **515**(7526), S94 (2014)
48. Piazza, R.: On house renovation and co-authoring – tricks of the trade to boost your h-index. Europhys. News **46**(1), 19–22 (2015)
49. Priem, J., Taraborelli, D., Groth, P., Neylon, C.: Altmetrics: a manifesto. altmetrics.org
50. Pritchard, A.: Statistical bibliography or bibliometrics? J. Doc. **25**(4), 348–349 (1969)
51. Publish or Perish: http://www.harzing.com/pop.htm
52. Science or Quackery: https://www.aspeninstitute.it/aspenia-online/article/international-university-rankings-science-or-quackery
53. Scientometrics: https://en.wikipedia.org/wiki/Scientometrics
54. SCIMAGO: http://www.scimagojr.com/
55. Scopus: http://www.scopus.com
56. Schekman, R.: https://www.theguardian.com/science/2013/dec/09/nobel-winner-boycott-science-journals
57. Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Generalized Hirsch *h*-index for disclosing latent facts in citation networks. Scientometrics **72**(2), 253–280 (2007)
58. Sidiropoulos, A., Katsaros, D., Manolopoulos, Y.: Ranking and identifying influential scientists vs. mass producers by the perfectionism index. Scientometrics **103**(1), 1–31 (2015)
59. SNIP: http://www.journalindicators.com
60. The Daily Californian: http://www.dailycal.org/2014/12/05/citations-sale/
61. Todeschini, R., Baccini, A.: Handbook of Bibliometric Indicators. Wiley (2016)
62. Tüür-Fröhlich, T.: The Non-trivial Effects of Trivial Errors in Scientific Communication and Evaluation. Verlag Werner Hülsbusch, Glückstadt (2016)
63. University Rankings: https://en.wikipedia.org/wiki/College_and_university_rankings
64. Van Bevern, R., Komusiewicz, C., Niedermeier, R., Sorge, M., Walsh, T.: H-index manipulation by merging articles: models, theory and experiments. Artif. Intell. **240**, 19–35 (2016)
65. van Noorden, R.: Impact factor gets heavyweight rival: citeScore uses larger database and gets different results. Nature **540**, 325–326 (2016)
66. Vardi, M.: Academic rankings considered harmful! Commun. ACM **59**(9), 5 (2016)
67. Vitanov, N.: Science Dynamics and Research Production. Springer, Cham (2016)
68. Web of Science: http://ipscience.thomsonreuters.com
69. Wildgaard, L., Schneider, J.W., Larsen, B.: A review of the characteristics of 108 author-level bibliometric indicators. Scientometrics **101**, 125–158 (2014)
70. Yan, Z., Wu, Q., Li, X.: Do Hirsch-type indices behave the same in assessing single publications? An empirical study of 29 bibliometric indicators. Scientometrics **109**(3), 1815–1833 (2016)