The Information Society, 30: 349–354, 2014
Published with license by Taylor & Francis
ISSN: 0197-2243 print / 1087-6537 online
DOI: 10.1080/01972243.2014.944731

Routledge
Taylor & Francis Group

PERSPECTIVE

Bibliometric Evaluation of Researchers in the Internet Age

Alberto Bartoli and Eric Medvet

Department of Engineering and Architecture, University of Trieste, Trieste, Italy

Downloaded by [University of Trieste] at 01:22 03 October 2014

Research evaluation, which is an increasingly pressing issue, invariably relies on citation counts. In this contribution we highlight two concerns that the research community needs to pay attention to. One, in the world of search engine facilitated research, factors such as ease of Web discovery, ease of access, and content relevance, rather than quality, influence what gets read and cited. Two, research evaluation based on citation counts works against many types of high-quality works. We also elaborate on the implications of these points by examining a recent nationwide evaluation of researchers performed in Italy. We focus on our discipline (computer science), but we believe that our observations have relevance for a broad audience.

Keywords academic search engines, bibliometrics, citation counts, Italy, research evaluation, researcher evaluation

There is an increasing pressure worldwide for research evaluation at different levels, from individual researchers to entire departments/research institutions. Bibliometric measures derived from citation counts are a key tool here. Their strengths and weaknesses have been widely analyzed in the literature (Bornmann and Daniel 2008) and are often the subject of vigorous debate (Meyer, Choppy, Staunstrup, and van Leeuwen 2009). In this Perspective article we highlight some issues that, in our opinion, are not adequately discussed and deserve to be brought to the attention of the community.

© Alberto Bartoli and Eric Medvet
Received 20 January 2014; accepted 26 June 2014
A preliminary version of this work was posted on Arxiv (http://arxiv.org/abs/1308.1946).
Address correspondence to Alberto Bartoli, Department of Engineering and Architecture, University of Trieste–DIA, Via Valerio 10, Trieste, Italy. E-mail: bartoli.alberto@univ.trieste.it

First, in the past researchers were forced to focus their reading efforts on a limited set of high-quality publication venues. Today, researchers have all venues at their fingertips and query search engines for the keyword sets they deem relevant. Content relevance has thus become a key criterion for bringing a paper to the attention of a researcher, largely independent of the rigor of the reviewing process, of the acceptance criteria enforced by editors, and so on. The likelihood of a paper getting cited today is critically dependent on ease of Web discovery, ease of access, and content relevance—features that are orthogonal to quality.

Second, low citation counts for high-quality works are extremely common. This assertion may perhaps sound obvious to people familiar with bibliometric research (Wallace, Larivière, and Gingras 2009). Yet research evaluation tends to value high citation counts. While a high citation count may be a good indicator of the quality of a paper, a low citation count tells essentially nothing about quality. Based on actual numbers from very high-level publication venues, we show that use of citation counts as a proxy for quality wipes away the vast majority of papers published in those venues.

We also elaborate on the implications of these points by examining a recent nationwide evaluation of researchers performed in Italy.

DISCOVERY OF RELEVANT WORKS: THE KEY ROLE OF SEARCH ENGINES

The bibliographic practices of researchers have changed radically in recent years, especially with regard to the discovery of relevant works (Schonfeld, and Housewright 2010). Younger researchers perhaps cannot fully appreciate how bibliographic research had to be carried out in the past, say up to the mid 1990s. The typical researcher had a subscription to just a bunch of journals and proceedings, which were delivered directly to her office. It was

Downloaded by [University of Trieste] at 01:22 03 October 2014

necessary to go to the library every now and then, to have a look at the new issues of other publications of interest. Only the latest issues of the most important journals and proceedings were in evidence; all the other material was archived in a specific part of a specific shelf of a specific corridor of a specific room. In many libraries the researcher could freely browse only the recent issues of a few publications; all other publications had to be fetched by the clerks who had to be told exactly which issue of which publication located in which shelf was of interest. Only a few papers could be examined while in the library, and only a few other papers could be photocopied and carried away.

The key point is that examining all recent works about a specific topic, say, "automatic generation of regular expressions," was simply not possible.

Prior to the diffusion of the Web, researchers were forced to focus their reading and search efforts on a well-defined and usually small set of journals and proceedings. As a result, papers published in these venues had clearly a much higher chance of being cited than papers published elsewhere. On the other hand, there was strong competition for publishing in these venues precisely because of the much higher chance to be read there and, thereby, cited. In fact, it is this logic that ultimately justifies the use of citation counts as a surrogate for measuring the quality of publication venues.

Today researchers still follow the few top-tier venues in their respective fields of interest, but they also increasingly rely on search engines, whether general-purpose (e.g., Google, Bing, Yahoo) or specialized for scholarly documents (e.g., Google Scholar, Microsoft Academic Search, CiteseerX). Researchers query search engines for the keyword sets they deem relevant. All papers related to the chosen keyword set have essentially equal chances of being included in the result set, irrespective of the rigor of the reviewing process, of the acceptance criteria enforced by editors, and so on. As a result, these factors no longer play a critical role in determining the set of papers that may potentially be cited.

Furthermore, the ranking algorithm of the search engine plays a critical role in determining which papers are actually brought to the attention of a researcher. Unfortunately, the role of quality—in the broad sense already sketched here—in this process is unclear; an algorithm for reliably quantifying quality of a scholar article has yet to be found. Result construction and ranking of search engines mix quality estimate with content relevance, but the details of this procedure are engine dependent and not publicly known. In practice, quality estimate may easily be obfuscated by content relevance (Gargouri, Hajjem, Lariviere, Gingras, Carr, Brody, and Harnad 2010).[1] Besides, the recent recommendation service by Google Scholar seems to be entirely content based. Citation counts appear to play a role in ranking algorithms, but their role with respect to content relevance is, again, engine dependent and not publicly known. Ranking may also be affected by feedback from users in the form of mere downloads of specific results. In summary, in the processes that shape what gets read, quality plays an increasingly reduced role. That is one example of how the introduction of a new communication or collaboration technology (in this case, search engines and the Web) in a research community disrupts the status quo (Grudin 2013).

The fact that the likelihood of getting cited is becoming decoupled from the quality of the publication venue is confirmed by recent bibliometric research: "Throughout most of the 20th century, papers' citation rates were increasingly linked to their respective journals' Impact Factors. However, since 1990, the advent of the digital age, . . . the proportion of highly cited papers coming from highly cited journals has been decreasing, and accordingly, the proportion of highly cited papers not coming from highly cited journals has also been increasing" (Lozano, Larivière, and Gingras 2012, 2140).

In effect, the ease of Web access rather than the imperative of a thorough literature review is influencing reading efforts. On the basis of an analysis that examined 27,000 articles published in nearly 2000 journals in the years 2000–2006, Gargouri et al. (2010) observe that papers made accessible in Open Access form are cited significantly more than papers in the same journal and year that have not been made Open Access (Gargouri et al. 2010). Similarly, back in 2001, Lawrence, after analyzing 120000 papers published in computer science conferences, noted: "The results are dramatic, showing a clear correlation between the number of times an article is cited and the probability that the article is (freely) online" (Lawrence 2001, 521). Morever, "If we assume that articles published in the same venue (proceedings for a given year) are of similar quality, then the analysis by venue suggests that online articles are more highly cited because of their easier availability" (Lawrence 2001, 521). Researchers need to become much more aware of these facts. Publishers have already acted on this issue—they increasingly allow forms of open access even for journals or proceedings requiring a subscription fee.

One might argue that authors are competent enough to select what is worth citing and what is not. What needs to be considered here is that authors often do not have the experience or specific competence of a reviewer. In the past, authors who focused their readings on publications in high-quality venues automatically benefitted from the informed filtering by editorial boards. Today, this form of quality certification is increasingly overwhelmed by radically different factors such as ease of Web discovery, ease of access, and content relevance.

Downloaded by [University of Trieste] at 01:22 03 October 2014

Of course, authors could exploit ease of Web access and actually read every potentially relevant work and then decide by themselves which works deserve to be cited. We believe it is fair to say that this approach is not practical: Elsevier, Wiley, and IEEE alone publish more than 320 computer science journals and the IEEEXplore library alone contains materials tagged "computing and processing" from 1189 conferences held in 2013.

Contemporary technology might allow novel forms of discovery of quality works, though. A cloud service storing the personal library of a large population of researchers could exploit several signals beyond content relevance for providing recommendations to its users. For example, the presence of a certain work in the personal library of many researchers could be a signal that correlates with the quality of that work. Another signal could be the number of times a researcher reads that work, as well as the time distribution of accesses—a work open only once or twice shortly after its insertion in the library is probably less useful than another work accessed many times across several weeks. Allowing researchers to rate works in their respective libraries could also be highly useful in this respect. Furthermore, the necessary parameter tuning for such recommendation engines could be tailored based on the number of researchers working in each sector; for example, presence of a certain work in a few dozens of personal libraries might be either irrelevant or a clear indication of high quality, depending on how many researchers actually work in that area. Services of this kind have already appeared in the form of citation managers,[2] and even Google Scholar now supports a notion of personal library. While such recommendation services have a huge potential, it is fair to say that their actual impact is currently quite marginal. In our estimation, such services would be truly useful only if they were the primary bibliographic tools for a large fraction of the population of researchers. Recommendations built upon a partial view of the library of a small fraction of researchers would hardly be very meaningful.

LOW CITATION COUNTS FOR HIGH-QUALITY WORK ARE VERY COMMON

Every year, a significant percentage of high-quality papers either never get cited at all or take just a bunch of citations. This assertion may perhaps sound obvious to people familiar with bibliometric research (Wallace, Larivière, and Gingras 2009), but looking at some actual numbers may be quite insightful for a broader audience.

We collected the citation counts for all papers published in some top-level venues of the respective fields (Table 1, first column). Certainly, the list is not exhaustive and not every paper published in these venues is groundbreaking. It is equally true, though, that every one of these papers has gone through a state-of-the-art reviewing process, that is, a careful analysis by several independent people who are recognized as experts by the relevant scientific community. Thus, it is fair to say that any reasonable assessment of research quality must necessarily assign to those publications a weight that is not negligible—note that we do not insist in providing a general definition of "quality," that is, a definition that may be applied to any paper in any publication venue. We focused on the years 2000–2009: a range sufficiently small that can be analyzed relatively easily and sufficiently large to filter out possible anomalies. It also extends into the past sufficiently to allow reasonable time for each paper to be cited. We collected the data from Microsoft Academic Search API.

The results are listed in Table 1. It is clear that a significant percentage of papers published in these high-level venues get a negligible amount of citations (the meaning of the last column will be explained in the next section). Different views of the same citation data are available at http://machinelearning.inginf.units.it/data-and-tools/paper-citations-for-important-cs-venues; such data show, in particular, that invariably every year just a few papers collect a nonnegligible amount of citations.

The implications of these data are quite odd when citation counts play an essential role in the evaluation of researchers, as discussed in the next section.

EVALUATION OF RESEARCHERS

The Italian government has recently established bibliometric-based conditions that researchers have to satisfy in order to be eligible for a faculty position.[3] These conditions are based on (we omit several details for brevity):

1. Number of journal publications.
2. Citation counts.
3. Contemporary h-index (a variant of the h-index in which the number of citations collected by a given paper is normalized, i.e., multiplied by 4 and divided by the number of years elapsed from the time of the publication).

The values of these parameters that need to be exceeded are the median values computed across all the researchers in a given discipline. For example, the values for becoming eligible for a position of associate professor in computer science are the median values across all the associate professors in computer science. The source of data are Scopus and Web of Science—as an aside, it has been argued that these sources are not adequate for computer science (Meyer, Choppy, Staunstrup, and van Leeuwen 2009). Each year, a nationwide evaluation is made in which a panel of five experts for each discipline assesses all the candidates (ASN, "Abilitazione Scientifica Nazionale,"

TABLE 1
Citation data for a few selected high-level computer science publication venues (four conferences, four journals), where c is the number of citations and c' is the normalized number of citations (see text)

		Percent of papers with		
Publication venue	Papers	$c = 0$	$c \leq 5$	$c \leq 6$
ACM Symposium on Operating Systems Principles (ACM SOSP)	159	21.4	40.9	49.7
Symposium on Principles of Distributed Computing (ACM PODC)	624	21.5	48.1	63.9
Internet Measurement Conference	203	7.4	25.6	38.9
IEEE Symposium on Security and Privacy	271	4.1	15.9	24.7
ACM Transactions on Computer Systems	134	3.7	20.1	36.6
ACM Transactions on Internet Technology	164	7.3	26.2	40.9
IEEE Transactions on Knowledge and Data Engineering	1311	14.5	38.1	58.6
IEEE Transactions on Software Engineering	851	17.9	31.5	45.7

Downloaded by [University of Trieste] at 01:22 03 October 2014

http://abilitazione.miur.it). Candidates assessed positively for a given position are entitled for 4 years to apply for that position at some university—if and when such a position becomes available. In the first evaluation, there were 260 and 413 candidates for full professor and associate professor in computer science,[4] of which 96 and 176 have been assessed positively.[5]

After the promulgation of the decree by the government, it was not fully clear whether satisfaction of these bibliometric conditions was indeed necessary for a candidate to be assessed positively. It was clarified that normally a positive evaluation should be reserved exclusively to candidates who satisfy at least two of these conditions, but each panel is free to assess positively also candidates who do not satisfy them, provided the panel gives a motivated extremely positive assessment of the candidate.

Concerning the first criterion, this short perspective article would count the same as, for example, an IEEE Transactions on Software Engineering paper. Papers published at top-level conferences like ACM SOSP and ACM PODC do not count.

More useful observations can be made based on the citation data from the previous section:

1. A significant percentage of papers contribute very little to citation counts, if at all (Table 1, third and fourth columns). Those papers might even contribute less than this short perspective article, in case this article was cited.
2. A large percentage of papers—for many venues, most of them—are completely useless for satisfying the third criterion, that is, contemporary h-index. Such a percentage is given in the last column of Table 1 and the numbers speak for themselves.[6]

It is important to note that we are considering papers published at top venues and that these odd outcomes are not a sort of rare or extreme event: They occur routinely for a large percentage of these papers.

Of course, the assessment of researchers takes into account several additional criteria in which the (panel-perceived) quality of a publication venue does play a role. The key point is, such criteria are of secondary value because bibliometric conditions are the essential requirement for participating in the game: They are necessary by default and may be neglected only in exceptional cases.

As an aside, it may be worth pointing out that the very same definition of the bibliometric conditions (median values across either associate professors or full professors in a given discipline) implies that a lot of the researchers who currently occupy a given position could not be assessed positively.

AN ECOSYSTEM WITH NECESSARY-BY-DEFAULT BIBLIOMETRIC MEASURES

An ecosystem with necessary-by-default bibliometric measures, with incentives heavily based on citation counts and where discovery of relevant literature is mostly based on content relevance and citation counts, could evolve along unexpected paths and challenge several of the assumptions traditionally taken for granted.

Submitting to a top-level journal may lead to several review rounds, each requiring a significant amount of work, with a turnaround time that may be years and that may conclude with a rejection. From a mere bibliometric point of view, submitting to such a journal may be an

Downloaded by [University of Trieste] at 01:22 03 October 2014

irrational move: high cost and high risk with little, if any, advantage in return. When playing a game with necessary-by-default bibliometric measures, there is definitely a strong advantage in submitting to journals with easier-to-satisfy editorial boards. As soon as a paper has been published on some "indexed" journal, it counts the same as any paper published on more prestigious journals; most importantly, the paper becomes ready to appear in search engine results and, thus, ready to be cited. Furthermore, citations to a paper presented in a poster session of a low-profile conference count the same as citations to a paper published in a top-level journal.

Based on these observations, editorial boards and commercial publishers could even have an incentive to lower the bar while maintaining a sort of "decent" quality. Indeed, bibliometric incentives are already playing a key role in the market opportunities for journals with little or no scrutiny,[7] a category including also journals from prestigious publishers and institutions (Technopolis 2009).

In other words, one could end up with pervasive incentives toward avoiding high-quality reviewing processes, which would have disrupting effects in the long term that can be imagined easily. Concerns about possible perverse incentives for academics to publish in weaker journals have been raised recently as a result of the bibliometrics pilot executed in the context of a nationwide evaluation of research institutions currently in progress in the United Kingdom (Technopolis 2009).

Spam and security problems would become a serious issue also in research platforms, as people are likely to manipulate publicly available pdf files for artificially increasing citation counts. Manipulating Google Scholar to this end is relatively straightforward (Labbe 2010). In 2010 Google Scholar was reporting 102 publications and an h-index of 94 for a fake author (Lozano, Larivière, and Gingras 2012). It is important to note that even a transitory artificial increase in the citation count of a real paper would suffice, as it would allow the fraudulently promoted paper to emerge in academic search engines from the ocean of low citation counts and thus to start collecting legitimate citations.

Finally, but not least importantly, diversity of research topics and freedom of exploration would greatly suffer, as research in areas that are not mainstream is unlikely to collect citations and thus would be penalized at evaluation time.

FINAL THOUGHTS

Research evaluation is a very difficult problem and there are countless proposals for approaching it. We believe that either promoting a specific proposal or suggesting yet another one would be a futile exercise on our part. We prefer to make some general remarks.

First, the IEEE Board of Directors (2013) recently issued recommendations for research evaluation, stating very clearly that "bibliometric performance indicators should be applied only as a collective group (and not individually)" (1). Furthermore, "while bibliometrics may be employed as a source of additional information ... the primary manner for assessment ... of an individual scientist should be peer review" (2). The same recommendation, that bibliometrics "must only be used to inform a peer-review process" (Technopolis 2009, 4) for the cited report, has been made in the earlier mentioned bibliometric pilot of the research evaluation in the United Kingdom. Evaluation of individual researchers in Italy has been instead built upon necessary-by-default bibliometric conditions.

Second, the evaluation strategy implemented in Italy effectively creates an environment where the global objective—improving the overall quality, in a broad sense, of the research system—differs from the objective of individual actors—satisfying necessary-by-default bibliometric conditions. The global objective may or may not be actually improved by actions driven by a different objective: accumulating publications at indexed journals, and collecting citations at indexed venues from indexed venues. As an aside, Nobel laureate Peter Higgs, after whom the Higgs boson particle is named, recently remarked that "today I wouldn't get an academic job. It's as simple as that. I don't think I would be regarded as productive enough" (Aitkenhead 2013, online). While Peter Higgs is an exceptional case, his observation is insightful and pertinent for our analytical purposes.

Third, the current system prompts actions at the level of individual researchers that are motivated mainly or solely by improvement of bibliometrics performance, including artificial joint authorship (a concern raised in Technopolis 2009), reciprocal citations, proliferation of short papers that share research in least publishable increments, speaker-only conferences, and market opportunities for venues of poor or dubious quality (see previous section). Actions of this sort certainly do not contribute toward improving the global objective and thus constitute, from a global point of view, a waste of resources.

Fourth, the scale matters. For example, in Italy's ASN there were 260 applications for full professor and each applicant could submit up to 20 papers for evaluation beyond the bibliometric data; there were also 413 applicants for associate professor who could submit up to 16 papers each. The panel of 5 experts thus had to assess more than 10,000 papers, potentially covering all subfields of computer science. In such a context, it is difficult to implement IEEE Board of Directors' recommendation that research evaluation be primarily based on peer review with bibliometrics merely serving as supplementary data.

Downloaded by [University of Trieste] at 01:22 03 October 2014

Fifth, the driving force behind the increased usage of bibliometrics is the trend toward reliance on research evaluation for allocation of resources at all levels, that is, from nationwide to university-wide, from grant assignments to hiring and promotion. Ranking, say, all universities nationwide based on some combination of bibliometric indicators is certainly much simpler, quicker, and cheaper than conducting an in-depth review of quality. Similarly, bibliometrics offer a straightforward way to compare performance of, say, different research groups or departments within the same university. Moreover, any possible misrepresentation caused by bibliometrics may also be seen as part of a broader trade-off between accuracy and cost of the evaluation itself.

Finally, while ranking all research institutions in a given country may be a necessity for informed political decisions about distribution of public funding, binary classification of all researchers in a given country is not a necessity. If and when a position opens at a university, a panel that is competent and follows the appropriate code of conduct would be capable of choosing, based on peer review, the best candidate. Such a strategy would not suffer of problem of scale discussed earlier and, most importantly, the research evaluation would be conducted without overwhelming reliance on bibliometric criteria.

We believe the research community should become fully aware of the actual meaning of bibliometric measures as well as of their implications when applied to the evaluation of individual researchers. We hope the issues we have raised in this perspective article will be useful in this respect.

NOTES

1. See also "Get found: Optimize your research articles for search engines," http://www.elsevier.com/connect/get-found-optimize-your-research-articles-for-search-engines

2. http://www.mendeley.com/, http://www.citeulike.org

3. Disclaimer: The authors do not satisfy these conditions.

4. In the Italian system there are actually two somewhat overlapping disciplines: "Sistemi per l'Elaborazione delle Informazione" and "Informatica." The results reported here correspond to the former. The results of the latter are similar.

5. Disclaimer: The first author participated in the evaluation for full professor and was not assessed positively.

6. To compute these values we normalized the number of citations taken by each paper as dictated by the definition of this index explained earlier; that is, we multiplied the number of citations c of a paper published at year Y by 4 and divided the result by $2012 - Y + 1$, where 2012 is the year at which the parameters of candidates have been computed.

7. List of predatory publishers 2014: http://scholarlyoa.com/2014/01/02/list-of-predatory-publishers-2014

REFERENCES

Aitkenhead, D. 2013. Peter Higgs: I wouldn't be productive enough for today's academic system. *The Guardian*, December 6. http://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system (accessed June 19, 2014).

Beel, J., G. Bela, and E. Wilde. 2010. Academic search engine optimization—ASEO. *Journal of Scholarly Publishing* 41(2): 176–90.

Bornmann, L., and H. D. Daniel. 2008. What do citation counts measure? A review of studies on citing behavior. *Journal of Documentation* 64(1): 45–80.

Gargouri, Y., C. Hajjem, V. Lariviere, Y. Gingras, L. Carr, T. Brody, and S. Harnad. 2010. Self-selected or mandated, open access increases citation impact for higher quality research. *PLoS One* 5(10): e13636.

Grudin, J. 2013. Varieties of conference experience. *The Information Society* 29(2): 71–77.

IEEE Board of Directors. 2013. Appropriate use of bibliometric indicators for the assessment of journals, research proposals, and individuals. http://www.ieee.org/publications_standards/publications/rights/ieee_bibliometric_statement_sept_2013.pdf (accessed June 19, 2014).

Labbe, C. 2010. Ike Antkare one of the great stars in the scientific firmament. *International Society for Scientometrics and Informetrics Newsletter* 6(2): 48–52.

Lawrence, S. 2001. Free online availability substantially increases a paper's impact. *Nature* 411: 521.

Lozano, G. A., V. Larivière, and Y. Gingras. 2012. The weakening relationship between the impact factor and papers' citations in the digital age. *Journal of the American Society for Information Science and Technology* 63(11): 2140–45.

Meyer, B., C. Choppy, J. Staunstrup, and J. van Leeuwen. 2009. Research evaluation for computer science. *Communication of the ACM* 52(4): 31–34.

Schonfeld, R., and R. Housewright. 2010. *Faculty survey 2009: Key insights for libraries, publishers, and societies*. http://www.sr.ithaka.org/research-publications/faculty-survey-2009 (accessed June 19, 2014).

Technopolis. 2009. Identification and dissemination of lessons learned by institutions participating in the Research Excellence Framework (REF) bibliometrics pilot: Results of the Round Two Consultation—Report to HEFCE by Technopolis. http://www.hefce.ac.uk/media/hefce/content/pubs/2009/rd1809/rd18_09.pdf (accessed June 25, 2014).

Wallace, M. L., V. Larivière, and Y. Gingras. 2009. Modeling a century of citation distributions. *Journal of Informetrics* 3(4): 296–303.