FDD3001 Homework 1
Bibliometry - a Pseudo-science?

Christian Smith
e-mail: ccs@kth.se

Abstract— **In this paper bibliometry as a scientific discipline is evaluated in order to see if it should actually be regarded as pseudo-science. An analysis of several quantative studies show that large parts of the field of bibliometry is not pseudo-science in the Popper sense of the world, as they can be shown to be clearly false.**

I. Introduction

Bibliometry has become an ever increasing factor in evaluation of scientific performance of not only journals, but of universities, research groups, and individual researchers as well. This has caused controversy, since not all researchers agree that common bibliometric measures are an accurate description of scientific value. There are even those that dispute the very existence of a meaningful metric on this subject [1]–[3].

The aim of the present paper is to examine the field of bibliometry itself, to see if it is based on the rigorous practices that should characterize all real science, or if it indeed belongs to the realm of pseudo-science. The method will be a meta-study of other studies, citing both critics and proponents.

II. Definitions

The first step of a rigorous analysis is to define the domains. In order to test the postulate that bibliometry is a pseudo-science we need a definition of both bibliometry and pseudo-science.

A. Pseudo-science

The term pseudo-science is first accredited to philosopher Karl Popper, who defines science as something that puts forth claims that can be tested by falsification. Consequently, a "pseudo-science" is a discipline that does not produce falsifiable claims [4], or in the words of physicist Wolfgang Pauli, "is not even wrong" [5]. This is a good working definition, and is the one that will be applied in the present paper, with the extension that if said claims are falsified, they should be revised or abandoned in order for a field to classify as science in practice.

It should also be made clear that applications of a supposedly scientific set of theories or hypothesis to problems not covered by the falsifiable claims also are to be considered pseudo-science. For example, even though Maxwell's laws of electrodynamics make scientific falsifiable claims regarding electric fields, it would be pseudoscience to apply them to economics.

B. Bibliometry

There is no commonly agreed-on definition of what exactly constitutes the discipline of bibliometry. In this context, however, we will adopt a working definition that is as close as possible to the practices applied to bibliometric evaluation of researchers and institutions. As a model, the metrics used in the 2008 KTH Research Assessment Exercise [6] are used. These include — but are not limited to — the following parts:

1) Journal Impact Factors (JIF): This is perhaps the most well known bibliometric indicator. It is a measure of the average number of of citations articles in a particular journal receive within two years of publication, where citations are defined as "citations found in database X". Here X is often the ISI Science Citation Index.

2) Citation Per Publication (CPP): This indicator is simply the average number of citations per publication, where citations are defined as "citations found in database X". Here X is often the ISI Science Citation Index, but can also be Scopus or Google Scholar.

3) Field Normalized Citation Score (NCSf): This indicator is basically the same as CPP, but normalized with the average number of citations for the research field.

4) Vitality: This is a measure of how recent the references cited by a paper are. A common way to define this is the average age of references normalized with the average for the research field.

C. Statistical Science

Bibliometry is commonly presented as a statistical science [1], [6]–[10], but mathematical statisticians do not seem to share this view, but note that normal statistical practice is rarely followed in bibliometry [11], [12]

D. Direct Bibliographical Measures

One major application of bibliometry, that is not very controversial, is to interpret all measures "as is". This means that we let the number of citations tell us how many times an article has been cited, the number of publications tell us how many articles an author has written, and co-author measures tell us who an author has collaborated with, and so on. These measures are truistic, and used in order to analyze the history of scientific fields or ideas or biographies of researchers. They will not be the main point of this paper, but are mentioned here for completeness.

III. Bibliometrical Claims

In this section we examine what (if any) falsifiable claims are put forth regarding the bibliometrical metrics. If the claims are falsifiable, we will try to see what attempts have been made to falsify these, and what outcome these attempts have had.

A. Journal Impact Factors (JIF)

When the JIF metric was first introduced by Garfield, the original motivation was to aid libraries in choosing what journals to subscribe to [7]–[9]. A proposed application of impact factors is exemplified as "Thus, we can say with reasonable certainty that any biochemistry librarian would be well advised to have Lowry's article on protein analysis available, since it is the most frequently cited paper in the field. On the other hand, this same information should be used with caution for personnel selection and evaluation" [8].

The claim made here is that the average number of citation per article is a measure of how useful a certain publication is to a scientific community, since the number of citations should correlate to the number of times a paper has contributed to other's work. This claim seems falsifiable, as long as there exists some objective measure of a publication usefulness that could be compared to the JIF for a correlation calculation.

An experiment that would falsify this claim of impact factors, is to show a significant body of articles in journals with a low impact factor that have significant usefulness. A simple demonstration of such an experiment is made in [3], where it is shown that the database most commonly used for calculating JIF, the "ISI Science Citation Index" has poor coverage of computer science, for example only 14% of the publications in computer science at ETH Zurich are indexed. Of the total publications, conferences and workshops make up 65% of the total, and almost none of these are indexed. Since these non-indexed venues were shown to have almost the same average amount of citations per paper (7.3 versus 7.5), the correlation between commonly calculated JIFs and publication quality does not seem strong, at least not for the field of computer science. Some indexes, such as Elsevier's Scopus[1], do take a small subset of conferences into account, but as it only contains 51 computer science conferences, or offers coverage of only 5 conferences in robotics, with only partial coverage of some years, there are clearly fields that are not decently covered. Thus, even if the original claim that the number of citations per paper correlate well with publication quality should hold, for some fields it should not be expected that a correct count of citations is performed.

However, for some fields, like the field of economics, the correlation coefficient of impact factor and peer ranking has been shown to be 0.93–0.98 [13], so it seems plausible to conclude that there exists fields where indexed journals have JIF that correlate strongly with other measures of journal quality. Thus, for these fields, using JIF to determine journal quality should be as good as any other measure.

Another more recent claim regarding the JIF is that there is a relevant correlation between the quality of an individual paper and the JIF of the journal where it is published [6]. If we accept citation counts as a measure of quality, it is of course by definition true in the statistical sense that there is a positive correlation, but there is no valid argumentation for why a poorly cited paper should be considered to be of higher quality if other papers appearing in the same journal are more highly cited than average [11], [14]. On the contrary, it would seem intuitive that a poorly cited paper in an obscure journal may have been poorly cited due to low readership, while a paper in a highly read journal should be highly cited if it is useful to other researchers, so that a low level of citation can not be blamed on anything but the paper's quality. As pointed out in [15], for an individual paper, the JIF of the journal where it was published correlates weakly (correlation coefficients of 0.21–0.40) with other measures of paper quality.

B. Citation Per Publication (CPP)

This brings us to the measure that is used axiomatically in JIF derivation, the raw counting of the number of citations a paper has recieved. This may be the total number of citations found anywhere, or limited to a certain time period or publication subset, but the main claim is consistently that the number of citations is a valid measure of the quality of a paper. If this claim is true, CPP should correlate well to other quality measures.

As pointed out in [16], there are several factors that determine what papers are cited, and the main factor is not necessarily the quality of the cited paper, but its "citability". For example, it is more convenient to cite a review paper than all the individual papers reviewed. Also, there is tendency to cite papers already cited previously, either by copying one's own citation list from an earlier paper, or by copying another paper's citations, thus ruling out the need to find the primary sources oneself. Apart from these effects, there are several other types of citations that make the CPP measure questionable, such as self or collegue citations and negative citations. It is difficult with current technology to treat these accurately in an automated manner, however, it is still argued that even with the noise added by these non-valid citations, the underlying correlations should still be statistically significant [6].

There is no authorative study on the correlation of the quality of a single paper and the number of citations it has received, but there are some studies on correlations of research quality and larger numbers of papers. According to the compilation presented in [17], several studies show a positive correlation between the aggregated number of citations for a unit and other quality measures. However, this correlation is weak for units where the number of papers is not very large. For example, evaluation of individual researchers correlate as low as 0.2 in two of the 7 studies cited, from 0.3 to 0.5 in 2 studies, and from 0.6 to 0.9 in the remaining 3 studies. For departments, the correlations vary from 0.67 to 0.85, showing that the huge sample base that should be available on departmental level is

[1] www.scopus.com

needed before citation counts start to give an indication of quality.

An example of the poor correlation between citations and peer-reviewed performance is the list of the recipients of the Turing Award. This award is given annualy by the ACM, and most computer scientists would agree that it is one of the most prestigious awards in computer science. However, when the list of award winners from the years 1984–2002 is cross-checked to the Citeseer "most cited" ranking list for the field of computer science, the average[2] *ranking* for award winners is 1542^{nd} place [3].

There are other factors that correlate more strongly to citation counts than perceived paper quality. For example, the number of authors per paper has a significant correlation of 0.84 to the number of citations [18].

Thus it can be assumed that there exists a correlation between the number of citations per paper and paper quality. However, it is so weak that it can only be used to evaluate large units, and is a very blunt instrument in evaluating individual researchers, and probably utterly meaningless in evaluating individual papers, since the number of citations needed to have a stasticially significant deviation from the average would only be attained by a diminishingly small number of researchers and/or papers, leaving the verdict for the majority of assessed units as "inconclusive".

C. Field Normalized Citation Score (NCSf)

One of the criticisms of citation counts is that they are not comparable between disciplines, as there are large differences in citation habits. The NCSf is one attempt to compensate for this by normalizing the citation counts with the average for that discipline [6]. The falsifiable claim here is that by doing this normalization, meaningful cross-discipline or cross-subdiscipline comparisons can be made, and also that, "Sum of NCSf indicates the total impact of the Unit of Assessment" [6].

However, this seem to be false for at least those fields that traditionally have lower CPP scores, like computer science. For these fields, the NCSf correlates much stronger with what neighboring fields with high CPP that lie close to a publication, since they will in essence be normalized by the wrong normalization factor.

As shown in [3], using this approach means that papers that border on biomedicine completely dominate the citation counts for computer science, since the average paper in biomedicine has more than 6 times as many citations as an average paper in computer science [18]. An illustration of this phenomenon is that only 15% of Scopus top cited computer science articles were actually core computer science papers. However, it is very difficult if not impossible to find an objective measure of to what extent a paper belongs to what fields of research, in order to find a correctly weighted normalization factor.

D. Vitality

The vitality measure is a measure of the average age of a paper's references, or alternatively the proportion of references newer than a certain limit. The claim made here is that "researchers which use the most recent references to their articles probably are the ones that are committed to participating at the forefront of science rather than on older science" [10].

The motivation for using this averaged reference age for vitality is based on studies that show that this correlates well with good performace for patents, while there are only heuristic arguments as to why this implies that it is also important for research publications. Also, vitality is often used together with normalization for research field averages, for which the same arguments as in the previous section apply.

The most obvious problem with the vitality measure is that it is trivial for an individual author to doctor this number at will. Normally, similar information is available in several sources, and it is more or less up to the individual author to decide which references to include. For instance, if the entire historical section is summed up with a reference to a recent review article, the average age would decrease significantly. Likewise, the number of references to contemporary work can easily be increased, as the decision line regarding what references are relevant enough to include is also up to the discretion of the author.

IV. Conclusion

The overall conclusion is that there seems to be scientific value in parts of the bibliometrical field. For instance, there are falsifiable claims that the number of citations correlate with other measures of quality that have withstood the test of falsification. Most notably there seems to be a reasonable agreement between JIF and the subjective evaluation of journals. However, there are publications — like monographs and conferences — that are not satisfactorily measured.

As for the other bibliometrical measures, it seems that the claims they make are falsifiable, but more or less false. The correlations that are claimed may exist, but they are too weak to have any meaningful application when used for smaller units. Also, there are other factors that correlate more strongly with these bibliometrical measures, and especially for low-citation fields such as computer science, it would seem that the noise-to-signal ratio is greater than 1 in several cases.

The general recommendation in many accounts of bibliometrical studies — even by their proponents [6], [9], [10], [17] — is to use bibliometrical measures with caution when evaluating research groups or individuals. It is not clear how they reach this conclusion. Given the statistical studies cited in the present paper, the only sane recommendation should be to only use bibliometry where it is statistically sound, for example as a measure of the expected utility of a journal to a library. Using bibliometry for other uses than this should be avoided. Any introductory course in business evaluation will teach you that having no measure is preferable to a poor measure, as the latter is likely to lead you to the

[2] The average was calculated leaving out the two award winners that were not even on the list.

wrong conclusions under the wrong assumption that you are right. The biblionetrical study in [6], for example, provides no measures of statistical significance whatsoever, making it impossible for a reader to evaluate the content.

REFERENCES

[1] J. Giske, "Benefitting from bibliometry," *Ethics in Science and Environmental Politics*, vol. 8, pp. 79–81, June 2008.

[2] P. Lawrence, "The mismeasurement of science," *Current Biology*, vol. 17, no. 15, pp. 583–585, 2007.

[3] F. Mattern, "Bibliometric evaluation of computer science," in *European Computer Science Summit (ECSS)*, 2008.

[4] K. Popper, *Conjectures and Refutations : the Growth of Scientific Knowledge*. London: Routhledge, 1963.

[5] R. E. Peierls, "Wolfgang Ernst Pauli. 1900–1958," *Biographical Memoirs of Fellows of the Royal Society*, vol. 5, no. 1, pp. 174–192, Feb 1960.

[6] E. Sandström and U. Sandström, "KTH RAE 2008 bibliometric study," Sep 2008.

[7] E. Garfield, "Citation indexes to science: a new dimension in documentation through association of ideas," *Science*, vol. 122, no. 3159, pp. 108–111, 1955.

[8] E. Garfield and I. H. Sher, "New factors in the evaluation of scientific literature through citation indexing," *American Documentation*, vol. 14, no. 3, pp. 195–201, 1963.

[9] E. Garfield, "The agony and the ecstacy — the history and meaning of the journal impact factor," in *International Congress on Peer Review And Biomedical Publication*, Chicago, Sep 2005.

[10] R. Klavans and K. Boyack, "Thought leadership: A new indicator for national and institutional comparison." *Scientometrics*, vol. 75, no. 2, pp. 239–252, 2008.

[11] R. Adler, J. Ewing, and P. Taylor, "Citation statistics," Joint Commitee on Quantitative Assessment of Research. Report from the International Mathematical Union (IMU) in cooperation with the International Council of Industrial and Applied Mathematics (ICIAM) and the Institute of Mathematical Statistics (IMS), Tech. Rep., 2008.

[12] R. L. Karandikar and V. S. Sunder, "On the impact of impact factors," *Current Science*, vol. 85, p. 235, Aug 2003.

[13] W. C. Bush, P. W. Hamelman, and R. J. Staaf, "A quality index for economics journals," *Review of Economics and Statistics*, vol. 56, pp. 123–125, Feb 1974.

[14] T. C. Ha, S. B. Tan, and K. C. Soo, "The journal impact factor: Too much of an impact?" *Ann Acad Med Singapore*, vol. 35, pp. 911–916, 2006.

[15] T. Berghmans, A. Meert, C. Mascaux, M. Paesmans, J. Lafitte, and J. Sculier, "Citation indexes do not reflect methodological quality in lung cancer randomised trials," *Annals of Oncology*, vol. 14, no. 5, pp. 715–721, May 2003.

[16] P. Seglen, "Citation rates and journal impact factors are not suitable for evaluation of research," *Acta Orthopaedia Scandinavia*, vol. 69, no. 3, pp. 224–229, 1998.

[17] F. Narin, *Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity.* New Jersey: Computer Horizons, Inc., 1976.

[18] M. Amin and M. Mabe, "Impact factor: use and abuse," *Perspectives in Publishing*, vol. 1, pp. 1–6, Oct 2000.