A simple back-of-the-envelope test for self-citations using Google Scholar author profiles

Frode Eika Sandnes[1*,2]

[1]Department of Computer Science, Faculty of Technology, Art and Design, Oslo Metropolitan University, P.O. Box 4, St. Olavs plass, 0130 Oslo, Norway

[2]Faculty of Technology, Kristiania University College, Kirkegata 24-26, 0153 Oslo, Noway

Email: frodes@oslomet.no

Tel: +47 466 97 592

ORCID: 0000-0001-7781-748X

MSC: N/A

JEL: I21, I23, I28

Abstract

The issue of self-citation has received much attention in academia. Widely used and accessible tools such as Google Scholar do not provide information about self-citations. Therefore, a simple and practical back-of-the-envelope test for identifying researchers with strategic self-citations is proposed without access to self-citation information. It is shown that the h-index squared divided by the number of citations predicts self-citations. The test is simple to apply based on Google Scholar author profiles. Bibliometric data for more than 100,000 researchers worldwide were used to assess the proposed test. Test values of 0.35 or more indicate high ratios of self-citation while test values below 0.2 suggest low ratios of self-citations.

Keywords: self-citations, test, researcher, h-index, publication metrics

Introduction

The issue of self-citations has received much attention in the academic community (Van Noorden and Singh 2019). Self-citations are often perceived negatively as they can result in misleading impressions of a researcher's impact. The debate becomes especially heated when publication metrics are used as incentives (Haugen and Sandnes 2016, Sandnes 2018). Indeed, claims of misleading impact hold if bibliometric data are not corrected for self-citations. On the other hand, one may argue that the use of citations to measure impact by itself is misleading. Self-citations can be a means for researchers to place their ongoing research in context of their prior research. Anne-Wil Harzing quite appropriately states that "What is a problem is a lack of non-self citations, i.e., the fact that other academics are not referring to these academics articles." (Harzing 2010). Hartley (2012) makes a convincing argument that self-citations should discriminate between those that are informative and those that are self-promoting for self-citation measures to be more credible. Unfortunately, such discrimination may be hard to achieve in practice.

Self-citation information can thus be used negatively to identify what some consider "cheating", but it can also be used positively to identify, say, young academics that are conscious about strategically promoting their research. Regardless of how the self-citation information is used, it can be non-trivial for researchers and students to obtain. Google Scholar, for instance, is a popular source for citation information as it has been shown to have the widest coverage (García-Pérez 2010; Gehanno,

Rollin, and Darmoni 2013). For certain disciplines Google Scholar is the only available source of citation information (Sandnes and Grønli 2018; Sandnes and Brevik 2019). However, Google Scholar author profiles show all citations without corrections for self-citations. Google scholar also provides the h-index (Hirsh 2005; Bornmann and Daniel 2005), and the Google scholar h-index is a popular yardstick among researchers.

To help overcome the lack of self-citation information, a simple test is proposed to help identify researchers with strategic self-citations. The test is solely based on a total citation count and h-index as provided by Google Scholar author profiles. It is intended as a simple back-of-the-envelope tool to help identify authors that are likely to cite their own works strategically. Researchers with high test values could be flagged for further analysis using more specialized bibliometric tools. As the test does not rely on self-citation data, the test does not actually provide evidence of self-citation.

Previous studies on self-citations have addressed the ratio of self-citations in original articles and review papers (Falagas and Kavvadia 2006), the relationship between h-index increases and self-citations (Gianoli and Molina-Montenegro 2009), and the relationship between self-citation rates and overall citation counts (Aksnes 2003; Medoff 2006), to mention a few.

The self-citation test

The proposed test uses the *h-index* and the total number of *citations* as provided by Google Scholar author profiles. The proposed test for self-citation is defined as

$$T = \frac{h\text{-}index^2}{citations}$$

Obviously, the test value *T* is a value in the range between 0 and 1. A high value indicates that the author has many self-citations, while a low value indicates few self-citations.

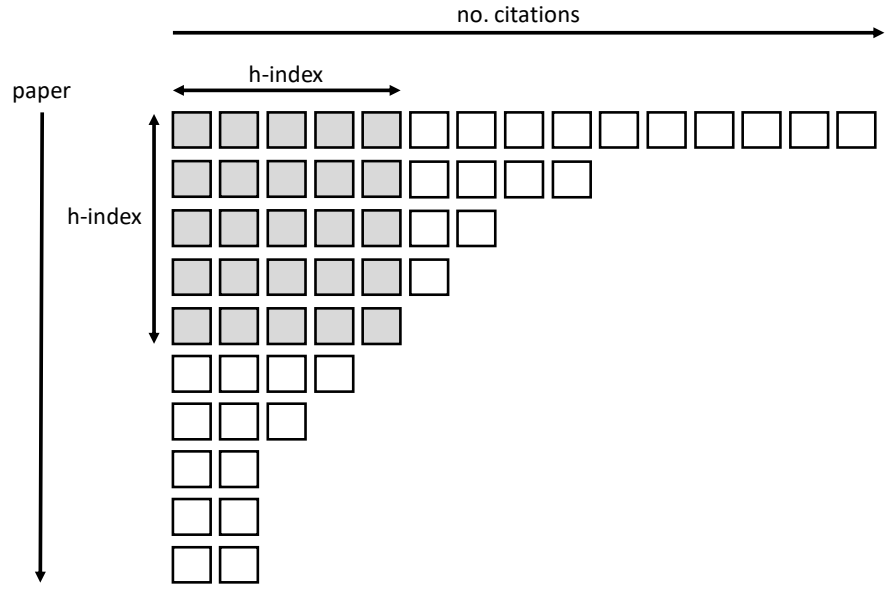

Figure 1. Test for self-citations as a ratio of strategic citations (grey) of all citations (white and grey).

Rationale for the test

The self-citation test assumes that the citation patterns of a typical researchers follow an exponential curve where one article is cited much more than the second most cited article, and so

forth. An author's citations will therefore be unevenly spread across the cited publications. Next, the definition of the h-index is that there must be at least *h* papers with at least *h* citations. Assuming someone did not have any non-self-citations and wanted to achieve a h-index to *h* with as few publications as possible, they would have to cite *h* of their publications *h* times resulting in $h \times h$ self-citations (see Figure 1). Therefore, dividing the observed total h-index squared by the total number of citations will yield a higher value if an author has strategically attempted to influence his or her h-index. A researcher who has not strategically inserted self-citations would yield a low ratio as the exponential-like citation patterns would not fit the quadratic h-index pattern.

False positives occur if a researcher with *h* publications managed to attract much attention and achieve close to *h* citations for each publication. This would generate a high ratio. False negatives occur if a researcher cites himself or herself but not in a systematic manner. The citation pattern may then resemble that of non-self-citations and yield a low ratio.

Empirical evaluation

To evaluate the self-citation test, a bibliometric dataset of 105,026 researchers worldwide based on Scopus data including self-citation information (Ioannidis, Baas, Klavans and Boyack 2019) was used. This dataset comprises highly cited researchers with low ratios of self-citations (approximately 0%) up to 93% self-citations. The researcher with 93% self-citations yields a self-citation test value of 0.47, while the maximum self-citation test value observed in this dataset is 0.61 for a researcher with 17% self-citations. The lowest test values are close to 0 (actually 0.00012).

The simple self-citation test values correlates moderately with both the ratio of self-citations ($r(105026) = 0.349$, $p < .001$, 95% CI [0.344, 0.354]) and the ratio of h-index without self-citations and the h-index based on all citations ($r(105026) = 0.294$, $p < .001$, 95% CI [0.300, 0.289]).

A table of self-citation test intervals to help with interpreting the self-citation test values was computed (see Table 1). The table lists the characteristics for different self-citation test value intervals, the median, max and min self-citation rates for the given range, the false negative and false positive rates, the proportion of researchers within the given range and a textual interpretation. The false negative rate is calculated using the number of records that are closer to the max than the median, and the false positive rate is calculated using the proportion of records in the range that are closer to the minimum than the median. The false negative rate represents exceptionally high self-citation rates that are associated with an underestimated self-citation ratio, and the false positive rate represents exceptionally low-self-citation rates that are associated with an overestimated self-citation ratio. Note that the false positive and false negative values are rough estimates as there are no absolute ground truths.

self-citation test	median %	max %	min %	false negatives	false positives	portion	Interpretation
> 0.45	21.31%	93.83%	4.94%	11.63%	13.95%	0.04%	High risk
0.40 - 0.44	17.82%	74.88%	1.07%	2.78%	12.39%	0.45%	
0.35 - 0.39	15.77%	82.18%	0.00%	0.65%	10.53%	4.27%	
0.30 - 0.34	14.75%	79.03%	0.06%	0.44%	9.97%	18.96%	Moderate risk
0.25 - 0.29	13.60%	88.94%	0.00%	0.23%	9.52%	34.06%	
0.20 - 0.24	11.65%	91.40%	0.00%	0.19%	11.64%	24.61%	
0.15 - 0.19	8.88%	62.74%	0.00%	1.16%	15.22%	10.76%	Low risk
0.10 - 0.14	6.01%	66.29%	0.00%	0.13%	19.64%	4.54%	
0.05 - 0.09	3.54%	27.20%	0.00%	1.63%	23.46%	1.81%	

0.00 - 0.04	1.03%	16.57%	0.00%	6.13%	29.37%	0.51%

Table 1. Interpreting the results of the self-citation test intervals and the corresponding median, max and min self-citation ratio, false negatives, false positives, portion of the population, and subjective interpretation.

Approximately half of the researchers yield a self-citation test value of 0.20 or lower, while those with a self-citation test value of 0.35 of higher represent about 5% of the researchers. Such cases should probably be flagged for further exploration. Next, the false negative rate is low for low self-citation test values. The chance of missing an exceptionally high self-citation record is therefore lower with low self-citation test values. Similarly, the chance of accidentally classify a record as having self-citation is higher for high self-citation ratios. One should thus not solely use the results of the self-citation test to conclude that a researcher has many strategic self-citations.

Some discrepancies may occur when applying these limits to Google scholar metrics as the dataset are extracted from Scopus records. However, such discrepancies may be limited since the test value is a ratio.

Conclusions

A simple test for strategic self-citations without self-citation information was proposed. A test ratio is computed as the ratio of the squared h-index over the total citation count. Empirical data show that the test is successful in identifying high levels of self-citations and that test values of 0.35 or more are contenders for further analysis. The test also produces false positives and the test must thus be used with caution. Items flagged as high need to be studied with actual measurements of self-citations using appropriate bibliometric tools.

References

Aksnes, D. W. (2003). A macro study of self-citation. Scientometrics, 56(2), 235-246.Bornmann, L., & Daniel, H. D. (2005). Does the h-index for ranking of scientists really work? Scientometrics, 65(3), 391–392.

Bornmann, L., & Daniel, H. D. (2005). Does the h-index for ranking of scientists really work? Scientometrics, 65(3), 391–392.

Falagas, M. E., & Kavvadia, P. (2006). “Eigenlob”: self-citation in biomedical journals. The FASEB Journal, 20(8), 1039-1042.

García-Pérez, M. A. (2010). Accuracy and completeness of publication and citation records in the Web of Science, PsycINFO, and Google Scholar: A case study for the computation of h indices in Psychology. Journal of the American society for information science and technology, 61(10), 2070-2085.

Gehanno, J. F., Rollin, L., & Darmoni, S. (2013). Is the coverage of Google Scholar enough to be used alone for systematic reviews. BMC medical informatics and decision making, 13(1), 7.

Gianoli, E., & Molina-Montenegro, M. A. (2009). Insights into the relationship between the h-index and self-citations. Journal of the American Society for Information Science and Technology, 60(6), 1283-1285.

Hartley, J. (2012). To cite or not to cite: author self-citations and the impact factor. Scientometrics, 92(2), 313-317.

Harzing, A. W. (2010). The publish or perish book. Tarma Software Research Pty Limited.

Haugen, K. K., & Sandnes, F. E. (2016). The new Norwegian incentive system for publication: from bad to worse. Scientometrics, 109(2), 1299-1306.

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. Proceedings of the National Academy of Sciences of the United States of America, 102(46), 16569–16572.

Ioannidis, J. P., Baas, J., Klavans, R., & Boyack, K. W. (2019). A standardized citation metrics author database annotated for scientific field. PLoS biology, 17(8), e3000384.

Medoff, M. H. (2006). The efficiency of self-citations in economics. Scientometrics, 69(1), 69-84.

Sandnes, F. E. (2018). Do Norwegian academics who publish more earn higher salaries?. Scientometrics, 115(1), 263-281.

Sandnes, F.E., & Grønli, T.M. (2018). Thirty years of NIK: A bibliometric study of pa-per impact and changes in publication patterns. In Proceedings of Norsk Informatikk-konferanse, NISK Stiftelsen, BIBSYS Open Journal System. https://ojs.bibsys.no/index.php/NIK/article/view/496. Accessed 20 Feb 2020.

Sandnes, F.E., & Brevik, E. (2019). Twenty-five years of NOKOBIT: A bibliometric study of impact. In Proceedings from of the annual NOKOBIT conference. NISK Stiftelsen, BIBSYS Open Journal System, 27(1). https://ojs.bibsys.no/index.php/Nokobit/article/view/661. Accessed 20date Feb 2020.

Van Noorden, R., & Singh, C. D. (2019). Hundreds of extreme self-citing scientists revealed in new database. Nature, 572(7771), 578.