Analysing Author Self-Citations in Computer Science Publications

Tobias Milz[1][0000−0003−3159−7666] and Christin Seifert[2][0000−0002−6776−3868]

[1] University of Passau, 94030 Passau, Germany
tobias.milz@uni-passau.de
[2] University of Twente, PO BOX 217, 7500 AE Enschede, The Netherlands
c.seifert@utwente.nl

Abstract. In scientific papers, citations refer to relevant previous work in order to underline the current line of argumentation, compare to other work and/or avoid repetition in writing. Self-citations, e.g. authors citing own previous work might have the same motivation but have also gained negative attention w.r.t. unjustified improvement of scientific performance indicators. Previous studies on self-citations do not provide a detailed analysis in the domain of computer science. In this work, we analyse the prevalence of self-citations in the DBLP, a digital library for computer science. We find, that approx. 10% of all citations are self-citations, while the rates vary with year after publication and the position of the author in the list as well as with the gender of the lead author. Further, we find that C-ranked venues have the highest incoming self-citation rate, while the outgoing rate is stable across all ranks.

Keywords: Citations · Self-Citations · Analysis · DBLP · CORE.

1 Introduction

Scientific work is iterative, findings and discoveries usually lead to new questions waiting to be answered [7]. Consequently, scientific publications, which are the communication of the scientific process and its findings are also iterative and built upon each other. This means, that self-citations, e.g. authors citing their own previous work in publications are an inherent property of the scientific process [10, 11]. They, however, also gained a negative connotation due to their abuse regarding performance indicators of scientists [20, 6, 2, 10, 11]. As a result, scores for detecting likely unfair self-citations have been developed e.g. [3]. In this paper, we characterise the prevalence of self-citations in the domain of computer science according to different factors such as publication year, author gender, conference/journal rank and author positions. We base our analysis on the DBLP computer science bibliography [17], a major bibliographic reference for computer science papers containing more than 3 million publication records as of October 2017. Additionally, we use the conference ranking provided by the Computing

Research and Education Association of Australasia (CORE rankings)[3] to enrich the DBLP data.

Previous work identified different reasons for self-citations (for a complete list see [5]). For instance, in scientific fields with incremental contributions, (own) previous work is cited to avoid duplicate material. Also, in small research areas with a limited number of publication outlets and researchers, citing one's own work is highly prevalent. Thus, we want to emphasise that we provide a quantitative analysis of self-citations, but we do not make an assessment whether a self-citation is justified or not.

The next section outlines some already available statistics, which are mostly available for domains other than computer science. Those statistics are impractical to compare, because of different quantifications. We define our notions in section 3 and continue with the description of our approach followed by the experiments, which are ought to answer questions similar to "how does factor X influence self-citation rates?"

2 Related Work

Different studies have analysed *self-citation rates and their influencing factors.* The most comprehensive study w.r.t to time coverage comprised approx. 1.8 million JSTOR papers from the period of 1779 to 2011 and found an average self-citation rate of 23% (counting outgoing author self-citations) [12]. The only study explicitly covering computer science found a rate of 24% of incoming author self-citations in Norway [1]. Self-citation rates have been found to decrease over time (e.g. [4]), and vary across countries [21, 22] and research fields [21, 22, 8, 12]. Also, collaboration was identified as an influencing factor (e.g. [18, 15, 9]), albeit it seems only important whether papers are single-author or multi-author papers and not how many authors collaborated [9]. While age does not seem to have an influence on self-citation rates [4] authors' gender has a clear impact. According to [8, 12] men cite themselves more often. Ghiasis et al [8] found a gender gap in self-citation behaviour of approx. 35% (with variance across scientific fields and time). The quality of publication venue has also been identified as influencing factor in a small-scale study (643 paper) in the domain of ecology: self-citation rate increased with the impact factor of journals [15]. Apart from the average citation rate of 24% in Norway [1] there is no analysis available for the domain of computer science. In this paper, we analyse the computer science bibliography to this end and compare the findings with results from previous studies in different domains.

3 Problem Statement

Following previous work [5, 9] we distinguish between author and paper self-citations as follows: Let s and t be papers with s citing t and let $A(s)$ and $A(t)$

[3] www.core.edu.au, accessed 2018-03-02

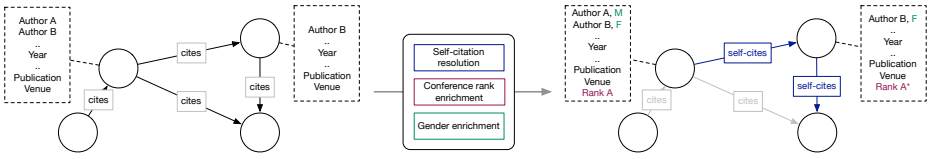

Fig. 1. Overview of the graph structures before and after self-citation resolution and enrichment. Changes are colored.

be the respective sets of authors. A *paper self-citation scp* is any citation for which the author sets of both papers overlap, i.e., A(s) $\bigcap A(t) \neq \emptyset$.

Let s and t be papers with s citing t and let a be an arbitrary author of paper s. A citation is an *author self-citation sca* if author a is also in the author list of paper t, i.e., $a \in A(s) \bigcap A(t)$. Author self-citations are also termed authorship self-citations [12] or self-citations at micro level [9]. We further distinguish between incoming and outgoing self-citations, also referred to as synchronous and asynchronous self-citations [14]. We define sc^{in} as the number of self-citations a paper receives and sc^{out} as the number of self-citations in its reference list. Further, we define self-citation rates as the number of self-citations normalized by the number of citations. Thus, $sc_r^{in} = \frac{sc^{in}}{in_G}$ and $sc_r^{out} = \frac{sc^{out}}{out_G}$, with in_G being the in-degree and out_G being the out-degree of the respective node in the citation graph G. Note, that both definitions can be applied to author self-citations *sca* in a similar fashion.

The goal of this paper is to analyse the occurrences of self-citations with respect to different characteristics of papers (publication year, number of authors, conference/journal rank) and authors (gender, position in author list).

4 Approach

For characterizing self-citations, we use the DBLP citation graph, identify self-citations, enrich the papers' metadata with ranking information from the CORE database and assign a gender attribute to the authors.

An overview of the procedure is depicted in figure 1. On the left, the structure of the DBLP citation graph is shown. The graph $G = (V, E)$ consists of nodes $v \in V$ representing papers, and directed edges $e \in E$ representing the "cite"-relationship. An edge (u, v) means, that paper u cites paper v. For all nodes, the following metadata is available: paper title, author list, publication venue and year[4]. The graph is directed and in general acyclic, however, the latter property can be violated, as some accepted publications might be referenced before they are published. After the identification of self-citations and the enrichment of conference/journal ranks we obtain a graph $G' = (V, E')$ with $E' \subset E$ and $e' \in E'$ representing the "self-cite"-relation. Additionally, the metadata for each node with conference/journal information either contains the corresponding rank

[4] Additional metadata is available, but not relevant for this work

or "unknown" if the publication venue cannot be found in the CORE database (cf. figure 1, right). Importing this graph into a graph-based database such as Neo4j allows easy extraction of both, paper self-citations and author self-citations according to the definitions in section 3.

4.1 Author name disambiguation

A core problem of citation analysis is to identify the person an author's name is referring to. Name homography, i.e., two equal strings referring to different persons, leads to mixed citation errors, while name variability, i.e. two different strings referring to the same person, leads to split citation errors [16]. The DBLP database uses a semiautomatic approach including feedback collection from the scientific community to author name disambiguation [19]. While 100% accuracy cannot be ensured, especially since > 1000 publications are added each day [19], we consider the remaining errors to be insignificant w.r.t. aggregated statistics, however, this assumption remains to be proven.

4.2 Rank enrichment

Most publications in the DBLP database contain the name of their publishing venue (conference/journal), which is used to determine the rank of the publication, based on the year it was published and the rating the corresponding venue had at that time. In order to match the correct venue from the CORE database, a string comparison method is used that takes abbreviations and small string differences into account. We found that in almost all non-trivial cases the DBLP database uses very similar, but smaller sub-strings of the actual venue name taken from the CORE database. As it is very important that these non-trivial matched do not add any false positives, a match is found only if each word of the sub-string is present and they account for 70% of the words of the actual name. Out of all 3,079,007 papers, 506,699 did not include any information about their publishing venue. From the remaining publications, a CORE conference rank was assigned to 437,613 (17.01%) (including the rank "Unranked").

4.3 Gender enrichment

In order to determine the gender of an author, we match the author's first name (given name) to country-specific name lists following the approach from [13]. First, all authors were matched to the list from the US Census[5], which contains 1219 male names and 4275 female names. Ambiguous names (names that appear in both lists) were labelled as unisex unless they are 10 times more frequently used for one gender. If an author's name is not present in the US census list, then it is matched with a comprehensive list of international names taken from Wikipedia and two popular baby name databases[6]. Combining both sets we

[5] www2.census.gov/topics/genealogy/1990surnames/, accessed 2018-04-04
[6] www.rrq.gouv.qc.ca/en, www.babycenter.com/baby-names, accessed 2018-04-04

Table 1. Data set after enrichment. Papers - P, Authors - A, binary gender - G (M/F)

# P	# A	# P w Rank	# A w G(M/F)	Time Period
3,079,007 (100%)	1,766,540 (100%)	440,356 (14.30%)	954,705 (54.04%)	1936 – 2018

gathered 223,428 international given names and their associated gender, including the unisex label for ambiguous names. After our enrichment we identified 266,584 (15.09%) female authors, 688,121 (38.95%) male authors and 287,784 (16.29%) unisex authors in our dataset. 524,051 (29.67%) author names could not be found in any list and their gender remained unknown. 154,497 (29.48%) of these names contained abbreviated first names or only initials. In total, 54% of all author names and 59% of non-abbreviated names could be assigned a binary gender. This is comparable to other work, where 56% of all authors could be assigned a binary gender [12]. Table 1 summarises the data set after enrichment.

5 Experiments

The leading questions for the experiments were:

Q-C: How does the number of self-citations relate to the total number of citations (C)?
Q-Y: How do self-citations depend on publication year (Y), and age (the year after publication)?
Q-R: How are self-citations influenced by venue rank (R)?
Q-A: How does the number of authors (A) relate to self-citations?
Q-P: Do first or last authors have more self-citations , i.e., what is the influence of author position (P)?
Q-G: Do self-citations differ across genders (G)?

We expect the number of self-citations to increase when the number of citations increases (*Q-C*), but the self-citation rate sc_r to be decreasing with the number of citations as found in other studies, e.g. [1]. For question *Q-Y* we expect an increase of the self-citation rates over time for two reasons: First, with the increasing complexity of the scientific discoveries research becomes more incremental. Second, the academic culture has changed over years and emphasizes citation counts and performance metrics over publications leading to phenomena like least-publishable units and publish-and-perish. With respect to rank (*Q-R*) we do not expect any differences in self-citations rates. On the one hand, lower ranked conferences are used to publish minor improvements and adaptations and therefore boost self-citations, on the other hand, higher ranked conferences publish larger, more comprehensive bodies of scientific work, which might also result in more self-citations. We expect multi-author papers to have a higher self-citation rate than single-author papers, while the number of authors does

not have much influence as has been found in other corpora [9]. Further, we expect last authors to have more self-citations than first or middle authors (Q-P) because in computer science last authors are usually supervisors, they are more senior and have published more papers themselves. For the gender-aspect of self-citations (Q-Y) we expect a higher self-citation rate for male authors than for females as indicated by previous work for other research fields [8, 12].

6 Results

In this section we show the results of the analysis organised by the previously introduced experiment questions.

6.1 Self-citations by total number of citations (Q-C)

In the corpus, 18.67% of all papers do not have any outgoing citations, 51.56% do not have any outgoing and 60.28% do not have any incoming self-citations. Figure 2 and 3 provide an overview of the relation between citations and self-citations. Both, the number of outgoing and incoming self-citations increases with the number of citations, while the outgoing ones show a steeper increase (see figure 2). On average, the rate of outgoing and incoming self-citations for papers with at most 50 outgoing citations is 14.16% and 14.75% respectively (9.44% and 11.40% for all papers), while both rates are not constant over the number of citations (cmp. figure 3). These numbers are considerably less than the numbers reported in a previous study (24% incoming), which included only 283 computer science papers from Norway [1]. Similar research on other domains has shown, that the number of outgoing self-citations varies significantly depending on the research field (e.g. 6.3% for law and 12.3% for math) [12].

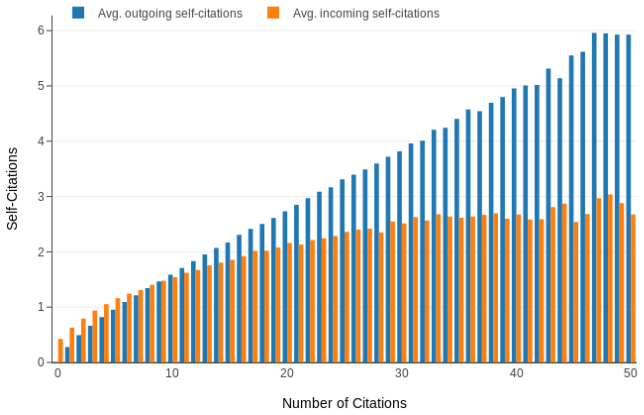

Fig. 2. Relation of number of citations and self-citation rate.

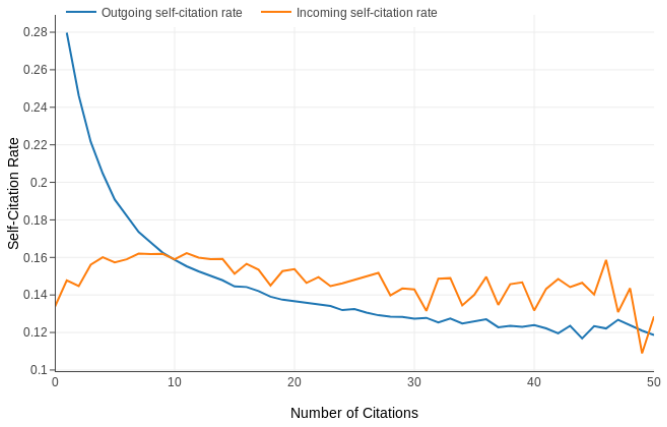

Fig. 3. Relation of self-citations and self-citation rate.

6.2 Self-citations by year (Q-Y)

From the 1960s onwards the rate of outgoing self-citations remains relatively constant over time. This period covers most of the publications in the database (cmp. Figure 4). On the contrary, the rate of incoming self-citations started to increase in approx. 2000. As figure 5 shows, papers receive most citations up to 10 years after publication, and the highest rates of self-citations in the first years, which is in line with previous observations in other domains [1, 9]. Surprisingly, we found papers that were cited before they were published. We investigated those papers in more detail and found that this is due to citations made to books or collections containing older publications. The DBLP database seems to misidentify very few of these publications as recent works, which however only affects 0.7% of all citations.

6.3 Self-citations by rank (Q-R) and author gender (Q-G)

The self-citation rates with respect to rank are shown in figure 6, "Other" aggregates the ranks *New*, *National:Spain*, *L* and *Regional*. The outgoing citation rate is relatively stable across all ranks, but the incoming self-citation rate varies, being 11.69% for A* conferences and 26.35% for C conferences. Figure 7 illustrates the outgoing and incoming self-citation rates with a first author of a specific gender. Publications with a male first author tend to have significantly more outgoing self-citations than papers with female lead authors. However, publications with a female first author receive more incoming self-citations. This observation is in line with findings in domains [9].

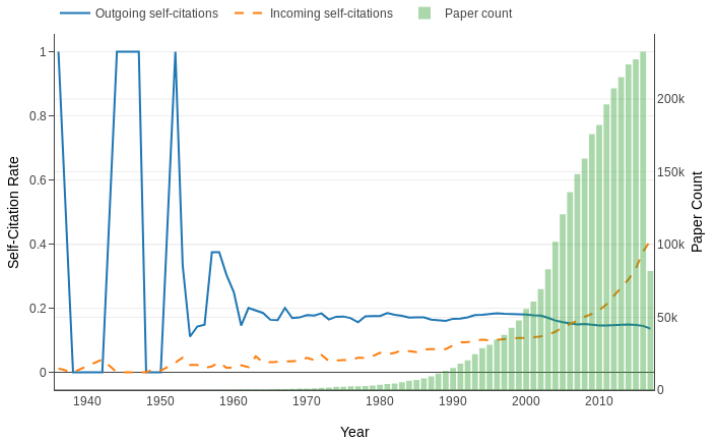

Fig. 4. Relation of number of self-citations and absolute time.

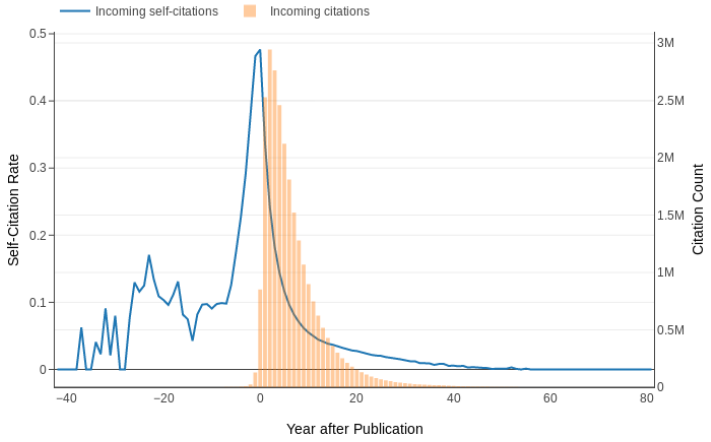

Fig. 5. Relation of number of self-citations and year after publication.

6.4 Self-citations and authors (Q-A and Q-P)

Figure 8 depicts the relation between the number of authors and the self-citation rates, showing papers with at most 10 authors (99.5% of all papers). As a tendency, the outgoing and incoming self-citation rate increases with the number of authors (denotes n_a). The Pearson correlation coefficients in regards to self-citations are $\rho(n_a, sc_r^{out}) = 0.89$, $\rho(n_a, sc_r^{in}) = 0.81$ and $\rho(n_a, c_r^{out}) = 0.50$, $\rho(n_a, c_r^{in}) = 0.48$ regarding the total number of citations (normalized by the number of papers). The general tendency is in line with previous studies in other domains [9], but we additionally observe a correlation between number of authors and (self-) citation rates.

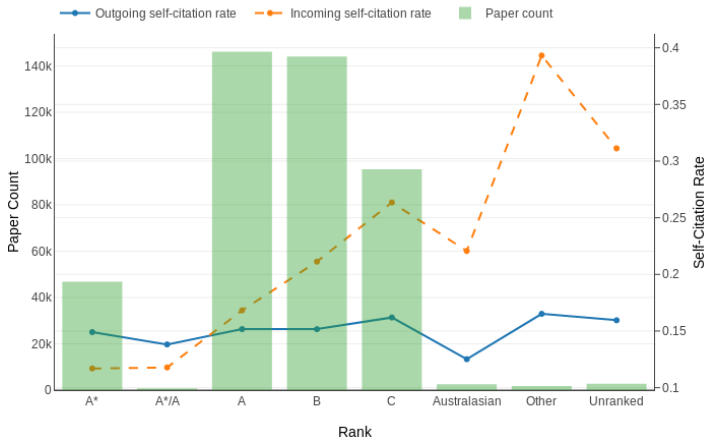

Fig. 6. Relation of venue rank and self-citation rate.

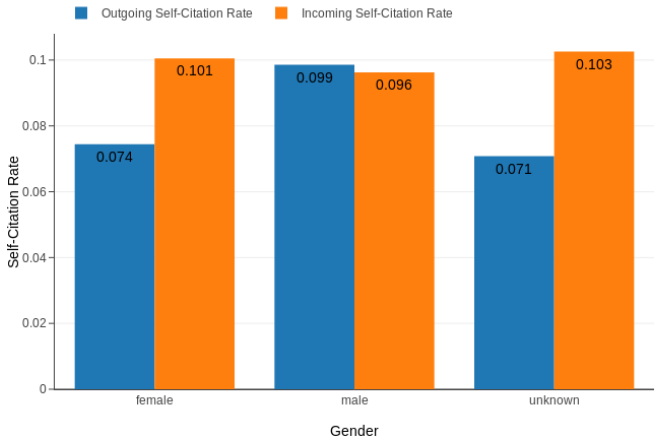

Fig. 7. Relation of gender and self-citation rate.

Figure 9 depicts the relation of author position and self-citation rate for papers with at least one outgoing self-citation. We depict author self-citations w.r.t. to the total number of outgoing citations (blue bars) and the total number of outgoing self-citations (orange bars) in the corresponding subset of the data. These subsets contain all single-author papers (condition "Single") and all papers with at least two authors (conditions "First" and "Last"). From all citations in single-author papers, 14.8% are self-citations. From all citations in multi-author papers, 7.7% are self-citations that include the first author and 9.1% that include the last author. With this, the first author is part of 50% of all self-citations in multi-author papers, while the last author is involved in 60% of all self-citations.

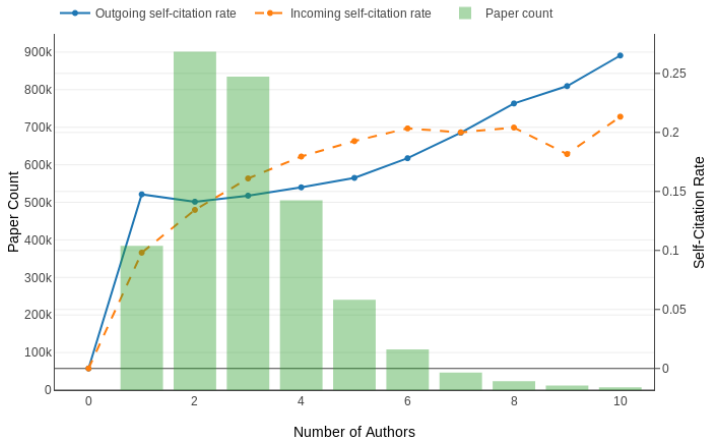

Fig. 8. Relation of number of authors and self-citation rate.

These values do not sum up to 100% as both authors can be part of the same self-citation. This means single authors do more self-citations than first or last authors, but last authors have more self-citations than first authors.

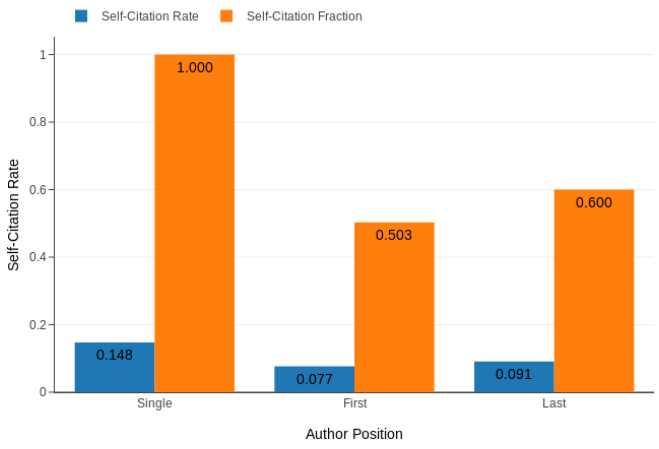

Fig. 9. Relation of author position and self-citation rate. Normalized by the number of citations in single and multi-author papers (blue) and by the number of self-citations (orange).

7 Summary & Future Work

In this paper, we analysed the prevalence of self-citations in the domain of computer science based on the DBLP citation graph. We found an average incoming and outgoing self-citation rate of 9.44% and 11.40%, respectively. As in other domains, self-citation rates were highest in the first years after publication. The outgoing self-citation rate is relatively stable across conference rank, but the incoming rate differs across ranks, with C-rated venues having the highest rate. Further, we found, that the more authors, the higher the rates of outgoing and incoming self-citations are, while last authors have higher self-citation rates than first authors for multi-author papers. Papers with male first authors have a higher outgoing self-citation rate, while papers with female first authors have a higher incoming self-citation rate.
In this study, we presented statistical observations without in-depth interpretation. Future work could address the questions that arise from these findings. One might consider investigating the reasoning behind the opposing self-citation rates of male and female lead authors or find arguments for the amount of incoming self-citation of C ranked venues. Furthermore, we considered every attributed (e.g., rank, year) on its own. Further analysis could reveal how attribute combinations influence the (self-) citation rates.

Acknowledgment

We would like to thank Moritz Grünbauer for his preliminary analysis and help with constructing the queries for the graph database.

References

1. Aksnes, D.W.: A macro study of self-citation. Scientometrics **56**(2), 235–246 (Feb 2003)
2. Alonso, S., Cabrerizo, F., Herrera-Viedma, E., Herrera, F.: h-index: A review focused in its variants, computation and standardization for different scientific fields. Journal of Informetrics **3**(4), 273 – 289 (2009)
3. Bartneck, C., Kokkelmans, S.: Detecting h-index manipulation through self-citation analysis. Scientometrics **87**(1), 85–98 (2010)
4. Costas, R., van Leeuwen, T.N., Bordons, M.: Self-citations at the meso and individual levels: effects of different calculation methods. Scientometrics (2010)
5. Ferrara, E., Romero, A.E.: Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the h-index. JASIST **64**(11), 2332–2339 (2013)
6. Fowler, J.H., Aksnes, D.W.: Does self-citation pay? Scientometrics **72**(3), 427–437 (Sep 2007)
7. Gauch Jr, H.G.: Scientific Method in Practice. Cambridge University Press (2002). https://doi.org/10.1017/CBO9780511815034
8. Ghiasi, G., Larivère, V., Sugimoto, C.R.: Gender differences in synchronous and diachronous self-citations. In: Proc. Intl. Conference on Science and Technology Indicaors (2016)

9. Glänzel, W., Debackere, K., Thijs, B., Schubert, A.: A concise review on the role of author self-citations in information science, bibliometrics and science policy. Scientometrics (2006)
10. Hemmat Esfe, M., Wongwises, S., Asadi, A., Karimipour, A., Akbari, M.: Mandatory and self-citation; types, reasons, their benefits and disadvantages. Science and Engineering Ethics (2015)
11. Ioannidis, J.P.: A generalized view of self-citation: Direct, co-author, collaborative, and coercive induced self-citation. Journal of Psychosomatic Research (2015)
12. King, M.M., Bergstrom, C.T., Correll, S.J., Jacquet, J., West, J.D.: Men set their own cites high: Gender and self-citation across fields and over time. Socius **3** (2017)
13. Larivire, V., Ni, C., Gingras, Y., Cronin, B., Sugimoto, C.: Bibliometrics: Global gender disparities in science **504** (12 2013)
14. Lawani, S.M.: On the heterogeneity and classification of author self-citations. Journal of the American Society for Information Science **33**(5), 281–284 (1982)
15. Leblond, M.: Author self-citations in the field of ecology. Scientometrics (2012)
16. Lee, D., On, B.W., Kang, J., Park, S.: Effective and scalable solutions for mixed and split citation problems in digital libraries. In: Proc. Intl. WS. on Information Quality in Information Systems. pp. 69–76. ACM, New York, NY, USA (2005)
17. Ley, M.: The dblp computer science bibliography: Evolution, research issues, perspectives. In: String Processing and Information Retrieval. pp. 1–10. Springer (2002)
18. Medoff, H.M.: The efficiency of self-citations in economics. Scientometrics (2006)
19. Müller, M.C., Reitz, F., Roy, N.: Data sets for author name disambiguation: an empirical analysis and a new resource. Scientometrics (2017). https://doi.org/10.1007/s11192-017-2363-5
20. Purvis, A.: The h index: playing the numbers game. Trends in Ecology & Evolution **21**(8), 422 (2006). https://doi.org/https://doi.org/10.1016/j.tree.2006.05.014
21. Schubert, A., Glänzel, W., Thijs, B.: The weight of author self-citations. a fractional approach to self-citation counting. Scientometrics (2006)
22. Thijs, B., Glänzel, W.: The influence of author self-citations on bibliometric meso-indicators. The case of European universities. Scientometrics (2006)