Who Cites What in Computer Science? - Analysing Citation Patterns across Conference Rank and Gender

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. Citations are a means to refer to previous, relevant scientific bodies of work. However, little is known about how citations behave with respect to venue reputation. Do A* papers get more often cited by C papers or vice versa? What is the source and sink of a citation in terms of venue reputation? In this work, we investigate this issue by analysing the DBLP database of computer science publications, utilizing rank information from the CORE database. Our analysis shows that authors tend to cite publications from the same or higher ranked venues more often than from lower tier venues. Self-citations, on the contrary, are especially focused on same-tier venues. The gender of the first author does not seem to have any impact on the citations from and to differently ranked mediums.

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

1 Introduction

Citations are a means to refer to previous scientific bodies of work, and are also used to calculate impact factors for journals [**?**,**?**] and performance measures for scientists [**?**] and thus have become a valuable commodity in science. Research has been concerned with finding influencing factors for citations (e.g. [**?**]), and most prominently to identify the influence of self-citations on citations and subsequently on indicators of scientific performance, e.g. [**?**,**?**]. Multi-authored, as well as papers with male first author have been found to have a higher self-citation rate [**?**,**?**,**?**], while self-citation rates generally vary over fields and countries [**?**]. To the best of our knowledge, the only study that investigated the relation of self-citations and the scientific reputation of the publication venue is in the economics domain [**?**]. The authors found that the proportion of self-citations increased with the impact factor of ecology journals.

This paper contributes to the knowledge of citation and self-citation by analysing the domain of computer science. Specifically, we investigate the DBLP computer science bibliography [**?**] w.r.t. ranking of the conferences/journals and gender of the first author.

2 Problem Statement

Citations can either be synchronous (outgoing) or diachronous (incoming) [**?**], the former refers to the number of publications a paper cites and the latter how often a publication gets cited. Analogously, outgoing and incoming self-citations are citations from and to publications of the same author. The self-citation rate is defined as the ratio of the self-citations normalized by the total number of citations and can be calculated for both, incoming and outgoing self-citations. In this paper, we analyse incoming and outgoing citations and self-citation rates with respect to the conference/journal rank. For instance, if paper P cites paper Q, and P was published at an A* conference while paper Q was published at a C conference, the citation counts as an outgoing citation for A* and incoming citation for C.

3 Method

For our analysis, we use the DBLP citation graph [**?**], supplemented with the paper's ranking information and a gender attribute for the authors. The rankings are extracted from the Computing Research and Education Association of Australasia (CORE) database[3] using a rule-based string matching method of the venue name. The focus of this method is to find the most likely match, but without introducing any false-positives in favour of Recall. The publication year of the papers is also considered in order to take rank changes of venues into account. We follow previously suggested methods to determine an author's gender by matching their first name (given name) to country-specific name lists [**?**]. For author identity, we rely on the quality of the DBLP citation graph, which already employs author name disambiguation approaches [**?**]. Out of all 3,079,007 papers in DBLP covering the publication and citation period from 1946–2018, 55.66% (1,744,449) were assigned a binary (female/male) gender based on the first author's inferred gender. A CORE rank was assigned to 14.15% (435,823), while both information could be assigned to 7.86% (242,096) of all papers.

4 Results

The heatmaps in figure 1 show the fraction of outgoing and incoming citations and self-citations for publications from each conference/journal rank. The initial theory is, that publications will more often cite highly ranked papers, as they have more visibility. According to the results, this hypothesis seems to hold true. For example, 93.6% of all outgoing citations from publications with a B rating, cite other publications with the same or higher rating (top left). Furthermore, A, B and C ranked papers receive more than half of all their incoming citations from publications of the same rank (top right). For self-citations, this effect is even more prominent especially for the categories C and Australasian, which

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

Fig. 1. Ratio of citations (top) and self-citations (bottom) from venues with specific rank. Rows indicate the source and columns the target of citations. Left: normalized by the total number of outgoing citations per rank; right: normalized by total incoming citations per rank.

Table 1. Comparison of citations by gender (M - male, F- female, X - unisex, ? - unknown) and conference/journal rank

			Conference/Journal Rank							$\sum$
		Papers	A*	A*/A	A	B	C	Austr.	Other	Citations
in	M	1,334,187	0.138	0.003	0.387	0.319	0.143	0.006	0.003	1,957,108
	F	410,262	0.126	0.001	0.398	0.325	0.143	0.005	0.003	417,655
	X	609,101	0.134	0.002	0.371	0.343	0.144	0.005	0.003	748,836
	?	725,453	0.117	0.001	0.355	0.346	0.174	0.005	0.003	676,809
out	M	1,334,187	0.234	0.008	0.427	0.239	0.087	0.004	0.002	1,355,908
	F	410,262	0.226	0.003	0.433	0.248	0.084	0.004	0.001	430,910
	X	609,101	0.231	0.004	0.432	0.244	0.084	0.004	0.001	733,721
	?	725,453	0.237	0.002	0.427	0.235	0.094	0.004	0.001	761,150

have much lower citation rates (35.1% and 9.3% respectively) than self-citation rates (61.2% and 33.2% respectively) towards same-tier publications (bottom). In other words, authors prefer to cite higher ranked publications, but self-citations are more commonly towards publications of the same conference/journal rank. Please note, that although a difference is observable in values for categories Australasia and Other, we abstain from an interpretation, since both categories only contain 4318 (0.9%) of the papers with an assigned rank.

Table 1 shows the statistics w.r.t. venue rank and gender of the first author. For example, out of all 1,957,108 outgoing citations towards papers with a male lead author, 13.8% are cited in publications from conferences/journals with an A* rating. This citation-rate indicates how citations from/to differently rated mediums are affected by the first author's gender of the cited/citing paper. The results show that despite the lower number of papers with female leading authors (410,262 papers with female and 1,334,187 with male lead author), the distribution of the incoming and outgoing citation rate stays the same. In other words, the gender of the leading author has no significant effect on the citations of papers when considering their identified rating.

Further studies are required to shed light on the reason for the difference in citation/self-citations behaviour w.r.t. rank. An interesting future question would be, whether a homophily property in citation behaviour can be observed, i.e., whether a specific gender tends to cite authors of the same gender

References