Paper Publication in Journals: Analyzing the Impact on Citations Rate

Roye Katzav and Dudi Biton

August 2023

Abstract

While numerous studies have delved into various determinants of citation counts, journal publication and its influence on citation rates have been largely unaddressed. This paper investigates the correlation between journal publication and citation rates. Our findings indicate an initial increase in citation rates post-publication, which tends to diminish over time and varies across different academic fields. These insights could help researchers understand the potential benefits and limitations of journal publication for increasing their work's visibility and impact.

1 Introduction

Citation counts serve as a critical indicator of the reach, impact, and perceived quality of academic research. They reflect the extent to which a particular piece of scholarly work has permeated its field, influencing the thought processes and subsequent work of other researchers. As such, a higher citation count often signifies a study's relevance, offering a tangible metric that measures a paper's contribution to its respective field.

Given the value and importance attributed to citation counts in academic circles, understanding the determinants of these counts, and how researchers can potentially increase their paper's citation rates, becomes a crucial task. This understanding directly aids in strategic academic publishing, potentially enhancing the visibility and impact of researchers' work.

The existing literature provides extensive insight into various factors that influence citation counts. Numerous variables have been investigated, such as the number of authors, suggesting that a collaborative effort may enhance visibility and reach [1]; the length of the manuscript, indicating that a more comprehensive exploration of a topic may attract more citations [2]; the accessibility and usage of articles, implying that more readily available and frequently used articles often receive higher citations [3]; and the nature of the paper's title, demonstrating that certain titling strategies may capture more attention and thus garner more citations [4]. Each of these studies contributes a valuable piece to the citation count puzzle.

However, one aspect remains conspicuously underinvestigated: the potential influence of publishing a paper in a journal on its citation count.

In this paper, we focus on understanding whether publishing a paper in a journal could have a discernible effect on its citation count. In other words, our guiding research question is: Does publishing a paper in a journal influence its citation rate?

Addressing this question could provide a new perspective in the ongoing discussion about the effect of the citation count. Moreover, it could offer practical insights for researchers regarding their publishing plans.

Through our analysis, we found an initial increase in citation rates post-publication, which tends to diminish over time. We also noted that this effect varies across different academic fields. These findings provide a nuanced view of the impact of journal publication on citation rates.

The rest of the paper is structured as follows: section 2 reviews related work in the field, section 3 describes our research methodology, section 4 describes the results, section 5 is the discussion, and section 6 is the conclusions and future work.

2 Related Work

In recent years, a multitude of factors have been examined to understand their impact on the citation count of papers.

Tahamtan et al. [1] present a comprehensive review of the literature, investigating the factors affecting the number of citations a paper receives. Their review offers a broad perspective, examining the multifaceted nature of the factors influencing citation counts.

Looking specifically into the field of ecology, Fox et al. [2] found that manuscript length, the number of authors, and the number of references cited in a paper increase the number of citations. Their work indicates that these factors may also influence citation counts in other scientific fields.

Meanwhile, the study by Kurtz et al. [3] explores the effect of the usage and accessibility of papers on their citation count. They argue that ease of access plays a significant role in a paper's citation frequency, providing an interesting avenue for future research on open-access publishing.

Lastly, Jamali and Nikzad [4] investigate the relationship between the type of a paper's title and its number of downloads and citations. This study suggests that even seemingly minor elements, such as the title, can have an impact on the paper's visibility and, ultimately, its citation count.

These studies collectively indicate that a wide variety of factors, ranging from accessibility and title type to manuscript length and the number of authors, can influence the citation count of papers. Further investigation into these variables may contribute to a more comprehensive understanding of the dynamics behind citation counts.

3 Methodology

In this section, we will present the experimental setup along with the data collection and preprocessing processes that we conducted for this work. Also, we present the scenarios which used to analyze the data.

3.1 Experimental Setup

The code was executed on a GPU cluster consisting of a few physical machines of RTX 2080 Ti, six cores, and 32 GB RAM.

3.2 Data Collection

During the initial stage of data collection, we downloaded the arXiv dataset [1] from Kaggle. To augment our research data, we performed a cross-reference of DOIs (Digital Object Identifiers) between the arXiv dataset and the OpenAlex dataset [2].

3.3 Data pre-processing

The data pre-processing involved several essential steps as follows:

Step one: We began by converting the JSON files of the arXiv dataset into a Pandas DataFrame object, this dataset contained 2,292,057 papers.

Step two: To ensure data quality, we removed all entries with missing DOI values (NaN) from the arXiv dataset. This step was necessary as the data in arXiv was incomplete, and we planned to complement it by extracting further information from the OpenAlex dataset through DOI comparisons.

Step three: During the DOI comparison process, we utilized API requests to match DOIs in OpenAlex with those in the arXiv dataset. However, some links in OpenAlex returned HTTP errors with code 403. We carefully identified these erroneous entries and eliminated them from our data.

Step four: Employing API calls in batches of 50 at a time, we compared DOIs between the arXiv and OpenAlex datasets. From OpenAlex, we extracted valuable information such as the number of citations by year, the publication date of a paper in the journal, the total number of citations, the category to which the paper belongs, papers cited by the specific paper, and the OpenAlex ID. Every 10,000 iterations, we saved all extracted information to separate CSV files (in order to avoid saving one large file that could cause a RAM error), and finally, merged all CSV files into one comprehensive dataset.

Step five: For consistency, we treated the initial publication date in arXiv as the actual publication date of a paper. The date was initially represented in the format "Fri, 30 Jul 2021 12:54:46 UTC," but we converted all dates to the format "Year-Month-Day," which aligns with the date format in OpenAlex.

Step six: At this stage, we expanded the data to include information on all the papers that a particular paper cites. In addition to the other columns that contain information about each paper, each line of the dataset now contained two essential columns: "src_id" for the citing paper and "id_referenced_works" for the cited paper.

Step seven: In order to facilitate further analysis, we created a dictionary where the keys represented paper IDs, and the corresponding values indicated the date the paper was cited according to its initial publication date in the first version of arXiv. Leveraging that dictionary, we enriched the original CSV with additional information for each paper.

Step eight: To ensure data consistency and accuracy, we resolved potential conflicts arising from the comparison between the two datasets. Specifically, we removed rows where the publication date in the journal was earlier than the initial publication date in arXiv.

As a result of these comprehensive data processing steps, our dataset now contains information on 842,082 distinct papers.

3.4 Data Analyzing Scenarios

We conduct data analysis using two different scenarios designed to reveal the impact of publishing a paper in a journal on citation rates from different perspectives.

3.4.1 Scenario One: Pre and Post-Journal Publication Citation Rates

For our initial analysis, we considered a subset of papers that had a specific time window both prior to and post their journal publication and examined the pre and post-citations. The time window was taken according to the accurate dates. This selection criterion was vital to ensure a balanced comparative timescale, thus avoiding the bias that could be introduced by an excessive span of post-publication years amassing citations. In addition, we only examined the top five categories of papers to make our analysis more refined, focused and relevant.

[1] https://www.kaggle.com/datasets/Cornell-University/arxiv

[2] https://openalex.org/

3.4.2 Scenario Two: Citation Rate Over the Years

While the first scenario provided insights into citation rate fluctuations before and after journal publication, it lacked a temporal dimension that could elucidate the effect of specific years on these rates. Thus, in our second scenario, we conducted a year-by-year analysis of the average citation rate from the year of first publication to 2023. First, We examined the overall effect without taking into account any specific categories. Afterward, just like in the first scenario, we focused on the top five categories based on the limitations of the scenario.

Both scenarios are integral to understanding the multifaceted dynamics of citation rates in relation to journal publication. They provide distinct yet complementary perspectives, with the first scenario enabling a direct before-and-after comparison, and the second bringing in a longitudinal dimension, charting the citation rate trends over time.

4 Results

The presented results were obtained by analyzing the two scenarios.

4.1 Pre and Post-Journal Publication Citation Rates

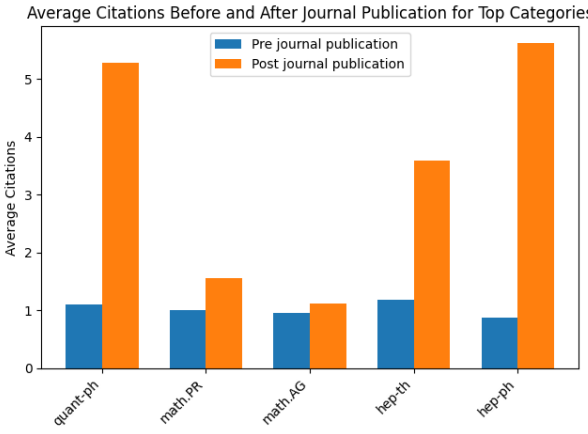

Figure 1: Pre and Post-Journal Publication Average Citation Rates

Figure 1 and Figure 2 represent the results for the first scenario. These figures examine a subset of papers that had a time window of 2.5 to 3 years both prior to and post their journal publication (other time windows can be found in the results file in our code [3]). The X-axis contains the top five different categories according to the scenario's limitations. The Y-axis in figure 1 and figure 2 represents the average and median citations count respectively. Note,

[3] https://github.com/dudi709/big-data-journal-citations/blob/main/Results_and_Visualizations.ipynb

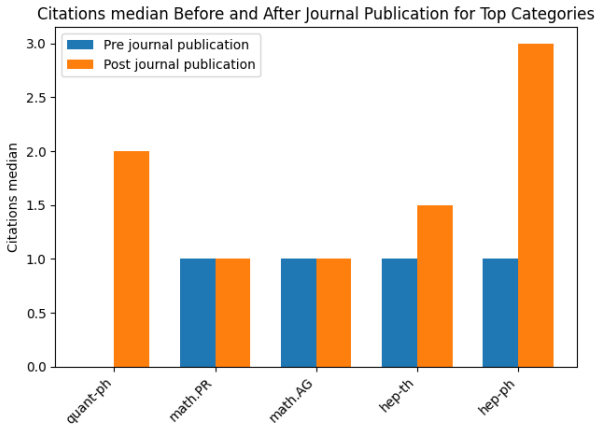

Figure 2: Pre and Post-Journal Publication Median Citation Rates

median citations count were analyzed to prevent a case in which only average results with possible extreme outliers are presented. The blue bar represents the pre-journal publication's citations count and the orange bar represents the post-journal publication's citations count. When examining the average and median citations count, we note that in most of the categories, there are more post-journal publication citations than pre. However, when examining the median citation count (figure 2) separately, we can see that math.PR (mathematics - probability) and math.AG (mathematics - algebraic geometry) categories have the same median citation count, which could overall indicate on similar citation count between pre and post-journal publications, in these categories.

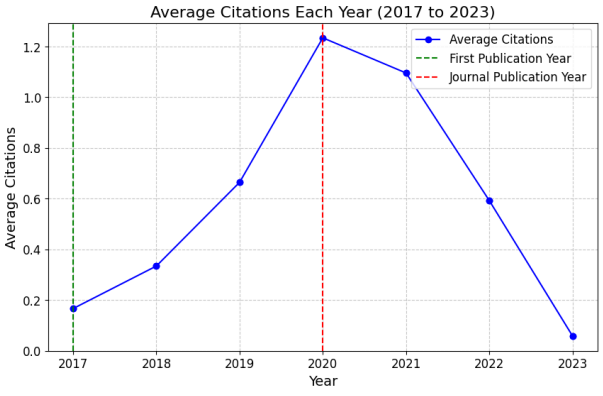

Figure 3: Overall Citation Rate Over the Years

4.2 Citation Rate Over the Years

Figure 3 and Figure 4 represent the results for the second scenario. These figures examine a subset of papers that were first published on arXiv in 2017 and subsequently published in a journal in 2020 (other analyses of arXiv and journal publication years interval can be found in the results file in our code[3]) Figure 3 represents the results for the second scenario

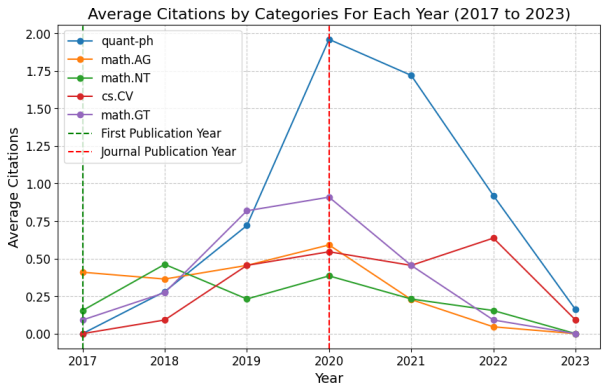

Figure 4: Citation Rate Over the Years for Top 5 Categories

without categories, while Figure 4 represents the results for the top five categories according to the scenario's limitations. The X-axis represents the years, from 2017 to 2023. Note, the information regarding the year 2023 is partial, which might be reflected in a smaller citation rate. The Y-axis represents the average citation rate. There are two dashed vertical lines for the arXiv (green) and journal (red) publication years. The rest of the lines represent the different categories (figure 4) or the overall results over the years (figure 3). From figure 3 we can see that overall the citation rate extremely increased in the journal publication year, while after two years the citation rate decreased to a similar value compared to the last year before published in the journal. From figure 4, which focuses on the more specific analysis (of categories), we can see that in most of the categories, the citation rate increased in the journal publication year, however, this increase is not significance in most of the categories, compared to the results in figure 3. Moreover, besides the cs.CV (computer vision and pattern recognition) category, there is a decrease in the citation rate in the years after the journal publication.

5 Discussion

Our results reveal intriguing insights into the effects of journal publication on citation rates. In the first scenario, where we analyzed citation rates in a specific time window of 2.5 to 3 years prior to and post to the journal publication year, the trend indicates that post-journal publication citation counts are generally higher. This suggests that journal publication indeed seems to enhance visibility and recognition, leading to a higher citation rate. However, it is important to note the exceptions we found in specific fields like mathematics - probability (math.PR) and algebraic geometry (math.AG), where the median citation count remained consistent pre and post-journal publication. This could suggest that certain fields may be less influenced by journal publications or that the peer communities in these fields source information more evenly from both preprints and journal publications.

The second scenario, which examined the trend in citation rates over the years, from the first publication on arXiv in 2017 to the year 2023, showed an increase in the citation rate in the year of journal publication. Yet, the two years following the publication year showed a decrease in citation rates, almost back to the pre-journal publication level. This suggests an ephemeral "publication effect", where the immediate attention gained from journal publication could wane in subsequent years. Furthermore, the category-specific analysis shows that the publication effect varies among different fields, with some seeing a more enduring boost in citation rates (such as computer vision and pattern recognition, cs.CV), while others experience a drop.

6 Conclusions and Future Work

In conclusion, our study provides nuanced insights into the effect of journal publication on citation rates. Our results demonstrate an initial increase in the citation rates post-publication, which may taper off in subsequent years. However, this effect is not uniform across different academic categories, suggesting that the influence of journal publication is contingent on the specific characteristics and citation habits within each field.

For future work, an extended analysis covering additional categories and a broader timeline could offer a more comprehensive understanding of these trends. Examining underlying factors that contribute to the differences among fields could further enhance our insights into citation habits.

References

[1] I. Tahamtan, A. Safipour Afshar, and K. Ahamdzadeh, "Factors affecting number of citations: a comprehensive review of the literature," *Scientometrics*, vol. 107, pp. 1195–1225, 2016.

[2] C. W. Fox, C. T. Paine, and B. Sauterey, "Citations increase with manuscript length, author number, and references cited in ecology journals," *Ecology and Evolution*, vol. 6, no. 21, pp. 7717–7726, 2016.

[3] M. J. Kurtz, G. Eichhorn, A. Accomazzi, C. Grant, M. Demleitner, E. Henneken, and S. S. Murray, "The effect of use and access on citations," *Information processing & management*, vol. 41, no. 6, pp. 1395–1402, 2005.

[4] H. R. Jamali and M. Nikzad, "Article title type and its relation with the number of downloads and citations," *Scientometrics*, vol. 88, no. 2, pp. 653–661, 2011.