Self-Citation Analysis of Semantic Scholar Dataset

Olga Dorabiala, Marta Wolfshorndl, Yang Zhou

Abstract— **In our project "Citation Analysis," we aim to produce a comprehensive analysis of citation attributes from the Semantic Scholar dataset with a particular focus on self-citation, which occurs when an author cites their own work in a publication. We calculate a standard H-Index for each author in the dataset and compare it to the H-Indexes calculated by other sources, we define and investigate an Adjusted H-Index that takes into account self-citation information and compare its effectiveness in measuring author influence to the standard H-Index, and we examine the self-citation rate among authors with respect to factors such as journal, field of study, and age. We analyze our results and draw conclusions based on the information we extract.**

I. Introduction

Citation metrics play a large role in scientists' careers, as they are often used to measure the impact of their work. A 2019 Nature news feature [1] found that self-citation, as defined by an author referencing their own work in a publication, is prevalent in the literature. Self citation incentivizes self promotion and allows scientists to artificially inflate their H-index [2].

Definition 1: An author is defined as having an H-index of h if h of his or her N total papers has at least h citations each.

We are interested in answering questions such as: 1) Who are the most influential authors based on metrics such as the H-index? 3)Does self-citation influence the H-index rankings in our dataset? 3) What is the self-citation rate per author and per paper? We will perform a self-citation analysis to determine if self-citation rates in the Semantic Scholar dataset match those of the literature, who the worst offenders are in terms of authors and publications, and how many self-citations to other citations there are per paper.

Semantic Scholar [3] is a search tool that uses machine learning to identify connections between papers and help scholars discover and understand new research. The Semantic Scholar dataset contains information for 22 attributes, including authors, number of citations, year, and journal name, for about 70 million publications in JSON format.

In the literature, the issue of self-citation is analyzed through a 2019 Nature news feature, which described how citation metrics in a database of around 100,000 researchers were analyzed, see [1], [4]. The paper determined that the median self-citation rate of researchers in their database was 12.7% and found examples of so-called extreme self-citers, the majority of whose citations come from either themselves, or their co-authors. The paper also described citation farms, where a small group of researchers continuously cite each other to increase their citation statistics. The authors' main focus was on creating a publicly available database and on standardizing the calculations of h-index and self-citation so that science as a whole can become more transparent about citations, given the importance they can play in scientists' careers.

Multiple papers have evaluated the effects of gender on self-citation. In 2017, King, et al analyzed 1.5 million papers in the JSTOR database and found that men were 56% more likely than women to self-cite, and that number rose to 70% more likely in the last two decades [5]. In addition, women were 10% more likely to not cite their previous work at all. However, another citation analysis paper published in 2018 also evaluated self-citation metrics using a database of 1.6 million papers from Authority, a version of PubMed, a database which is primarily focused on biomedical studies [6]. Their findings on gender contradicted the results from King, et al., finding that the gender effect on self-citation largely disappeared when correcting for previous citation counts. They went further and tested a wide variety of parameters, including byline position, affiliation, ethnicity, collaboration size, time lag, subject-matter novelty, reference/citation counts, publication type, language, and venue, and found that gender had the weakest effect of the parameters tested. Their results also found that self-citation rate was about 13%, which agreed with the findings from Ioannis et al 2019 [4]. One of their claims was that self-citation is a hallmark of a productive author with a lot of publications.

In 2017, Flatt, Blasimme, and Vayena [2] argued that

there should be a self-citation index ('s-index') similar to an H-Index that is calculated for each author, and that H-Index should be recalculated to take self-citation into account. They provided many arguments as to why self-citation can be a big problem for science as a whole, including that it incentivizes self promotion and scientists can artificially inflate their H-Index scores with no negative impact. The s-index they proposed was defined as: "A scientist has a self-citation index s equal to the total number of s papers that he or she has published that have at least the same amount of s self-citations." They provided a compelling example case for researchers whose h-index would dramatically decrease if self-citation were taken into account.

II. Methods

Our approach to analyzing the Semantic Scholar dataset was threefold. First, we calculated an H-Index for every author in order to compare it to H-Indexes from other sources for the same authors. This gave us an idea about the completeness and scope of our data. Next, we moved on to examining self-citation by calculating both an Adjusted H-Index that took into account self-citation information and a self-citation rate, which differed slightly from that used by [2]. We defined our Adjusted H-Index and self citation rate as,

Definition 2: An author is defined as having an Adjusted H-index of h if h of his or her N total papers has at least h citations coming from other authors each.

Definition 3: The self-citation rate is defined as

$$\text{Self-Citation Rate} = \frac{\#\text{ of self-citations}}{\#\text{of total citations}} \times 100$$

We were interested in comparing self-citation rate to H-Index to determining if, as proposed by [6], more highly productive researchers were also more likely to self-cite. Our data set did not contain demographic information, such as gender, which was used for study in some aforementioned papers, but we examined the effect of factors such as journal, field of study, and age on self-citation. We compared our results to the findings of the papers mentioned above, with a particular interest in seeing if we found a self-citation rate of $\sim 13\%$ in our database.

To answer the questions in our project, we took three steps: upload the data to and clean it using Snowflake, query the data, and then analyze and plot the results using Python. The major challenge of this project was the size of the data set, about 100 GB in total. This greatly complicated our ability to manipulate and work with the data.

To split up the work involved in this project, we each took on different parts. Before we resolved to use Snowflake, we were attempting to download the data to our personal computers, clean it using Python, and then upload it to PostgreSQL. Yang wrote the script for cleaning the data and converting it to our Schema in Python and since initially Marta was the only one who was able to download all of the Semantic Scholar data onto her personal computer, she worked on running the script and uploading the cleaned data to PostgreSQL. In the end, this attempt failed, and we had to re-evaluate. Olga then figured out how to upload the data to Snowflake and cleaned the data and formatted it into our schema directly there. Then Yang carried out H-Index queries, Olga investigated the adjusted H-index, and Marta completed all queries concerning self-citation rate. Each group member analyzed and created plots for their respective queries and contributed to their sections of the presentations and final paper.

III. Data Upload and Cleaning

The first step of our project revolved around uploading and cleaning the Semantic Scholar data. Initially, we attempted to use PostgreSQL for this step, but quickly realized that the system was inefficient at data ingestion. Therefore, we decided to use Snowflake, a cloud data platform, due to its superior ability to handle and manipulate large magnitude data. To put it into perspective, uploading a subset of the data to PostgreSQL took a number of hours, whereas uploading the entirety of the data to Snowflake took less than five minutes.

Once we had the data ingested and were able to work with it, our next objective was to clean it and format it into our schema, shown in Figure 1. Our schema contains three entities: Author, Publication, and Journal. There exists a many to many relationship Authored between Author and Journal, a many to one relationship PublishedIn between Publication and Journal, and two many to one relationships for in-citations and out-citations between Publication and itself. Bold attributes denote primary keys. In this project, in-citations refer to papers referencing the publication in question and out-citations refer to papers being referenced by the publication in question. Of the original twenty-two attributes contained in the raw Semantic Scholar data we kept author ID, author name, publication ID, year of

publication, publication venue, fields of study contained in the publication, and journal name. In addition, we use out-citations as the citation relationship between publications instead of in-citations and provide the reason in Section VI.

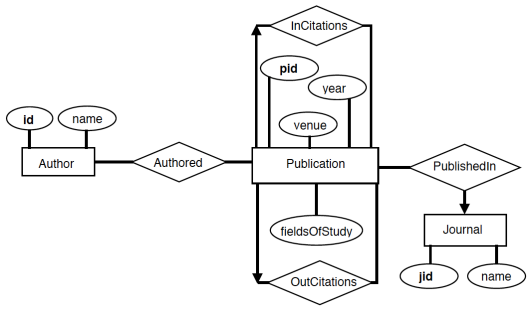

Fig. 1. The schema of our data. There are three entities: Author, Publication, and Journal. There exists a many to many relationship Authored between author and journal, a many to one relationship PublishedIn between Publication and Journal, and two many to one relationships for inCitations and out Citations between Publication and itself. Bold attributes denote primary keys.

IV. Querying the Data

In this section, we discuss the queries for our analysis. The first question we hoped to address dealt with the most influential authors based on H-Index. This calculation was a bit involved. As shown below, in order to calculate the H-index, we query table Authored, which stores the relationship between authors and publications, and table Citation, which stores the citation relationship between publications. For each author, the inner query selects their name, their publications, the number of citations, and the citation ranking within all their publications. Then, the outer query counts the number of publications that have more citations than its ranking, and we can show that it is equal to the author's H-Index.

```
SELECT authid, count(*) AS h_index
FROM
    (SELECT a.authid, c.outcitations,
    COUNT(c.pubid) AS citations_number,
    RANK() over (
        PARTITION BY a.authid
        ORDER BY COUNT(c.pubid)
        DESC)
    AS publication_citation_ranking
    FROM Authored a, outcitations c
    WHERE a.pubid = c.outcitations
    GROUP BY a.authid,
        c.outcitations) T
WHERE publication_citation_ranking
    <= citations_number
GROUP BY authid;
```

Our second question involved defining an Adjusted H-Index and investigating whether self-citation had a strong influence on author rankings with respect to H-Index. Adjusted H-Index was found the same way as H-Index, with the only difference being that the number of self-citations was subtracted from citation count. Once this calculation was completed, we looked at the ratio of Adjusted to standard H-Index for all authors to see how much on average the Adjusted index decreased their score. An example query for calculating this ratio can be found below. In it, we select the standard H-Index and the Adjusted H-Index and author ID for each author from the H-Index and Adjusted H-Index tables joined on author ID. We also calculate the ratio of Adjusted to standard H-Index and select that as our final column.

```
SELECT x.authid, x.h_index,
    y.adj_h_index,
    y.adj_h_index/x.h_index AS h_ratio
FROM H_Index x, Adjusted_H_Index y
WHERE x.authid = y.authid;
```

Finally, we investigated of self-citation rate. Recall that self-citation is defined as an author referencing their own work in a publication. In this query we count the total number of citations an author has for all publications, the total number of self-citations, where they are an author on the citing publication, and then the ratio between the two counts for each author. An example query for calculating self-citation rate can be found below.

```
SELECT sc1.authid,
    COUNT (outcitations) AS all_cite,
    self_cite, self_cite/all_cite AS ratio
FROM outcitations, authored,
    (SELECT a1.authid,
        COUNT(outcitations) AS self_cite
    FROM outcitations out,
        authored a1,
        authored a2
    WHERE out.pubid = a1.pubid
    AND out.outcitations = a2.pubid
    AND a1.authid = a2.authid
    GROUP BY a1.authid) AS sc1
WHERE outcitations.pubid = authored.pubid
```

```
AND sc1.authid = authored.authid
GROUP BY sc1.authid, self_cite;
```

V. Results and Discussion

First, we calculated and investigated H-Index within our dataset. Figure 2 shows the top twenty authors with the highest calculated H-Indexes and their corresponding H-Indexes according to Google Scholar.

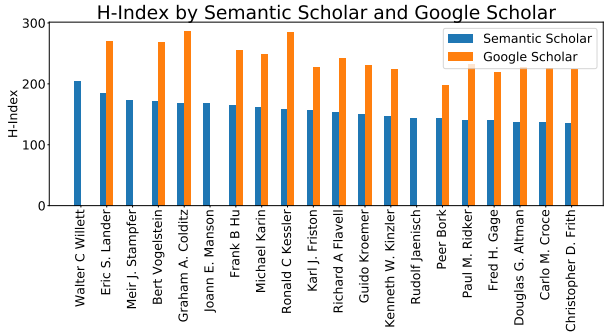

Fig. 2. Top 20 scholars with highest H-Index in the Semantic Scholar database compared to the H-Index of those same authors in Google Scholar.

We note that there is a difference between the two H-Indexes for each author. Some authors in the Semantic Scholar dataset did not have information available on Google Scholar. For those that did, in all cases, Google Scholar calculated a higher H-Index for each author than we did. This was likely due to our dataset being smaller and containing different information than the one used by Google.

Next, we wanted to investigate whether an H-Index adjusted to include self-citation information was a better indicator of citation impact than standard H-Index. Figure 3 plots H-Index and Adjusted H-Index for the 100 top authors in the data set, ranked by citation impact according to H-Index (i.e. Author 1 has the highest H-Index in the data set and Author 100 has the 100th highest).

What we see is that at least for the most influential authors, Adjusted H-Index does not vary much from standard H-Index, and author ranking would not appear to change drastically if Adjusted-H index were used as the metric instead. Some authors even have equivalent standard and Adjusted H-index values. This could be because the Semantic Scholar dataset is incomplete, and there was not enough information on self-citations available. On the other hand, it could be that because these authors have so many citations on their publications, subtracting the self-citations genuinely does not affect their overall scores. It is important to

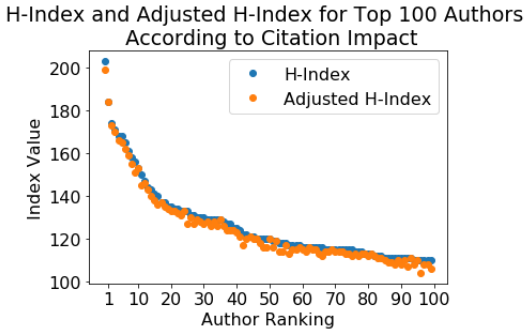

Fig. 3. A comparison of standard and Adjusted H-index for the top 100 authors in terms of citation impact.

point out that some authors have Adjusted H-Indexes significantly lower than their standard scores, but there are only a few of these outliers. This is further shown in Figure 4, where we plot a histogram of the ratio of Adjusted to standard H-Index for all authors in the dataset.

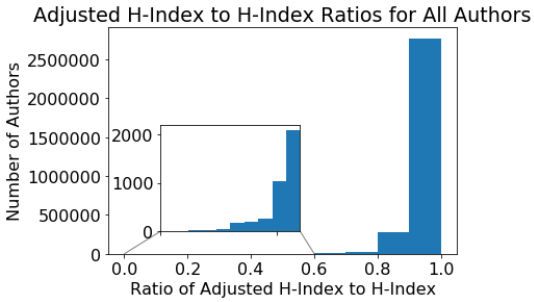

Fig. 4. A histogram of the ratio of Adjusted to standard H-index for all authors in the dataset.

From Figure 4, we see that the vast majority of authors have ratios of Adjusted to Standard H-Index of 0.80 or higher. In fact, the average ratio value was 0.96. This suggests that for most authors in our dataset, the difference between Adjusted and standard H-Index was minimal. Although there were authors with ratio values 0.60 or lower, as shown in the zoomed in version of the histogram, the number was somewhere in the thousands, which was almost insignificant compared to the millions whose ratio values were larger. The maximum ratio of Adjusted to standard H-Index was 1.00, meaning that some authors had no known self-citations and the minimum ratio value was 0.00, meaning that some authors had publications that only cited themselves. Overall, however, we concluded that an Adjusted H-Index does not have a large impact on author ranking and would not necessarily be a better indicator of author influence than H-Index.

Finally, we investigated self-citation in our dataset. We calculated the average self-citation rate across all authors in the dataset in order to get a comparison between Semantic Scholar and the results in the literature. We found that the average number of total citations per author is 340.50, the average number of self-citations per author is 14.6, and the average self-citation rate of all authors in the Semantic Scholar database is 4.7%. We also calculated the median and found it to be even lower at 2.5% This is significantly lower than the $\sim$ 13% calculated average and median self-citation rates reported in the literature [4], [6]. This could suggest that our dataset is not representative in terms of containing all of the publications for every author. However, the Semantic Scholar dataset has over 26 million unique authors and 70 million publications, and therefore is significantly larger than the datasets that were used by [4], [6], which contained 100,000 researchers and 1.5 million publications. Therefore, it is instead possible that our database is actually more representative of the general literature than those used by previous researchers. Yet another possibility is that [6], for example, used PubMed as a data source, which focused on fewer fields than ours than Semantic Scholar and could have skewed their data. Furthermore, we analyzed how self-citation rate is influenced by field of study, journal of publication, and age of author. For Figure 5, we calculated the average self-citation rate per author in each field and plotted in a bar graph in order of increasing self-citation rate.

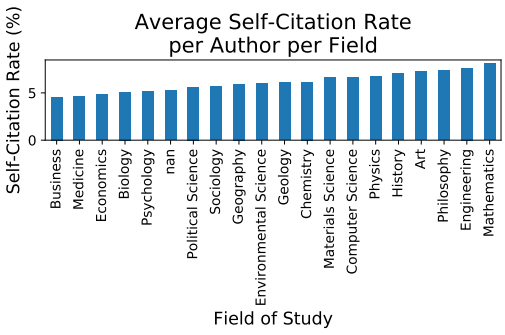

Fig. 5. A bar plot of average self-citations per author in each field.

From Figure 5, we can see that Mathematics has the highest self-citation rate of any field (8%) and business has the lowest (4.5%). There is a linear relationship between self-citation rate and the number of total publications or total authors that are in a particular field (not shown), so the results in this plot seem to generally be due to the size of each field. One claim made by the literature was that more productive authors (authors with more papers and citations overall) have a higher self-citation rate [6]. At least on a per-field average basis, our data does not corroborate this claim. The average number of self-citations per author in each field has a linear relationship with the average number of total citations per author. However, the self-citation rate is not related to number of overall citations.

Another area of investigation was how journals compared to each other in terms of self-citation rates, in this case measuring self-citation on average per each paper in each journal. In Figure 6, we plotted the self-citation rates for the 25 journals that have the most publications in the dataset, ranked by number of publications (i.e. Journal 1 has the most publications and Journal 25 has the fewest in the top 25) . Figure 7 contains the names of the journals corresponding to those referenced in Figure 6.

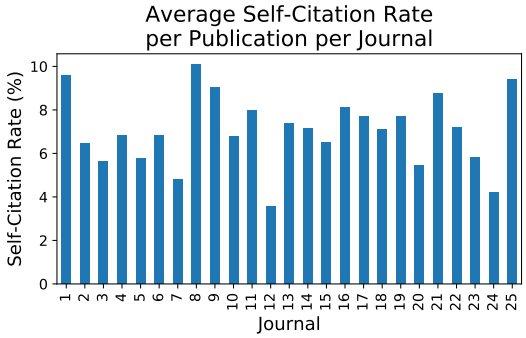

Fig. 6. A bar plot of average self-citations per publication in the 25 journals with the most publications.

As can be seen in Figure 6, the self-citation rates are not determined by the number of publications, and they vary quite a bit. If we compare two journals, Nature and Science, which have similar impact factors, 43.07 and 41.06 respectively, and are both multidisciplinary, Nature has a self-citation rate per paper of 5.4% and Science has a self-citation rate per paper of 8.8%, almost double. Although Nature has many more papers in the data set than Science, which isn't among the top 25, that should not be causing the difference. It could potentially be due to the types of authors who publish in one journal versus another.

Another factor we analyzed was author age. Because our dataset did not include demographic information, we used the year that an author published their first paper as a proxy for their age. In Figure 8, we plot average number of total and self-citations for authors

Number	JournalName
1	ArXiv
2	PNAS
3	Journal of Neuroscience
4	Cancer research
5	Blood
6	Journal of virology
7	NeuroImage
8	Journal of bacteriology
9	Optics express
10	Applied and environmental microbiology
11	The Journal of clinical investigation
12	ACS applied materials & interfaces
13	The Journal of Cell Biology
14	Investigative ophthalmology & visual science
15	Environmental science & technology
16	The Journal of general virology
17	Experimental Brain Research
18	Journal of neurophysiology
19	Infection and immunity
20	Nature
21	Circulation
22	Psychopharmacology
23	Applied Microbiology and Biotechnology
24	Environmental Science and Pollution Research
25	Plant physiology

Fig. 7. A key for the names of the journals from Figure 6

whose first papers were published in a given year.

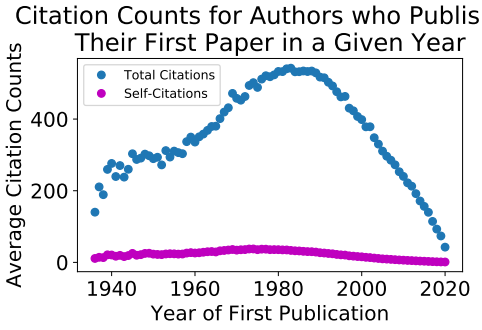

Fig. 8. A scatter plot of the average total and self citations for authors who published their first paper in a given year (used as a proxy for age).

From Figure 8, we see that the average total number of citations for authors who began publishing in 1980 is the highest overall, and there is a slight bump in the average number of self-citations for these authors as well. It makes sense that authors who began publishing more recently have fewer citations overall because their careers are shorter. In Figure 9, we further investigate this phenomenon by plotting the average self-citation rates for authors who published their first paper in a given year.

In Figure 9, we find that self-citation rates have been decreasing steadily over time, since the first year in the dataset. This is a surprising result because a previous study found that self-citation was more prevalent among younger authors, who have more of an incentive to boost their H-Index, as they are starting

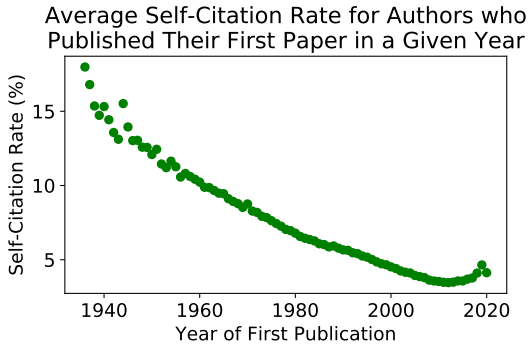

Fig. 9. A scatter plot showing the self-citation rate for these same authors

out their careers and job searching. Furthermore, since younger authors have fewer citations overall, one would expect self-citations to have more of an impact on their H-Indexes [5]. However, our data shows the opposite result, until the last decade, where we do see a slight increase in self-citation rate. Again, we also see that more productive authors don't have higher self-citation rates - they do on average have higher self-citation counts, but these are proportional to their total citation counts.

VI. Conclusions and Future Directions

Overall, our results do not agree with previous studies from the literature measuring self-citation rates and H-Indexes. The dataset we used is much larger than those used in previous studies [1], [4], but also likely smaller than that used by Google Scholar. This leaves it still unknown as to whether our dataset is more or less representative of existing publications than previous ones.

It would be nice to have a "sanity check" where we could compare our calculated H-Index to one calculated by Semantic Scholar, but Semantic Scholar calculates an "influence score" instead of an H-Index for each of their authors. Their influence score "measures the impact of one author's publications on another author's work...The score is based on a weighted combination of citations and Highly Influential Citations" [3]. Unfortunately, the exact calculation of Semantic Scholars' influence score is not explicitly provided, and therefore it is hard to check if our calculations from the Semantic Scholar data agree with theirs.

One potential source of error in our dataset involved the raw data columns labeled as 'incitations' and 'outcitations', which represent publications citing a particular paper and publications being cited by that same paper,

respectively. In theory, these columns should contain redundant information, since if paper A is cited by B in 'incitations', then paper B should cite A in 'outcitations.' However, we found that these columns did not match when we first cleaned the Semantic Scholar data. In the end we chose to work with 'outcitations,' because that column contained more information, but we have questions about why the columns were unequal to begin with and if it had something to do with a flaw in the data collection. We would like to study combining information from both columns further, perhaps by taking the union of 'incitations' and 'outcitations' and seeing if this changes our calculations. Another direction that we would like to explore is to calculate a citation metric for journals (analogous to H-Index for an author). One example would be an Impact Factor per journal, which is equal to the frequency with which the average article in a particular journal has been cited in a given year. We could calculate this metric for our dataset, then calculate an Adjusted Impact Factor to take into account self-citations. This would both enable us to compare our dataset to other sources that calculate impact factor as well as further investigate the role of self-citations in promoting journals.

REFERENCES

[1] R. Van Noorden and C. D. Singh, "Hundreds of extreme self-citing scientists revealed in new database." *Nature*, vol. 572, no. 7771, p. 578, 2019.

[2] J. W. Flatt, A. Blasimme, and E. Vayena, "Improving the measurement of scientific success by reporting a self-citation index," *Publications*, vol. 5, no. 3, p. 20, 2017.

[3] W. Ammar, D. Groeneveld, C. Bhagavatula, I. Beltagy, M. Crawford, D. Downey, J. Dunkelberger, A. Elgohary, S. Feldman, V. Ha *et al.*, "Construction of the literature graph in semantic scholar," *arXiv preprint arXiv:1805.02262*, 2018.

[4] J. P. Ioannidis, J. Baas, R. Klavans, and K. W. Boyack, "A standardized citation metrics author database annotated for scientific field," *PLoS biology*, vol. 17, no. 8, p. e3000384, 2019.

[5] M. M. King, C. T. Bergstrom, S. J. Correll, J. Jacquet, and J. D. West, "Men set their own cites high: gender and self-citation across fields and over time," *Socius: Sociological Research for a Dynamic World*, vol. 3, 2017.

[6] S. Mishra, B. D. Fegley, J. Diesner, and V. I. Torvik, "Self-citation is the hallmark of productive authors, of any gender," *Plos One*, vol. 13, no. 9, p. e0195773, 2018.