Citation Analysis

Olga Dorabiala, Marta Wolfshorndl, Yang Zhou

In our project "Citation Analysis", we aim to produce a comprehensive analysis of citation attributes from the Semantic Scholar dataset. We are interested in answering questions such as: 1) A self citation analysis determining the measurement of self-citation per paper for each author. It was recently reported in a paper published in 2019 that self-citation is a hidden problem in many scientific fields, artificially inflating an author?s citation statistics. We will determine if this is the case in the Semantic Scholar data, who the worst offenders are in terms of authors and publications, and how many self-citations to other citations there are per paper. 2) How many citations away from a Turing award winner is the average author (in different fields)? We think it would be interesting to determine how many citations or coauthors each author is away from famous individuals in their field or in other fields. 3) What are the most influential papers and journals based on metrics such as the H-index?

The literature that we were able to find for self-citation research has all been relatively recent. We were initially aware of the issue of self-citation through a 2019 Nature news feature, which describes a paper in which citation metrics in a database of around 100,000 researchers were analyzed (Noorden and Chawla 2019; Ioannidis, et al 2019). The paper determined that the median self-citation rate of researchers in their database was 12.7 %. There are examples of so-called extreme self-citers, the majority of whose citations come from either themselves, or their co-authors. The paper also describes 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 a scientist's career.

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. 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 Author-ity (Mishra, et al 2018). Their findings on gender contradicted the results from King, et al., finding that the gender effect on self-citation largely disappears when correcting for previous citation counts. They also 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 on the parameters tested. Their results also found that self-citation rate is about 13%, which agrees with the findings from Ioannis et al 2019. 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 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. Our approach to examining self-citation will be to calculate our own s-index, which may be similar to that of Flatt, Blasimme, and Vayena. We are also planning on calculating an h-index, and we are interested in comparing the s-index to h-index and determining if, as proposed by Mishra, et al., more highly productive researchers are also more likely to self-cite. Our data set doesn?t have the demographic data, such as gender, that was used in other papers, but we are interested in looking at such factors as venue and field of study in relation to self-citation. We are interested to see if we can replicate some of the findings of these papers, including whether we will see a self-citation rate of 13% in our database.

Our approach to answering our project questions involves four steps: clean the data using Python, upload the data to PostgreSQL, query the data, and then analyze the results. The major challenge of this project is the size of the data set, about 100 GB in total. This greatly complicates our ability to manipulate and work with the data. So far, we have cleaned the data and are working on uploading it.

In cleaning the data set, we considered two possibilities. We could either directly import the semi-structured data in its JSON format to pSQL or manipulate it in a Python wrapper first. The original dataset from Semantic Scholar was 100 GB and contained twenty attributes, not all of which we intended to use, and due to its size, uploading directly to pSQL seemed infeasible. The latter approach allowed us to clean the data, remove fields we deemed unimportant for our analysis, and reduce the size of the data set before upload. Of the original twenty attributes, we kept only the id of the publication, the names and unique ids of the authors, the year the paper was published, the publication venue, the journal name, the fields of study the paper in question addresses, all of the in citations, and all of the out citations. These attributes were enough for us to form our schema. In the end, we decided to have a relation Author with attributes id (the key) and Name; a relation Publication with attributes pid (the key), Year, fieldsOfStudy, publishedIn, and Venue; and a relation Journal with attributes journalId (the key) and journalName. We also have a one to one relationship PublishedIn between Publication and Journal, two one to many relationships from Publication to itself to keep track of in and out citations, and a many to many relationship Authored from Author to Publication. See Figure 1 for the Entity Relationship diagram.

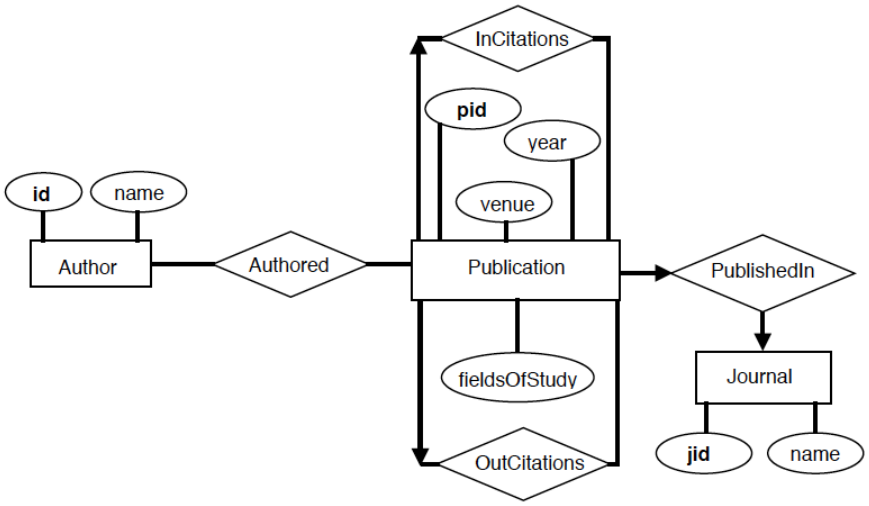

FIGURE 1. Entity Relationship Diagram for the Semantic Scholar dataset

Currently, we are in the process of uploading the data to PostgreSQL. Although our data cleaning allowed us to reduce the size of the data set by about half, we are still working with around 50 GB of data, and we anticipate the upload to be a very slow process. We are considering if there is any other way for us to further reduce the size of the data. For one, if the size of the in-citation table is consistent with the size of the out-citation table, we will only upload one overall citation table to the database to save time and space because in-citations and out-citations should ultimately contain the same information. However, if the table sizes are inconsistent, this will not be an option. We now also face the added challenge of only being able to have one person run queries and produce results, as only one group member was able to download the initial data set and will have to run the script to clean all the data and import it into pSQL.

For the rest of the quarter, we will focus on the final two steps of our project, querying the data and analyzing the results. Although all group members do not have access to the data, we plan to use the fact that the original data set from Semantic Scholar comes in batches to our advantage. We hope to individually experiment on these smaller, cleaned batches to write queries that can answer our project questions. For instance, one of our goals is to determine a measurement for self-citation. Determining this self-citation metric will involve not only counting the number of out-citations that come from the author's own papers, but comparing this number to the average number of self-citations among various authors and the frequency of in-citations from other authors. Once we have a metric, we can determine which authors are the worst offenders of citing themselves and which journals publish the most self-cited papers. Another goal is to determine the h-index of various authors. A scientist having an index of h is defined as h of his or her N_p total papers having

at least h citations each, and the other $(N_p - h)$ papers each having less than h citations (Hirsch 2005). The query to determine h-index will have to count the number of papers associated with each author, the number of citations each paper has, and then calculate the maximum number of papers h that each have at least h citations. Once we execute this query, we will compare the value of our calculated h-index to that of the h-index found on google scholar to see how closely our calculations agree. We will also consider how and if self-citations impact h-index by removing self-citations from the total number of citations each paper reports, and by calculating a self-citation index for each author. Finally, if time allows, we will investigate how many citations or coauthors each author is away from famous individuals in their field. To do so, we will have to construct a citation graph and count the number of edges that constitute the shortest path from one author to a famous one. We are still investigating what kind of method we would use to carry out this query, and whether such a query is even possible in pSQL due to its seemingly recursive structure.

References:

Flatt J, Blasimme A, Vayena E (2017) Improving the Measurement of Scientific Success by Reporting a Self-Citation Index. Publications 5(3):20.

Hirsch, Jorge E. "An index to quantify an individual's scientific research output." Proceedings of the National academy of Sciences 102.46 (2005): 16569-16572.

Ioannidis, John PA, et al. "A standardized citation metrics author database annotated for scientific field." PLoS biology17.8 (2019): e3000384.

King MM, Bergstrom CT, Correll SJ, Jacquet J, West JD (2017) Men Set Their Own Cites High: Gender and Self-citation across Fields and over Time. Socius Sociol Res a Dyn World 3:237802311773890.

Mishra S, Fegley BD, Diesner J, Torvik VI (2018) Self-citation is the hallmark of productive authors, of any gender. PLoS One 13(9):e0195773.

Van Noorden, Richard, and Chawla D. Singh. "Hundreds of extreme self-citing scientists revealed in new database." Nature 572.7771 (2019): 578.