UNIVERSITÄT PASSAU

BACHELOR'S THESIS

Analysing Characteristics of Self-Citations in Computer Science

Author:
Moritz GRÜNBAUER

Supervisor:
Dr. Christin SEIFERT

A thesis submitted in fulfillment of the requirements
for the degree of Bachelor of Science

at the

Lehrstuhl für Data Science
Fakultät für Informatik und Mathematik

August 19, 2017

Declaration of Authorship

I, Moritz GRÜNBAUER, declare that this thesis titled, "Analysing Characteristics of Self-Citations in Computer Science" and the work presented in it are my own. I confirm that:

- This work was done wholly while in candidature for a bachelor's degree at the University of Passau.
- This work was never submitted to another university in this form or any similar one.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

_____

Date:

_____

Universität Passau

Abstract

Fakultät für Informatik und Mathematik

Lehrstuhl für Data Science

Bachelor of Science

Analysing Characteristics of Self-Citations in Computer Science

by Moritz GRÜNBAUER

When writing scientific papers, authors commonly reference other works. At times they even cite their own previous work, which is called an author self-citation.
This thesis analyses a dataset of over three million computer science papers and their citations to each other, in order to identify self-citations among them, using a graph based database. Then, the frequency of self-citations is investigated, to find out whether there are any correlations between it and various characteristics of papers. The examined characteristics include publishing year, number of authors, position of authors, rank of publishing venue and outgoing as well as incoming citations.
The best indicator for a paper being self-citing that was noticed, was the number of citations a paper is a part of. A trend for highly prestigious papers to contain self-citations more often than less prestigious papers was discovered when evaluating correlations between self-citation frequency and conference rank as well as the number of received citations.

Acknowledgements

First and foremost, I thank Dr. Christin Seifert for her advice and supervision of this thesis. I would also like to thank Prof. Dr. Granitzer and Prof. Dr. Kosch for supervising the seminar where this thesis was presented.
Additionally, I thank Emmanuel Berndl for advising me on using graph databases, and Sebastian Böhm for managing the server on which my calculations were run.

Contents

List of Figures

List of Tables

To my family, and everyone who supported me during my studies.

Chapter 1

Introduction

1.1 Defining Self-Citations

In order to examine self-citations, a definition of the term has to be given first. In general, one defines a citation between two papers that have some characteristic in common, as a self-citation. This characteristic can be anything quantifiable, such as the country, journal or conference where this paper was published, for example.

In this thesis author self-citation are examined, which are the most commonly researched types of self-citations. In a specification of the general definition, an author self-citation is a citation between papers that share one or more authors.
For the rest of this thesis, the terms self-citation and author self-citation are used interchangeably, unless explicitly stated otherwise.

1.2 Motivation

The core question behind this thesis is how characteristics of papers correlate with the frequency at which self-citations occur. Out of this data, trends for how specific characteristics influence the likelihood of a paper being self-citing can be attempted to be extracted.
This can be useful to learn what causes self-citations and where one can find self-citing papers with high likelihood.
Among the most interesting characteristics of papers analyzed in this thesis, the rank of the publishing conference and the number of incoming citations can be used to

estimate how prestigious papers are. With this, this thesis can attempt to make a statement about whether more prestigious papers tend to be more self-citing than others or not.

The results presented in this thesis could be useful for future research into self-citations, to learn whether self-citation is a good practice, for example.

1.3 Related Work

In the past, a lot of research has gone into how self-citations affect popular ranking methods of journals and conferences, such as the h-index, which can be inflated by self-citations (Zhivotovsky and Krutovsky, 2008), and how these effects can be detected (Bartneck and Kokkelmans, 2010) and corrected for (Schreiber, 2007).
This thesis, however, uses the manually ranked CORE conference rank (CORE, 2017), which is not purely based on the number of received citations. Also, this thesis investigates the relation between venue rank and self-citations in the opposite direction, trying to gather how rank influences self-citation frequency.

In a similar study to this thesis, Aksnes (2003) performed an analysis of over 45,000 Norwegian papers and how self-citations within them correlate with various characteristics of papers. This was later expanded on by Fowler and Aksnes (2007). What separates this thesis from the mentioned studies, is that the dataset of papers is much larger in scope, and also international.

Another point of research concerning self-citations is their use in different disciplines (Snyder and Bonzi, 1998), time frames (Rousseau, 1999) and motivations for them (Bonzi and Snyder, 1991).

1.4 Approach of this Thesis

In the following chapters of this thesis, the used datasets will be introduced and an explanation as to why they where chosen will be given. Additionally, some issues with the usage of this data will be discussed. After that, the process of reformatting the data and importing it into the graph database will be described. Also, it will be explained how the data is modified in the database, and finally exported from it.
After the methodology, in which the results were acquired, has been shown, the evaluation of those results includes general statistics and a specific investigation of the correlation between each paper characteristic and the frequency of self-citation. Then, the effects of a bug within the Aminer dataset are evaluated, and finally overarching trends in the results are identified and summarized.

Chapter 2

Methodology

In this chapter, the methodology of of the performed research is shown and the choices behind it are explained. The used datasets are introduced and issues in their usage are shown and treatment of those issues is explained. Then, the performed actions to identify the self-citations within the dataset are shown, including pre- and post-processing of the data.

All Java applications and cypher Neo4j scripts used and described here are available online in this thesis' github repository[1].

2.1 Used Datasets

For the research performed for this thesis, two datasets are utilized. One, the Aminer citation graph, carries the information about the papers and their citations to each other, whereas the other, the CORE conference rank, contains a ranking of international computer science conferences. Explanation as to why each of those datasets were chosen is also given.

2.1.1 Aminer Citation Graph

For information about papers and their citations to each other, this thesis uses the Aminer citation graph (Tang et al., 2008). The dataset can be found online[2] in a number of different versions. For this thesis, the DBLP V8 version is used, which is dated July 14, 2016, even though on July 3, 2017 the DBLP V9 version was released.

[1] https://github.com/MoritzGr/AnalysingSelfCitations
[2] https://aminer.org/citation

TABLE 2.1: Metadata about the number of entries in the Aminer citation graph

Number of Papers	3,272,911
Number of Citations	8,466,859
Number of Unique Author Names	1,752,443
Number of Authorship Relations	9,235,190
Number of Unique Venue Names	12,626

This was, however, at a point in time when the calculations were already running for the DBLP V8 version, so the new dataset was not utilized. Though, since the formatting of the dataset stayed consistent, the methodology described in this thesis can be applied to this new version, and any future, similarly formatted version of the dataset.

This citation graph is very expansive compared to other available datasets (See table 2.1 for exact number of entries). For every paper within the dataset, information about the following characteristics are given:

- **Title:** A clear text string representation of the title is given for every paper within the dataset. Since it is difficult to quantify this characteristic and find correlations to self-citation frequency, the title is not evaluated further and dropped from the dataset in order to reduce size.
- **Authors:** A list of comma separated names of authors in order of author position is given. From this, author self-citations can be identified, and characteristics for number of authors and position of authors are extracted.
- **Publishing Year:** A single number is given to identify the year in which the paper was originally published. This is utilized as a characteristic to investigate the distribution of self-citations across the years.
- **Publication Venue:** A clear text string is given to identify the venue in which the paper was originally published. This is combined with another dataset for conference ranks in order to extract a characteristic of conference rank.
- **Citations:** For information about citations between the paper and other papers, each paper has an index assigned to them. Also, every paper has a list

of indices of the papers which it cites. This information is used to identify citations, and self-citations, and can be utilized as characteristics for the number of incoming and outgoing citations of a paper.

- **Abstract:** A clear text string representation of the abstract is given for some papers. Similarly to the title field, this information is not easily quantized into a characteristic to investigate, and is therefore dropped from the dataset to reduce its size.

There are, however, two issues with this dataset. One is specific to the Aminer dataset, and one affects most other datasets of the same kind.

Duplicate Citation indices

Due to a bug in the Aminer software, the indices used to represent citations are not unique to papers. There are 173,051 papers, which is about ~5.3% of all papers, that have indices not unique to them, which makes any citations to them not clearly defined. It is impossible to distinguish which one of the papers is supposed to receive a citation when an index is given that is used across multiple papers.
When asked about this, Dr. Jie Tang, who manages the Aminer project, stated that this bug will be fixed in future versions of the dataset, but said that these entries could be safely ignored.
Since the papers affected by this bug can be easily identified by their duplicate indices, the data is evaluated in two ways. Once, while ignoring all papers that have duplicate indices, and a second time while including papers with duplicate indices. Then the two evaluations will be compared to find whether there are any trends introduced by the duplicate indices and whether overarching trends still hold.
Once a fixed version of the dataset is released, the methodology described in this paper can be repeated on that dataset in order to get more precise results.

Author Name Ambiguity

A problem that will affect most datasets and approaches of similar work is author name ambiguity. It is core to identifying author self-citations, since identifying whether two authors of two papers are the same person is the basic problem to be

solved. But since information about the authors is limited to their name, this identification proves difficult. Author names may be ambiguous, which manifests itself as two sides of the same issue, the homonym and the synonym problem.
When multiple authors go by the same name, one can falsely identify them as the same person, which is the homonym problem.
When one author goes by multiple names, one can falsely identify them as two different persons, which is the synonym problem.
Both of these problems interfere heavily with the identification of author self-citations, but their treatment proves difficult. While there are multiple approaches to disambiguating author names algorithmically, none of them have perfect accuracy. Combined with the fact, that there is no ground truth for the Aminer citation graph for author name ambiguity, this means that there is no guarantee that a disambiguation algorithm will improve the dataset's accuracy. Also, with access only to the Aminer dataset, an abstraction of the DBLP database, which is itself only an abstraction of the full text papers, valuable information for author name disambiguation, such as field of research, author nationality or publication location, among others, is not available.
Due to these factors no disambiguation is performed on author names. Since accuracy on a paper level, rather than on a histogram level, is not important to the research that is being done in this thesis, trends within the correlation between self-citations and paper characteristics should still hold. This is because the ambiguity problem affects all papers equally.

It proved to be very difficult, however, to provide an estimate on how much of an effect this issue has on the number of identified self-citations. This is due to the lack of ground truth in the dataset, which is something that only exists in manually labeled datasets, which are commonly used as training data. The issue is, that these datasets are constructed to maximize the errors within the dataset, to provide more efficient training, which would result in a skewed error estimate, if used as an error estimate for this thesis.

TABLE 2.2: Ranks within the CORE ranking system

Descriptions of ranks are taken from CORE (2017)

Rank	Description
A*	flagship conference, a leading venue in a discipline area
A	excellent conference, and highly respected in a discipline area
B	good conference, and well regarded in a discipline area
C	other ranked conference venues that meet minimum standards
Australasian	A conference for which the audience is primarily Australians and New Zealanders
National	A conference which is not international an therefore not ranked
Unranked	A conference for which no ranking decision has been made
L	No documentation for this rank (very rare)

Advantages of this dataset

Despite the issues mentioned above, the Aminer dataset still proved to be the most applicable dataset for this thesis. This is because the characteristics of papers within it are extremely accurate, since this dataset bases itself on DBLP[3], where the meta data of papers is rigorously checked and manually verified. The combination of this fact and the size of the dataset is why it was chosen.

2.1.2 CORE Conference Rank

To be able to quantify the information about publishing venues by assigning a rank to conferences, the CORE conference rank is consulted (CORE, 2017). It is a ranking of 1597 conferences in the domain of computer science, created with the goal to rank all international conferences in the field. It categorizes each conference into one of a number of ranks (See table 2.2). For this thesis the 2017 rank was used, which can be found online[4]. Similar to the citation graph, once a new version is released, this methodology can be repeated for the new data, provided its formatting did not change.

Since not all of the venues appearing in the Aminer citation graph are conferences, and not all conferences appear in the CORE ranking, only a fraction of papers can be assigned a CORE conference rank. 798,014 Papers out of 3,272,911 (~24.4%)

[3] http://dblp.uni-trier.de/
[4] http://portal.core.edu.au/conf-ranks/

appear in a conference for which a CORE rank could be found.

Advantages of this dataset

This dataset was chosen, since a manual ranking of conferences is more sensible as a characteristic than an algorithmic one. Most algorithmic ranking methods, such as the h-index, use solely the number of citations to papers within that venue for ranking, which means that larger conferences are heavily favored to be highly ranked. Since the data the Aminer datasets provides is only a portion of all papers ever published, which becomes more noticeable, the older a paper is, the ranks calculated from it would be skewed. Also the more interesting features of conferences, such as reputation, are not clearly reflected in those ranking methods.
Therefore a dataset that is independent of the citation graph and manually ranked, was chosen. The CORE conference rank is one the most well known and maintained datasets out of the ones that fit these criteria.

2.2 Pre-Processing

In order for the data to be analyzed in a database, first it has to be pre-processed and reformatted into a format which can easily be imported into that database. For this purpose, a piece of Java code was written that parses the Aminer citation graph line by line and writes its information into multiple comma separated value files.
In addition to this, a header for the comma separated values file for the CORE conference rank is manually inserted, to simplify import into the database.

In this step, the values for title and abstract of papers are dropped, in order to increase performance and decrease file size, since they are not evaluated in this thesis.

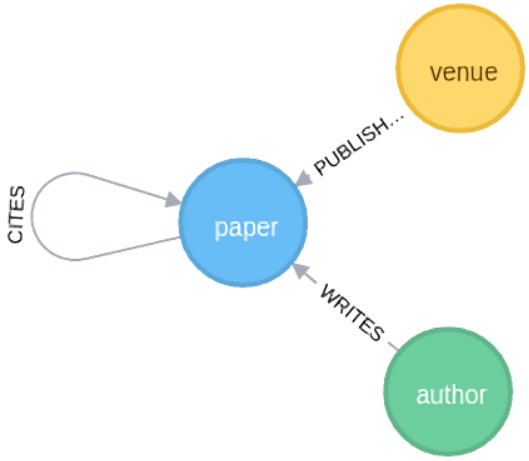

FIGURE 2.1: Simplified meta-graph for the neo4j database

This figure shows the metagraph for the used database. It includes three node labels: paper, venue, author. These nodes are connected via three relationships: PUBLISHES connects venues to papers, WRITES connects authors to papers, CITES connects papers to papers.

2.3 Graph Database

Because of the graph nature of the investigated data, Neo4j[5], a graph database, was chosen to efficiently analyze and match patterns in the citation graph. The advantage this approach provides, over relational databases, is that the computationally expensive carthesian products between entries can be saved as edges in a graph database, whereas in a relational database, every query would need a "JOIN" operator.

2.3.1 Import

In a series of import queries the comma separated value files are imported into the Neo4j database, and saved as either nodes or edges. The resulting meta-graph of the database can be seen in Figure 2.1. Following that import, papers with duplicate indices due to the bug mentioned in chapter 2.1.1, are identified and labeled, so they can be separately evaluated.

2.3.2 Identification of Self-Citations

Once the graph has been completely built, self-citations are identified via a query that matches a pattern where an author node has "WRITES" edges with paper nodes that are connected by a "CITES" edge. This pattern (See figure 2.2) constitutes an

[5] https://neo4j.com/

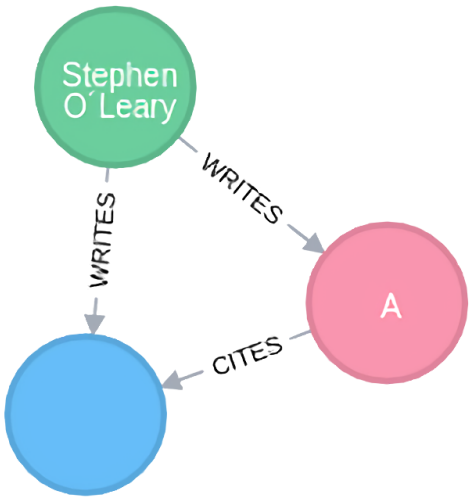

FIGURE 2.2: A self-citation pattern within the graph

The self-citing paper is marked pink. The self-citing paper has the conference rank "A", whereas the other paper is not ranked.

TABLE 2.3: Example of the format of exported data from the graph database

This example shows the amount of papers of each label combination for the characteristic of a paper being published in the year 2016. In 2016, there were 46773 papers published, that were neither self-citing nor duplicate, 2887 which were papers with a duplicate index, 4 which were self-citing and 113 which were both.

Publishing year	Paper labels	Paper count
2016	""paper""	46773
2016	""paper"",""duplicatePaper""	2887
2016	""paper"",""selfCitingPaper""	4
2016	""paper"",""selfCitingPaper"",""duplicatePaper""	113

author self-citation and, once it is identified, the citing papers are labeled as self-citing. Together with the labels for duplicate indices, this causes our meta graph to evolve from the one shown in figure 2.1 into the one shown in figure 2.3.

2.3.3 Export

After all self-citations are identified, exportation of the data is done via counting the number of papers with different label combinations for each characteristic value (See table 2.3). Since labeling of relationships is not supported in neo4j, it is instead emulated via two boolean values, when exporting the data for the author position characteristic, which is a property of author relations. All the export is done to files of comma separated value format.

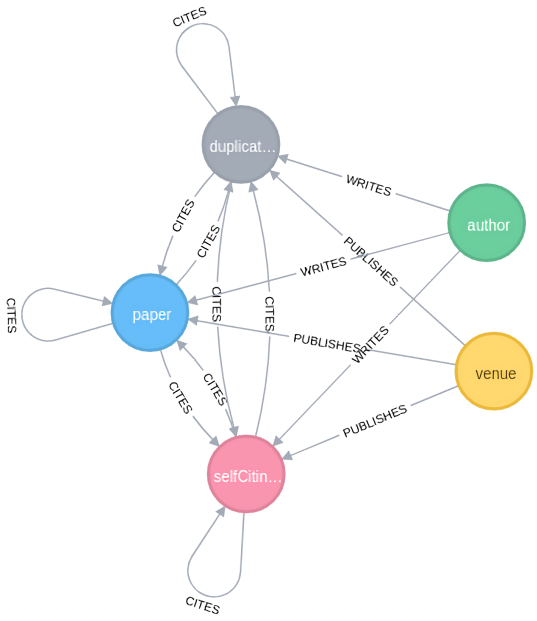

FIGURE 2.3: The fully expanded meta-graph for the neo4j database

Self-citing papers are pink and duplicate papers are grey. They both are always papers too, but Neo4j handles every label for a node as if it could exist alone.

TABLE 2.4: Example of the reformatted exported data

This table shows how the results are formatted after reformatting the raw output. The generic "value" field always describes the value of the characteristic outlined in the name of the file. Important to note is that, contrary to the original formatting outlined in table 2.3, the count of papers is cumulative for the paper types. This means that this tables selfCitingPapers column constitutes the sum of all papers which are self-citing regardless of whether they are duplicate index papers or not.

value	selfCitingPapers	papers	selfCitingDuplicatePapers	duplicatePapers
2016	117	49777	113	3000

2.4 Post-Processing

The evaluation of data in the format described in table 2.3 is inconvenient, due to the fact that for every distinct value of a characteristic, up to four separate rows can exist. For ease of later use, it is desirable to have every characteristic value only appear in one row, containing every piece of information about it (See table 2.4). For this purpose a Java application was written that parses the exported files and reformats them appropriately.

Chapter 3

Evaluation

In this chapter, the data acquired via the methodology described in chapter 2 is evaluated. First, general statistics are shown and after that, every available characteristic and its correlation to the frequency of self-citations is investigated. Then the effect of the duplicate index bug described in chapter 2.1.1 is evaluated.

3.1 General Statistics

After labeling all 3,272,991 papers, the distribution of paper types seen in table 3.1 is reached. This means that the frequency of self-citations over the whole data set is ~11.5% both when disregarding and including records with duplicate indices.

Similar statistics for authorships instead of papers can be found in table 3.2. Here, a self-citing authorship means that this specific author was self-cited within the paper that is connected to the author via this authorship relation. The average frequency of authorship relations being self-citing is ~8.2% both when disregarding papers with duplicate indices and when including them.

When investigating citation relationships in the same way in figure 3.3, it is important to note, that the duplicate index bug directly influences the number of citations in the graph. The dataset is supposed to only have 8,466,859 citations in it, but due to the duplicate index bug a number of them have no certain target, so a citation to every target was introduced, for a total of 580,386 additional relations. The likelihood for a citation to be a self-citation is about ~10.0% both when disregarding and when including duplicate indices.

TABLE 3.1: Number of identified papers of each type

Self-citing	Duplicate index	Number of Papers	Percentage
No	No	2,742,260	~83.8%
Yes	No	357,680	~10.9%
No	Yes	154,590	~4.7%
Yes	Yes	18,461	~0.6%
		3,272,991	100%

TABLE 3.2: Number of identified authorships of each type

Self-citing	Duplicate index	Number of Authorships	Percentage
No	No	8,057,808	~87.3%
Yes	No	717,451	~7.8%
No	Yes	422,548	~4.6%
Yes	Yes	37,383	~0.4%
		9,235,190	100%

3.2 Correlations of Self-Citation Frequency with Paper Characteristics

In the following part of this thesis, correlations between self-citation frequency and paper characteristics are investigated. The figures were plotted in R using ggplots2, the scripts to do so can be found in this thesis' github repository[1], where the raw data can also be found. These figures disregard papers with duplicate indices due to the bug described in chapter 2.1.1. Later, the effect of the false citations introduced into the data by this bug is investigated.

[1] https://github.com/MoritzGr/AnalysingSelfCitations

TABLE 3.3: Number of identified citations of each type

Self-citing	Duplicate index	Number of Citations	Percentage
No	No	7,189,997	~79.5%
Yes	No	796,429	~8.8%
No	Yes	953,721	~10.5%
Yes	Yes	107,098	~1.2%
		9,047,245	100%

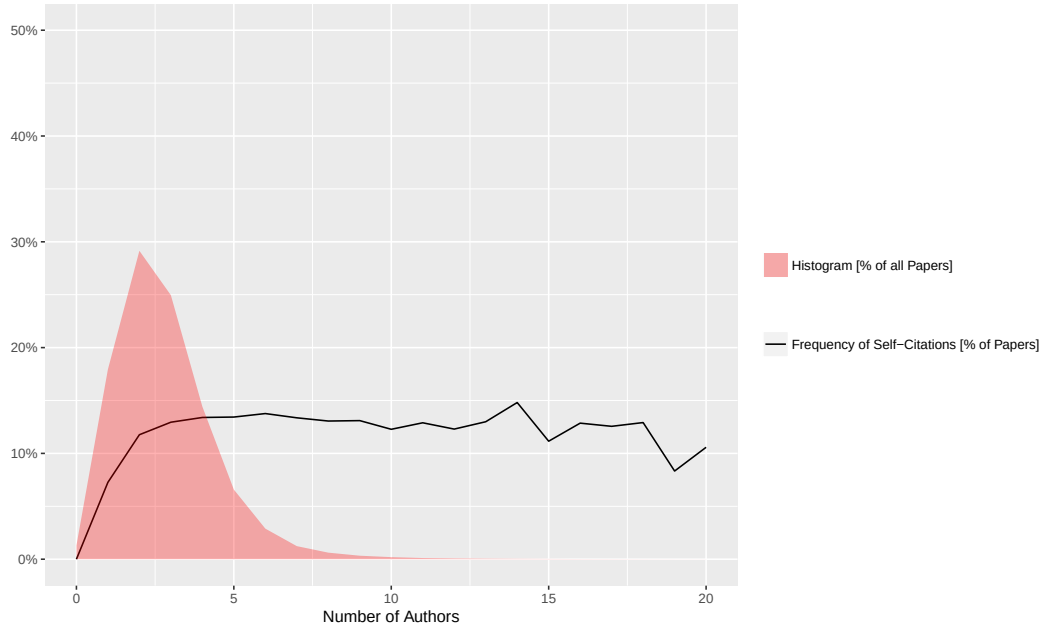

FIGURE 3.1: Number of authors plotted against self-citation frequency

3.2.1 Characteristic: Number of Authors

Figure 3.1 shows the distribution of self-citation frequency across papers with different numbers of authors. It is evident that once a papers have more than 3 authors, the frequency of self-citations is more or less constant. Papers with fewer authors than that, however, show a clear correlation, seeing as the frequency of self-citations nearly increases by about 60% from one author (~7.3%) to two authors (~11.8%). Papers with no authors, which are relatively rare with only 45,611 examples, can obviously not be author self-citing, so their self-citation frequency is 0%. The histogram peaks at 2 authors and diminishes rapidly after that causing high variance in the parts of the plot that have low sample sizes. This makes results more unreliable the higher the author count gets.

The correlation between the number of authors and self-citations at low author numbers can be easily explained by co-authorship practices. A primary author in a student or research assistant role is often incentivised to cite their supervisors work, since their own work is often derivative or dependent of it. In these cases the supervisors often also take a co-author role, which can explain the jump in self-citation frequency from one to two authors. Once the supervisor is part of the co-author

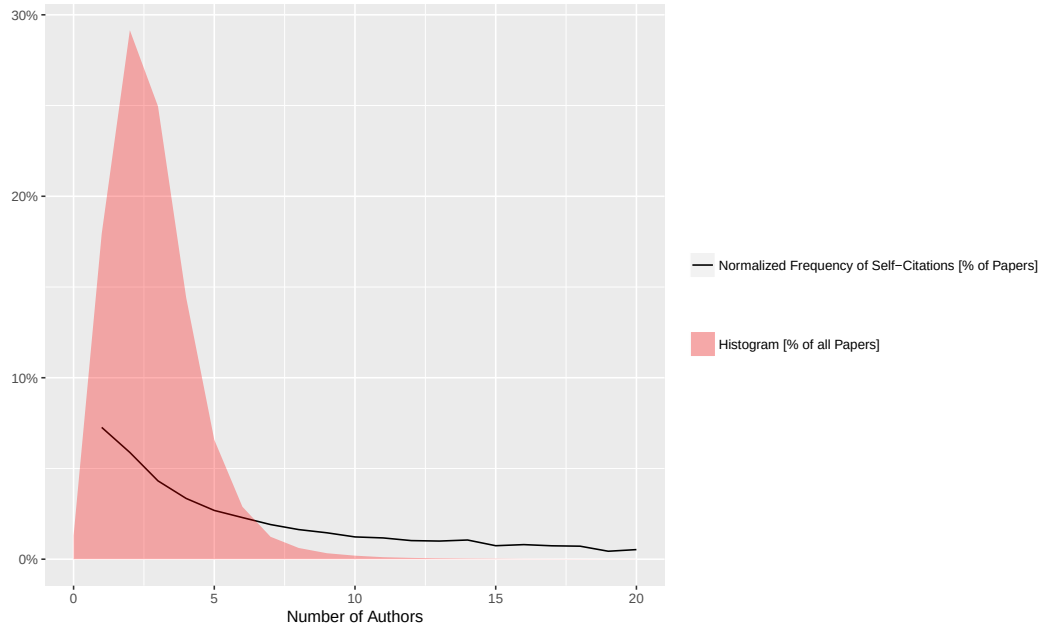

FIGURE 3.2: Number of authors plotted against self-citation frequency normalized for number of authors

group, and the paper has more than two authors, the frequency of self-citation stays constant, due to this effect not scaling with the number of authors.

This interaction is due to the fact that the number of authors directly impacts how many possible targets there are for self-citations. Therefore it is also interesting to look at a figure normalized by the number of authors (See figure 3.2). In it, it is evident that increasing the number of authors does not give individual authors a higher likelihood to be target of a self-citation. In fact, the opposite is true, the more authors collaborate on a paper, the less likely any of them are to be target of self-citations.

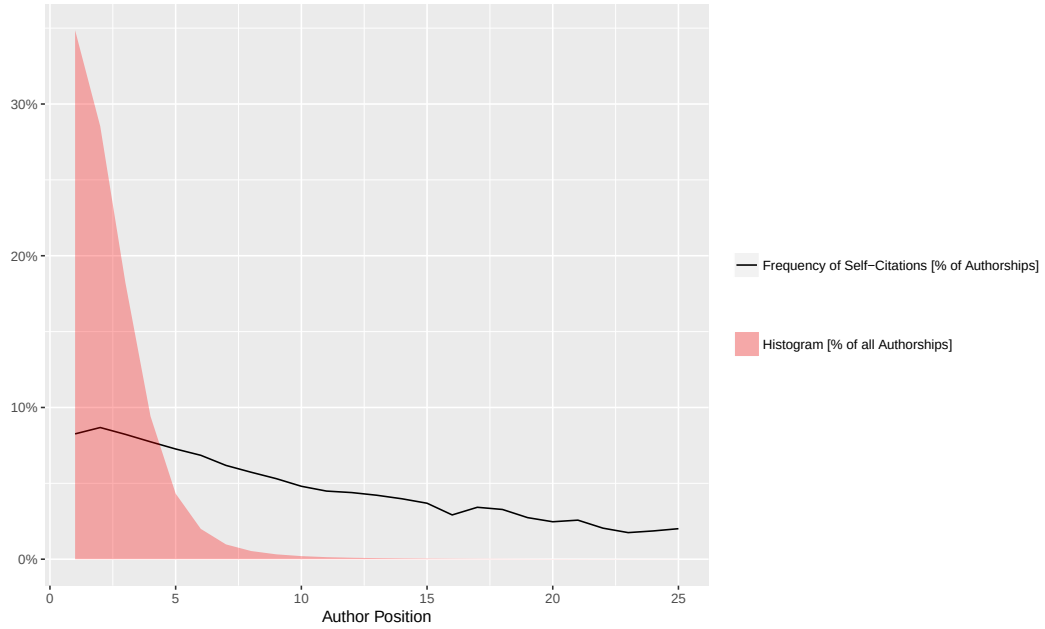

FIGURE 3.3: Author position plotted against self-citation frequency

Note: This figure, unlike the others in these chapters, does not depict the self-citation frequency of papers, but rather the frequency of authorship relations between a paper and an author who is self-cited by that paper.

3.2.2 Characteristic: Positions of Authors

The next investigated characteristic also pertains to authors. In figure 3.3 the position of an author is plotted against the frequency of self-citations to authors at that position. The general trend here is, that the further down the list of authors a given author is, the less likely they are to be targets of self-citations. There is a small increase in self-citation frequency at the second position compared to the first, which could be an indication of direct supervisors of papers receiving self-citation of their previous work.

After the second position, the frequency of self-citations to authors at that position steadily declines, with only a few outliers due to the variance caused by small sample sizes.

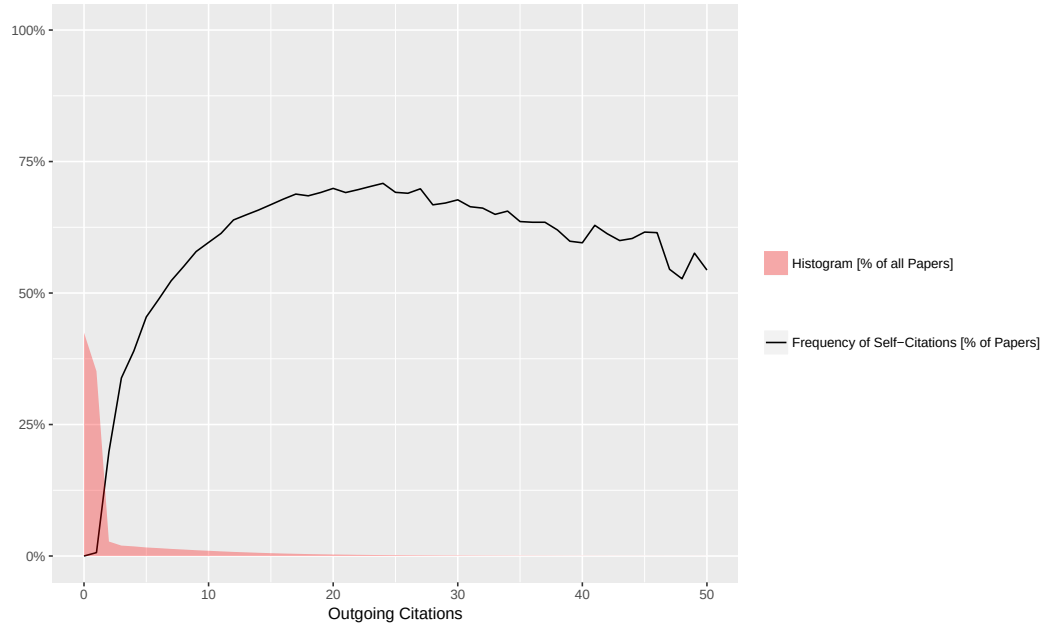

FIGURE 3.4: Outgoing citations plotted against self-citation frequency

3.2.3 Characteristic: Outgoing Citations

The number of outgoing citations, as in citations to other papers that a paper makes, show a very heavy correlation with self-citation frequency (See figure 3.4). While the majority of papers only cite a small number of other papers, as can be seen in the histogram, the few papers that cite more than 7 papers are more likely to have a self-citation among those citations than not. At around 20 outgoing citations a self-citation frequency of ~70% is reached. After that the self-citation frequency slowly declines and its variance increases due to smaller sample sizes.

What is also very interesting is that papers with only one outgoing citation are extremely unlikely to be self-citing. Only ~0.6% of all papers with one outgoing citation are self-citing. This chance increases by ~3109.6% to ~19.6% once a paper has two outgoing citations[2]. The vast majority of papers, however, have either one or no outgoing citations within the dataset.

[2] ~77.9% (2,548,945 papers of 3,272,991) have less than two outgoing citations

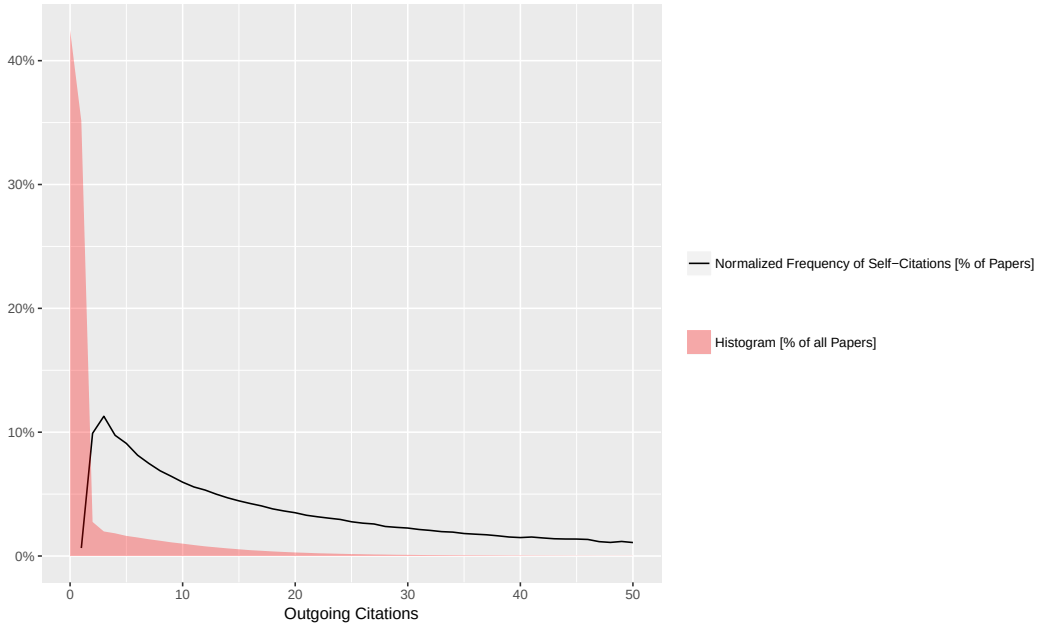

FIGURE 3.5: Outgoing citations plotted against self-citation frequency normalized for number of outgoing citations

These phenomena can be partially explained by the fact that self-citations are outgoing citations themselves. Therefore the number of outgoing citations directly impacts the number of citations that could be self-citations. Because of that, it is sensible to normalize this plot by the number of outgoing citations, to get a statement about how likely a single outgoing citation of a paper with a given number of outgoing citations is to be a self-citation (See figure 3.5).

From that figure one can gather that the chance for a singular citation to be a self-citation is maximal when it is in a paper with a total number of three outgoing citations. After that, the chances for singular citations to be self-citing constantly decrease.

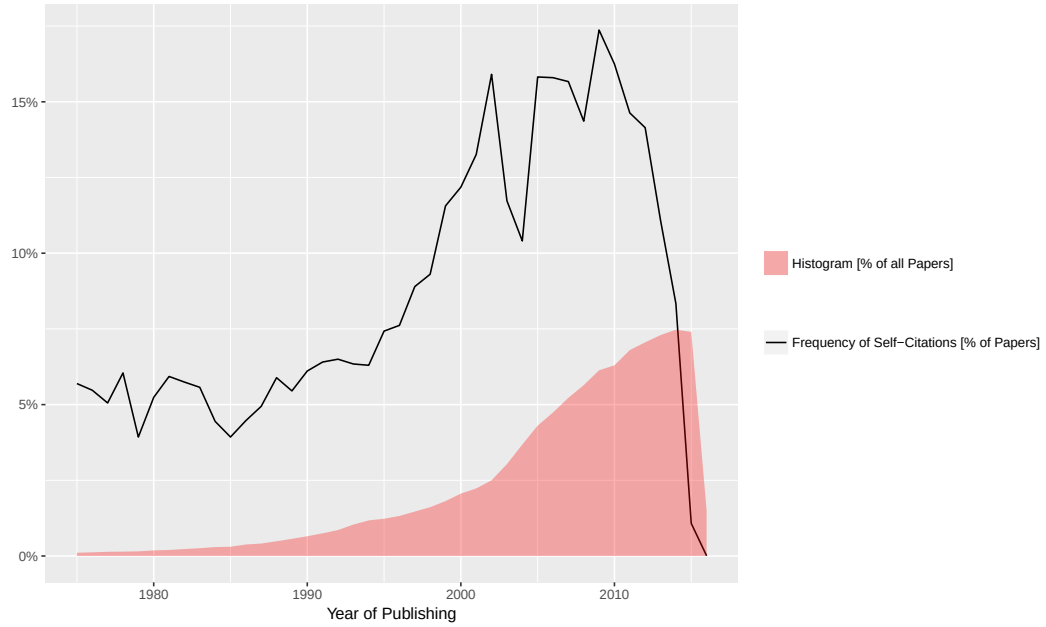

FIGURE 3.6: Publishing Year plotted against Self-Citation Frequency

3.2.4 Characteristic: Publishing Year

The figure 3.6 seems to show a drastic increase in self-citation rate after 1995, but this is actually misleading. The publishing year has a very unique interaction with another characteristic of papers, namely outgoing citations. Papers only ever cite papers that were published in the past, and even among those, recent papers are preferred. This means that due to the lower sample size of older papers, there are less outgoing citations in the older parts of the dataset. If one examines how the average number of outgoing citations changes over the years (See figure 3.7), it is very easily noticed that the topology of that plot is very similar to the one shown in figure 3.6. Since, as discovered in chapter 3.2.3, the number of outgoing citations very heavily correlates with self-citation rates, it is sensible to normalize for that characteristic (See figure 3.8).

Here the self-citation rate per outgoing citation is more or less constant, until it rapidly drops after 2013. It seems unlikely that the practice of self-citation nearly vanished in the last few years, which leads to believe that there is some intrinsic property to this dataset that leads to decreased numbers of outgoing citations in recent years. It proved difficult to identify exactly why this is, however one approach to check for this would be to repeat the methodology described in this thesis for

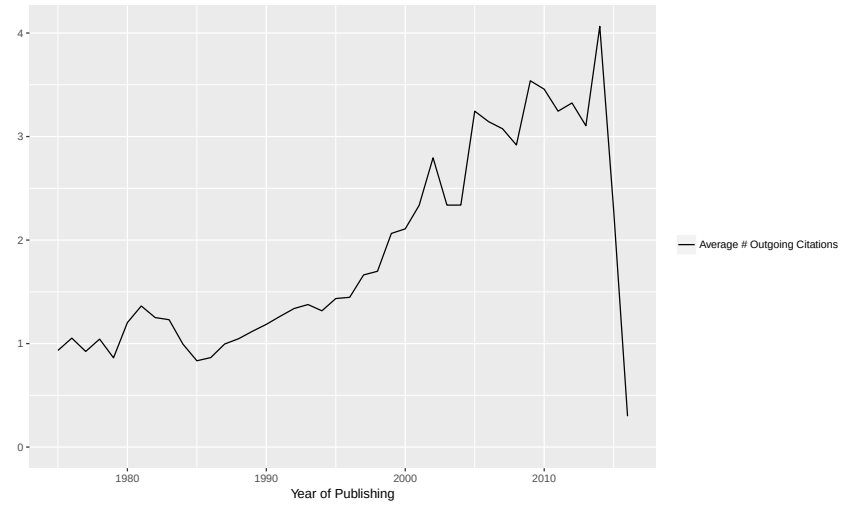

FIGURE 3.7: The average number of outgoing citations of papers plotted against publishing year

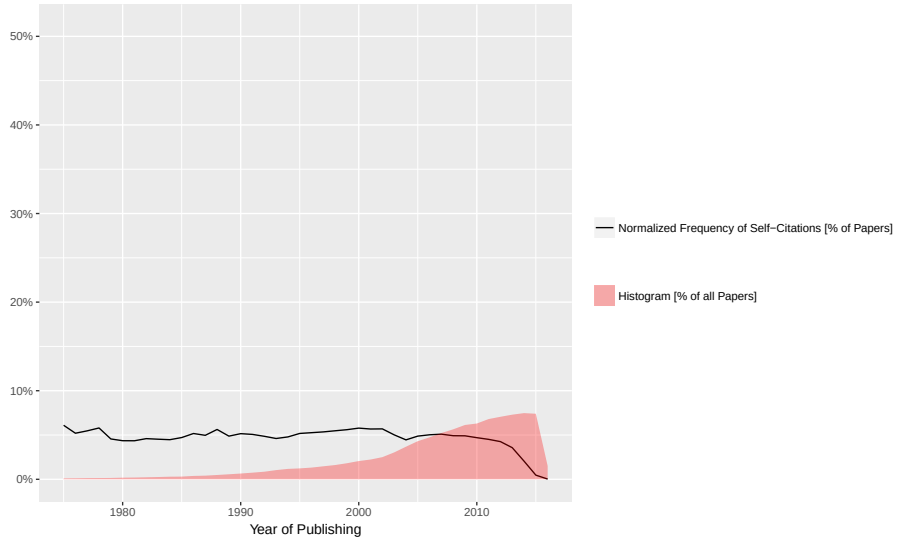

FIGURE 3.8: Publishing year plotted against self-citation frequency normalized for average number of outgoing citations

every Aminer dataset since the first one in 2010. If this drop off error persists and moves with the end of the graph, then this is, in fact, a problem with the dataset. Because of this uncertainty, there is no clear statement about correlation of publishing year and self-citation frequency to be made.

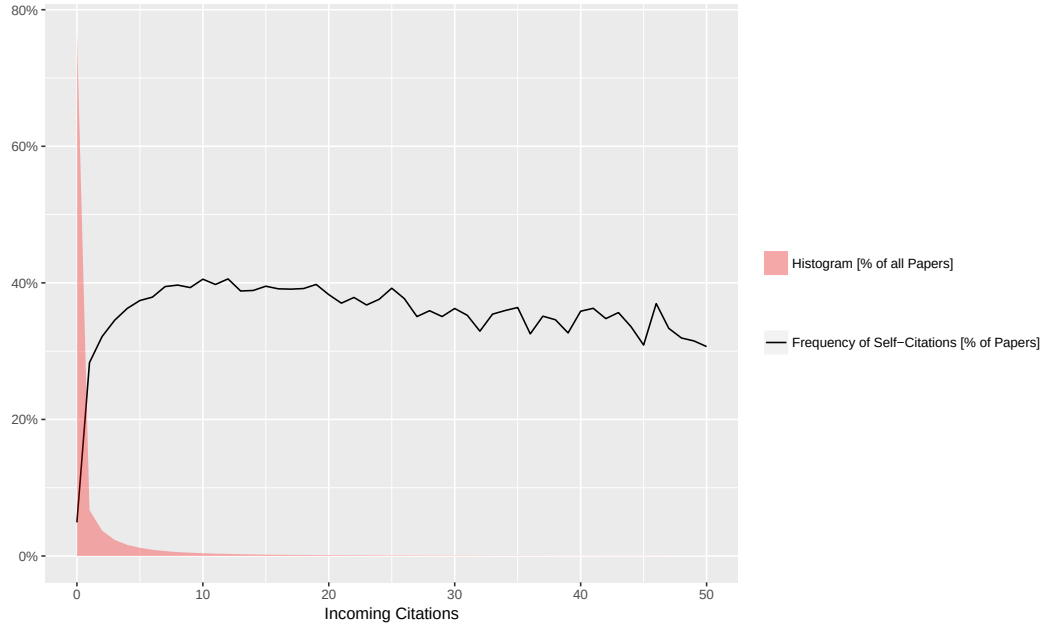

FIGURE 3.9: Incoming citations plotted against self-citation frequency

3.2.5 Characteristic: Incoming Citations

In figure 3.9 the correlation between incoming citations to a paper, meaning the number of citations that a paper receives within the dataset, and self-citation frequency is explored. As can be seen in the histogram, the vast majority of papers never receive a single citation. These papers show a low self-citation rate of ~4.9%. This increases to ~27.7% once papers receive a single citation. This trend of increasing frequency continues, until a maximum is reached at papers with ten incoming citations, which have a self-citation rate of ~40.6%, after which the rate stays relatively constant and only shows outliers when the sample size vanishes.

What is interesting to note, is that incoming citations have no effect on the writing process of a paper, since they are only ever performed after the papers have been already published. So this correlation seems to show a trend for papers that receive a lot of attention in the form of citations, to be self-citing. This could be because self-citing papers are often continuing the authors previous work, which implies that this previous work was fruitful and interesting enough to rectify a continuation. So it is reasonable to assume that these continuing, self-citing papers are also fruitful and interesting, meaning they are more likely to be cited.

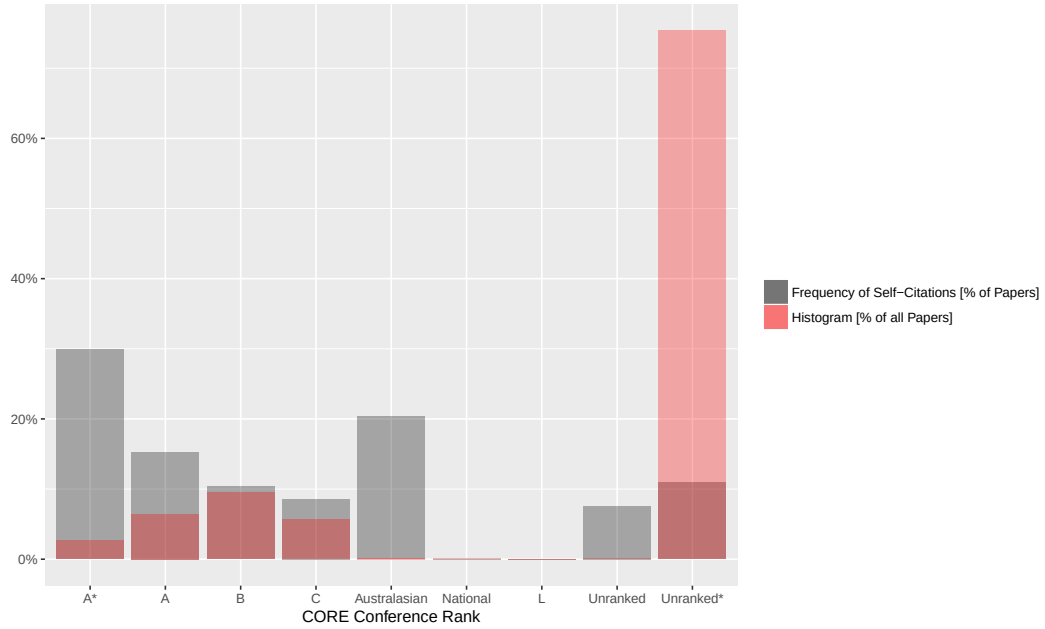

FIGURE 3.10: CORE conference rank plotted against self-citation frequency

Note: *Unranked** is different from *Unranked*. *Unranked* is a rank given out by CORE to conferences that appear in their database, but have no rank, whereas *Unranked** are all conferences that couldn't be assigned a CORE rank, due to them not appearing in that database.

3.2.6 Characteristic: CORE Conference Rank

For quantifying the conference property of papers, the CORE conference rank was used, as described in chapter 2.1.2, where a description of the ranks can also be found. When investigating how that rank correlates with self-citation frequency, it seems that the higher the rank of a conference is, the higher the rate of self-citation of papers within that conference is. The ranks *A**, *A*, *B* and *C* have strictly diminishing self-citation frequencies. *Australasian* conferences self-cite with a frequency between that of *A** and *A* conferences. Due to the minuscule sample size of the ranks *National* and *L*, not a single self-citing paper within these conferences was found. *Unranked* conferences seem to be less likely to self-cite than the *Unranked** conferences, which are conferences that could not be found in the CORE database.

This data strengthens the theory, which was established before, that prestigious papers are more likely to be self-citing .

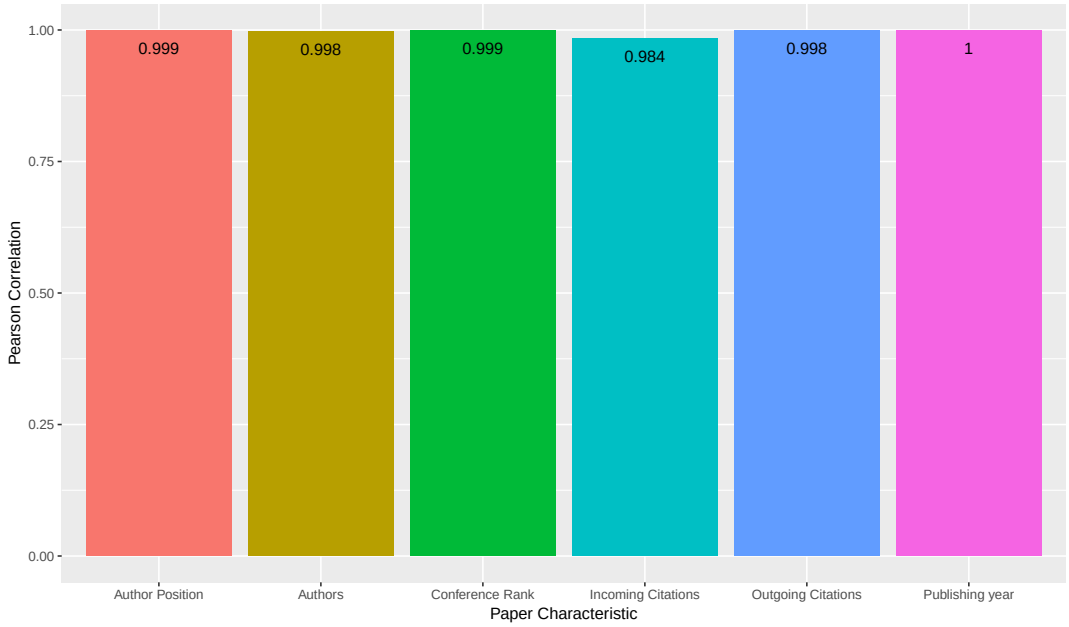

FIGURE 3.11: Pearson correlation coefficients for the self-citation rates with and without duplicate papers across characteristics

3.3 Effect of Duplicate indices

Due to the bug described in chapter 2.1.1, not all papers were evaluated in the previous part of this thesis, since the duplicate papers contain false information. It is, however still interesting to know whether and how the correlations would change if these duplicate index papers were to be included. For this purpose the Pearson correlation coefficients between the plots in chapter 3.2 and corresponding figures that include duplicate index papers were calculated (See figure 3.11). For all characteristics other than incoming citations, the coefficient is almost exactly one, meaning there is almost no change when including duplicate papers, since the plots are very linearly correlated. The outlier, the plot for incoming citations, has a still very high value of ~0.984. This is likely caused by the extremely skewed histogram, as seen in figure 3.9. When one only considers papers with ten or fewer incoming citations, which constitute ~95.5% of all papers, the Pearson correlation coefficient is ~0.999.

In summary, the duplicate index bug only heavily affects results that have a low sample size and are already uncertain because of that, to begin with.

3.4 Overarching Trends

When looking at all the data that was examined in this chapter, one notices a few trends.

It seems that the best indicator of whether a paper is self-citing or not, is the number of citations it is a part of. This is true for both incoming and outgoing citations, suggesting that connectivity within the citation graph is a high impact characteristic on whether a paper self-cites or not.

When examining author statistics, it is noticeable that being part of a large cooperation of authors on a paper is detrimental to a singular author's chance to be target of a self-citation. This is because even though the number of authors grows, the chance of a paper being self-citing does not grow significantly after two authors.

The most profound discovery, however, is that more prestigious papers are more likely to be self-citing that less prestigious ones. This is evident from the analysis on the characteristics of conference rank and the number of incoming citations.

Chapter 4

Conclusion

4.1 Summary

At the start of this thesis, the used datasets, the Aminer citation graph (Tang et al., 2008) and the CORE conference rank (CORE, 2017), were introduced and it was discussed why they were chosen. The Aminer dataset proved to be the most applicable dataset, because of its size and accuracy of paper characteristics, even though there were issues with using the dataset, namely author name ambiguity, which affects all datasets of this kind, and a bug within the Aminer software. The CORE conference rank was useful because of the fact that its ranks are manually assigned and not algorithmically computed.

After that, the pre-processing necessary to convert the Aminer citation graph into comma separated value files was explained. How this data was imported into the Neo4j database, which was chosen because of its underlying graph structure, improving performance, was explained afterwards, as was the identification and export process within the database. The export queries, however, had formatting which was difficult to use, so a reformatting process was described.

Once the acquisition of the results was completely explained, these results were evaluated, first via general statistics, showing that about ~10.5% of papers were self-citing, and then, for each characteristic of papers, the correlation between it and self-citation rates were examined. Noteworthy trends within these correlations are, that connectivity via citations is a strong indicator for self-citations, also that the number of authors only increases the self-citation rate drastically when going from one to two authors, and finally that more prestigious papers with high conference

ranks and high numbers of incoming citations are more likely to be self-citing than others.

4.2 Recommendations

The results presented in this thesis can be improved in accuracy in multiple ways in future research.
First of all, after the bug in the Aminer citation graph was reported, Dr. Jie Tang has stated that future citation graphs will not have this bug, so repeating this research with the debugged dataset is very sensible.
Also, the author name ambiguity problem can be attempted to be treated via various disambiguation methods, although the increase in accuracy is hard to measure. Or a correctly disambiguated dataset can be used, which seems very hard to acquire with the same size as the Aminer citation graph.
Since there are a lot of venues within the Aminer dataset which had no corresponding CORE rank, it may be sensible to combine multiple manually assigned rankings into one larger rank database. One could even compute algorithmic ranks for the venues, to see if there are different correlations to be discovered.

Of course, since the datasets used in this thesis are annually updated, the methodology presented in this thesis can be repeated whenever new datasets are available.

Bibliography

Aksnes, Dag (2003). "A macro study of self-citation". In: *Scientometrics* 56.2, pp. 235–246.

Bartneck, Christoph and Servaas Kokkelmans (2010). "Detecting h-index manipulation through self-citation analysis". In: *Scientometrics* 87.1, pp. 85–98.

Bonzi, Susan and Herbert Snyder (1991). "Motivations for citation: A comparison of self citation and citation to others". In: *Scientometrics* 21.2, pp. 245–254.

CORE (2017). *The CORE Conference DB*. http://www.core.edu.au/conference-portal. Retrieved May 1, 2017.

Fowler, James H and Dag W Aksnes (2007). "Does self-citation pay?" In: *Scientometrics* 72.3, pp. 427–437.

Rousseau, Ronald (1999). "Temporal differences in self-citation rates of scientific journals". In: *Scientometrics* 44.3, pp. 521–531.

Schreiber, Michael (2007). "Self-citation corrections for the Hirsch index". In: *EPL (Europhysics Letters)* 78.3, p. 30002.

Snyder, Herbert and Susan Bonzi (1998). "Patterns of self-citation across disciplines (1980-1989)". In: *Journal of Information Science* 24.6, pp. 431–435.

Tang, Jie et al. (2008). "ArnetMiner: Extraction and Mining of Academic Social Networks". In: *KDD'08*, pp. 990–998.

Zhivotovsky, Lev and Konstantin Krutovsky (2008). "Self-citation can inflate h-index". In: *Scientometrics* 77.2, pp. 373–375.