An analysis of ranking algorithms and their susceptibility to being manipulated on a citation network

Kevin Derrane

Submitted in Partial Fulfilment of the Requirements for the Degree of
Masters of Science in Computer Science (Data Analytics)

College of Engineering & Informatics
National University of Ireland, Galway

September 2019

Supervisor: Professor Michael G. Madden

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification.

_____

Kevin Derrane
12409118
September 2019

Acknowledgements

I would like to thank my research supervisor, Professor Michael Madden for his time, help and support throughout this project. I'd also like to give a special thanks to my year head Dr. Enda Howley and a special thanks to Dr. Michael Schukat for their support over the year. Finally, to my family, girlfriend, and friends, thank you for your support, encouragement, and guidance throughout these last few months.

Abstract

Citation analysis is an important tool used to evaluate researchers and their scientific work. The most common evaluation metrics used today are the *impact factor* for journals and the *h*-index for authors. In recent years a trend has emerged where these evaluation metrics are increasingly being used to determine whether or not a researcher gets considered for a job, gets a promotion, or even gets considered for a government grant. The issue here is that these evaluation metrics are easily manipulated by self-citations and the more serious recent emergence of citation cartels. On the one hand, self-citations are easy to spot but on the other hand, citation cartels are not. This research project introduces alternative approaches, which are based on Google's PageRank algorithm, to evaluate researchers and journals. A citation dataset composed by Valcav Belák, ArnetCite, was used. How these algorithms ranked papers compared to raw citation counts was first looked at. The robustness of these algorithms against author self-citations was then determined. After this, four of the lowest ranking papers in both algorithms were chosen and a citation cartel was formed by creating synthetic citation data with cartel features by modifying existing entries. The performance of the algorithms is measured in terms of how robust they are after their scores were recalculated when the cartel was created. The methodologies and the results of the algorithms are discussed, and future work and limitations are also provided.

Table of Contents

1 Introduction

This research project focuses on challenges in the areas of citation networks, visualisation of citation networks, transforming and cleaning of data, ranking algorithms, and their susceptibility to being manipulated.

1.1 Background

As described by Fister & Perc (2016), the number of citations a journal receives is one of the most important measures of its impact and influence within a scientific community.

Gross & Gross (1927, cited in Walters, W, 2017) first proposed the idea of using citation counts as an evaluation metric for the importance of academic journals. Due to the problem of capable library facilities and the increasing size and scope of academic fields, Gross & Gross (1927) saw the need for college libraries to accurately rank academic journals to help students decide which periodicals to take.

It wasn't again until the 1950's that citation counting entered the mainstream of scientific work and became a valued evaluation metric. Garfield (1955, cited in Varin et al, 2015) proposed the idea that the number of times a journal receives a citation should be normalised to the number of citable items in that journal. As described by Varin et al (2015), this idea was the first reference to what eventually became the *impact factor*, which is one of the most popular evaluation metrics for ranking scientific journals used today. Developed in the early 1960s by Garfield, the 'Science Citation Index' lead to the creation of the *impact factor*. Garfield (1963, cited in Garfield, 2006) proposed that when looking for new journals to add to the Science Citation Index, you could not solely depend on citation counts as small but important journals could be left out. The *impact factor* of a journal deals with this issue by calculating the average number of citations that the papers published in a journal have received over the previous two years (Else 2019). For example, an *impact factor* of five for a journal would mean that for a journal with five published papers in 2010, there would be five citations of these papers in 2012.

Since the 1960s a lot of research has been done on how to accurately measure the productivity and impact of authors. One of the most widely used metrics is the *h*-index. Developed by Hirsch (2005, cited in Bornmann & Hans-Dieter, 2008), the *h*-index attempts to measure the research output of an author. As described by Bornmann & Hans-Dieter (2008), the is *h*-index is based on citation counts and

is based on the number of papers an author has published and the number of times they have been cited in other publications. For example, an *h*-index of five would mean an author has published at least five papers and that each paper has been cited at least five times.

In recent years, citation cartels have emerged as a serious problem in scientific publishing. Fister & Perc (2016) describe citation cartels as groups of authors that cite each other excessively more than other authors in the same field. As a result, these published journals have higher *impact factor* scores and the authors higher *h*-index scores (Haley, 2017).

When the World Wide Web became publicly available in 1991, thousands of web pages were being created every day. Determining which web pages were important and which web pages appeared in a search result was not an easy task. These results could also be easily manipulated. Karch (2019) describes that early search engines linked to web pages based on keyword density, web pages with the highest keyword density were ranked first. Subsequently, this meant that websites could easily manipulate the system by repeating keyword phrases over and over to place higher on the search results. As a result, early search engines were unreliable. It wasn't until the release of the search engine Google in 1998 by Brin and Page, which used PageRank to rank the results, that search engines became reliable. As described by Strickland (2019), PageRank is a vote by all web pages to determine the importance of a web page, how important a web page is, depends on the number and quality of links a target web page receives.

The research presented in this research project focuses on ranking algorithms that are based on PageRank that have been adapted for use in a citation network. The focus of this research is based on the comparison of these ranking algorithm's and their susceptibility to being manipulated.

1.2 Motivation

Evaluating academic works is becoming an important aspect of the scientific community. With the birth of the World Wide Web in 1991, the way researchers have conducted their research has changed. Researchers can access, read and cite other publications and authors easier. They are no longer limited by their libraries resources.

As described by Fister & Perc (2016) and Else (2019), today the most widely used and important metrics to rank scientific papers and journals are still based around citation counts and the *impact factor*, which have been in use since the 1920s and 1960s respectively.

Evaluating researchers based on their *h*-index and the *impact factor* of their publications has become an increasing trend in academic institutions application processes. Some universities won't even consider candidates until they have met a certain threshold.

Making sure that a researcher, paper or journal is ranked fairly and accordingly is an important task. A paper of journal that has been ranked too highly could be taking the place of another paper or journal that would have been more deserving. A researcher that is ranked too highly might get offered a role that they are not suited for and take the position of a better yet lower-ranked candidate. Journals, papers, and authors that are ranked too low can have the opposite effect.

The emergence of citation cartels has posed a big threat to these evaluation metrics. Enago Academy (2018), describes that from 2013-2018, Thomas-Reuters, the organisation that determines the *impact factor* for journals has banned eighty-four journals for suspected citation stacking. Citation cartels can inflate the *impact factor* and *h*-index scores of journals and papers (Haley, 2017). They are essentially gaming the system. Both of these evaluation metrics are susceptible to these citation cartels.

Just like how the early search engines were manipulated to place websites that were less important first, this research project looks at the application of PageRank based algorithms to see if they are a good evaluation metric for scientific papers. It provides further insight into whether or not they are vulnerable to scientific articles with citation cartel like features. With the change in research methods and the evaluation of researchers, how scientific articles are evaluated should change too.

1.3 Problem Statement

The problem statement for this research project is concerned with the analysis of ranking algorithms and their susceptibility to being manipulated on a citation network. As described above, the current most widely used and important evaluation metrics for ranking scientific journals is the *impact factor* and the most used for authors is the *h*-index. These evaluation metrics do not take into consideration the recent emergence of citation cartels. This research focuses on PageRank based ranking algorithms for ranking papers and compares it to raw citation counts. It then looks at the effects the emergence of these citation cartels can have on these ranking algorithms.

Based on the points outlined above, the following research questions have been identified for this research project:

1. How do ranking algorithms such as PageRank and ArticleRank compare with each other and raw citation counts in terms of ranking papers on this dataset?

2. How robust are PageRank and ArticleRank to cartel-like behaviors?

1.4 Methodology

The methodology describes how the research questions will be answered. It can be broken up into two parts, one for each research question. The methodology involves the following:

Research question one:

- Transforming the data.
- Cleaning and understanding of the citation data and its relationships.
- Loading the citation data into Neo4J.
- Calculating the top papers based on their citation count.
- Calculating the PageRank & ArticleRank scores of the papers using the built-in Neo4J algorithms.

Research question two:

- Understanding what a citation cartel is and what its features are.
- Creating synthetic citation data with cartel features by modifying existing entries.
- Recalculating the PageRank & ArticleRank scores of these modified entries to see how robust they are.

1.5 Thesis Overview

This thesis consists of 6 chapters, they are the following:

1. Chapter 1 provides an introduction to the project, the motivation behind it, the research questions being explored, and how these research questions will be answered.

2. Chapter 2 reviews existing literature in the areas of networks, citation networks, the history of citation analysis, what citation counts measure, the impact of self-citing, citation cartels, ranking methods, and ranking algorithms.

3. Chapter 3 gives an understanding of the dataset and describes how the data was transformed, how the data was cleaned, what additions were made, and how the data was loaded into Neo4J.

4. Chapter 4 looks at experiments, methodology and the results relating to the first research question.

5. Chapter 5 looks at the creation of synthetic citation data with cartel features from existing entries. It explores the experiments, methodology and results relating to the second research question.

6. Chapter 6 reflects on the results from the previous two chapters with the research questions in mind and provides a recommendation of potential future work building on top of this thesis and any limitations encountered.

2 Literature Review

This chapter provides an understanding of what networks and citation networks are. It looks at the history of citation analysis and its use as an evaluation metric. It provides an insight into what citation counts measure and looks at the impact of self-citations and citation cartels. Finally, it provides a review of existing literature in the areas of ranking papers, journals, and researchers.

2.1 Networks and their Characteristics

As described by Barabasi (2016), complex systems are systems where it is difficult to derive their collective behaviour by just looking at their components. Examples of complex systems include our climate and the communication infrastructure that allows us to communicate with one another in a variety of different ways.

Barabasi (2016) states that in order to understand a complex system we need an understanding of the networks behind them and how its components interact with each other. Today, networks can be seen everywhere, from the world wide web to social networks such as Facebook and Twitter.

Barabasi (2016) describes a network as a collection of a systems components called nodes or vertices with the interactions between them, called links or edges. Networks are powerful in their ability to model the relationships between nodes. There are also hubs which are nodes with an exceptional number of links that exceeds the average. An example of a hub based on a sample Twitter follower dataset (which I created and plotted as part of a Web & Network Science assignment) can be seen below in Figure 2.1. A real-life example of hubs are celebrities on Twitter. Donald Trump, for example, has 63 million Twitter followers, whereas the average Twitter user has 707 as of 2016 (Kickfactory, 2016). There are two types of networks, directed and undirected. A directed network is where all the edges are directed from one node to another. An undirected network is where the edges have no direction. An example of each can be seen below in Figure 2.2 based on code taken from Ognyanova (2019). A real-life example of a directed graph could be someone following someone on Instagram with that person not necessary following them back. In the directed graph below, this can be seen as Sam following Rugile, Rugile following Ciara, and Ciara following Sam. They don't necessarily have to follow each other back. A real-life example of an undirected graph could be LinkedIn, where to follow someone they have to follow you back. In the directed graph below, this can be seen as John following Jamie and Jamie following John. They have a bidirectional relationship (Nykamp, 2019).

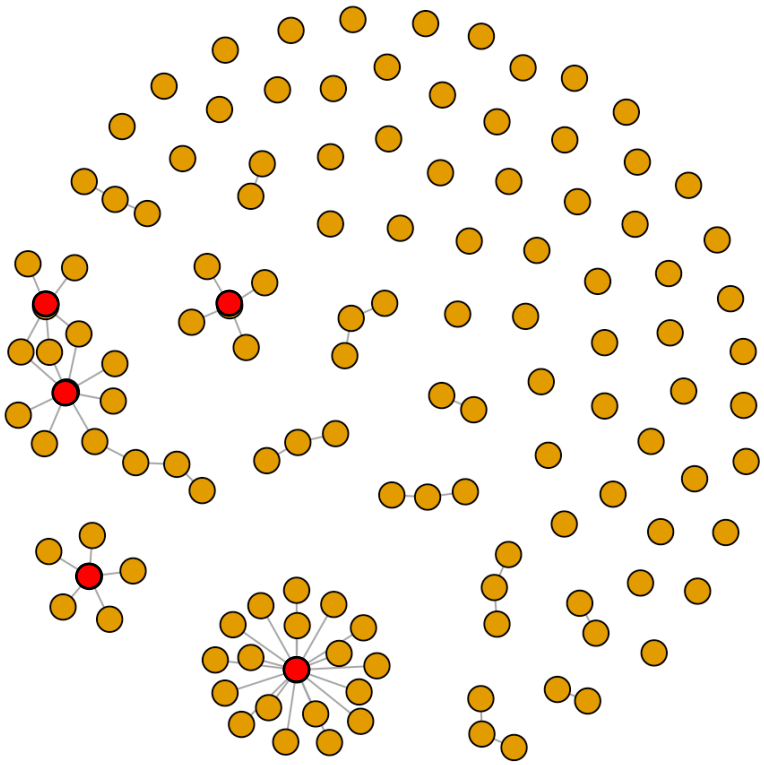

Figure 2.1. An example of hubs, (coloured red) on a sample Twitter follower dataset. These users (nodes) have followers (links) exceptionally greater than the average.

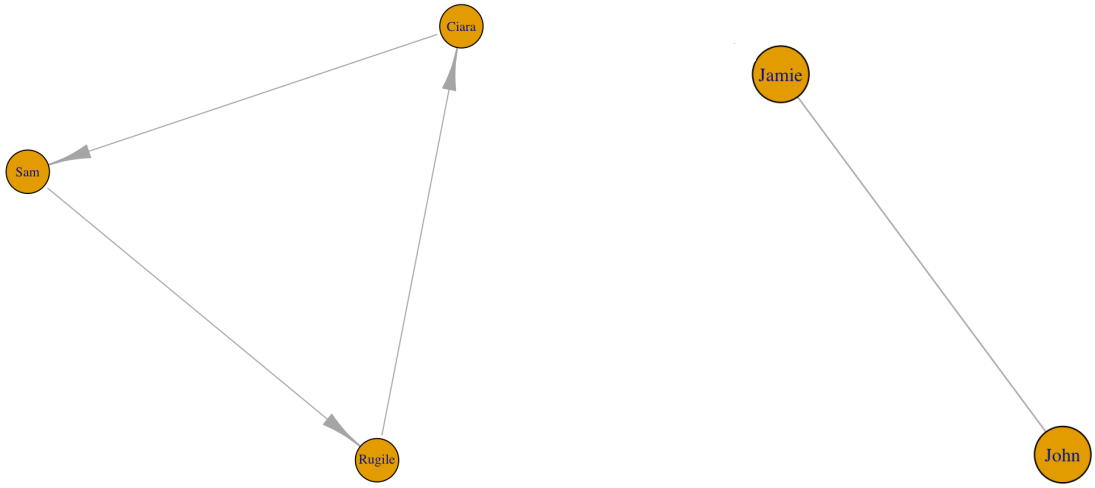

Figure 2.2. An example of an directed network (left) and an undirected network (right). Based on code taken from Ognyanova (2019).

2.1.2 Citations and Citation Networks

A citation provides a way to credit others when their work is used in research. It is a reference to the source of information used in research in the body of that research (LibGuides, 2019).

As described by Fister & Perc (2016), citation networks are the relationships between researchers and papers connected by citation relationships. Egghe & Rousseau (1990) cited in the Wikipedia article about Citation Networks describe a citation network as when a document d_i from a collection D, cites a document d_j, the relationship between the two can be shown by an arrow going from the node d_i to the node d_j, (Wikipedia Contributors, 2018). Here, the documents from the collection D form a directed graph which is called a citation network. In a citation network, the documents would be the nodes and the links between these nodes would be the citations.

An example of an Artificial Intelligence based citation network can be seen below in Figure 2.3 (which I created in Neo4J based on citation data for this research project, more information can be seen below in Chapter 3). Here the scientific papers are the nodes and the links between the papers are the citations. The relationship between these nodes can be seen on the edges.

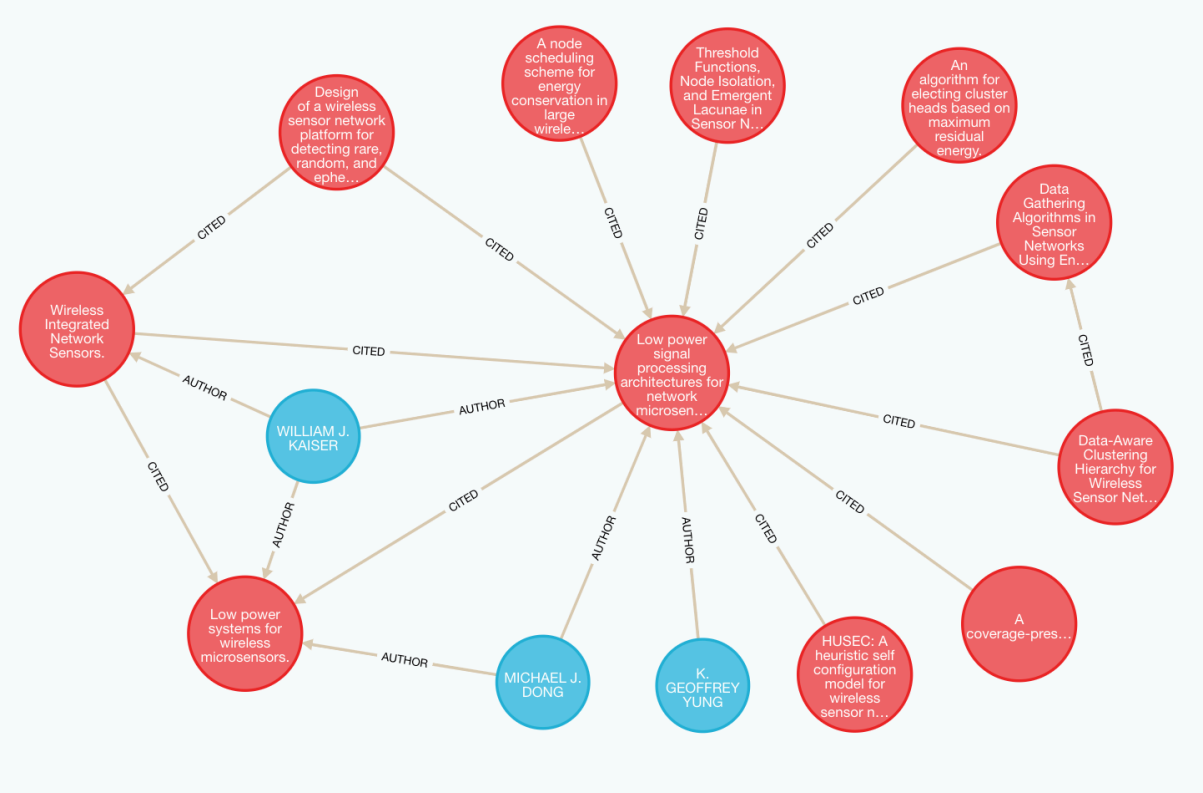

Figure 2.3. An example citation network in Neo4J.

2.2 The History of Citation Analysis

As described by Moed (2010), citation analysis is the creation and application of a series of indicators of the impact, influence, and quality of scientific work derived from citation data. Chikate & Patil (2008) describe citation analysis as to reference in one scientific article to another scientific article, with information on how that scientific article can be found.

The first record of citation analysis and its use as an evaluation metric dates back to 1927. In 1927, a paper published by Gross & Gross (1927, cited in Walters, W, 2017) was the first to mention and use citation counts as an evaluation metric for the importance of scientific journals. Gross & Gross (1927), saw the need for smaller libraries in colleges like their own, to identify important scientific journals to cover a wide range of periodicals for its students. With finite resources, smaller libraries had to make sure they choose the appropriate journals to acquire.

Gross & Gross (1927) first looked at 'The Journal of the American Chemical Society' and tabulated a list of its references. From this, they proposed the idea that the journals that were most frequently cited would be the most valuable for their library to purchase. Their method was based on a couple of key principles; the use of a periodical is in direct proportion to the number of times its articles are cited, and the journals they chose were representative of the entire field (Smith, 2009).

Brodman (1944, cited in Smith, D, 2009) was the first to suggest the shortcomings of the Gross & Gross method. While examining methods for choosing physiology journals, she began to question and noticed shortcomings in the key principles of the Gross & Gross method. Brodman (1944) concluded that although the Gross & Gross method had helped help libraries build up journal collections for many diverse fields, it appeared to be unscientific and unscholarly and gave untrustworthy results. Despite her conclusion, she failed to mention any alternative or any solution to the shortcomings of the Gross & Gross method. As she described, it had been accepted uncritically for so long based on the assumption that any method is better than no method.

It wasn't until the mid-fifties that a new method was proposed to deal with the shortcomings of the Gross & Gross method. This method was the *impact factor* which eventually came to fruition in the late sixties (Garfield, 2006). Despite being over fifty years old, the *impact factor* is still the most widely used evaluation metric today. It is further described below in Section 2.6.1.

2.3 What Citation Counts Measure

Given the increasing importance of evaluating scientific journals and papers, what exactly citation counts measure is an important issue (Bornmann & Hans-Dieter, 2006). Aksnes et al (2019) state that citations are assumed to measure the impact and or quality of research. As described by Smith (1984, cited in Bornmann & Hans-Dieter, 2006), citation counts are an attractive record for evaluating scientific journals, as they are easy to obtain measures and they do not depend on the cooperation of a respondent and do not bias the response.

Garfield (1979) looks at the idea of using citation counts to measure the quality or impact of scientific work. He found that, rather than citation counts being a good measure of the quality of scientific work, they are a utility. A highly cited paper is a paper that has been found useful rather than important by a large number of people. Garfield (1979) proposed the issue with the number of citations papers from Albert Einstein and Oliver H. Lowry received. The fact that Einstein's paper about his unified field theory received significantly fewer citations than Lowry's paper on protein determination does not indicate that Lowry's contribution is more significant than Einstein's. It just showed that more people were interested in Lowry's paper than Einstein's. Garfield (1979) concluded that the only claim citation counts have is that they are an aid in evaluating authors and they provide a measure of the utility or impact of a paper. They do not measure the of a quality of a paper. He also questions the impact and role self-citing had, this is further described below in Section 2.4.

Lindsey (1988) looks at the idea that citation counts have widely become an accepted measure to judge the quality of a scientific contribution. Cicourel & Franzen (1965, cited in Lindsey, 1988) describe issues with this approach, they remarked that you may too often be measuring what is measurable rather than what is quality. Lindsey (1988) says that using citation counts to judge the quality of scientific work derives from the notion that if an author's work is of quality it will be used by others. She remarks that the strongest evidence in support of citations as a measure of quality lies in the various validation studies that have indicated that citation counts correlate quite highly with measures of quality. She concludes and argues that we may just use citation counts to judge the quality of a paper because they are the most convenient and requires the least amount of work to determine what is a quality paper.

There is also the question of what defines a high quality and or impactful paper (Lindsey, 1988). Is a paper that has won multiple awards a high quality and impactful paper? Who decides what a high-quality paper is? A paper can be of high quality but have a low impact. Conversely, a paper can have a high impact but be a low quality paper.

2.4 Impact of Self-Citing

Self-citing is the term usually used to describe the act of when a paper *A* is cited in paper *B*, and paper *A* and *B* have at least one distinct author in common (Uwe, 2017). Self-citing also occurs at the journal and university level. Self-citing at the author level will be discussed below.

The argument of whether or not self-citations should be included in citation analysis has been going on since the early days and is still ongoing today (Shema, 2012). Some researchers state that they have their place as it shows that researchers are building upon their own work. Other researchers believe that they are used to manipulate citation rates. Garfield (1979) in his paper, 'Is citation analysis a legitimate evaluation tool?' mentions self-citations as one of the criticisms when using citation counts to measure the quality or impact of scientific work.

Aksnes (2003, cited in Shema, 2012) studied more than 45,000 publications by Norwegian authors and found that 36% of all citations represent author self-citations. Aksnes concluded that self-citations should preferably be removed before making comparisons (Aksnes, 2003). Fowler & Aksnes (2007), in their paper, 'Does self-citation pay?' did a further study on the impact of self-citing. They discovered that for each self-citation, an additional 3.65 citations can be garnered after 10 years. This meant that an additional 40% of total citations could be created indirectly from self-citations. They concluded that a few self-citations could be the difference between funding and promotional decisions.

Carley et al (2012), state that the role self-citing plays in measuring research output is considerable. She mentions that the citing of one's work can have a significant influence on a wide range of evaluation metrics including, *impact factor,* total citation counts, and *h*-index. Bartneck & Kokkelmans (2010) showed that authors could easily inflate their *h*-index by strategically citing their own publications.

Mishra et al (2018) reported that men self-cite >50% more often than women across a wide variety of fields. If self-citing is positively manipulating the different evaluation metrics for male authors, then female authors are being adversely affected.

Although the act of self-citing has had a considerable impact on the evaluation of scientific articles, it is easy to detect. In recent years there emerged another gaming tactic that is way more harmful and difficult to detect, as discussed next in Section 2.5 (Davis, 2012).

2.5 Citation Stacking and Citation Cartels

Citation stacking is the process of concentrating bursts of citations from one journal to another (Van Noorden, 2013). The authors and journals who practice this citation stacking form the citation cartels (Enago Academy, 2019). The concept of a citation cartel was first unmasked by Franck (1999, cited in Fister & Perc, 2016). Here, cartels were described as the event of authors and journals working closely together for mutual benefit. Citation cartels just like self-citations are used to boost the *impact factor* of journals and the *h*-index of authors (Haley, 2017).

As shown by Fister & Perc (2016), citation cartels can be represented on a multilayer graph of a citation network. This can be seen below in Figure 2.4 taken from Fister & Perc (2016). They show how an original paper citation network, described on the bottom layer, can be represented as a direct graph with the nodes representing the papers and the edges representing the relationship between these papers, i.e. the citations. The author citation network can be seen in the middle layer. Here the nodes represent the authors and the edges represent the relationships between these, based on whether or not an author cites another author. The top layer is the citation cartel, derived from the middle layer. Authors $A2$ and $A3$ can be seen citing each other and their papers more often compared to the other authors. Fister & Perc (2016) describe the cartel as where the number of inter-citations between two nodes needs to exceed some threshold value in a clique of order two. They concluded that the discovery of citation cartels in networks is just like the discovery of communities in networks, except that members of a citation cartel do their best to stay undetected.

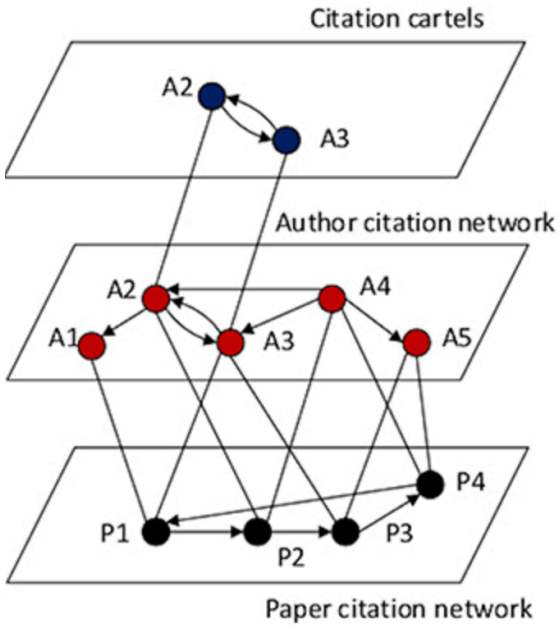

Figure 2.4. A citation network represented in a multilayer network. Fister & Perc (2016).

From 2009 to June 2013, Mauricio Rocha-e-Silva and his team engaged in the process of citation stacking to elevate their 2011 journals *impact factor* (Van Noorden, 2013). It wasn't until the 19th of June 2013 that the pattern was discovered. Thomson-Reuters was able to detect the presence thanks to a new algorithm which they developed in 2012 due to a tip-off regarding the emergence of citation cartels. They banned the journal along with four other Brazilian journals. In total 14 journals were banned that year (Van Noorden, 2013). The process the Brazilians used can be seen below in Figure 2.5 taken from Van Noorden (2013). Here, there the 2011 journals referring to papers from 2009-2010 can be seen. From this, the journal's 2011 *impact factors* was derived. This raised their *impact factor* score.

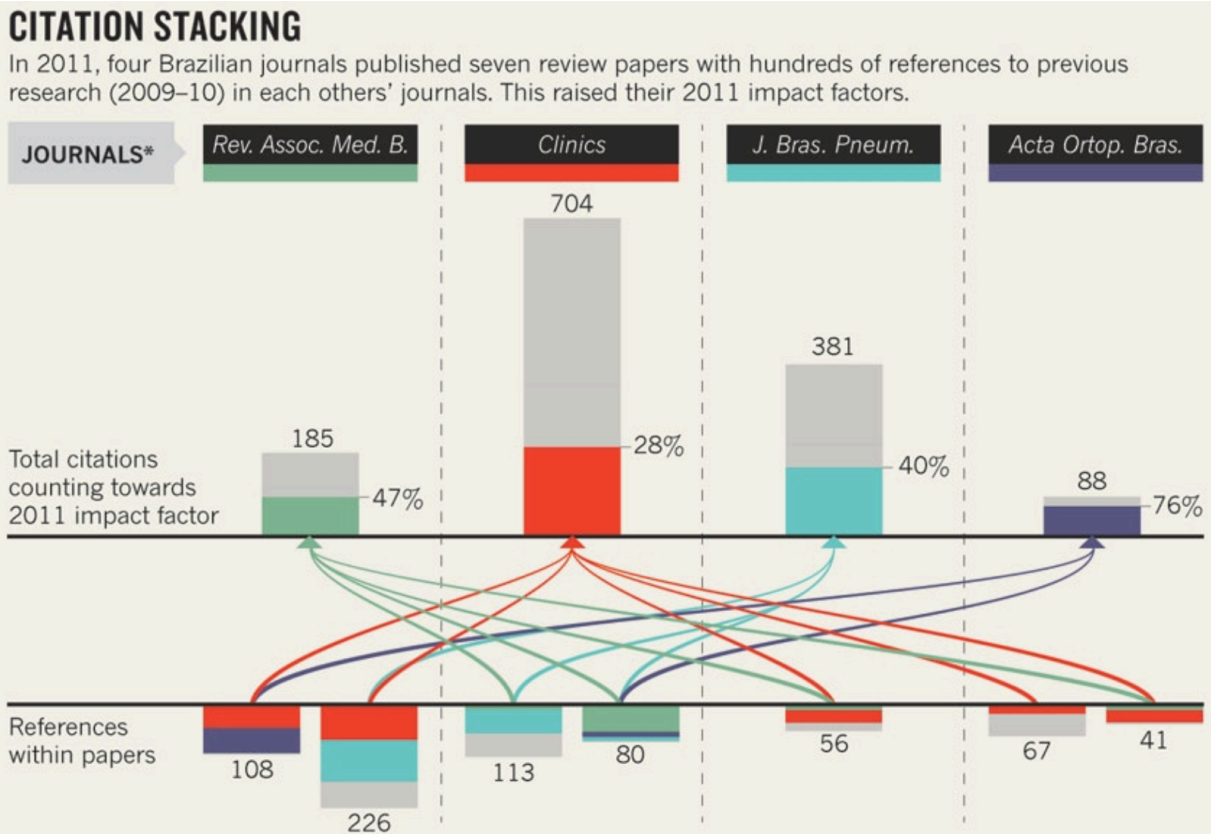

Figure 2.5. The Brazilian citation stacking scheme. Van Noorden (2013).

Citation cartels leave a huge impression on the *impact factor*. As described by Chorus & Waltman (2016, cited in Enago Academy, 2018), citation cartels are damaging the validity of the *impact factor*. Davis (2012) looked at the impact citation cartels had on a journals *impact factor*. He found that when removing two review articles that attributed more than 500 citations from a suspected citation cartels journal, the journals *impact factor* dropped from 6.204 to 4.082.

A study done by Haley (2017) showed the effects of citation cartels on an author's *h*-index. He demonstrated an example citation process that could be implemented by colluding authors. The initial *h*-index for this author was 8 and the goal was to increase it to 9. Haley (2017) found that this could be easily accomplished with a 1.7% increase in total citation counts, granted it was done in a specific way for a 12.5% or 1 increase in *h*-index.

Citation cartels by their very nature are hard to detect (Davis, 2016). In 2016 Thomas-Reuters suspended two journals that they accused of forming a citation cartel. The cartel between these two journals can be seen below in Figure 2.6 taken from Davis (2016). The journals are the Methods of Information in Medicine (red) and the Applied Clinical Informatics (blue). Here, four main papers stand out, two papers by Lehman (red) and two papers by Haux (blue). The author surname and year the paper was published is shown on each paper. On the surface, these papers look like they could just be hubs, but on closer inspection, you can notice that each one of these papers cites a large number of papers in the others journal within the two years prior. This is within the timeframe to calculate the journals *impact factor*. The information provided and visualization alone are not enough to guarantee that these authors are engaged in a citation cartel. Further knowledge is required to be sure. However, it gives a good representation of what a citation cartel looks like (Davis, 2016).

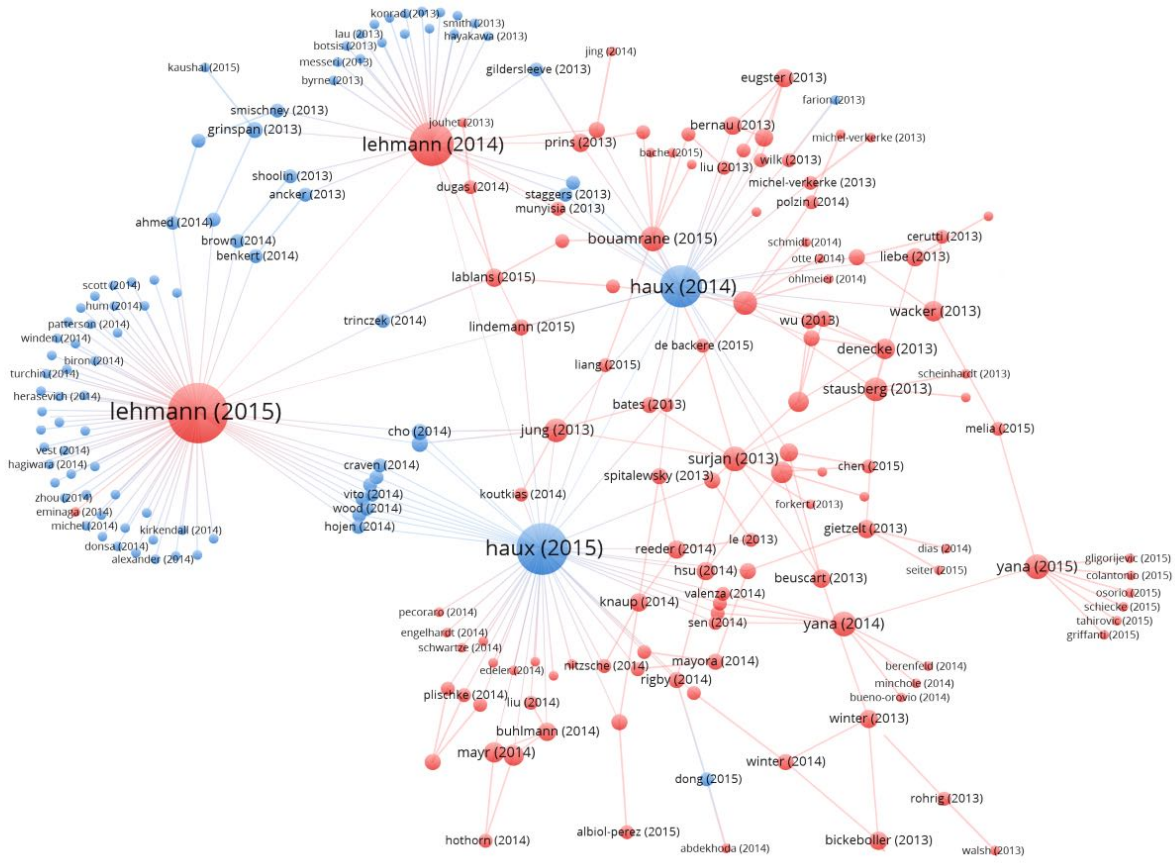

Figure 2.6. Visual representation of a citation cartel. Davis (2016).

2.6 Ranking Methods

Ranking can be defined as a list of things that are ordered according to quality, impact, etc (Merriam Webster 2019). Today, ranking publications and authors is more important than ever. Ranking publications fairly and correctly is a major issue. Katerattanakul et al (2003) state, that ranking journals is of particular interest to researchers since publications in a prestigious journal can have a significant influence on tenure, promotional decisions, and compensation increases. Researchers should also take an interest in ensuring that their articles are ranked highly on the different academic search engines such as Microsoft Academic Search, Google Scholar, and PubMed (Beel et al 2010). Korobkin (1999) describes the human fascination that we just want to know who is on top and who is not.

As mentioned above, the most commonly used ranking methods for journals are citation counts and the *impact factor*. The most commonly used method for authors is the *h*-index. Citation counts have been described in detail above in Sections 2.1.2, 2.3, and 2.4.

2.6.1 Impact Factor

As described by the Journal Citation Reports, the *impact factor* of a journal is the measure of the frequency in which a 'citable item' in a journal has been cited in a particular year or period (Web of Science Group, 1994). As mentioned above in Section 1.1, Garfield (2006) created the *impact factor* to alleviate the issues of selecting journals just based on their citation count. He found that smaller important journals may not be left out.

As mentioned above in Sections 2.4 and 2.5, the *impact factor* does not take into consideration self-citations and it is vulnerable to citation cartels. Seglen (1997) in his paper 'Why the impact factor of journals should not be used for evaluating research', describes many other issues associated with the *impact factor*. These issues include: the citation rate of a journal determines its *impact factor* and not the other way around, citations to 'non-citable items' are often included, and that a journals *impact factor* is determined by mechanics unrelated to the scientific quality of their articles. He concludes that high *impact factors* are more likely in journals covering large areas due to rapidly expanding but short-lived papers.

The PLoS Medicine Editors (2006) describe other issues associated with the *impact factor*. They describe the issue of how the *impact factor* has grown beyond its control and is used in many inappropriate ways. The *impact factor* has been used to decide whether or not authors get promoted, offered a position in a

department, or are awarded a grant. In some countries, government funding for institutions is decided based on its publications having high *impact factors*.

The formula for calculating a journals *impact factor* can be seen below in Equation 1 taken from Phdontrack (2019). The *impact factor* (IF) of a journal is based on two elements; the numerator which is the ratio of the number of citations (A) in the current year, to articles published in the previous two years, and the denominator, which is the number of citable articles (B) published in the same two years (Phdontrack 2019).

$$IF\ for\ year\ X = \frac{Citations\ in\ X\ to\ articles\ published\ in\ X-1\ and\ X-2}{Articles\ published\ in\ X-1\ and\ X-2}$$

Equation 1

Saha et al (2003) give the following example to work out a journals impact factor: Z = Total citations in 2010 to articles published in Journal X, A= 2010 citations to items published in Journal X in 2000-2001(subset of Z), and B = Number of citable articles published in Journal X in 2000-2001. IF =A/B.

2.6.2 *h*-index

The *h*-index is used to quantify a researchers scientific output (UIC 2018). As described by Gann (2018) the *h*-index is a number used to represent both the productivity and the impact of a researcher (Gann 2018). Kreiner (2016) describes how the *h*-index has become regarded as a 'magic tool', that is used to measure what is unmeasurable, the quality of science.

As mentioned above in Sections 2.4 and 2.5, the *h*-index is susceptible to both self-citations and citation cartels. There are also many other issues present. Kreiner (2016) describes the problem of comparing researchers during different stages of their careers. Researchers with shorter academic careers are at a disadvantage as the *h*-index does not decrease over time. If a researcher was to stop contributing their score would never go down. Bornmann & Hans-Dieter (2008) conclude that if the *h*-index is to be used for the evaluation of a researchers scientific output, it should always be taken into account that it is dependent on the length of a researchers career and the field of study in which the papers are published and cited.

There are many different variants of the *h*-index such as the *g*-index and *i10*-index. These all aim to fix the shortcomings of the *h*-index. However, Bornmann

& Hans-Dieter (2008) states that research on different variants should stop and instead test the validity of the existing h-index. Once the validity of the h-index has been confirmed, it can be used to assess scientific work.

The formula for calculating a researchers h-index can be seen below in Equation 2 taken from Bornmann & Hans-Dieter (2018). The h-index of a researcher is calculated based on the total number of citations and the total number of publications of their work. The higher the citation count, and the higher the publication count, the higher the h-index The lower the citation count, and lower the publication count, the lower the h-index.

$$\text{A scientist has index } h \text{ if } h \text{ of his or her N}p \text{ papers have at least } h \text{ citations each and the other (N}p - h) \text{ papers have fewer than } \leq h \text{ citations each}$$

Equation 2

As described by Bornmann & Hans-Dieter (2008), the h-index captures only a part of the publication and citation data if the distribution is right-skewed. Researchers with very different citation frequencies can, therefore, have the same h-index. This issue can be seen in Figure 2.7 below taken from Bornmann & Hans-Dieter (2008). For example, in the A graph, this researcher has a high number of citations and publications. In graph B, this researcher has a low citation count but high publication count. Therefore, they both have the same h-index score and they both have the scientific output, even though one researcher may be contributing more than another.

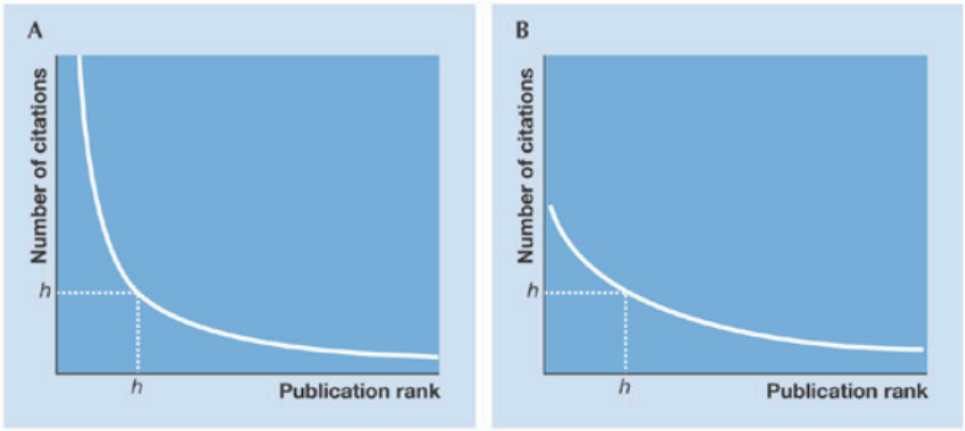

Figure 2.7. Graph of h-index for distributions of citation frequencies (the publications are sorted in graphs A–B by number of citations). Bornmann & Hans-Dieter (2008).

2.6.3 Ranking Algorithms

As described by Walker et al (2007), many information networks have become hard to navigate due to their large size and rapid growth. Navigation is especially tough without any sort of ranking scheme. Walker et al (2007) give the example of the World Wide Web, a network of pages connected by hyperlinks. A successful solution to this problem was the use of Google's PageRank algorithm.

2.6.3.1 PageRank

As described by Page et al (1998), the importance of a web page is a subjective matter, which depends on the reader's interests, knowledge, and attitudes. The PageRank algorithm was created by Brin and Page in 1998 to compute a ranking for every web page on the web. PageRank calculates the importance of a web page by looking at the quantity and quality of other pages that link to it. The underlying assumption is that pages that have a low out-degree are more important than pages with a higher out-degree (Neo4J(1), 2019). An example of this can be seen below in Figure 2.8 taken from Roberts (2006). Here web page C has a higher PageRank compared to web page E, despite having fewer links. Since web page C is linked by only one web page, page A, and page A has a high PageRank score, web page C will have a high score. Web page E is linked by many web pages that have a low score and its score will be lower as a result.

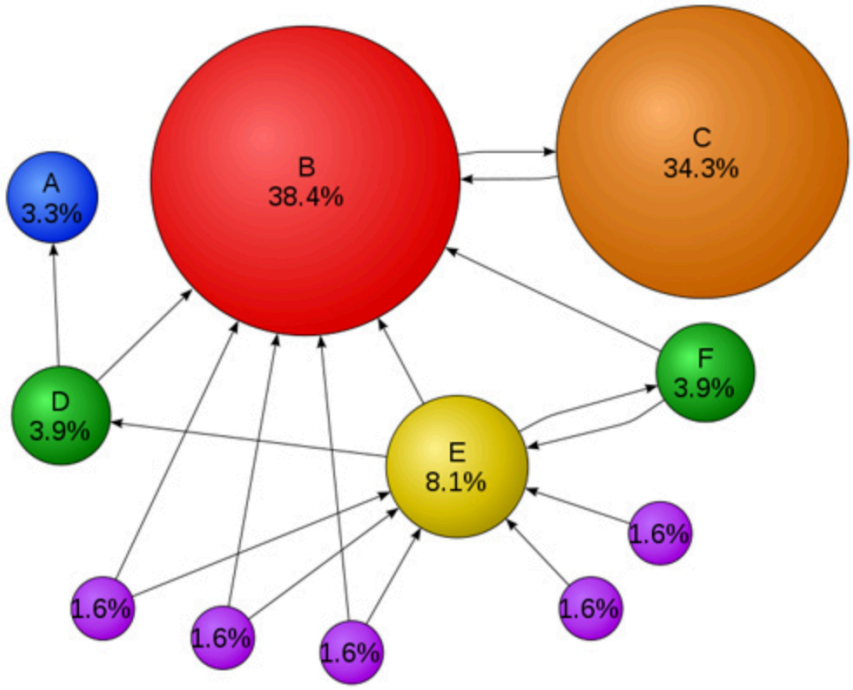

Figure 2.8. Graph showing how PageRank works. Roberts (2006).

Page et al (1998) describe the creation of the PageRank algorithm and liken it to citation analysis. Page describes it as a citation importance ranking tool. Their idea was inspired by the way scientists gauge the importance of scientific papers, by looking at the number of other papers referencing them (Soulo, 2018).

Early search engines at the time relied on either keyword density or the number of hyperlinks (citations). However, just like citation counts, these search engines were easily manipulated. Page et al (1998) explain that there are many cases where citation counts do not correspond to our common sense of importance.

The formula for calculating the PageRank score of a web page can be seen below in Equation 3 taken from Brin & Page (1998). They define PageRank as the following: we assume page A has pages T1 to Tn which point to it (i.e they are citations). d is the dampening factor which is set between 0-1 but is generally 0.85. The dampening factor is the probability that a random user will continue clicking on links as they browse. The probability of a user clicking a link on the first page you visit is high but on the next page, it is lower and so on. This is assumed to decrease with each click. As a result, the total score is manipulated by the dampening factor. C(A) is defined as the number of links going out of page A. The PageRank of a page A is as follows:

$$\mathrm{PR(A)} = (1\text{-}d) + d\,(\mathrm{PR(T1)/C(T1)} + \ldots + \mathrm{PR(T}n\mathrm{)/C(T}n\mathrm{)})$$

Equation 3

As described by Roberts (2006), PageRank calculates the score for the web in seven steps.

1. Start with a sample set of webpages in a graph.

2. Crawl the graph to determine their link structure.

3. Assign each page an initial PageRank of 1 / N.

4. Update the PageRank of each page by adding up the weight of every page that links to it divided by the number of links coming out from the referring page.

5. If a page has no outward links, redistribute is PageRank equally among the other pages in the graph.

6. Repeat this process until the PageRank stabilises.

7. The dampening factor is added at each stage to reproduce the fact that a user will eventually stop searching.

PageRank has been used in citation analysis. As described by Walker et al (2007), current methods for ranking publications based on citations are rather crude. They treat all citations as equal and ignore differences in the importance of the citing papers. They describe that one of the advantages of the PageRank algorithm is that it implicitly accounts for the importance of the citing article in a self-consistent fashion. Bollen et al (2006) applied the PageRank algorithm for ranking publications and found that the PageRank algorithm can be used to obtain a metric that reflects importance. As described by Chen et al (cited in Walker et al, 2007), they successfully applied the PageRank algorithm to papers published in the 'American Physical Society Journals' and found it helped to discover a set of highly influential papers. Ma et al (2008) state that one of the advantages of using PageRank for citation analysis is that it could largely eliminate the flattery of academic influence caused by self-citations. They also found that PageRank is an excellent way to rank the influence of publications suffering from low citation counts. Walker et al (2007) conclude that there are significant differences between the World Wide Web and citation networks, enough to suggest a modification to the PageRank algorithm.

2.6.3.2 ArticleRank

ArticleRank is an algorithm derived from Google's PageRank algorithm to measure the influence of scientific articles as an alternative to citation counts (Li & Willet, 1998). As described by Neo4J(1) (2019), ArticleRank differs slightly to PageRank. PageRank assumes that the relationships from nodes that have a low out-degree are more important than the relationships from nodes with a higher out-degree. ArticleRank weakens this assumption.

ArticleRank provides an easy way of classifying articles with equal numbers of citations and boosts the ranking of articles that are cited by articles that have a considerable impact in their own right (Li & Willet, 1998). They conclude that ArticleRank is an interesting alternative to citation counts but it requires significantly more computing power the larger the number of papers.

The formula for calculating the PageRank score of a web page can be seen below in Equation 4 taken from Li & Willet (1998). They define ArticleRank as the following: we assume page A has pages T1 to Tn which point to it (i.e they are citations). d is the dampening factor which is set between 0-1 but is generally 0.85. C(A) is defined as the number of links going out of page A. C(AVG) is described as the average number of links going out of all pages. The ArticleRank of a page A is as follows:

$$\mathrm{AR(A)} = (1\text{-}d) + d(\mathrm{AR(T1)}/(\mathrm{C(T1)} + \mathrm{C(AVG)}) + \ldots + \mathrm{AR(T}n\mathrm{)}/(\mathrm{C(T}n\mathrm{)} + \mathrm{C(AVG)})$$

Equation 4

2.7 Chapter Summary

This chapter provides a review of literature in the areas of networks, citations and citation networks, the history of citation analysis, what citations measure, the impacts of self-citations and citation cartels, the various ranking methods such as the h-index for authors and the impact factor for journals, and the algorithms PageRank and ArticleRank.

It shows that when evaluating papers, citation counts should only be used as an aid to provide a measure of the utility and how useful a paper is. It should not be used to measure the quality of a paper.

It follows on with the impact that author self-citing has had on citation counts and the other ranking methods, the h-index for authors and *impact factor* for journals. Self-citations are one of the primary reasons the Gross & Gross methods is shunned upon.

In recent years a new practice has emerged that is much more harmful and difficult to detect compared to self-citations. Citation cartels are groups of authors working closely together with other elect groups of authors to boost their h-index and journals *impact factor*.

The two most commonly used evaluation metrics today are the *impact factor* for journals and h-index for authors, two methods that do not take into account the recent emergence of citation cartels and both of which are easily manipulated by citation cartels.

PageRank, which was created by Brin & Page in 1998 to compute the ranking for every web page on the internet as an alternative to early search engines. Early search engines were easily gamed by web pages through techniques such as keyword density. ArticleRank a variation of PageRank used for ranking articles on a citation network is then explored. ArticleRank weakens the main assumption of PageRank, this assumption is that the quality of a web page is dependent on the quality and not the number of its links, ArticleRank weakens this and it is more dependent on the quantity of the links.

3 Data Analysis & Neo4J

This chapter looks at the process of transforming the dataset (see Section 3.1) to an easier to use format. It explores the tables and each of their fields and looks at the relationships between the tables. It describes any data cleaning and data additions that have been performed. It looks at how the data can be loaded into Neo4J and what graph schema is used. The analysis, understanding, visualisation, and cleaning of this dataset are key in helping to answer both research questions.

3.1 Transforming the Data

The ArnetCite dataset for this project was composed by Valcav Belák. ArnetCite is citation dataset that is focused on publications from venues associated with the field of Artificial Intelligence. It is based on merged records from Arnet-Miner and CiteSeerX, both of which are digital libraries for scientific papers and articles (Belák & Hayes, 2015).

This dataset was provided in the format of a PostgreSQL back up file, arnet.psql. To get the data in a readable format for analysis, the backup had to be restored. This process can be seen below in Figure 3.1.

```
admins-MacBook-Pro:~ postgres$ psql restored_database < /private/tmp/arnet.psql
Password for user postgres:
```

Figure 3.1. Restoring the arnet.psql file.

Once the data was restored back into PostgreSQL, the necessary tables were downloaded in CSV format. The tables deemed necessary were Author, AuthorShip, Citation, and Paper. A snapshot of the Paper table in Microsoft Excel can be seen below in Figure 3.2. The tables and their relationships are further described below in Sections 3.2 and 3.2.1.

	A	B	C	D	E	F	G	H
1	title	summary	year	citation_count	arnet_id	id	venue	csx_id
2	Silent Data Corruption - Myth or reality?		2008	3	162048	162043	1294	NULL
3	A recurrence-relation-based reward model for perfor		2008	2	162055	162050	1294	NULL
4	Foundations of Measurement Theory Ap	Increasing in	2007	26	162058	162053	1294	NULL
5	Automatic security assessment of critical cyber-infras		2008	8	162049	162044	1294	10152784
6	Byzantine replication under attack.		2008	41	162053	162048	1294	9045592
7	Automated duplicate detection for bug tracking syste		2008	75	162051	162046	1294	9232760
8	Scheduling for performance and availability in system		2008	10	162057	162052	1294	9215026
9	Security through redundant data diversity.		2008	24	162059	162054	1294	9208961
10	Confidence: Its Role in Dependability Ca	Society is inc	2007	27	162056	162051	1294	3770136

Figure 3.2. Snapshot of the Paper table.

3.2 Data Understanding

The tables and the fields composed by Belák are as follows:

- **Author:**
  - ***id***: A unique numbered identifier for each author.
  - ***name***: The name of the author.
- **AuthorShip:**
  - ***paperid***: A unique numbered identifier for each paper.
  - ***authorid***: A unique numbered identifier for each author.
- **Citation:**
  - ***id***: A unique numbered identifier for each citation pairing.
  - ***citing***: The paper in question.
  - ***cited***: The paper that the citing paper cites.
- **Paper:**
  - ***id***: A unique numbered identifier for each paper.
  - ***title***: The title of the paper.
  - ***summary***: A summary/abstract of the paper.
  - ***year***: The year in which the paper was published. This ranges from 2013 back to 1936.
  - ***citation_count***: The number of times the paper has been cited by other papers outside of this dataset. An issue was encountered with this column while running the PageRank and ArticleRank algorithms. The top 10 papers were returned with high scores but some had 0 citation counts. This shouldn't be possible because these scores are dependent on the citation counts. On further inspection, it showed that these papers with 0 citation counts had been cited within the dataset itself but the ***citation_count*** equalled 0. The assumption here is that these citation counts must have been sourced from somewhere else, more than likely the CiteSeerX database. These citation counts are based on the number of times a paper is cited in their database. Since this dataset is a merger of two databases, the ***citation_count*** column cannot be a representation of the number of times a paper has been cited within this dataset. It's more of a global citation count rather than a citation count representative of this dataset. As a result, a new citation count column will be created based on the number of times the papers are cited in the dataset. Gross & Gross looked at the number of times a paper was cited in a journal, not the number of times they were cited globally. This process is described further below in Section 3.4.
  - ***arnet_id***: A unique arnet identifier for each paper
  - ***venue***: A unique numbered identifier for each venue.
  - ***csx_id***: A unique identifier for csx.

3.2.1 Data Relationships

The relationships between the tables is very important to understand. They give a good understanding of how the tables are linked together and what fields are dependent on each other. The entity relationship diagram below in Figure 3.3 (which I created in DBeaver), gives a good visualisation of the relationships between these tables.

The relationships are as follows:

- **Author to Authorship:**
  - Author ***id*** links to ***authorid*** in the Authorship table.

- **Authorship to Paper:**
  - Authorship ***paperid*** links to ***id*** in the Paper table.

- **Citation to Paper:**
  - Citation ***citing*** and ***cited*** link to ***id*** in the Paper table. The ***id*** in the Citation table is unused.

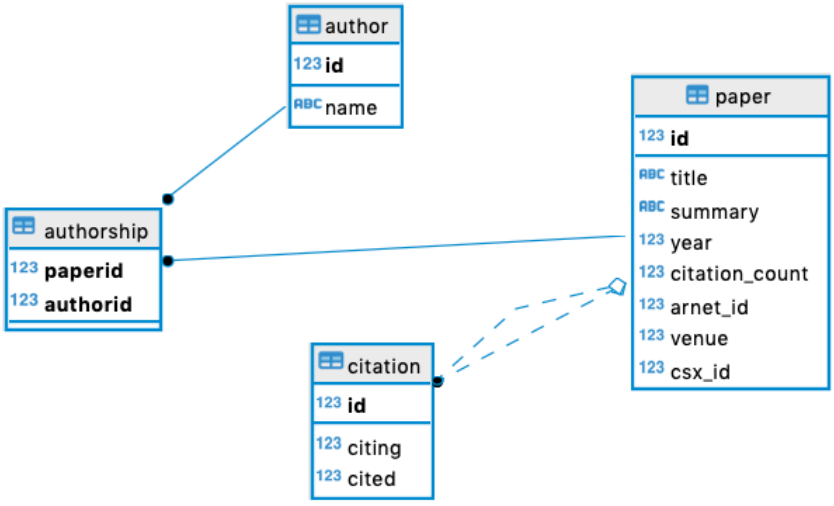

Figure 3.3. An entity relationship diagram of the tables.

3.2.2 Data Statistics

In total, we have 2,212,589 unique Papers, 1,264,926 unique authors, and 3,389,536 unique citations.

3.3 Data Cleaning

After gaining an understanding of the datasets fields and relationships, the data needed to be cleaned. The data was cleaned using the r programming language. Since the data is the combination of records from Arnet-Miner and CiteSeerX, there was going to be some issues. One issue encountered regarding the *citation_count* column has been discussed above in Section 3.2. Another issue that was encountered was due to the size of the dataset, there are over 2 million unique papers and the Paper table alone is also over 700mb in size. The loading and calculating of the algorithm scores on a dataset of this size would require a lot of processing power and time. There have been two main cleaning processes carried out in R Studio on the dataset.

They are outlined below:

1. **Removal of unnecessary columns:**
   - **Paper table:**
     - ***summary***: The ***summary*** column was removed as a large portion of the papers had no summary. These summary column also attributed most to the size of this file. By just removing the summary column the file size went from 740mb to just 107mb. A reduction of 85%.
     - ***arnet_id***: The ***arnet_id*** was not used and it was of no importance. It is related to the Arnet-Miner website and how they create and label there papers in the citation datasets.
     - ***venue***: The ***venue*** column was removed as the venue table itself was not used.
     - ***csx_id:*** Just like the ***arnet_id*** column, the ***csx_id*** was unused and irrelevant.

2. **Removal of special characters:**
   - **Paper table:**
     - ***title***: One issue faced when loading the data into Neo4J was an error with special character such as ‘ and “. As a result, they were removed from the ***title*** column.

After completing the above steps, the size of the tables combined went from nearly 1GB to 260mb. A reduction of almost 75%. This is quite a significant decrease. The data is now ready to be loaded into Neo4J for analysis.

3.4 Data Additions

As described above in Section 3.2, there is an issue with the *citation_count* column. As a result, two additional columns were added to the Paper table.

They are outlined below:

- **Addition of two new columns for the Paper table:**
  - **Paper table:**
    - ***num_citations***: This column is the number of times a paper has been cited by another paper in this dataset only. If a paper has been cited 10 times by other papers in this dataset, then the ***num_citations*** will be 10. It is a better representation of the papers in this dataset compared to the ***citation_count*** column. This column will be helpful when comparing our algorithms in the next chapter below.
    - ***num_papers_cited***: This column is the number of times a paper cites other papers. For example, if a paper cites 20 other papers within this dataset then ***num_papers_cited*** will be 20. This column will be helpful when identifying and modifying papers to have 'cartel' features in the next chapter below.

3.5 Loading Data into Neo4J & Defining Graph Schema

Neo4J is a graph-based database management system. It uses cypher, a declarative graph query language which is of similar structure to SQL. There are two different ways to import CSV files into Neo4J. One is using the LOAD CSV function and the other is to use the import tool. The LOAD CSV function was originally used but it was very time-consuming. It was taking over three-quarters of an hour to load just the relationships. As a result, the import tool was used.

Before loading the data in Neo4J using the import tool, the columns had to be renamed and the relationships between the tables described. An example of the Paper table, cleaned with renamed columns can be seen below in Figure 3.4 based on code taken from Neo4J(2) (2019). Note the columns *year* described as type integer and the *id* column described as the ID and tagged as the Paper-ID.

```
title,year:int,citation_count:int,id:ID(Paper-ID)
Silent Data Corruption - Myth or reality?,2008,3,162043
A recurrence-relation-based reward model for performability evaluation of embedded systems.,2008,2,162050
Foundations of Measurement Theory Applied to the Evaluation of Dependability Attributes.,2007,26,162053
Automatic security assessment of critical cyber-infrastructures.,2008,8,162044
Byzantine replication under attack.,2008,41,162048
Automated duplicate detection for bug tracking systems.,2008,75,162046
Scheduling for performance and availability in systems with temporal dependent workloads.,2008,10,162052
Security through redundant data diversity.,2008,24,162054
Confidence: Its Role in Dependability Cases for Risk Assessment.,2007,27,162051
```

Figure 3.4. Snapshot of the Paper table with renamed columns. Neo4J(2) (2019).

The command used to load this data and their relationships can be seen below in Figure 3.5 based on commands taken from Neo4J(2) (2019). Note the – *nodes:* function, this means that there will be two types of nodes in the graph, nodes for the Author and nodes for the Papers. The relationships between these are defined by the *–relationship:* function, this means that on the relationship (edge) between these papers and authors (nodes), their relationship will be displayed. This can be seen in the graph schema defined below in Figure 3.7.

```
bash-3.2$ bin/neo4j-admin import --nodes:Author=import/
author.csv --nodes:Paper=import/papercleaned.csv --rela
tionships:CITED=import/citation.csv --relationships:AUT
HOR=import/authorship.csv --ignore-missing-nodes
```

Figure 3.5. Loading the csv files and their relationships into the database. Neo4J(2) (2019).

The result from the command in Figure 3.5 above can be seen below in Figure 3.6. The import tool was very quick to import the data.

```
IMPORT DONE in 19s 321ms.
Imported:
  3477515 nodes
  9483342 relationships
  15805386 properties
```

Figure 3.6. Time taken to import the data and relationships. Neo4J(2) (2019).

The graph schema for this dataset is outlined below in Figure 3.7 (which I created in Neo4J based on the citation data above using the db.schema() command). Here the Papers and Authors are the nodes. The relationships between the Paper and Author nodes is whether or not the author wrote the paper. The relationship between the Paper nodes is whether or not a paper cites or is cited by another paper.

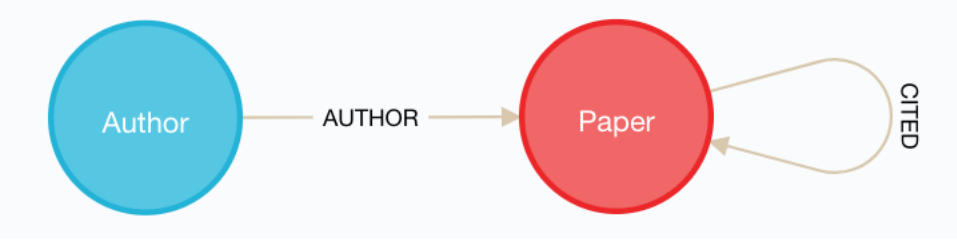

Figure 3.7. Graph Schema.

In total there were 3,477,515 nodes with two labels (Author & Paper), and 9,483,342 relationships with two labels (AUTHOR & CITED), loaded into the Neo4J graph database.

3.6 Chapter Summary

In this chapter, the dataset for this research project was explored. First, the data was restored and transformed from is PostgreSQL format to CSV files. The data was then explored and the relationships between each table understood. An issue with the c*itation_count* column was noted and rectified. The data was then thoroughly cleaned using the r programming language. In total it was reduced by almost 75% or over 750mb. To fix the issue with the *citation_count* column, two new columns were added to the Paper table, these were *num_citations* which is the number of times a paper is cited within this dataset and, *num_papers_cited* which is the number of times a paper cites another paper in this dataset. The data was then loaded into Neo4J using its built-in import tool and the graph schema was defined for the relationships between the papers and the authors.

4 Experimentation: Comparison of raw Citation Counts, PageRank, and ArticleRank

This chapter looks at the experiments carried out to compare how raw citation counts, PageRank, and ArticleRank, rank papers within this dataset. These experiments aim to observe how the top 10 papers change between these three evaluation metrics. Some of the methodology for these experiments have already been carried out, they are described in detail above in Chapter 3.

These experiments explore the first research question which is outlined below:

1. How do ranking algorithms such as PageRank and ArticleRank compare with each other and raw citation counts in terms of ranking papers on this dataset?

4.1 Why is this Important?

As discussed above in Chapter 2, Section 2.4, self-citations are one of the primary methods used to boost citation counts and therefore boosting a journals *impact factor* and or an authors *h*-index. Since PageRank and ArticleRank are based more on the quality and not just the quantity of the citations, it could be assumed that they are somewhat robust to the art of self-citing. This work in this chapter checks the influence citation counts have on these ranking algorithms while comparing the ranking of these algorithms to the raw citation counts. If these ranking algorithms can be resilient to this practice, then they could become a viable evaluation alternative to the commonly used *impact factor* for journals and the *h*-index for authors, all the while without being influenced by self-citations

4.2 Experiment 1. Citation Counts

The first experiment carried out was to rank the top 10 most important papers based on their citation count within this dataset. Citation counts are one of the easiest and most convenient ways to rank papers. Although what they measure is still up for debate, citation counts at a minimum provide a measure of the utility and the usefulness of a paper. The use of citation counts, what they measure, and their limitations are described in detail above in Sections 2.1.2 and 2.3.

4.2.1 Methodology

- After the data was loaded as described above in Section 3.5, an index was created on the Paper *title* column. In Neo4J, a database index is a sample copy of some of the data in the database to make the searches and ranking of titles quicker and more efficient (Neo4J(3) 2019).
- In order to return the top 10 most important papers by citation counts, a cypher query was written. This query returned the top 10 papers based on the *num_citations* column outlined above in Section 3.4. These papers were returned along with their title, the year in which they were published, and their citation count.

4.2.2 Results

The results for this cypher query can be seen below in Table 4.1. Here the top 10 papers are shown in descending order based on their citation count.

Table 4.1. Top 10 papers based on the number of citations.

Title	Year	Num Citations	Num Cited
Fast Algorithms for Mining Association Rules in Large Databases.	1994	1593	7
Mining Association Rules between Sets of Items in Large Databases.	1993	1374	6
Abstract Interpretation: A Unified Lattice Model for Static Analysis of Programs by Construction or Approximation of Fixpoints.	1977	1205	5
Distinctive Image Features from Scale-Invariant Keypoints.	2004	1204	15
Graph-Based Algorithms for Boolean Function Manipulation.	1986	1181	0
The Anatomy of a Large-Scale Hypertextual Web Search Engine.	1998	1177	0
Chord: A scalable peer-to-peer lookup service for internet applications.	2001	1097	7
Time, Clocks, and the Ordering of Events in a Distributed System.	1978	1071	0
Ad-hoc On-Demand Distance Vector Routing.	1999	1012	2
A decision-theoretic generalization of on-line learning and an application to boosting.	1995	960	8

The highest-ranked paper, “Fast Algorithms for Mining Association Rules in Large Databases” can be seen below in Figure 4.1. Each paper is a red-coloured node and the number of citations that these papers have is displayed in the middle. A sample of 100 papers (Neo4J limitation) that share a relationship (cited this paper or cited by this paper) with this paper is displayed below.

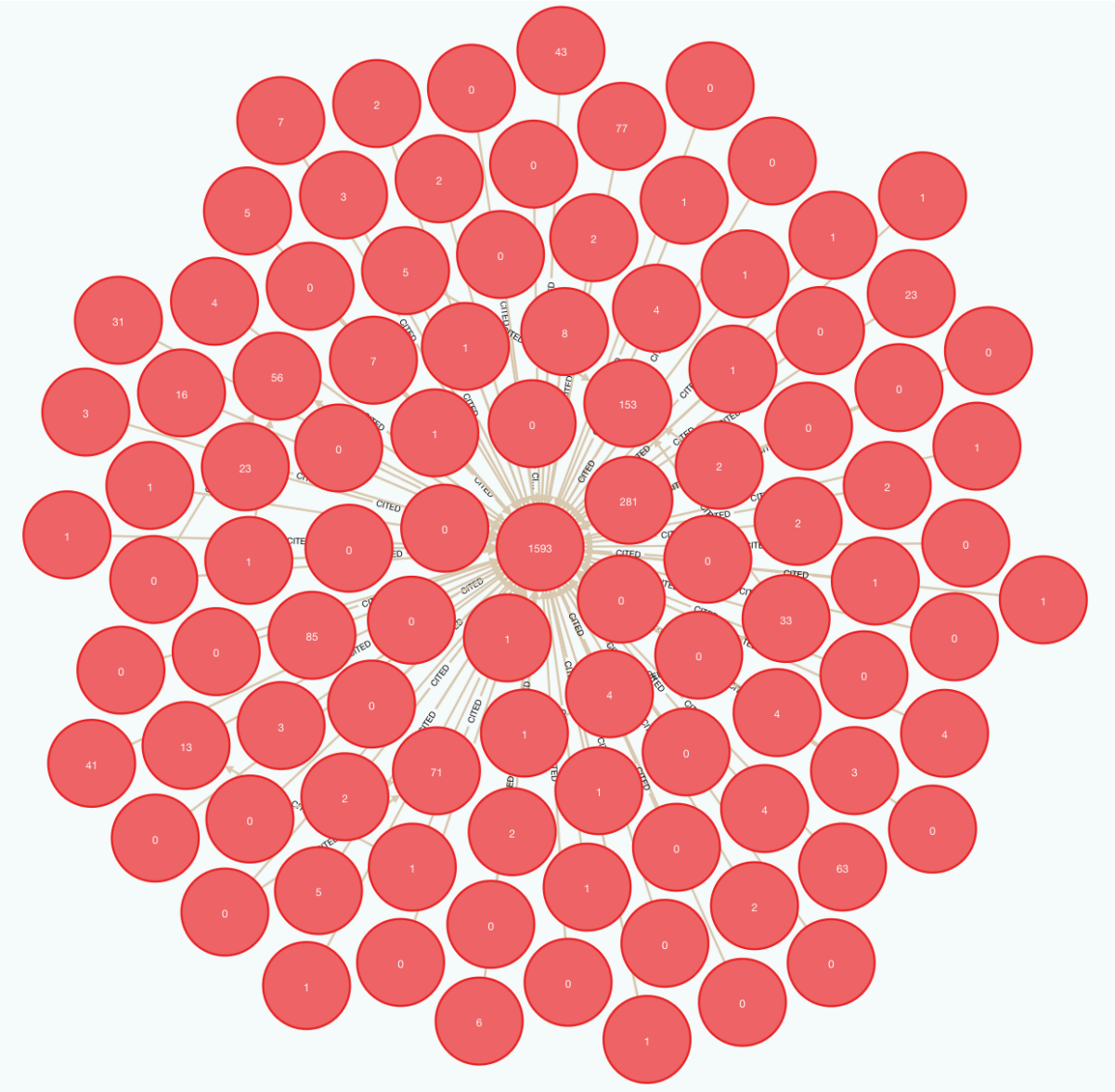

Figure 4.1. Snapshot of the highest ranked citation count paper “Fast Algorithms for Mining Association Rules in Large Databases” with 1593 citation counts in the centre.

4.2.2.1 Insights gained

- Judging from the table above, it can be seen that the papers that have the most citations are the the older papers. Older papers would have received more citations compared to newer papers so this is expected.
- This correlation is one of the main issues with using raw citation counts. This reason is one of the reasons the *impact factor* for journals was created. When looking to add new journals to the Science Citation Index, you could not depend solely on raw citation counts as small but important journals could be left out. The same can be said here for papers, newer and potentially important papers are left out.
- The median year that these papers were published is 1995, the median citation count is 1179, and the median papers cited is 5.5.
- Looking at the graph above in Figure 4.1 of our highest ranked paper, it can be seen with the papers that cite this paper, the number of citations these papers have on average is quite low. There are several 0's and 1's.
- This shows that even if a paper has a high citation count, the papers that cite this paper are generally not that important. It shows that citation counts are not a good representation of the importance of a paper and it shows that citation counts can be easily manipulated.

4.3 Experiment 2. PageRank Algorithm

The second experiment carried out was to rank the top 10 most important papers based on their PageRank score. As described above in Section 2.6.3.1, PageRank crawls through this graph and assigns each paper a PageRank score. PageRank calculates the importance of a paper by looking at the quantity and quality of other papers that cite it. The underlying assumption here is that papers that have a low number of outgoing citations are more important than those with a high number of outgoing citations. PageRank determines a score based on these metrics.

4.3.1 Methodology

- The data has already been loaded and the index already created.
- Next, the PageRank algorithm was run using the built-in package in Neo4J. The PageRank algorithm calculates the score for the Papers nodes based on their relationship CITED as described above in Section 3.4. The algorithm then writes the PageRank score to each paper under a new property *page_rank.*
- After, a cypher query was written to return the top 10 most important papers based on their *page_rank* score. The title, the year it was published, the number of citations, and the number of papers it cites was also returned.

4.3.2 Results

The results for cypher query can be seen below in Table 4.2. The top 10 papers are displayed in descending order based off of their PageRank score.

Table 4.2. Top 10 papers based on the PageRank algorithm.

Title	Year	Num Citations	Num Cited	PR Score
A Unified Approach to Functional Dependencies and Relations.	1975	15	1	197.22
On the Semantics of the Relational Data Model.	1975	50	1	192.47
Database Abstractions: Aggregation and Generalization.	1977	260	1	189.16
A Characterization of Ten Hidden-Surface Algorithms.	1974	114	2	131.59
An algorithm for hidden line elimination.	1969	10	0	130.72
A Computing Procedure for Quantification Theory.	1960	585	0	121.12
Congestion avoidance and control.	1988	587	6	118.46
A Machine-Oriented Logic Based on the Resolution Principle."	1965	610	3	111.82
A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.	1988	224	0	102.33
Programming semantics for multiprogrammed computations.	1966	125	3	96.84

The highest-ranked PageRank paper ‘A Unified Approach to Functional Dependencies and Relations’ and its relationships with other papers can be seen below in Figure 4.2, Figure 4.3, Figure 4.4, and Figure 4.5. The title, number of citations these papers have received, their PageRank score, and the number of times they cite other papers are displayed in the centre of each node respectively. The top-ranked paper can be seen as the middle node in each figure.

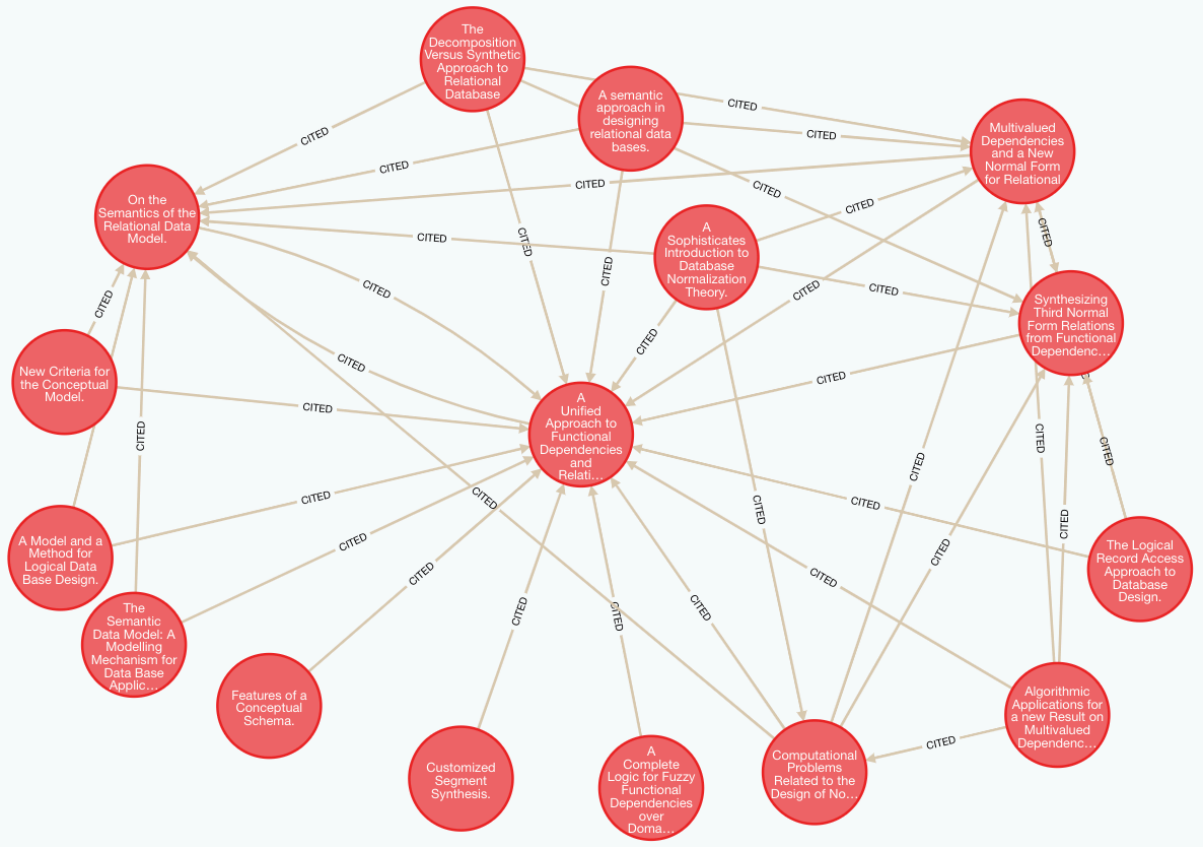

Figure 4.2. Snapshot of the highest ranked PageRank Paper "A Unified Approach to Functional Dependencies and Relations" showing the Paper and its relationships.

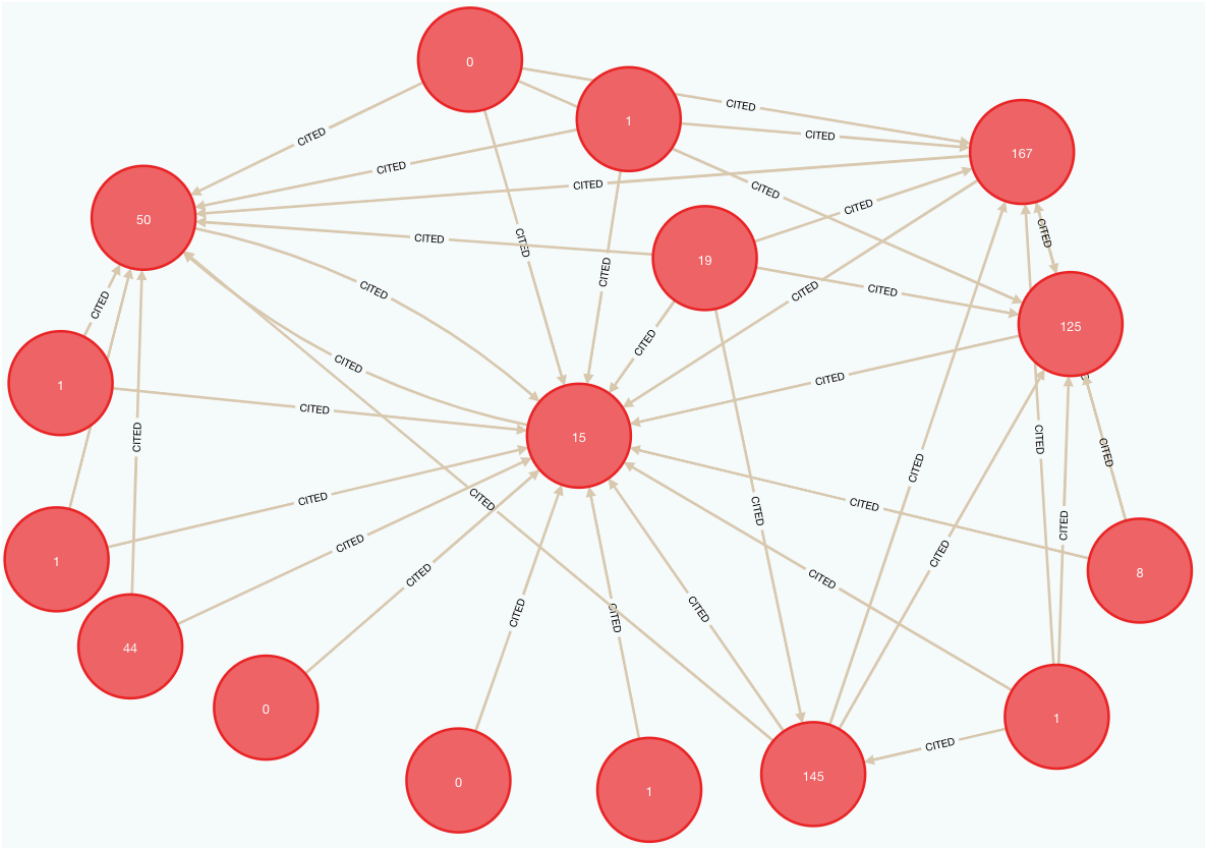

Figure 4.3. Snapshot showing the Papers, their relationships, and their citation counts.

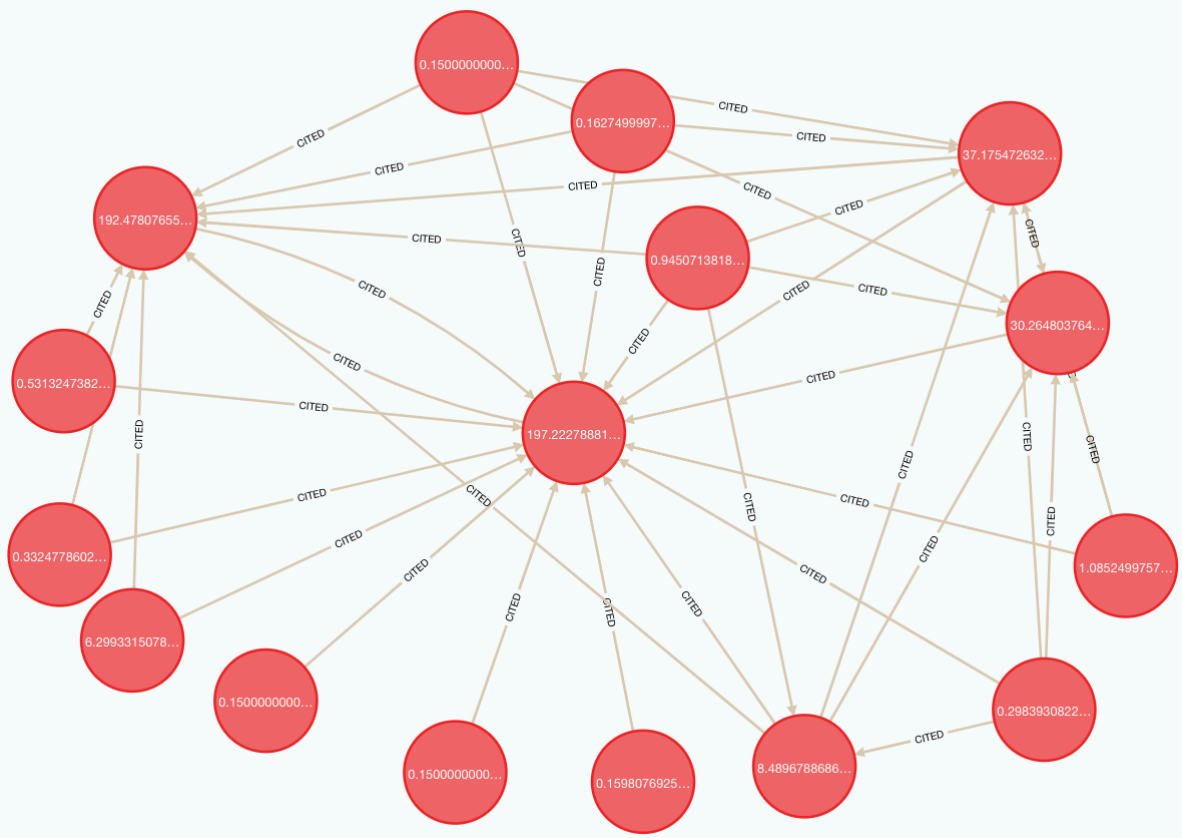

Figure 4.4. Snapshot showing the Papers, their relationships, and their PageRank scores.

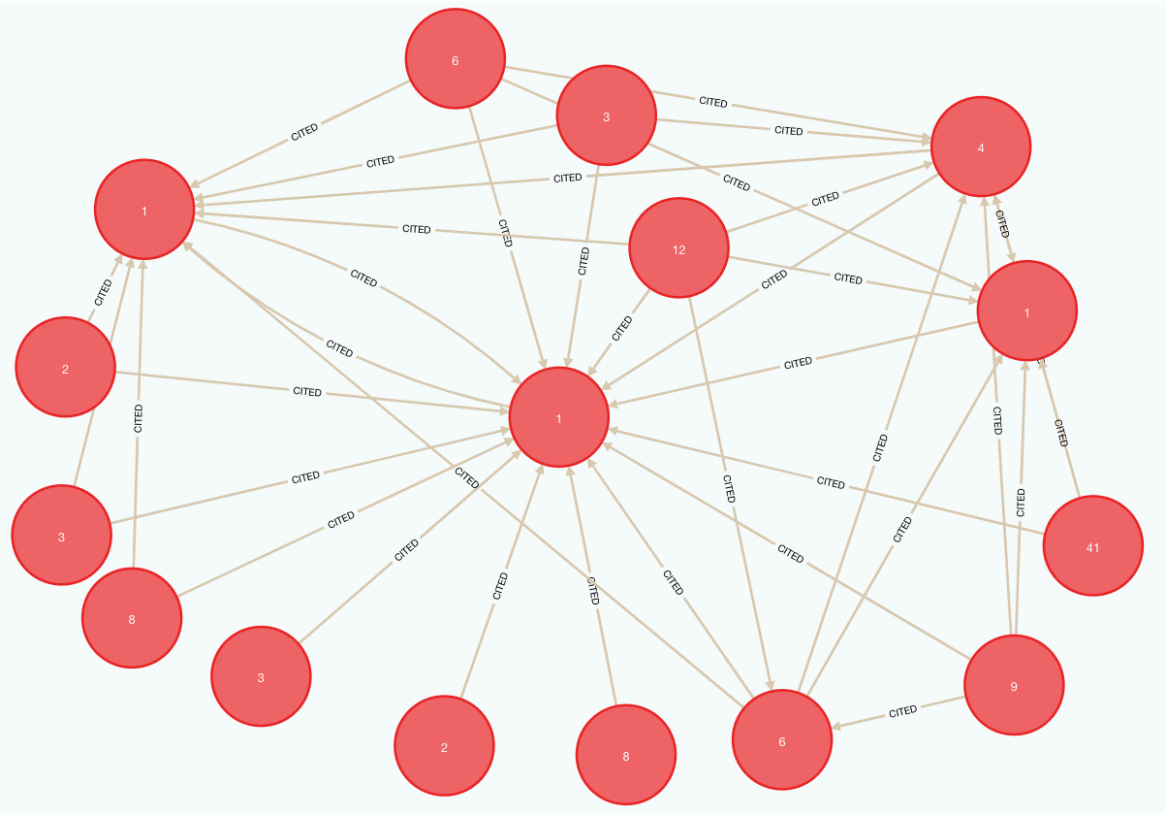

Figure 4.5. Snapshot showing the Papers, their relationships and their number of papers they cite.

4.3.2.1 Insights gained

- Looking at the table above, the papers with the highest score are the older papers. The age of a paper correlates with its PageRank score.
- The highest-ranked paper has 15 citation counts. The number of citation counts a paper has does not seem to be a big factor in determining its PageRank score. As described above in Section 4.3, it is the quality and quantity of the citation counts not just the quantity.
- The median year that these papers were published is 1975, the median citation count is 174.5, the median papers cited is 1, and the media PageRank score is 125.92.
- Papers with higher scores don't tend to cite many other papers.
- Figure 4.3: the PageRank assumption can be seen here. The PageRank score is not heavily dependent on the number of citations it receives but rather the quality of the citations it receives.
- Figure 4.4: The papers that cite this paper have relatively high PageRank scores. This again verifies the assumption and the same result can be seen above in Figure 2.8 in Section 2.6.3.1.
- Figure 4.5: Out of the 15 citations this paper received, five of these papers citing it has over 40 citations each, with three of them having over 120 citations each.
- Figure 4.4 & 4.5: The top two ranked paper only cite each other. Again the same result can be seen above in Figure 2.8 in Section 2.6.3.1.

4.4 Experiment 3. ArticleRank Algorithm

The final experiment carried out was to rank the top 10 most important papers based on their ArticleRank score. ArticleRank is a variation of the PageRank algorithm. As described above in Section 2.6.3.1, the underlying assumption PageRank assumption is that papers that have a low out-degree are more important than those with a higher out-degree. ArticleRank weakens this assumption. Since ArticleRank weakens the PageRank assumption you would expect the results to have higher citation counts. ArticleRank and its algorithm have been discussed in detail above in section 2.6.3.2.

4.4.1 Methodology

- Again, the data has already been loaded and the index defined.
- The methodology for this section is very similar to section 4.3.1 above. The main difference is the use of the built-in ArticleRank algorithm in Neo4J and the writing of these ArticleRank scores to a new property *article_rank.*

4.4.2 Results

The results can be seen below in Table 4.3. The top 10 papers are displayed in descending order based on their ArticleRank score.

Table 4.3. Top 10 papers based on the ArticleRank algorithm.

Title	Year	Num Citations	Num Cited	AR Score
Congestion avoidance and control.	1988	587	6	50.11
Time, Clocks, and the Ordering of Events in a Distributed System.	1978	1071	0	48.91
Highly Dynamic Destination-Sequenced Distance-Vector Routing (DSDV) for mobile computers.	1994	753	3	46.65
Graph-Based Algorithms for Boolean Function Manipulation.	1986	1181	0	44.56
A Machine-Oriented Logic Based on the Resolution Principle.	1965	610	3	42.7
A Computing Procedure for Quantification Theory.	1960	585	0	42.15
Mining Association Rules between Sets of Items in Large Databases.	1993	1374	6	40.6
Fast Algorithms for Mining Association Rules in Large Databases.	1994	1593	7	39.5
Induction of Decision Trees.	1986	956	0	39.46
A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.	1988	224	0	37.36

4.4.2.1 Insights gained

- Looking at the table above, the highest-ranked papers are the older papers. It can be said that the year in which the paper was published correlates to its ArticleRank score.
- The papers with the highest score have some of the highest citation counts. Citation counts seem to influence the ArticleRank score. This is expected as ArticleRank weakens the main assumption of PageRank about the quality and quantity of a papers citations. As a result, higher cited papers will have a higher score.
- Papers with higher scores don't tend to cite many other papers.
- The median year that these papers were published is 1987, the median citation count is 855, the median papers cited is 1.5, and the media ArticleRank score is 42.43.

4.5 Comparison

Between the three tables above, several papers appeared in two tables but there was not one paper that appeared in all three. These comparisons are further discussed below.

4.5.1 Citation Counts & PageRank

There were no papers that were in the top 10 for both Citation Counts and PageRank. This was an interesting finding and shows that PageRank is not heavily dependent on citation counts. This is in line with the underlying assumption of the PageRank algorithm outlined above in Section 4.3

4.5.2 Citation Counts & ArticleRank

As described above in Section 4.4, ArticleRank is a variation of the PageRank algorithm but weakens its underlying assumption. Since the assumption is weakened the number of citations a paper has, has a greater influence on the ArticleRank score. As expected, several papers were present in both tables. In total there were four.

These four papers can be seen below in Table 4.4 along with the positions on each table. The higher ranked papers on the Citation Count table are on the lower end of the ArticleRank table. This is expected based on how the ArticleRank algorithm works.

Table 4.4. Comparison of Citation Counts & ArticleRank

Title	Year	CC Pos	AR Pos	Num Citations	Num Cited	AR Score
Time, Clocks, and the Ordering of Events in a Distributed System.	1978	7	2	1071	0	48.91
Graph-Based Algorithms for Boolean Function Manipulation.	1986	5	4	1181	0	44.56
Mining Association Rules between Sets of Items in Large Databases.	1993	2	7	1374	6	40.6
Fast Algorithms for Mining Association Rules in Large Databases.	1994	1	8	1593	7	39.5

4.5.3 PageRank & ArticleRank

Similar to above there were four papers in the top 10 PageRank and ArticleRank tables. Since ArticleRank is derived from PageRank, similar results it both are to be expected. In total four papers were matched in each table. Their citation counts are in the mid-range. ArticleRank tends to rank the papers matched higher.

The results can be seen below in Table 4.5.

Table 4.5. Comparison of PageRank & ArticleRank

Title	Year	AR Pos	PR Pos	Num Citations	Num Cited	AR	PR
Congestion avoidance and control.	1988	1	7	587	6	50.11	118.46
A Machine-Oriented Logic Based on the Resolution Principle.	1986	5	8	610	3	42.7	111.82
A Computing Procedure for Quantification Theory.	1993	6	6	585	0	42.15	121.12
A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text.	1994	10	9	224	0	37.36	102.33

4.6 Chapter Summary

In this chapter, three approaches to rank papers are presented and compared. These are the raw citation counts, PageRank, and Article Rank. The shortcomings of using the raw citations counts are discussed. It shows that citation counts only really measure the utility or popularity of a paper, rather than its quality. When using PageRank to rank papers, the year the paper was published and the quality of the citations is more important than the number of citations. ArticleRank is somewhere in-between PageRank and the raw citation counts when ranking papers. Since it weakens the PageRank assumption it is depends more on the quantity of the citations. For papers present in all three tables, these paper didn't cite many other papers.

5 Experimentation: Robustness of the Ranking Algorithms to Papers with Cartel like Features

This chapter first looks at the definition of a citation cartel and aims to modify existing entries in the data to have these cartel-like features. It then looks at the experiments carried out to compare the robustness of PageRank and ArticleRank against these papers with citation cartel features. Some of the methodology for these experiments have been carried out, they are described in detail above in Chapter 3.

These experiments explore the second research question which is outlined below:

2. How robust are PageRank and ArticleRank to cartel like behaviors?

5.1 Why is this Important?

As can see from the results above in Chapter 4, Sections 4.3.2 and 4.4.2, PageRank and ArticleRank don't just rank papers based on the number of their citation counts. Since PageRank and ArticleRank are based more on the quality and not just the quantity of the citations, it could be said that they are somewhat robust to the practice of self-citations. PageRank depends less on the number of citations and therefore it can be considered more resilient. However, as discussed above in Chapter 2, Section 2.5, there is a practice that is more damaging and harder to detect than author self-citing. This practice is the practice of citation cartels. Just like self-citations, citation cartels are used to boost an authors *h*-index and a journals *impact factor*. If PageRank and ArticleRank can also be robust to this practice, then they could become a viable evaluation alternative to the commonly used *impact factor* for journals and the *h*-index for authors.

5.2 Definition of a Citation Cartel

Citation cartels have been discussed in detail above in Chapter 2, Section 2.5. It is known that they are the event of authors citing each other disproportionately more than other groups of authors for their mutual benefit. Although there is not an exact formula that constitutes what a citation cartel is, there are some examples. A visual example of a known cartel is shown above in Figure 2.6. This example is what the modifications of these papers will be based on. It is discussed in Section 5.3 below.

5.3 Creating Synthetic Citation Data with 'Cartel' Features by Modifying Existing Entries.

Since there is no exact formula for what a citation cartel, Figure 2.6 above will be used as a guideline. This is a known example of a citation cartel and as a result, this journal got suspended by Thomas-Reuters. Four of the bottom-ranked papers in both metrics were chosen at random from two authors to try and boost their PageRank and ArticleRank scores. Each paper has 0 citations and cites no other papers. The lowest score for a paper with no relationships for both PageRank and ArticleRank in the graph is 0.15. These papers and their scores can be seen in Table 5.1 below, and these papers and their relationships can be seen in below in Neo4J in Figure 5.1.

Table 5.1. Four papers with low PageRank & ArticleRank scores.

Title	Author	Year	Num Citations	AR Score	PR Score
Silent Data Corruption - Myth or reality?	Sarah Michalak & R. Harper	2008	0	0.15	0.15
Randomized Selection on the GPU	Sarah Michalak	2011	0	0.15	0.15
From tele to human : the pragmatic construction of the human in communications systems research.	R. Harper	2009	0	0.15	0.15
Correcting computer-based assessments for guessing.	R. Harper	2003	0	0.15	0.15

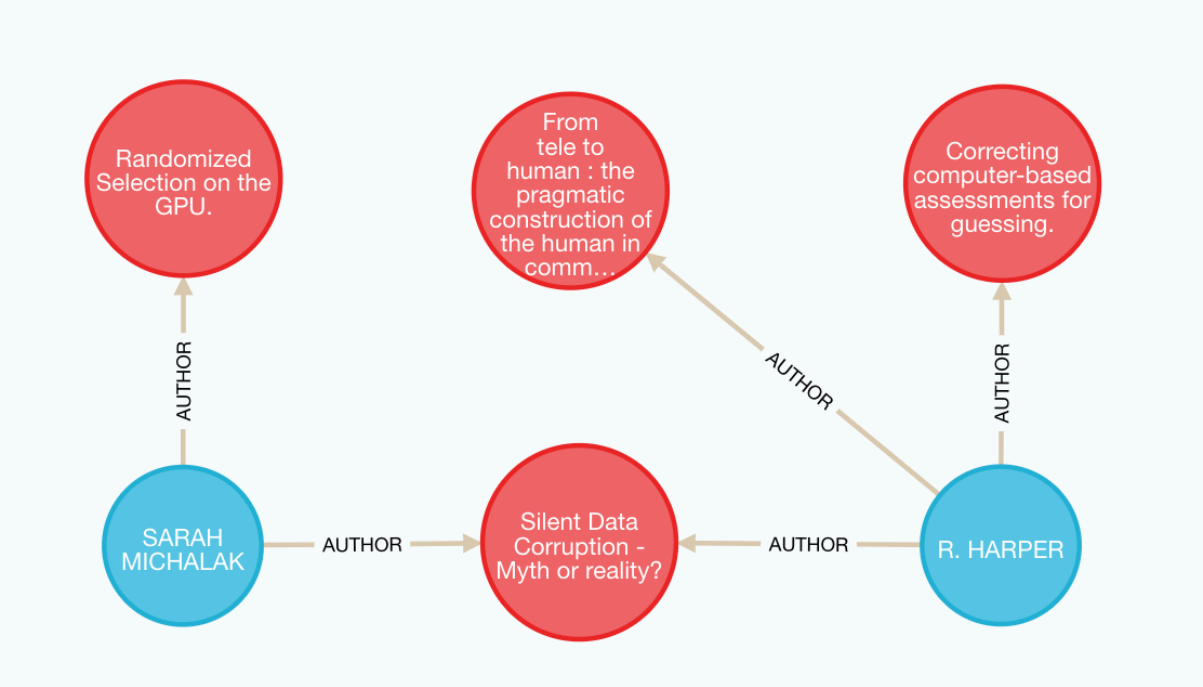

Figure 5.1. Snapshot of the four papers outlined above in Table 5.1 in Neo4J.

5.3.1 Modifying the Data

The goal here is to increase the PageRank and ArticleRank scores of these papers and look at the effect of synthetic data by modifying existing entries creating synthetic citation data. A combination of Microsoft Excel and MacOS TextEdit was used to modify the data. Both had to be used due to a Microsoft Excel limit where it cannot load more than 1,048,576 rows at a time and the Citation table has over 3 million records.

Papers with 0 citations within this dataset will be used to cite these four papers. 252 papers with 0 citations will be taken and added to the *cited* column and these four papers will be added to the *citing* column in the Citation table. Each of these four papers will have 63 citations. 63 citations were chosen because 5 of the top 10 ranked PageRank papers with lower citation counts have on average a citation count of 63. Three of the top 10 PageRank papers also have citation counts less than 63. Looking at the citation cartel above in Figure 2.6, each of these four main papers has around 60 citations each.

Just the Citation table will be modified to create the cartel. This table will be modified twice, once containing the modified synthetic data, and another time with this data engaging in cartel-like behaviour. The Paper id's of these papers are added to the *cited* and *citing* column in the Citation table. These tables and columns are described in detail above in Chapter 3, Section 3.2.

After the 63 citations have been added to each of these papers, the cartel can now be created. To create the cartel, these four papers will, in turn, cite one another, engaging in cartel-like behaviour.

Once the Citation table was modified, the data was reloaded twice into Neo4J following the guideline above in Section 3.5.

5.3.2 Synthetic Citation Data in Neo4J

The synthetic citation loaded into Neo4J can be seen below in Figure 5.2. Here, the four papers from Table 5.1 have been modified to have their 63 citations each. At the moment these papers, just like any other papers have a moderate number of citations. They haven't engaged in citation cartel behaviour yet.

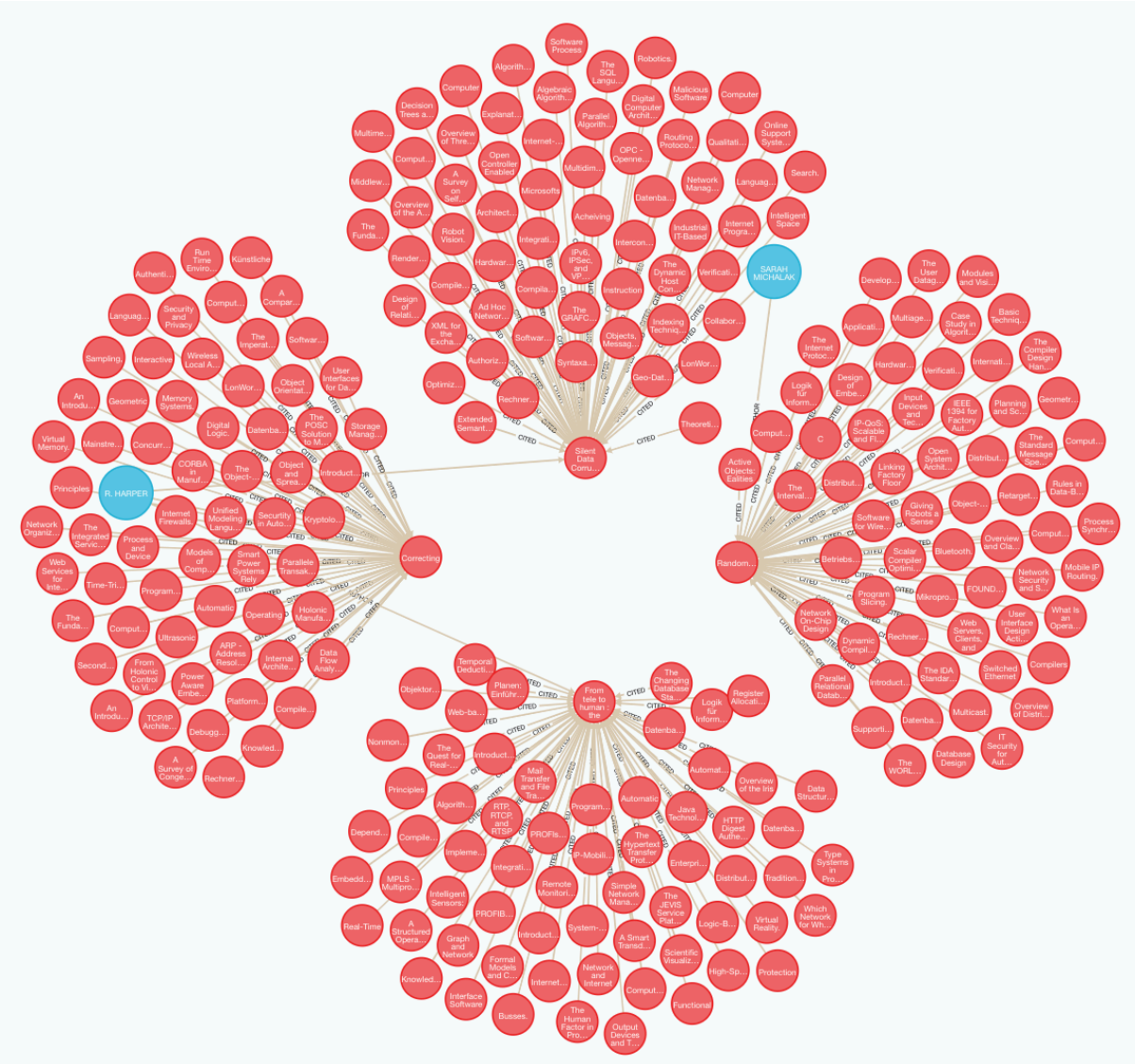

Figure 5.2. Snapshot of the synthetic citation data before it becomes a cartel in Neo4J.

5.3.3 Citation Cartel in Neo4J

The citation cartel loaded into Neo4J can be seen below in Figure 5.3. Here, the four papers from Table 5.1 have been modified to have cartel-like features. Each paper has 63 citations and each of the four papers cites one another.

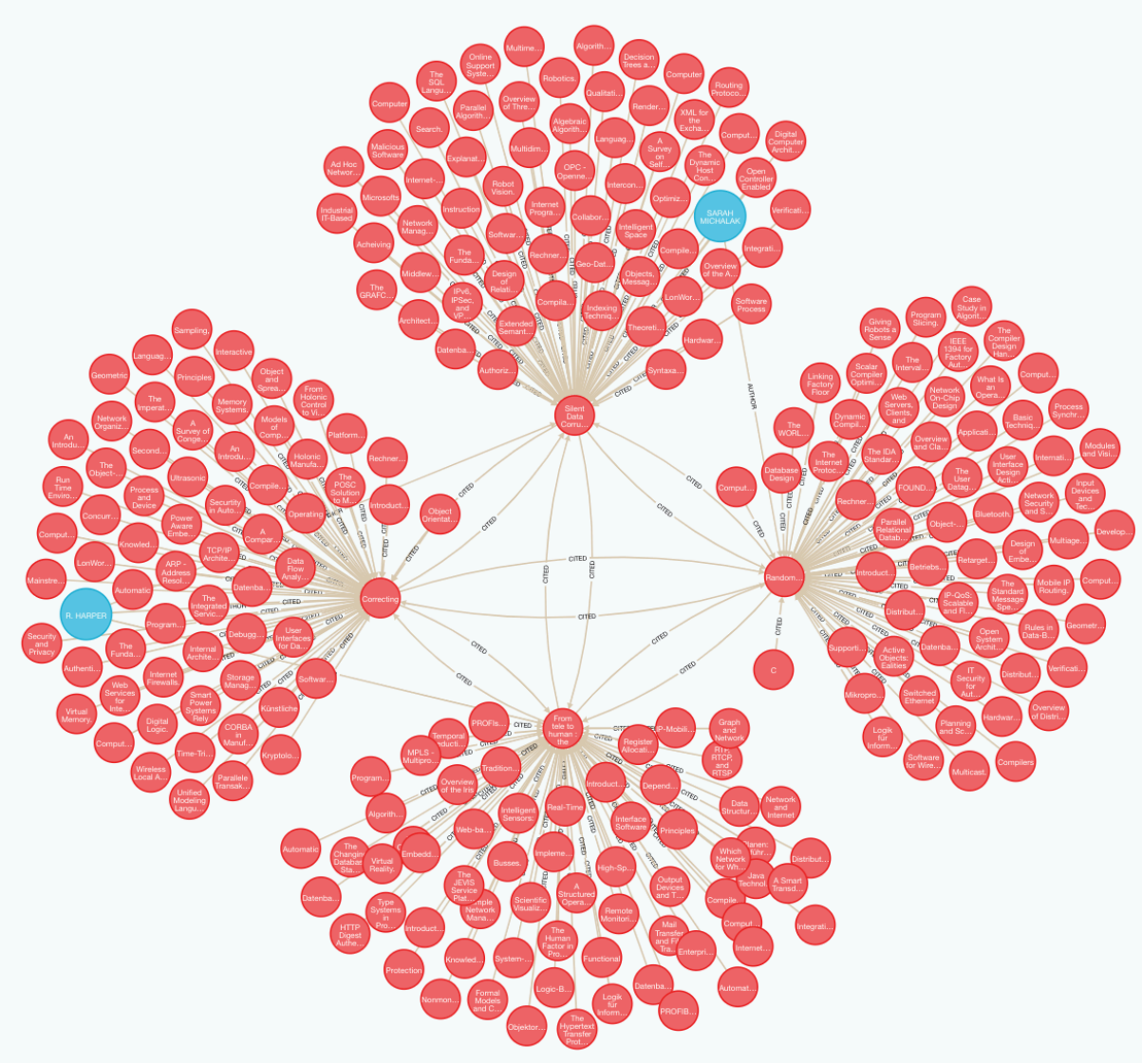

Figure 5.3. Snapshot of the synthetic citation data with citation cartel features in Neo4J.

5.4 Experiment 1. Effect on PageRank

The PageRank experiment outlined above in Chapter 4, Section 4.3 was rerun on both datasets. The methodology remains the same. The only difference here is the newly modified citation data.

5.4.1 Results

The results for this experiment on the four papers can be seen below in Table 5.2 and 5.3. Table 5.2 looks at the results before the data engages in cartel-like behaviour and is just based on increasing the papers citation counts. Table 5.3 looks at the results after the cartel has been created. The scores for the 5 papers directly above and below these papers can be seen in Table 5.4.

Table 5.2. PageRank scores of papers after the synthetic data.

Title	Author	Year	Num Citations	PR Before	PR After	Data Rank
Silent Data Corruption - Myth or reality?	Sarah Michalak & R. Harper	2008	63	0.15	7.47	1820
Randomized Selection on the GPU	Sarah Michalak	2011	63	0.15	7.92	1645
From tele to human : the pragmatic construction of the human in communications systems research.	R. Harper	2009	63	0.15	8.01	1610
Correcting computer-based assessments for guessing.	R. Harper	2003	63	0.15	7.89	1651

5.4.1.1 Insights Gained

- Looking at Table 5.2 above, there was an increase in the PageRank score of each paper from 0.15 to around 7.75.
- The papers ranked a bit higher after the synthetic data was created. Before this, the dataset ranking for each paper would have been around the bottom, around the 2 million mark. Now they are averaging around 1700th in the dataset.
- Looking at the five papers directly above and below these four papers on the dataset, these papers have a median citation count of 50 so it is quite close to having the same correlation as the papers nearby meaning that these citation counts have had a slightly higher impact on the papers PageRank scores compared to the other papers nearby.

Table 5.3. PageRank scores of papers after the citation cartel.

Title	Author	Year	Num Citations	PR Before	PR After	Data Rank
Silent Data Corruption - Myth or reality?	Sarah Michalak & R. Harper	2008	63	7.47	49.88	51
Randomized Selection on the GPU	Sarah Michalak	2011	63	7.92	50.22	49
From tele to human : the pragmatic construction of the human in communications systems research.	R. Harper	2009	63	8.01	50.30	48
Correcting computer-based assessments for guessing.	R. Harper	2003	63	7.89	50.20	50

5.4.1.2 Insights Gained

- Looking at Table 5.3 above, there was an increase in PageRank score of each paper from 7.47 to around 50. It can be seen that the PageRank score is heavily influenced by the citation cartel.
- Compared to the highest PageRank score of 197, the scores are around 25% of the highest score.
- The papers also ranked very highly after the cartel was created. Before this, the dataset ranking for each paper would have been around the bottom, around the 1700th mark. Now they are averaging around 50th in the dataset. This is a massive improvement.
- Looking at five Papers directly above and below these PageRank scores on the dataset, these papers have a median citation count of 300. The citation cartel was able to achieve similar results to papers with over five times as many citations. These papers directly above and below these papers can be seen below in Table 5.4.
- By just letting each of the four papers cite each other, giving them a total of three papers cited, we were able to massively increase their PageRank score and Data Rank score.

Table 5.4. PageRank scores of papers above and below our papers.

Title	Year	Num Citations	PR Score	Data Rank
Fast Algorithms for Mining Association Rules in Large Databases.	1994	1593	55.62	44
A contribution to the development of ALGOL.	1966	38	55.55	45
Ad-hoc On-Demand Distance Vector Routing.	1999	1012	53.76	46
An Overview of the Basic Principles of the Q-Coder Adaptive Binary Arithmetic Coder.	1988	21	52.35	47
R-Trees: A Dynamic Index Structure for Spatial Searching.	1984	843	50.49	52
The Active Badge Location System.	1992	330	48.23	53
Optimal Surface Reconstruction from Planar Contours.	1977	60	47.60	54
A Computational Approach to Edge Detection.	1986	625	45.54	55
Texture and Reflection in Computer Generated Images.	1976	123	45.19	56
Hypertext: An Introduction Survey	1987	271	45.08	57

5.5 Experiment 2. Effect on ArticleRank

The ArticleRank experiment outlined above in Chapter 4, Section 4.4 was rerun. The methodology remains the same. Again, the only difference here is the newly modified citation data.

5.5.1 Results

The results for this experiment on the four papers can be seen below in Table 5.5 and 5.6. Table 5.5 looks at the results before the data engages in cartel-like behaviour and is just based on increasing the papers citation counts. Table 5.6 looks at the results after the cartel has been created. The scores for the 5 papers directly above and below these papers can be seen in Table 5.7.

Table 5.5. ArticleRank scores of Papers after modification of data.

Title	Author	Year	Num Citations	AR Before	AR After	Data Rank
Silent Data Corruption - Myth or reality?	Sarah Michalak & R. Harper	2008	63	0.15	2.89	2677
Randomized Selection on the GPU	Sarah Michalak	2011	63	0.15	3.03	2435
From tele to human : the pragmatic construction of the human in communications systems research.	R. Harper	2009	63	0.15	3.06	2400
Correcting computer-based assessments for guessing.	R. Harper	2003	63	0.15	3.04	2433

5.5.1.1 Insights Gained

- Looking at Table 5.5 above, there was an increase in the ArticleRank score of each paper from 0.15 to around 2.95.
- The papers ranked a bit higher after the synthetic data was created. Before this, the dataset ranking for each paper would have been around the bottom, around the 2 million mark. Now they are averaging around 2500th in the dataset.
- Looking at five papers directly above and below these four papers on the dataset, these papers have a median citation count of 64 so it has nearly the exact same correlation as the papers nearby meaning that the increase in citation counts has had the same increase in the papers ArticleRank score compared to the other papers nearby with the same citation counts.

Table 5.6. ArticleRank scores of Papers after modification of data.

Title	Author	Year	Num Citations	AR Before	AR After	Data Rank
Silent Data Corruption - Myth or reality?	Sarah Michalak & R. Harper	2008	63	2.89	6.383	542
Randomized Selection on the GPU	Sarah Michalak	2011	63	3.03	6.499	517
From tele to human : the pragmatic construction of the human in communications systems research.	R. Harper	2009	63	3.06	6.520	510
Correcting computer-based assessments for guessing.	R. Harper	2003	63	3.04	6.501	515

5.5.1.2 Insights Gained

- Looking at Table 5.6 above, there was an increase in the ArticleRank scores of each paper from 3 to 6.5. The citation cartel has a moderate effect on the ArticleRank scores of each paper.
- Compared to the highest PageRank score of 50, the scores are around 13% of the highest score.
- The papers also ranked highly after the cartel was created. Prior to this, the dataset ranking for each paper would have been around the bottom, around the 2500th mark. Now they are averaging around 500th in the dataset. This is a moderate improvement.
- Looking at five Papers directly above and below these ArticleRank scores on the dataset, these papers have a median citation count of 157. The citation cartel was able to achieve similar results to papers with over 2.5 times as many citations. These papers directly above and below these papers can be seen below in Table 5.7.
- By just letting each of the four papers cite each other, giving them a total of three papers cited, we were able to moderately increase their ArticleRank score and Data Rank score.

Table 5.7. PageRank scores of papers above and below our papers.

Title	Year	Num Citations	AR Score	Data Rank
GloMoSim: A Library for Parallel Simulation of Large-Scale Wireless Networks.	1998	266	6.57	505
GIBIS: A Hypertext Tool for Exploratory Policy Discussion.	1988	172	6.57	506
The World-Wide Web.	1994	126	6.56	507
On Self-Organizing Sequential Search Heuristics	1974	44	6.54	508
The nesC language: A holistic approach to networked embedded systems.	2003	329	6.52	509
Footprint evaluation for volume rendering.	1990	142	6.37	543
The LOCUS Distributed Operating System.	1983	102	6.37	544
Enhanced Hypertext Categorization Using Hyperlinks.	1998	218	6.37	545
An Efficient Context-Free Parsing Algorithm.	1970	190	6.36	546
Inside-Outside Re estimation from Partially Bracketed Corpora.	1992	65	6.36	547

5.6 Chapter Summary

This chapter looks at what a citation cartel is and shows the process of modifying our citation dataset to have citation cartel features. The effects both synthetic citation data and synthetic citation data with citation cartel features have on both the PageRank and ArticleRank ranking algorithms are looked at. By just increasing a papers citation counts, it has a moderate effect on their PageRank and ArticleRank scores. These papers positions in the data set ranking for both metrics are in line with the other papers around them in terms of citation counts. However, if you engage these papers in cartel-like behaviour then there is a big change in both their PageRank and ArticleRank scores, the correlation with citation counts compared to other papers, and there position in the dataset.

6 Conclusion, Future Work, & Limitations

This chapter looks at the conclusions derived from the research work carried out in this research project. It then gives a research summary of the project and its findings, it suggests some avenues for future work, and some limitations encountered.

6.1 Conclusions

The research questions and conclusions are drawn from this research project are outlined below:

The first research question was the following:

1. How do ranking algorithms such as PageRank and ArticleRank compare with each other and raw citation counts in terms of ranking papers on this dataset?

The raw citation counts themselves give a good idea of the most useful papers in this dataset. The top 10 papers in the raw citation counts had a median year of 1995.

PageRank ranked the papers based on the quality of the citations and not just the quantity. It didn't have any papers in common with the top 10 papers from the raw citation counts. The top 10 PageRank papers had a median year of 1975, 20 years less than the citation count median. PageRank is more dependent on the year in which the paper was published rather than the number of citations a paper has. As a result of this, it could be said that PageRank is somewhat robust to the practice of author self-citation.

ArticleRank is somewhere in the middle between the raw citation counts and PageRank. The top 10 ArticleRank papers had a median year of 1987, in the middle of both the PageRank and citation count medians. ArticleRank is more dependent on the quantity of the citations and not just the quality. As a result, ArticleRank is less resilient compared to PageRank on the practice of author self-citation but ranks papers closer to how raw citation counts do. The fact that the raw citation counts featured some papers in the ArticleRank results, shows that citation counts do have some merit for ranking papers.

The second research question further explored the first research question and looked at the robustness of these algorithms to citation cartels, it is the following:

2. How robust are PageRank and ArticleRank to cartel-like behaviors?

Interestingly the roles have reversed for this experiment compared to the first experiment, where ArticleRank was less robust to the practice of author self-citation.

The PageRank scores were less robust to the citation cartels compared to the ArticleRank scores. We were able to achieve PageRank scores similar to papers with over 5 times the number of citations. Since PageRank is less dependent on the number of the citations but rather the quality of the citations, it is easier for a citation cartel to manipulate it. These four papers all cite each other and, in turn, are cited by 63 papers each. Since these four papers cite each other and they all have a high PageRank score thanks to the other citation, their score is further increased. As a result, PageRank is not that robust to papers with cartel-like behaviours.

Since ArticleRank weakens the PageRank assumption, a higher number of citation counts are required to have a bigger influence on the ArticleRank scores. We were able to achieve ArticleRank scores similar to papers with over 2.5 times the number of citations. It's almost twice as robust as PageRank. As a result, it can be said that ArticleRank is moderately robust to papers with cartel-like features.

6.1.2 Research Summary

This research thesis has shown the benefits of using PageRank based algorithms on a citation network. It showed that the algorithms can be used to rank papers in a citation network and showed that they are not dependent on the number of citation counts, something the *impact factor* for journals and h-index for authors are. This research project explored a new angle, the robustness of PageRank based algorithms to the recement emergence of citation cartels, a practice that traditional methods the i*mpact factor* and h-index do not consider. It showed that the ArticleRank algorithm, which is derived from the PageRank algorithm is moderately robust to the practice of citation cartels when ranking papers in a citation network. PageRank, on the other hand, was not very robust. By modifying our data and creating synthetic data with citation cartel behaviours, we were able to get a PageRank score similar to that of papers with 5 times as many citations. Regarding ArticleRank, we were able to get scores similar to papers with less than 2.5 times the number of citations. As a result, ArticleRank could potentially be used to evaluate academic journals and scientific researchers in the future, while at the same time being robust to the practice of citation cartels.

6.2 Limitations

There were several limitations encountered throughout this research project. They are regarding the dataset, the graph database management system Neo4J, and the ranking and evaluation of the papers.

Regarding the dataset, there was not a lot of information available for it. Many columns were not used and had no information available. The *citation_count* column also had an issue, this is outlined above in detail in Chapter 3, Section 3.2. The number of citations in this column was not representative of the number of times a paper was cited in this dataset. As a result, when the top 10 papers were returned based on their citation count, they were wrong. The scores for some of the top 10 papers in the PageRank and ArticleRank algorithms were coming back with 0 citations even though this shouldn't be possible. As a result, a new *num_citations* column was created.

The second limitation encountered was within Neo4J itself. One of the original requirements of this project was to be able to visualise the data and show papers that were closely related based on a ranking score. However, as the project evolved and the requirements changed, Neo4J was still used. Here, we were limited to the built-in algorithms PageRank and ArticleRank, and as a result, other algorithms such as CiteRank, NewRank, the *impact factor*, and *h*-index could not be worked out.

The final limitation encountered was to do with the ranking and evaluation of the papers and not the authors and the journals. The PageRank, ArticleRank, h-index, and impact factor were not calculated and compared for these. As a result, we cannot determine whether or not PageRank and or ArticleRank can replace the commonly used *impact factor* for journals and *h*-index for authors today. This could be an idea implemented in future revisions of this project.

6.3 Future Work

An important and potential very useful output from this research is the application of PageRank and ArticleRank algorithms in the evaluation of academic journals and scientific researchers. The application of these algorithms on the ranking of papers and their robustness to citation cartels has already been discussed above. These algorithms could potentially replace the current widely used evaluation metrics which are the *impact factor* for journals and the *h*-index for authors. To do this, the *h*-index of the researchers and the *impact factor* of the journals in this dataset would have to be calculated. A table of the top 10 researchers and journals would be composed. After this, the PageRank and ArticleRank scores for the top

10 journals and researchers would be created. The two tables would then be compared to see if these algorithms can rank journals and researchers similar to the *impact factor* and the h-index. If they can, and the fact that it has already been shown that ArticleRank is robust to the practice of citation cartels then they could potentially replace these evaluation metrics. If they can't then other PageRank based algorithms could be evaluated. There are many different PageRank based algorithms specifically used on citation networks such as CiteRank, NewRank, and SceasRank.

Bibliography

Aksnes, D. (2003). A macro study of self-citation. Scientometrics, 56(2), pp.235-246.

Aksnes, D., Langfeldt, L. and Wouters, P. (2019). Citations, Citation Indicators, and Research Quality: An Overview of Basic Concepts and Theories. SAGE Open, 9(1), pp.1-17

Barabasi, A. (2016). Network science. Cambridge: Cambridge University Press.

Bartneck, C. and Kokkelmans, S. (2010). Detecting h-index manipulation through self-citation analysis. Scientometrics, 87(1), pp.85-98.

Beel, J., Gipp, B. and Wilde, E. (2010). Academic Search Engine Optimization (ASEO). Journal of Scholarly Publishing, 41(2), pp.176-190.

Belák, V., Hayes, C. (2015). The Risks of Introspection: A Quantitative Analysis of Influence Between Scientific Communities. FLAIRS Conference.

Bollen, J., Rodriquez, M. and Van de Sompel, H. (2006). Journal status. Scientometrics, 69(3), pp.669-687.

Brin, S. and Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30(1-7), pp.107-117.

Brodman, E. 1944. Choosing physiology journals. Bull Med Libr Assn, 32: pp.479–483.

Bornmann, L. and Hans-Dieter, D. (2006). What do citation counts measure? A review of studies on citing behavior. Journal of Documentation, 64(1), pp.45-80.

Bornmann, L. and Hans-Dieter, D. (2008). The state of h index research. Is the h index the ideal way to measure research performance?. EMBO reports, 10(1), pp.2-6.

Carley, S., Porter, A. and Youtie, J. (2012). Toward a more precise definition of self-citation. Scientometrics, 94(2), pp.777-780.

Chen, P., Xie, H., Maslov, S. and Redner, S. (2007). Finding scientific gems with Google's PageRank algorithm. Journal of Informetrics, 1(1), pp.8-15.

Chikate, R.V. and Patil, S.K. (2008). Citation Analysis of Theses in Library and Information Science Submitted to University of Pune: A Pilot Study. Library Philosophy and Practice (e-journal). 222.

Chorus, C. and Waltman, L. (2016). A Large-Scale Analysis of Impact Factor Biased Journal Self-Citations. PLOS ONE, 11(8), p.e0161021.

Cicourel, A. and Franzen, R. (1965). Method and Measurement in Sociology. Journal of Marketing Research, 2(2).

Davis, P. (2012). The Emergence of a Citation Cartel. [online] The Scholarly Kitchen. Available at: https://scholarlykitchen.sspnet.org/2017/03/09/citation-cartel-or-editor-gone-rogue/ [Accessed 12 Aug. 2019].

Davis, P. (2016). Visualizing Citation Cartels. [online] The Scholarly Kitchen. Available at: https://scholarlykitchen.sspnet.org/2016/09/26/visualizing-citation-cartels/ [Accessed 12 Aug. 2019].

Else, H. (2019). Impact factors are still widely used in academic evaluations. [online] Nature.com. Available at: https://www.nature.com/articles/d41586-019-01151-4 [Accessed 3 Aug. 2019].

Enago Academy. (2019). An Introduction to Citation Stacking - Enago Academy. [online] Enago Academy. Available at: https://www.enago.com/academy/what-is-citation-stacking/ [Accessed 12 Aug. 2019].

Enago Academy. (2018). Citation Cartels: The Mafia of Scientific Publishing - Enago Academy. [online] Available at: https://www.enago.com/academy/citation-cartels-the-mafia-of-scientific-publishing/ [Accessed 3 Aug. 2019].

Fister, I. (2017). How to spot a “citation cartel”. [online] Retraction Watch. Available at: https://retractionwatch.com/2017/01/18/spot-citation-cartel/ [Accessed 22 Aug. 2019].

Fister, I. and Perc, M. (2016). Toward the Discovery of Citation Cartels in Citation Networks. Frontiers in Physics, 4.

Fowler, J. and Aksnes, D. (2007). Does self-citation pay?. Scientometrics, 72(3), pp.427-437.

Gann, L. (2018). What is an h-index? How do I find the h-index for a particular author? - LibAnswers. [online] Mdanderson.libanswers.com. Available at: http://mdanderson.libanswers.com/faq/26221 [Accessed 15 Aug. 2019].

Garfield, E. (1955). Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. Science, 122(3159), pp.108-111.

Garfield, E. (1979). Is citation analysis a legitimate evaluation tool?. Scientometrics, 1(4), pp.359-375.

Garfield, E. (2006). The History and Meaning of the Journal Impact Factor. JAMA, 295(1), pp.90–93.

Haley, M. (2017). On the inauspicious incentives of the scholar-level h-index: an economist's take on collusive and coercive citation, Applied Economics Letters, 24:2, 85-89

Katerattanakul, P., Han, B. and Hong, S. (2003). Objective quality ranking of computing journals. Communications of the ACM, 46(10), pp.111-114.

Kickfactory. (2016). The Average Twitter User Now has 707 Followers - Science of Social Sales. [online] Available at: https://kickfactory.com/blog/average-twitter-followers-updated-2016/ [Accessed 18 Aug. 2019].

Korobkin, R. (1999). Ranking Journals: Some Thoughts on Theory and Methodology Symposium. Florida State University Law Review, pp.851-877.

Kreiner G. (2016). The Slavery of the h-index-Measuring the Unmeasurable. Frontiers in human neuroscience, 10(556).

Li, J. and Willett, P. (2009). ArticleRank: a PageRank-based alternative to numbers of citations for analysing citation networks. Aslib Proceedings, 61(6), pp.605-618.

LibGuides. (2019). What Is a Citation?. [online] Subjectguides.esc.edu. Available at: https://subjectguides.esc.edu/researchskillstutorial/citationparts [Accessed 10 Aug. 2019].

Lindsey, D. (1988). USING CITATION COUNTS AS A MEASURE OF QUALITY IN SCIENCE MEASURING WHAT'S MEASURABLE RATHER THAN WHAT'S VALID. Scientometrics, 15(3-4), pp.189-203.

Ma, N., Guan, J. and Zhao, Y. (2008). Bringing PageRank to the citation analysis. Information Processing & Management, 44(2), pp.800-810.

Merriam Webster (2019). Definition of RANKING. [online] Available at: https://www.merriam-webster.com/dictionary/ranking [Accessed 13 Aug. 2019].

Mishra, S., Fegley, B., Diesner, J. and Torvik, V. (2018). Self-citation is the hallmark of productive authors, of any gender. PLOS ONE, 13(9).

Moed, H. (2010). Citation analysis in research evaluation. Dordrecht: Springer.

Neo4J(1). (2019). 5.2. The ArticleRank algorithm - Chapter 5. Centrality algorithms. [online] Available at: https://neo4j.com/docs/graph-algorithms/current/algorithms/article-rank/ [Accessed 16 Aug. 2019].

Neo4J(2). (2019). B.2. Use the Import tool - Appendix B. Tutorials. [online] Available at: https://neo4j.com/docs/operations-manual/current/tutorial/import-tool/ [Accessed 19 Aug. 2019].

Neo4J(3). (2019). 5.1. Indexes - Chapter 5. Schema. [online] Available at: https://neo4j.com/docs/cypher-manual/current/schema/index/ [Accessed 21 Aug. 2019].

Nykamp, D. (2019). Undirected graph definition - Math Insight. [online] Mathinsight.org. Available at: https://mathinsight.org/definition/undirected_graph [Accessed 10 Aug. 2019].

Ognyanova, K. (2019). Network Analysis and Visualization with R and igraph. [online] Kateto.net. Available at: http://www.kateto.net/netscix2016 [Accessed 10 Aug. 2019]

Page, L., Brin, S., Motwani, R. & Winograd, T. (1998), The PageRank citation ranking: Bringing order to the Web, in 'Proceedings of the 7th International World Wide Web Conference' , pp. 161-172 .

Phdontrack.net. (2019). Citation impact. [online] Available at: https://www.phdontrack.net/share-and-publish/citation-impact/#The_h-index [Accessed 14 Aug. 2019].

PLoS Medicine Editors. (2006). The Impact Factor Game. PLoS Medicine, 3(6), p.e291.

Roberts, E. (2006). [online] Web.stanford.edu. Available at: https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf [Accessed 21 Aug. 2019].

Saha, S., Saint, S., & Christakis, D. A. (2003). Impact factor: a valid measure of journal quality?. Journal of the Medical Library Association : JMLA, 91(1), 42–46.

Seglen, P. (1997). Why the impact factor of journals should not be used for evaluating research. BMJ, 314(7079), pp.498-513.

Shema, H. (2012). On Self-Citation. [online] Scientific American Blog Network. Available at: https://blogs.scientificamerican.com/information-culture/on-self-citation/ [Accessed 12 Aug. 2019].

Smith, D. (2009). A 30-Year Citation Analysis of Bibliometric Trends at the Archives of Environmental Health, 1975–2004. Archives of Environmental & Occupational Health, 64(sup1), pp.43-54.

Soulo, T. (2018). Google PageRank is NOT Dead: Why It Still Matters. [online] SEO Blog by Ahrefs. Available at: https://ahrefs.com/blog/google-pagerank/ [Accessed 16 Aug. 2019].

Strickland, J. (2019). Why is the Google algorithm so important?. [online] HowStuffWorks. Available at: https://computer.howstuffworks.com/google-algorithm1.htm [Accessed 8 Aug. 2019].

UIC. (2018). Measuring Your Impact: Impact Factor, Citation Analysis, and other Metrics: Citation Analysis. [online] Available at: https://researchguides.uic.edu/c.php?g=252299&p=1683205 [Accessed 15 Aug. 2019].

Uwe (2017). How to define “self-citation”?. [online] Academia Stack Exchange. Available at: https://academia.stackexchange.com/questions/95736/how-to-define-self-citation [Accessed 12 Aug. 2019].

Van Noorden, R. (2013). Brazilian citation scheme outed. Nature, 500(7464), pp.510-511.

Varin, C., Cattelan, M. and Firth, D. (2015). Statistical modelling of citation exchange between statistics journals. Journal of the Royal Statistical Society: Series A (Statistics in Society), [online] 179(1), pp.1-63. Available at: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/rssa.12124 [Accessed 6 Aug. 2019].

Walker, D., Xie, H., Yan, K. and Maslov, S. (2007). Ranking scientific publications using a model of network traffic. Journal of Statistical Mechanics: Theory and Experiment, 2007(06), pp.P06010-P06010.

Walters, W. (2017). *Citation-Based Journal Rankings: Key Questions, Metrics, and Data Sources - IEEE Journals & Magazine*. [online] Ieeexplore.ieee.org.

Available at: https://ieeexplore.ieee.org/document/8063396 [Accessed 6 Aug. 2019].

Web of Science Group. (1994). The Clarivate Analytics Impact Factor - Web of Science Group. [online] Available at: https://clarivate.com/essays/impact-factor/ [Accessed 15 Aug. 2019].

Wikipedia Contributors. (2018). Citation network. [online] En.wikipedia.org. Available at: https://en.wikipedia.org/wiki/Citation_network [Accessed 10 Aug. 2019]

Thesis & Code

This thesis, the Neo4J commands used, and the r programming code used to clean the data are available at the following GitHub repository:

https://github.com/kevinderrane/Data-Analytics-Thesis